Provable Tensor-Train Format Tensor Completion by Riemannian Optimization

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

November 11, 2021

Abstract

The tensor train (TT) format enjoys appealing advantages in handling structural high-order tensors. The recent decade has witnessed the wide applications of TT-format tensors from diverse disciplines, among which tensor completion has drawn considerable attention. Numerous fast algorithms, including the Riemannian gradient descent (RGrad) algorithm, have been proposed for the TT-format tensor completion. However, the theoretical guarantees of these algorithms are largely missing or sub-optimal, partly due to the complicated and recursive algebraic operations in TT-format decomposition. Moreover, existing results established for the tensors of other formats, for example, Tucker and CP, are inapplicable because the algorithms treating TT-format tensors are substantially different and more involved. In this paper, we provide, to our best knowledge, the first theoretical guarantees of the convergence of RGrad algorithm for TT-format tensor completion, under a nearly optimal sample size condition. The RGrad algorithm converges linearly with a constant contraction rate that is free of tensor condition number without the necessity of re-conditioning. We also propose a novel approach, referred to as the sequential second-order moment method, to attain a warm initialization under a similar sample size requirement. As a byproduct, our result even significantly refines the prior investigation of RGrad algorithm for matrix completion. Numerical experiments confirm our theoretical discovery and showcase the computational speedup gained by the TT-format decomposition.

1 Introduction

An m-th order tensor is an m-dimensional array, i.e, a matrix is a 2nd-order tensor. Tensor completion refers to the task of recovering the whole tensor by observing only a *small* subset of

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

its entries. Of course, this is possible only when the underlying tensor possesses certain structural conditions such that the tensor of interest actually lies in a low-dimensional space. Throughout this paper, we assume that the tensor of interest is *low-rank*. There are diverse applications that drive the research of tensor completion: visual data in-painting (Liu et al., 2012; Li et al., 2017), medical imaging (Gandy et al., 2011; Semerci et al., 2014; Cheng et al., 2017), seismic data analysis (Kreimer et al., 2013; Ely et al., 2013), personalized medicine (Soroushmehr and Najarian, 2016; Pawlowski, 2019), to name a few. It is a natural generalization of the well-explored matrix completion (Candès and Recht, 2009; Candès and Tao, 2010; Cai et al., 2010; Recht, 2011; Davenport and Romberg, 2016; Xia and Yuan, 2021; Chen et al., 2019; Cai et al., 2016; Xia and Koltchinskii, 2016). While seeming, at least conceptually, a straightforward extension of matrix completion, the multi-linear nature of tensors poses unprecedented challenges to tensor completion from multiple fronts. For instance, convex relaxation by matrix nuclear norm is a prevailing approach for low-rank matrix completion, whereas tensor nuclear norm, also a convex function, is generally NP-hard (Hillar and Lim, 2013) to compute. This computation hardness exists in many tensor-related convex functions such as tensor operator norm. As a result, the trick of convex relaxation for tensor completion can be computationally infeasible in some cases. Moreover, another phenomenon making tensor completion fundamentally different from matrix completion is the gap between the information-theoretical sample complexity (the number of observed entries for example) and that required by polynomial-time algorithms. Indeed, it is known that matrix completion is solvable by fast algorithms with a nearly optimal sample complexity (Gross, 2011; Candès and Recht, 2009). However, evidence (Barak and Moitra, 2016) has been found showing that the sample size required by polynomial-time algorithms for tensor completion is significantly larger than the information-theoretical sample complexity. This phenomenon is also observed in other tensor-related problems such as tensor PCA (Zhang and Xia, 2018; Brennan et al., 2018) and tensor clustering (Luo and Zhang, 2020).

Since a multi-linear array can always be re-arranged into a matrix, tensor completion can also be re-formulated as matrix completion. Along this direction are some representable works (Liu et al., 2012; Song et al., 2019; Yuan and Zhang, 2017; Gandy et al., 2011). These methods, while easy to implement (borrowing state-of-art conclusions from matrix completion), suffer from unnecessarily high computational cost because the intrinsic tensor structure is abandoned and the resultant matrix lies in a much higher dimensional space. To overcome these issues, methods for tensor completion better act on the low-rank tensor structure *directly*. Unlike matrices whose ranks are universally defined, there exist multiple widely-accepted definitions of the rank(s) for tensors and, hence, different formats of tensor decomposition. Existing literature of tensor completion, especially those with theoretical investigations, mainly focus on the CP format and Tucker format. The CP decomposition of a tensor seeks a representation by the sum of a minimal number, called

the CP rank, of rank-one tensors. Under the assumption that the underlying tensor is CP decomposable with r *orthogonal* components, Jain and Oh (2014) proposed a fast alternating minimization algorithm for *exactly* completing a $d \times d \times d$ tensor using merely $O(\kappa_0^4 r^5 d^{3/2} \cdot \text{Polylog(d)})$ randomly observed entries, where κ_0 denotes the tensor condition number (see formal definition in Section 2). Later, Barak and Moitra (2016) introduced a semi-definite programming, referred to as the sum-of-squares (SOS) hierarchy, and demonstrated that this polynomial-time method can *approximately* recover the tensor using a similar number of observed entries, even if the CP components are not orthogonal. Then, in the work of Potechin and Steurer (2017), the authors further proved that, with orthogonal CP components, SOS method can *exactly* recover the tensor by observing $O(rd^{3/2} \cdot \text{Polylog(d)})$ randomly sampled entries. Even though SOS is a *provably* polynomial-time method, it usually runs very slowly making it impractical for real-world applications. Meanwhile, the orthogonal decomposability is a rather restrictive assumption. Still restricted to the CP format, a spectral algorithm was proposed by Montanari and Sun (2018) which *approximately* recovers the tensor under a sample size requirement comparable to Potechin and Steurer (2017). The spectral algorithm runs very fast, so it is more scalable to large tensors. More recently, based on many sample splittings, Liu and Moitra (2020) proposed an algorithm, combining both non-convex and convex ideas, to *exactly* recover a tensor with robust linearly independent components. And there are many other representable works (Bi et al., 2018; Ibriga and Sun, 2021; Sun et al., 2017; Cai et al., 2020).

The Tucker rank of a tensor refers to the rank of the matrices obtained by tensor *unfolding*. An m-th order tensor of size $d \times \cdots \times d$ admits m ways of unfolding, so the Tucker rank consists of a collection of m matrix ranks. Tucker decomposition factorizes the tensor into the multilinear product of a *core tensor* and m orthonormal matrices, usually referred to as the singular vectors. The core tensor is usually small-sized but still of order m. Tucker decomposition is *always* attainable via higher-order singular value decomposition (HOSVD). Readers are suggested to the work of Kolda and Bader (2009) for more details. Presuming low Tucker ranks, Huang et al. (2015) designed a polynomial-time algorithm, by tensor unfolding and (matrix) nuclear norm minimization, for tensor completion. Due to the unbalanced unfolding of odd-order tensors, their algorithm requires observing a random sample of $O(rd^{\lceil \frac{m}{2} \rceil} \cdot \text{Polylog(d)})$ entries for completing an m-th order tensor. Zhang (2019) proposed a special sampling scheme showing that $O(rd + r^m)$ entries suffice to exactly recover the unknown tensor. Convex approach by tensor nuclear norm minimization, albeit being computationally infeasible, was studied in the work of Yuan and Zhang (2016) which completes the tensor by using $O(r^{1/2} d^{m/2})$ randomly sampled entries. Later, Xia and Yuan (2019) investigated a polynomial time algorithm for exact tensor completion with a sample complexity $O(\kappa_0^2 r^m d^{m/2} \cdot \text{Polylog(d)}))$ via gradient descent on the Grassmannian manifold,

but the iteration complexity was not provided. More recently, Tong et al. (2021) introduced a scaled gradient descent algorithm and proved the iteration complexity that is free of the condition number. In the work of Xia et al. (2021), a fast higher-order orthogonal iteration algorithm was proposed for noisy tensor completion achieving a statistically optimal rate with a sample complexity $O(r^{m/2}d^{m/2} \cdot \text{Polylog(d)})$.

Despite the popularity of CP format and Tucker format in applications and theories, both these two formats have their pros and cons. The CP decomposition is more friendly for interpreting the principal components of tensors, and the required degree of freedom $O(mrd)$ grows linearly with respect to the order m of a tensor of size $d \times \cdots \times d$ and CP rank r. Unfortunately, the set of tensors of a fixed CP rank is not even closed (Kolda and Bader, 2009), implying that the existence of a best rank r approximation is not even guaranteed. Moreover, it is generally computationally NP-hard to determine the CP rank (Hillar and Lim, 2013) of a given tensor. In contrast, the Tucker rank and decomposition of a tensor can be *always* and *easily* determined by HOSVD, and the tensors with Tucker ranks bounded by a constant constitute a manifold. However, when representing an m-way tensor of Tucker rank $\leq (r, \ldots, r)$, the required number of parameters is $O(r^m + mdr)$ and grows exponentially fast with respect to the order m. As a result, Tucker decomposition consumes a great deal of memory and computation resources for the tensors of very high orders.

The recent decade has witnessed an increasing attraction in a new tensor format (Oseledets, 2009, 2011), referred to as the *tensor train* (TT, see formal definition in Section 2), which enjoys the advantages of both CP and Tucker formats. The TT format was inspired by the matrix product state (MPS, Perez-Garcia et al. 2006), an extremely powerful method to represent the state of a large quantum system (Gross, 2011; Koltchinskii and Xia, 2015). The model complexity of TT formats grows *linearly* with respect to the tensor order. For instance, the parameters needed to store an m-th order tensor of TT rank $\leq (r, \ldots, r)$ is $O(mdr^2)$, saving significant space than the Tucker format. More importantly, TT rank is always and easily attainable like the Tucker rank. Indeed, similarly as the Tucker decomposition, an algorithm based on the sequential SVDs, named as tensor train SVD (TTSVD, Oseledets 2011), is applicable to decide the TT rank, and hence the TT decomposition, of a tensor. TTSVD can also be viewed as a quasi-optimal approximation of a given tensor (Oseledets, 2011). Since the tensors of fixed TT ranks construct a manifold (Holtz et al., 2012), numerous manifold-based algorithms are readily adaptable to the TT formats. Due to the aforementioned advantages of TT formats, there has emerged a vast literature in tensor computation, application and theory exploring the TT-format decomposition. The earliest appearance of TT format or MPS can be traced back to the seminal works in physics, specifically in the simulations of quantum dynamics for very large systems (Vidal, 2003, 2004; Perez-Garcia et al., 2006). The formal definition of TT ranks is later proposed by Oseledets (2011), which has inspired a great

many works for the computation and applications of low TT-rank tensors. Bengua et al. (2017) introduced the nuclear norm for TT-format tensors based on the unfolded matrices, and proposed simultaneous matrix factorization for tensor completion, showcasing its superior performances in the recovery of color images and videos. Inspired by their ideas and motivated by the local smoothness of image data, Ding et al. (2019) further proposed a total variation regularization for the image and video inpainting problems. Based on tensor factorization, Wang et al. (2016) investigated low TT-rank tensor completion by alternating minimization which updates the estimated components sequentially. This is a non-convex approach where the good initialization plays a critical role, and they adopted the spectral initialization by TTSVD. A gradient descent algorithm was proposed in the work of Yuan et al. (2019) for TT-format tensor completion with a random initialization, which often performs poorly when the sample size is small. More recently, Ko et al. (2020) provided a novel but heuristic initialization method for the applications in recovering images and videos that is efficient in numerical experiments. These prior works mostly focus on the methodology and algorithm designs without, or with rather limited, theoretical justification. Towards that end, Imaizumi et al. (2017) introduced a new convex relaxation by the Schatten norm of matrices obtained by the separations (see definition in Section 2) of a TT-format tensor. They investigated a convex method for tensor completion showing that the TT-format tensor is *approximately* recovered by observing $O(rd^{\lceil m/2 \rceil} \cdot \mathrm{Polylog(d)}))$ randomly sampled entries. More recently, Zhou et al. (2020) established the statistically optimal convergence rates of tensor SVD by the fast higher-order orthogonal iteration algorithm in the TT-format.

A particularly important class of efficient algorithms for learning low-rank tensor decomposition is based on the Riemannian optimization. The gist of these algorithms is to view the tensor of interest as a point on the Riemannian manifold (Holtz et al., 2012), for example, the collection of tensors with a bounded Tucker-rank or TT-rank., and then to adapt the Riemannian gradient descent algorithm (RGrad) for minimizing the associated objective function. An incomplete list of representable works of RGrad for matrix and tensor applications includes the works of Steinlechner (2016); Wei et al. (2016b,a); Kressner et al. (2014); Cai et al. (2021b). Similarly, the TT-format tensor completion can be recast to an unconstrained problem over the Riemannian manifold and is numerically solvable via RGrad (Wang et al., 2019). This algorithm is similar to most non-convex algorithms in the sense that it starts from a good initial point on the manifold, iteratively updates the new estimate by descending along the *Riemannian gradient* and retracts it back to the target manifold by TTSVD. Here Riemannian gradient is simply the projection of vanilla gradient onto the tangent space of the manifold. The Riemannian gradient is low-rank and hence can significantly speedup the downstream task of TTSVD. Fortunately, Riemannian gradient, as shown in the works of Lubich et al. (2015); Steinlechner (2016), can be fast computed using QR

decomposition and recursive SVD. All these foregoing properties of computational efficiency make RGrad a perfect choice for TT-format tensor completion, especially for tensors of a very high order. Interestingly, we also observe considerable time saving by TT-format tensor completion compared with the Tucker-format tensor completion, even when the Riemannian gradient descent algorithm is applied for both scenario. See the numerical experiments in Section 6.

1.1 Our Contributions

Despite the rich literature in algorithm designs for TT-format tensor completion and their empirical efficiency, the theoretical understanding, for example, sample size requirement, initialization condition, convergence behaviour and recovery guarantee, of those algorithms is relatively scarce. While abundant results are available for the CP-format and Tucker-format tensor completion, they cannot be directly translated into the case of TT-format for, at least, four reasons. First, the gap of model complexity between TT-format and other formats suggests that the sample size requirement can be different. Secondly, another fundamental condition making tensor completion possible is the so-called *incoherence condition*. It can be straightforwardly defined by the components of tensor decomposition in the CP-format and Tucker-format. Since the decomposition of a TT-format tensor is recursive, a suitable adaptation of the incoherence condition is necessary which causes additional complications in the theoretical understanding. Thirdly, the algorithm design (for instance, RGrad) for TT-format tensor completion is quite special, also due to the recursive nature of TT decomposition, involving the recursive reshapes by tensor separation. It poses extra challenges in analyzing the convergence behaviour of any iterative algorithms for TT-format tensor completion. Finally, obtaining a warm initialization is crucial. The naive spectral initialization suggested by Wang et al. (2016) requires a large sample size observed by empirical simulations. In addition, the second-order moment method (Xia et al., 2021) of spectral initialization for Tucker-format is not directly applicable for the tensors of TT-format.

In this manuscript, we investigate the Riemannian gradient descent algorithm for TT-format tensor completion and provide, to our best knowledge, the first theoretical guarantees of this algorithm under a nearly optimal sample size requirement. More specifically, for an m-th order tensor $\mathcal{T}^*$ of size $d \times \cdots \times d$ in the TT-format with TT rank bounded by r, the RGrad algorithm can exactly recover $\mathcal{T}^*$ by observing $O(\kappa_0^{4m-4} r^{(5m-9)/2} d^{m/2} \cdot \mathrm{Polylog(d)} + \kappa_0^{4m+8} r^{3m-3} d \cdot \mathrm{Polylog(d)})$ randomly sampled entries where κ_0 denotes the condition number of $\mathcal{T}^*$. Our contributions can be summarized into three folds. First of all, by a more sophisticated approach of analysis, we show that the RGrad algorithm for tensor completion converges linearly as long as the initialization is just reasonably good, namely $\|\mathcal{T}_0 - \mathcal{T}^*\|_{\mathrm{F}} = o(1) \cdot \|\mathcal{T}^*\|_{\mathrm{F}}$. This significantly improves existing results (Wei et al., 2016b,a) on the RGrad algorithm for matrix completion ($m = 2$) which require

a stringent initialization condition $\|\mathcal{T}_0 - \mathcal{T}^*\|_{\mathrm{F}} = o(n^{1/2}/d) \cdot \|\mathcal{T}^*\|_{\mathrm{F}}$. Secondly, we prove that the error of the iterates produced by RGrad algorithm contracts at a constant rate that is strictly smaller than 1, under a nearly optimal sample size condition. The contract rate is free of the tensor condition number, improving over prior works (Jain and Oh, 2014; Cai et al., 2021a; Xia and Yuan, 2019) the iteration complexity for completing an ill-conditioned tensor. The attained iteration complexity matches the best one achieved by a recently proposed scaled gradient descent algorithm (Tong et al., 2021). Finally, inspired by the idea in Xia and Yuan (2019), we propose a novel initialization, based on a *sequential* second order moment method, as the input of RGrad algorithm for tensor completion. This approach, unlike the one-pass spectral estimate in the work of Xia and Yuan (2019), involves a recursive estimate of the components of TT decomposition, posing additional challenges in the proof. Our method guarantees a warm initialization with a nearly optimal sample size requirement. Our result fills the void of guaranteed initialization for TT-format tensor completion. Existing initialization approaches (Ko et al., 2020; Wang et al., 2016; Yuan et al., 2019) are either heuristic or miss theoretical justifications.

For readers' convenience, we showcase our theoretical results in Table 1 and compare with representable literature of tensor completion with respect to the tensor formats, sample complexity and iteration complexity. For ease of exposition, Table 1 only focus on third order tensors with $m = 3$ and of size $d \times d \times d$.

1.2 Organization of the Paper

The rest of this manuscript is organized as follows. Section 2 reviews the basics of TT-format tensors, its decomposition and TTSVD. The formulation of tensor completion and the incoherence of TT-format tensors are presented in Section 3. In Section 4, we explain in details the RGrad algorithm for TT-format tensor completion, and the sequential second order moment method for initialization. Section 5 presents the main theorems regarding the convergence of RGrad algorithm and the performance bound of spectral initialization, and establishes the specific sample size requirement. Comprehensive numerical experiments are displayed in Section 6.

1.3 Notations

Throughout this manuscript, we shall use the calligraphic letters $(\mathcal{T}, \mathcal{X})$ to denote tensors of size $d_1 \times \cdots \times d_m$, the capital letters (T, M) to denote the components (see formal definition in Section 2) of TT tensors or matrices and blackboard bold-face letters $(\mathbb{R}, \mathbb{M})$ for sets, the lower case bold-face letters $(\boldsymbol{x}, \boldsymbol{y})$ to denote vectors. The j-th canonical basis vector is denoted by e_j, and we omit the ambient space it lies in whenever the context is clear. For a positive integer d, denote

Data type	Algorithms	Sample complexity	Iteration complexity
CP format with *orthogonal components*	Alternating minimization (Jain and Oh, 2014)	$d^{3/2}r^5\kappa_0^4\log^4(d)$	$\log(\frac{r\kappa_0}{\epsilon})$
CP format	Vanilla gradient descent (Cai et al., 2021a)	$C_{\kappa_0}d^{3/2}r^4\log^4 d$	$\kappa_0^{8/3}\log(\frac{1}{\epsilon})$
Tucker	Grassmannian gradient descent (Xia and Yuan, 2019)	$d^{3/2}r^{7/2}\kappa_0^4\log^{7/2}(d)$	N/A
Tucker	Scaled gradient descent (Tong et al., 2021)	$d^{3/2}r^2\kappa_0(\sqrt{r}\vee\kappa_0^2)\log^3(d)$	$\log(\frac{1}{\epsilon})$
Tensor Train	Riemannian gradient descent (**this paper**)	$d^{3/2}r^3\kappa_0^8\log^5(d)$	$\log(\frac{1}{\epsilon})$

Table 1: Comparisons between TT-format RGrad algorithm and prior algorithms for tensor completion with respect to the tensor formats, sample complexity and iteration complexity. For ease of exposition, we only present the results for 3rd-order tensor $\mathcal{T}^*$ of size $d\times d\times d$ with CP rank r, Tucker rank (r,r,r) and TT rank (r,r), respectively. For sample complexity, we state the number required in terms of d,r only for clarity. For the iteration complexity, we consider the number of iterations needed to output $\widehat{\mathcal{T}}$ such that $\|\widehat{\mathcal{T}}-\mathcal{T}^*\|_{\mathrm{F}}\le\epsilon\underline{\sigma}$. Here κ_0 and $\underline{\sigma}$ are the condition number and minimum singular value of $\mathcal{T}^*$, and may be defined in different ways for different formats. See more details for the case of TT-format in Section 2. We note that, in the work of Cai et al. (2021a), the condition number κ_0 is assumed $\asymp 1$. In fact, their sample complexity and iteration complexity, after checking the proof, indeed depend on κ_0. Moreover, Theorem 1.1 in the work of Jain and Oh (2014) states the iteration complexity as $\log(\sqrt{r}\|\mathcal{T}^*\|_{\mathrm{F}}/(\epsilon\underline{\sigma}))$, where $\|\mathcal{T}^*\|_{\mathrm{F}}/\underline{\sigma}$ has an order $\sqrt{r}\kappa_0$.

$[d]:=\{1,\cdots,d\}$. The standard basis in the tensor space $\mathbb{R}^{d_1\times\cdots\times d_m}$ is denoted by $\{\mathcal{E}_\omega:\omega\in[d_1]\times\cdots\times[d_m]\}$, where $\mathcal{E}_\omega$ is a binary tensor of size $d_1\times\cdots\times d_m$ with only the ω-th entry being 1. We use $\mathcal{T}(\boldsymbol{x}),\mathcal{T}(\omega)$ for $\omega,\boldsymbol{x}\in[d_1]\times\cdots\times[d_m]$ as the entry of $\mathcal{T}$.

Let $\|\cdot\|_{\mathrm{F}}$ denote the Frobenius norm of tensors or matrices. We use $\|\cdot\|_{\ell_p}$ to denote the ℓ_p norm of vectors for $0<p\le\infty$ and $\|\cdot\|_{\ell_0}$ to represent the number of nonzero entries. The notations $C,C_1,\ldots$ are reserved for some positive and absolute constants which do not depend on the related parameters of the problem, but their actual values may change from line to line. Sometimes, these constants may depend on the tensor order m and we shall write them as $C_m,C_{m,1},\cdots$. For positive integers $r_1,\cdots,r_{m-1}$, we denote $\overline{d}:=\max_{i=1}^m d_i,\overline{r}:=\max_{i=1}^{m-1}r_i$ and $\underline{r}:=\min_{i=1}^{m-1}r_i$. Moreover,

define $d^* := d_1 \cdots d_m$ and $r^* := r_1 \cdots r_{m-1}$.

2 Preliminaries of TT-format Tensors

We now briefly review the basic ideas of tensor trains and the various formats frequently used in the TT-format tensor representation. They play a critical role in the motivation of the TT-format RGrad algorithm. Interested readers can refer to the works of Oseledets (2011) and Holtz et al. (2012) for more details and examples.

TT-format. For an m-th order tensor $\mathcal{T} \in \mathbb{R}^{d_1 \times \cdots \times d_m}$, the TT-format rewrites it as a product of m 3-way tensors, called the TT-format *components* and denoted by $T_1, \ldots, T_m$, where the i-th component $T_i \in \mathbb{R}^{r_{i-1} \times d_i \times r_i}$ with the convention $r_0 = r_m = 1$, such that for all $\boldsymbol{x} = (x_1, \ldots, x_m) \in [d_1] \times \ldots \times [d_m]$,

$$\mathcal{T}(\boldsymbol{x}) = \sum_{k_1, \ldots, k_{m-1}} T_1(x_1, k_1) T_2(k_1, x_2, k_2) \cdots T_m(k_{m-1}, x_m),$$

where the auxiliary index variables k_i runs from 1 to r_i. The representation can be simplified. If we view each T_i as a matrix valued function, usually called a *component function*, defined by

$$T_i : [d_i] \to \mathbb{R}^{r_{i-1} \times r_i}, \quad T_i(x_i) = T_i(:, x_i, :).$$

Here we follow the Matlab syntax to denote $T_i(:, x_i, :)$ the sub-matrix of T_i with the second index being fixed at x_i. Now for $\boldsymbol{x} = (x_1, \ldots, x_m)$, the value $\mathcal{T}(\boldsymbol{x})$ can be compactly written in the matrix product form:

$$\mathcal{T}(\boldsymbol{x}) = T_1(x_1) \cdots T_m(x_m), \tag{1}$$

where $T_1(x_1)$ is a row vector and $T_m(x_m)$ is a column vector. To simplify the notations, we often write the TT-format in short as $\mathcal{T} = [T_1, \ldots, T_m]$.

Separation and TT rank. The dimension r_i's of the component T_i are called the TT ranks of $\mathcal{T}$. They are defined by the ranks of matrices obtained from the so-called *separation* of $\mathcal{T}$. More specifically, the i-th separation of $\mathcal{T}$, denoted by $\mathcal{T}^{\langle i \rangle}$, is a matrix of size $(d_1 \cdots d_i) \times (d_{i+1} \cdots d_m)$ and defined by

$$\mathcal{T}^{\langle i \rangle}(x_1 \cdots x_i, x_{i+1} \cdots x_m) = \mathcal{T}(\boldsymbol{x}).$$

Then, r_i is defined by the rank of $\mathcal{T}^{\langle i \rangle}$. The collection $\boldsymbol{r} = (r_1, \cdots, r_{m-1})$ is called the TT-rank of $\mathcal{T}$. For ease of exposition, we denote $\mathsf{rank_{tt}}(\mathcal{T}) = \boldsymbol{r}$. As proved in Theorem 1 in the work of

Holtz et al. (2012), the TT rank is well defined for any tensor. Meanwhile, a TT decomposition $\mathcal{T} = [T_1, \ldots, T_m]$ is always attainable with $T_i \in \mathbb{R}^{r_{i-1} \times d_i \times r_i}$ by the fast TTSVD algorithm, to be introduced in subsequent paragraphs. See Algorithm 1.

Left and right unfoldings. Note that the TT decomposition (1) for a given tensor is not unique. Indeed, one can always multiply $T_i(x_i)$ from right with any invertible matrix A and meanwhile multiply $T_{i+1}(x_{i+1})$ from left with the inverse matrix A^{-1}, rendering a distinct TT decomposition. For identifiability, we impose additional conditions on the TT-format components of $\mathcal{T}$. To that end, we first define the *left* and *right unfolding* of a 3-rd order tensor, reformatting the tensor into matrices. For any $\mathcal{U} \in \mathbb{R}^{p_1 \times p_2 \times p_3}$ be a 3-way tensor, the left and right unfolding linear operators $L : \mathbb{R}^{p_1 \times p_2 \times p_3} \to \mathbb{R}^{(p_1 p_2) \times p_3}$, $R : \mathbb{R}^{p_1 \times p_2 \times p_3} \to \mathbb{R}^{p_1 \times (p_2 p_3)}$ are defined as follows

$$L(\mathcal{U})(jx, k) = \mathcal{U}(j, x, k), \quad \text{and} \quad R(\mathcal{U})(j, xk) = \mathcal{U}(j, x, k).$$

We say the component T_i is *left-orthogonal* if $L(T_i)^\top L(T_i)$ is an identity matrix. Similarly, T_i is said *right-orthogonal* if $R(T_i)^\top R(T_i)$ is identity.

For identifiability of the TT-format components, we assume that $T_1, \cdots, T_{m-1}$ in eq. (1) are *all* left-orthogonal. The resultant TT decomposition is called the *left orthogonal decomposition* of $\mathcal{T}$. Note that no condition is required for the last component T_m. The left orthogonal decomposition of $\mathcal{T}$ can be easily obtained using Algorithm 1 (TTSVD) without the truncation step. By Theorem 1 in the work of Holtz et al. (2012), such a decomposition of a TT-format tensor is unique up to the insertions of orthogonal matrices: for any two left-orthogonal decompositions of $\mathcal{T}$ satisfying $\mathcal{T} = [T_1, \ldots, T_m] = [\widetilde{T}_1, \ldots, \widetilde{T}_m]$, there exist orthogonal matrices $Q_1, \ldots, Q_{m-1}$ with $Q_i \in \mathbb{R}^{r_i \times r_i}$ such that

$$T_1(x_1)Q_1 = \widetilde{T}_1(x_1), Q_{m-1}^\top T_m(x_m) = \widetilde{T}_m(x_m) \text{ and } Q_{i-1}^\top T_i(x_i) Q_i = \widetilde{T}_i(x_i) \text{ for } 2 \le i \le m-1.$$

For any TT rank $\boldsymbol{r} = (r_1, \cdots, r_{m-1})$, define $\mathbb{M}_{\boldsymbol{r}}^{\mathsf{tt}} = \{\mathcal{T} \in \mathbb{R}^{d_1 \times \cdots \times d_m} : \mathsf{rank}_{\mathsf{tt}}(\mathcal{T}) \le \boldsymbol{r}\}$ the set of tensors with TT rank $\le \boldsymbol{r}$. Holtz et al. (2012) proves that $\mathbb{M}_{\boldsymbol{r}}^{\mathsf{tt}}$ is a manifold of dimension $\sum_{i=1}^m r_{i-1} d_i r_i - \sum_{i=1}^{m-1} r_i^2$.

Left and right parts of a TT-format tensor. The low-rank factorization of the separation $\mathcal{T}^{\langle i \rangle}$ is frequently used throughout the algorithm design and technical proof. It is attainable from the TT decomposition (1) of $\mathcal{T}$. To be exact, define the matrix $T^{\le i}$ of size $(d_1 \cdots d_i) \times r_i$, known as the i-th *left part* of $\mathcal{T}$, row-wisely by

$$T^{\le i}(x_1 \cdots x_i, :) = T_1(x_1) T_2(x_2) \cdots T_i(x_i)$$

Similarly, we define the matrix $T^{\geq i}$ of size $r_i \times (d_i \cdots d_m)$, know as the i-th *right part* of $\mathcal{T}$, column-wisely by

$$T^{\geq i}(:, x_i \cdots x_m) = T_i(x_i)T_{i+1}(x_{i+1}) \cdots T_m(x_m)$$

By default, we set $T^{\leq 0} = T^{\geq m+1} = [1]$. With these notations, the i-th separation of $\mathcal{T}$ can be factorized as

$$\mathcal{T}^{\langle i \rangle} = T^{\leq i} T^{\geq i+1}.$$

By definition, there exists a recursive relations between the left parts of $\mathcal{T}$ given by $T^{\leq i} = (T^{\leq i-1} \otimes I_{d_i})L(T_i)$ and, similarly, between the right parts of $\mathcal{T}$ given by $T^{\geq i} = R(T_i)(I_{d_i} \otimes T^{\geq i+1})$. Here $\otimes$ denotes the Kronecker product such that $T^{\leq i-1} \otimes I_{d_i}$ is a matrix of size $(d_1 \cdots d_i) \times (r_i d_i)$. Another useful fact is when the TT decomposition (1) is a left orthogonal decomposition, its left parts are also orthogonal. See Lemma 1. This can be proved by induction and the recursive equation $T^{\leq i} = (T^{\leq i-1} \otimes I)L(T_i)$.

Lemma 1. *Let $\mathcal{T} = [T_1, \ldots, T_m] \in \mathbb{M}_{\boldsymbol{r}}^{\boldsymbol{tt}}$ be a left orthogonal decomposition. Then we have $T^{\leq i\top} T^{\leq i} = I_{r_i}$ for all $i = 1, \cdots, m-1$.*

TTSVD. Given an arbitrary m-th order tensor $\mathcal{T}$, we are often interested in obtaining an approximation of $\mathcal{T}$ by a TT-format tensor with TT rank $\leq \boldsymbol{r} = (r_1, \cdots, r_{m-1})$ for some pre-determined positive integers r_i's. Meanwhile, the desired low TT-rank approximation better be readily representable with a left orthogonal decomposition. This can be obtained by the TTSVD algorithm proposed by Oseledets (2009). For completeness, we here restate the algorithm as in Algorithm 1. At first glance Algorithm 1 may seem slightly different from the original one proposed by Oseledets (2009). But they are indeed equivalent due to the following fact:

$$\mathsf{reshape}(\widehat{T}^{\leq i-1} \mathcal{T}^{\langle i-1 \rangle}, [r_{i-1} d_i, d_{i+1} \cdots d_m]) = (\widehat{T}^{\leq i-1} \otimes I_{d_i})^T \mathcal{T}^{\langle i \rangle}$$

An immediate question is whether the output $\widehat{\mathcal{T}}$ by Algorithm 1 is the best low TT rank-$\boldsymbol{r}$ approximation of $\mathcal{T}$. Unfortunately, this is generally untrue. In fact, based on existing evidence (Hillar and Lim, 2013), computing the best low TT rank-$\boldsymbol{r}$ approximation of an arbitrary tensor is generally NP-hard.

Another technical issue, frequently met in the proof, is that when $\mathcal{T} = \mathcal{T}^* + \mathcal{E}$ where $\mathcal{T}^* \in \mathbb{M}_{\boldsymbol{r}}^{\mathsf{tt}}$ and $\mathcal{E}$ is a *small* but arbitrary perturbation, then how accurately does the output $\widehat{\mathcal{T}}$ from Algorithm 1 approximate the true low-rank $\mathcal{T}^*$? Interestingly, using the spectral representation formula (Xia, 2021), we prove that $\|\widehat{\mathcal{T}} - \mathcal{T}^*\| \leq (1 + o(1)) \cdot \|\mathcal{E}\|_{\mathrm{F}}$ in Lemma 18 with a sharp leading constant 1

Algorithm 1 TT-SVD

Input: an arbitrary $\mathcal{T} \in \mathbb{R}^{d_1 \times \cdots \times d_m}$ and desired TT rank $\boldsymbol{r} = (r_1, \ldots, r_{m-1})$.

Set $\widehat{T}^{\leq 0} = [1]$

for $i = 1, \ldots, m-1$ **do**

Let $L(\widehat{T}_i)$ be the top r_i left singular vectors of the matrix $(\widehat{T}^{\leq i-1} \otimes I_{d_i})^\top \mathcal{T}^{\langle i \rangle}$

Set $\widehat{T}^{\leq i} = (\widehat{T}^{\leq i-1} \otimes I_{d_i}) L(\widehat{T}_i)$

end for

$\widehat{T}_m = (\widehat{T}^{\leq m-1})^\top \mathcal{T}^{\langle m-1 \rangle}$.

Output: $\widehat{\mathcal{T}} = [\widehat{T}_1, \ldots, \widehat{T}_m] \in \mathbb{M}^{\mathsf{tt}}_{\boldsymbol{r}}$.

and being free of the condtion number of $\mathcal{T}^*$, which might be of independent interest. To our best knowledge, ours is the first result of this kind.

Tensor condition number. The signal strength of a TT-format tensor $\mathcal{T}$ with TT-rank $\boldsymbol{r} = (r_1, \cdots, r_{m-1})$ is defined by the smallest singular value among all the matrices obtained from separation, that is,

$$\underline{\sigma}(\mathcal{T}) := \min_{i=1}^{m-1} \sigma_{r_i}(\mathcal{T}^{\langle i \rangle}),$$

where $\sigma_k(\cdot)$ returns the k-th singular value of a matrix. Similarly, the largest singular value of $\mathcal{T}$ is defined by $\overline{\sigma}(\mathcal{T}) := \max_{i=1}^{m-1} \sigma_1(\mathcal{T}^{\langle i \rangle})$. Then the condition number of $\mathcal{T}$ is defined by $\kappa(\mathcal{T}) := \underline{\sigma}(\mathcal{T})^{-1} \overline{\sigma}(\mathcal{T})$. The condition number plays a critical role in the convergence behaviour of many iterative algorithms for tensor completion. See, for instance, Table 1.

3 TT-format Tensor Completion and Incoherence Condition

The goal of tensor completion is to (exactly) recover an underlying tensor by only observing a small subset of its entries. Denote by $\mathcal{T}^*$ the true underlying tensor of size $d_1 \times \cdots \times d_m$ with TT rank $\boldsymbol{r} = (r_1, \cdots, r_{m-1})$ which, for simplicity, is assumed known beforehand and satisfy $r_i \ll d_i$. The observed entries of $\mathcal{T}^*$ are assumed to be uniformly sampled with replacement, a prevailing sampling scheme in the literature (Koltchinskii et al., 2011; Xia and Yuan, 2019; Xia et al., 2021) for its convenience in modelling randomness. More exactly, let $\Omega = \{\omega_i : i = 1, \cdots, n\}$ where ω_i is independently and uniformly sampled from the set of collections $[d_1] \times \cdots \times [d_m]$. By observing only the entries $\{\mathcal{T}^*(\omega_i)\}_{i=1}^n$, we aim to design computationally efficient methods to recover the whole tensor $\mathcal{T}^*$. Intuitively, tensor completion becomes easier when more entries are observed. Oftentimes, the number of observed entries n, known as sample size, is significantly smaller than $d^* := d_1 \cdots d_m$.

Note that the problem can be ill-posed if $\mathcal{T}^*$ has, for instance, only one entry that is non-zero, then it is impossible to recover $\mathcal{T}^*$ unless this non-zero entry is indeed observed. To resolve this issue, it is usually assumed that the information $\mathcal{T}^*$ carries spreads fairly among almost all its entries. One concept characterizing this information spread is by the spikiness of $\mathcal{T}^*$ (Yuan and Zhang, 2016; Cai et al., 2021b) defined by

$$\mathsf{Spiki}(\mathcal{T}^*) := (d^*)^{(1/2)}\|\mathcal{T}^*\|_{\ell_\infty}/\|\mathcal{T}^*\|_{\mathrm{F}}.$$

If the spikiness of $\mathcal{T}^*$ is upper bounded by a constant, it means that most of its entries have comparable magnitudes. Oftentimes, another related concept characterizing the information spread, called incoherence condition, is more frequently used. The exact definition of incoherence condition usually relies on the tensor formats. See the definitions for CP-format tensors in the works of Jain and Oh (2014); Cai et al. (2021a) and for Tucker-format tensors in the works of Yuan and Zhang (2016); Xia et al. (2021). To formalize the definition of incoherence for TT-format tensors, we begin with reviewing the incoherence condition of a matrix.

For a matrix M of size $p_1 \times p_2$ and rank r whose compact SVD is given by $M = U\Sigma V^\top$, the incoherence of M is defined by

$$\mathsf{Incoh}(M) := \max\left\{\sqrt{p_1/r}\cdot \max_{i\in[p_1]}\|U^\top e_i\|_{\ell_2}, \sqrt{p_2/r}\cdot \max_{j\in[p_2]}\|V^\top e_j\|_{\ell_2}\right\}.$$

We say M is incoherent with a constant μ if $\mathsf{Incoh}(M) \le \mu^{1/2}$. Note that incoherence is defined through the singular subspace of M. Therefore, the incoherence of M can be equivalently obtained from any low-rank decomposition $M = U_1\Sigma_1 V_1^\top$ satisfying $U_1U_1^\top = UU^\top$ and $V_1V_1^\top = VV^\top$. This simple property will ease our understanding of incoherence for TT-format tensors. Now the incoherence of the TT-format tensor $\mathcal{T}^* \in \mathbb{M}_{\boldsymbol{r}}^{\mathsf{tt}}$ is defined by

$$\mathsf{Incoh}(\mathcal{T}^*) := \max\left\{\mathsf{Incoh}(\mathcal{T}^{*\langle i\rangle}) : i = 1, 2, \cdots, m-1\right\}.$$

Similarly, we say $\mathcal{T}^*$ is incoherent with a constant μ if $\mathsf{Incoh}(\mathcal{T}^*) \le \mu^{1/2}$. We write the left orthogonal decomposition of $\mathcal{T}^*$ by $\mathcal{T}^* = [T_1^*, \cdots, T_m^*]$. Then for $1 \le i \le m-1$, the i-th separation $\mathcal{T}^{*\langle i\rangle}$ can be written as $\mathcal{T}^{*\langle i\rangle} = T^{*\le i}\Lambda_{i+1}^* V_{i+1}^{*\top}$ where Λ_{i+1}^* is invertible and of size $r_i \times r_i$, and $T^{*\le i\top}T^{*\le i} = V_{i+1}^{*\top}V_{i+1}^* = I_{r_i}$. Then the condition $\mathsf{Incoh}(\mathcal{T}^*) \le \mu^{1/2}$ implies that

$$\max_{i=1}^{m-1}\left\{\max_{k\in[d_1\ldots d_i]}\|T^{*\le i\top}e_k\|_{\ell_2}(d_1\ldots d_i/r_i)^{1/2}, \max_{k\in[d_{i+1}\ldots d_m]}\|V_{i+1}^{*\top}e_k\|_{\ell_2}(d_{i+1}\ldots d_m/r_i)^{1/2}\right\} \le \sqrt{\mu}.$$

The spikiness and incoherence of TT-format tensors are closely related and summarized in the following lemma.

Lemma 2 (Spikiness implies incoherence). *Let $\mathcal{T}^* \in \mathbb{M}_{\boldsymbol{r}}^{tt}$ satisfy $\mathit{Spiki}(\mathcal{T}^*) \le \nu$. Then we have*

$$\mathit{Incoh}(\mathcal{T}^*) \le \nu\kappa_0,$$

where $\mathit{Incoh}(\mathcal{T}^*)$ is the incoherence parameter of $\mathcal{T}^*$ and κ_0 is the condition number of $\mathcal{T}^*$ defined by $\kappa_0 = \overline{\sigma}(\mathcal{T}^*)/\underline{\sigma}(\mathcal{T}^*)$.

Now that assuming the incoherence property of the $\mathcal{T}^*$, the problem of tensor completion is well-posed. To recover $\mathcal{T}^*$, the natural idea is to search for a low TT rank tensor that is consistent with the observed entries $\{\mathcal{T}^*(\omega_i)\}_{i=1}^m$. This can be formalized as a non-convex optimization program of a least squares estimator

$$\min_{\mathcal{T}\in\mathbb{R}^{d_1\times\cdots\times d_m}} f_\Omega(\mathcal{T}) := \frac{1}{2}\langle \mathcal{T}-\mathcal{T}^*, \mathcal{P}_\Omega(\mathcal{T}-\mathcal{T}^*)\rangle = \frac{1}{2}\sum_{\omega\in\Omega}\big(\mathcal{T}(\omega)-\mathcal{T}^*(\omega)\big)^2 \qquad (2)$$
$$\text{such that } \mathsf{rank}_{\mathsf{tt}}(\mathcal{T}) \le \boldsymbol{r},$$

where the operator $\mathcal{P}_\Omega$ is defined by $\mathcal{P}_\Omega(\mathcal{T}) := \sum_{i=1}^n \mathcal{T}(\omega_i)\cdot\mathcal{E}_{\omega_i}$ and the inner product of any two tensors $\mathcal{T}_1, \mathcal{T}_2$ of the same size $d_1\times\cdots\times d_m$ is defined by $\langle\mathcal{T}_1,\mathcal{T}_2\rangle := \sum_{\boldsymbol{x}\in[d_1]\times\cdots\times[d_m]}\mathcal{T}_1(\boldsymbol{x})\mathcal{T}_2(\boldsymbol{x})$. We remark that, due to the sampling with replacement, $f_\Omega(\mathcal{T})$ may not be equal to $\|\mathcal{P}_\Omega(\mathcal{T}-\mathcal{T}^*)\|_{\mathrm{F}}^2/2$. This is slightly different from the setting of sampling without replacement. To simplify the notation, we shall just write $f(\cdot)$ in short for $f_\Omega(\cdot)$.

The optimization program (2) is highly non-convex due to the constraint of TT rank, which is usually solvable only locally. Since the objective function in (2) is smooth, the major concern in algorithm design is usually placed on the effective enforcement of the rank constraint. A particularly popular class of algorithms is based on the projected gradient descent including the singular value projection (SVP, Meka et al. 2009) and iterative hard thresholding (IHT, Goldfarb and Ma 2011). These algorithms consist of mainly two steps: (1) update the estimate along the vanilla gradient descent direction and (2) retract the new estimate to the target tensor/matrix manifold by low-rank approximation, such as HOSVD or TTSVD. Oftentimes, these method suffer from high computational cost because the vanilla gradient is often full-rank and so is the resultant new estimate. Consequently, the retraction in the second step relies on the SVD of a very large and full-rank matrix, at each iteration.

4 Riemannian Gradient Descent and Spectral Initialization

As explained, the vanilla gradient descent algorithm often suffers from high computational burdens. To reduce the computational costs, a modified algorithm, named as the Riemannian gradient descent, was proposed by Edelman et al. (1998); Kressner et al. (2014); Vandereycken (2013), which

explores the local geometry structure of low-rank tensors/matrices. The essential modification is to take advantage of the manifold structure and project the vanilla gradient onto the tangent space leading to the so-called Riemannian gradient. Compared with the vanilla gradient, the Riemannian gradient is also low-rank, rendering considerable speedup in the downstream task of retraction. Inspired by this idea, the Riemannian gradient descent algorithm has been widely applied for matrix/tensor completion (Wei et al., 2016b; Kressner et al., 2014), generalized generalized low-rank tensor estimation (Cai et al., 2021b) and etc. We are in position to explain how RGrad can be adapted to the TT-format tensor completion.

4.1 TT-format Riemannian Gradient Descent

RGrad is an iterative algorithm and, given the current estimate $\mathcal{T}_l \in \mathbb{M}_{\boldsymbol{r}}^{\mathsf{tt}}$ at an iteration, the algorithm updates the estimate by three steps: (1) compute the Riemannian gradient; (2) descent along the Riemannian gradient and (3) retract back to the TT-format tensor manifold. See the pseudocodes in Algorithm 2.

In the first step, the Riemannian gradient (Absil et al., 2009) is obtained via projecting the vanilla gradient $\nabla f(\mathcal{T}_l)$ onto the tangent space, denoted by $\mathbb{T}_l$, of $\mathbb{M}_{\boldsymbol{r}}^{\mathsf{tt}}$ at the point $\mathcal{T}_l$. The tangent space $\mathbb{T}_l$ has an explicit form so that the projection $\mathcal{P}_{\mathbb{T}_l}(\nabla f(\mathcal{T}_l))$ onto $\mathbb{T}_l$ is attainable by fast computations. For cleaner presentation here, we sink the detailed explanation of $\mathcal{P}_{\mathbb{T}_l}(\cdot)$ to Section 4.3. The follow-up gradient descent step is easy, and we demonstrate that a fixed stepsize $\alpha = 0.12n^{-1}d^*$ suffices for convergence where $d^* = d_1 \cdots d_m$ and the constant 0.12 is slightly adjustable in practice. The last step, *retraction*, is of crucial importance. The updated estimate $\mathcal{W}_l = \mathcal{T}_l - \alpha_l \cdot \mathcal{P}_{\mathbb{T}_l}(\nabla f(\mathcal{T}_l))$ after the first two steps is generally not an element in $\mathbb{M}_{\boldsymbol{r}}^{\mathsf{tt}}$, and in fact, the TT-rank of $\mathcal{W}_l$ is larger than the desired TT-rank $\boldsymbol{r}$, violating the rank constraints. The retraction procedure enforces the TT-rank constraint by projecting $\mathcal{W}_l$ back to $\mathbb{M}_{\boldsymbol{r}}^{\mathsf{tt}}$. This can be done by TTSVD, denoted by $\mathsf{SVD}_{\boldsymbol{r}}^{\mathsf{tt}}$ in Algorithm 2.

Due to technical reasons, Algorithm 2 involves an additional procedure, called *trimming*. The trimming operator Trim_ζ which acts entry-wisely on a tensor is defined as follows:

$$\mathsf{Trim}_\zeta(\mathcal{T}) = \begin{cases} \zeta \cdot \mathrm{sign}(\mathcal{T}(\boldsymbol{x})), & \text{if } |\mathcal{T}(\boldsymbol{x})| \geq \zeta, \\ \mathcal{T}(\boldsymbol{x}), & \text{otherwise} \end{cases}$$

Basically, it trims those large entries of $\mathcal{T}$ and thus maintains an acceptable spikiness or incoherence. The trimming step is necessary to guarantee the incoherence property of the new estimate $\mathcal{T}_l$ at each iteration, playing a critical role in proving the local convergence of Algorithm 2. It might be possible to directly prove through a more sophisticated analysis (Cai et al., 2021a) that the incoherence property holds automatically without trimming. We indeed observe in numerical experiments that

the RGrad algorithm without trimming runs nearly the same as the trimmed version, implying the automatic validity of incoherence. However, to directly prove the incoherence without trimming is very challenging. Instead, we resort to trimming for simplicity. This procedure is completely for convenience of the technical proof and alters almost nothing in our numerical experiments.

Algorithm 2 TT-format Riemannian Gradient Descent (RGrad)

Initialization: $\mathcal{T}_0 \in \mathbb{M}_{\boldsymbol{r}}^{\mathsf{tt}}$, spikiness parameter ν and maximum iterations l_{max}

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

$\mathcal{G}_l = \mathcal{P}_\Omega(\mathcal{T}_l - \mathcal{T}^*)$

$\alpha_l = 0.12\frac{d^*}{n}$

$\mathcal{W}_l = \mathcal{T}_l - \alpha_l \cdot \mathcal{P}_{\mathbb{T}_l}\mathcal{G}_l$

Set $\widetilde{\mathcal{W}}_l = \mathsf{Trim}_{\zeta_l}(\mathcal{W}_l)$ with $\zeta_l = \frac{10\|\mathcal{W}_l\|_{\mathrm{F}}}{9\sqrt{d^*}}\nu$

$\mathcal{T}_{l+1} = \mathsf{SVD}_{\boldsymbol{r}}^{\mathsf{tt}}(\widetilde{\mathcal{W}}_l)$

end for

We further remark on the choice of spikiness parameter ν in the algorithm. Note that here ν is not necessary the true spikiness parameter of $\mathcal{T}^*$. It suffices to take ν as a tuning parameter that is slightly larger than $\mathsf{Spiki}(\mathcal{T}^*)$, assuming $\mathsf{Spiki}(\mathcal{T}^*)$ is relatively small. For instance, when one is confident that the true spikiness is bounded by $O(1)$, then this tuning parameter can be set as $\nu \asymp \log \bar{d}$ in the numerical implementation.

Theorem 4 in Section 5 demonstrates that, under mild conditions on the initialization and suitable sample size requirement, Algorithm 2 guarantees that $\|\mathcal{T}_{l+1} - \mathcal{T}^*\|_{\mathrm{F}} \leq \gamma \cdot \|\mathcal{T}_l - \mathcal{T}^*\|_{\mathrm{F}}$ for an absolute constant $\gamma \in (0, 1)$, implying a linear convergence of Algorithm 2. However, it requires the initialization $\mathcal{T}_0$ to be close enough to the underlying tensor $\mathcal{T}^*$. Existing literature (Wang et al., 2016) suggests a naive spectral initialization by the observed tensor $n^{-1}d^*\mathcal{P}_\Omega(\mathcal{T}^*)$. It turns out that a naive spectral initialization performs poorly and only works when the sample size n is exceedingly large. We now propose a novel approach, inspired by Xia et al. (2021) and called sequential second-order moment method, to produce a good initialization requiring only an almost optimal sample size.

4.2 Sequential Spectral Initialization

Recall that the left orthogonal decomposition of $\mathcal{T}^*$ is written as $\mathcal{T}^* = [T_1^*, \cdots, T_m^*]$. Our initialization method yields good estimates for $T_1^*, \ldots, T_m^*$ up to orthogonal rotations. For ease of illustration, we denote by $\widetilde{\mathcal{T}} = n^{-1}d^*\mathcal{P}_\Omega(\mathcal{T}^*)$ the scaled observed tensor. Due to the uniform sampling scheme, it is clear that $\mathbb{E}\widetilde{\mathcal{T}} = \mathcal{T}^*$. Since T_1^* contains the left singular vectors of $\mathcal{T}^{*\langle i\rangle}$, the naive spectral initialization, suggested by Wang et al. (2016), takes the top r_1 left singular vectors

of $\widetilde{\mathcal{T}}^{\langle 1\rangle}$ as an estimation for T_1^*. Unfortunately, this method usually performs poorly because the matrix $\mathcal{T}^{*\langle 1\rangle}$ is of size $d_1 \times (d_2 \cdots d_m)$ whose number of columns can be significantly larger than d_1. While we are only interested in a parameter from an d_1-dimensional space, the quality of spectral estimate from $\widetilde{\mathcal{T}}^{\langle 1\rangle}$ is affected by the larger dimension between d_1 and $d_2 \cdots d_m$.

The second-order spectral moment method is inspired (Xia and Yuan, 2019) by the observation that T_1^* also contains the eigenvectors of $\mathcal{T}^{*\langle 1\rangle}\mathcal{T}^{*\langle 1\rangle\top}$ which is a square matrix of size $d_1 \times d_1$. Therefore, Xia and Yuan (2019) proposed a U-statistic to estimate the square matrix $\mathcal{T}^{*\langle 1\rangle}\mathcal{T}^{*\langle 1\rangle\top}$ directly and then applied the spectral initialization. Unlike the Tucker-format where the components can be computed independent of each other (Section 4 of Xia and Yuan 2019), the computation of components in TT-format decomposition depends recursively on each other, that is, $\widehat{T}_{i+1}$ relies on the availability of $\widehat{T}_i$, see the TTSVD procedure in Algorithm 3. Indeed, $L(T_{i+1}^*)$ is the top r_{i+1} left singular vectors of the matrix $(T^{*\leq i} \otimes I)^\top \mathcal{T}^{*\langle i+1\rangle}$, implying that we shall aim to estimate the eigenvectors of $(T^{*\leq i} \otimes I)^\top \mathcal{T}^{*\langle i+1\rangle}\mathcal{T}^{*\langle i+1\rangle\top}(T^{*\leq i} \otimes I)$. Conceptually, there is no difficulty to generalize the second-order moment method for this purpose. However, on the technical front, the dependence of $\widehat{T}^{\leq i}$ on the original data creates substantial challenges in establishing a sharp spectral perturbation bound. For simplicity, we resort to the trick of sample splitting.

For ease of illustration, some notations are necessary. We randomly split Ω into $2m-1$ disjoint groups $\Omega_1, \ldots, \Omega_{2m-1}$, without loss of generality, of the same size $|\Omega_i| \equiv n_0$ [1]. Basically, for each $i = 1, \cdots, m-1$, we use the sub-samples Ω_{2i-1} and Ω_{2i}, together with the estimates $\widehat{T}_1, \cdots, \widehat{T}_{i-1}$, to estimate T_i^*. Finally, the last sub-sample Ω_{2m-1} is used to estimate T_m^*. Now, for each $i = 1, 2, \cdots, m-1$, define a $(d_1 \cdots d_i) \times (d_1 \cdots d_i)$ matrix

$$N_i = \frac{1}{2n_0^2}\left(\mathcal{P}_{\Omega_{2i-1}}(\mathcal{T}^*)^{\langle i\rangle}\big(\mathcal{P}_{\Omega_{2i}}(\mathcal{T}^*)^{\langle i\rangle}\big)^\top + \mathcal{P}_{\Omega_{2i}}(\mathcal{T}^*)^{\langle i\rangle}\big(\mathcal{P}_{\Omega_{2i-1}}(\mathcal{T}^*)^{\langle i\rangle}\big)^\top\right)$$

Due to the independence between Ω_{2i-1} and Ω_{2i}, one can easily verify $\mathbb{E}N_i = \mathcal{T}^{*\langle i\rangle}\mathcal{T}^{*\langle i\rangle\top}$ so that N_i is an unbiased estimator. Note that the dimension of N_i becomes larger when i increases. It turns out that a direct spectral initialization by N_i still performs poorly unless the sample size is greater than $d_1 \cdots d_i$. Instead, we multiply the left and right hand side of N_i by the estimated left part $\widehat{T}^{\leq i-1} \otimes I_{d_i}$ and its transpose, respectively. The resultant symmetric matrix is then used for estimating the i-th component T_i^*. See the details in Algorithm 3. So the initialization is proceeded in an iterative fashion. Note that an additional truncation procedure is applied to guarantee the incoherence property of the estimates $\widehat{T}_i$'s.

After obtaining the estimates $\widehat{T}_1, \ldots, \widehat{T}_{m-1}$ and the respectively constructed left part $\widehat{T}^{\leq m-1}$,

[1] We note that the *minimal* sample size requirement for ensuring the effectiveness of N_i is distinct for different i's. This is reasonable because the dimensions of N_i's change with respect to i. For ease of exposition, we set all the n_i's to be equal.

the last component $\widehat{T}_m$ is estimated by the minimizer of

$$\min_{T_m} \left\|\widehat{T}^{\leq m-1} T_m - n_0^{-1} d^* \mathcal{P}_{\Omega_{2m-1}}(\mathcal{T}^*)\right\|_{\mathrm{F}},$$

whose solution is explicitly given by $\widehat{T}_m := n_0^{-1} d^* T^{\leq m-1\top}(\mathcal{P}_{\Omega_{2m-1}}(\mathcal{T}^*))^{\langle m-1\rangle}$. Finally, a trimming treatment is implemented on the reconstructed low TT-rank tensor $\widehat{\mathcal{T}}$ to ensure the desired spikiness condition. We again remark that the numbers μ and ν are not necessarily the true spikiness and incoherence parameter of $\mathcal{T}^*$, and they are treated as tuning parameters of Algorithm 3.

Algorithm 3 Initialization by Sequential Second-order Moment Method

Input: Spikiness parameter ν and incoherence parameter μ

Set $\widetilde{T}_1$ be the top r_1 left singular vectors of N_1

Truncation: $\overline{T}_1^i = \frac{\widetilde{T}_1^i}{\|\widetilde{T}_1^i\|_{\ell_2}} \cdot \min\{\|\widetilde{T}_1^i\|_{\ell_2}, (\mu r_1/d_1)^{1/2}\}$

Re-normalization: $\widehat{T}_1 = \overline{T}_1 (\overline{T}_1^\top \overline{T}_1)^{-1/2}$

for $i = 2, \ldots, m-1$ **do**

Set $L(\widetilde{T}_i)$ to be the top r_i left singular vectors of $(\widehat{T}^{\leq i-1} \otimes I)^\top N_i (\widehat{T}^{\leq i-1} \otimes I)$

Truncation: $L(\overline{T}_i)^j = \frac{L(\widetilde{T}_i)^j}{\|L(\widetilde{T}_i)^j\|_{\ell_2}} \cdot \min\{\|L(\widetilde{T}_i)^j\|_{\ell_2}, \sqrt{\mu r_i/d_i}\}$

Re-normalization: $L(\widehat{T}_i) = L(\overline{T}_i)\big(L(\overline{T}_i)^\top L(\overline{T}_i)\big)^{-1/2}$

end for

The last component: $\widehat{T}_m = (\widehat{T}^{\leq m-1})^\top \left(\frac{d^*}{n_0} \mathcal{P}_{\Omega_{2m-1}}(\mathcal{T}^*)\right)^{\langle m-1\rangle}$

Reconstruction: $\widehat{\mathcal{T}} = [\widehat{T}_1, \ldots, \widehat{T}_m]$

Output: $\mathcal{T}_0 = \mathsf{SVD}^{\mathsf{tt}}_{\boldsymbol{r}}(\mathsf{Trim}_\zeta(\widehat{\mathcal{T}}))$ with $\zeta = \frac{10\|\widehat{\mathcal{T}}\|_{\mathrm{F}}}{9\sqrt{d^*}}\nu$

4.3 Computation of Riemannian Gradient

In this section, we provide more details regarding the computation of Riemannian gradient, that is, $\mathcal{P}_{\mathbb{T}}(\mathcal{A})$, where $\mathcal{A}$ is a given tensor of size $d_1 \times \cdot \times d_m$ and $\mathbb{T}$ is the tangent space of the TT-format tensor manifold $\mathbb{M}_{\boldsymbol{r}}^{\mathsf{tt}}$ at the point $\mathcal{T}$ with a left orthogonal decomposition $\mathcal{T} = [T_1, \cdots, T_m]$. By definition, $\mathcal{P}_{\mathbb{T}}(\mathcal{A})$ is the projection of $\mathcal{A}$ onto the tangent space $\mathbb{T}$.

Let us begin with parametrizing the tangent space $\mathbb{T}$. By Theorem 2 of Holtz et al. (2012), the parametrization of $\mathbb{T}$ depends on a gauge sequence. For simplicity and ease of exposition, we take the gauge sequence as a sequence of identity matrices. By Holtz et al. (2012), for any element $\mathcal{X} \in \mathbb{T}$, there exist a sequence of tensors $X_1, \cdots, X_m$ with X_i being a tensor of size $r_{i-1} \times d_i \times r_i$ such that $\mathcal{X}$ can be explicitly written in the form

$$\mathcal{X} = \sum_{i=1}^{m} \delta\mathcal{X}_i \quad \text{where the TT-format tensor } \delta\mathcal{X}_i = [T_1, \cdots, T_{i-1}, X_i, T_{i+1}, \cdots, T_m]. \tag{3}$$

The tensor X_i should satisfy that $L(T_i)^\top L(X_i)$ is an all-zero matrix of size $r_i \times r_i$ for all $i = 1, \cdots, m-1$. There is no constraint on the last component X_m.

Based on the parametrization form (3), for an arbitrary tensor $\mathcal{A}$ of size $d_1 \times \cdots \times d_m$, the Riemannian gradient $\mathcal{P}_{\mathbb{T}}(\mathcal{A})$ must be of the following form

$$\mathcal{P}_{\mathbb{T}}(\mathcal{A}) = \delta\mathcal{A}_1 + \ldots + \delta\mathcal{A}_m \quad \text{where } \delta\mathcal{A}_i = [T_1, \cdots, T_{i-1}, A_i, T_{i+1}, \cdots, T_m]$$

for some tensor A_i of size $r_{i-1} \times d_i \times r_i$ satisfying $L(T_i)^\top L(A_i)$ is an all-zero matrix for all $i = 1, \cdots, m-1$. This suggests that, for all $i \neq j$, the tensor $\delta\mathcal{A}_i$ is orthogonal to $\delta\mathcal{A}_j$, proved in the following lemma.

Lemma 3. *Let $\mathcal{T} = [T_1, \ldots, T_m] \in \mathbb{M}^{\boldsymbol{tt}}_{\boldsymbol{r}}$ be a left orthogonal decomposition of $\mathcal{T}$. For an arbitrary $\mathcal{A}$ of size $d_1 \times \cdots \times d_m$, the components $\delta\mathcal{A}_i$'s of $\mathcal{P}_{\mathbb{T}}(\mathcal{A})$ satisfy $\langle \delta\mathcal{A}_i, \delta\mathcal{A}_j \rangle = 0$ for all $1 \leq i \neq j \leq m$.*

Proof. Without loss of generality, assume $i < j$. Then we have

$$\langle \delta\mathcal{A}_i, \delta\mathcal{A}_j \rangle = \langle (\delta\mathcal{A}_i)^{\langle i \rangle}, (\delta\mathcal{A}_j)^{\langle i \rangle} \rangle = \langle (T^{\leq i-1} \otimes I) L(A_i) T^{\geq i+1}, (T^{\leq i-1} \otimes I) L(T_i) \widetilde{A}_j^{\geq i+1} \rangle = 0,$$

where the last equality is due to the facts that $(T^{\leq i-1} \otimes I)$ has orthonormal columns and $L(A_i)^\top L(T_i)$ is an all-zero matrix. Here, for simplicity, we denote $(\delta\mathcal{A}_j)^{\langle i \rangle} = T^{\leq i} \widetilde{A}_j^{\geq i+1}$ for some matrix $\widetilde{A}_j^{\geq i+1}$. □

Due to this orthogonality property of Lemma 3, for all $i \in [m-1]$, determining $\delta\mathcal{A}_i$ is equivalent to solving the following individual optimization problem

$$\min_{A_i} \|\mathcal{A} - \delta\mathcal{A}_i\|_{\mathrm{F}}, \quad \text{s.t. } \delta\mathcal{A}_i = [T_1, \ldots, A_i, \ldots, T_m] \;\text{ and }\; L(A_i)^\top L(T_i) = 0 \tag{4}$$

For the last component, it suffices to solve

$$\min_{A_m} \|\mathcal{A} - \delta\mathcal{A}_m\|_{\mathrm{F}}, \quad \text{s.t. } \delta\mathcal{A}_m = [T_1, \ldots, T_{m-1}, A_m] \tag{5}$$

Finally, based on (4) and (5), we can obtain a closed-form solution of $A_i, i \in [m]$, represented in the form of $L(A_i)$, by

$$L(A_i) = \begin{cases} (I - L(T_i)L(T_i)^\top)(T^{\leq i-1} \otimes I)^\top \mathcal{A}^{\langle i \rangle} (T^{\geq i+1})^\top (T^{\geq i+1}(T^{\geq i+1})^\top)^{-1}, & i \in [m-1] \\ (T^{\leq m-1} \otimes I)^\top \mathcal{A}^{\langle m \rangle}, & \text{if } \; i = m. \end{cases} \tag{6}$$

This yields the way of computing the Riemannian gradient $\mathcal{P}_{\mathbb{T}}(\mathcal{A})$.

5 Exact Recovery and Convergence Analysis

In this section, we prove the validity of the sequential second-order spectral initialization Algorithm 3 and linear convergence behaviour of the RGrad Algorithm 2 so that the underlying tensor can be exactly recovered with an almost optimal sample size of observed entries. For ease of exposition, we assume $r_1, \ldots, r_{m-1}$ are of the same order and denote $\overline{r}$ an upper bound for them. The sample size requirement in more general cases of r_i's can be found from Theorem 6 and Theorem 11 in the Appendix. Recall the notations $\overline{d} = \max_j d_j$ and $d^* = d_1 \cdots d_m$. The smallest singular value and condition number of $\mathcal{T}^*$ are denoted by $\underline{\sigma}$ and κ_0, respectively. Our main theorem of exact recovery is described as follows.

Theorem 1. *Suppose that $\mathcal{T}^*$ is of size $d_1 \times \cdots \times d_m$ with a TT-rank $\boldsymbol{r} = (r_1, \ldots, r_{m-1})$ tensor whose spikiness is bounded by $\mathsf{Spiki}(\mathcal{T}^*) \le \nu$. Let $\mathcal{T}_0$ be initialized by the sequential second-order method as Algorithm 3 and $\{\mathcal{T}_l\}_{l=1}^{l_{\mathsf{max}}}$ be the iterates produced by Algorithm 2 where l_{max} is the maximum number of iterations, and the stepsize $\alpha = 0.12 n^{-1} d^*$. There exist absolute constants $C_{m,1}, C_{m,2} > 0$ depending only on m such that if the sample size n satisfies*

$$n \ge C_{m,1} \cdot \big(\kappa_0^{4m-4} \nu^{m+3} (d^*)^{1/2} \overline{r}^{(5m-9)/2} \log^{m+2} \overline{d} + \kappa_0^{4m+8} \nu^{2m+2} \overline{d} \overline{r}^{3m-3} \log^{2m+4} \overline{d}\big),$$

then with probability at least $1 - (2m+4)\overline{d}^{-m}$, for any $\varepsilon \in (0,1)$, after $l_{\mathsf{max}} = \lceil C_{m,2} \cdot \log(\underline{\sigma}/\varepsilon) \rceil$ iterations, the final output achieves error $\|\mathcal{T}_{l_{\mathsf{max}}} - \mathcal{T}^*\|_{\mathrm{F}} \le \varepsilon$.

If the tensor dimension is balanced $d_j \asymp d$ for all j, rank $\overline{r} = O(1)$ and $\mathcal{T}^*$ is will conditioned in that $\kappa_0, \nu = O(1)$, the sample size requirement of Theorem 1 simplifies to $O_m(d^{m/2} \cdot \mathsf{Polylog(d)})$. This improves the existing result (Imaizumi et al., 2017) based on matricization and matrix nuclear norm penalization. Moreover, for the case $m = 3$, Barak and Moitra (2016) conjectures that, based on the reduction to Boolean satisfiability problem, $O(d^{3/2})$ is a lower bound for the sample size such that polynomial-time algorithm exists for exact tensor completion. Therefore, the sample size requirement of Theorem 1 is likely optimal, at least for $m = 3$, up the logarithmic factors if only polynomial-time algorithms are sought.

We note that, in the sample size requirement, the exponent on rank r still depends on the order of tensor, due to technical difficulties. The recursive nature of computing a TT decomposition substantially complicates the theoretical analysis, involving repeated appearances of the incoherence parameters and TT rank. Interestingly, the required number of iterations l_{max} is free of the condition number, which often appears in decomposition-based algorithms (Cai et al., 2021a; Han et al., 2020), except the recently proposed scaled gradient descent algorithm (Tong et al., 2021).

The proof of Theorem 1 relies on two essential parts: the local convergence of warm-initialized Algorithm 2 and the validity of initialization by Algorithm 3, which are separately dealt with in

the subsequent sections.

5.1 Local Convergence of Riemannian Gradient Descent

For the local convergence of an iterative algorithm, we study its behaviour in a small neighbourhood of the global minimizer, namely $\mathcal{T}^*$ for our case. Lemma 4 dictates that Algorithm 2 converges linearly to $\mathcal{T}^*$ as long as the initialization, be it obtained from our proposed Algorithm 3 or not, $\mathcal{T}_0$ is sufficiently close to $\mathcal{T}^*$ and a sample size condition holds. The proof of Lemma 4 is postponed to the Appendix.

Lemma 4. *Suppose the conditions on $\mathcal{T}^*$ from Theorem 1 hold, and the initialization $\mathcal{T}_0$ satisfies*

$$\|\mathcal{T}_0 - \mathcal{T}^*\|_{\mathrm{F}} \le \frac{\underline{\sigma}}{Cm\kappa_0 \overline{r}^{1/2}} \quad \text{and} \quad \mathit{Incoh}(\mathcal{T}_0) \le 2\kappa_0^2 \nu$$

for a sufficiently large but absolute constant $C > 0$. There exists an absolute constant $C_m > 0$ depending only on m such that if the sample size n satisfies

$$n \ge C_m \cdot \Big(\kappa_0^{2m+4} \nu^{m+1} (d^*)^{1/2} \overline{r}^{\frac{m+3}{2} \vee (m-1/2)} \log^{m+2} \overline{d} + \kappa_0^{4m+8} \nu^{2m+2} \overline{d}\, \overline{r}^{(m+3)\vee(2m-1)} \log^{2m+4} \overline{d} \Big),$$

then with probability at least $1 - (m+4)\overline{d}^{-m}$, the sequence $\{\mathcal{T}_l\}_{l=1}^{\infty}$ generated by Algorithm 2 with a constant step size $\alpha = 0.12 n^{-1} d^*$ satisfy

$$\|\mathcal{T}_l - \mathcal{T}^*\|_{\mathrm{F}}^2 \le 0.975 \cdot \|\mathcal{T}_{l-1} - \mathcal{T}^*\|_{\mathrm{F}}^2.$$

for all $l = 1, 2, \cdots$.

Lemma 4 dictates that the error contracts at a constant rate which is strictly smaller than 1. One interesting fact of our results is that the contraction rate is independent of the condition number κ_0, which is the reason that l_{max} in Theorem 1 is free of κ_0. It suggests that the Riemannian gradient descent algorithm converges fast even for very ill-conditioned tensors, improving the existing results (Jain and Oh, 2014) and (Cai et al., 2021a). Recently, Tong et al. (2021) introduced a scaled gradient descent algorithm to remove the dependence on the condition number, where the rescaling procedure plays a role of re-conditioning. Interestingly, Riemannian gradient descent algorithm automatically achieves this performance without the need to re-scaling. This is perhaps an intrinsic advantage of the manifold-type algorithms (see also Cai et al. 2021b). We note that the contraction rate 0.975 is improvable but no further efforts are made for that purpose.

In the case $\overline{r}, \kappa_0, \nu = O(1)$, the sample size required by Lemma 4 is$O_m(d^{m/2} \cdot \mathsf{Polylog(d)})$, which matches that required by Theorem 1. However, if $\overline{r}$ grows with $\overline{d}$, the exponent on $\overline{r}$ is slightly better than that of Theorem 1.

If $m = 2$, a TT-format tensor is merely a matrix and the left orthogonal decomposition reduces a decomposition with an orthogonal matrix on left hand side. The convergence analysis of Riemannian gradient descent algorithm for matrix completion was investigated by Wei et al. (2016b), showing that the algorithm converges linearly if the initialization is so good that $\|\mathcal{T}_0 - \mathcal{T}^*\|_{\mathrm{F}} = o\big((n/d^*)^{1/2}\big) \cdot \underline{\sigma}$. This is very restrictive since, oftentimes, the desired sample size n is only of order $d^{*1/2}$. It is speculated that this gap is due to technical reasons, deriving from the special form of Riemannian gradient where the existing strategy (Candès and Recht, 2009) simply fails, but the issue has never been really resolved. Taking a more sophisticated approach, by integrating prior tools from Xia and Yuan (2019); Cai et al. (2021b) and Tong et al. (2021), we finally provide an affirmative answer that the initialization condition can indeed be relaxed to the typical ones required by other rivalry algorithms. Lemma 4 suggests that, if $m, \overline{r}, \kappa_0$ are bounded, we only require the initialization satisfies $\|\mathcal{T}_0 - \mathcal{T}^*\| \le c \cdot \underline{\sigma}$ for a small enough but absolute constant $c > 0$. This holds for higher-order TT-format tensors and is not restricted to matrices.

5.2 Initialization by Sequential Second-Order Spectral Method

As stated in Lemma 4, the warm initialization is of crucial importance to ensure the convergence of Algorithm 2. Now we show that our proposed sequential second-order spectral initialization, stated in Algorithm 3, can indeed, with high probability, deliver an estimate close to $\mathcal{T}^*$ under a nearly optimal sample size condition. While our algorithm is inspired by Xia and Yuan (2019), the theoretical investigate turns out to be more challenging due to the recursive nature of Algorithm 3. Compared with Xia and Yuan (2019), a major challenge in our proof is to establish the concentration of $(\widehat{T}^{\le i-1} \otimes I)^\top N_i (\widehat{T}^{\le i-1} \otimes I)$ rather than the concentration of N_i itself. Actually, the concentration of N_i is poor because its dimension can be quite large. By multiplying both sides with the incoherent matrix $\widehat{T}^{\le i-1}$, the resultant smaller matrix enjoys much better concentration. The proof of Lemma 5 is relegated to the Appendix.

Lemma 5. *Suppose the conditions of $\mathcal{T}^*$ from Theorem 1 hold. For any absolute constant $C > 0$, there exists an absolute constant $C_m > 0$ depending only on m such that if*

$$n \ge C_m \nu^{m+3} \kappa_0^{4m-4} \big((d^*)^{1/2} \overline{r}^{(5m-9)/2} + \overline{d}\,\overline{r}^{3m-4}\big) \log^2 \overline{d},$$

then with probability at least $1 - m\overline{d}^{-m}$, the output of Algorithm 3 satisfies

$$\|\mathcal{T}_0 - \mathcal{T}^*\|_{\mathrm{F}} \le \frac{\underline{\sigma}}{C m \kappa_0^2 \overline{r}^{1/2}} \quad \text{and} \quad \mathit{Incoh}(\mathcal{T}_0) \le 2\kappa_0^2 \nu.$$

Based on Lemma 5, the output of Algorithm 3 indeed satisfies the initialization condition required by Lemma 4. If the tensor dimension is balanced $d_j \asymp d$ and $\nu, \kappa_0, \overline{r} = O(1)$, the sample

size requirement for warm initialization matches, with a slightly better dependence on $\log \overline{d}$, that required by the algorithmic convergence of Lemma 4. Thus our initialization method is valid requiring a nearly optimal sample size.

Now Theorem 1 can be readily proved by combining Lemma 4 and Lemma 5.

Proof. Under the sample size condition of Theorem 1, both Lemma 4 and Lemma 5 hold. Therefore, from the contraction property in Lemma 4, we get

$$\|\mathcal{T}_{l_{\mathsf{max}}} - \mathcal{T}^*\|_{\mathrm{F}} \leq (0.975)^{l_{\mathsf{max}}} \cdot \|\mathcal{T}_0 - \mathcal{T}^*\|_{\mathrm{F}} \leq (0.975)^{l_{\mathsf{max}}} \cdot \underline{\sigma},$$

where the last inequality is guaranteed by Lemma 5 and the above bound holds with probability at least $1 - (2m = 4)\overline{d}^{-m}$. Then, for any $\varepsilon > 0$, there exists a constant $C_2 > 0$ such that if $l_{\mathsf{max}} = \lceil C_2 \log(\underline{\sigma}/\varepsilon) \rceil$, the final output achieves the error $\|\mathcal{T}_{l_{\mathsf{max}}} - \mathcal{T}^*\| \leq \varepsilon$, which concludes the proof. □

6 Numerical Experiments

In this section, we conduct several numerical experiments for both synthetic data and real data. We also compare the RGrad-TT algorithm with RGrad-Tucker format to show the computation efficiency of the tensor train format. Throughout this section, we shall use relative error frequently and it is defined in the following way:

$$\mathsf{relative\ error} = \frac{\|\widehat{\mathcal{T}} - \mathcal{T}^*\|_{\mathrm{F}}}{\|\mathcal{T}^*\|_{\mathrm{F}}}$$

where $\widehat{\mathcal{T}}$ is the output of the algorithm and $\mathcal{T}^*$ is the original tensor. And all for the numerical experiments, the stopping criteria is chosen when $\|\mathcal{T}_{l+1} - \mathcal{T}_l\|_{\mathrm{F}} / \|\mathcal{T}_l\|_{\mathrm{F}} \leq 0.001$.

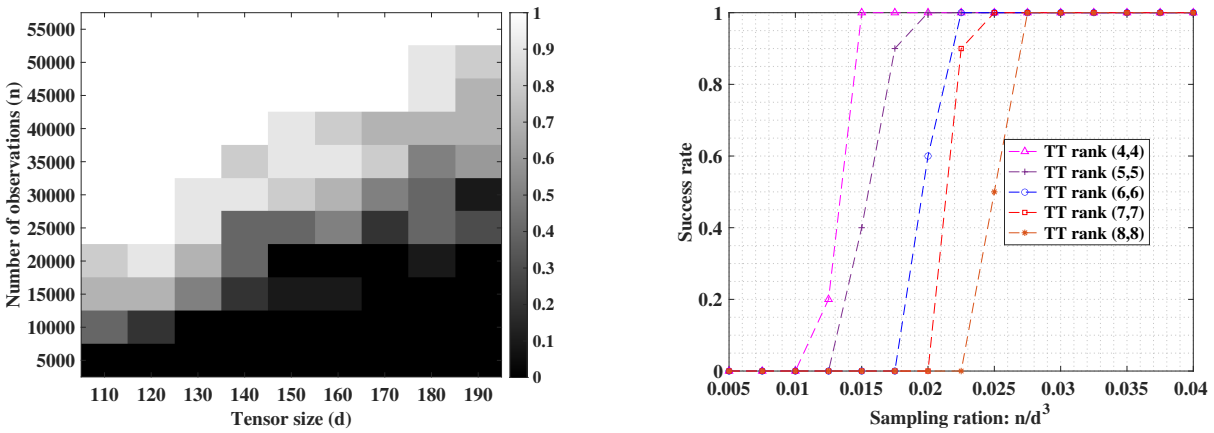
6.1 Synthetic Data

In this section we present some synthetic experiments.

6.1.1 Phase Transition

The number of measurements required for an algorithm to reliably rebuild a low TT rank tensor is an essential question in tensor completion. We explore the recovery abilities of the proposed algorithm in the framework of phase transition, which compares the number of measurements, n, the size of a cubic $d \times d \times d$ tensor of TT rank (r, r).

In the first test, we fix the rank $(r, r) = (2, 2)$ and change the size of the tensor d and the number of measurements n. For each (d, n) tuple, we test 10 random instances. The test low TT



(a) Fix $r = 2$ (b) Fix $d = 100$

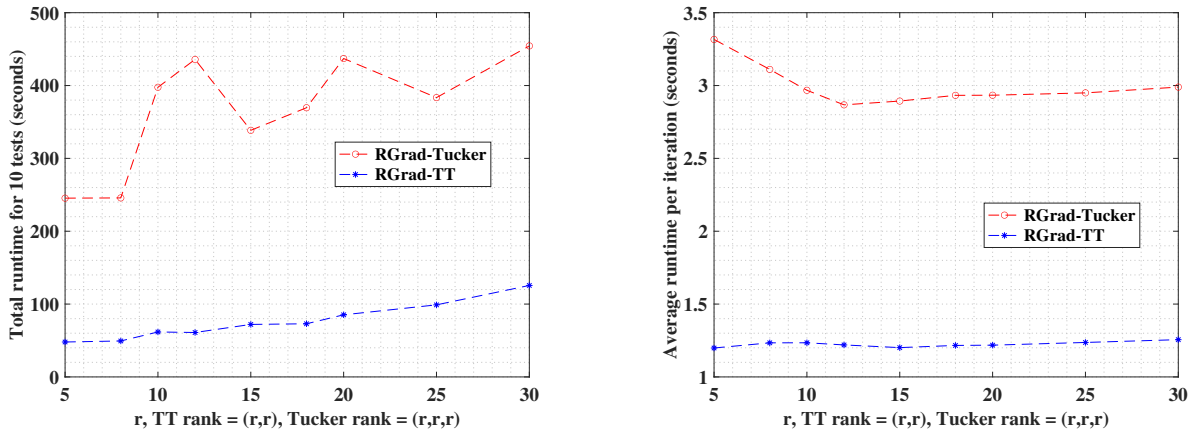
Figure 1: Left: Empirical phase transition curve using RGrad-TT. Successful recovery rate of tensors with a fixed rank $(r, r) = (2, 2)$ from exact data is shown. White denotes successful recovery in all ten random tests, and black denotes failure in all tests. Right: Successful recovery rate of tensors with fix dimension $d = 100$ of different TT rank.

rank (r, r) tensor $\mathcal{T}^*$ is generated from truncating a random Gaussian tensor using TT-SVD. The measurements tensor $\mathcal{P}_\Omega(\mathcal{T}^*)$ is obtained by sampling n entries of $\mathcal{T}^*$ uniformly at random. A test is considered to be successful if the returned tensor $\mathcal{T}$ satisfies $\|\mathcal{T} - \mathcal{T}^*\|_{\mathrm{F}}/\|\mathcal{T}^*\|_{\mathrm{F}} \leq 0.01$. The dimension of the tensor are ranging from 110 to 190, and the measurements are from 5000 to 55000. The probabilities of successful recovery for the RGrad-TT is displayed in Figure 1a. In this figure, white color means that the algorithm can recover all 10 randomly drawn test tensors, whereas the black color shows that the algorithm cannot recover any of the tensors. A clear phase transition can be seen from the figure.

In the second test, we fix the dimension $d = 100$. For each (n, r) tuple, we conduct 10 random instances and a test is considered to be successful if the returned tensor $\widehat{\mathcal{T}}$ satisfies $\|\widehat{\mathcal{T}} - \mathcal{T}^*\|_{\mathrm{F}}/\|\mathcal{T}^*\|_{\mathrm{F}} \leq 0.01$. We plot the curves of successful recovery rate against the sampling ration n/d^3 of tensors with TT ranks from $(4, 4)$ to $(8, 8)$ in Figure 1b.

6.1.2 Efficiency of RGrad-TT

We also conduct experiments to illustrate the efficiency of RGrad-TT against RGrad-Tucker. In this test, we fix the dimension $d = 300$ and we consider cubic tensor $\mathcal{T}^*$ of size $d \times d \times d$. We modify the parameter r and consider random tensors of TT rank (r, r) and Tucker rank (r, r, r). The TT rank (r, r) tensor is generated by applying TT-SVD to a random Gaussian tensor and the Tucker



(a) Total runtime of 10 instances

(b) Average runtime per iteration

Figure 2: Left: Total runtime of 10 instances; Right: average runtime for per iteration. In both cases we fix the tensor of size $300\times300\times300$.

rank (r, r, r) tensor is generated by applying HOSVD to a random Gaussian tensor. And we use Algorithm 3 for initialization for TT format and use the second order moment method proposed by Xia and Yuan (2019) as initialization for Tucker format. To eliminate the impact of the randomness, for each fixed r, we conduct experiments on 10 instances and we count the total runtime and runtime for per iteration. The stopping criteria is satisfied when $\|\mathcal{T}_{l+1} - \mathcal{T}_l\|_{\mathrm{F}}/\|\mathcal{T}_l\|_{\mathrm{F}} \leq 0.001$. The results are shown in Figure 2.

From the figures, we can see that when the TT rank is (r, r) and Tucker rank is (r, r, r), RGrad-TT is much faster in terms of both total runtime and per iteration runtime.

6.2 Real Data: Video Completion

We consider a video of a tomato of size (242,320,167), where the third dimension is the number of frames. Since the video is an RGB one, we concatenate one channel after another along the third direction and the size of the true tensor $\mathcal{T}^*$ is (242,320,501).

To demonstrate the represent-ability of TT format against Tucker format, we apply TTSVD and HOSVD to the original video with TT rank (r, r) and Tucker rank (r, r, r). The approximation error is measured in the relative error and is plotted in the red and pink curves. From these curves, we can see that TTSVD has a better performance in approximating this video.

Then we conduct video completion for this data in both TT format and Tucker format. Suppose 90% of the pixels are missing and we would like to recover the original video. We use RGrad-TT (Algorithm 2) with Algorithm 3 as initialization. To make the initialization comparable, we use

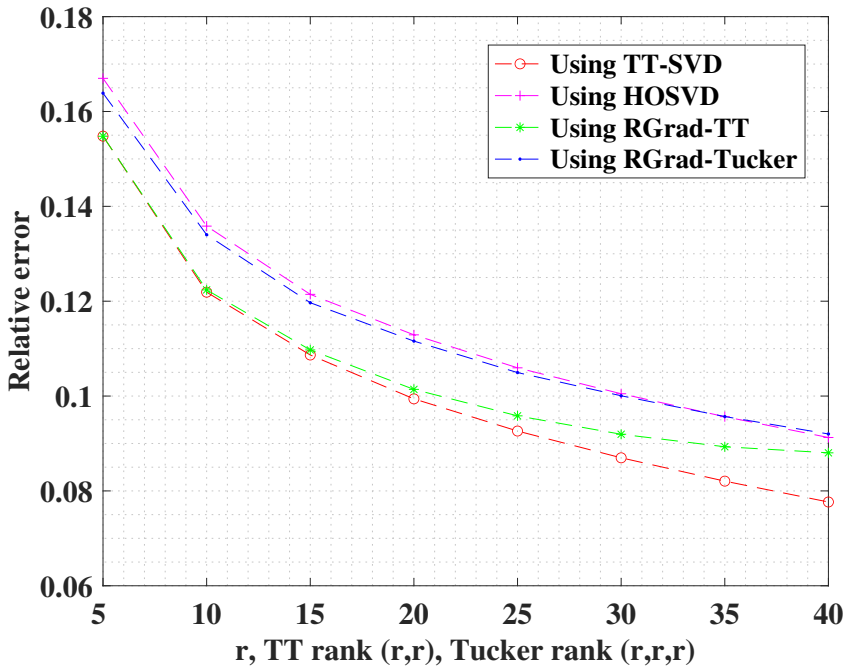


Figure 3: Results using TTSVD, HOSVD to approximate the original video with the given rank and using RGrad-TT, RGrad-Tucker to solve the tensor completion problem in the corresponding format with given rank and 10 percent of observed entries.

second order method introduced in Xia and Yuan (2019) for Tucker format. We change the rank parameter r, and the corresponding TT rank is (r, r) and the Tucker rank is (r, r, r). The recovered accuracy is measured in terms of relative error as shown in Figure 3[2]. From this result, we can see when we fix TT rank to be (r, r) and Tucker rank to be (r, r, r), the accuracy using RGrad-TT is better than RGrad-Tucker. Also, we can see that the recovered accuracy is almost the same as the approximation using TTSVD, which is a quasi-optimal approximation as shown in Oseledets (2011).

[2]The results using RGrad-Tucker is better than using HOSVD for small r since HOSVD is only a quasi-optimal approximation. We show the results for HOSVD only for reference.

References

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

Boaz Barak and Ankur Moitra. Noisy tensor completion via the sum-of-squares hierarchy. In *Conference on Learning Theory*, pages 417–445. PMLR, 2016.

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

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

Matthew Brennan, Guy Bresler, and Wasim Huleihel. Reducibility and computational lower bounds for problems with planted sparse structure. In *Conference On Learning Theory*, pages 48–166. PMLR, 2018.

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

Changxiao Cai, Gen Li, H Vincent Poor, and Yuxin Chen. Nonconvex low-rank tensor completion from noisy data. *Operations Research*, 2021a.

Jian-Feng Cai, Emmanuel J Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on optimization*, 20(4):1956–1982, 2010.

Jian-Feng Cai, Jingyang Li, and Dong Xia. Generalized low-rank plus sparse tensor estimation by fast riemannian optimization. *arXiv preprint arXiv:2103.08895*, 2021b.

Tianxi Cai, T Tony Cai, and Anru Zhang. Structured matrix completion with applications to genomic data integration. *Journal of the American Statistical Association*, 111(514):621–633, 2016.

Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009.

Emmanuel J Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

Yuxin Chen, Jianqing Fan, Cong Ma, and Yuling Yan. Inference and uncertainty quantification for noisy matrix completion. *Proceedings of the National Academy of Sciences*, 116(46):22931–22937, 2019.

Joseph Y Cheng, Tao Zhang, Marcus T Alley, Martin Uecker, Michael Lustig, John M Pauly, and Shreyas S Vasanawala. Comprehensive multi-dimensional mri for the simultaneous assessment of cardiopulmonary anatomy and physiology. *Scientific reports*, 7(1):1–15, 2017.

Mark A Davenport and Justin Romberg. An overview of low-rank matrix recovery from incomplete observations. *IEEE Journal of Selected Topics in Signal Processing*, 10(4):608–622, 2016.

Meng Ding, Ting-Zhu Huang, Teng-Yu Ji, Xi-Le Zhao, and Jing-Hua Yang. Low-rank tensor completion using matrix factorization based on tensor train rank and total variation. *Journal of Scientific Computing*, 81(2):941–964, 2019.

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

Gregory Ely, Shuchin Aeron, Ning Hao, and Misha E Kilmer. 5d and 4d pre-stack seismic data completion using tensor nuclear norm (tnn). In *SEG Technical Program Expanded Abstracts 2013*, pages 3639–3644. Society of Exploration Geophysicists, 2013.

Shmuel Friedland and Lek-Heng Lim. Computational complexity of tensor nuclear norm. *arXiv preprint arXiv:1410.6072*, 2014.

Silvia Gandy, Benjamin Recht, and Isao Yamada. Tensor completion and low-n-rank tensor recovery via convex optimization. *Inverse problems*, 27(2):025010, 2011.

Evarist Giné and Richard Nickl. *Mathematical foundations of infinite-dimensional statistical models*. Cambridge university press, 2021.

Donald Goldfarb and Shiqian Ma. Convergence of fixed-point continuation algorithms for matrix rank minimization. *Foundations of Computational Mathematics*, 11(2):183–210, 2011.

David Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory*, 57(3):1548–1566, 2011.

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

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

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

Shenglong Hu. Relations of the nuclear norm of a tensor and its matrix flattenings. *Linear Algebra and its Applications*, 478:188–199, 2015.

Bo Huang, Cun Mu, Donald Goldfarb, and John Wright. Provable models for robust low-rank tensor completion. *Pacific Journal of Optimization*, 11(2):339–364, 2015.

Hilda S Ibriga and Will Wei Sun. Covariate-assisted sparse tensor completion. *arXiv preprint arXiv:2103.06428*, 2021.

Masaaki Imaizumi, Takanori Maehara, and Kohei Hayashi. On tensor train rank minimization: Statistical efficiency and scalable algorithm. *arXiv preprint arXiv:1708.00132*, 2017.

Prateek Jain and Sewoong Oh. Provable tensor factorization with missing data. In *Advances in Neural Information Processing Systems*, pages 1431–1439, 2014.

Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from a few entries. *IEEE transactions on information theory*, 56(6):2980–2998, 2010.

Ching-Yun Ko, Kim Batselier, Lucas Daniel, Wenjian Yu, and Ngai Wong. Fast and accurate tensor completion with total variation regularized tensor trains. *IEEE Transactions on Image Processing*, 29:6918–6931, 2020.

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

Vladimir Koltchinskii and Dong Xia. Optimal estimation of low rank density matrices. *J. Mach. Learn. Res.*, 16(53):1757–1792, 2015.

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

Nadia Kreimer, Aaron Stanton, and Mauricio D Sacchi. Tensor completion based on nuclear norm minimization for 5d seismic data reconstruction. *Geophysics*, 78(6):V273–V284, 2013.

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

Xutao Li, Yunming Ye, and Xiaofei Xu. Low-rank tensor completion with total variation for visual data inpainting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

Allen Liu and Ankur Moitra. Tensor completion made practical. *arXiv preprint arXiv:2006.03134*, 2020.

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

Christian Lubich, Ivan V Oseledets, and Bart Vandereycken. Time integration of tensor trains. *SIAM Journal on Numerical Analysis*, 53(2):917–941, 2015.

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

Raghu Meka, Prateek Jain, and Inderjit S Dhillon. Guaranteed rank minimization via singular value projection. *arXiv preprint arXiv:0909.5457*, 2009.

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

Ivan Oseledets. Compact matrix form of the d-dimensional tensor decomposition. *IEICE Proceedings Series*, 43(B2L-C2), 2009.

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

Colin Pawlowski. *Machine learning for problems with missing and uncertain data with applications to personalized medicine*. PhD thesis, Massachusetts Institute of Technology, 2019.

David Perez-Garcia, Frank Verstraete, Michael M Wolf, and J Ignacio Cirac. Matrix product state representations. *arXiv preprint quant-ph/0608197*, 2006.

Aaron Potechin and David Steurer. Exact tensor completion with sum-of-squares. In *Conference on Learning Theory*, pages 1619–1673. PMLR, 2017.

Benjamin Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12(12), 2011.

Oguz Semerci, Ning Hao, Misha E Kilmer, and Eric L Miller. Tensor-based formulation and nuclear norm regularization for multienergy computed tomography. *IEEE Transactions on Image Processing*, 23(4):1678–1693, 2014.

Qingquan Song, Hancheng Ge, James Caverlee, and Xia Hu. Tensor completion algorithms in big data analytics. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 13(1):1–48, 2019.

SM Reza Soroushmehr and Kayvan Najarian. Transforming big data into computational models for personalized medicine and health care. *Dialogues in clinical neuroscience*, 18(3):339, 2016.

Michael Steinlechner. Riemannian optimization for high-dimensional tensor completion. *SIAM Journal on Scientific Computing*, 38(5):S461–S484, 2016.

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

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

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

Guifré Vidal. Efficient classical simulation of slightly entangled quantum computations. *Physical review letters*, 91(14):147902, 2003.

Guifré Vidal. Efficient simulation of one-dimensional quantum many-body systems. *Physical review letters*, 93(4):040502, 2004.

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

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

Ke Wei, Jian-Feng Cai, Tony F Chan, and Shingyu Leung. Guarantees of riemannian optimization for low rank matrix completion. *arXiv preprint arXiv:1603.06610*, 2016a.

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

Dong Xia. Normal approximation and confidence region of singular subspaces. *Electronic Journal of Statistics*, 15(2):3798–3851, 2021.

Dong Xia and Vladimir Koltchinskii. Estimation of low rank density matrices: bounds in schatten norms and other distances. *Electronic Journal of Statistics*, 10(2):2717–2745, 2016.

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

Dong Xia and Ming Yuan. Statistical inferences of linear forms for noisy matrix completion. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(1):58–77, 2021.

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

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

Ming Yuan and Cun-Hui Zhang. On tensor completion via nuclear norm minimization. *Foundations of Computational Mathematics*, 16(4):1031–1068, 2016.

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

Anru Zhang. Cross: Efficient low-rank tensor completion. *The Annals of Statistics*, 47(2):936–964, 2019.

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

Yuchen Zhou, Anru R Zhang, Lili Zheng, and Yazhen Wang. Optimal high-order tensor svd via tensor-train orthogonal iteration. *arXiv preprint arXiv:2010.02482*, 2020.

A Proofs of Main Lemmas

In this section, we will present the proofs for Lemma 4 and Lemma 5. Before we start the proof, we shall introduce some notations and definitions that will be used throughout in the proof.

Following Yuan and Zhang (2016), we define the spectral norm of $\mathcal{T} \in \mathbb{R}^{d_1 \times \cdots \times d_m}$ as

$$\|\mathcal{T}\| := \sup_{u_i \in \mathbb{S}^{d_i - 1}} \langle \mathcal{T}, u_1 \otimes \cdots \otimes u_m \rangle,$$

where $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d : \|x\|_{\ell_2} = 1\}$. The nuclear norm is the dual of spectral norm:

$$\|\mathcal{T}\|_* := \max_{\|\mathcal{Y}\| \le 1} \langle \mathcal{T}, \mathcal{Y} \rangle.$$

The relation between nuclear norm and Frobenius norm for tensors of low TT rank is summarized in Lemma 17.

We shall use the semicolon ; to separate the row and column indices of a matrix, for example, we shall write $\mathcal{T}^{\langle i \rangle}(x_1, \ldots, x_i; x_{i+1}, \ldots, x_m)$. The operator $\mathsf{reshape}(M, [p_1, \ldots, p_m])$ reshapes the data M of size $p_1 \cdots p_m$ to a tensor of size $p_1 \times \cdots \times p_m$ in the Matlab way.

We also define some norms for matrices. For a matrix $A \in \mathbb{R}^{p_1 \times p_2}$, the $\|\cdot\|_{2,\infty}$ of A is defined as $\|A\|_{2,\infty} = \max_{i=1}^{p_1} \|A(i,:)\|_{\ell_2}$ and the $\|\cdot\|_{\ell_\infty}$ norm of A is defined to be $\|A\|_{\ell_\infty} = \max_{i,j} |A(i,j)|$.

We also define the projection distance and the chordal distance between two orthogonal matrices $U, V \in \mathbb{R}^{p \times r}$ as

$$d_p(U,V) = \|UU^T - VV^T\|_{\mathrm{F}}, \quad d_c(U,V) = \min_{Q \in \mathbb{O}_r} \|UQ - V\|_{\mathrm{F}}.$$

Then we have $\frac{1}{\sqrt{2}} d_c(U,V) \le d_p(U,V) \le d_c(U,V)$.

A.1 Proof of Lemma 4

We first restate a more detailed version of the lemma.

Lemma 6 (Restate of Lemma 4)**.** *Suppose the conditions on $\mathcal{T}^*$ from Theorem 1 hold, and the initialization $\mathcal{T}_0$ satisfies*

$$\|\mathcal{T}_0 - \mathcal{T}^*\|_{\mathrm{F}} \le \frac{\sigma}{Cm\kappa_0 \overline{r}^{1/2}} \quad \text{and} \quad \mathit{Incoh}(\mathcal{T}_0) \le 2\kappa_0^2 \nu$$

for a sufficiently large but absolute constant $C > 0$. There exists an absolute constant $C_m > 0$ depending only on m such that if the sample size n satisfies

$$n \ge C_m \Big(\kappa_0^{2m+4} \nu^{m+1} \log^{m+2}(\overline{d}) \cdot (d^*)^{1/2} ((r^*)^{1/2} \overline{r}^2 \vee r^* \overline{r}^{1/2}) + \kappa_0^{4m+8} \nu^{2m+2} \log^{2m+4}(\overline{d}) \cdot \overline{d}(r^* \overline{r}^4 \vee (r^*)^2 \overline{r}) \Big),$$

then with probability at least $1-(m+4)\overline{d}^{-m}$*, the sequence* $\{\mathcal{T}_l\}_{l=1}^{\infty}$ *generated by Algorithm 2 with a constant step size* $\alpha = 0.12n^{-1}d^*$ *satisfy*

$$\|\mathcal{T}_l - \mathcal{T}^*\|_{\mathrm{F}}^2 \leq 0.975 \cdot \|\mathcal{T}_{l-1} - \mathcal{T}^*\|_{\mathrm{F}}^2.$$

for all $l = 1, 2, \cdots$.

Now we present the proof of this lemma. First we introduce several events that will be useful in the proof. And the randomness of these events are from the sampling set Ω.

$$\begin{aligned}
\boldsymbol{\mathcal{E}}_1 &= \{\|\frac{d^*}{n}\mathcal{P}_{\mathbb{T}^*}\mathcal{P}_{\Omega}\mathcal{P}_{\mathbb{T}^*} - \mathcal{P}_{\mathbb{T}^*}\| \leq \frac{1}{2}\},\\
\boldsymbol{\mathcal{E}}_2 &= \{\max_{x\in[d^*]}\sum_{i=1}^{n} I(\omega_i = x) \leq 2m\log(\overline{d})\},\\
\boldsymbol{\mathcal{E}}_3 &= \{\|(\mathcal{P}_{\Omega} - \frac{n}{d^*}\mathcal{I})(\mathcal{J})\| \leq C_m\left(\sqrt{\frac{n\overline{d}}{d^*}}+1\right)\log^{m+2}(\overline{d})\},
\end{aligned}$$

where $\mathcal{I} \in \mathbb{R}^{d_1\times\cdots\times d_m}$ is the tensor with all its entries one and $\mathcal{I}$ is the identity operator from $\mathbb{R}^{d_1\times\cdots\times d_m}$ to $\mathbb{R}^{d_1\times\cdots\times d_m}$. From Lemma 21, $\boldsymbol{\mathcal{E}}_1$ holds with probability exceeding $1-\overline{d}^{-m}$. From Lemma 23 $\boldsymbol{\mathcal{E}}_2$ holds with probability exceeding $1-\overline{d}^{-m}$. From Lemma 22, $\boldsymbol{\mathcal{E}}_3$ holds with probability exceeding $1-\overline{d}^{-m}$.

Also we consider the following empirical process:

$$\beta_n(\gamma_1,\gamma_2) := \sup_{\mathcal{A}\in\mathbb{K}_{\gamma_1,\gamma_2}}\left|\langle\mathcal{P}_{\Omega}\mathcal{A},\mathcal{A}\rangle - \frac{n}{d^*}\|\mathcal{A}\|_{\mathrm{F}}^2\right|, \tag{7}$$

where

$$\mathbb{K}_{\gamma_1,\gamma_2} = \{\mathcal{A}\in\mathbb{R}^{d_1\times\ldots\times d_m} : \|\mathcal{A}\|_{\mathrm{F}} \leq 1, \|\mathcal{A}\|_{\ell_\infty} \leq \gamma_1, \|\mathcal{A}\|_* \leq \gamma_2\}.$$

The following lemma states gives the upper bound for $\beta_n(\gamma_1,\gamma_2)$ with high probability.

Lemma 7. *Given* $0 < \delta_1^- < \delta_1^+, 0 < \delta_2^- < \delta_2^+$ *and* $t \geq 1$*, let*

$$t = s + \log 2 + \log(\log_2(\frac{\delta_1^+}{\delta_1^-})) + \log(\log_2(\frac{\delta_2^+}{\delta_2^-})).$$

Then there exists a universal constant $C_m > 0$ *such that with probability at least* $1-e^{-s}$*, the following bound holds for all* $\gamma_1 \in [\delta_1^-, \delta_1^+]$ *and all* $\gamma_2 \in [\delta_2^-, \delta_2^+]$*,*

$$\beta_n(\gamma_1,\gamma_2) \leq C_m\gamma_1\gamma_2\left(\sqrt{\frac{n\overline{d}}{d^*}}+1\right)\log^{m+2}(\overline{d}) + 4\gamma_1\sqrt{\frac{nt}{d^*}} + 8\gamma_1^2 t.$$

Now we consider for any $\mathcal{A} \in \mathbb{R}^{d_1 \times \ldots \times d_m}$, we have $\frac{\|\mathcal{A}\|_{\ell_\infty}}{\|\mathcal{A}\|_{\mathrm{F}}} \in [1/d^*, 1]$. Also from (Hu 2015, Lemma 5.1), $\frac{\|\mathcal{A}\|_*}{\|\mathcal{A}\|_{\mathrm{F}}} \in [1, \overline{d}^{(m-1)/2}]$. So we use Lemma 7 with $\delta_1^- = 1/d^*, \delta_1^+ = 1, \delta_2^- = 1, \delta_2^+ = \overline{d}^{(m-1)/2}$ and $s = \alpha \log(\overline{d})$, then $t = m\log(\overline{d}) + \log 2 + \log(\log_2(d^*)) + \log(\log_2(\overline{d}^{(m-1)/2})) \leq 4m\log(\overline{d})$. So with probability exceeding $1 - \overline{d}^{-m}$, for all $\gamma_1 \in [1/d^*, 1]$ and $\gamma_2 \in [1, \overline{d}^{(m-1)/2}]$,

$$\beta_n(\gamma_1, \gamma_2) \lesssim_m \gamma_1\gamma_2 \left(\sqrt{\frac{n\overline{d}}{d^*}} + 1 \right) \log^{m+2}(\overline{d}) + \gamma_1 \sqrt{\frac{n\log(\overline{d})}{d^*}} + \gamma_1^2 \log(\overline{d}).$$

Denote this event by $\boldsymbol{\mathcal{E}}_4$.

Now we denote the event $\boldsymbol{\mathcal{E}}_5^i = \left\{ \|\mathcal{P}^{(i)}(\mathcal{P}_\Omega - \frac{n}{d^*}\mathcal{I})\mathcal{P}^{(i)}\| \leq C_m \sqrt{\frac{\mu^2\overline{r}^2\overline{d}n\log(\overline{d})}{(d^*)^2}} \right\}$ for all $i \in [m]$, where $\mathcal{P}^{(i)} : \mathbb{R}^{d_1 \times \ldots \times d_m} \to \mathbb{R}^{d_1 \times \ldots \times d_m}$ in terms of its i-th separation:

$$(\mathcal{P}^{(i)}\mathcal{X})^{\langle i \rangle} = (T^{*\leq i-1}T^{*\leq i-1T} \otimes I)\mathcal{X}^{\langle i \rangle}V_{i+1}^*V_{i+1}^{*T}.$$

It is easy to see that $\mathcal{P}^{(i)}$ is a projection and this operator is independent of the choice of left orthogonal representation of $\mathcal{T}^*$. Set $\boldsymbol{\mathcal{E}}_5 = \cap_{i=1}^m \boldsymbol{\mathcal{E}}_5^i$, then $\boldsymbol{\mathcal{E}}_5$ holds with probability exceeding $1 - m\overline{d}^{-m}$ from the following lemma.

Lemma 8. *Suppose that $\mathcal{T}^* = [T_1^*, \ldots, T_m^*] \in \mathbb{M}_{\boldsymbol{r}}^{\boldsymbol{tt}}$ satisfies $\mathsf{Incoh}(\mathcal{T}^*) \leq \sqrt{\mu}$. And Ω is sampled uniformly with replacement such that $|\Omega| = n$. Then we have with probability exceeding $1 - \overline{d}^{-m}$,*

$$\|\mathcal{P}^{(i)}(\mathcal{P}_\Omega - \frac{n}{d^*}\mathcal{I})\mathcal{P}^{(i)}\| \leq C_m \sqrt{\frac{\mu^2\overline{r}^2\overline{d}n\log(\overline{d})}{(d^*)^2}}$$

holds as long as $n \geq C\mu^2\overline{r}^2\overline{d}\log(\overline{d})$.

When $\mathcal{A} = [T_1^*, \ldots, T_{i-1}^*, A, T_{i+1}^*, \ldots, T_m^*], \mathcal{B} = [T_1^*, \ldots, T_{i-1}^*, B, T_{i+1}^*, \ldots, T_m^*]$, then we have $\mathcal{P}^{(i)}\mathcal{A} = \mathcal{A}$ and $\mathcal{P}^{(i)}\mathcal{B} = \mathcal{B}$. Applying Cauchy-Schwartz leads to the following corollary of Lemma 8.

Corollary 1. *For any $i \in [m]$, set*

$$\mathcal{A} = [T_1^*, \ldots, T_{i-1}^*, A, T_{i+1}^*, \ldots, T_m^*], \quad \mathcal{B} = [T_1^*, \ldots, T_{i-1}^*, B, T_{i+1}^*, \ldots, T_m^*]$$

for arbitrary $A, B \in \mathbb{R}^{r_{i-1} \times d_i \times r_i}$. Then under the event $\boldsymbol{\mathcal{E}}_5^i$, we have

$$\langle \mathcal{A}, (\mathcal{P}_\Omega - \frac{n}{d^*}\mathcal{I})\mathcal{B} \rangle \leq C_m \sqrt{\frac{\mu^2\overline{r}^2\overline{d}n\log(\overline{d})}{(d^*)^2}} \|\mathcal{A}\|_{\mathrm{F}}\|\mathcal{B}\|_{\mathrm{F}}.$$

And we denote $\boldsymbol{\mathcal{E}} = \boldsymbol{\mathcal{E}}_1 \cap \boldsymbol{\mathcal{E}}_2 \cap \boldsymbol{\mathcal{E}}_3 \cap \boldsymbol{\mathcal{E}}_4 \cap \boldsymbol{\mathcal{E}}_5$. Then $\boldsymbol{\mathcal{E}}$ holds with probability exceeding $1 - (m+4)\overline{d}^{-m}$. Now we proceed assuming $\boldsymbol{\mathcal{E}}$ holds.

Using the idea of induction, we start the proof assuming $\|\mathcal{T}_l - \mathcal{T}^*\|_{\mathrm{F}} \leq \frac{\underline{\sigma}}{600000m\kappa_0\sqrt{r}}$ and $\mathsf{Incoh}(\mathcal{T}_l) \leq 2\kappa_0^2\nu$. For simplicity we drop the subscript and denote $\mathcal{T} = \mathcal{T}_l$ and $\mathbb{T} = \mathbb{T}_l$ in the following.

Now suppose we fix a left orthogonal decomposition of $\mathcal{T} = [T_1, \ldots, T_m]$, we choose a left orthogonal decomposition for $\mathcal{T}^*$ accordingly. First let $\mathcal{T}^* = [T_1', \ldots, T_m']$ be a left orthogonal decomposition. Define $R_1 = \arg\min_{R \in \mathbb{O}_{r_1}} \|T_1 - T_1' R\|_{\mathrm{F}}$. Now suppose we obtain $R_1, \ldots, R_{i-1}$, define

$$R_i = \arg\min_{R \in \mathbb{O}_{r_i}} \|T^{\leq i} - T'^{\leq i} R\|_{\mathrm{F}}.$$

In this way we obtain $R_1, \ldots, R_{m-1}$. And we define $L(T_i^*) = (R_{i-1} \otimes I)^T L(T_i') R_i$ for $i \in [m-1]$ using the convention $R_0 = [1]$ and $T_m^* = R_{m-1}^T T_m'$. Now we can prove by induction that $[T_1', \ldots, T_m'] = [T_1^*, \ldots, T_m^*]$ and $T'^{\leq i} R_i = T^{*\leq i}$ So we take $\mathcal{T}^* = [T_1^*, \ldots, T_m^*]$ to be the left orthogonal decomposition, and it is the one such that $T^{*\leq i}$ and $T^{\leq i}$ are aligned in the sense that $d_c(T^{\leq i}, T^{*\leq i}) = \|T^{\leq i} - T^{*\leq i}\|_{\mathrm{F}}$. And we can write $T^{\geq j+1} = \Lambda_{j+1} V_{j+1}^T$ and $T^{*\geq j+1} = \Lambda_{j+1}^* V_{j+1}^{*T}$ such that $\Lambda_{j+1}, \Lambda_{j+1} \in \mathbb{R}^{r_j}$ are invertible and V_{j+1}, V_{j+1}^* are orthogonal and $d_c(V_{j+1}, V_{j+1}^*) = \|V_{j+1} - V_{j+1}^*\|_{\mathrm{F}}$.

As a result from the above alignment process and Wedin's theorem, we have for all $i \in [m-1]$,

$$\max\{\|T^{\leq i} - T^{*\leq i}\|_{\mathrm{F}}, \|V_{i+1} - V_{i+1}^*\|_{\mathrm{F}}\} \leq \frac{2\|\mathcal{T} - \mathcal{T}^*\|_{\mathrm{F}}}{\underline{\sigma}}. \tag{8}$$

We first derive the upper bounds for the terms in $\|\mathcal{T}_{l+1} - \mathcal{T}^*\|_{\mathrm{F}}^2$ in the following subsections.

A.1.1 Estimation of $\langle \mathcal{T} - \mathcal{T}^*, \mathcal{P}_\Omega(\mathcal{T} - \mathcal{T}^*) \rangle$.

Since the operator $\mathcal{P}_\Omega$ is SPD, we have

$$\langle \mathcal{T} - \mathcal{T}^*, \mathcal{P}_\Omega(\mathcal{T} - \mathcal{T}^*) \rangle \geq \frac{1}{2} \langle \mathcal{P}_\Omega \mathcal{P}_{\mathbb{T}^*}(\mathcal{T} - \mathcal{T}^*), \mathcal{P}_{\mathbb{T}^*}(\mathcal{T} - \mathcal{T}^*) \rangle - \langle \mathcal{P}_\Omega \mathcal{P}_{\mathbb{T}^*}^\perp(\mathcal{T}), \mathcal{P}_{\mathbb{T}^*}^\perp(\mathcal{T}) \rangle. \tag{9}$$

Since $\boldsymbol{\mathcal{E}}$ holds, we have

$$\begin{aligned} \langle \mathcal{P}_\Omega \mathcal{P}_{\mathbb{T}^*}(\mathcal{T} - \mathcal{T}^*), \mathcal{P}_{\mathbb{T}^*}(\mathcal{T} - \mathcal{T}^*) \rangle &\geq \frac{n}{2d^*} \|\mathcal{P}_{\mathbb{T}^*}(\mathcal{T} - \mathcal{T}^*)\|_{\mathrm{F}} = \frac{n}{2d^*} \|\mathcal{T} - \mathcal{T}^*\|_{\mathrm{F}}^2 - \frac{n}{2d^*} \|\mathcal{P}_{\mathbb{T}^*}^\perp(\mathcal{T})\|_{\mathrm{F}}^2 \\ &\overset{\text{Coro. 2}}{\geq} \frac{n}{2d^*} \|\mathcal{T} - \mathcal{T}^*\|_{\mathrm{F}}^2 - \frac{n}{d^*} \frac{200m^2 \|\mathcal{T} - \mathcal{T}^*\|_{\mathrm{F}}^4}{\underline{\sigma}^2}. \end{aligned} \tag{10}$$

On the other hand, we consider the upper bound for $\langle \mathcal{P}_\Omega \mathcal{P}_{\mathbb{T}^*}^\perp \mathcal{T}, \mathcal{P}_{\mathbb{T}^*}^\perp \mathcal{T} \rangle$. Since $\mathcal{T}$ and Ω are dependent, we need to consider the empirical process (7). First notice that

$$\langle \mathcal{P}_\Omega \mathcal{P}_{\mathbb{T}^*}^\perp \mathcal{T}, \mathcal{P}_{\mathbb{T}^*}^\perp \mathcal{T} \rangle \leq \frac{n}{d^*} \|\mathcal{P}_{\mathbb{T}^*}^\perp \mathcal{T}\|_{\mathrm{F}}^2 + \|\mathcal{P}_{\mathbb{T}^*}^\perp \mathcal{T}\|_{\mathrm{F}}^2 \beta_n \left(\frac{\|\mathcal{P}_{\mathbb{T}^*}^\perp \mathcal{T}\|_{\ell_\infty}}{\|\mathcal{P}_{\mathbb{T}^*}^\perp \mathcal{T}\|_{\mathrm{F}}}, \frac{\|\mathcal{P}_{\mathbb{T}^*}^\perp \mathcal{T}\|_*}{\|\mathcal{P}_{\mathbb{T}^*}^\perp \mathcal{T}\|_{\mathrm{F}}} \right).$$

Under $\boldsymbol{\mathcal{E}}_4$, we have for any $\mathcal{A} \in \mathbb{R}^{d_1 \times \ldots \times d_m}$,

$$\|\mathcal{A}\|_{\mathrm{F}}^2 \beta_n \left(\frac{\|\mathcal{A}\|_{\ell_\infty}}{\|\mathcal{A}\|_{\mathrm{F}}}, \frac{\|\mathcal{A}\|_*}{\|\mathcal{A}\|_{\mathrm{F}}} \right) \lesssim_m \|\mathcal{A}\|_{\ell_\infty} \|\mathcal{A}\|_* \left(\sqrt{\frac{n\bar{d}}{d^*}} + 1 \right) \log^{m+2}(\bar{d}) \\ + \|\mathcal{A}\|_{\ell_\infty} \|\mathcal{A}\|_{\mathrm{F}} \sqrt{\frac{n \log(\bar{d})}{d^*}} + \|\mathcal{A}\|_{\ell_\infty}^2 \log(\bar{d}).$$

And this implies that

$$\langle \mathcal{P}_\Omega \mathcal{A}, \mathcal{A} \rangle \leq \frac{n}{d^*} \|\mathcal{A}\|_{\mathrm{F}}^2 + C_m \|\mathcal{A}\|_{\ell_\infty} \|\mathcal{A}\|_* \left(\sqrt{\frac{n\bar{d}}{d^*}} + 1 \right) \log^{m+2}(\bar{d}). \tag{11}$$

Now we focus on $\|\mathcal{P}_{\mathbb{T}^*}^{\perp} \mathcal{T}\|_{\ell_\infty}, \|\mathcal{P}_{\mathbb{T}^*}^{\perp} \mathcal{T}\|_*$.

Estimation of $\|\mathcal{P}_{\mathbb{T}^*}^{\perp} \mathcal{T}\|_{\ell_\infty}$. Notice that we have $\mathsf{Incoh}(\mathcal{T}) \leq 2\kappa_0^2 \nu$ and $\mathsf{Incoh}(\mathcal{T}^*) \leq \kappa_0 \nu$ from Lemma 2. Using triangle inequality, we have $\|\mathcal{P}_{\mathbb{T}^*}^{\perp} \mathcal{T}\|_{\ell_\infty} \leq \|\mathcal{T}\|_{\ell_\infty} + \|\mathcal{P}_{\mathbb{T}^*} \mathcal{T}\|_{\ell_\infty}$. Meanwhile, for all $1 \leq i \leq m-1$, we have

$$\sigma_{\max}(\Lambda_{i+1}) \leq \sigma_{\max}(\Lambda_{i+1}^*) + \|\mathcal{T} - \mathcal{T}^*\|_{\mathrm{F}} \leq \frac{11}{10} \overline{\sigma}. \tag{12}$$

As a result of the above inequality and Lemma 13, we have

$$\|\mathcal{T}\|_{\ell_\infty} = \|\mathcal{T}^{\langle i \rangle}\|_{\ell_\infty} = \|T^{\leq i} \Lambda_{i+1} V_{i+1}^T\|_{\ell_\infty} \leq \frac{11}{10} \overline{\sigma} \|T^{\leq i}\|_{2,\infty} \|V_{i+1}\|_{2,\infty} \leq \frac{22}{5} \overline{\sigma} \frac{\kappa_0^4 \nu^2 r_i}{\sqrt{d^*}}.$$

As this holds for all $i \in [m-1]$, we have

$$\|\mathcal{T}\|_{\ell_\infty} \leq \frac{22}{5} \frac{\underline{r} \kappa_0^4 \nu^2}{\sqrt{d^*}} \overline{\sigma}. \tag{13}$$

On the other hand, we have $\mathcal{P}_{\mathbb{T}^*} \mathcal{T} = \delta \mathcal{T}_1 + \ldots + \delta \mathcal{T}_m$, where

$$\delta \mathcal{T}_i = [T_1^*, \ldots, T_{i-1}^*, X_i, T_{i+1}^*, \ldots, T_m^*]$$

and the expression of X_i are give in (6). Now we would like to estimate $\|\delta \mathcal{T}_i\|_{\ell_\infty}$. Since the reshape operation remains the infinity norm unchanged, we have for all $i \in [m-1]$,

$$\begin{aligned} \|\delta \mathcal{T}_i\|_{\ell_\infty} &= \|\delta \mathcal{T}_i^{\langle i \rangle}\|_{\ell_\infty} \\ &= \|(T^{*\leq i-1} \otimes I)(I - L(T_i^*) L(T_i^*)^T)(T^{*\leq i-1} \otimes I)^T \mathcal{T}^{\langle i \rangle} V_{i+1}^* V_{i+1}^{*T}\|_{\ell_\infty} \\ &\leq \|(T^{*\leq i-1} \otimes I)(T^{*\leq i-1} \otimes I)^T \mathcal{T}^{\langle i \rangle} V_{i+1}^* V_{i+1}^{*T}\|_{\ell_\infty} + \|T^{*\leq i} T^{*\leq iT} \mathcal{T}^{\langle i \rangle} V_{i+1}^* V_{i+1}^{*T}\|_{\ell_\infty}, \end{aligned}$$

and for $i = m$,

$$\|\delta\mathcal{T}_m\|_{\ell_\infty} = \|\delta\mathcal{T}_m^{\langle m\rangle}\|_{\ell_\infty} = \|(T^{*\le m-1}\otimes I)(T^{*\le m-1}\otimes I)^T\mathcal{T}^{\langle m\rangle}\|_{\ell_\infty}$$
$$\overset{\text{Lemma 12}}{=} \|T^{*\le m-1}T^{*\le m-1T}\mathcal{T}^{\langle m-1\rangle}\|_{\ell_\infty}.$$

We check the term $\|(T^{*\le i-1}\otimes I)(T^{*\le i-1}\otimes I)^T\mathcal{T}^{\langle i\rangle}V_{i+1}^*V_{i+1}^{*T}\|_{\ell_\infty}$.

$$\begin{aligned}
&\|(T^{*\le i-1}\otimes I)(T^{*\le i-1}\otimes I)^T\mathcal{T}^{\langle i\rangle}V_{i+1}^*V_{i+1}^{*T}\|_{\ell_\infty}\\
\overset{(a)}{\le}\ &\sqrt{d_{i+1}\dots d_m}\|(T^{*\le i-1}T^{*\le i-1T}\otimes I)\mathcal{T}^{\langle i\rangle}\|_{\ell_\infty}\|V_{i+1}^*\|_{2,\infty}\\
\overset{(b)}{\le}\ &\|T^{*\le i-1}T^{*\le i-1T}\mathcal{T}^{\langle i-1\rangle}\|_{\ell_\infty}\kappa_0\nu\sqrt{r_i},
\end{aligned}\tag{14}$$

where in (a) we use Lemma 13 and $\|V_{i+1}^*\|_{2,\infty} = \|V_{i+1}^*V_{i+1}^{*T}\|_{2,\infty}$; in (b) we use Lemma 12 and $\mathsf{Incoh}(\mathcal{T}^*) \le \kappa_0\nu$. Also, when $2 \le i \le m$

$$\begin{aligned}
\|T^{*\le i-1}T^{*\le i-1T}\mathcal{T}^{\langle i-1\rangle}\|_{\ell_\infty} &= \|T^{*\le i-1}T^{*\le i-1T}T^{\le i-1}\Lambda_iV_i^T\|_{\ell_\infty}\\
&\le \sigma_{\max}(\Lambda_i)\|V_i\|_{2,\infty}\cdot\max_j\|T^{\le i-1T}T^{*\le i-1}T^{*\le i-1T}e_j\|_{\ell_2}\\
&\overset{(a)}{\le} \sigma_{\max}(\Lambda_i)\|V_i\|_{2,\infty}\|T^{\le i-1}\|_{2,\infty}\|T^{*\le i-1}\|_{2,\infty}\sqrt{d_1\dots d_{i-1}}\\
&\overset{(b)}{\le} \frac{22}{5}\frac{\kappa_0^5\nu^3r_{i-1}^{3/2}}{\sqrt{d^*}}\overline{\sigma},
\end{aligned}$$

where in (a) we use Lemma 13 and in (b) we use $\mathsf{Incoh}(\mathcal{T}^*) \le \kappa_0\nu, \mathsf{Incoh}(\mathcal{T}) \le 2\kappa_0^2\nu$ and (12). And when $i = 1$,

$$\|T^{*\le 0}T^{*\le 0T}\mathcal{T}^{\langle 0\rangle}\|_{\ell_\infty} = \|\mathcal{T}\|_{\ell_\infty} \overset{(13)}{\le} \frac{22}{5}\frac{\underline{r}\kappa_0^4\nu^2}{\sqrt{d^*}}\overline{\sigma}.$$

Combine these with (14) and we get

$$\begin{aligned}
\|(T^{*\le i-1}\otimes I)(T^{*\le i-1}\otimes I)^T\mathcal{T}^{\langle i\rangle}V_{i+1}^*V_{i+1}^{*T}\|_{\ell_\infty} &\le \begin{cases}\frac{22}{5}\frac{\kappa_0^5\nu^3\underline{r}r_1^{1/2}}{\sqrt{d^*}}\overline{\sigma}, & i=1\\ \frac{22}{5}\frac{\kappa_0^6\nu^4r_{i-1}^{3/2}r_i^{1/2}}{\sqrt{d^*}}\overline{\sigma}, & 2\le i\le m\end{cases}\\
&\le \frac{22}{5}\frac{\kappa_0^6\nu^4\overline{r}^2}{\sqrt{d^*}}\overline{\sigma}.
\end{aligned}$$

Now for all $i \in [m-1]$, we check $\|T^{*\le i}T^{*\le iT}\mathcal{T}^{\langle i\rangle}V_{i+1}^*V_{i+1}^{*T}\|_{\ell_\infty}$.

$$\begin{aligned}
\|T^{*\le i}T^{*\le iT}\mathcal{T}^{\langle i\rangle}V_{i+1}^*V_{i+1}^{*T}\|_{\ell_\infty} &= \max_{j,k}|e_j^TT^{*\le i}T^{*\le iT}T^{\le i}\Lambda_{i+1}V_{i+1}^TV_{i+1}^*V_{i+1}^{*T}e_k|\\
&\le \max_{j,k}\sigma_{\max}(\Lambda_{i+1})\|T^{\le iT}T^{*\le i}T^{*\le iT}e_j\|_{\ell_2}\|V_{i+1}^TV_{i+1}^*V_{i+1}^{*T}e_k\|_{\ell_2}\\
&\le \sigma_{\max}(\Lambda_{i+1})\|T^{*\le i}\|_{2,\infty}\|T^{\le i}\|_{2,\infty}\|V_{i+1}\|_{2,\infty}\|V_{i+1}^*\|_{2,\infty}\sqrt{d^*}\\
&\le \frac{22}{5}\frac{\kappa_0^6\nu^4r_i^2}{\sqrt{d^*}}\overline{\sigma}.
\end{aligned}$$

Put these together and we have

$$\|\mathcal{P}_{\mathbb{T}^*}^{\perp}\mathcal{T}\|_{\ell_\infty} \le \frac{44m}{5}\frac{\kappa_0^6\nu^4\overline{r}^2}{\sqrt{d^*}}\overline{\sigma}. \tag{15}$$

Estimation of $\|\mathcal{P}_{\mathbb{T}^*}^{\perp}\mathcal{T}\|_*$. First notice $\mathcal{P}_{\mathbb{T}^*}^{\perp}\mathcal{T} = \mathcal{T} - \mathcal{P}_{\mathbb{T}^*}\mathcal{T}$. From Lemma 16 and $\mathsf{rank}_{\mathsf{tt}}(\mathcal{T}) \le (r_1,\dots,r_{m-1})$, we know $\mathsf{rank}_{\mathsf{tt}}(\mathcal{P}_{\mathbb{T}^*}^{\perp}\mathcal{T}) \le (3r_1,\dots,3r_{m-1})$. Now we use Lemma 17 and we get

$$\|\mathcal{P}_{\mathbb{T}^*}^{\perp}\mathcal{T}\|_* \le 3^{(m-1)/2}\sqrt{r_1\dots r_{m-1}}\|\mathcal{P}_{\mathbb{T}^*}^{\perp}\mathcal{T}\|_{\mathrm{F}}. \tag{16}$$

Now we combine (15), (16), (11) and Lemma 19, since $\|\mathcal{T}-\mathcal{T}^*\|_{\mathrm{F}} \le \frac{\underline{\sigma}}{20m}$,

$$\begin{aligned}\langle \mathcal{P}_\Omega\mathcal{P}_{\mathbb{T}^*}^{\perp}\mathcal{T}, \mathcal{P}_{\mathbb{T}^*}^{\perp}\mathcal{T}\rangle \le& 400m^2\frac{n}{d^*}\frac{\|\mathcal{T}-\mathcal{T}^*\|_{\mathrm{F}}^4}{\underline{\sigma}^2}\\ &+ C_m\frac{\kappa_0^7\nu^4\overline{r}^2}{\sqrt{d^*}}\sqrt{r_1\dots r_{m-1}}\|\mathcal{T}-\mathcal{T}^*\|_{\mathrm{F}}^2\left(\sqrt{\frac{n\overline{d}}{d^*}}+1\right)\log^{m+2}(\overline{d}).\end{aligned} \tag{17}$$

So as long as $n \ge C_m\left(\kappa_0^7\nu^4\sqrt{d^*}\overline{r}^2(r^*)^{1/2}\log^{m+2}(\overline{d}) + \kappa_0^{14}\nu^8\overline{d}\overline{r}^4 r^*\log^{2m+4}(\overline{d})\right)$, we have from (10) and (17), we have

$$\langle \mathcal{P}_\Omega\mathcal{P}_{\mathbb{T}^*}^{\perp}\mathcal{T}, \mathcal{P}_{\mathbb{T}^*}^{\perp}\mathcal{T}\rangle \le \frac{1}{100}\langle \mathcal{P}_\Omega\mathcal{P}_{\mathbb{T}^*}(\mathcal{T}-\mathcal{T}^*), \mathcal{P}_{\mathbb{T}^*}(\mathcal{T}-\mathcal{T}^*)\rangle.$$

Together with (9) and (10), we get

$$\begin{aligned}\langle \mathcal{P}_\Omega\mathcal{P}_{\mathbb{T}^*}\mathcal{T}, \mathcal{P}_{\mathbb{T}^*}\mathcal{T}\rangle &\ge \frac{49}{100}\langle \mathcal{P}_\Omega\mathcal{P}_{\mathbb{T}^*}(\mathcal{T}-\mathcal{T}^*), \mathcal{P}_{\mathbb{T}^*}(\mathcal{T}-\mathcal{T}^*)\rangle\\ &\ge \frac{49}{100}\left(\frac{n}{2d^*}\|\mathcal{T}-\mathcal{T}^*\|_{\mathrm{F}}^2 - \frac{n}{d^*}\frac{200m^2\|\mathcal{T}-\mathcal{T}^*\|_{\mathrm{F}}^4}{\underline{\sigma}^2}\right)\\ &\ge \frac{6n}{25d^*}\|\mathcal{T}-\mathcal{T}^*\|_{\mathrm{F}}^2,\end{aligned} \tag{18}$$

where the last inequality holds since $\|\mathcal{T}-\mathcal{T}^*\|_{\mathrm{F}} \le \frac{1}{600m}\underline{\sigma}$.

A.1.2 Estimation of $\|\mathcal{P}_{\mathbb{T}}\mathcal{P}_\Omega(\mathcal{T}-\mathcal{T}^*)\|_{\mathrm{F}}^2$

First notice that we have both $\mathcal{T}$ and $\mathcal{T}^*$ are of TT rank $(r_1,\dots,r_{m-1})$. And $\mathsf{Incoh}(\mathcal{T}) \le 2\kappa_0^2\nu$, $\mathsf{Incoh}(\mathcal{T}^*) \le \kappa_0\nu \le 2\kappa_0^2\nu =: \sqrt{\mu}$. First notice that

$$\|\mathcal{P}_{\mathbb{T}}\mathcal{P}_\Omega(\mathcal{T}-\mathcal{T}^*)\|_{\mathrm{F}}^2 \le 1001\|\mathcal{P}_{\mathbb{T}}(\mathcal{P}_\Omega - \frac{n}{d^*}\mathcal{I})(\mathcal{T}-\mathcal{T}^*)\|_{\mathrm{F}}^2 + 1.001\frac{n^2}{(d^*)^2}\|\mathcal{P}_{\mathbb{T}}(\mathcal{T}-\mathcal{T}^*)\|_{\mathrm{F}}^2. \tag{19}$$

Now we check $\|\mathcal{P}_{\mathbb{T}}(\mathcal{P}_\Omega - \frac{n}{d^*}\mathcal{I})(\mathcal{T}-\mathcal{T}^*)\|_{\mathrm{F}}$. From the variational representation of Frobenius norm, we can write it as

$$\|\mathcal{P}_{\mathbb{T}}(\mathcal{P}_\Omega - \frac{n}{d^*}\mathcal{I})(\mathcal{T}-\mathcal{T}^*)\|_{\mathrm{F}} = \langle(\mathcal{P}_\Omega - \frac{n}{d^*}\mathcal{I})(\mathcal{T}-\mathcal{T}^*), \mathcal{P}_{\mathbb{T}}(\mathcal{X}_0)\rangle,$$

for some $\mathcal{X}_0$ with $\|\mathcal{X}_0\|_{\mathrm{F}} \le 1$. Now we set $\mathcal{P}_{\mathbb{T}}(\mathcal{X}_0) = \delta\mathcal{X}_1 + \ldots + \delta\mathcal{X}_m$, with $\delta\mathcal{X}_i = [T_1, \ldots, X_i, \ldots, T_m]$. For all $i \in [m]$, we consider the bound for $\langle (\mathcal{P}_\Omega - \frac{n}{d^*}\mathcal{I})(\mathcal{T} - \mathcal{T}^*), \delta\mathcal{X}_i \rangle$. We can decompose $\mathcal{T} - \mathcal{T}^*$ as

$$\begin{aligned}\mathcal{T} - \mathcal{T}^* &= [T_1^*, \ldots, \Delta_i, \ldots, T_m^*] + \sum_{j=1}^{i-1} [T_1^*, \ldots, \Delta_j, \ldots, T_i, T_{i+1}^*, \ldots, T_m^*] \\ &\qquad + \sum_{j=i+1}^{m} [T_1, \ldots, T_i, T_{i+1}, \ldots, \Delta_j, \ldots, T_m^*] \\ &=: \mathcal{Y}_{i,i} + \sum_{j=1}^{i-1} \mathcal{Y}_{i,j} + \sum_{j=i+1}^{m} \mathcal{Y}_{i,j}, \end{aligned} \tag{20}$$

where $\Delta_j = T_j - T_j^*$. Before the estimation, we need the following lemmas whose proofs are presented in the Section C.

Lemma 9. *Suppose that Ω is the set sampled uniformly with replacement with size $|\Omega| = n$. Then under the event $\boldsymbol{\mathcal{E}}_3$, we have for any tensors $\mathcal{A}, \mathcal{B}$ with TT rank $(r_1, \ldots, r_{m-1})$,*

$$\begin{aligned}&|\langle (\mathcal{P}_\Omega - \frac{n}{d^*}\mathcal{I})\mathcal{A}, \mathcal{B} \rangle| \\ &\le C_m \left(\sqrt{\frac{n\bar{d}}{d^*}} + 1 \right) \log^{m+2}(\bar{d}) \cdot \prod_{i=1}^{m} \left(\max_{x_i} \|A_i(:, x_i, :)\|_{\mathrm{F}} \cdot \|B_i\|_{\mathrm{F}} \wedge \max_{x_i} \|B_i(:, x_i, :)\|_{\mathrm{F}} \cdot \|A_i\|_{\mathrm{F}} \right),\end{aligned}$$

where $\mathcal{A} = [A_1, \ldots, A_m]$ and $\mathcal{B} = [B_1, \ldots, B_m]$ can be arbitrary decompositions such that $A_i, B_i \in \mathbb{R}^{r_{i-1} \times d_i \times r_i}$.

Lemma 10. *Let $\mathcal{T}$ be a tensor of rank $(r_1, \ldots, r_{m-1})$ such that $\mathsf{Incoh}(\mathcal{T}) \le \sqrt{\mu}$, and it has a left orthogonal decomposition $\mathcal{T} = [T_1, \ldots, T_m]$. Then we have*

$$\max_{x_i} \|T_i(:, x_i, :)\|_{\mathrm{F}}^2 \le \frac{\mu r_i}{d_i}, \quad \|T_i\|_{\mathrm{F}} = \sqrt{r_i}, i \in [m-1],$$

$$\max_{x_m} \|T_m(:, x_m)\|_{\mathrm{F}}^2 \le \sigma_{\max}^2(\mathcal{T}) \frac{\mu r_{m-1}}{d_m}, \quad \|T_m\|_{\mathrm{F}} = \|\mathcal{T}\|_{\mathrm{F}} \le \sqrt{\underline{r}}\sigma_{\max}(\mathcal{T}).$$

Now we present some bounds related to Δ_j and X_j.

Properties for Δ_j. For all $j \in [m-1]$, we estimate $\|\Delta_j\|_{\mathrm{F}} = \|L(T_j) - L(T_j^*)\|_{\mathrm{F}}$ as follows. Notice

$\mathcal{T}^{\langle j\rangle} = (T^{\le j-1}\otimes I)L(T_j)T^{\ge j+1}, (\mathcal{T}^*)^{\langle j\rangle} = (T^{*\le j-1}\otimes I)L(T_j^*)T^{*\ge j+1}$. So we have,

$$\begin{aligned}
\|L(T_j) - L(T_j^*)\|_{\mathrm{F}} &= \|(T^{\le j-1}\otimes I)^T\mathcal{T}^{\langle j\rangle}V_{j+1}\Lambda_{j+1}^{-1} - (T^{*\le j-1}\otimes I)^T(\mathcal{T}^*)^{\langle j\rangle}V_{j+1}^*(\Lambda_{j+1}^*)^{-1}\|_{\mathrm{F}} \\
&\le \|((T^{\le j-1}\otimes I)^T - (T^{*\le j-1}\otimes I)^T)\mathcal{T}^{\langle j\rangle}V_{j+1}\Lambda_{j+1}^{-1}\|_{\mathrm{F}} \\
&\quad + \|(T^{*\le j-1}\otimes I)^T(\mathcal{T}^{\langle j\rangle} - (\mathcal{T}^*)^{\langle j\rangle})V_{j+1}\Lambda_{j+1}^{-1}\|_{\mathrm{F}} \\
&\quad + \|(T^{*\le j-1}\otimes I)^T(\mathcal{T}^*)^{\langle j\rangle}(V_{j+1}\Lambda_{j+1}^{-1} - V_{j+1}^*(\Lambda_{j+1}^*)^{-1})\|_{\mathrm{F}} \\
&\overset{(a)}{\le} \sigma_{\min}^{-1}(\mathcal{T})\frac{\|\mathcal{T}-\mathcal{T}^*\|_{\mathrm{F}}}{2\sigma_{\min}(\mathcal{T}^*)}\sqrt{r_j}\sigma_{\max}(\mathcal{T}) + \sigma_{\min}^{-1}(\mathcal{T})\|\mathcal{T}-\mathcal{T}^*\|_{\mathrm{F}} \\
&\quad + \sqrt{r}\sigma_{\max}(\mathcal{T}^*)\frac{12\kappa_0\|\mathcal{T}-\mathcal{T}^*\|_{\mathrm{F}}}{\sigma_{\min}^2(\mathcal{T}^*)} \\
&\overset{(b)}{\le} 20\sqrt{r_j}\frac{\kappa_0^2\|\mathcal{T}-\mathcal{T}^*\|_{\mathrm{F}}}{\sigma_{\min}(\mathcal{T}^*)},
\end{aligned}$$

where in (a) we use (8) and the bound

$$\|V_{j+1}\Lambda_{j+1}^{-1} - V_{j+1}^*(\Lambda_{j+1}^*)^{-1}\|_{\mathrm{F}} \le \|(V_{j+1}-V_{j+1}^*)(\Lambda_{j+1}^*)^{-1}\|_{\mathrm{F}} + \|V_{j+1}(\Lambda_{j+1}^{-1} - (\Lambda_{j+1}^*)^{-1})\|_{\mathrm{F}}$$

and Lemma 14 and in (b) we use $|\sigma_{\max}(\mathcal{T}) - \sigma_{\max}(\mathcal{T}^*)| \vee |\sigma_{\min}(\mathcal{T}) - \sigma_{\min}(\mathcal{T}^*)| \le \frac{1}{10}\sigma_{\min}(\mathcal{T}^*)$. When $i = m$, we have

$$\begin{aligned}
\|T_m - T_m^*\|_{\mathrm{F}} &= \|T^{\le m-1T}\mathcal{T}^{\langle m-1\rangle} - T^{*\le m-1T}(\mathcal{T}^*)^{\langle m-1\rangle}\|_{\mathrm{F}} \\
&\le \|(T^{\le m-1T} - T^{*\le m-1T})(\mathcal{T}^*)^{\langle m-1\rangle}\|_{\mathrm{F}} + \|T^{\le m-1T}(\mathcal{T}^{\langle m-1\rangle} - (\mathcal{T}^*)^{\langle m-1\rangle})\|_{\mathrm{F}} \\
&\le 2\sqrt{r_{m-1}}\kappa_0\|\mathcal{T}-\mathcal{T}^*\|_{\mathrm{F}} + \|\mathcal{T}-\mathcal{T}^*\|_{\mathrm{F}} \le 3\sqrt{r_{m-1}}\kappa_0\|\mathcal{T}-\mathcal{T}^*\|_{\mathrm{F}}.
\end{aligned}$$

Therefore, we have

$$\|\Delta_i\|_{\mathrm{F}} \le \begin{cases} 20\sigma_{\min}^{-1}(\mathcal{T}^*)\sqrt{r_i}\kappa_0^2\|\mathcal{T}-\mathcal{T}^*\|_{\mathrm{F}}, i \in [m-1] \\ 3\sqrt{r_{m-1}}\kappa_0\|\mathcal{T}-\mathcal{T}^*\|_{\mathrm{F}}, i = m. \end{cases} \tag{21}$$

On the other hand, from Lemma 10, we have

$$\max_{x_i}\|\Delta_i(:,x_i,:)\|_{\mathrm{F}} \le \begin{cases} 2\sqrt{\mu r_i d_i^{-1}}, i \in [m-1] \\ 2\sqrt{\mu r_{m-1}d_m^{-1}}\cdot\sigma_{\max}(\mathcal{T}^*), i = m. \end{cases} \tag{22}$$

Upper bound for $\|X_j\|_{\mathrm{F}}$. For all $X_i, i \in [m-1]$, we have $L(X_i) = (I - L(T_i)L(T_i)^T)(T^{\le i-1}\otimes I)^T\mathcal{X}_0^{\langle i\rangle}T^{\ge i+1T}(T^{\ge i+1}T^{\ge i+1T})^{-1}$, thus $\|X_i\|_{\mathrm{F}} \le \sigma_{\min}^{-1}(\mathcal{T}) \le 2\sigma_{\min}^{-1}(\mathcal{T}^*)$. For $i = m$, $L(X_m) = (T^{\le m-1}\otimes I)^T\mathcal{X}_0^{\langle m\rangle}$, so $\|X_m\|_{\mathrm{F}} \le 1$. Therefore, we obtain

$$\|X_i\|_{\mathrm{F}} \le 2\sigma_{\min}^{-1}(\mathcal{T}^*), i \in [m-1], \quad \|X_m\|_{\mathrm{F}} \le 1. \tag{23}$$

Now we consider the upper bound for $\langle(\mathcal{P}_\Omega - \frac{n}{d^*}\mathcal{I})(\mathcal{T}-\mathcal{T}^*), \delta\mathcal{X}_i\rangle$.

When $i \in [m-1]$. Due to (20), we can write it as

$$\langle(\mathcal{P}_\Omega - \frac{n}{d^*}\mathcal{I})(\mathcal{T}-\mathcal{T}^*), \delta\mathcal{X}_i\rangle = \sum_{j=1}^{m}\langle(\mathcal{P}_\Omega - \frac{n}{d^*}\mathcal{I})\mathcal{Y}_{i,j}, \delta\mathcal{X}_i\rangle.$$

Now for all $j \neq i, m$, we consider $\langle(\mathcal{P}_\Omega - \frac{n}{d^*}\mathcal{I})(\mathcal{Y}_{i,j}), \delta\mathcal{X}_i\rangle$. By setting $\mathcal{A} = \mathcal{Y}_{i,j}$, $\mathcal{B} = \delta\mathcal{X}_i$ in Lemma 9, together with Lemma 10, (21), (22), (23) and we have under the event $\boldsymbol{\mathcal{E}}_3$,

$$|\langle(\mathcal{P}_\Omega - \frac{n}{d^*}\mathcal{I})(\mathcal{Y}_{i,j}), \delta\mathcal{X}_i\rangle| \leq C_m\left(\sqrt{\frac{n\bar{d}}{d^*}}+1\right)\log^{m+2}(\bar{d})\kappa_0^4\mu^{m/2}\frac{r^*\cdot r_{m-1}}{\sqrt{r_i d^*}}\|\mathcal{T}-\mathcal{T}^*\|_{\mathrm{F}}.$$

For $j = m$, we have under the event $\boldsymbol{\mathcal{E}}_3$,

$$|\langle(\mathcal{P}_\Omega - \frac{n}{d^*}\mathcal{I})(\mathcal{Y}_{i,m}), \delta\mathcal{X}_i\rangle| \leq C_m\left(\sqrt{\frac{n\bar{d}}{d^*}}+1\right)\log^{m+2}(\bar{d})\kappa_0^2\mu^{m/2}\frac{r^*\cdot r_{m-1}}{\sqrt{r_i d^*}}\|\mathcal{T}-\mathcal{T}^*\|_{\mathrm{F}}.$$

And when $j = i$, we estimate $|\langle(\mathcal{P}_\Omega - \frac{n}{d^*}\mathcal{I})(\mathcal{Y}_{i,i}), \delta\mathcal{X}_i\rangle|$. We can write

$$\begin{aligned}\delta\mathcal{X}_i &= [T_1^*, \ldots, X_i, \ldots, T_m^*] + [\Delta_1, T_2, \ldots, X_i, \ldots, T_m] + \ldots + [T_1^*, \ldots, X_i, T_{i+1}^*, \ldots, \Delta_m] \\ &=: \mathcal{X}_{i,0} + \sum_{k=1, k\neq i}^{m}\mathcal{X}_{i,k}. \end{aligned} \tag{24}$$

Then $\langle(\mathcal{P}_\Omega - \frac{n}{d^*}\mathcal{I})(\mathcal{Y}_{i,i}), \delta\mathcal{X}_i\rangle = \langle(\mathcal{P}_\Omega - \frac{n}{d^*}\mathcal{I})(\mathcal{Y}_{i,i}), \mathcal{X}_{i,0}\rangle + \sum_{k=1,k\neq i}^{m}\langle(\mathcal{P}_\Omega - \frac{n}{d^*}\mathcal{I})(\mathcal{Y}_{i,i}), \mathcal{X}_{i,k}\rangle$. And here the first term can be bounded using Corollary 1,

$$|\langle(\mathcal{P}_\Omega - \frac{n}{d^*}\mathcal{I})(\mathcal{Y}_{i,i}), \mathcal{X}_{i,0}\rangle| \leq C_m\kappa_0^4\frac{\mu\bar{r}\sqrt{n\bar{d}\log(\bar{d})}}{d^*}\sqrt{r_i}\|\mathcal{T}-\mathcal{T}^*\|_{\mathrm{F}} \leq C_m\kappa_0^4\frac{\mu\bar{r}^{3/2}\sqrt{n\bar{d}\log(\bar{d})}}{d^*}\|\mathcal{T}-\mathcal{T}^*\|_{\mathrm{F}}.$$

When $k \neq i, m$, we have

$$|\langle(\mathcal{P}_\Omega - \frac{n}{d^*}\mathcal{I})(\mathcal{Y}_{i,i}), \mathcal{X}_{i,k}\rangle| \leq C_m\left(\sqrt{\frac{n\bar{d}}{d^*}}+1\right)\log^{m+2}(\bar{d})\mu^{m/2}\frac{r^*\cdot r_{m-1}}{\sqrt{r_i d^*}}\kappa_0^4\|\mathcal{T}-\mathcal{T}^*\|_{\mathrm{F}}.$$

When $k = m$, we have

$$|\langle(\mathcal{P}_\Omega - \frac{n}{d^*}\mathcal{I})(\mathcal{Y}_{i,i}), \mathcal{X}_{i,m}\rangle| \leq C_m\left(\sqrt{\frac{n\bar{d}}{d^*}}+1\right)\log^{m+2}(\bar{d})\mu^{m/2}\frac{r^*\cdot r_{m-1}}{\sqrt{r_i d^*}}\kappa_0^2\|\mathcal{T}-\mathcal{T}^*\|_{\mathrm{F}}.$$

In summary, under the event $\boldsymbol{\mathcal{E}}_3$, we have

$$|\langle(\mathcal{P}_\Omega - \frac{n}{d^*}\mathcal{I})(\mathcal{Y}_{i,i}), \delta\mathcal{X}_i\rangle| \leq C_m\kappa_0^4\mu^{m/2}\log^{m+2}(\bar{d})\left(\sqrt{\frac{n\bar{d}}{d^*}}+1\right)\frac{r^*\cdot r_{m-1}r_i^{-1/2}}{\sqrt{d^*}}\|\mathcal{T}-\mathcal{T}^*\|_{\mathrm{F}}.$$

Together with the estimation when $j \neq i$, we see that under the event $\boldsymbol{\mathcal{E}}_3$,

$$\langle (\mathcal{P}_\Omega - \frac{n}{d^*}\mathcal{I})(\mathcal{T}-\mathcal{T}^*), \delta\mathcal{X}_i\rangle \leq C_m\kappa_0^4\mu^{m/2}\log^{m+2}(\bar{d})\left(\sqrt{\frac{n\bar{d}}{d^*}}+1\right)\frac{r^*\cdot r_{m-1}r_i^{-1/2}}{\sqrt{d^*}}\|\mathcal{T}-\mathcal{T}^*\|_{\mathrm{F}}.$$

When $i = m$. Similarly, we write

$$\langle (\mathcal{P}_\Omega - \frac{n}{d^*}\mathcal{I})(\mathcal{T}-\mathcal{T}^*), \delta\mathcal{X}_m\rangle = \sum_{j=1}^{m}\langle (\mathcal{P}_\Omega - \frac{n}{d^*}\mathcal{I})\mathcal{Y}_{m,j}, \delta\mathcal{X}_m\rangle.$$

First when $j \in [m-1]$, we have

$$|\langle (\mathcal{P}_\Omega - \frac{n}{d^*}\mathcal{I})(\mathcal{Y}_{m,j}), \delta\mathcal{X}_m\rangle| \leq C_m\left(\sqrt{\frac{n\bar{d}}{d^*}}+1\right)\log^{m+2}(\bar{d})\kappa_0^3\mu^{m/2}\frac{r^*\cdot r_{m-1}^{1/2}}{\sqrt{d^*}}\|\mathcal{T}-\mathcal{T}^*\|_{\mathrm{F}}. \tag{25}$$

And when $j = m$, we consider $\langle (\mathcal{P}_\Omega - \frac{n}{d^*}\mathcal{I})(\mathcal{Y}_{m,m}), \delta\mathcal{X}_m\rangle = \langle (\mathcal{P}_\Omega - \frac{n}{d^*}\mathcal{I})(\mathcal{Y}_{m,m}), \mathcal{X}_{m,0}\rangle + \sum_{k=1}^{m-1}\langle (\mathcal{P}_\Omega - \frac{n}{d^*}\mathcal{I})(\mathcal{Y}_{m,m}), \mathcal{X}_{m,k}\rangle$ as in (24). Similarly, using Lemma 8, we have

$$\begin{aligned}\langle (\mathcal{P}_\Omega - \frac{n}{d^*}\mathcal{I})(\mathcal{Y}_{m,m}), \mathcal{X}_{m,0}\rangle &\leq C_m\kappa_0\frac{\mu\bar{r}\sqrt{n\bar{d}\log(\bar{d})}}{d^*}\sqrt{r_{m-1}}\|\mathcal{T}-\mathcal{T}^*\|_{\mathrm{F}}\\ &\leq C_m\kappa_0\frac{\mu\bar{r}^{3/2}\sqrt{n\bar{d}\log(\bar{d})}}{d^*}\|\mathcal{T}-\mathcal{T}^*\|_{\mathrm{F}}.\end{aligned}$$

And when $k \in [m-1]$, we have

$$\langle (\mathcal{P}_\Omega - \frac{n}{d^*}\mathcal{I})(\mathcal{Y}_{m,m}), \mathcal{X}_{m,k}\rangle \leq C_m\kappa_0^3\mu^{m/2}\log^{m+2}(\bar{d})\left(\sqrt{\frac{n\bar{d}}{d^*}}+1\right)\frac{r^*\cdot r_{m-1}^{1/2}}{\sqrt{d^*}}\|\mathcal{T}-\mathcal{T}^*\|_{\mathrm{F}}.$$

So under the event $\boldsymbol{\mathcal{E}}_3$, we have

$$|\langle (\mathcal{P}_\Omega - \frac{n}{d^*}\mathcal{I})(\mathcal{Y}_{m,m}), \delta\mathcal{X}_m\rangle| \leq C_m\kappa_0^3\mu^{m/2}\log^{m+2}(\bar{d})\left(\sqrt{\frac{n\bar{d}}{d^*}}+1\right)\frac{r^*\cdot r_{m-1}^{1/2}}{\sqrt{d^*}}\|\mathcal{T}-\mathcal{T}^*\|_{\mathrm{F}},$$

and

$$\langle (\mathcal{P}_\Omega - \frac{n}{d^*}\mathcal{I})(\mathcal{T}-\mathcal{T}^*), \delta\mathcal{X}_m\rangle \leq C_m\kappa_0^3\mu^{m/2}\log^{m+2}(\bar{d})\left(\sqrt{\frac{n\bar{d}}{d^*}}+1\right)\frac{r^*\cdot r_{m-1}^{1/2}}{\sqrt{d^*}}\|\mathcal{T}-\mathcal{T}^*\|_{\mathrm{F}}. \tag{26}$$

Now we conclude from (25), (26) and under the event $\boldsymbol{\mathcal{E}}_3$,

$$\|\mathcal{P}_{\mathbb{T}}(\mathcal{P}_\Omega - \frac{n}{d^*}\mathcal{I})(\mathcal{T}-\mathcal{T}^*)\|_{\mathrm{F}} \leq C_m\kappa_0^4\mu^{m/2}\log^{m+2}(\bar{d})\left(\sqrt{\frac{n\bar{d}}{d^*}}+1\right)\frac{r^*r_{m-1}\cdot\sum_{i=1}^{m-1}r_i^{-1/2}}{\sqrt{d^*}}\|\mathcal{T}-\mathcal{T}^*\|_{\mathrm{F}}.$$

Together with (19), as long as

$$n \geq C_m \kappa_0^8 \mu^m \log^{2m+4}(\overline{d}) \cdot \overline{d}(r^*)^2 r_{m-1}^2 (\sum_{i=1}^{m-1} r_i^{-1}) + C_m \kappa_0^4 \mu^{m/2} \log^{m+2}(\overline{d}) \cdot (d^*)^{1/2} r^* r_{m-1} (\sum_{i=1}^{m-1} r_i^{-1/2}),$$

we have under the event $\boldsymbol{\mathcal{E}}_3$,

$$\|\mathcal{P}_{\mathbb{T}}\mathcal{P}_{\Omega}(\mathcal{T}-\mathcal{T}^*)\|_{\mathrm{F}}^2 \leq 1.002 \frac{n^2}{(d^*)^2} \|\mathcal{T}-\mathcal{T}^*\|_{\mathrm{F}}^2. \tag{27}$$

A.1.3 Estimation of $\langle \mathcal{P}_{\mathbb{T}}^{\perp}(\mathcal{T}-\mathcal{T}^*), \mathcal{P}_{\Omega}(\mathcal{T}-\mathcal{T}^*)\rangle$

Now we derive the upper bound for $\langle \mathcal{P}_{\mathbb{T}}^{\perp}(\mathcal{T}-\mathcal{T}^*), \mathcal{P}_{\Omega}(\mathcal{T}-\mathcal{T}^*)\rangle$. First we have

$$\langle \mathcal{P}_{\mathbb{T}}^{\perp}(\mathcal{T}-\mathcal{T}^*), \mathcal{P}_{\Omega}(\mathcal{T}-\mathcal{T}^*)\rangle = \langle \mathcal{P}_{\mathbb{T}}^{\perp}(\mathcal{T}-\mathcal{T}^*), (\mathcal{P}_{\Omega} - \frac{n}{d^*}\mathcal{I})(\mathcal{T}-\mathcal{T}^*)\rangle + \frac{n}{d^*}\|\mathcal{P}_{\mathbb{T}}^{\perp}\mathcal{T}^*\|_{\mathrm{F}}^2.$$

As in (20), we write $\mathcal{T}-\mathcal{T}^*$ as $\mathcal{T}-\mathcal{T}^* = \sum_{j=1}^m \mathcal{Y}_{m,j}$. Notice as estimated above (16), we have $\mathsf{rank}_{\mathsf{tt}}(\mathcal{P}_{\mathbb{T}}^{\perp}(\mathcal{T}-\mathcal{T}^*)) \leq (3r_1, \ldots, 3r_{m-1})$. So there exists a TT decomposition of $\mathcal{P}_{\mathbb{T}}^{\perp}(\mathcal{T}-\mathcal{T}^*) = [Y_1, \ldots, Y_{m-1}, Y_m]$ such that $\|Y_i\|_{\mathrm{F}} = \sqrt{3r_i}, i \in [m-1]$ and $\|Y_m\|_{\mathrm{F}} = \|\mathcal{P}_{\mathbb{T}}^{\perp}(\mathcal{T}-\mathcal{T}^*)\|_{\mathrm{F}}$.

When $j \in [m-1]$, we estimate $\langle \mathcal{P}_{\mathbb{T}}^{\perp}(\mathcal{T}-\mathcal{T}^*), (\mathcal{P}_{\Omega} - \frac{n}{d^*}\mathcal{I})\mathcal{Y}_{m,j}\rangle$. Using Lemma 9, Lemma 10, (21), (22), (23) and Lemma 19, and we have under the event $\boldsymbol{\mathcal{E}}_3$,

$$|\langle \mathcal{P}_{\mathbb{T}}^{\perp}(\mathcal{T}-\mathcal{T}^*), (\mathcal{P}_{\Omega} - \frac{n}{d^*}\mathcal{I})\mathcal{Y}_{m,j}\rangle| \leq C_m \left(\sqrt{\frac{n\overline{d}}{d^*}} + 1\right) \log^{m+2}(\overline{d}) \mu^{m/2} \kappa_0 \frac{r^* \cdot r_{m-1}^{1/2}}{\sqrt{d^*}} \|\mathcal{T}-\mathcal{T}^*\|_{\mathrm{F}}^2.$$

And when $j = m$, we have under the event $\boldsymbol{\mathcal{E}}_3$,

$$|\langle \mathcal{P}_{\mathbb{T}}^{\perp}(\mathcal{T}-\mathcal{T}^*), (\mathcal{P}_{\Omega} - \frac{n}{d^*}\mathcal{I})\mathcal{Y}_{m,m}\rangle| \leq C_m \left(\sqrt{\frac{n\overline{d}}{d^*}} + 1\right) \log^{m+2}(\overline{d}) \mu^{m/2} \kappa_0 \frac{r^* \cdot r_{m-1}^{1/2}}{\sqrt{d^*}} \|\mathcal{T}-\mathcal{T}^*\|_{\mathrm{F}}^2.$$

So we conclude under the event $\boldsymbol{\mathcal{E}}_3$,

$$\begin{aligned}
&|\langle \mathcal{P}_{\mathbb{T}}^{\perp}(\mathcal{T}-\mathcal{T}^*), \mathcal{P}_{\Omega}(\mathcal{T}-\mathcal{T}^*)\rangle| \\
&\leq C_m \left(\sqrt{\frac{n\overline{d}}{d^*}} + 1\right) \log^{m+2}(\overline{d}) \mu^{m/2} \kappa_0 \frac{r^* \cdot r_{m-1}^{1/2}}{\sqrt{d^*}} \|\mathcal{T}-\mathcal{T}^*\|_{\mathrm{F}}^2 + \frac{n}{d^*} \frac{300m^2 \|\mathcal{T}-\mathcal{T}^*\|_{\mathrm{F}}^4}{\sigma_{\min}^2(\mathcal{T}^*)} \\
&\leq \frac{n}{1000d^*} \|\mathcal{T}_l - \mathcal{T}^*\|_{\mathrm{F}}^2,
\end{aligned} \tag{28}$$

where the last inequality holds as long as

$$n \geq C_m \log^{m+2}(\overline{d}) \mu^{m/2} \kappa_0 r^* \overline{r}^{1/2} \sqrt{d^*} + C_m \overline{d}\overline{r}(r^*)^2 \mu^m \kappa_0^2 \log^{2m+4}(\overline{d}).$$

A.1.4 Contraction

Now we consider the error $\|\mathcal{T}_{l+1}-\mathcal{T}^*\|_{\mathrm{F}}$ assuming $\boldsymbol{\mathcal{E}}$ holds. From Algorithm 2, we have

$$\|\mathcal{T}_{l+1}-\mathcal{T}^*\|_{\mathrm{F}}^2=\|\mathsf{SVD}_{\boldsymbol{r}}^{\mathsf{tt}}(\widetilde{\mathcal{W}}_l)-\mathcal{T}^*\|_{\mathrm{F}}^2\overset{\text{Lemma 18}}{\leq}\|\widetilde{\mathcal{W}}_l-\mathcal{T}^*\|_{\mathrm{F}}^2+\frac{600m\|\widetilde{\mathcal{W}}_l-\mathcal{T}^*\|_{\mathrm{F}}^3}{\underline{\sigma}}. \tag{29}$$

From the way we choose ζ_l, we have $\|\widetilde{\mathcal{W}}_l-\mathcal{T}^*\|_{\mathrm{F}}\leq\|\mathcal{W}_l-\mathcal{T}^*\|_{\mathrm{F}}$. Now we estimate $\|\mathcal{W}_l-\mathcal{T}^*\|_{\mathrm{F}}^2$.

$$\begin{aligned}\|\mathcal{W}_l-\mathcal{T}^*\|_{\mathrm{F}}^2&=\|\mathcal{T}_l-\mathcal{T}^*-\alpha_l\mathcal{P}_{\mathbb{T}_l}\mathcal{P}_\Omega(\mathcal{T}_l-\mathcal{T}^*)\|_{\mathrm{F}}^2\\&=\|\mathcal{T}_l-\mathcal{T}^*\|_{\mathrm{F}}^2-2\alpha_l\underbrace{\langle\mathcal{T}_l-\mathcal{T}^*,\mathcal{P}_\Omega(\mathcal{T}_l-\mathcal{T}^*)\rangle}_{\mathrm{I}}\\&\quad+2\alpha_l\underbrace{\langle\mathcal{P}_{\mathbb{T}_l}^{\perp}(\mathcal{T}_l-\mathcal{T}^*),\mathcal{P}_\Omega(\mathcal{T}_l-\mathcal{T}^*)\rangle}_{\mathrm{II}}+\alpha_l^2\underbrace{\|\mathcal{P}_{\mathbb{T}_l}\mathcal{P}_\Omega(\mathcal{T}_l-\mathcal{T}^*)\|_{\mathrm{F}}^2}_{\mathrm{III}}.\end{aligned}$$

From (18),(28) and (27), when

$$\begin{aligned}n\geq C_m\Big(&\kappa_0^{2m+4}\nu^{m+1}\log^{m+2}(\overline{d})\cdot(d^*)^{1/2}((r^*)^{1/2}\overline{r}^2\vee r^*r_{m-1}\underline{r}^{-1/2}\vee r^*\overline{r}^{1/2})\\&+\kappa_0^{4m+8}\nu^{2m+2}\log^{2m+4}(\overline{d})\cdot\overline{d}(r^*\overline{r}^4\vee(r^*)^2r_{m-1}^2\underline{r}^{-1}\vee(r^*)^2\overline{r})\Big),\end{aligned}$$

we obtain

$$\mathrm{I}\geq\frac{6n}{25d^*}\|\mathcal{T}_l-\mathcal{T}^*\|_{\mathrm{F}}^2,\quad|\mathrm{II}|\leq\frac{n}{1000d^*}\|\mathcal{T}_l-\mathcal{T}^*\|_{\mathrm{F}}^2,\quad|\mathrm{III}|\leq1.002\frac{n^2}{(d^*)^2}\|\mathcal{T}_l-\mathcal{T}^*\|_{\mathrm{F}}^2.$$

From these estimations and we get,

$$\|\mathcal{W}_l-\mathcal{T}^*\|_{\mathrm{F}}^2\leq\|\mathcal{T}_l-\mathcal{T}^*\|_{\mathrm{F}}^2(1-0.239\alpha_l\frac{n}{d^*}+1.002\alpha_l^2\frac{n^2}{(d^*)^2}).$$

Now we set $\alpha_l=0.12\frac{d^*}{n}$, and we get $\|\mathcal{W}_l-\mathcal{T}^*\|_{\mathrm{F}}\leq0.986\|\mathcal{T}_l-\mathcal{T}^*\|_{\mathrm{F}}$. When $\|\mathcal{T}_l-\mathcal{T}^*\|_{\mathrm{F}}\leq\frac{\underline{\sigma}}{600000m}$, we have from (29)

$$\|\mathcal{T}_{l+1}-\mathcal{T}^*\|_{\mathrm{F}}^2\leq0.975\|\mathcal{T}_l-\mathcal{T}^*\|_{\mathrm{F}}^2.$$

And this implies $\|\mathcal{T}_{l+1}-\mathcal{T}^*\|_{\mathrm{F}}\leq\frac{\underline{\sigma}}{600000m\kappa_0\sqrt{\overline{r}}}$ and $\mathsf{Incoh}(\mathcal{T}_{l+1})\leq2\kappa_0^2\nu$ is implied by Lemma 20. So we finish the proof of Lemma 6.

A.2 Proof of Lemma 5

Before we start the proof, we give a detailed version of the theorem, notice here we set $|\Omega_i|=n_i$.

Lemma 11 (Restate of Lemma 5)**.** *Suppose the conditions of $\mathcal{T}^*$ from Theorem 1 hold. For any absolute constant $C > 0$, there exists an absolute constant $C_m > 0$ depending only on m such that if*

$$
\begin{aligned}
n = \sum_{i=1}^{2m-1} n_i \geq C_m \nu^{m+3} \kappa_0^{4m-4} \log^2(\overline{d}) \Bigg(& (d^*)^{1/2} (r^* \overline{r} \underline{r}^2)^{1/2} + \overline{d} r^* \overline{r} \underline{r}^2 \\
& + \sum_{k=1}^{m-2} \left((d_k \cdots d_m)^{1/2} (r_1 \cdots r_{k-1})^3 r_k (r_{k+1} \cdots r_{m-1})^{1/2} (\overline{r} \underline{r}^2)^{1/2} + \overline{d} (r_1 \cdots r_{k-1})^2 r^* \overline{r} \underline{r}^2 \right) \\
& + (\overline{d} r^* \underline{r} \overline{r} r_{m-1})^{1/2} + \overline{d} \overline{r} \underline{r} r_{m-1} + r^* \underline{r} \overline{r} r_{m-1} \Bigg),
\end{aligned}
$$

then with probability at least $1 - m\overline{d}^{-m}$, the output of Algorithm 3 satisfies

$$
\|\mathcal{T}_0 - \mathcal{T}^*\|_{\mathrm{F}} \leq \frac{\underline{\sigma}}{C m \kappa_0^2 \overline{r}^{1/2}} \qquad \textit{and} \qquad \mathsf{Incoh}(\mathcal{T}_0) \leq 2\kappa_0^2 \nu.
$$

Step 0: We denote $R_i = \arg\min_{R \in \mathbb{O}_{r_i}} \|\widehat{T}^{\leq i} - T^{*\leq i} R\|_{\mathrm{F}}$ and we set $\sqrt{\mu} = 2\kappa_0^2 \nu$.

Step 1: When $i = 1$.
Firstly from Wedin's $\sin\Theta$ theorem, and from Lemma 24, we see that when $n_1, n_2 \geq C_m \nu^2 \kappa_0^{2m-2} \cdot (d^*)^{1/2} (r^* \overline{r})^{1/2} \underline{r} \log^2(\overline{d}) + C_m \nu^4 \kappa_0^{4m-4} \overline{d} r^* \overline{r} \underline{r}^2$, we have

$$
d_p(\widetilde{T}_1, T_1^*) \leq \frac{2\sqrt{r_1} \|N_1 - N_1^*\|}{\underline{\sigma}^2} \leq (C_m \kappa_0^{2m-4} (r_2 \cdots r_{m-1})^{1/2} \overline{r}^{1/2})^{-1}.
$$

Now from Lemma 27, we know $\mathsf{Incoh}(\widehat{T}_1) \leq \sqrt{3\mu}$ and $d_c(\widehat{T}_1, T_1^*) \leq (C_m \kappa_0^{2m-4} (r_2 \cdots r_{m-1})^{1/2} \overline{r}^{1/2})^{-1}$.

Step 2: When $2 \leq i \leq m-1$.
Suppose we already have $\mathsf{Incoh}(\widehat{T}^{\leq i-1}) \leq (3\mu)^{(i-1)/2} (r_1 \cdots r_{i-2})^{3/2} =: \sqrt{\mu_{i-1}}$ and $d_c(\widehat{T}^{\leq i-1}, T^{*\leq i-1}) \leq (C_{i-1} m^2 \kappa_0^{2t_{i-1}} \sqrt{\overline{r} \cdot r_i})^{-1}$. From Lemma 25, we see that,

$$
\begin{aligned}
& d_p(L(\widetilde{T}_i), (R_{i-1} \otimes I)^T L(T_i^*)) \\
\leq & \frac{2\sqrt{r_i} \|(\widehat{T}^{\leq i-1} \otimes I)^T N_i (\widehat{T}^{\leq i-1} \otimes I) - (T^{*\leq i-1} R_{i-1} \otimes I)^T N_i^* (T^{*\leq i-1} R_{i-1} \otimes I)\|}{\underline{\sigma}^2}.
\end{aligned}
$$

Notice that

$$
\begin{aligned}
& \|(\widehat{T}^{\leq i-1} \otimes I)^T N_i (\widehat{T}^{\leq i-1} \otimes I) - (T^{*\leq i-1} R_{i-1} \otimes I)^T N_i^* (T^{*\leq i-1} R_{i-1} \otimes I)\| \\
\leq & \|(\widehat{T}^{\leq i-1} \otimes I)^T (N_i - N_i^*) (\widehat{T}^{\leq i-1} \otimes I)\| + 2\|\widehat{T}^{\leq i-1} - T^{*\leq i-1} R_{i-1}\| \cdot \overline{\sigma}^2.
\end{aligned}
$$

So we have

$$\begin{aligned} d_p(L(\widetilde{T}_i),(R_{i-1}\otimes I)^T L(T_i^*)) &\le 2\sqrt{r_i}\underline{\sigma}^{-2}\|(\widehat{T}^{\le i-1}\otimes I)^T(N_i-N_i^*)(\widehat{T}^{\le i-1}\otimes I)\| \\ &\quad + 4\sqrt{r_i}\kappa_0^2 d_c(\widehat{T}^{\le i-1},T^{*\le i-1}) \\ &=: b_{i-1}+4\sqrt{r_i}\kappa_0^2 x_{i-1}. \end{aligned} \tag{30}$$

Now we derive the chordal distance between $\widehat{T}^{\le i}$ and $T^{*\le i}$. Notice that

$$\begin{aligned} d_c(\widehat{T}^{\le i},T^{*\le i}) &\le \sqrt{r_i}d_c(\widehat{T}^{\le i-1},T^{*\le i-1}) + d_c(L(\widehat{T}_i),(R_{i-1}\otimes I)^T L(T_i^*)) \\ &\le \sqrt{r_i}d_c(\widehat{T}^{\le i-1},T^{*\le i-1}) + \sqrt{2}d_p(L(\widehat{T}_i),(R_{i-1}\otimes I)^T L(T_i^*)) \\ &\overset{\text{Lemma 27}}{\le} \sqrt{r_i}d_c(\widehat{T}^{\le i-1},T^{*\le i-1}) + 4\sqrt{2}\pi\cdot d_p(L(\widetilde{T}_i),(R_{i-1}\otimes I)^T L(T_i^*)) \end{aligned}$$

Together with (30), and denote by $q_{i-1}=80\sqrt{r_i}\kappa_0^2$, we have $x_i \le q_{i-1}x_{i-1}+b_{i-1}$. Sum this up and we have

$$x_{m-1} \le q_{m-1}\cdots q_1 x_1 + \sum_{k=1}^{m-2} q_{m-2}\cdots q_{k+1}b_k.$$

From Lemma 25, as long as

$$\begin{aligned} n_{2k+1},n_{2k+2} \ge C_m\nu^{k+2}\kappa_0^{2m}\log^2(\overline{d})\cdot(d_k\cdots d_m)^{1/2}(r_1\cdots r_{k-1})^3 r_k(r_{k+1}\cdots r_{m-1})^{1/2}(\overline{r}\underline{r}^2)^{1/2} \\ + C_m\nu^{k+4}\kappa_0^{4m-2k}\log(\overline{d})\cdot\overline{d}(r_1\cdots r_{k-1})^3 r_k r_{k+1}\cdots r_{m-1}(\overline{r}\underline{r}^2), \end{aligned}$$

we have $q_{m-2}\cdots q_{k+1}b_k \le \frac{1}{Cm^2\kappa_0^2\sqrt{\overline{r}}}$. Together with the estimation in Step 1, we have

$$d_c(\widehat{T}^{\le m-1},T^{*\le m-1}) = x_{m-1} \le (Cm\kappa_0^2\sqrt{\overline{r}})^{-1}. \tag{31}$$

Step 3: When $i=m$. We have

$$\begin{aligned} \|\mathcal{T}^*-\widehat{\mathcal{T}}\|_{\mathrm{F}} &= \|T^{*\le m-1}R_{m-1}R_{m-1}^T T_m^* - \widehat{T}^{\le m-1}\widehat{T}_m\|_{\mathrm{F}} \\ &\le \|(T^{*\le m-1}R_{m-1}-\widehat{T}^{\le m-1})R_{m-1}^T T_m^*\|_{\mathrm{F}} + \|R_{m-1}^T T_m^* - \widehat{T}_m\|_{\mathrm{F}} \\ &\le d_c(T^{*\le m-1},\widehat{T}^{\le m-1})\cdot\overline{\sigma} + \|R_{m-1}^T T_m^* - \widehat{T}_m\|_{\mathrm{F}}. \end{aligned}$$

Notice $d_c(T^{*\le m-1},\widehat{T}^{\le m-1})$ is estimated in (31). On the other hand, we have

$$\begin{aligned} \|R_{m-1}^T T_m^* - \widehat{T}_m\|_{\mathrm{F}} &= \|(T^{*\le m-1}R_{m-1})^T(\mathcal{T}^*)^{\langle m-1\rangle} - (\widehat{T}^{\le m-1})^T(\frac{d^*}{n_{2m-1}}\mathcal{P}_{\Omega_{2m-1}}(\mathcal{T}^*))^{\langle m-1\rangle}\|_{\mathrm{F}} \\ &\le \|(\widehat{T}^{\le m-1})^T((\mathcal{T}^*)^{\langle m-1\rangle} - (\frac{d^*}{n_{2m-1}}\mathcal{P}_{\Omega_{2m-1}}(\mathcal{T}^*))^{\langle m-1\rangle})\|_{\mathrm{F}} \\ &\quad + \|(T^{*\le m-1}R_{m-1}-\widehat{T}^{\le m-1})^T(\mathcal{T}^*)^{\langle m-1\rangle}\|_{\mathrm{F}} \\ &\le \|(\widehat{T}^{\le m-1})^T((\mathcal{T}^*)^{\langle m-1\rangle} - (\frac{d^*}{n_{2m-1}}\mathcal{P}_{\Omega_{2m-1}}(\mathcal{T}^*))^{\langle m-1\rangle})\|_{\mathrm{F}} \\ &\quad + d_c(T^{*\le m-1},\widehat{T}^{\le m-1})\cdot\overline{\sigma}. \end{aligned} \tag{32}$$

Together with Lemma 26, we have as long as

$$n_{2m-1} \geq C_m \nu^{(m+1)/2} \kappa_0^{m+1} \log(\overline{d}) \cdot (\overline{d} r^* \underline{r} \overline{r} r_{m-1})^{1/2} + C_m \nu \kappa_0^4 \log(\overline{d}) \cdot \overline{d} \underline{r} \overline{r} r_{m-1} + C_m \nu^m \kappa_0^{2m+2} \log(\overline{d}) \cdot r^* \underline{r} \overline{r} r_{m-1},$$

we have

$$\|(\widehat{T}^{\leq m-1})^T((\mathcal{T}^*)^{\langle m-1 \rangle} - (\frac{d^*}{n_{2m-1}} \mathcal{P}_{\Omega_{2m-1}}(\mathcal{T}^*))^{\langle m-1 \rangle})\|_{\mathrm{F}} \leq \frac{\underline{\sigma}}{Cm\kappa_0 \sqrt{\overline{r}}}.$$

From (31) - (32) and we conclude with probability exceeding $1 - m\overline{d}^{-m}$,

$$\|\mathcal{T}^* - \widehat{\mathcal{T}}\|_{\mathrm{F}} \leq \frac{\underline{\sigma}}{Cm\kappa_0 \sqrt{\overline{r}}}.$$

Finally, together with Lemma 20, we conclude that the output $\mathcal{T}_0$ satisfies

$$\|\mathcal{T}_0 - \mathcal{T}^*\|_{\mathrm{F}} \leq \frac{\underline{\sigma}}{Cm\kappa_0^2 \sqrt{\overline{r}}} \text{ and } \mathsf{Incoh}(\mathcal{T}_0) \leq 2\kappa_0^2 \nu.$$

And this finishes the proof of the lemma.

B Technical Lemmas

In this section, we provide some technical lemmas. Some proofs for the lemmas are placed to the next section.

B.1 Lemmas about linear algebra

Lemma 12. *Let $N \in \mathbb{R}^{d_N \times d_1 \cdots d_i}$ and $M \in \mathbb{R}^{d_{i+2} \ldots d_m \times d_M}$, then we have the following relations:*

$$\begin{aligned}
&\mathsf{reshape}(N\mathcal{T}^{\langle i \rangle}, [d_N d_{i+1}, d_{i+2} \cdots d_m]) = (N \otimes I_{d_{i+1}})\mathcal{T}^{\langle i+1 \rangle}, \\
&\mathsf{reshape}\left((N \otimes I_{d_{i+1}})\mathcal{T}^{\langle i+1 \rangle}, [d_N, d_{i+1} \cdots d_m]\right) = N\mathcal{T}^{\langle i \rangle}, \\
&\mathsf{reshape}(\mathcal{T}^{\langle i+1 \rangle} M, [d_1 \ldots d_i, d_{i+1} d_M]) = \mathcal{T}^{\langle i \rangle}(I_{d_{i+1}} \otimes M), \\
&\mathsf{reshape}(\mathcal{T}^{\langle i \rangle}(I_{d_{i+1}} \otimes M), [d_1 \ldots d_{i+1}, d_M]) = \mathcal{T}^{\langle i+1 \rangle} M.
\end{aligned}$$

Proof. We first show the first equation. For all $x_N \in [d_N]$ and $x_j \in [d_j]$, we have

$$\begin{aligned}
N\mathcal{T}^{\langle i \rangle}(x_N; x_{i+1} \cdots x_m) &= \sum_{x_1, \cdots, x_i} N(x_N; x_1, \cdots, x_i)\mathcal{T}(x_1, \cdots, x_m) \\
&= \sum_{x_1, \cdots, x_i, x'_{i+1}} N(x_N; x_1, \cdots, x_i) I(x_{i+1}, x'_{i+1})\mathcal{T}(x_1, \cdots, x'_{i+1}, \cdots, x_m) \\
&= \sum_{x_1, \cdots, x_i, x'_{i+1}} (N \otimes I_{d_{i+1}})(x_N, x_{i+1}; x_1, \cdots, x_i, x'_{i+1})\mathcal{T}(x_1, \cdots, x'_{i+1}, \cdots, x_m).
\end{aligned}$$

So from this we get the desired result. Now the second equation is just the inverse statement of the first one. And the third and fourth equations are similar to the first one. □

Lemma 13. *For any matrix $A \in \mathbb{R}^{n\times m}, x \in \mathbb{R}^n$, we have*

$$\|A^T x\|_{\ell_2} \le \sqrt{n}\|A\|_{2,\infty}\|x\|_{\ell_2}.$$

Let another matrix $B \in \mathbb{R}^{p\times m}$, then we have

$$\|AB^T\|_{\ell_\infty} \le \sqrt{m}\|A\|_{\ell_\infty}\|B\|_{2,\infty}.$$

When we have another matrix $X \in \mathbb{R}_{m\times m}$ and $m \le p, m \le n$, we have

$$\|AXB^T\|_{\ell_\infty} \le \|X\| \cdot \|A\|_{2,\infty}\|B\|_{2,\infty}.$$

Proof. Set $A = \begin{bmatrix} a_1^T \\ \vdots \\ a_n^T \end{bmatrix} \in \mathbb{R}^{n\times m}$, then $A^T x = \sum_{i=1}^n x_i a_i$. And from Cauchy-Schwartz inequality, we have

$$\|A^T x\|_{\ell_2}^2 \le \|x\|_{\ell_2}^2 \cdot \sum_{i=1}^n \|a_i\|_{\ell_2}^2 \le n\|x\|_{\ell_2}^2\|A\|_{2,\infty}^2.$$

Also set $B = \begin{bmatrix} b_1^T \\ \vdots \\ b_p^T \end{bmatrix}$, then

$$\|AB^T\|_{\ell_\infty} = \max_{j,k} |a_j^T b_k| \le \max_{j,k} \|a_j\|_{\ell_\infty}\|b_k\|_{\ell_1} \le \sqrt{m}\|A\|_{\ell_\infty} \max_k \|b_k\|_{\ell_2} \le \sqrt{m}\|A\|_{\ell_\infty} \cdot \|B\|_{2,\infty}.$$

Finally, let $X = U\Sigma V^T$ be its SVD,

$$\begin{aligned}\|AXB^T\|_{\ell_\infty} &\le \|AX\|_{2,\infty}\|B\|_{2,\infty} = \|AU\Sigma\|_{2,\infty}\|B\|_{2,\infty} \\ &\le \|X\|\|AU\|_{2,\infty}\|B\|_{2,\infty} = \|X\|\|A\|_{2,\infty}\|B\|_{2,\infty}.\end{aligned}$$

□

Lemma 14. *Let $A_1, A_2 \in \mathbb{R}^{m\times n}$ be two rank r matrices with the decomposition $A_1 = U_1\Sigma_1 V_1^T, A_2 = U_2\Sigma_2 V_2^T$ such that $U_1^T U_1 = U_2^T U_2 = V_1^T V_1 = V_2^T V_2 = I_r$ and $\Sigma_1, \Sigma_2 \in \mathbb{R}^{r\times r}$ be invertible and U_1, U_2, V_1, V_2 are well aligned in the sense that $d_c(U_1, U_2) = \|U_1 - U_2\|_{\mathrm{F}}$ and $d_c(V_1, V_2) = \|V_1 - V_2\|_{\mathrm{F}}$. Suppose that $\|A_1 - A_2\|_{\mathrm{F}} \le \frac{1}{10}\min\{\sigma_{\min}(A_1), \sigma_{\min}(A_2)\}$, then we have*

$$\|\Sigma_1 - \Sigma_2\|_{\mathrm{F}} \le 4\sqrt{r}\kappa\|A_1 - A_2\|,$$

where $\kappa = \max\{\kappa(A_1), \kappa(A_2)\}$ and $\sigma_{\min}(A)$ is the smallest nonzero singular value of A.

Proof. We can decompose $\Sigma_1 - \Sigma_2$ as follows,

$$\begin{aligned}\|\Sigma_1 - \Sigma_2\|_{\mathrm{F}} &\le \|(U_1 - U_2)^T A_1 V_1\|_{\mathrm{F}} + \|U_2^T (A_1 - A_2) V_1\|_{\mathrm{F}} + \|U_2^T A_2 (V_1 - V_2)\|_{\mathrm{F}} \\ &\le (\sqrt{r} + \sqrt{2r}\kappa(A_1) + \sqrt{2r}\kappa(A_2))\|A_1 - A_2\| \le 4\sqrt{r}\kappa\|A_1 - A_2\|,\end{aligned}$$

where the first inequality in the second line follows Wedin's sinΘ theorem. □

Lemma 15. *Let $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{np\times q}$ be such that $A^TA = I_n, B^TB = I_q$ and $\mathsf{Incoh}(A) \le \sqrt{\mu_1}, \mathsf{Incoh}(B) \le \sqrt{\mu_2}$. Then we have $\mathsf{Incoh}((A \otimes I_p)B) \le \sqrt{\mu_1\mu_2 n}$.*

Proof. Consider for any $k \in [mp]$, $\|e_k^T(A \otimes I_p)B\|_{\ell_2}$, denote by a_k^T the k-th row of $A \otimes I_p$, and thus $\|a_k\|_{\ell_0} \le n$. So we have

$$\|e_k^T(A \otimes I_p)B\|_{\ell_2} = \|a_k^T B\|_{\ell_2} \le \sqrt{\frac{\mu_1 n}{m}}\sqrt{n\frac{\mu_2 q}{np}} = \sqrt{\mu_1\mu_2 n\frac{q}{mp}}.$$

□

B.2 Lemmas concerning TT format

Lemma 16 (Facts about TT rank)**.** *1. Let $\mathcal{A}, \mathcal{B} \in \mathbb{R}^{d_1\times\ldots\times d_m}$ be two tensors satisfies $\mathsf{rank}_{tt}(\mathcal{A}) = (r_1, \ldots, r_{m-1})$ and $\mathsf{rank}_{tt}(\mathcal{B}) = (s_1, \ldots, s_{m-1})$. Then we have*

$$\mathsf{rank}_{tt}(\mathcal{A} + \mathcal{B}) \le (r_1 + s_1, \ldots, r_{m-1} + s_{m-1}).$$

2. Let $\mathcal{T}^* \in \mathbb{M}_{\boldsymbol{r}}^{tt}$ and $\mathbb{T}^*$ be the corresponding tangent space. Let $\mathcal{T} \in \mathbb{R}^{d_1\times\ldots\times d_m}$ be an arbitrary tensor. Then the rank of $\mathcal{P}_{\mathbb{T}^*}(\mathcal{T})$ satisfies $\mathsf{rank}_{tt}(\mathcal{P}_{\mathbb{T}^*}(\mathcal{T})) \le (2r_1, \ldots, 2r_{m-1})$.

Proof. 1. It follows from $\mathsf{rank}((\mathcal{A} + \mathcal{B})^{\langle i\rangle}) \le \mathsf{rank}(\mathcal{A}^{\langle i\rangle}) + \mathsf{rank}(\mathcal{B}^{\langle i\rangle}) = r_i + s_i$.

2. Since $\mathcal{P}_{\mathbb{T}^*}(\mathcal{T}) = \delta\mathcal{T}_1 + \ldots + \delta\mathcal{T}_m$, where $\delta\mathcal{T}_i = [T_1^*, \ldots, T_{i-1}^*, X_i, T_{i+1}^*, \ldots, T_m^*]$ and the expression of X_i are give in (6). Now we consider the $i-th$ separation rank of $\mathcal{P}_{\mathbb{T}^*}(\mathcal{T})$. Notice that for all $j \le i$, we have $\delta\mathcal{T}_j^{\langle i\rangle} = T_j^{*\le i}T^{*\ge i+1}$, where $T_j^{*\le i} \in \mathbb{R}^{d_1\ldots d_i\times r_i}$ is the matrix generated by $T_1^*, \ldots, X_j, \ldots, T_i^*$. And for $j \ge i+1$, we have $\delta\mathcal{T}_j^{\langle i\rangle} = T^{*\le i}T_j^{*\ge i+1}$, where $T_j^{*\ge i} \in \mathbb{R}^{r_i\times d_{i+1}\ldots d_m}$ is the matrix generated by $T_{i+1}^*, \ldots, X_j, \ldots, T_m^*$. So we have

$$\mathcal{P}_{\mathbb{T}^*}(\mathcal{T})^{\langle i\rangle} = (\sum_{j=1}^{i} T_j^{*\le i})T^{*\ge i+1} + T^{*\le i}(\sum_{j=i+1}^{m} T_j^{*\ge i+1}).$$

And thus $\mathsf{rank}(\mathcal{P}_{\mathbb{T}^*}(\mathcal{T})^{\langle i\rangle}) \le 2r_i$.

□

Lemma 17. *Let $\mathcal{T} \in \mathbb{M}_{\boldsymbol{r}}^{\boldsymbol{tt}}$ be a tensor of TT rank $(r_1, \ldots, r_{m-1})$ with a left orthogonal decomposition $\mathcal{T} = [T_1, \ldots, T_m]$. Then we have*

$$\|\mathcal{T}\|_* \leq \sqrt{r_1 \ldots r_{m-1}} \|\mathcal{T}\|_{\mathrm{F}}.$$

Proof. Using the alternative definition for the tensor nuclear norm from in Friedland and Lim (2014), we have

$$\|\mathcal{T}\|_* = \min\{\sum_{i=1}^{s} |\lambda_i| : \mathcal{T} = \sum_{i=1}^{s} \lambda_i u_{1,i} \otimes \ldots \otimes u_{m,i}, \|u_{l,i}\|_{\ell_2} = 1, l \in [m], i \in [s], s \in \mathbb{N}\}. \tag{33}$$

So we can write $\mathcal{T}$ as sum of rank one tensors in the following form

$$\mathcal{T} = \sum_{k_1=1}^{r_1} \cdots \sum_{k_{m-1}=1}^{r_{m-1}} T_1(\cdot, k_1) \otimes T_2(k_1, \cdot, k_2) \otimes \ldots \otimes T_m(k_{m-1}, \cdot),$$

where each $T_i(k_{i-1}, \cdot, k_i) \in \mathbb{R}^{d_i}$ is a vector. For each fixed $k_1, \ldots, k_{m-1}$, we have

$$\|T_1(\cdot, k_1) \otimes T_2(k_1, \cdot, k_2) \otimes \ldots \otimes T_m(k_{m-1}, \cdot)\|_{\mathrm{F}} = \prod_{i=1}^{m} \|T_i(k_{i-1}, \cdot, k_i)\|_{\ell_2}.$$

And

$$\begin{aligned}
\Big(\sum_{k_1=1}^{r_1} \cdots \sum_{k_{m-1}=1}^{r_{m-1}} \prod_{i=1}^{m} \|T_i(k_{i-1}, \cdot, k_i)\|_{\ell_2}\Big)^2 &\leq r_1 \ldots r_{m-1} \sum_{k_1=1}^{r_1} \cdots \sum_{k_{m-1}=1}^{r_{m-1}} \prod_{i=1}^{m} \|T_i(k_{i-1}, \cdot, k_i)\|_{\ell_2}^2 \\
&\overset{(a)}{=} r_1 \ldots r_{m-1} \cdot \sum_{k_1, \ldots, k_{m-1}} \prod_{i=2}^{m} \|T_i(k_{i-1}, \cdot, k_i)\|_{\ell_2}^2 \\
&= \ldots \\
&\overset{(b)}{=} r_1 \ldots r_{m-1} \cdot \sum_{k_{m-1}} \|T_m(k_{m-1}, \cdot)\|_{\ell_2}^2 = r_1 \ldots r_{m-1} \|\mathcal{T}\|_{\mathrm{F}}^2,
\end{aligned}$$

where (a) holds since we have $\|T_1(\cdot, k_1)\|_{\ell_2} = 1$ since $L(T_1)L(T_1)^T = I$ and (b) holds for similar reason and we use $L(T_i)^T L(T_i) = I_{r_i}$ for all $i \in [m-1]$. Together with $\|T_m\|_{\mathrm{F}} = \|\mathcal{T}\|_{\mathrm{F}}$ and (33), we finish the proof. □

Lemma 18 (Perturbation bound for TT SVD)**.** *Let $\mathcal{T}^* \in \mathbb{M}_{\boldsymbol{r}}^{\boldsymbol{tt}}$ be the m-way tensor and $\underline{\sigma} := \min_{i=1}^{m-1} \sigma_{\min}(\mathcal{T}^*)^{\langle i \rangle}$. And we denote the tensor $\mathcal{T} = \mathcal{T}^* + \mathcal{D}$. Then suppose $\underline{\sigma} \geq C_m \|\mathcal{D}\|_{\mathrm{F}}$ for some constant $C_m \geq 500m$ depending only on m, we have*

$$\|\mathsf{SVD}_{\boldsymbol{r}}^{\boldsymbol{tt}}(\mathcal{T}) - \mathcal{T}^*\|_{\mathrm{F}}^2 \leq \|\mathcal{D}\|_{\mathrm{F}}^2 + \frac{600m\|\mathcal{D}\|_{\mathrm{F}}^3}{\underline{\sigma}}.$$

Proof. See Section C.6. □

Lemma 19. *Let $\mathcal{T}, \mathcal{T}^* \in \mathbb{M}_{\boldsymbol{r}}^{tt}$ be two TT-rank $\boldsymbol{r}$ tensors. Suppose we have $8\|\mathcal{T}-\mathcal{T}^*\|_{\mathrm{F}} \le \underline{\sigma}$, then we have*

$$\|\mathcal{P}_{\mathbb{T}}^{\perp}(\mathcal{T}^*)\|_{\mathrm{F}} \le \frac{12\sqrt{2}m\|\mathcal{T}-\mathcal{T}^*\|_{\mathrm{F}}^2}{\underline{\sigma}},$$

where $\mathbb{T}$ is the tangent space at the point $\mathcal{T}$.

Proof. See Section C.7. □

Interchanging the roles of $\mathcal{T}$ and $\mathcal{T}^*$ in the theorem and using Weyl's inequality and we get the following corollary.

Corollary 2. *Under the setting of Lemma 19, let $\mathbb{T}^*$ be the corresponding tangent space at $\mathcal{T}^*$, then we have*

$$\|\mathcal{P}_{\mathbb{T}^*}^{\perp}(\mathcal{T})\|_{\mathrm{F}} \le \frac{20m\|\mathcal{T}-\mathcal{T}^*\|_{\mathrm{F}}^2}{\underline{\sigma}}.$$

Lemma 20 (TTSVD + Trim implies incoherence)**.** *Let $\mathcal{T}^* \in \mathbb{M}_{\boldsymbol{r}}^{tt}$ be such that $\mathsf{Spiki}(\mathcal{T}^*) \le \nu$. Suppose that $\mathcal{W}$ satisfies $\|\mathcal{W}-\mathcal{T}^*\|_{\mathrm{F}} \le \frac{\underline{\sigma}}{600m\sqrt{\bar{r}}\kappa_0}$, then we have $\mathsf{Incoh}\left(\mathsf{SVD}_{\boldsymbol{r}}^{tt}(\mathsf{Trim}_{\zeta}(\mathcal{W}))\right) \le 2\kappa_0^2\nu$ if we choose $\zeta = \frac{10\|\mathcal{W}\|_{\mathrm{F}}}{9\sqrt{d^*}}\nu$. Furthermore,*

$$\|\mathsf{SVD}_{\boldsymbol{r}}^{tt}(\mathsf{Trim}_{\zeta}(\mathcal{W})) - \mathcal{T}^*\|_{\mathrm{F}} \le \sqrt{2}\|\mathcal{W}-\mathcal{T}^*\|_{\mathrm{F}}.$$

Proof. See Section C.8. □

B.3 Concentration inequalities

First we introduce some operators for the following lemma. For all $x \in [d_1] \times \ldots \times [d_m]$, let the operator $\mathcal{P}_x : \mathbb{R}^{d_1 \times \ldots \times d_m} \to \mathbb{R}^{d_1 \times \ldots \times d_m}$ be defined by $\mathcal{P}_x(\mathcal{X}) = \langle \mathcal{X}, \mathcal{E}_x \rangle \mathcal{E}_x$. And let $\Omega = \{\omega_1, \ldots, \omega_n\}$ be the sampling set and define $\mathcal{P}_\Omega : \mathbb{R}^{d_1 \times \ldots \times d_m} \to \mathbb{R}^{d_1 \times \ldots \times d_m}$ by $\mathcal{P}_\Omega(\mathcal{X}) = \sum_{i=1}^m \langle \mathcal{X}, \mathcal{E}_{\omega_i} \rangle \mathcal{E}_{\omega_i}$.

Lemma 21 (Concentration inequality)**.** *Suppose Ω with $|\Omega| = n$ is a set of indices sampled independently and uniformly with replacement. Suppose $\mathsf{Spiki}(\mathcal{T}^*) \le \nu$ for some $\nu > 0$. When $n \ge C_m(\nu\kappa_0)^4 \bar{d}\bar{r}^2 \log(d^*)$ for some large constant $C_1 > 0$, with probability exceeding $1 - (\bar{d})^{-m}$, we have*

$$\|\frac{d^*}{n}\mathcal{P}_{\mathbb{T}^*}\mathcal{P}_\Omega\mathcal{P}_{\mathbb{T}^*} - \mathcal{P}_{\mathbb{T}^*}\| \le \frac{1}{2}.$$

Proof. First we define the operators:

$$\mathcal{Z}_x := \frac{d^*}{n}\mathcal{P}_{\mathbb{T}^*}\mathcal{P}_x\mathcal{P}_{\mathbb{T}^*} - \frac{1}{n}\mathcal{P}_{\mathbb{T}^*}, \quad x \in [d_1] \times \ldots \times [d_m].$$

Then $\frac{d^*}{n}\mathcal{P}_{\mathbb{T}^*}\mathcal{P}_\Omega\mathcal{P}_{\mathbb{T}^*} - \mathcal{P}_{\mathbb{T}^*} = \sum_{i=1}^n \mathcal{Z}_{\omega_i}$. We first estimate an upper bound for $\|\mathcal{Z}_{\omega_i}\|$. First notice

$$(\mathcal{P}_{\mathbb{T}^*}\mathcal{P}_x\mathcal{P}_{\mathbb{T}^*})^2(\mathcal{X}) = \|\mathcal{P}_{\mathbb{T}^*}(\mathcal{E}_x)\|_{\mathrm{F}}^2(\mathcal{P}_{\mathbb{T}^*}\mathcal{P}_x\mathcal{P}_{\mathbb{T}^*})(\mathcal{X}).$$

This implies $\|\mathcal{P}_{\mathbb{T}^*}\mathcal{P}_x\mathcal{P}_{\mathbb{T}^*}\| \le \|\mathcal{P}_{\mathbb{T}^*}(\mathcal{E}_x)\|_{\mathrm{F}}^2$. And

$$\|\mathcal{Z}_x\| \le \frac{d^*}{n}\|\mathcal{P}_{\mathbb{T}^*}\mathcal{P}_x\mathcal{P}_{\mathbb{T}^*}\| + \frac{1}{n} \le \frac{d^*}{n}\|\mathcal{P}_{\mathbb{T}^*}(\mathcal{E}_x)\|_{\mathrm{F}}^2 + \frac{1}{n} \le \frac{2m(\nu\kappa_0)^4\bar{d}\bar{r}^2}{n},$$

where the last inequality we use $\max_x \|\mathcal{P}_{\mathbb{T}^*}(\mathcal{E}_x)\|_{\mathrm{F}}^2 \le \frac{m(\nu\kappa_0)^4\bar{d}\bar{r}^2}{n}$, which comes from Lemma 2. Now we bound $\|\mathbb{E}\sum_{i=1}^n \mathcal{Z}_{\omega_i}^2\|$. Simple calculations show that

$$\mathbb{E}\sum_{i=1}^n \mathcal{Z}_{\omega_i}^2 = \frac{d^*}{n}\sum_{x\in[d_1]\times\ldots\times[d_m]}(\mathcal{P}_{\mathbb{T}^*}\mathcal{P}_x\mathcal{P}_{\mathbb{T}^*})^2 - \frac{1}{n}\mathcal{P}_{\mathbb{T}^*}.$$

And this implies

$$\begin{aligned}\|\mathbb{E}\sum_{i=1}^n \mathcal{Z}_{\omega_i}^2\| &\le \frac{d^*}{n}\|\sum_x(\mathcal{P}_{\mathbb{T}^*}\mathcal{P}_x\mathcal{P}_{\mathbb{T}^*})^2\| + \frac{1}{n}\\ &\overset{(a)}{\le} \frac{d^*}{n}\max_x\|\mathcal{P}_{\mathbb{T}^*}(\mathcal{E}_x)\|_{\mathrm{F}}^2\cdot\|\sum_x\mathcal{P}_{\mathbb{T}^*}\mathcal{P}_x\mathcal{P}_{\mathbb{T}^*}\| + \frac{1}{n}\\ &\le \frac{2m(\nu\kappa_0)^4\bar{d}\bar{r}^2}{n},\end{aligned}$$

where in (a) we use the fact that $(\mathcal{P}_{\mathbb{T}^*}\mathcal{P}_x\mathcal{P}_{\mathbb{T}^*})^2 \le \|\mathcal{P}_{\mathbb{T}^*}(\mathcal{E}_x)\|_{\mathrm{F}}^2\mathcal{P}_{\mathbb{T}^*}\mathcal{P}_x\mathcal{P}_{\mathbb{T}^*}$, where $\mathcal{A}\le\mathcal{B}$ means $\mathcal{B}-\mathcal{A}$ is an SPD operator. So we conclude using operator Bernstein inequality,

$$\|\frac{d^*}{n}\mathcal{P}_{\mathbb{T}^*}\mathcal{P}_\Omega\mathcal{P}_{\mathbb{T}^*} - \mathcal{P}_{\mathbb{T}^*}\| \le C\left(\sqrt{\frac{m(\nu\kappa_0)^4\bar{d}\bar{r}^2\log(d^*)}{n}} + \frac{m(\nu\kappa_0)^4\bar{d}\bar{r}^2\log(d^*)}{n}\right)$$

with probability exceeding $1-(\bar{d})^{-m}$. When $n \ge C_m(\nu\kappa_0)^4\bar{d}\bar{r}^2\log(d^*)$, the right hand side is less than $1/2$. And this finishes the proof of the lemma. □

Lemma 22 (Xia et al. 2021, Theorem 1)**.** *Suppose Ω is the set sampled uniformly with replacement with size $|\Omega| = n$, then we have with probability exceeding $1-(\bar{d})^{-m}$, the following holds*

$$\|(\mathcal{P}_\Omega - \frac{n}{d^*}\mathcal{I})(\mathcal{J})\| \le C_m\left(\sqrt{\frac{n\bar{d}}{d^*}}+1\right)\log^{m+2}(\bar{d}),$$

where $\mathcal{J}\in\mathbb{R}^{d_1\times\cdots\times d_m}$ is the tensor with all its entries 1.

Lemma 23. *Let $\Omega = \{\omega_i : i = 1,\cdots,n\}$ where ω_i is independently and uniformly sampled from the set of collections $[d_1]\times\cdots\times[d_m]$. With probability at least $1-\bar{d}^{-m}$, the maximum number of repetitions of any entry in Ω is less than $2m\log(\bar{d})$.*

Proof. This is a result of standard Chenroff bound. See for example (Recht 2011, Proposition 5) for a proof. □

B.4 Lemmas for initialization

Lemma 24 (Xia and Yuan 2019, Theorem 2). *Let $M \in \mathbb{R}^{p_1 \times p_2}$ and $X_i = p_1 p_2 \mathcal{P}_{\omega_i} M$, $Y_j = p_1 p_2 \mathcal{P}_{\omega_j'} M$, where $\omega_i \in \Omega_1, \omega_j' \in \Omega_2$ are independently and uniformly sampled from $[p_1] \times [p_2]$ and $|\Omega_1| = |\Omega_2| = n$. Denote by $N = MM^T$ and $\widetilde{N} = \frac{1}{2n^2} \sum_{i,j} (X_i Y_j^T + Y_j X_i^T)$, then with probability exceeding $1 - p^{-\alpha}$ with $p = \max\{p_1, p_2\}$, we have*

$$\|\widetilde{N} - N\| \le C\alpha^2 \frac{p_1^{3/2} p_2^{3/2} \log(p)}{n} \left[\left(1 + \frac{p_1}{p_2}\right)^{1/2} + \frac{p_1^{1/2} p_2^{1/2}}{n} + \left(\frac{n}{p_2 \log(p)}\right)^{1/2} \right] \cdot \|M\|_\infty^2.$$

Lemma 25. *Let $M \in \mathbb{R}^{p_1 \times p_2}$ and $X_i = p_1 p_2 \mathcal{P}_{\omega_i} M, Y_j = p_1 p_2 \mathcal{P}_{\omega_j'} M$, where $\omega_i \in \Omega_1, \omega_j' \in \Omega_2$ are independently and uniformly sampled from $[p_1] \times [p_2]$ and $|\Omega_1| = |\Omega_2| = n$. Let $U \in \mathbb{R}^{p_1 \times r}$ be the orthogonal matrix such that $\mathsf{Incoh}(U) \le \sqrt{\mu}$. Then with probability exceeding $1 - p^{-\alpha}$, we have*

$$\begin{aligned} &\|\frac{1}{2n^2} \sum_{i,j} (U^T X_i Y_j^T U + U^T Y_j X_i^T U) - U^T M M^T U\| \\ &\le C\alpha^2 \log^2(p) \frac{p_1 p_2 \|M\|_\infty^2}{n} \left(\mu r p_2^{1/2} + \frac{\mu r p_2}{n} + (\frac{\mu r n}{\log^3(p)})^{1/2} \right). \end{aligned}$$

Proof. See Section C.9. □

When $\mathsf{Spiki}(M) \le \nu$, we have

$$\begin{aligned} &\|\frac{1}{2n^2} \sum_{i,j} (U^T X_i Y_j^T U + U^T Y_j X_i^T U) - U^T M M^T U\| \\ &\le C\alpha^2 \log^2(p) \nu^2 \|M\|_{\mathrm{F}}^2 \left(\frac{\mu r p_2^{1/2}}{n} + \frac{\mu r p_2}{n^2} + (\frac{\mu r}{n \log^3(p)})^{1/2} \right) \end{aligned}$$

holds with probability exceeding $1 - p^{-\alpha}$.

Lemma 26. *Let $M \in \mathbb{R}^{p_1 \times p_2}$ and $X_i = p_1 p_2 \mathcal{P}_{\omega_i}(M)$, where $\omega_i \in \Omega$ is independently and uniformly sampled in $[p_1] \times [p_2]$ and $|\Omega| = n$. Let $U \in \mathbb{R}^{p_1 \times r}$ be the orthogonal matrix such that $\mathsf{Incoh}(U) \le \sqrt{\mu}$. Then with probability exceeding $1 - p^{-\alpha}$, we have*

$$\|U^T (\frac{p_1 p_2}{n} \mathcal{P}_\Omega(M) - M)\| \le C\alpha \left(\frac{\sqrt{p_1} p_2 \sqrt{\mu r} \|M\|_\infty \log(p)}{n} + \sqrt{\frac{p_1 p_2 \|M\|_\infty^2 (\mu r \vee p_2) \log(p)}{n}} \right).$$

Proof. See appendix C.10. □

Lemma 27 (Keshavan et al. 2010, Remark 6.2). *Let $U, X \in \mathbb{R}^{p \times r}$ be orthogonal and $\mathsf{Incoh}(U) \le \sqrt{\mu_0}$ and $d_p(U, X) \le \delta \le \frac{1}{16\pi}$. Then $\bar{X}$ satisfies $\mathsf{Incoh}(\widehat{X}) \le \sqrt{3\mu_0}$ and $d_p(\widehat{X}, U) \le 4\pi\delta$, where*

$$\overline{X}^i = \frac{X^i}{\|X^i\|_{\ell_2}} \cdot \min\{\|X^i\|_{\ell_2}, \sqrt{\frac{\mu r}{p}}\}, \quad \widehat{X} = \overline{X}(\overline{X}^T \overline{X})^{-1/2}.$$

C Proofs of technical lemmas

In this section, we provide the proofs for the technical lemmas.

C.1 Proof of Lemma 2

Since $(\mathcal{T}^*)^{\langle i\rangle} = T^{*\leq i}\Lambda^*_{i+1}V^{*T}_{i+1}$, we have $T^{*\leq i} = (\mathcal{T}^*)^{\langle i\rangle}V^*_{i+1}(\Lambda^*_{i+1})^{-1}$. And thus for any $k \in [d_1 \dots d_i]$,

$$\|e_k^T T^{*\leq i}\|_{\ell_2} = \|e_k^T(\mathcal{T}^*)^{\langle i\rangle}V^*_{i+1}(\Lambda^*_{i+1})^{-1}\|_{\ell_2} \leq \frac{1}{\sigma_{\min}(\Lambda^*_{i+1})}\|e_k^T(\mathcal{T}^*)^{\langle i\rangle}\|_{\ell_2} \leq \frac{\sqrt{d_{i+1}\dots d_m}}{\sigma_{\min}(\Lambda^*_{i+1})}\|\mathcal{T}^*\|_{\ell_\infty}.$$

And the spikiness condition implies $\|\mathcal{T}^*\|_{\ell_\infty} \leq \frac{\nu}{\sqrt{d^*}}\|\mathcal{T}^*\|_{\mathrm{F}}$, and together with $\|\mathcal{T}^*\|_{\mathrm{F}} \leq \sqrt{r_i}\sigma_1(\Lambda^*_{i+1})$, we obtain

$$\|e_k^T T^{*\leq i}\|_{\ell_2}\frac{\sqrt{d_1\dots d_i}}{\sqrt{r_i}} \leq \nu\kappa_0.$$

The bound for $\|e_k^T V^*_{i+1}\|_{\ell_2}$ can be similarly derived. And this finishes the proof.

C.2 Proof of Lemma 7

For simplicity denote the random tensor which is uniformly distributed in $\{\mathcal{E}_\omega\}_{\omega\in[d^*]}$ by $\mathcal{E}$ and let $\mathcal{E}_1, \dots, \mathcal{E}_n$ be n i.i.d. copies of $\mathcal{E}$. Also define $\delta_{1,j} = 2^j\delta_1^-, j = 0, \dots, \lceil\log_2(\frac{\delta_1^+}{\delta_1^-})\rceil =: j_0$, $\delta_{2,k} = 2^k\delta_2^-, k = 0, \dots, \lceil\log_2(\frac{\delta_2^+}{\delta_2^-})\rceil =: k_0$. For each j, k, we derive an upper bound for $\beta_n(\gamma_1, \gamma_2)$ with $\gamma_1 = \delta_{1,j}, \gamma_2 = \delta_{2,k}$.

We observe that

$$\sup_{\mathcal{A}\in\mathbb{K}_{\gamma_1,\gamma_2}} \left|\langle\mathcal{A},\mathcal{E}\rangle^2 - \mathbb{E}\langle\mathcal{A},\mathcal{E}\rangle^2\right| \leq \gamma_1^2,$$

and

$$\sup_{\mathcal{A}\in\mathbb{K}_{\gamma_1,\gamma_2}} \mathrm{Var}\langle\mathcal{A},\mathcal{E}\rangle^2 \leq \sup_{\mathcal{A}\in\mathbb{K}_{\gamma_1,\gamma_2}} \langle\mathcal{A},\mathcal{E}\rangle^4 \leq \frac{\gamma_1^2\|\mathcal{A}\|_{\mathrm{F}}^2}{d^*} \leq \frac{\gamma_1^2}{d^*}.$$

Now apply Bousquet's version of Talagrand concentration inequality (see Giné and Nickl 2021, Theorem 3.3.9), and we get with probability at least $1 - e^{-t}$ for any $t > 0$,

$$\beta_n(\gamma_1,\gamma_2) \leq 2\mathbb{E}\beta_n(\gamma_1,\gamma_2) + 2\gamma_1\sqrt{\frac{nt}{d^*}} + 2\gamma_1^2 t.$$

Using symmetric inequality, we have

$$\mathbb{E}\beta_n(\gamma_1,\gamma_2) \leq 2n\mathbb{E}\sup_{\mathcal{A}\in\mathbb{K}_{\gamma_1,\gamma_2}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i\langle\mathcal{A},\mathcal{E}_i\rangle^2\right|,$$

where $\epsilon_1, \dots, \epsilon_n$ are i.i.d. Rademacher random variables. Since $|\langle\mathcal{A},\mathcal{E}\rangle| \leq \gamma_1$, we have from contraction inequality

$$\mathbb{E}\beta_n(\gamma_1,\gamma_2) \leq 4n\gamma_1\mathbb{E}\sup_{\mathcal{A}\in\mathbb{K}_{\gamma_1,\gamma_2}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i\langle\mathcal{A},\mathcal{E}_i\rangle\right|.$$

Now we denote $\mathcal{Y} = \frac{1}{n}\sum_{i=1}^{n}\epsilon_i\mathcal{E}_i$. Then we have

$$\mathbb{E}\sup_{\mathcal{A}\in\mathbb{K}_{\gamma_1,\gamma_2}}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\langle\mathcal{A},\mathcal{E}_i\rangle\right| \le \mathbb{E}\sup_{\mathcal{A}\in\mathbb{K}_{\gamma_1,\gamma_2}}\|\mathcal{Y}\|\|\mathcal{A}\|_* \le \gamma_2\mathbb{E}\|\mathcal{Y}\|.$$

The estimation for $\|\mathcal{Y}\|$ derived in the Theorem 1 of Xia et al. (2021) gives

$$\mathbb{E}\|\mathcal{Y}\| \le C_m\left(\sqrt{\frac{\overline{d}}{nd^*}}+\frac{1}{n}\right)\log^{m+2}(\overline{d}).$$

As a result, with probability exceeding $1-e^{-t}$, we have

$$\beta_n(\gamma_1,\gamma_2) \le C_m\gamma_1\gamma_2\left(\sqrt{\frac{n\overline{d}}{d^*}}+1\right)\log^{m+2}(\overline{d}) + 2\gamma_1\sqrt{\frac{nt}{d^*}} + 2\gamma_1^2 t. \tag{34}$$

Now we take union bound and we get with probability exceeding $1-2\log_2\left(\frac{\delta_1^+}{\delta_1^-}\right)\log_2\left(\frac{\delta_2^+}{\delta_2^-}\right)e^{-t}$, and for all $\gamma_1 \in \{\delta_{1,0},\ldots,\delta_{1,j_0}\}, \gamma_2 \in \{\delta_{2,0},\ldots,\delta_{2,k_0}\}$, (34) holds. Now we consider arbitrary $\gamma_1 \in [\delta_1^-,\delta_1^+], \gamma_2 \in [\delta_2^-,\delta_2^+]$. Then there exists some j,k, such that $\gamma_1 \in [\delta_{1,j-1},\delta_{1,j}], \gamma_2 \in [\delta_{2,k-1},\delta_{2,k}]$. Together with the fact that $\delta_{1,j} \le 2\gamma_1$ and $\delta_{2,k} \le 2\gamma_2$ we get

$$\beta_n(\gamma_1,\gamma_2) \le \beta_n(\delta_{1,j},\delta_{2,k}) \le C_m\gamma_1\gamma_2\left(\sqrt{\frac{n\overline{d}}{d^*}}+1\right)\log^{m+2}(\overline{d}) + 4\gamma_1\sqrt{\frac{nt}{d^*}} + 8\gamma_1^2 t.$$

This finishes the proof of the lemma.

C.3 Proof of Lemma 8

Let $\omega \in [d_1]\times\ldots\times[d_m]$, then

$$\|\mathcal{P}^{(i)}\mathcal{E}_\omega\|_{\mathrm{F}}^2 \le \begin{cases} \frac{\mu r_1 d_1}{d^*}, & i=1 \\ \frac{\mu^2 r_{i-1}r_i d_i}{d^*}, & 2\le i\le m-1 \\ \frac{\mu r_{m-1}d_m}{d^*}, & i=m \end{cases} \le \frac{\mu^2\overline{r}^2\overline{d}}{d^*}.$$

Set $\mathcal{Z}_j = \mathcal{P}^{(i)}(\mathcal{P}_{\omega_j}-\frac{1}{d^*}\mathcal{I})\mathcal{P}^{(i)}$ for all $j\in[n]$, then $\mathcal{P}^{(i)}(\mathcal{P}_\Omega-\frac{n}{d^*}\mathcal{I})\mathcal{P}^{(i)} = \sum_{j=1}^{n}\mathcal{Z}_j$. First for arbitrary $\mathcal{X}$, we have

$$(\mathcal{P}^{(i)}\mathcal{P}_\omega\mathcal{P}^{(i)})^2\mathcal{X} = \|\mathcal{P}^{(i)}\mathcal{E}_\omega\|_{\mathrm{F}}^2\cdot\mathcal{P}^{(i)}\mathcal{P}_\omega\mathcal{P}^{(i)}\mathcal{X}.$$

Therefore $\|\mathcal{P}^{(i)}\mathcal{P}_\omega\mathcal{P}^{(i)}\| \le \|\mathcal{P}^{(i)}\mathcal{E}_\omega\|_{\mathrm{F}}^2$ and this implies that

$$\|\mathcal{Z}_j\| \le \max_\omega\|\mathcal{P}^{(i)}\mathcal{E}_\omega\|_{\mathrm{F}}^2 + \frac{1}{d^*} \le \frac{2\mu^2\overline{r}^2\overline{d}}{d^*}.$$

On the other hand, since $\mathcal{Z}_j$ is an symmetric operator, we consider $\mathbb{E}\sum_{j=1}^n \mathcal{Z}_j^2$.

$$\|\mathbb{E}\sum_{j=1}^n \mathcal{Z}_j^2\| = n\|\frac{1}{d^*}\sum_{\omega}(\mathcal{P}^{(i)}\mathcal{P}_\omega\mathcal{P}^{(i)})^2 - \frac{1}{(d^*)^2}\mathcal{P}^{(i)}\|$$
$$\leq \frac{n}{d^*}\max_{\omega}\|\mathcal{P}^{(i)}\mathcal{E}_\omega\|_{\mathrm{F}}^2 + \frac{n}{(d^*)^2} \leq \frac{2\mu^2\overline{r}^2 n\overline{d}}{(d^*)^2}.$$

Now using operator Bernstein inequality, with probability exceeding $1-\overline{d}^{-m}$,

$$\|\mathcal{P}^{(i)}(\mathcal{P}_\Omega - \frac{n}{d^*}\mathcal{I})\mathcal{P}^{(i)}\| \leq C_m\left(\frac{\mu^2\overline{r}^2\overline{d}\log(\overline{d})}{d^*} + \sqrt{\frac{\mu^2\overline{r}^2\overline{d}n\log(\overline{d})}{(d^*)^2}}\right) \leq C_m\sqrt{\frac{\mu^2\overline{r}^2\overline{d}n\log(\overline{d})}{(d^*)^2}},$$

where the last inequality holds as long as $n \geq C\mu^2\overline{r}^2\overline{d}\log(\overline{d})$.

C.4 Proof of Lemma 9

Notice that

$$\langle(\mathcal{P}_\Omega - \frac{n}{d^*}\mathcal{I})\mathcal{A},\mathcal{B}\rangle = \langle(\mathcal{P}_\Omega - \frac{n}{d^*}\mathcal{I})\mathcal{J},\mathcal{A}\odot\mathcal{B}\rangle \leq \|(\mathcal{P}_\Omega - \frac{n}{d^*}\mathcal{I})\mathcal{J}\|\cdot\|\mathcal{A}\odot\mathcal{B}\|_*.$$

where $\mathcal{J}$ is the tensor with all its entries one. And from the definition of nuclear norm in (33), we have

$$\|\mathcal{A}\odot\mathcal{B}\|_*^2 \leq \left(\sum_{\substack{k_1,\ldots,k_{m-1}\\ k_1',\ldots,k_{m-1}'}} \|A_1(:,k_1)\odot B_1(:,k_1')\|_{\ell_2}\ldots\|A_m(k_{m-1},:)\odot B_1(k_{m-1}',:)\|_{\ell_2}\right)^2$$
$$\leq \sum_{x_1}\|A_1(x_1,:)\|_{\mathrm{F}}^2\|B_1(x_1,:)\|_{\mathrm{F}}^2\ldots\sum_{x_m}\|A_m(:,x_m)\|_{\mathrm{F}}^2\|B_m(:,x_m)\|_{\mathrm{F}}^2,$$

where the last inequality comes from using Cauchy-Schwartz inequality $m-1$ times. Since $\boldsymbol{\mathcal{E}}_3$ holds and

$$\sum_{x_i}\|A_i(:,x_i,:)\|_{\mathrm{F}}^2\|B_i(:,x_i,:)\|_{\mathrm{F}}^2 \leq \max_{x_i}\|A_i(:,x_i,:)\|_{\mathrm{F}}^2\cdot\|B_i\|_{\mathrm{F}}^2 \wedge \max_{x_i}\|B_i(:,x_i,:)\|_{\mathrm{F}}^2\cdot\|A_i\|_{\mathrm{F}}^2$$

we get the desired result.

C.5 Proof of Lemma 10

Consider for all $i\in[m-1]$, notice that $L(T_i) = (T^{\leq i-1}\otimes I)^T T^{\leq i}$. We have

$$L(T_i)(k_{i-1},x_i;k_i) = \sum_{y_{i-1}\in[d_1\cdots d_{i-1}]} T^{\leq i-1}(y_{i-1},k_i)T^{\leq i}(y_{i-1},x_i;k_i).$$

This implies

$$\|T_i(:,x_i,:)\|_{\mathrm{F}}^2 = \|(T^{\leq i-1})^T \cdot T^{\leq i}(:,x_i,:)\|_{\mathrm{F}}^2 \leq \|T^{\leq i}(:,x_i,:)\|_{\mathrm{F}}^2 \leq \frac{\mu r_i}{d_i},$$

where $T^{\leq i}(:,x_i,:)$ is viewed as a matrix of size $d_1\cdots d_{i-1} \times r_i$ by extracting $d_1\cdots d_{i-1}$ rows of $T^{\leq i}$. And since the decomposition is left orthogonal, we have $\|T_i\|_{\mathrm{F}} = \|L(T_i)\|_{\mathrm{F}} = r_i$.

When $i = m$, since $T_m = \Lambda_m R_m^T$ for some $\Lambda_m \in \mathbb{R}^{r_{m-1}\times r_{m-1}}$ invertible and orthogonal R_m with $\mathsf{Incoh}(R_m) \leq \sqrt{\mu}$. So we have

$$\max_{x_m} \|T_m(:,x_m)\|_{\mathrm{F}}^2 \leq \sigma_{\max}^2(\mathcal{T})\frac{\mu r_{m-1}}{d_m}.$$

And $\|\mathcal{T}\|_{\mathrm{F}} = \|\mathcal{T}^{\langle m-1\rangle}\|_{\mathrm{F}} = \|T^{\leq m-1}T_m\|_{\mathrm{F}} = \|T_m\|_{\mathrm{F}}$ since $T^{\leq m-1}$ is orthogonal.

C.6 Proof of Lemma 18

Denote $\widehat{\mathcal{T}} = \mathsf{SVD}_{\boldsymbol{r}}^{\mathsf{tt}}(\mathcal{T}) = [\widehat{T}_1,\ldots,\widehat{T}_m]$. Let $\mathcal{T}^* = [T_1^{**},\ldots,T_m^{**}]$ be a TT decomposition that is left orthogonal and $R_i = \arg\min_{R\in\mathbb{O}_{r_i}} \|(T^{**})^{\leq i}R - \widehat{T}^{\leq i}\|_{\mathrm{F}}$. Now we set $L(T_i^*) = (R_{i-1}\otimes I)^T L(T_i^{**})R_i$. Then $\mathcal{T}^* = [T_1^*,\ldots,T_m^*]$ is another left orthogonal decomposition so that $T^{*\leq i}$ and $\widehat{T}^{\leq i}$ are well aligned in terms of chordal distance. Also, let $(\mathcal{T}^*)^{\langle i\rangle} = T^{*\leq i}\Lambda_{i+1}V_{i+1}^{*T}$ be such that $V_{i+1}^{*T}V_{i+1}^* = I_{r_i}$ and $\Lambda_{i+1}\in\mathbb{R}^{r_i\times r_i}$ be invertible.

From Algorithm 1, we have $(\widehat{T}^{\leq m-1})^T\mathcal{T}^{\langle m-1\rangle} = \widehat{T}_m$. Now using the notations $\widehat{\mathcal{P}}_i = \widehat{T}^{\leq m-1}\widehat{T}^{\leq m-1T}$ and $\mathcal{P}_i^* = T^{*\leq i}T^{*\leq iT}$, we have

$$\begin{aligned}
\|\mathsf{SVD}_{\boldsymbol{r}}^{\mathsf{tt}}(\mathcal{T}) - \mathcal{T}^*\|_{\mathrm{F}} &= \|\widehat{\mathcal{T}}^{\langle m-1\rangle} - (\mathcal{T}^*)^{\langle m-1\rangle}\|_{\mathrm{F}} = \|(\widehat{\mathcal{P}}_{m-1}-I)(\mathcal{T}^*)^{\langle m-1\rangle} + \widehat{\mathcal{P}}_{m-1}\mathcal{D}^{\langle m-1\rangle}\|_{\mathrm{F}} \\
&\overset{(a)}{=} \|(\widehat{\mathcal{P}}_{m-1}-\mathcal{P}_{m-1}^*)(\mathcal{T}^*)^{\langle m-1\rangle} + \mathcal{P}_{m-1}^*\mathcal{D}^{\langle m-1\rangle} + (\widehat{\mathcal{P}}_{m-1}-\mathcal{P}_{m-1}^*)\mathcal{D}^{\langle m-1\rangle}\|_{\mathrm{F}} \\
&\overset{(b)}{=} \| \underbrace{(I-\mathcal{P}_{m-1}^*)\Delta_{m-1}V_m^*V_m^{*T} + \mathcal{P}_{m-1}^*\mathcal{D}^{\langle m-1\rangle}}_{\text{(I.1)}} \\
&\quad + \underbrace{H(\widehat{\mathcal{P}}_{m-1},\mathcal{P}_{m-1}^*)(\mathcal{T}^*)^{\langle m-1\rangle} + (\widehat{\mathcal{P}}_{m-1}-\mathcal{P}_{m-1}^*)\mathcal{D}^{\langle m-1\rangle}}_{\text{high order terms=:(I.2)}}\|_{\mathrm{F}},
\end{aligned} \tag{35}$$

where in (a) we use the fact that $(I-\mathcal{P}_{m-1}^*)(\mathcal{T}^*)^{\langle m-1\rangle} = 0$ and (b) follows from the following equation when $i = m-1$, which is a result of Lemma 22,

$$\widehat{\mathcal{P}}_i - \mathcal{P}_i^* = T^{*\leq i}(\Lambda_{i+1}^*)^{-1}(V_{i+1}^*)^T\Delta_i^T(I-\mathcal{P}_i^*) + (I-\mathcal{P}_i^*)\Delta_i V_{i+1}^*(\Lambda_{i+1}^*)^{-1}(T^{*\leq i})^T + H(\widehat{\mathcal{P}}_i,\mathcal{P}_i^*),$$

and since $T^{*\leq i}$ is the top r_i left singular vectors of $(\mathcal{T}^*)^{\langle i\rangle}$ and $\widehat{T}^{\leq i}$ is the top r_i left singular vectors

of $(\widehat{\mathcal{P}}_{i-1} \otimes I)\mathcal{T}^{\langle i \rangle}$, so $\Delta_i = (\widehat{\mathcal{P}}_{i-1} \otimes I)\mathcal{T}^{\langle i \rangle} - (\mathcal{P}^*_{i-1} \otimes I)(\mathcal{T}^*)^{\langle i \rangle}$, and thus

$$\begin{aligned}
\Delta_i &= \left((\widehat{\mathcal{P}}_{i-1} - \mathcal{P}^*_{i-1}) \otimes I_{d_i}\right)(\mathcal{T}^*)^{\langle i \rangle} + \left(\widehat{\mathcal{P}}_{i-1} \otimes I_{d_i}\right)\mathcal{D}^{\langle i \rangle} \\
&= \left((\widehat{\mathcal{P}}_{i-1} - \mathcal{P}^*_{i-1}) \otimes I_{d_i}\right)(\mathcal{T}^*)^{\langle i \rangle} + \left(\mathcal{P}^*_{i-1} \otimes I_{d_i}\right)\mathcal{D}^{\langle i \rangle} + \left((\widehat{\mathcal{P}}_{i-1} - \mathcal{P}^*_{i-1}) \otimes I_{d_i}\right)\mathcal{D}^{\langle i \rangle} \\
&= \left((I - \mathcal{P}^*_{i-1})\Delta_{i-1}V_i^*(\Lambda_i^*)^{-1}(T^{*\leq i-1})^T \otimes I_{d_i}\right)(\mathcal{T}^*)^{\langle i \rangle} + \left(\mathcal{P}^*_{i-1} \otimes I_{d_i}\right)\mathcal{D}^{\langle i \rangle} \\
&\quad + \underbrace{\left(H(\widehat{\mathcal{P}}_{i-1}, \mathcal{P}^*_{i-1}) \otimes I_{d_i}\right)(\mathcal{T}^*)^{\langle i \rangle} + \left((\widehat{\mathcal{P}}_{i-1} - \mathcal{P}^*_{i-1}) \otimes I_{d_i}\right)\mathcal{D}^{\langle i \rangle}}_{\text{high order terms}=:H_i} \\
&=: L_i + H_i, \qquad (36)
\end{aligned}$$

here we use L_i and H_i to denote the leading terms and high order terms of Δ_i respectively. Now we derive first the upper bound for $\|H_i\|_{\mathrm{F}}$.

Bound for $\|H_i\|_{\mathrm{F}}$. Notice by triangle inequality, we have

$$\|H_i\|_{\mathrm{F}} \leq \| \left(H(\widehat{\mathcal{P}}_{i-1}, \mathcal{P}^*_{i-1}) \otimes I_{d_i}\right)(\mathcal{T}^*)^{\langle i \rangle}\|_{\mathrm{F}} + \| \left((\widehat{\mathcal{P}}_{i-1} - \mathcal{P}^*_{i-1}) \otimes I_{d_i}\right)\mathcal{D}^{\langle i \rangle}\|_{\mathrm{F}}.$$

From the assumption we have $\sigma_{\min}(\Lambda_i^*) \geq 8\|\Delta_{i-1}\|_{\mathrm{F}}$, we have

$$\| \left(H(\widehat{\mathcal{P}}_{i-1}, \mathcal{P}^*_{i-1}) \otimes I_{d_i}\right)(\mathcal{T}^*)^{\langle i \rangle}\|_{\mathrm{F}} \leq \frac{12\|\Delta_{i-1}\|_{\mathrm{F}}^2}{\sigma_{\min}(\Lambda_i^*)},$$

and

$$\| \left((\widehat{\mathcal{P}}_{i-1} - \mathcal{P}^*_{i-1}) \otimes I_{d_i}\right)\mathcal{D}^{\langle i \rangle}\|_{\mathrm{F}} \leq \frac{4\|\mathcal{D}\|_{\mathrm{F}}\|\Delta_{i-1}\|_{\mathrm{F}}}{\sigma_{\min}(\Lambda_i^*)}.$$

So combine these two estimations, and we have

$$\|H_i\|_{\mathrm{F}} \leq \frac{12\|\Delta_{i-1}\|_{\mathrm{F}}^2}{\sigma_{\min}(\Lambda_i^*)} + \frac{4\|\mathcal{D}\|_{\mathrm{F}}\|\Delta_{i-1}\|_{\mathrm{F}}}{\sigma_{\min}(\Lambda_i^*)}. \qquad (37)$$

Upper bound for $\|\Delta_i\|_{\mathrm{F}}$. We first show the following for all $2 \leq i \leq m-1$.

$$\begin{aligned}
&\|(I - \mathcal{P}^*_i)\Delta_i\|_{\mathrm{F}}^2 - \|(I - \mathcal{P}^*_{i-1})\Delta_{i-1}\|_{\mathrm{F}}^2 \\
&\leq \|(T^{*\leq i-1} \otimes I_{d_i})(I - L(T_i^*)L(T_i^*)^T)((T^{*\leq i-1})^T \otimes I_{d_i})\mathcal{D}^{\langle i \rangle}\|_{\mathrm{F}}^2 + 2\|(I - \mathcal{P}^*_i)L_i\|_{\mathrm{F}}\|H_i\|_{\mathrm{F}} + \|H_i\|_{\mathrm{F}}^2. \qquad (38)
\end{aligned}$$

As a consequence of $\mathcal{P}^*_i = (U^{*\leq i-1} \otimes I_{d_i})L(T_i^*)L(T_i^*)^T((T^{*\leq i-1})^T \otimes I_{d_i})$, we have

$$\mathcal{P}^*_i \left((I - \mathcal{P}^*_{i-1})\Delta_{i-1}V_i^*(\Lambda_i^*)^{-1}(T^{*\leq i-1})^T \otimes I_{d_i}\right)(\mathcal{T}^*)^{\langle i \rangle} = 0.$$

So the leading term of $(I-\mathcal{P}_i^*)\Delta_i$ is

$$\begin{aligned}
(I-\mathcal{P}_i^*)L_i &= \left((I-\mathcal{P}_{i-1}^*)\Delta_{i-1}V_i^*(\Lambda_i^*)^{-1}(T^{*\le i-1})^T\otimes I_{d_i}\right)(\mathcal{T}^*)^{\langle i\rangle} + (I-\mathcal{P}_i^*)\left(\mathcal{P}_{i-1}^*\otimes I_{d_i}\right)\mathcal{D}^{\langle i\rangle} \\
&\overset{(a)}{=} \left((I-\mathcal{P}_{i-1}^*)\Delta_{i-1}V_i^*(\Lambda_i^*)^{-1}(T^{*\le i-1})^T\otimes I_{d_i}\right)(\mathcal{T}^*)^{\langle i\rangle} \\
&\quad + (T^{*\le i-1}\otimes I_{d_i})(I-L(T_i^*)L(T_i^*)^T)((T^{*\le i-1})^T\otimes I_{d_i})\mathcal{D}^{\langle i\rangle} \qquad (39)\\
&\overset{(b)}{=} \mathsf{reshape}\left((I-\mathcal{P}_{i-1}^*)\Delta_{i-1}V_i^*V_i^{*T}\right) + (T^{*\le i-1}\otimes I_{d_i})(I-L(T_i^*)L(T_i^*)^T)((T^{*\le i-1})^T\otimes I_{d_i})\mathcal{D}^{\langle i\rangle},
\end{aligned}$$

where in (a) we use

$$\begin{aligned}
&(I-\mathcal{P}_i^*)\left(\mathcal{P}_{i-1}^*\otimes I_{d_i}\right) \\
&= \left(I-(T^{*\le i-1}\otimes I_{d_i})L(T_i^*)L(T_i^*)^T((T^{*\le i-1})^T\otimes I_{d_i})\right)\left(T^{*\le i-1}(T^{*\le i-1})^T\otimes I_{d_i}\right) \\
&= T^{*\le i-1}(T^{*\le i-1})^T\otimes I_{d_i} - (T^{*\le i-1}\otimes I_{d_i})L(T_i^*)L(T_i^*)^T(T^{*\le i-1}\otimes I_{d_i})^T \\
&= (T^{*\le i-1}\otimes I_{d_i})(I-L(T_i^*)L(T_i^*)^T)(T^{*\le i-1}\otimes I_{d_i})^T.
\end{aligned}$$

And in (b) we use Lemma 12. Notice the two terms in (39) are mutually orthogonal, so we have

$$\begin{aligned}
&\|(I-\mathcal{P}_i^*)L_i\|_{\mathrm{F}}^2 \\
&= \|(I-\mathcal{P}_{i-1}^*)\Delta_{i-1}V_i^*V_i^{*T}\|_{\mathrm{F}}^2 + \|(T^{*\le i-1}\otimes I_{d_i})(I-L(T_i^*)L(T_i^*)^T)((T^{*\le i-1})^T\otimes I_{d_i})\mathcal{D}^{\langle i\rangle}\|_{\mathrm{F}}^2 \\
&\le \|(I-\mathcal{P}_{i-1}^*)\Delta_{i-1}\|_{\mathrm{F}}^2 + \|(T^{*\le i-1}\otimes I_{d_i})(I-L(T_i^*)L(T_i^*)^T)((T^{*\le i-1})^T\otimes I_{d_i})\mathcal{D}^{\langle i\rangle}\|_{\mathrm{F}}^2. \qquad (40)
\end{aligned}$$

Meanwhile, from (36), we have

$$\|(I-\mathcal{P}_i^*)\Delta_i\|_{\mathrm{F}}^2 \le \|(I-\mathcal{P}_i^*)L_i\|_{\mathrm{F}}^2 + 2\|(I-\mathcal{P}_i^*)L_i\|_{\mathrm{F}}\|H_i\|_{\mathrm{F}} + \|H_i\|_{\mathrm{F}}^2. \qquad (41)$$

Combine (40), (41) and we get (38).

Upper bound for $\|(I-\mathcal{P}_k^*)\Delta_k\|_{\mathrm{F}}^2 + \|\mathcal{P}_k^*\mathcal{D}^{\langle k\rangle}\|_{\mathrm{F}}^2$. From the recurrence relation (38), we have

$$\begin{aligned}
&\|(I-\mathcal{P}_k^*)\Delta_k\|_{\mathrm{F}}^2 + \|\mathcal{P}_k^*\mathcal{D}^{\langle k\rangle}\|_{\mathrm{F}}^2 \\
&\le \|\mathcal{P}_k^*\mathcal{D}^{\langle k\rangle}\|_{\mathrm{F}}^2 + \sum_{i=2}^{k}\|(T^{*\le i-1}\otimes I_{d_i})(I-L(T_i^*)L(T_i^*)^T)((T^{*\le i-1})^T\otimes I_{d_i})\mathcal{D}^{\langle i\rangle}\|_{\mathrm{F}}^2 \\
&\quad + \|(I-\mathcal{P}_1^*)\mathcal{D}^{\langle 1\rangle}\|_{\mathrm{F}} + \underbrace{\sum_{i=2}^{k}\left(2\|(I-\mathcal{P}_i^*)L_i\|_{\mathrm{F}}\|H_i\|_{\mathrm{F}} + \|H_i\|_{\mathrm{F}}^2\right)}_{\text{high order terms}=:\xi_{k,2}} \\
&=: \xi_{k,1} + \xi_{k,2}.
\end{aligned}$$

Now we first show $\xi_{k,1} = \|\mathcal{D}\|_{\mathrm{F}}^2$. The key point of the proof lies in the following equation:

$$\|\mathcal{P}_i^*\mathcal{D}^{\langle i\rangle}\|_{\mathrm{F}}^2 + \|(T^{*\le i-1}\otimes I_{d_i})(I-L(T_i^*)L(T_i^*)^T)((T^{*\le i-1})^T\otimes I_{d_i})\mathcal{D}^{\langle i\rangle}\|_{\mathrm{F}}^2 = \|\mathcal{P}_{i-1}^*\mathcal{D}^{\langle i-1\rangle}\|_{\mathrm{F}}^2. \qquad (42)$$

Suppose this holds, then we have the left hand side is equal to $\|\mathcal{P}_1^*\mathcal{D}^{\langle 1\rangle}\|_{\mathrm{F}}^2+\|(I-\mathcal{P}_1^*)\mathcal{D}^{\langle 1\rangle}\|_{\mathrm{F}}^2=\|\mathcal{D}\|_{\mathrm{F}}^2$ and this finishes the proof. So now we verify (42).

$$\begin{aligned}\text{LHS of (42)} &= \|(T^{*\leq i-1}\otimes I_{d_i})L(T_i^*)L(T_i^*)^T((T^{*\leq i-1})^T\otimes I_{d_i})\mathcal{D}^{\langle i\rangle}\|_{\mathrm{F}}^2 \\ &\quad + \|(T^{*\leq i-1}\otimes I_{d_i})(I-L(T_i^*)L(T_i^*)^T)((T^{*\leq i-1})^T\otimes I_{d_i})\mathcal{D}^{\langle i\rangle}\|_{\mathrm{F}}^2 \\ &= \|(T^{*\leq i-1}\otimes I_{d_i})((T^{*\leq i-1})^T\otimes I_{d_i})\mathcal{D}^{\langle i\rangle}\|_{\mathrm{F}}^2 \\ &\overset{(a)}{=} \|\mathcal{P}_{i-1}^*\mathcal{D}^{\langle i-1\rangle}\|_{\mathrm{F}}^2,\end{aligned}$$

where in (a) we use Lemma 12.

On the other hand, for $\xi_{k,2}$, we have from (40), $\|(I-\mathcal{P}_i^*)L_i\|_{\mathrm{F}}\leq 3\|\mathcal{D}\|_{\mathrm{F}}$. And from (37), we have

$$\|H_i\|_{\mathrm{F}}\leq\frac{12\|\Delta_{i-1}\|_{\mathrm{F}}^2}{\sigma_{\min}(\Lambda_i^*)}+\frac{4\|\mathcal{D}\|_{\mathrm{F}}\|\Delta_{i-1}\|_{\mathrm{F}}}{\sigma_{\min}(\Lambda_i^*)}\leq\frac{56\|\mathcal{D}\|_{\mathrm{F}}^2}{\sigma_{\min}(\Lambda_i^*)}.$$

Combine the above two estimations, and we have

$$\xi_{k,2}\leq\sum_{i=2}^{k}\left(2\cdot 3\|\mathcal{D}\|_{\mathrm{F}}\frac{56\|\mathcal{D}\|_{\mathrm{F}}^2}{\sigma_{\min}(\Lambda_i^*)}+\frac{56^2\|\mathcal{D}\|_{\mathrm{F}}^4}{\sigma_{\min}^2(\Lambda_i^*)}\right)\leq\frac{350(k-1)\|\mathcal{D}\|_{\mathrm{F}}^3}{\underline{\sigma}}.$$

And thus

$$\|(I-\mathcal{P}_k^*)\Delta_k\|_{\mathrm{F}}^2+\|\mathcal{P}_k^*\mathcal{D}^{\langle k\rangle}\|_{\mathrm{F}}^2\leq\|\mathcal{D}\|_{\mathrm{F}}^2+\frac{350(k-1)\|\mathcal{D}\|_{\mathrm{F}}^3}{\underline{\sigma}}. \tag{43}$$

Upper bound for $\|\Delta_i\|_{\mathrm{F}}$. We shall bound $\|\Delta_i\|_{\mathrm{F}}$ by induction. For the base case when $i=1$, we have $\|\Delta_1\|_{\mathrm{F}}=\|\mathcal{D}^{\langle 1\rangle}\|_{\mathrm{F}}=\|\mathcal{D}\|_{\mathrm{F}}\leq 2\|\mathcal{D}\|_{\mathrm{F}}$. Now suppose we have the bound for $\|\Delta_i\|_{\mathrm{F}}\leq 2\|\mathcal{D}\|_{\mathrm{F}}$ for all $1\leq i\leq k$. Then from the definition of L_{k+1}, we have

$$\begin{aligned}\|L_{k+1}\|_{\mathrm{F}}^2 &= \|(I-\mathcal{P}_k^*)\Delta_k V_{k+1}^*V_{k+1}^{*T}\|_{\mathrm{F}}^2+\|(\mathcal{P}_k^*\otimes I_{d_{k+1}})\mathcal{D}^{\langle k+1\rangle}\|_{\mathrm{F}}^2\leq\|(I-\mathcal{P}_k^*)\Delta_k\|_{\mathrm{F}}^2+\|\mathcal{P}_k^*\mathcal{D}^{\langle k\rangle}\|_{\mathrm{F}}^2 \\ &\overset{(a)}{\leq}\|\mathcal{D}\|_{\mathrm{F}}^2+\frac{350(k-1)\|\mathcal{D}\|_{\mathrm{F}}^3}{\sigma_{\min}(\Lambda_k^*)},\end{aligned}$$

where (a) follows from (43). And the upper bound for $\|H_{k+1}\|_{\mathrm{F}}$ is already derived as in (37), since $\sigma_{\min}(\Lambda_{k+1}^*)\geq C_1\|\mathcal{D}\|_{\mathrm{F}}\geq 8\|\Delta_k\|_{\mathrm{F}}$, we have

$$\|H_{k+1}\|_{\mathrm{F}}\leq\frac{12\|\Delta_k\|_{\mathrm{F}}^2}{\sigma_{\min}(\Lambda_{k+1}^*)}+\frac{4\|\mathcal{D}\|_{\mathrm{F}}\|\Delta_k\|_{\mathrm{F}}}{\sigma_{\min}(\Lambda_{k+1}^*)}\leq\frac{56\|\mathcal{D}\|_{\mathrm{F}}^2}{\sigma_{\min}(\Lambda_{k+1}^*)}.$$

From (36), we have

$$\|\Delta_{k+1}\|_{\mathrm{F}}=\|L_{k+1}+H_{k+1}\|_{\mathrm{F}}\leq\sqrt{1+\frac{350(k-1)\|\mathcal{D}\|_{\mathrm{F}}}{\sigma_{\min}(\Lambda_{k+1}^*)}}\|\mathcal{D}\|_{\mathrm{F}}+\frac{56\|\mathcal{D}\|_{\mathrm{F}}}{\sigma_{\min}(\Lambda_{k+1}^*)}\|\mathcal{D}\|_{\mathrm{F}}\leq 2\|\mathcal{D}\|_{\mathrm{F}}.$$

And this finishes the induction. So it holds for all $i \in [m-1]$ that $\|\Delta_i\|_{\mathrm{F}} \leq 2\|\mathcal{D}\|_{\mathrm{F}}$.

Estimation of $\|\mathsf{SVD}^{\mathsf{tt}}_{\boldsymbol{r}}(\mathcal{T}) - \mathcal{T}^*\|_{\mathrm{F}}$. Now we go back to (35) and we bound $\|(\mathrm{I}.1)\|_{\mathrm{F}}$ and $\|(\mathrm{I}.2)\|_{\mathrm{F}}$ separately. For the term $\|(\mathrm{I}.2)\|_{\mathrm{F}}$, we have

$$\|(\mathrm{I}.2)\|_{\mathrm{F}} = \|H(\widehat{\mathcal{P}}_{m-1}, \mathcal{P}^*_{m-1})(\mathcal{T}^*)^{\langle m-1\rangle} + (\widehat{\mathcal{P}}_{m-1} - \mathcal{P}^*_{m-1})\mathcal{D}^{\langle m-1\rangle}\|_{\mathrm{F}} \overset{(a)}{=} \|H_m\|_{\mathrm{F}},$$

where (a) follows from (36) and Lemma 12. So from (37), we have

$$\|(\mathrm{I}.2)\|_{\mathrm{F}} \leq \frac{12\|\Delta_{m-1}\|_{\mathrm{F}}^2}{\sigma_{\min}(\Lambda^*_m)} + \frac{4\|\mathcal{D}\|_{\mathrm{F}}\|\Delta_{m-1}\|_{\mathrm{F}}}{\sigma_{\min}(\Lambda^*_m)} \leq \frac{56\|\mathcal{D}\|_{\mathrm{F}}^2}{\sigma_{\min}(\Lambda^*_m)}.$$

And for the term $\|(\mathrm{I}.1)\|_{\mathrm{F}}$, we have from (43),

$$\|(\mathrm{I}.1)\|_{\mathrm{F}}^2 \leq \|(I - \mathcal{P}^*_{m-1})\Delta_{m-1}\|_{\mathrm{F}}^2 + \|\mathcal{P}^*_{m-1}\mathcal{D}^{\langle m-1\rangle}\|_{\mathrm{F}}^2 \leq \|\mathcal{D}\|_{\mathrm{F}}^2 + \frac{350(m-2)\|\mathcal{D}\|_{\mathrm{F}}^3}{\sigma_{\min}(\Lambda^*_{m-1})}$$

So together with (35), we have

$$\begin{aligned}
\|\mathsf{SVD}^{\mathsf{tt}}_{\boldsymbol{r}}(\mathcal{T}) - \mathcal{T}^*\|_{\mathrm{F}}^2 &\leq \|(\mathrm{I}.1)\|_{\mathrm{F}}^2 + 2\|(\mathrm{I}.1)\|_{\mathrm{F}}\|(\mathrm{I}.2)\|_{\mathrm{F}} + \|(\mathrm{I}.2)\|_{\mathrm{F}}^2 \\
&\leq \|\mathcal{D}\|_{\mathrm{F}}^2 + \frac{350(m-2)\|\mathcal{D}\|_{\mathrm{F}}^3}{\sigma_{\min}(\Lambda^*_{m-1})} + \frac{250\|\mathcal{D}\|_{\mathrm{F}}^3}{\sigma_{\min}(\Lambda^*_m)} \\
&\leq \|\mathcal{D}\|_{\mathrm{F}}^2 + \frac{600m\|\mathcal{D}\|_{\mathrm{F}}^3}{\underline{\sigma}}.
\end{aligned}$$

And this finishes the proof of the lemma.

C.7 Proof of Lemma 19

First we introduce some notations,

$$\mathcal{T} = [T_1, \ldots, T_m], \mathcal{T}^* = [T_1^*, \ldots, T_m^*], \text{ and } (\mathcal{T}^*)^{\langle i\rangle} = T^{*\leq i}\Lambda^*_{i+1}V^*_{i+1}, \mathcal{T}^{\langle i\rangle} = T^{\leq i}\Lambda_{i+1}V_{i+1}, i \in [m-1].$$

Also we denote $\mathcal{P}_i = T^{\leq i}T^{\leq iT}$, $\mathcal{Q}_{i+1} = V_{i+1}V_{i+1}^T$ as shorthand notations, and define similar notations for $\mathcal{T}^*$. Now we take $\mathcal{A} = \mathcal{T}^*$ and denote $\mathcal{P}_{\mathbb{T}}(\mathcal{T}^*) = \delta\mathcal{T}_1 + \ldots + \delta\mathcal{T}_m$, then we have for all $i \in [m-1]$

$$\begin{aligned}
(\delta\mathcal{T}_i)^{\langle i\rangle} &= (T^{\leq i-1} \otimes I)(I - L(T_i)L(T_i)^T)(T^{\leq i-1} \otimes I)^T(\mathcal{T}^*)^{\langle i\rangle}V_{i+1}V_{i+1}^T \\
&= (I - \mathcal{P}_i)(\mathcal{P}_{i-1} \otimes I)(\mathcal{T}^*)^{\langle i\rangle}\mathcal{Q}_{i+1}.
\end{aligned}$$

And

$$\begin{aligned}
-\|\delta\mathcal{T}_i\|_{\mathrm{F}}^2 &= -\langle (I-\mathcal{P}_i)(\mathcal{P}_{i-1}\otimes I)(\mathcal{T}^*)^{\langle i\rangle}\mathcal{Q}_{i+1}, (I-\mathcal{P}_i)(\mathcal{P}_{i-1}\otimes I)(\mathcal{T}^*)^{\langle i\rangle}\mathcal{Q}_{i+1}\rangle \\
&= -\langle (\mathcal{T}^*)^{\langle i\rangle}, (\mathcal{P}_{i-1}\otimes I)(I-\mathcal{P}_i)(\mathcal{P}_{i-1}\otimes I)(\mathcal{T}^*)^{\langle i\rangle}\mathcal{Q}_{i+1}\rangle \\
&= -\langle (\mathcal{T}^*)^{\langle i\rangle}, (\mathcal{P}_{i-1}\otimes I)(\mathcal{T}^*)^{\langle i\rangle}\mathcal{Q}_{i+1}\rangle + \langle (\mathcal{T}^*)^{\langle i\rangle}, \mathcal{P}_i(\mathcal{T}^*)^{\langle i\rangle}\mathcal{Q}_{i+1}\rangle \\
&= -\langle (\mathcal{T}^*)^{\langle i\rangle}, (\mathcal{P}_{i-1}\otimes I)(\mathcal{T}^*)^{\langle i\rangle}(\mathcal{Q}_{i+1}-\mathcal{Q}_{i+1}^*)\rangle + \langle (\mathcal{T}^*)^{\langle i\rangle}, \mathcal{P}_i(\mathcal{T}^*)^{\langle i\rangle}(\mathcal{Q}_{i+1}-\mathcal{Q}_{i+1}^*)\rangle \\
&\quad - \langle (\mathcal{T}^*)^{\langle i\rangle}, (\mathcal{P}_{i-1}\otimes I)(\mathcal{T}^*)^{\langle i\rangle}\rangle + \langle (\mathcal{T}^*)^{\langle i\rangle}, \mathcal{P}_i(\mathcal{T}^*)^{\langle i\rangle}\rangle \\
&= -\langle (\mathcal{T}^*)^{\langle i\rangle}, ((\mathcal{P}_{i-1}-\mathcal{P}_{i-1}^*)\otimes I)(\mathcal{T}^*)^{\langle i\rangle}(\mathcal{Q}_{i+1}-\mathcal{Q}_{i+1}^*)\rangle \\
&\quad + \langle (\mathcal{T}^*)^{\langle i\rangle}, (\mathcal{P}_i-\mathcal{P}_i^*)(\mathcal{T}^*)^{\langle i\rangle}(\mathcal{Q}_{i+1}-\mathcal{Q}_{i+1}^*)\rangle \\
&\quad - \langle (\mathcal{T}^*)^{\langle i\rangle}, (\mathcal{P}_{i-1}\otimes I)(\mathcal{T}^*)^{\langle i\rangle}\rangle + \langle (\mathcal{T}^*)^{\langle i\rangle}, \mathcal{P}_i(\mathcal{T}^*)^{\langle i\rangle}\rangle
\end{aligned}$$

Meanwhile, when $i=m$, we have

$$-\|\delta\mathcal{T}_m\|_{\mathrm{F}}^2 = -\|(\mathcal{P}_{m-1}\otimes I)(\mathcal{T}^*)^{\langle m\rangle}\|_{\mathrm{F}}^2 = -\|\mathcal{P}_{m-1}(\mathcal{T}^*)^{\langle m-1\rangle}\|_{\mathrm{F}}^2,$$

where the last equality is from Lemma 12. Now we first estimate

$$\begin{aligned}
&\sum_{i=1}^{m-1}\Big(-\langle (\mathcal{T}^*)^{\langle i\rangle}, (\mathcal{P}_{i-1}\otimes I)(\mathcal{T}^*)^{\langle i\rangle}\rangle + \langle (\mathcal{T}^*)^{\langle i\rangle}, \mathcal{P}_i(\mathcal{T}^*)^{\langle i\rangle}\rangle\Big) - \|\mathcal{P}_{m-1}(\mathcal{T}^*)^{\langle m-1\rangle}\|_{\mathrm{F}}^2 \\
=&\sum_{i=1}^{m-1}\Big(-\|(\mathcal{P}_{i-1}\otimes I)(\mathcal{T}^*)^{\langle i\rangle}\|_{\mathrm{F}}^2 + \|\mathcal{P}_i(\mathcal{T}^*)^{\langle i\rangle}\|_{\mathrm{F}}^2\Big) - \|\mathcal{P}_{m-1}(\mathcal{T}^*)^{\langle m-1\rangle}\|_{\mathrm{F}}^2 \\
=&\sum_{i=1}^{m-1}\Big(-\|\mathcal{P}_{i-1}(\mathcal{T}^*)^{\langle i-1\rangle}\|_{\mathrm{F}}^2 + \|\mathcal{P}_i(\mathcal{T}^*)^{\langle i\rangle}\|_{\mathrm{F}}^2\Big) - \|\mathcal{P}_{m-1}(\mathcal{T}^*)^{\langle m-1\rangle}\|_{\mathrm{F}}^2 = -\|\mathcal{T}^*\|_{\mathrm{F}}^2.
\end{aligned}$$

And now we estimate $\langle (\mathcal{T}^*)^{\langle i\rangle}, (\mathcal{P}_i-\mathcal{P}_i^*)(\mathcal{T}^*)^{\langle i\rangle}(\mathcal{Q}_{i+1}-\mathcal{Q}_{i+1}^*)\rangle$ first,

$$\begin{aligned}
&\langle (\mathcal{T}^*)^{\langle i\rangle}, (\mathcal{P}_i-\mathcal{P}_i^*)(\mathcal{T}^*)^{\langle i\rangle}(\mathcal{Q}_{i+1}-\mathcal{Q}_{i+1}^*)\rangle \\
=&\langle (\mathcal{P}_i-\mathcal{P}_i^*)(\mathcal{T}^*)^{\langle i\rangle}, (\mathcal{T}^*)^{\langle i\rangle}(\mathcal{Q}_{i+1}-\mathcal{Q}_{i+1}^*)\rangle \\
=&\langle \mathcal{P}_i^*\Delta_i(I-\mathcal{Q}_{i+1}^*) + (\mathcal{T}^*)^{\langle i\rangle}H(\mathcal{Q}_{i+1},\mathcal{Q}_{i+1}^*), (I-\mathcal{P}_i^*)\Delta_i\mathcal{Q}_{i+1}^* + H(\mathcal{P}_i,\mathcal{P}_i^*)(\mathcal{T}^*)^{\langle i\rangle}\rangle \\
=&\langle (\mathcal{T}^*)^{\langle i\rangle}H(\mathcal{Q}_{i+1},\mathcal{Q}_{i+1}^*), H(\mathcal{P}_i,\mathcal{P}_i^*)(\mathcal{T}^*)^{\langle i\rangle}\rangle \\
\le&\|(\mathcal{T}^*)^{\langle i\rangle}H(\mathcal{Q}_{i+1},\mathcal{Q}_{i+1}^*)\|_{\mathrm{F}}\|H(\mathcal{P}_i,\mathcal{P}_i^*)(\mathcal{T}^*)^{\langle i\rangle}\|_{\mathrm{F}}
\end{aligned}$$

Notice here we use

$$\mathcal{P}_i-\mathcal{P}_i^* = T^{*\le i}(\Lambda_{i+1}^*)^{-1}(V_{i+1}^*)^T\Delta_i^T(I-\mathcal{P}_i^*) + (I-\mathcal{P}_i^*)\Delta_i V_{i+1}^*(\Lambda_{i+1}^*)^{-1}(T^{*\le i})^T + H(\mathcal{P}_i,\mathcal{P}_i^*),$$

and

$$\mathcal{Q}_{i+1} - \mathcal{Q}^*_{i+1} = V^*_{i+1}(\Lambda^*_{i+1})^{-1}(T^{*\leq i})^T\Delta_i(I - \mathcal{Q}^*_{i+1}) + (I - \mathcal{Q}^*_{i+1})\Delta_i^T T^{*\leq i}(\Lambda^*_{i+1})^{-1}V^{*T}_{i+1} + H(\mathcal{Q}_{i+1}, \mathcal{Q}^*_{i+1}),$$

where $\Delta_i = (\mathcal{T} - \mathcal{T}^*)^{\langle i\rangle}$ and $H(\cdot,\cdot)$ denotes high order terms.

Similarly for $\langle(\mathcal{T}^*)^{\langle i\rangle}, ((\mathcal{P}_{i-1} - \mathcal{P}^*_{i-1}) \otimes I)(\mathcal{T}^*)^{\langle i\rangle}(\mathcal{Q}_{i+1} - \mathcal{Q}^*_{i+1})\rangle$, we have

$$\begin{aligned}
&\langle(\mathcal{T}^*)^{\langle i\rangle}, ((\mathcal{P}_{i-1} - \mathcal{P}^*_{i-1}) \otimes I)(\mathcal{T}^*)^{\langle i\rangle}(\mathcal{Q}_{i+1} - \mathcal{Q}^*_{i+1})\rangle \\
&= \langle(\mathcal{T}^*)^{\langle i\rangle}H(\mathcal{Q}_{i+1}, \mathcal{Q}^*_{i+1}), (H(\mathcal{P}_{i-1}, \mathcal{P}^*_{i-1}) \otimes I)(\mathcal{T}^*)^{\langle i\rangle}\rangle \\
&\leq \|(\mathcal{T}^*)^{\langle i\rangle}H(\mathcal{Q}_{i+1}, \mathcal{Q}^*_{i+1})\|_{\mathrm{F}}\|(H(\mathcal{P}_{i-1}, \mathcal{P}^*_{i-1}) \otimes I)(\mathcal{T}^*)^{\langle i\rangle}\|_{\mathrm{F}} \\
&= \|(\mathcal{T}^*)^{\langle i\rangle}H(\mathcal{Q}_{i+1}, \mathcal{Q}^*_{i+1})\|_{\mathrm{F}}\|H(\mathcal{P}_{i-1}, \mathcal{P}^*_{i-1})(\mathcal{T}^*)^{\langle i-1\rangle}\|_{\mathrm{F}}
\end{aligned}$$

Now as long as $\underline{\sigma} \geq 8\|\mathcal{T} - \mathcal{T}^*\|_{\mathrm{F}}$, we have

$$\begin{aligned}
\|H(\mathcal{P}_i, \mathcal{P}^*_i)(\mathcal{T}^*)^{\langle i\rangle}\|_{\mathrm{F}} &\leq \frac{12\|\mathcal{T} - \mathcal{T}^*\|_{\mathrm{F}}^2}{\underline{\sigma}}, \\
\|(\mathcal{T}^*)^{\langle i\rangle}H(\mathcal{Q}_{i+1}, \mathcal{Q}^*_{i+1})\|_{\mathrm{F}} &\leq \frac{12\|\mathcal{T} - \mathcal{T}^*\|_{\mathrm{F}}^2}{\underline{\sigma}}.
\end{aligned}$$

Finally, we have

$$\begin{aligned}
\|\mathcal{P}^{\perp}_{\mathbb{T}}(\mathcal{T}^*)\|_{\mathrm{F}}^2 = \|\mathcal{T}^*\|_{\mathrm{F}}^2 - \sum_{i=1}^m \|\delta\mathcal{T}_i\|_{\mathrm{F}}^2 &= \sum_{i=1}^{m-1}\big(-\langle(\mathcal{T}^*)^{\langle i\rangle}, ((\mathcal{P}_{i-1} - \mathcal{P}^*_{i-1}) \otimes I)(\mathcal{T}^*)^{\langle i\rangle}(\mathcal{Q}_{i+1} - \mathcal{Q}^*_{i+1})\rangle \\
&\quad + \langle(\mathcal{T}^*)^{\langle i\rangle}, (\mathcal{P}_i - \mathcal{P}^*_i)(\mathcal{T}^*)^{\langle i\rangle}(\mathcal{Q}_{i+1} - \mathcal{Q}^*_{i+1})\rangle\big) \\
&\leq \frac{288m^2\|\mathcal{T} - \mathcal{T}^*\|_{\mathrm{F}}^4}{\underline{\sigma}^2}.
\end{aligned}$$

And this finishes the proof.

C.8 Proof of Lemma 20

We begin the proof introducing some notations. For simplicity, we denote $\widetilde{\mathcal{W}} = \mathsf{Trim}_{\zeta}(\mathcal{W})$. And $\mathsf{SVD}^{\mathsf{tt}}_{\boldsymbol{r}}(\widetilde{\mathcal{W}}) = \widehat{\mathcal{W}} = [\widehat{W}_1, \ldots, \widehat{W}_m]$. And we denote for all $i \in [m-1]$, $\widehat{\mathcal{W}}^{\langle i\rangle} = \widehat{W}^{\leq i}\widehat{W}^{\geq i+1} = \widehat{W}^{\leq i}\Sigma_{i+1}N_{i+1}^T$ where $\Sigma_{i+1} \in \mathbb{R}^{r_i\times r_i}$ is invertible and $N_{i+1}^TN_{i+1} = I_{r_i}$. And we also introduce some notations for the process of TTSVD. From Algorithm 1, we know $(\widehat{W}^{\leq i-1} \otimes I)^T\widetilde{\mathcal{W}}^{\langle i\rangle} = L(\widehat{W}_i)S_{i+1}R_{i+1}^T + E_i$ is how we get the estimation $\widehat{W}_i$ once we know $\widehat{W}_1, \ldots, \widehat{W}_{i-1}$, where $S_{i+1} \in \mathbb{R}^{r_i\times r_i}$ is invertible and $R_{i+1}^TR_{i+1} = I_{r_i}$. To estimate the incoherence of $\widehat{\mathcal{W}}$, we check $\max_j \|e_j^T\widehat{W}^{\leq i}\|_{\ell_2}$ and $\max_j \|N_{i+1}^Te_j\|_{\ell_2}$.

We first estimate $\sigma_{\min}(\Sigma_{i+1})$ and $\sigma_{\min}(S_{i+1})$. From the Algorithm 1, we know

$$\begin{aligned}\sigma_{\min}(S_{i+1}) &= \sigma_{\min}((\widehat{W}^{\leq i-1}\widehat{W}^{\leq i-1T}\otimes I)\widetilde{\mathcal{W}}^{\langle i\rangle}) \\ &\geq \sigma_{\min}((\mathcal{T}^*)^{\langle i\rangle}) - \|(\mathcal{T}^*)^{\langle i\rangle} - (\widehat{W}^{\leq i-1}\widehat{W}^{\leq i-1T}\otimes I)\widetilde{\mathcal{W}}^{\langle i\rangle}\|_{\mathrm{F}}.\end{aligned}$$

So we need to derive an upper bound for $\|(\mathcal{T}^*)^{\langle i\rangle} - (\widehat{W}^{\leq i-1}\widehat{W}^{\leq i-1T}\otimes I)\widetilde{\mathcal{W}}^{\langle i\rangle}\|_{\mathrm{F}}$.

$$\begin{aligned}(\widehat{W}^{\leq i-1}\widehat{W}^{\leq i-1T}\otimes I)\widetilde{\mathcal{W}}^{\langle i\rangle} - (\mathcal{T}^*)^{\langle i\rangle} &= (\widehat{W}^{\leq i-1}\widehat{W}^{\leq i-1T} - T^{*\leq i-1}T^{*\leq i-1T})\otimes I\cdot(\widetilde{\mathcal{W}}^{\langle i\rangle} - (\mathcal{T}^*)^{\langle i\rangle}) \\ &\quad + T^{*\leq i-1}T^{*\leq i-1T}\otimes I\cdot(\widetilde{\mathcal{W}}^{\langle i\rangle} - (\mathcal{T}^*)^{\langle i\rangle}) \\ &\quad + (\widehat{W}^{\leq i-1}\widehat{W}^{\leq i-1T} - T^{*\leq i-1}T^{*\leq i-1T})\otimes I\cdot(\mathcal{T}^*)^{\langle i\rangle}. \qquad (44)\end{aligned}$$

Since $\|\mathcal{T}^*\|_{\ell_\infty} = \mathsf{Spiki}(\mathcal{T}^*)\frac{\|\mathcal{T}^*\|_{\mathrm{F}}}{\sqrt{d^*}} \leq \nu\frac{\|\mathcal{T}^*\|_{\mathrm{F}}}{\sqrt{d^*}} \leq \frac{10\nu}{9}\frac{\|\mathcal{W}\|_{\mathrm{F}}}{\sqrt{d^*}} = \zeta$, we have $\|\widetilde{\mathcal{W}} - \mathcal{T}^*\|_{\mathrm{F}} \leq \|\mathcal{W} - \mathcal{T}^*\|_{\mathrm{F}}$. And the bound for $\|\widehat{W}^{\leq i-1}\widehat{W}^{\leq i-1T} - T^{*\leq i-1}T^{*\leq i-1T}\|$ goes as follows. First we notice that from Lemma 18,

$$\|\widehat{\mathcal{W}} - \mathcal{T}^*\|_{\mathrm{F}}^2 \leq \|\widetilde{\mathcal{W}} - \mathcal{T}^*\|_{\mathrm{F}}^2 + \frac{600m\|\widetilde{\mathcal{W}} - \mathcal{T}^*\|_{\mathrm{F}}^3}{\underline{\sigma}} \leq 2\|\mathcal{W} - \mathcal{T}^*\|_{\mathrm{F}}^2, \qquad (45)$$

where the last inequality is from the assumption $\|\widetilde{\mathcal{W}} - \mathcal{T}^*\|_{\mathrm{F}} \leq \|\mathcal{W} - \mathcal{T}^*\|_{\mathrm{F}} \leq \frac{1}{600m}\underline{\sigma}$. So we have from Wedin's theorem,

$$\|\widehat{W}^{\leq i-1}\widehat{W}^{\leq i-1T} - T^{*\leq i-1}T^{*\leq i-1T}\| \leq \frac{\sqrt{2}\|\widehat{\mathcal{W}} - \mathcal{T}^*\|_{\mathrm{F}}}{\sigma_{\min}(\Lambda_i^*) - \|\widetilde{\mathcal{W}} - \mathcal{T}^*\|_{\mathrm{F}}} \leq \frac{4\|\mathcal{W} - \mathcal{T}^*\|_{\mathrm{F}}}{\underline{\sigma}} \leq \frac{1}{150m\sqrt{\bar{r}}\kappa_0},$$

where in the penultimate inequality we use $\sigma_{\min}(\Lambda_i^*) - \|\widetilde{\mathcal{W}} - \mathcal{T}^*\|_{\mathrm{F}} \geq \frac{\underline{\sigma}}{2}$ and (45). This together with (44) gives

$$\begin{aligned}&\|(\mathcal{T}^*)^{\langle i\rangle} - (\widehat{W}^{\leq i-1}\widehat{W}^{\leq i-1T}\otimes I)\widetilde{\mathcal{W}}^{\langle i\rangle}\|_{\mathrm{F}} \\ &\qquad \leq \frac{4\|\mathcal{W} - \mathcal{T}^*\|_{\mathrm{F}}}{\underline{\sigma}}\cdot\|\mathcal{W} - \mathcal{T}^*\|_{\mathrm{F}} + \|\mathcal{W} - \mathcal{T}^*\|_{\mathrm{F}} + \frac{4\|\mathcal{W} - \mathcal{T}^*\|_{\mathrm{F}}}{\underline{\sigma}}\cdot\|\mathcal{T}^*\|_{\mathrm{F}} \\ &\qquad \leq \frac{\underline{\sigma}}{10}.\end{aligned}$$

So we conclude that $\sigma_{\min}(S_{i+1}) \geq \frac{9}{10}\underline{\sigma}$.

Meanwhile, for $\sigma_{\min}(\Sigma_{i+1})$ and $\sigma_{\max}(\Sigma_{i+1})$, we have

$$\begin{aligned}\sigma_{\max}(\Sigma_{i+1}) = \sigma_{\max}(\widehat{\mathcal{W}}^{\langle i\rangle}) &\leq \sigma_{\max}((\mathcal{T}^*)^{\langle i\rangle}) + \|\mathcal{T}^* - \mathcal{W}\|_{\mathrm{F}} \leq \frac{11}{10}\overline{\sigma}, \\ \sigma_{\min}(\Sigma_{i+1}) = \sigma_{\min}(\widehat{\mathcal{W}}^{\langle i\rangle}) &\geq \sigma_{\min}((\mathcal{T}^*)^{\langle i\rangle}) - \|\mathcal{T}^* - \mathcal{W}\|_{\mathrm{F}} \geq \frac{9}{10}\underline{\sigma}.\end{aligned}$$

With these preparations, we are ready to bound the incoherence of $\widehat{\mathcal{W}}$. For all $j \in [d_1 \dots d_i]$,

$$\begin{aligned}
\|e_j^T \widehat{W}^{\leq i}\|_{\ell_2} &= \|e_j^T (\widehat{W}^{\leq i-1} \widehat{W}^{\leq i-1T} \otimes I) \widetilde{\mathcal{W}}^{\langle i \rangle} R_{i+1} S_{i+1}^{-1}\|_{\ell_2} \\
&\leq \frac{\|f_j^T \widetilde{\mathcal{W}}^{\langle i \rangle}\|_{\ell_2}}{\sigma_{\min}(S_{i+1})} \overset{(a)}{\leq} \frac{10\|f_j\|_{\ell_2}\sqrt{d_{i+1}\dots d_m}\|\widetilde{\mathcal{W}}\|_{\ell_\infty}}{9\underline{\sigma}} \\
&\overset{(b)}{\leq} \frac{100}{81\underline{\sigma}} \frac{\nu\|\mathcal{W}\|_{\mathrm{F}}}{\sqrt{d_1 \dots d_i}} \leq \frac{100\nu}{81\underline{\sigma}} \frac{\|\mathcal{T}^*\|_{\mathrm{F}} + \|\mathcal{W} - \mathcal{T}^*\|_{\mathrm{F}}}{\sqrt{d_1 \dots d_i}} \\
&\leq \frac{100\nu}{81}\kappa_0 \frac{\sqrt{r_i}}{\sqrt{d_1 \dots d_i}} + \frac{10\nu}{81} \frac{1}{\sqrt{d_1 \dots d_i}} \leq \frac{110\nu\kappa_0}{81} \frac{\sqrt{r_i}}{\sqrt{d_1 \dots d_i}},
\end{aligned} \tag{46}$$

where $f_j = (\widehat{W}^{\leq i-1} \widehat{W}^{\leq i-1T} \otimes I) e_j$ and in (a) we use the estimation for $\sigma_{\min}(S_{i+1})$ and $\|Wx\|_{\ell_2} \leq \sqrt{n}\|W\|_{\ell_\infty}\|x\|_{\ell_2}$ for some matrix $W \in \mathbb{R}^{n \times m}$; in (b) we use $\|\widetilde{\mathcal{W}}\|_{\ell_\infty} \leq \zeta = \frac{10\|\mathcal{W}\|_{\mathrm{F}}}{9\sqrt{d^*}}\nu$ and $\|f_j\|_{\ell_\infty} \leq 1$.

On the other hand, we have

$$\widehat{W}^{\geq i+1} = R(\widehat{W}_{i+1})(I \otimes \widehat{W}^{\geq i+2}). \tag{47}$$

Here we use the convention that $\widehat{W}^{m+1} = [1]$. And from $L(\widehat{W}_{i+1}) = (\widehat{W}^{\leq i} \otimes I)^T \widetilde{\mathcal{W}}^{\langle i+1 \rangle} R_{i+2} S_{i+2}^{-1}$, we obtain

$$R(\widehat{W}_{i+1}) = (\widehat{W}^{\leq i})^T \mathsf{reshape}(\widetilde{\mathcal{W}}^{\langle i+1 \rangle} R_{i+2} S_{i+2}^{-1}) = (\widehat{W}^{\leq i})^T \widetilde{\mathcal{W}}^{\langle i \rangle} (I \otimes R_{i+2} S_{i+2}^{-1}). \tag{48}$$

Combine (47) and (48), we have

$$\widehat{W}^{\geq i+1} = (\widehat{W}^{\leq i})^T \widetilde{\mathcal{W}}^{\langle i \rangle} (I \otimes R_{i+2} S_{i+2}^{-1} \widehat{W}^{\geq i+2}),$$

which implies

$$N_{i+1}^T = \Sigma_{i+1}^{-1} (\widehat{W}^{\leq i})^T \widetilde{\mathcal{W}}^{\langle i \rangle} (I \otimes R_{i+2} S_{i+2}^{-1} \widehat{W}^{\geq i+2}).$$

Now for any $j \in [d_{i+1} \dots d_m]$, we have

$$\begin{aligned}
\|N_{i+1}^T e_j\|_{\ell_2} &= \|\Sigma_{i+1}^{-1} (\widehat{W}^{\leq i})^T \widetilde{\mathcal{W}}^{\langle i \rangle} (I \otimes R_{i+2} S_{i+2}^{-1} \widehat{W}^{\geq i+2}) e_j\|_{\ell_2} \\
&\leq \frac{1}{\sigma_{\min}(\Sigma_{i+1})} \|(\widehat{W}^{\leq i})^T \widetilde{\mathcal{W}}^{\langle i \rangle} (I \otimes R_{i+2} S_{i+2}^{-1} \widehat{W}^{\geq i+2}) e_j\|_{\ell_2} \\
&\leq \frac{1}{\sigma_{\min}(\Sigma_{i+1})} \|\widetilde{\mathcal{W}}^{\langle i \rangle} (I \otimes R_{i+2} S_{i+2}^{-1} \widehat{W}^{\geq i+2}) e_j\|_{\ell_2} \\
&\leq \frac{\sqrt{d_1 \dots d_i}}{\sigma_{\min}(\Sigma_{i+1})} \|\widetilde{\mathcal{W}}\|_{\ell_\infty} \|(I \otimes R_{i+2} S_{i+2}^{-1} \widehat{W}^{\geq i+2}) e_j\|_{\ell_2} \\
&\leq \frac{\sqrt{d_1 \dots d_i}}{\sigma_{\min}(\Sigma_{i+1})} \|\widetilde{\mathcal{W}}\|_{\ell_\infty} \frac{\sigma_{\max}(\Sigma_{i+2})}{\sigma_{\min}(S_{i+2})} \\
&\leq \frac{1100}{729\underline{\sigma}} \kappa_0 \frac{\nu\|\mathcal{W}\|_{\mathrm{F}}}{\sqrt{d_{i+1} \dots d_m}} \\
&\leq \frac{1210}{729} \kappa_0^2 \nu \frac{\sqrt{r_i}}{\sqrt{d_{i+1} \dots d_m}}.
\end{aligned} \tag{49}$$

From (46) and (49), we have

$$\mathsf{Incoh}(\widehat{\mathcal{W}}) \le 2\kappa_0^2\nu.$$

Now we consider the second part, from Lemma 18, we see that

$$\|\mathsf{SVD}^{\mathsf{tt}}(\widetilde{\mathcal{W}}) - \mathcal{T}^*\|_{\mathrm{F}}^2 \le \|\widetilde{\mathcal{W}} - \mathcal{T}^*\|_{\mathrm{F}}^2 + \frac{600m\|\widetilde{\mathcal{W}} - \mathcal{T}^*\|_{\mathrm{F}}^3}{\underline{\sigma}} \le 2\|\widetilde{\mathcal{T}} - \mathcal{T}^*\|_{\mathrm{F}}^2 \le 2\|\mathcal{W} - \mathcal{T}^*\|_{\mathrm{F}}^2.$$

And this finishes the proof of the lemma.

C.9 Proof of Lemma 25

For simplicity, we denote

$$S_{1k} = \sum_{i=1}^{k} U^T X_i, \quad S_{2k} = \sum_{i=1}^{k} U^T Y_i \quad \text{and} \quad \Delta_{1k} = S_{1k} - kU^T M, \quad \Delta_{2k} = S_{2k} - kU^T M.$$

First we estimate $\|\frac{1}{n}S_{1n} - U^T M\|$. Notice that

$$\frac{1}{n}S_{1n} - U^T M = \frac{1}{n}\sum_{i=1}^{n}\left(p_1p_2U^T\mathcal{P}_{\omega_i}(M) - U^T M\right) =: \frac{1}{n}\sum_{i=1}^{n} Z_i.$$

And the uniform bound for $p_1p_2U^T\mathcal{P}_{\omega_i}(M)$ is given by

$$\|p_1p_2U^T\mathcal{P}_{\omega_i}(M)\| \le p_1p_2\|M\|_\infty\sqrt{\frac{\mu r}{p_1}} = \sqrt{\mu p_2 r}\sqrt{p_1p_2}\|M\|_\infty.$$

and $\|U^T M\| \le \|M\| \le \sqrt{p_1p_2}\|M\|_\infty$. So $\|Z_i\| \le 2\sqrt{\mu p_2 r}\sqrt{p_1p_2}\|M\|_\infty$. On the other hand we have

$$\begin{aligned}
\mathbb{E}Z_iZ_i^T &\le p_1p_2 \sum_{(j,k)\in[p_1]\times[p_2]} M_{jk}^2 U^T e_j e_j^T U \\
&\le p_1p_2\|M\|_\infty^2 \sum_{(j,k)\in[p_1]\times[p_2]} U^T e_j e_j^T U \\
&= p_1p_2^2\|M\|_\infty^2 I_r
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}Z_i^TZ_i &\le p_1p_2 \sum_{(j,k)\in[p_1]\times[p_2]} M_{jk}^2 e_k e_j^T UU^T e_j e_k^T \\
&\le p_1p_2\|M\|_\infty^2 \sum_k\sum_j \|U^T e_j\|_{\ell_2}^2 e_k e_k^T \\
&\le p_1p_2\|M\|_\infty^2 \mu r I_{p_2}.
\end{aligned}$$

Therefore we have $\max\{\|\sum_{i=1}^n Z_i Z_i^T\|, \|\sum_{i=1}^n Z_i^T Z_i\|\} \leq np_1p_2(\mu r \vee p_2)\|M\|_\infty^2$. And from operator Bernstein inequality, with probability exceeding $1-p^{-\alpha}$,

$$\|\sum_{i=1}^n Z_i\| \leq \sqrt{\frac{8(\alpha+1)\mu p_1 p_2^2 r\log(p)}{3n}}\|M\|_\infty.$$

Now set the event $\boldsymbol{\mathcal{E}}_1$ as

$$\left\{\|\Delta_{1n}\| \leq \sqrt{\frac{8n(\alpha+1)\mu p_1 p_2^2 r\log(p)}{3}}\|M\|_\infty\right\} \cap \left\{\|\Delta_{2n}\| \leq \sqrt{\frac{8n(\alpha+1)\mu p_1 p_2^2 r\log(p)}{3}}\|M\|_\infty\right\}$$

and we know $\mathbb{P}(\boldsymbol{\mathcal{E}}_1) \geq 1-2p^{-\alpha}$. Also, define the event

$$\begin{aligned}\boldsymbol{\mathcal{E}}_2 = &\left\{\max_{j\in[p_2]}\sum_{\omega\in\Omega_1} 1(\omega_2=j) \leq (3\alpha+7)(\frac{n}{p_2}+\log(p))\right\} \\ &\cap \left\{\max_{j\in[p_2]}\sum_{\omega\in\Omega_2} 1(\omega_2=j) \leq (3\alpha+7)(\frac{n}{p_2}+\log(p))\right\}.\end{aligned}$$

From Chernoff bound, we know $\boldsymbol{\mathcal{E}}_2$ holds with probability exceeding $1-2p^{-\alpha}$. From triangular inequality, we have

$$\begin{aligned}&\|\frac{1}{2n^2}\sum_{i,j}(U^T X_i Y_j^T U + U^T Y_j X_i^T U) - U^T M M^T U\| \\ \leq &\frac{1}{2n^2}\|\Delta_{1n}\Delta_{2n}^T + \Delta_{2n}\Delta_{1n}^T\| + \frac{1}{2n}\|\Delta_{2n}M^T U + U^T M\Delta_{2n}^T\| + \frac{1}{2n}\|\Delta_{1n}M^T U + U^T M\Delta_{1n}^T\| \\ =: &A_1 + A_2 + A_3.\end{aligned}$$

Using Markov's inequality and Golden-Thompson inequality repeatedly, we get

$$\begin{aligned}&\mathbb{P}\left(\{\|A_1\| > t\} \cap \boldsymbol{\mathcal{E}}\right) \\ \leq &r\cdot e^{-\lambda t}\mathbb{E}\left(\left\|\mathbb{E}\left\{\exp\left[\frac{\lambda}{2n^2}[\Delta_{1n}(Y_n-M)^T U + U^T(Y_n-M)\Delta_{1n}^T]\right]1_{\boldsymbol{\mathcal{E}}}\middle|S_{1n}\right\}\right\|^n\right).\end{aligned}$$

From triangular inequality, we have

$$\|\frac{\lambda}{2n^2}\Delta_{1n}(Y_n-M)^T U + U^T(Y_n-M)\Delta_{1n}^T\| \leq \frac{\lambda}{n^2}\left(\|\Delta_{1n}Y_n^T U\| + \|\Delta_{1n}M^T U\|\right).$$

Under the event $\boldsymbol{\mathcal{E}}_2$, we have

$$\begin{aligned}&\|\Delta_{1n}Y_n^T U\| \leq \|S_{1n}Y_n^T U\| + n\|U^T M Y_n^T U\| \\ \leq &(3\alpha+7)p_1p_2\|M\|_\infty^2(\frac{n}{p_2}+\log(p))\mu r p_2 + n\sqrt{\mu r p_2}p_1p_2\|M\|_\infty^2.\end{aligned}$$

On the other hand, under the event $\boldsymbol{\mathcal{E}}_1$,

$$\|\Delta_{1n}M^TU\| \le \|\Delta_{1n}\|\|M^TU\| \le \sqrt{\frac{8(\alpha+1)\mu p_1 p_2^2 r\log(p)n}{3}}\|M\|_\infty\sqrt{p_1p_2}\|M\|_\infty.$$

As long as $n \ge \frac{8}{3}(\alpha+1)\log(p)$, we have $\|\Delta_{1n}M^TU\| \le n\sqrt{\mu p_2 r}p_1p_2\|M\|_\infty^2$ and thus

$$\begin{aligned}
&\frac{\lambda}{2n^2}\|\Delta_{1n}(Y_n-M)^TU + U^T(Y_n-M)\Delta_{1n}^T\| \\
\le\ &\frac{\lambda}{n^2}\left((3\alpha+7)p_1p_2\|M\|_\infty^2(\frac{n}{p_2}+\log(p))\mu r p_2 + 2n\sqrt{\mu r p_2}p_1p_2\|M\|_\infty^2\right).
\end{aligned}$$

Therefore for any

$$\lambda \le n^2\left((3\alpha+7)p_1p_2\|M\|_\infty^2(\frac{n}{p_2}+\log(p))\mu r p_2 + 2n\sqrt{\mu r p_2}p_1p_2\|M\|_\infty^2\right)^{-1},$$

we have

$$\begin{aligned}
&\mathbb{E}\left\{\exp\left[\frac{\lambda}{2n^2}[\Delta_{1n}(Y_n-M)^TU + U^T(Y_n-M)\Delta_{1n}^T]\right]1_{\boldsymbol{\mathcal{E}}}\middle|S_{1n}\right\} \\
\le\ & I_r + \mathbb{E}\left\{\left[\frac{\lambda^2}{4n^4}[\Delta_{1n}(Y_n-M)^TU + U^T(Y_n-M)\Delta_{1n}^T]^2\right]1_{\boldsymbol{\mathcal{E}}}\middle|S_{1n}\right\} \\
\le\ & I_r + \mathbb{E}\left\{\left[\frac{\lambda^2}{4n^4}[\Delta_{1n}Y_n^TU + U^TY_n\Delta_{1n}^T]^2\right]1_{\boldsymbol{\mathcal{E}}}\middle|S_{1n}\right\} \\
\le\ & I_r + \frac{\lambda^2p_1p_2\|M\|_\infty^2}{4n^4}\left[(\mu r+2)\Delta_{1n}\Delta_{1n}^T + \mathrm{tr}(\Delta_{1n}\Delta_{1n}^T)I_r\right],
\end{aligned}$$

where in the first inequality we use $\exp(A) \le I+A+A^2$ for $\|A\| \le 1$. And notice that $\mathrm{tr}(\Delta_{1n}\Delta_{1n}^T) \le r\|\Delta_{1n}\Delta_{1n}^T\|$, we obtain

$$\begin{aligned}
&\left\|\mathbb{E}\left\{\exp\left[\frac{\lambda}{2n^2}[\Delta_{1n}(Y_n-M)^TU + U^T(Y_n-M)\Delta_{1n}^T]\right]1_{\boldsymbol{\mathcal{E}}}\middle|S_{1n}\right\}\right\| \\
\le\ & 1 + \frac{\lambda^2p_1p_2\|M\|_\infty^2\mu r}{2n^4}\|\Delta_{1n}\Delta_{1n}^T\| \\
\le\ & 1 + \frac{\lambda^2p_1p_2\|M\|_\infty^2\mu r}{2n^4}\frac{8n(\alpha+1)\mu p_1p_2^2r\log(p)}{3}\|M\|_\infty^2 \\
=\ & 1 + \frac{4(\alpha+1)\lambda^2\mu^2r^2p_2\log(p)}{3n^3}(p_1p_2\|M\|_\infty^2)^2.
\end{aligned}$$

Therefore we have

$$\mathbb{P}\left(\{\|A_1\| > t\}\cap\boldsymbol{\mathcal{E}}\right) \le r\cdot e^{-\lambda t}\exp\left(\frac{4(\alpha+1)\lambda^2\mu^2r^2p_2\log(p)}{3n^2}(p_1p_2\|M\|_\infty^2)^2\right).$$

Taking

$$\begin{aligned}
\lambda = \min\Big\{&\frac{3n^2t}{8(\alpha+1)\mu^2r^2p_1^2p_2^3\|M\|_\infty^4\log(p)}, \frac{n^2}{(6\alpha+14)p_1p_2^2\mu r\|M\|_\infty^2(\frac{n}{p_2}+\log(p))}, \\
&\frac{n}{4\sqrt{\mu r p_2}p_1p_2\|M\|_\infty^2}\Big\}
\end{aligned}$$

yields

$$\mathbb{P}\left(\{\|A_1\| > t\} \cap \boldsymbol{\mathcal{E}}\right) \leq r \cdot \exp\big(- \min\big\{ \frac{3n^2t^2}{16(\alpha+1)\mu^2 r^2 p_1^2 p_2^3 \|M\|_\infty^4 \log(p)}, \\ \frac{n^2 t}{(12\alpha+28)p_1 p_2^2 \mu r \|M\|_\infty^2 (\frac{n}{p_2} + \log(p))}, \frac{nt}{8\sqrt{\mu r p_2} p_1 p_2 \|M\|_\infty^2} \big\}\big).$$

Now we bound A_2 and A_3. Due to the symmetry, we shall consider A_2 only. In a similar fashion, we have

$$\mathbb{P}\left(\{\|A_2\| > t\} \cap \boldsymbol{\mathcal{E}}\right) \leq r \cdot \left\| \mathbb{E}\left\{ \exp\left[\frac{\lambda}{2n} (U^T M (Y_n - M)^T U + U^T (Y_n - M M^T U)) \right] 1_{\boldsymbol{\mathcal{E}}} \right\} \right\|^n.$$

Simple calculations show that

$$\| \frac{\lambda}{2n} (U^T M (Y_n - M)^T U + U^T (Y_n - M) M^T U) \| \leq \frac{2\mu r p_1 p_2 \|M\|_\infty^2 \lambda}{n}.$$

So as long as $\lambda \leq \frac{n}{2\mu r p_1 p_2 \|M\|_\infty^2}$, we have

$$\begin{aligned} &\left\| \mathbb{E}\left\{ \exp\left[\frac{\lambda}{2n} (U^T M (Y_n - M)^T U + U^T (Y_n - M) M^T U) \right] 1_{\boldsymbol{\mathcal{E}}} \right\} \right\| \\ \leq\ & 1 + \left\| \mathbb{E}\left\{ \left[\frac{\lambda}{2n} (U^T M (Y_n - M)^T U + U^T (Y_n - M) M^T U) \right]^2 \cdot 1_{\boldsymbol{\mathcal{E}}} \right\} \right\| \\ \leq\ & 1 + \left\| \mathbb{E}\left\{ \left[\frac{\lambda}{2n} (U^T M Y_n^T U + U^T Y_n M^T U) \right]^2 \cdot 1_{\boldsymbol{\mathcal{E}}} \right\} \right\| \\ \leq\ & 1 + \frac{\lambda^2 \mu r p_1^2 p_2^2 \|M\|_\infty^4}{n^2}. \end{aligned}$$

Now take $\lambda = \min\left\{ \frac{n}{2\mu r p_1 p_2 \|M\|_\infty^2}, \frac{nt}{2\mu r p_1^2 p_2^2 \|M\|_\infty^2} \right\}$, then

$$\mathbb{P}\left(\{\|A_2\| > t\} \cap \boldsymbol{\mathcal{E}}\right) \leq r \cdot \exp\left\{ - \min\{ \frac{nt}{4\mu r p_1 p_2 \|M\|_\infty^2}, \frac{nt^2}{4\mu r p_1^2 p_2^2 \|M\|_\infty^4} \} \right\}.$$

So we have

$$\begin{aligned} &\mathbb{P}\left\{ \| \frac{1}{2n^2} \sum_{i,j} (U^T X_i Y_j^T U + U^T Y_j X_i^T U) - U^T M M^T U \| > t \right\} \\ \leq\ & \sum_{k=1}^3 \mathbb{P}(\{\|A_k\| > t/3\} \cap \boldsymbol{\mathcal{E}}) + \mathbb{P}(\boldsymbol{\mathcal{E}}^c). \end{aligned}$$

By taking

$$t = C\alpha^2 \log^2(p) \frac{p_1 p_2 \|M\|_\infty^2}{n} \left(\mu r p_2^{1/2} + \frac{\mu r p_2}{n} + (\frac{\mu r n}{\log^3(p)})^{1/2} \right),$$

we conclude that

$$\mathbb{P}\left\{ \| \frac{1}{2n^2} \sum_{i,j} (U^T X_i Y_j^T U + U^T Y_j X_i^T U) - U^T M M^T U \| > t \right\} \leq 7p^{-\alpha}.$$

And the proof is finalized by replacing α with $\alpha + \log_p(7)$ and adjusting the constant C accordingly.

C.10 Proof of Lemma 26

First denote $Z_i = U^T(p_1p_2\mathcal{P}_{\omega_i}(M) - M)$. Then we have $U^T(\frac{p_1p_2}{n}\mathcal{P}_\Omega(M) - M) = \frac{1}{n}\sum_{i=1}^n Z_i$. Notice that $\|p_1p_2U^T\mathcal{P}_{\omega_i}(M)\| \le \sqrt{p_1}p_2\sqrt{\mu r}\|M\|_\infty$. And this implies that

$$\|Z_i\| \le 2\sqrt{p_1}p_2\sqrt{\mu r}\|M\|_\infty.$$

On the other hand, we have

$$\mathbb{E}Z_iZ_i^T \le p_1p_2U^T(\sum_{i,j} M_{ij}^2 e_ie_i^T)U \le p_1p_2^2\|M\|_\infty^2 I_r,$$

and

$$\mathbb{E}Z_i^TZ_i \le \mu r p_1p_2\|M\|_\infty^2 I_{p_2}.$$

So $\|\mathbb{E}\sum_{i=1}^n Z_iZ_i^T\| \vee \|\mathbb{E}\sum_{i=1}^n Z_i^TZ_i\| \le np_1p_2\|M\|_\infty^2(\mu r \vee p_2)$. Using matrix Bernstein inequality and we have with probability exceeding $1 - p^{-\alpha}$,

$$\|U^T(p_1p_2\mathcal{P}_{\omega_i}(M) - M)\| \le C\alpha\left(\frac{\sqrt{p_1}p_2\sqrt{\mu r}\|M\|_\infty\log(p)}{n} + \sqrt{\frac{p_1p_2\|M\|_\infty^2(\mu r \vee p_2)\log(p)}{n}}\right).$$