Low-tubal-rank Tensor Completion using Alternating Minimization

Xiao-Yang Liu, Shuchin Aeron, Vaneet Aggarwal, Xiaodong Wang

Abstract

The low-tubal-rank tensor model has been recently proposed and recognized superior for modeling real-world multidimensional data. In this paper, we study the low-tubal-rank tensor completion problem, i.e., to recover a third-order tensor by observing a subset of elements selected uniform at random. We propose a fast iterative algorithm, called *Tubal-Alt-Min*, that is inspired by similar approach for low-rank matrix completion. The unknown low-tubal-rank tensor is parameterized as the product of two much smaller tensors with the low-tubal-rank property being automatically incorporated, and Tubal-Alt-Min alternates between estimating those two tensors using tensor least squares minimization. First, we note that tensor least squares minimization is different from its counterpart and nontrivial as the circular convolution operator of the low-tubal-rank tensor model is intertwined with the sub-sampling operator, and we introduce a routine to carry out this operation. Second, the performance guarantees are of interest while challenging since Tubal-Alt-Min is iterative and nonconvex in nature. We prove that 1) Tubal-Alt-Min guarantees exponential convergence to the global optima, and 2) for an $n \times n \times k$ tensor with tubal-rank $r \ll n$, the required sampling complexity is $O(nr^2k \log n)$ and the computational complexity is $O(n^2rk^2 \log^2 n)$. Third, on both synthetic data and real-world video data, evaluation results show that compared with tensor nuclear norm minimization (TNN-ADMM), Tubal-Alt-Min improves the recovery error dramatically (by orders of magnitude). Tubal-Alt-Min's geometric convergence rate is estimated to be 0.3612 that is much faster than TNN-ADMM's 0.9285, and the running time can be accelerated by more than 5 times for a $200 \times 200 \times 20$ tensor.

X.-Y. Liu and X. Wang are with the Department of Electrical Engineering, Columbia University.

X.-Y. Liu is with the Department of Computer Science and Engineering, Shanghai Jiao Tong University.

S. Aeron is with the Department of Electrical and Computer Engineering, Tufts University.

V. Aggarwal is with the School of Industrial Engineering, Purdue University.

I. INTRODUCTION

The big data era calls for efficient algorithms to analyze the enormous amount of data generated by high-resolution sensors, mobile devices, online merchants, and social networks [1]. Such real-world data/signals[1] are naturally represented as multidimensional arrays [2], namely, vectors, matrices, high-order tensors or tensor networks. Signal recovery from partial measurements [3], exploiting the redundancy property modeled as sparse or low-rank, receives wide attention in various research and engineering communities. We are interested in fast algorithms for the problem of multilinear data completion when the measurement procedure being modelled as a simple downsampling operation. Exemplar applications include MRI imaging [3], signal processing [2], big data analysis with missing entries [4], data privacy [5], network engineering [6–9], Internet of Things [10, 11], machine learning [12], computer vision [13, 14], recommender system [15] and system identification [16].

Such diverse applications motivate/justify the developments of compressive sensing (vector case) [3, 17], matrix completion and matrix sensing [18, 19], and higher-order tensor completion [20–22]. Compressive sensing [3, 17] advocated relaxing the original NP-hard problem to its convex surrogate, i.e., replacing the ℓ_0-norm with ℓ_1-norm. Similarly, researchers introduced nuclear norm [23]/tensor-nuclear norm [24] to approximate the combinatorial *rank* function for the low-rank matrix/tensor completion problem[2]. Those two relaxation approaches achieve optimal results with computational sacrifice, mainly because of the time-consuming SVD (singular value decomposition) or tensor-SVD operations [24, 25].

Alternating minimization approaches have been proposed and widely used for the matrix completion problem [27–29]. First, it is both computation- and storage-efficient in implementation. The unknown low-rank matrix $M \in \mathbb{R}^{m\times n}$ is factorized into two much smaller matrices X and Y of size $m \times r$ and $n \times r$, respectively, i.e., $M = XY^\dagger$, and rank $r \ll \min(m, n)$ implying $(m + n)r \ll mn$, thus requiring much less computation and memory to optimize. Second, this factorization approach enables easier modeling. Besides the low-rank property, this factorization approach allows one to impose extra constraints on the target matrix M or factors (X, Y). For example, Sparse PCA [26] seeks for a low-rank M that is the product of *sparse* X and Y. Third, it converges to the global optima at a geometric rate, such theoretic results are available only very recently [27–29].

However, extending existing alternating minimization algorithms (originally designed for the matrix case) [7, 10, 27–29] to higher-order tensors is impeded by three major challenges: 1) there exists different

[1]In the following, we use the words "signal" and "data" interchangeably.

[2]A vector is a first-order tensor while a matrix is a second-order tensor.

definitions for tensor operators and thus lead to different low-rank tensor models, i.e., the CP-rank tensor [x], the Tuker-rank tensor [x] and the low-tubal-rank tensor [x], 2) existing approaches would be rather inefficient for higher-order tensors due to the curse of dimensionality, and 3) those algorithms do not guarantee good theoretical performance.

In this paper, we address those challenges for the third-order tensor completion problem. More specifically, we are interested in the low-tubal-rank tensor model that shares a similar algebraic structure with the low-rank matrix model. Our goal is to design a fast algorithm under the alternating minimization framework, and hopefully it will have good theoretical performance. We believe this approach would be a breakthrough point for higher-order tensors due to the following three perspectives:

- The low-tubal-rank tensor model [21, 22] is recently proposed and recognized superior for modeling multilinear real-world data, such as WiFi fingerprints [6], images [30], videos [25], seismic data [31], and machine learning [30]. There is a "spatial-shifting" property in those data, and we believe it is ubiquitous in real-world data arrays. The low-tubal-rank tensor model is superior in capturing this kind of characteristic.
- Although being iterative and nonconvex in nature, alternating minimization can be much faster than its convex counterparts. The potential efficiency comes from the fact that it automatically incorporates the low-rank property, resulting in massive dimension reduction. Note that efficiency is critical for processing big data.
- It has been proved that alternating minimization achieves global optimal at a geometric rate for matrix completion [27–29]. Obeying similar algebra laws, such performance guarantees may hold for higher-order tensors.

First, we propose a fast alternating minimization algorithm, *Tubal-Alt-Min*, for the low-tubal-rank tensor completion problem. A key novelty is solving a least squares minimization for tensors by defining a new set of operators, which can be of independent interest. Tensor least squares minimization is different from the standard least squares minimization because the circular convolution operator of the low-tubal-rank tensor model is intertwined with the sub-sampling operator. Therefore, the tensor completion problem is essentially different from matrix completion, implying that existing alternating minimization algorithms [27–29] cannot be extended straightforwardly to our problem.

Second, although being non-convex, the proposed alternating minimization-based approach can be much faster than its convex counterpart, i.e., the tensor nuclear norm minimization with alternating direction method of multipliers (TNN-ADMM) [24, 25]. We prove that 1) the proposed algorithm guarantees

convergence to the global optima at a geometric rate, which is much faster than TNN-ADMM, and 2) for a tensor of size $n \times n \times k$ and tubal-rank $r \ll n$, the required sampling complexity is $O(nr^2k \log n)$ and the computational complexity is $O(n^2rk^2 \log^2 n)$. Please note that there is no constraint on the size of the third-dimension. The proof is done by exploiting an injective mapping between the *circulant algebra* and the *circular matrix space*.

Third, we evaluate Tubal-Alt-Min on both synthetic data and real-world video data. The performances are measured in terms of recovery error, convergence rate, and running time. Compared with the convex relaxation-based algorithm TNN-ADMM [24, 25], Tubal-Alt-Min improves the recovery error by one order at sampling rate 50% for synthetic data, and three orders for the video data, respectively. Tubal-Alt-Min converges to the global optima with a constant number of iterations and the convergence rate is estimated to be 0.3612, which is much faster than TNN-ADMM's rate 0.9285. The running time can be accelerated by more than 5 times for a $200 \times 200 \times 20$ tensor.

The reminder of the paper is organized as follows. In Section II, we present the low-tubal-rank tensor model and the required preliminaries of the circulant algebra. Section III describes the low-tubal-rank tensor completion problem and the proposed *Tubal-Alt-Min* algorithm, including a novel routine to carry out the key subproblem: tensor least squares minimization. Section IV provides performance guarantees of the Tubal-Alt-Min algorithm, while detailed proofs are given in the Appendix. In Section V, we evaluate the proposed scheme on both synthetic data and real-world video data. The conclusions and future works are given in Section VI.

II. Notations and Preliminaries

We begin by first outlining the notations, the algebraic models and preliminary results for third-order tensors as in [21, 22, 32, 34]. Note that Definition 9, 15, 16 and Lemma 2, 4 are newly introduced. The first subsection presents the low-tubal-rank tensor model, used for problem formulation and algorithm design in Section III. The second part presents preliminaries of the circulant algebra, used for performance analysis in Section IV and the Appendix.

For computation and sample complexity, the following asymptotic notations are used throughout this paper. Given non-negative functions $f(n)$ and $g(n)$[3]:

- $f(n) = \Theta(g(n))$ means for two constants $0 < c_1 < c_2$, $c_1 g(n) \leq f(n) \leq c_2 g(n)$ for sufficiently large n.

[3]Note that Θ and Ω are reused in the following, whose meanings will be clear from the context.

- $f(n) = O(g(n)))$ means $\lim_{n\to\infty} \frac{f(n)}{g(n)} \to \infty$.
- $f(n) = \Omega(g(n)))$ means $\lim_{n\to\infty} \frac{f(n)}{g(n)} = 0$.

A. Low-tubal-rank Tensor Model

Throughout the paper, we will focus on real value third-order tensors in the space $\mathbb{R}^{m\times n\times k}$. We use $m,\ n,\ k,\ r$ for tensor dimensions, $x,\ y \in \mathbb{R}$ for scalar variables, **x, y** $\in \mathbb{R}^n$ for vectors, and $X,\ Y \in \mathbb{R}^{m\times n}$ for matrices. Tensors are denoted by calligraphic letters and their corresponding circular matrices (defined in Section II-B) are tagged with the superscript c, i.e., $\mathcal{T} \in \mathbb{R}^{m\times n\times k},\ \mathcal{X} \in \mathbb{R}^{m\times r\times k},\ \mathcal{Y} \in \mathbb{R}^{n\times r\times k}$ and $T^c \in \mathbb{R}^{mk\times nk},\ X^c \in \mathbb{R}^{mk\times rk},\ Y^c \in \mathbb{R}^{nk\times rk}$.

Let $X^\dagger$ denote the transpose of matrix X. We use i, j, κ to index the first, second and third dimension of a tensor, and s, t for temporary indexing. $[n]$ denotes the set $\{1, 2, ..., n\}$. Usually, $i \in [m], j \in [n], \kappa \in [k]$ unless otherwise specified. For tensor $\mathcal{T} \in \mathbb{R}^{m\times n\times k}$, the (i, j, κ)-th entry is $\mathcal{T}(i, j, \kappa)$, or concisely represented as $\mathcal{T}_{ij\kappa}$. The ℓ_2-norm of a vector is defined as $||\mathbf{x}||_2 = \sqrt{\sum_{i=1} \mathbf{x}_i^2}$, while the Frobenius norm of a matrix X is $||X||_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n X_{ij}^2}$ and a tensor is $||\mathcal{T}||_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n \sum_{\kappa=1}^k \mathcal{T}_{ij\kappa}^2}$.

Tubes/fibers, and slices of a tensor: A *tube* (also called a fiber) is a 1-D section defined by fixing all indices but one, while a *slice* is a 2-D section defined by fixing all but two indices. We use $\mathcal{T}(:, j, \kappa),\ \mathcal{T}(i, :, \kappa),\ \mathcal{T}(i, j, :)$ to denote the mode-1, mode-2, mode-3 tubes, which are vectors, and $\mathcal{T}(:, :, \kappa),\ \mathcal{T}(:, j, :),\ \mathcal{T}(i, :, :)$ to denote the frontal, lateral, horizontal slices, which are matrices. For easy representation in some formulas, we denote $\mathcal{T}^{(i)} = \mathcal{T}(:, :, i)$.

Tensor transpose and frequency domain representation: $\mathcal{T}^\dagger$ is obtained by transposing each of the frontal slices and then reversing the order of transposed frontal slices 2 through k, i.e., for $2 \le \kappa \le k$, $\mathcal{T}^\dagger(:, :, \kappa) = \mathcal{T}^{\dagger(\kappa)} = (\mathcal{T}(:, :, k + 2 - \kappa))^\dagger$ (the transpose of matrix $\mathcal{T}(:, :, k + 2 - \kappa)$). For reasons to become clear soon, we define a tensor $\widetilde{\mathcal{T}}$, which is the representation in the frequency domain and is obtained by taking the Fourier transform along the third mode of $\mathcal{T}$, i.e., $\widetilde{\mathcal{T}}(i, j, :) = \text{fft}(\mathcal{T}(i, j, :))$. In MATLAB notation, $\widetilde{\mathcal{T}} = \text{fft}(\mathcal{T}, [\,], 3)$, and one can also compute $\mathcal{T}$ from $\widetilde{\mathcal{T}}$ via $\mathcal{T} = \text{ifft}(\widetilde{\mathcal{T}}, [\,], 3)$.

We now define the linear algebraic development [22] for the low-tubal-rank tensor model. It rests on defining a tensor-tensor product between two 3-D tensors, referred to as the t-product as defined below. For two tubes/vectors of the same size, i.e., **a, b** $\in \mathbb{R}^k$, let **a** $*$ **b** denote the *circular convolution* between these two vectors, which preserves the size. Next, we describe the definition of tensor product and following definitions under this construction.

Definition 1. ***t-product****. The tensor-product $\mathcal{C} = \mathcal{A} * \mathcal{B}$ of $\mathcal{A} \in \mathbb{R}^{n_1\times n_2\times k}$ and $\mathcal{B} \in \mathbb{R}^{n_2\times n_3\times k}$ is a tensor*

of size $n_1 \times n_3 \times k$, $\mathcal{C}(i,j,:) = \sum_{s=1}^{n_2} \mathcal{A}(i,s,:) * \mathcal{B}(s,j,:)$, *for* $i \in [n_1]$ *and* $j \in [n_3]$.

A 3-D tensor of size $n_1 \times n_2 \times k$ can be viewed as an $n_1 \times n_2$ matrix of tubes which lie in the third-dimension. So the t-product of two tensors can be regarded as a matrix-matrix multiplication, except that the operation between scalars is replaced by circular convolution between two tubes. Therefore, the two operators, i.e, element-wise addition and the t-product, and the space $\mathbb{R}^{n\times n\times k}$ together define an *Abelian group* [34].

Definition 2. ***Identity tensor***. *The identity tensor* $\mathcal{I} \in \mathbb{R}^{n\times n\times k}$ *is a tensor whose first frontal slice* $\mathcal{I}(:,:,1)$ *is the* $n \times n$ *identity matrix and all other frontal slices* $\mathcal{I}^{(i)},\ (i = 2, ..., k)$ *are zero.*

Definition 3. ***Orthogonal tensor***. *A tensor* $\mathcal{Q} \in \mathbb{R}^{n\times n\times k}$ *is orthogonal if it satisfies* $\mathcal{Q}^\dagger * \mathcal{Q} = \mathcal{Q} * \mathcal{Q}^\dagger = \mathcal{I}$.

Definition 4. ***Inverse***. *The inverse of a tensor* $\mathcal{T} \in \mathbb{R}^{n\times n\times k}$ *is written as* $\mathcal{T}^{-1} \in \mathbb{R}^{n\times n\times k}$ *and satisfies* $\mathcal{T}^{-1} * \mathcal{T} = \mathcal{T} * \mathcal{T}^{-1} = \mathcal{I}$.

Definition 5. ***Block diagonal form of third-order tensor***. *Let* $\overline{\mathcal{A}}$ *denote the block-diagonal matrix representation of the tensor* $\mathcal{A}$ *in the Fourier domain, i.e.,*

$$\overline{\mathcal{A}} \triangleq \textit{blkdiag}(\hat{\mathcal{A}}) \triangleq \begin{bmatrix} \hat{\mathcal{A}}^{(1)} & & & \\ & \hat{\mathcal{A}}^{(2)} & & \\ & & \cdots & \\ & & & \hat{\mathcal{A}}^{(k)} \end{bmatrix} \in \mathbb{C}^{mk\times nk}. \tag{1}$$

It is easy to verify that the block diagonal matrix of $\mathcal{A}^\dagger$ *is equal to the transpose of the block diagonal matrix of* $\mathcal{A}$, *i.e.,* $\overline{\mathcal{A}^\dagger} = \overline{\mathcal{A}}^\dagger$.

Remark 1. *The following fact will be used throughout the paper for calculating tensor products and also tensor inverse. For tensors* $\mathcal{A} \in \mathbb{R}^{n_1\times n_2\times k}$ *and* $\mathcal{B} \in \mathbb{R}^{n_2\times n_3\times k}$, *we have*

$$\mathcal{A} * \mathcal{B} = \mathcal{C} \Longleftrightarrow \overline{\mathcal{A}\mathcal{B}} = \overline{\mathcal{C}}. \tag{2}$$

Definition 6. ***f-diagonal tensor***. *A tensor is called f-diagonal if each frontal slice of the tensor is a diagonal matrix, i.e.,* $\Theta(i,j,\kappa) = 0$ *for* $i \neq j, \forall \kappa$.

Using all these definitions, one can obtain the t-SVD [21] (tensor singular value decomposition) for dimensionality reduction of third order data. Please see Fig. 1 for a graphical representation of the (reduced) t-SVD in Remark 2.

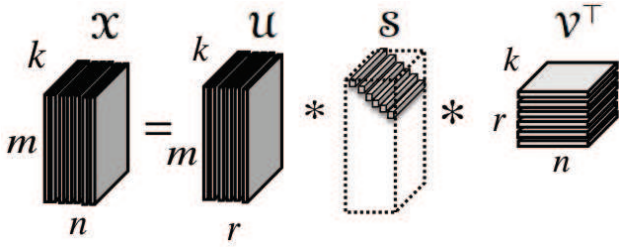

Fig. 1. The (reduced) t-SVD of an $m \times n \times k$ tensor of tubal-rank r.

Definition 7. ***t-SVD***. *For $\mathcal{T} \in \mathbb{R}^{m\times n\times k}$, the t-SVD of $\mathcal{T}$ is given by $\mathcal{T} = \mathcal{U} * \Theta * \mathcal{V}^\dagger$, where $\mathcal{U}$ and $\mathcal{V}$ are orthogonal tensors of sizes $m \times m \times k$ and $n \times n \times k$, respectively. Θ is a f-diagonal tensor of size $m \times n \times k$ and the tubes are called the eigentubes of $\mathcal{T}$. This is depicted in Fig. 1. An algorithm for computing the t-SVD is outlined as Alg. 1.*

Algorithm 1 t-SVD

Input: $\mathcal{X} \in \mathbb{C}^{n_1\times n_2\times n_3}$

$\widetilde{\mathcal{X}} \leftarrow \texttt{fft}(\mathcal{X}, [\,], 3)$; %Take DFT along 3-rd direction

for $i = 1$ to n_3 **do**

$[\hat{U}, \hat{S}, \hat{V}] = \texttt{SVD}(\widetilde{\mathcal{X}}(:, :, i))$;

$\widetilde{\mathcal{U}}^{(i)} = \hat{U}; \widetilde{\mathcal{S}}^{(i)} = \hat{S}; \widetilde{\mathcal{V}}^{(i)} = \hat{V}$;

end for

$\mathcal{U} \leftarrow \texttt{ifft}(\widetilde{\mathcal{U}}, [\,], 3)$; $\mathcal{S} \leftarrow \texttt{ifft}(\widetilde{\mathcal{S}}, [\,], 3)$; $\mathcal{V} \leftarrow \texttt{ifft}(\widetilde{\mathcal{V}}, [\,], 3)$.

Definition 8. ***Tensor tubal-rank***. *The tensor tubal-rank of a third-order tensor $\mathcal{T}$ is the number of non-zero tubes of Θ in the t-SVD, denoted as r.*

Remark 2. *Suppose $\mathcal{T}$ has tubal-rank r, then the reduced t-SVD of $\mathcal{T}$ is given by $\mathcal{T} = \mathcal{U} * \Theta * \mathcal{V}^\dagger$, where $\mathcal{U} \in \mathbb{R}^{m\times r\times k}$ and $\mathcal{V} \in \mathbb{R}^{n\times r\times k}$ satisfying $\mathcal{U}^\dagger * \mathcal{U} = \mathcal{I}$, $\mathcal{V}^\dagger * \mathcal{V} = \mathcal{I}$, and Θ is a f-diagonal tensor of size $r \times r \times k$. This reduced version of t-SVD will be used throughout the paper unless otherwise noted.*

Lemma 1. ***Best rank-r approximation***. *Let the t-SVD of $\mathcal{T} \in \mathbb{R}^{m\times n\times k}$ be $\mathcal{T} = \mathcal{U} * \Theta * \mathcal{V}^\dagger$ and positive integer r, define $\mathcal{T}_r = \sum_{s=1}^{r} \mathcal{U}(:, s, :) * \Theta(s, s, :) * \mathcal{V}^\dagger(:, s, :)$, then $\mathcal{T}_r = \arg \min_{\overline{\mathcal{T}}\in\mathbb{T}} ||\mathcal{T} - \overline{\mathcal{T}}||_F$, where $\mathbb{T} = \{\mathcal{X} * \mathcal{Y}^\dagger | \mathcal{X} \in \mathbb{R}^{m\times r\times k}, \mathcal{Y} \in \mathbb{R}^{n\times r\times k}\}$.*

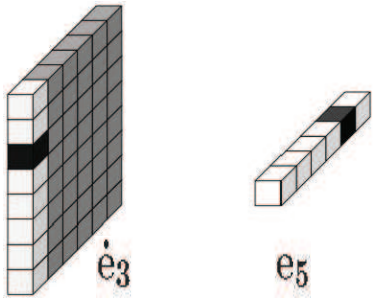

Fig. 2. The column basis $\dot{e}_3$ and tubal basis e_5. The black entries are 1, gray and white entries are 0. The white entries are those that could be 1.

Note that Θ in t-SVD is organized in a decreasing order, i.e., $||\Theta(1,1,:)||_2 \geq ||\Theta(2,2,:)||_2 \geq ...$, which is implicitly defined in [22]. Therefore, the best rank-r approximation of tensors is similar to PCA (principal component analysis).

For easy description of the proposed Tubal-Alt-Min algorithm, we define the *symmetric square tensor* in the following. Similar to the matrix case [29], our results generalize straightforwardly to *rectangular tensors*.

Definition 9. ***Square tensor, rectangular tensor, symmetric square tensor***. *A tensor* $\mathcal{T} \in \mathbb{R}^{n\times n\times k}$ *is a* square tensor, *and a tensor* $\mathcal{X} \in \mathbb{R}^{n\times r\times k}$ *is a* rectangular tensor. *A* symmetric square tensor $\mathcal{T}$ *is a tensor whose frontal slices are symmetric square matrices, i.e.,* $\mathcal{T} \in \mathbb{T} = \{\mathcal{X} * \mathcal{X}^\dagger | \mathcal{X} \in \mathbb{R}^{n\times r\times k}\}$.

Definition 10. ***Tensor basis and the corresponding decomposition***. *We introduce two tensor bases [24]. The first one is called* column basis $\dot{e}_i$ *of size* $n \times 1 \times k$ *with only one entry equaling* 1 *and the rest equaling* 0. *Note that this nonzero entry* 1 *will only appear at the* i*-th entry of the first frontal slice of* $\dot{e}_i$. *Naturally, its transpose* $\dot{e}_i^\dagger$ *is called* row basis. *The second tensor basis is called* tubal basis e_i *of size* $1 \times 1 \times k$ *with one entry equaling to* 1 *and rest equaling* 0. *Fig. 2 illustrates these two bases.*

With the above two bases, one can obtain a unit tensor $\mathcal{E}$ with only non-zero entry $\mathcal{E}_{ij\kappa}$ equaling 1 via the following:

$$\mathcal{E} = \dot{e}_i * e_j * \dot{e}_\kappa^\dagger. \tag{3}$$

Given any third order tensor $\mathcal{T} \in \mathbb{R}^{m\times n\times k}$, we have the following decomposition

$$\mathcal{X} = \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{\kappa=1}^{k} \mathcal{X}_{ij\kappa}\dot{e}_i * e_j * \dot{e}_\kappa^\dagger. \tag{4}$$

Definition 11. ***ℓ_{2^*}, $\ell_{\infty,2^*}$-norm of tensor column***. *Let $\mathcal{X}$ be an $m \times n \times k$ tensor column, we define an ℓ_{2^*}-norm of its j-th lateral slice $\mathcal{X}(:,j,:)$ as follows*

$$||\mathcal{X}(:,j,:)||_{2^*} = \sqrt{\sum_{i=1}^{n}\sum_{\kappa=1}^{k}\mathcal{X}_{ij\kappa}^2}, \quad ||\mathcal{X}||_{\infty,2^*} = \max_{j\in[n]}||\mathcal{X}(:,j,:)||_{2^*}. \tag{5}$$

Moreover, we have the following relationship between the ℓ_{2^*} norm of $\mathcal{X}(:,j,:)$ and its FFT along the third dimension $\mathcal{X}(:,j,:)$,

$$||\mathcal{X}(:,j,:)||_{2^*} = \frac{1}{\sqrt{k}}||\widetilde{\mathcal{X}}(:,j,:)||_{2^*}, \tag{6}$$

where $1/\sqrt{k}$ is the normalization constant.

Definition 12. ***Tensor spectral norm***. *The tensor spectral norm $||\mathcal{X}||$ of a third-order tensor $\mathcal{X} \in \mathbb{R}^{m\times n\times k}$ is defined as the largest singular value of $\mathcal{X}$*

$$||\mathcal{X}|| = \sup_{\mathcal{L}\in\mathbb{R}^{n\times n\times k},\ ||\mathcal{L}||_F\leq 1} ||\mathcal{X} * \mathcal{L}||_F. \tag{7}$$

Lemma 2. *The tensor spectral norm of $\mathcal{X}$ equals to the matrix spectral norm of $\overline{\mathcal{X}}$, and also that of the circular matrix X^c (defined in Section II-B), as follows:*

$$||\mathcal{X}|| = ||\overline{\mathcal{X}}|| = ||X^c||. \tag{8}$$

Definition 13. ***Tensor infinity norm***. *The tensor infinity norm $||\mathcal{A}||_\infty$ is defined as follows:*

$$||\mathcal{A}||_\infty = \max_{i,j,\kappa}|\mathcal{A}_{ij\kappa}|, \tag{9}$$

which is the entry with the largest absolute value of $\mathcal{A}$.

We define the corresponding *incoherence* of third-order tensors in our case which is used to guarantee unique solutions. The concept of incoherence is first introduced by [18, 19] for matrices and then a standard assumption for low-rank matrix/tensor completion problems.

Definition 14. ***Tensor Incoherent Condition***. *Given the t-SVD of a tensor $\mathcal{T} = \mathcal{U} * \Theta * \mathcal{V}^\dagger$ with tubal-rank r, $\mathcal{T}$ is said to satisfy the tensor incoherent condition, if there exists $\mu_0 > 0$ such that for $\kappa \in [k]$.*

$$\begin{aligned} &\textit{(Tensor-column incoherence)} \quad \mu(\mathcal{U}) \triangleq \frac{m}{r}\max_{i\in[m]}\left\|\mathcal{U}^\dagger * \dot{e}_i\right\|_{2^*}^2 \leq \mu_0, \\ &\textit{(Tensor-row incoherence)} \quad \mu(\mathcal{V}) \triangleq \frac{n}{r}\max_{j\in[n]}\left\|\mathcal{V}^\dagger * \dot{e}_j\right\|_{2^*}^2 \leq \mu_0, \end{aligned} \tag{10}$$

where $\dot{e}_i$ is the $m \times 1 \times k$ column basis with $\mathbf{e}_{i11} = 1$ and $\dot{e}_j$ is the $n \times 1 \times k$ column basis $\mathbf{e}_{j11} = 1$.

Remark 3. *The smallest μ_0 is 1 achieved by the case when each tensor column $\mathcal{U}(:,i,:)$ ($i \in [m]$) has elements with magnitude $\frac{1}{\sqrt{mk}}$, or each tensor column $\mathcal{U}(:,j,:)$ ($j \in [n]$) has elements with magnitude $\frac{1}{\sqrt{nk}}$. The largest possible value of μ_0 is $\min(m,n)/r$ when one of the tensor columns of $\mathcal{U}$ is equal to the standard tensor column basis $\dot{e}_i$. With low μ_0, each element of $\mathcal{T}$ is supposed to carry approximately same amount of information.*

Under t-SVD in Definition 7, a tensor-column subspace of $\mathcal{T}$ is the space spanned by the lateral slices of $\mathcal{U}$ under the t-product, i.e., the set generated by r-linear combinations, t-span$(\mathcal{U}) = \{\mathcal{X} = \sum_{s=1}^{r} \mathcal{U}(:,s,:) * \mathbf{c}_s \in \mathbb{R}^{n\times 1\times k},\ \mathbf{c}_s \in \mathbb{R}^{1\times 1\times k}\}$, where r denotes the tensor tubal-rank.

For simple description of the implementation of the tensor least squares minimization, we need to define three kinds of new products as follows

Definition 15. ***Element-wise tensor product, tube-wise circular convolution, frontal-slice-wise tensor product*** *Let $\mathcal{T} \in \mathbb{R}^{m\times n\times k}, \mathcal{T}_1,\ \mathcal{T}_2 \in \mathbb{R}^{m\times n\times k}$. The element-wise tensor product $\odot$ operates on two same-sized tensors and results in a same-size tensor, i.e., $\mathcal{T} = \mathcal{T}_1 \odot \mathcal{T}_2$. The tube-wise circular convolution $\cdot\otimes$ operates on two same-sized tensors and results in a same-size tensor, i.e., $\mathcal{T} = \mathcal{T}_1 \cdot \otimes \mathcal{T}_2$. The frontal-slice-wise tensor product $\cdot\S$ is defined as performing matrix multiplication on the corresponding frontal slices of two tensors.*

Here, the frontal-slice-wise tensor product is introduced to have a concise third-order tensor representation of Remark 1, avoiding the block diagonal form representations as in Definition 5. The operator $\cdot\S$ is introduced to replace the block diagonal matrix in [24][25] since we need to preserve the three-way data array structure in Section III-D.

B. Circulant Algebra

The circulant algebra is recently introduced to generalize matrix algebra to the third-order tensor case. We borrow some notations and operations from [32, 34], meanwhile we propose several new definitions to facilitate our analysis in the Appendix.

Throughout the paper, circulants are denoted by underlined letters. We define *tubal scalar*, *tubal vector*, and *tubal matrix* in the following. Note that they are one-dimension higher than their counterparts in traditional linear algebra. In circulant algebra, a tubal scalar is a vector of length k. $\mathbb{K}_k$ denotes the space of length-k *tubal scalars*, $\mathbb{K}_k^n$ denotes the space of *tubal vectors* where each element is a tubal scalar in $\mathbb{K}_k$, and $\mathbb{K}_k^{m\times n}$ denotes the space of $m \times n$ *tubal matrices* where each element is a tubal scalar in $\mathbb{K}_k$.

We use $\underline{\alpha}, \underline{\beta} \in \mathbb{K}_k$ for tubal scalars, $\underline{\mathbf{x}}, \underline{\mathbf{y}} \in \mathbb{K}_k^n$ for tubal vectors, and $\underline{A},\ \underline{B} \in \mathbb{K}_k^{m\times n}$ for tubal matrices. Their corresponding circular matrices are tagged with the superscript c, i.e., $\underline{\alpha}^c, \underline{\beta}^c, \underline{\mathbf{x}}^c, \underline{A}^c$.

D. Gleich, *et al* [32] introduced the operator $\text{circ}(\cdot)$ to map circulants to their corresponding circular matrix representations. For tubal scalar $\alpha = \{a_1, a_2, ..., a_k\} \in \mathbb{K}_k$, tubal vector $\underline{\mathbf{x}} \in \mathbb{K}_k^n$, and tubal matrix $\underline{A} \in \mathbb{K}_k^{m\times n}$, we use the notation $\leftrightarrow$ to denote this mapping as follows:

$$\begin{aligned}
\underline{\alpha} \quad &\leftrightarrow \quad \underline{\alpha}^c = \text{circ}(\underline{\alpha}) = \begin{bmatrix} \alpha_1 & \alpha_n & ... & \alpha_2 \\ \alpha_2 & \alpha_1 & ... & ... \\ ... & ... & ... & \alpha_k \\ \alpha_k & ... & \alpha_2 & \alpha_1 \end{bmatrix}, \qquad \underline{\mathbf{x}} \quad \leftrightarrow \quad \underline{\mathbf{x}}^c = \text{circ}(\underline{\mathbf{x}}) = \begin{bmatrix} \text{circ}(\underline{x}_1) \\ \vdots \\ \text{circ}(\underline{x}_n) \end{bmatrix}, \\
\underline{A} \quad &\leftrightarrow \quad \underline{A}^c = \text{circ}(\underline{A}) = \begin{bmatrix} \text{circ}(\underline{A}_{1,1}) & ... & \text{circ}(\underline{A}_{1,n}) \\ \vdots & \vdots & \vdots \\ \text{circ}(\underline{A}_{m,1}) & ... & \text{circ}(\underline{A}_{m,n}) \end{bmatrix}.
\end{aligned} \tag{11}$$

Lemma 3. *[34]* $(\mathbb{K}_k^n, \pm, *)$ *is a commutative ring with unity, where* $\pm$ *and* $*$ *denotes the addition/subtraction and circular convolution. We have:*

$$\begin{aligned}
\underline{\alpha} + \underline{\beta} \quad &\leftrightarrow \quad \underline{\alpha}^c + \underline{\beta}^c = \textit{circ}(\underline{\alpha}) + \textit{circ}(\underline{\beta}), \\
\underline{\alpha} * \underline{\beta} \quad &\leftrightarrow \quad \underline{\alpha}^c \underline{\beta}^c = \textit{circ}(\underline{\alpha})\ \textit{circ}(\underline{\beta}), \\
\underline{\boldsymbol{x}} * \underline{\alpha} \quad &\leftrightarrow \quad \underline{\boldsymbol{x}}^c \underline{\alpha}^c = \textit{circ}(\underline{\boldsymbol{x}})\ \textit{circ}(\underline{\alpha}), \\
\underline{A} * \underline{\boldsymbol{x}} \quad &\leftrightarrow \quad \underline{A}^c \underline{\boldsymbol{x}}^c = \textit{circ}(\underline{A})\ \textit{circ}(\underline{\boldsymbol{x}}),
\end{aligned} \tag{12}$$

where $\underline{1} = \{1\ 0\ ...\ 0\}$ *is the multiplicative identity.*

Note that the tensor-column subspace iteration in Alg. 7 relies on definitions of inverse, angle function, inner products, norm, conjugate, and also the circulant Fourier transform [32]. We describe the inverse and the circulant Fourier transform while omit the rest since they are implicitly used in our paper. The inverse of $\underline{\alpha} \in \mathbb{K}_k$ is $\alpha^{-1} \ \leftrightarrow\ \text{circ}(\underline{\alpha})^{-1}$, where $\text{circ}(\underline{\alpha})^{-1}$ is also a circulant. The circulant Fourier transforms, $\text{cft} : \underline{\alpha} \in \mathbb{K}_k \mapsto \mathbb{C}^{k\times k}$ and its inverse $\text{icft} : \mathbb{C}^{k\times k} \mapsto \mathbb{K}_k$, are defined as follows:

$$\text{cft}(\underline{\alpha}) \equiv= \begin{bmatrix} \hat{\alpha}_1 & & \\ & ... & \\ & & \hat{\alpha}_k \end{bmatrix} = \mathbf{F}^* \text{circ}(\underline{\alpha}) \mathbf{F}, \qquad \text{icft}\left(\begin{bmatrix} \hat{\alpha}_1 & & \\ & ... & \\ & & \hat{\alpha}_k \end{bmatrix} \right) \equiv \underline{\alpha} \leftrightarrow \mathbf{F}\text{cft}(\underline{\alpha}) \mathbf{F}^*, \tag{13}$$

where $\hat{\alpha}_\ell$ are the eigenvalues of $\text{circ}(\alpha)$ as produced in the Fourier transform order, $\mathbf{F}$ is the $k \times k$ discrete Fourier transform matrix, and $\mathbf{F}^*$ denotes the (circulant) conjugate [32].

Eigentubes and Eigenslices [32]: As in [32], we describe the eigentubes and eigenslices. The existence of an eigentube $\underline{\lambda} \in \mathbb{K}_k$ implies the existence of a corresponding eigenslice $\underline{A} \in \mathbb{K}_k^{n\times n}$, satisfying $\underline{A} * \underline{x} = \underline{x} * \underline{\lambda}$. The corresponding Fourier transforms $\text{cft}(\underline{A}), \text{cft}(\underline{x}), \text{cft}(\underline{\lambda})$ satisfy:

$$\begin{aligned} \text{cft}(\underline{A} * \underline{x}) &= \text{cft}(\underline{x} * \underline{\lambda}) \\ \text{cft}(\underline{A})\text{cft}(\underline{x}) &= \text{cft}(\underline{x})\text{cft}(\underline{\lambda}). \end{aligned} \tag{14}$$

Throughout this paper, we view a tensor in the space $\mathbb{R}^{m\times n\times k}$ as a tubal matrix in the space $\mathbb{K}_k^{m\times n}$. The tensors $\mathcal{T},\ \mathcal{X},\ \mathcal{Y}$ have circulant representations $\underline{T},\ \underline{X},\ \underline{Y}$ and circular matrix representations $T^c,\ X^c,\ Y^c$. We define the the Frobenius norm of a circulant as follows:

Lemma 4. *The Frobenius norm of a circulant equals to that of its tensor presentation, if $\mathcal{T} = \mathcal{X} * \mathcal{Y}$, then*

$$||\underline{T}||_F = ||\mathcal{T}||_F = ||\mathcal{X} * \mathcal{Y}||_F = \frac{1}{\sqrt{k}}||X^c Y^c||_F. \tag{15}$$

Definition 16. ***Tubal-wise transpose, circulant transpose**. Let $\mathcal{X}^T$ denote the the tube-wise transpose of $\mathcal{X}$, i.e., $\mathcal{X}^T(i,j,:) = \mathcal{X}(j,i,:)$. Similarly, let $\underline{X}^T$ denote the circulant transpose of $\underline{X}$, i.e., $\underline{X}^T(i,j) = \underline{X}(j,i)$, which can be viewed as the transpose of a matrix of vectors.*

III. Problem Statement and Proposed Algorithm

We first describe the low-tubal-rank tensor completion problem. Then, we present our Tubal-Alt-Min algorithm followed by detailed descriptions of key routines. Finally, we provide a way to implement the tensor least squares minimization.

A. Problem Statement

We consider the problem of completing a 3-D tensor under the assumption that the 3-D tensor has low-tubal-rank. Specifically, assume that the data tensor $\mathcal{T} \in \mathbb{R}^{m\times n\times k}$ has tubal-rank $r \ll \min(m,n)$. By observing a set $\Omega \subset [m]\times[n]\times[k]$ of $\mathcal{T}$'s elements, our aim is to recover $\mathcal{T}$. That is, knowing the elements $\mathcal{T}_{ij\ell}$ for $(i,j,\ell) \in \Omega$, we want to estimate the elements outside of Ω as accurately as possible.

Let $\mathcal{P}_\Omega(\cdot)$ denote the projection of a tensor onto the observed set Ω, such that

$$[\mathcal{P}_\Omega(\mathcal{T})]_{ij\ell} = \begin{cases} \mathcal{T}_{ij\ell}, & \text{if } (i,j,\ell) \in \Omega, \\ 0, & \text{otherwise,} \end{cases}$$

where the (i,j,ℓ)-th entry of $\mathcal{P}_\Omega(\mathcal{X})$ equals to $\mathcal{X}_{ij\ell}$ if $(i,j,\ell) \in \Omega$ and zero otherwise. Since $\mathcal{T}$ is known to be a low-tubal-rank tensor and the estimated $\hat{\mathcal{T}}$ should be in consistent with $\mathcal{T}$ on the set Ω, the *low-tubal-rank tensor completion problem* is formulated as the following optimization problem:

$$\begin{aligned} \widehat{\mathcal{T}} = \arg \min_{\mathcal{X} \in \mathbb{R}^{m\times n\times k}} \ & \text{rank}(\mathcal{X}) \\ s.t. \ \ & \mathcal{P}_\Omega(\mathcal{X}) = \mathcal{P}_\Omega(\mathcal{T}) \end{aligned} \tag{16}$$

where $\mathcal{X} \in \mathbb{R}^{m\times n\times k}$ is the decision variable, and the function rank$(\cdot)$ refers to the tensor tubal-rank. Problem (16) is NP-hard since rank induces combinatorial complexity and existing works [24, 25] seek to relax the rank function to its convex surrogate, namely, the tensor-nuclear norm. In [24], it was shown that given a sufficient number of observations ($|\Omega| = O(\max\{m,n\}kr \log mn)$), it was proved that tensor-nuclear norm minimization results in exact recovery under random sampling if the tensors satisfy certain tensor incoherency conditions.

However, the computational cost of the algorithm in [24] is relatively high due to two key factors: 1) each iteration requires computations of SVD for large block diagonal matrices, and 2) the iterations are jointly done in both the time domain and frequency domain, thus involving frequent and large number of Fourier and inverse Fourier transforms. Therefore, in the following section, we will introduce an alternating minimization algorithm, inspired by the alternating minimization approach's empirical and theoretical successes in low-rank matrix completion [27–29].

Taking into consideration of the noise, the linear constraint in (16) is replaced by a quadratical constraint $||\mathcal{P}^\Omega(\mathcal{X}-\mathcal{T})||_F \leq \varrho$ where ϱ is a user-given prarameter. The *noisy low-tubal-rank tensor completion problem* is formulated as follows:

$$\begin{aligned} \widehat{\mathcal{T}} = \arg \min_{\mathcal{X} \in \mathbb{R}^{m\times n\times k}} \ & \text{rank}(\mathcal{X}) \\ s.t. \ \ & \| ||\mathcal{P}^\Omega(\mathcal{X}-\mathcal{T})||_F \leq \varrho \end{aligned} \tag{17}$$

Note that this noisy case is inherently included in our alternating minimization algorithm since its key process, the tensor least squares minimization, deals with noise.

B. The Alternating Minimization Algorithm for Low-tubal-rank Tensor Completion

The pseudo code of our alternating minimization algorithm is given in Alg. 2. In essence, it is an approximate solution for the following nonconvex problem:

$$\widehat{\mathcal{X}} = \arg \min_{\mathcal{X} \in \mathbb{R}^{m\times r\times k},\ \mathcal{Y} \in \mathbb{R}^{n\times r\times k}} ||\mathcal{P}_\Omega(\mathcal{T}) - \mathcal{P}_\Omega(\mathcal{X} * \mathcal{Y}^\dagger)||_F^2, \tag{18}$$

Algorithm 2 Alternating Minimization: Tubal-Alt-Min$(\mathcal{P}_\Omega(\mathcal{T}), \Omega, L, r, t, \epsilon, \mu)$

Input: Observation set $\Omega \in [m] \times [n] \times [k]$ and the corresponding elements $\mathcal{P}_\Omega(\mathcal{T})$, number of iterations L, error parameter $\epsilon > 0$, target tubal-rank r, coherence parameter μ.

1: $(\Omega_0,\ \Omega_+) \leftarrow$ Split$(\Omega, 2)$,

2: $(\Omega_1, ...,\ \Omega_L) \leftarrow$ Split(Ω_+, L),

3: $\mathcal{X}_0 \leftarrow$ Initialize$(\mathcal{P}_{\Omega_0}(\mathcal{T}), \Omega_0, r, \mu)$,

4: For $\ell = 1$ to L

5: $\quad \mathcal{Y}_\ell \leftarrow$ MedianLS-Y$(\mathcal{P}_{\Omega_\ell}(\mathcal{T}), \Omega_\ell, \mathcal{X}_{\ell-1}, r, t)$,

6: $\quad \mathcal{Y}_\ell \leftarrow$ SmoothQR$(\mathcal{Y}_\ell, \epsilon, \mu)$,

7: $\quad \mathcal{X}_\ell \leftarrow$ MedianLS-X$(\mathcal{P}_{\Omega_\ell}(\mathcal{T}), \Omega_\ell, \mathcal{Y}_\ell, r, t)$,

8: $\quad \mathcal{X}_\ell \leftarrow$ SmoothQR$(\mathcal{X}_\ell, \epsilon, \mu)$,

Output: Tensor pair $(\mathcal{X}_L, \mathcal{Y}_L)$.

which finds a target tensor $\mathcal{T} \in \mathbb{R}^{m\times n\times k}$ that is parameterized as $\mathcal{T} = \mathcal{X} * \mathcal{Y}^\dagger$, $\mathcal{X} \in \mathbb{R}^{m\times r\times k}$, $\mathcal{Y} \in \mathbb{R}^{n\times r\times k}$, and r is the target tubal-rank. According to Lemma 1, we know that (18) is equivalent to the original problems, both (16) and (17), if there exists a unique tubal-rank-r tensor. Note that the uniqueness is guaranteed through the incoherence condition and the random sampling assumption [24, 25]. Note that the low-tubal-rank property is automatically incorporated in those two factors. However, it is challenging to analyze since the algorithm is iterative and nonconvex in nature. For simplicity, we use symmetric square tensor for analysis in the following, i.e., $\mathcal{T} \in \mathbb{R}^{n\times n\times k}$ and $\mathcal{T} = \mathcal{X} * \mathcal{X}^\dagger$, while our results can be easily extended to general tensors, following a similar way in [29].

The general flow of Alg. 2 is as follows.

- Line 1-2: Throughout the algorithm, the samples Ω are utilized for two purposes: initialize $\mathcal{X}_0$ as a "good" starting point, and to update $\mathcal{X}$ and $\mathcal{Y}$ in the iterations. A helper function, the Split function, first splits Ω into two same-sized subsets Ω_0, Ω_+ and then splits Ω_+ into L subsets of roughly equal size, while preserving the distributional assumption that our theorem uses.
- Line 3: Using the observed entries Ω_0, the Intialize procedure generates a good starting point that is relatively close to the optimal solution. This is required for analysis purpose, since from this starting point we are able to prove convergence.
- Line 4-8: The MedianLS minimization procedure relies on several basic *tensor least squares minimizations* as implemented in Section. III-D. For general tensors, each iteration includes two con-

secutive MedianLS minimizations for $\mathcal{X}$ and $\mathcal{Y}$, respectively. Note that $\mathcal{X}$ and $\mathcal{Y}$ need to be treated differently because of the t-product, which is not the case for the matrix completion [27–29].

An intuitive understanding of the proposed algorithm is that it alternates between finding the best $\mathcal{X}$ and the best $\mathcal{Y}$. Each step in isolation is essentially a tensor least squares update which is convex and tractable. In Section IV, we will prove that separating the nonconvex minimization (18) into two consecutive convex minimization subproblmes will also return the optimal solution of (16) and (17).

C. Key Routines

The key routines of Alg. 2 are: the Split function, the Initialize procedure, the median least squares minimization in Alg. 4, the tensor least squares minimization in Alg. 5, and the SmoothQR factorization in Alg. 6. In the following, we describe each one in details.

1) Splitting the Samples: The procedure Split(Ω, t) takes the sample set Ω and splits it into t independent samples $\Omega_1, ..., \Omega_t$ that preserve the uniform distribution assumption. The distribution that we assumed is simply each element is included independently with equal probability. For example, each elment of Ω belongs to one of those t sample sets by sampling with replacement. This split function is the same to the matrix case [27–29] since essentially they are both set operations.

Algorithm 3 Initialization Algorithm: Initialize$(\mathcal{P}_\Omega, \Omega, r, \mu)$

Input: observation set $\Omega \in \mathbb{R}^{[n]\times[n]\times[k]}$ and elements $\mathcal{P}_\Omega(\mathcal{T})$, target dimension r, coherence parameter $\mu \in \mathbb{R}$.

Compute the first r eigenslices $\mathcal{W} \in \mathbb{R}^{n\times r\times k}$ of $\mathcal{P}_\Omega(\mathcal{T})$,

$\mathcal{Z} \leftarrow \mathcal{W} * \mathcal{O}$ where $\mathcal{O} \in \mathbb{R}^{r\times r\times k}$ is a random orthonormal tensor,

$\mathcal{Z}' \leftarrow \text{Truncate}_{\mu'}(\mathcal{Z})$ with $\mu' = \sqrt{8\mu \log n/n}$, where $\text{Truncate}_{\mu'}$ scales the coefficients of a tube $\mathcal{Z}(i, j, :)$ (with Frobenius norm larger than μ') by $\mu'/||\mathcal{Z}(i, j, :)||_F$,

$\mathcal{X} \leftarrow \text{QR}(\mathcal{Z}')$,

Output: Orthonormal tensor $\mathcal{X} \in \mathbb{R}^{n\times r\times k}$.

2) Finding a Good Starting Point: Alg. 3 describes the procedure for finding a good starting point. Since it is unclear how well will the least squares minimization converges from a random initial tensor, we start with an initial tensor that has bounded distance from the optimal result. The algorithm serves a fast initialization procedure for our main algorithm. It computes the top-r eigenslices of $\mathcal{P}_\Omega(\mathcal{T})$, and truncates them in order to ensure incoherence. Note that the truncation for our tensor problem scales the

coefficients of a tube, which is different from the matrix case [29] that truncate an element. We use a random orthonormal transformation to spread out the tubes of the eigenslices before truncation.

3) Tensor Least Squares Minimization: In this section, we focus on describing the MedianLS iteration of $\mathcal{Y}$. Although the iteration for $\mathcal{X}$ is different from $\mathcal{Y}$, essentially it can be computed in a similar way as shown in Section III-D. Each median-LS minimization relies on the basic tensor least squares minimizations. Partitioning Ω_+ into $t = O(\log n)$ subsets, then performing the least squares minimization on each subset and than taking the median of the returned tensors. The median operation is taken in an elementwise manner.

Algorithm 4 Median Least Squares: MedianLS-Y$(\mathcal{P}_\Omega(\mathcal{T}), \Omega, \mathcal{X}, r, t)$

Input: target tubal-rank r, observation set $\Omega \in [n] \times [n] \times [k]$ and elements $\mathcal{P}_\Omega(\mathcal{T})$, orthonormal tensor $\mathcal{X} \in \mathbb{R}^{[m]\times[r]\times[k]}$.

$(\Omega_1, ..., \Omega_t) \leftarrow$ Split(Ω, t) for $t = O(\log n)$,

$\mathcal{Y}_i =$ LS$(\mathcal{P}_{\Omega_i}(\mathcal{T}), \Omega_i, \mathcal{X}, r)$ for $i \in [t]$,

Output: median$(\mathcal{Y}_1, ..., \mathcal{Y}_t)$.

Algorithm 5 Tensor Least Squares Minimization: LS$(\mathcal{P}_\Omega(\mathcal{T}), \Omega, \mathcal{X}, r)$

Input: target dimension r, observation set $\Omega \in [n] \times [n] \times [k]$ and elements $\mathcal{P}_\Omega(\mathcal{T})$, orthonormal tensor $\mathcal{X} \in \mathbb{R}^{[n]\times[r]\times[k]}$.

$\mathcal{Y} = \arg\min_{\mathcal{Y}\in\mathbb{R}^{[n]\times[r]\times[k]}} ||\mathcal{P}_\Omega(\mathcal{T} - \mathcal{X} * \mathcal{Y}^\dagger)||_F^2$,

Output: $\mathcal{Y}$.

4) Smooth QR: For the main theorem to hold, it is required that each iterates $\mathcal{X}_\ell$ and $\mathcal{Y}_\ell$ have coherence less than μ. To achieve this, we adopt the smooth operation, as shown in Alg. 6. Note the QR$(\cdot)$ and GS$(\cdot)$ operation for third-order tensors are defined in [22]. The QR$(\cdot)$ process returns the orthogonal projector while GS$(\cdot)$ make it to be orthonormal tensor-column space. The Gaussian perturbation $\mathcal{H}_\ell$ is added to $\mathcal{Y}_\ell$ to ensure small coherence. In our context, the tensor $\mathcal{X}_\ell$ and $\mathcal{Y}_\ell$ are the outcomes of a noisy operation $\mathcal{X}_\ell = \mathcal{A} * \mathcal{Y}_{\ell-1} + \mathcal{G}_\ell$ (in Appendix B), and so there is no harm in actually adding a Gaussian noise tensor $\mathcal{H}_\ell$ to $\mathcal{X}_\ell$ provided that the norm of that tensor is no larger than that of $\mathcal{G}_\ell$.

D. Implementation of Tensor Least Squares Minimization

For simplicity, denote $\mathcal{T}_\Omega = \mathcal{P}_\Omega(\mathcal{T})$ and let $\mathcal{P}_\Omega$ be the sampling tensor with ones at places where the tensor is sampled and zero otherwise. Then we have $\mathcal{T}_\Omega = \mathcal{P}_\Omega \odot \mathcal{T}$ where $\odot$ is the element-wise

Algorithm 6 Smooth QR factorization: SmoothQR$(\mathcal{Y}, \epsilon, \mu)$

Input: $\mathcal{Y} \in \mathbb{R}^{n\times r\times k}$, parameters $\mu, \epsilon > 0$

$\mathcal{Z} \leftarrow \text{QR}(\mathcal{Y}), \mathcal{H} \leftarrow 0, \sigma \leftarrow \varsigma||\mathcal{Y}||/n.$

While $\mu(\mathcal{Z}) > \mu$ and $\sigma \leq ||\mathcal{Y}||$

$\mathcal{Z} \leftarrow \text{GS}(\mathcal{Y} + \mathcal{H})$ where $\mathcal{H} \backsim \text{N}(0, \varsigma^2/n)$.

$\varsigma \leftarrow 2\varsigma$.

Output: $\mathcal{Z}$.

multiplication of same size arrays. For the (i, j)-th tube, $\mathcal{T}_{ij} = [\mathcal{P}_\Omega]_{ij} \odot \mathcal{T}_{ij}$. According to the *Convolution Theorem*, we can transform the least squares minimization in Alg. 5 to the following frequency domain version:

$$\widetilde{\mathcal{Y}} = \underset{\widetilde{\mathcal{Y}}\in\mathbb{R}^{r\times n\times k}}{\arg\min} ||\widetilde{\mathcal{T}}_\Omega - \widetilde{\mathcal{P}}_\Omega \cdot \otimes(\widetilde{\mathcal{X}} \ \cdot \S \ \widetilde{\mathcal{Y}})||_F^2, \tag{19}$$

where $\cdot\otimes$ denotes tube-wise circular convolution, and $\cdot\S$ denotes front-slice-wise matrix multiplication. The operator $\cdot\S$ is introduced to replace the block diagonal matrix in [24, 25] since we need to preserve the three-way data array structure. In the following, we will show how to compute (19) by providing methods to transform (19) into n separate standard least squares minimization problem. In other words, each lateral slice of $\widetilde{\mathcal{Y}}$ is a standard least squares minimization problem and will be estimated separately. Note that the above problem (19) can be split into n separate subproblems:

$$\widetilde{\mathcal{Y}}(:, j, :) = \underset{\widetilde{\mathcal{Y}}(:,j,:)\in\mathbb{R}^{r\times 1\times k}}{\arg\min} ||\widetilde{\mathcal{T}}_\Omega(:, j, :) - \widetilde{\mathcal{P}}_\Omega(:, j, :) \cdot \otimes(\widetilde{\mathcal{X}} \ \cdot \S \ \widetilde{\mathcal{Y}}(:, j, :))||_F^2, \tag{20}$$

where each subproblem corresponds to estimating a lateral slice $\widetilde{\mathcal{Y}}(:, j, :),\ j \in [n]$. Similarly, we can estimate $\widetilde{\mathcal{X}}$ in the following way:

$$\widetilde{\mathcal{X}} = \underset{\widetilde{\mathcal{X}}\in\mathbb{R}^{n\times r\times k}}{\arg\min} ||\widetilde{\mathcal{T}}_\Omega^T - \widetilde{\mathcal{P}}_\Omega^T \cdot \otimes(\widetilde{\mathcal{Y}}^T \ \cdot \S \ \widetilde{\mathcal{X}}^T)||_F^2, \tag{21}$$

where $\mathcal{X}^T$ denotes the tube-wise transpose (the transpose of a matrix of vectors). To solve this one performs the following steps.

1) A lateral slice, $\widetilde{\mathcal{T}}_\Omega(:, j, :)$ of size $n \times 1 \times k$, is squeezed into a vector b of size $nk \times 1$ in the following way:

$$b = [\text{squeeze}(\widetilde{\mathcal{T}}_\Omega(1, j, :)); \text{squeeze}(\widetilde{\mathcal{T}}_\Omega(2, j, :)); ...; \text{squeeze}(\widetilde{\mathcal{T}}_\Omega(n, j, :))], \tag{22}$$

where squeeze$(\widetilde{\mathcal{T}}_\Omega(i, j, :))$ squeezes the i-th tube of the j-th lateral slice of $\widetilde{\mathcal{T}}_\Omega$ into vector of size $k \times 1$.

Similarly $\widetilde{\mathcal{Y}}_\Omega(:,j,:)$ is transformed into a vector $\mathbf{x}$ of size $rk \times 1$:

$$\mathbf{x} = [\text{squeeze}(\widetilde{\mathcal{Y}}_\Omega(1,j,:)); \text{squeeze}(\widetilde{\mathcal{Y}}_\Omega(2,j,:)); ...; \text{squeeze}(\widetilde{\mathcal{Y}}_\Omega(n,j,:))]; . \tag{23}$$

2) $\widetilde{\mathcal{X}}$ is transformed into a block diagonal matrix A_1 of size $nk \times rk$ like so,

$$A_1 = \begin{bmatrix} \widetilde{\mathcal{X}}(:,:,1) & & & \\ & \widetilde{\mathcal{X}}(:,:,2) & & \\ & & \ldots & \\ & & & \widetilde{\mathcal{X}}(:,:,k) \end{bmatrix}. \tag{24}$$

3) The j-th lateral slice $\widetilde{\mathcal{P}}_\Omega(:,j,:)$ is transformed into a tensor A_2 of size $k \times k \times n$ first, and then into a matrix A_3 of size $nk \times nk$.

$$A_2(:,:,\ell) = \text{circ}(\widetilde{\mathcal{P}}_\Omega(\ell,j,:)),\ \ell \in [n], \tag{25}$$

where $\text{circ}(\mathbf{v})$ is a square circulant matrix formed from a vector $\mathbf{v}$, where the i-th column is a circularly shifted version of vector $\mathbf{v}$, whose elements are circularly shifted down by amount i.

$$A_3 = \begin{bmatrix} \text{diag}(A_2(1,1,:)) & \text{diag}(A_2(1,2,:)) & \ldots & \text{diag}(A_2(1,k,:)) \\ \text{diag}(A_2(2,1,:)) & \text{diag}(A_2(2,2,:)) & \ldots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ \text{diag}(A_2(k,1,:)) & \ldots & \ldots & \text{diag}(A_2(k,k,:)) \end{bmatrix}, \tag{26}$$

where the operator $\text{diag}(\cdot)$ transform a tube into a diagonal matrix by putting the elements in the diagonal.

Therefore, the estimation of the j-th lateral slice is transformed into the following standard least squares minimization problem:

$$\widehat{\mathbf{x}} = \underset{\mathbf{x} \in \mathbb{R}^{rk \times 1}}{\arg\min} ||b - A_3 A_1 \mathbf{x}||_F^2. \tag{27}$$

IV. Performance of the Proposed Algorithm

We first describe a counter example to show that the low-tubal-rank tensor completion problem is essentially different from the conventional matrix completion problem. Then, we present the analytical results for the performance guarantees.

A. Why Different from Matrix Case

One would naturally ask that is tensor completion in essence equal to matrix completion? It appears to be true but in fact wrong. Therefore, a tensor completion problem should be treated differently from a matrix completion problem. On the one hand, in Section II-B we introduce the operation $\text{circ}(\cdot)$ to establish a mapping between tensor product and matrix multiplication. However, such a mapping is injective, and we use it for easier understanding and do not mean equivalence. On the other hand, in Appendix A1 we express the least squares update step as $\mathcal{Y} = \mathcal{T} * \mathcal{X} + \mathcal{G}$ and then transform tensor operations to matrix operations on the corresponding circular matrices. Since the mapping is injective, this transformation is for analysis purpose, which holds only if the operations in the tensor forms hold. Still it is a necessary condition (not sufficient condition) resulted from the injective mapping between tensor product and its circular matrix multiplication.

The projection $\mathcal{P}_\Omega(\mathcal{T})$ can be viewed as an element-wise multiplication between two third-order tensors, i.e., $\mathcal{P}_\Omega(\mathcal{T}) = \mathcal{P}_\Omega \odot \mathcal{T}$, where $\mathcal{P}_\Omega$ is treated as a third-order tensor and defined as follows:

$$[\mathcal{P}_\Omega]_{ijl} = \begin{cases} 1, & \text{if } (i,j,\ell) \in \Omega, \\ 0, & \text{otherwise,} \end{cases}$$

We define a corresponding projection $\mathcal{P}_{\Omega'}(\cdot)$ for T^c as follows: $\Omega' = \text{circ}(\Omega), P_{\Omega'} = \text{circ}(\mathcal{P}_\Omega)$, i.e., $\mathcal{P}_{\Omega'}(T^c) = P_{\Omega'} \odot T^c$.

For the least squares update $\mathcal{Y} = \arg\min_{\mathcal{Y} \in \mathbb{R}^{[n]\times[r]\times[k]}} ||\mathcal{P}_\Omega(\mathcal{T} - \mathcal{X} * \mathcal{Y}^\dagger)||_F^2$, its circular form would be $Y = \arg\min_{Y \in \mathbb{R}^{[nk]\times[rk]}} \frac{1}{\sqrt{k}} ||\mathcal{P}_{\Omega'}(T^c - X^c Y^\dagger)||_F^2$. If Y is circular we can transform Y back to a tensor, then these two problems are equivalent. Therefore, the original question becomes the following one: will this circular least squares minimization output a circular matrix Y? In the following, we give a negative answer by presenting a counter example.

Assume that $\Omega = [n] \times [n] \times [k]$, then the circular least squares minimization is equivalent to the following optimization problem:

$$\min_Y \quad ||G||_F^2, \qquad s.t. \quad \mathcal{P}_{\Omega'}(T^c) = \mathcal{P}_{\Omega'}(X^c Y + G), \tag{28}$$

where G is a noise term. Without loss of generality, considering the following simple example with Ω being the whole set:

$$\min \quad ||G||_F^2, \qquad s.t. \quad \begin{bmatrix} T_1 & T_2 \\ T_2 & T_1 \end{bmatrix} = \begin{bmatrix} X_1 & X_2 \\ X_2 & X_1 \end{bmatrix} \begin{bmatrix} Y_{11} & Y_{21} \\ Y_{12} & Y_{22} \end{bmatrix} + \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix}, \tag{29}$$

where $\mathcal{T}, \mathcal{X} \in \mathbb{R}^{1\times1\times2}$ with $T^c, X^c \in \mathbb{R}^{2\times2}$, and $Y, G \in \mathbb{R}^{2\times2}$. The constraint in (29) can be transformed to the following four linear equations:

$$T_1 = X_1Y_{11} + X_2Y_{12} + G_{11} \tag{30}$$

$$T_1 = X_1Y_{22} + X_2Y_{21} + G_{22} \tag{31}$$

$$T_2 = X_1Y_{21} + X_2Y_{22} + G_{12} \tag{32}$$

$$T_2 = X_1Y_{12} + X_2Y_{11} + G_{21} \tag{33}$$

Considering the first two equations (30) and (31), if $G_{11} = G_{22}$ (and also $G_{12} = G_{21}$ in (32) and (33)), then the solution Y is a circular matrix. Given $G_{11} = -G_{22}$ that ensuring $G_{11}^2 = G_{22}^2$, then the solution Y is not circular matrix. Therefore, the problem in (28) does not guarantee to output a circular matrix.

B. Sampling Complexity and Computational Complexity

Low-tubal-rank tensor completion problem. We begin with our result on the *exact Low-tubal-rank tensor completion problem* where the goal is to recover an unknown tubal-rank-r tensor $\mathcal{T} \in \mathbb{R}^{n\times n\times k}$ from a subset Ω of its elements where each element is included independently with probability p. Here and in the following, we will assume that $\mathcal{T} = \mathcal{U} * \Theta * \mathcal{U}^\dagger$ ($\mathcal{U} \in \mathbb{R}^{n\times r\times k}, \Theta \in \mathbb{R}^{r\times r\times k}$) is a symmetric square tensor with eigentubes $\lambda_1, \lambda_2, ..., \lambda_r$[4] such that $||\lambda_1||_F \geq ||\lambda_2||_F \geq ... \geq ||\lambda_r||_F$. Accordingly, the block diagonal matrix $\overline{\mathcal{X}}$ has rank rk and we denote those rk singular values as $\overline{\sigma}_1, \overline{\sigma}_2, ..., \overline{\sigma}_{rk}$ such that $\overline{\sigma}_1 > \overline{\sigma}_2 > ... > \overline{\sigma}_{rk}$

Our Tubal-Alt-Min algorithm in Alg. 2 will output a pair of tensors $(\widehat{\mathcal{X}}, \widehat{\mathcal{Y}})$ such that $\widehat{\mathcal{X}} \in \mathbb{R}^{n\times r\times k}$ is an orthonormal tensor that approximates $\mathcal{U}$ in the strong sense that $||(\mathcal{I} - \mathcal{U} * \mathcal{U}^\dagger) * \widehat{\mathcal{X}}|| \leq \epsilon$, and the estimated tensor $\widehat{\mathcal{T}} = \widehat{\mathcal{X}} * \widehat{\mathcal{Y}}^\dagger$ is close to $\mathcal{T}$ in the Frobenius norm. We denote that coherence of $\mathcal{U}$ as $\mu(\mathcal{U}) \doteq \max_{i\in[n]} \frac{n}{r}||\mathcal{U}^\dagger * \dot{e}_i||_{2^*}^2$ where $\dot{e}_i$ is the i-th tensor-column basis.

Theorem 1. *Given a sample set Ω of size $O(pn^2k)$, each element randomly drawn with probability $p \geq O(\frac{r^2 \log n}{n})$ from an unknown $n \times n \times k$ (symmetric square) tensor $\mathcal{T} = \mathcal{U} * \Theta * \mathcal{U}^\dagger$ of tubal-rank r and incoherence μ. Then, Alg. 2 will output $(\widehat{\mathcal{X}}, \widehat{\mathcal{Y}})$ such that $||(\mathcal{I} - \mathcal{U} * \mathcal{U}^\dagger) * \widehat{\mathcal{X}}|| \leq \epsilon$ and $||\widehat{\mathcal{X}} * \widehat{\mathcal{Y}}^\dagger - \mathcal{T}||_F \leq \epsilon||\mathcal{T}||_F$ at a geometric rate.*

Please note that this theorem is reduced from Theorem 2 that deals with the noisy case.

[4]Note that the eigentubes $\lambda_1, \lambda_2, ..., \lambda_r$ correspond to $\Theta(1, 1, :), \Theta(2, 2, :), ..., \Theta(r, r, :)$ in Definition 7.

Remark 4. *Compared with the TNN-ADMM's [24] sampling complexity $O(nkr\log n)$, Tubal-Alt-Min requires an extra r factor. The better performance revealed in Section V, i.e., lower recovery error with less samples, indicates that 1) this extra r is not necessary and may be removed in the future, and 2) the information of the tubal-rank value plays a key role in Tubal-Alt-Min.*

Noisy Low-tubal-rank tensor completion problem. We assume that the unknown symmetric square tensor $\mathcal{T} \in \mathbb{R}^{n\times n\times k}$ is approximately low-tubal-rank in Frobenius norm. According to Lemma 1, the noisy tensor $\mathcal{T}$ has the form $\mathcal{T} = \mathcal{M} + \mathcal{N}$ where $\mathcal{M} = \mathcal{U} * \Theta * \mathcal{U}^\dagger$ ($\mathcal{U} \in \mathbb{R}^{n\times r\times k}, \Theta \in \mathbb{R}^{r\times r\times k}$) is a symmetric square tensor with eigentubes $\lambda_1, \lambda_2, ..., \lambda_r$ as before, and $\mathcal{N} = (\mathcal{I} - \mathcal{U} * \mathcal{U}^\dagger) * \widehat{\mathcal{M}}$ is the part of $\mathcal{T}$ not captured by the dominant eigentubes. Here, we assume that $\mathcal{N}$ can be an arbitrary deterministic tensor that satisfies the following constraints:

$$\max_{i\in[n]} ||\mathcal{N}^\dagger * \dot{e}_i||_F^2 \leq \frac{\mu_N}{n}\overline{\sigma}_{rk}^2, \quad \text{and} \quad \max_{i,j,\kappa} |\mathcal{N}_{ij\kappa}| \leq \frac{\mu_N}{n}||\mathcal{M}||_F, \tag{34}$$

where $\dot{e}_i$ denotes the i-th tensor-column basis so that $||\mathcal{N}^\dagger * \dot{e}_i||_F^2$ is the Frobenius norm of the i-th horizontal slice of $\mathcal{N}$. These constraint state that no element and no horizontal slice of $\mathcal{N}$ should be too large compared to the Frobenius norm of $\mathcal{N}$. One can think of the parameter μ_N as an analog to the coherence parameter $\mu(\mathcal{U})$ that we saw earlier.

Let $\mu^* = \max\{\mu(\mathcal{U}, \mu_N, \log n)\}$, then we have the following theorem.

Theorem 2. *Given a sample set Ω of size $O(pn^2k)$, each element randomly drawn with probability $p \geq O(\frac{r^2\log n}{n})$ from an unknown $n \times n \times k$ (symmetric square) tensor $\mathcal{T} = \mathcal{M} + \mathcal{N}$ where $\mathcal{M} = \mathcal{U} * \Theta * \mathcal{U}^\dagger$ of tubal-rank r and $\mathcal{N}$ satisfies condition (34). Then, Alg. 2 will output $(\widehat{\mathcal{X}}, \widehat{\mathcal{Y}})$ such that $||(\mathcal{I} - \mathcal{U} * \mathcal{U}^\dagger) * \widehat{\mathcal{X}}|| \leq \epsilon$ and $||\widehat{\mathcal{X}} * \widehat{\mathcal{Y}}^\dagger - \mathcal{T}||_F \leq \epsilon||\mathcal{T}||_F$ at a geometric rate.*

Proof: This theorem is a direct result of Theorem 7, and the detailed proof is presented in Appendix E. Here, we briefly describe the high-level proof structure for understanding.

Initialization is analyzed in Appendix C. We prove in Lemma 16 that initializing $\mathcal{X}_0$ as the top-r left orthogonal eigenslices of $\frac{1}{p}\mathcal{P}_\Omega(\mathcal{T})$ will result in (clear) bounded distance to the optimum tensor $\mathcal{T}$, i.e., $||\mathcal{P}_\Omega(\mathcal{T}) - \mathcal{T}|| \leq \sqrt{k}||\mathcal{T}||_F^2$ with high probability. This is necessary because the key step in Alg. 2, iterating the tensor least squares minimization, does not provide information about how well it converges from a random initial tensor. Note that this initialization is primary for analysis purpose and is not necessary in implementation, since random initialization is empirically successful[5] in implementation.

[5]Both in our testing and in the matrix scenarios reported by many researchers.

With the initialization guarantee, our analysis is stated as follows:

- Express the least squares update step as an update step of the noisy tensor-column subspace iteration which has the form $\mathcal{Y} = \mathcal{T} * \mathcal{X} + \mathcal{G}$.
- Analyze the local convergence of Alg. 5 by proving that the noisy tensor-column subspace iteration will converge at a geometric rate in Appendix B.
- Prove that the Frobenius norm of the error term $\mathcal{G}$, i.e., $||\mathcal{G}||_F$, is bounded in Lemma 9.

Alternating tensor least squares minimization is analyzed in Appendix A1. The key step of Alg. 2 is a tensor least squares minimization $\mathcal{Y}_\ell = \arg\min_{\mathcal{Y}\in\mathbb{R}^{n\times r\times k}} ||\mathcal{P}_\Omega(\mathcal{T} - \mathcal{X}_{\ell-1} * \mathcal{Y}^\dagger)||_F^2$. We will show that the solution to this has the form of $\mathcal{Y}_\ell = \mathcal{T} * \mathcal{X}_{\ell-1} + \mathcal{G}_\ell$. The error term $||\mathcal{G}_\ell||$ depends on the quantity $||\mathcal{V}^\dagger * \mathcal{X}_{\ell-1}||$ which coincides with the sine of the largest principal angle between $\mathcal{U}$ and $\mathcal{X}_{\ell-1}$. As the algorithm begins to converge, the norm of the error term starts to diminish. ■

Computation Complexity.

Theorem 3. *The Tubal-Alt-Min algorithm in Alg. 2 has computational complexity* $O(n^2rk^2\log n)$ *if* $k \geq \sqrt{\frac{n}{Lr\log n}}$, *and* $O(n^3)$ *otherwise.*

Proof: We characterize the computational complexity by counting the multiplications in Alg. 2. The basic building block is the complexity of the tensor least squares minimization, corresponding to one step in Section III-D. This complexity is dominated by (27), which has complexity $O(nrk^2)$. More specifically, calculating A_3A_1 has complexity $O(rk)$ exploiting the sparseness of both A_3 and A_1, and the least squares minimization (27) requires $O(nrk^2)$ multiplications which also takes into consideration of the diagonal structure of A_3A_1. Each lateral slice are treated separately, thus a tensor least squares minimization has complexity $O(n^2rk^2)$.

The initialization process includes two sub-processes: the t-SVD process and the DFT process. The t-SVD process in Alg. 1 has complexity $O(n^3)$, while the DFT process has complexity $O(n^2k\log k)$. The initialization process has total complexity $O(n^3 + n^2k\log k) = O(n^3)$ as we assume that $k\log k < n$. Since for large k being comparable to n, the implementation in (27) will incur the out-of-memory error.

Note that each median operation calls $t = O(\log n)$ tensor least squares minimization, and there is $L = \theta(\gamma_{rk}^{-1}\log(n/\epsilon))$ iterations. Therefore, the total complexity of Alg. 2 is $O(Ln^2rk^2\log n + n^3)$. Therefore, the computation complexity is $O(n^2rk^2\log^2 n)$, if $k \geq \sqrt{\frac{n}{Lr\log n}}$. If $k < \sqrt{\frac{n}{Lr\log n}}$, then the total complexity is $O(n^3)$, dominated by the SVD process. Please note that Tubal-Alt-Min has geometric convergence rate (also verified by our experiments in Section V), and require a constant number of iterations, being about $L = O(\log n) \approx 10$. ■

V. EVALUATION

We evaluate our alternating minimization algorithm (Tubal-Alt-Min) on both synthetic and real video data. The synthetic data, generated according to our low-tubal-rank tensor model, serves as well-controlled inputs for testing and understanding Tubal-Alt-Min's performance over the convex algorithm TNN-ADMM. The real video data tests the applicability of our low-tubal-rank tensor model, comparing with other tensor models.

A. Experiment Setup

We use Matlab installed on a server with Linux operating system. The parameters of the server is: Intel® Xeon® Processor E5-2650 v3, 2.3 GHz clock speed, 2 CPU each having 10 physical cores, virtually maximum 40 threads, 25 MB cache, and 64 GB memory.

For synthetic data, we compare our Alter-Min algorithm with tensor-nuclear norm (TNN-ADMM) [25], since both are designed for low-tubal-rank tensors, to show the advantages of our non-convex approach over its convex counterpart. We conduct experiments to recover a third-order tensor of different size $n_1 \times n_2 \times k$ and tubal-ranks r, from $|\Omega|$ observed elements. Three metrics are adopted for comparison, e.g., the recovery error, the running time, and the convergence speed.

- For recovery error, we adopt the relative square error metric, defined as RSE$= ||\hat{\mathcal{T}} - \mathcal{T}||_F / ||\mathcal{T}||_F$.
- For running time, varying the tensor size and fixing other parameters, we measure cpu time in seconds.
- For the convergence speed, we measure the decreasing rate of the RSE across the iterations by linearly fitting the measured RSEs. We include those plots due to three reasons: 1) both algorithms are iterative, 2) our theoretical analysis predicts geometric convergence rate, and 3) the decreasing speed of the RSE provides explanations for the observed performance of the recovery error and the running time.

For real dataset, we choose a basketball video of size $144 \times 256 \times 40$ (source: YouTube, as used in [25]), with a non-stationary panning camera moving from left to right horizontally following the running players. Besides the low-tubal-rank tensor model, there are two widely used tensor models: low CP-rank tensor [20] and low Tuker-rank tensor [14].

The compared algorithms are briefly described as follows:

- TNN-ADMM [25] introduced a convex norm, tensor-nuclear norm [24], to approximate the tubal-rank and proposed an Alternating Direction Method of Multipliers (ADMM) algorithm. It is a convex

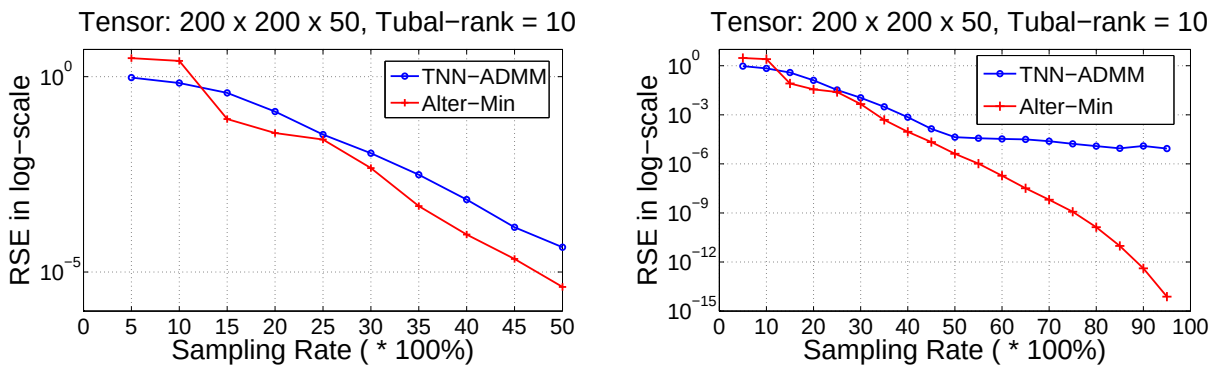

Fig. 3. Recovery error RSE in log scale for different sampling rates.

approach and shares the alternating and iterative property. We include TNN-ADMM [25] to show that the predicted advantages on the synthetic data in Section V-B also hold for real datasets.

- The alternating minimization algorithm (CP-Altmin) [37] under CANDECOMP/PARAFAC decomposition. The CP decomposition models a tensor as the outer product of r (the CP-rank) vector components. The CP-Altmin algorithm [37] alternates among the leat squares minimization subproblems for each component.
- The gradient-type algorithm Tuker-Gradient [38] under Tuker decomposition. Tuker decomposition matricizes a tensor from n modes, while the Tuker-Gradient algorithm [38] iteratively estimate each matricized component based on the conjugate gradient information. Essentially, Tuker-Gradient does not fall into the alternating minimization approach. However, we feel it necessary to compare with since both Tubal-Alt-Min and Tuker-Gradient iteratively estimate each factor and are nonconvex.

B. Synthetic Data

For recovery error, our input is a low-tubal-rank tensor of size $200 \times 200 \times 50$ and tubal-rank 10. We first generate two Guassian random tensors of sizes $200 \times 10 \times 50$ and $10 \times 200 \times 50$ and then perform tensor product in Definition 1 to get the input tensor of size $200 \times 200 \times 50$. We set the maximum iteration number to be 10 for Alter-Min and 500 for TNN-ADMM, respectively. Varying the sampling rate as $5\%,\ 10\%, ..., 95\%$ by uniform randomly select entries, we test each sampling rate 5 times and then plot the average results.

Fig. 3 shows the recovery error performance (RSE in $\log$ scale) of Alter-Min and TNN-ADMM for varying sampling rates. For a clear comparison, we draw two plots for sampling rate $\leq 50\%$ and

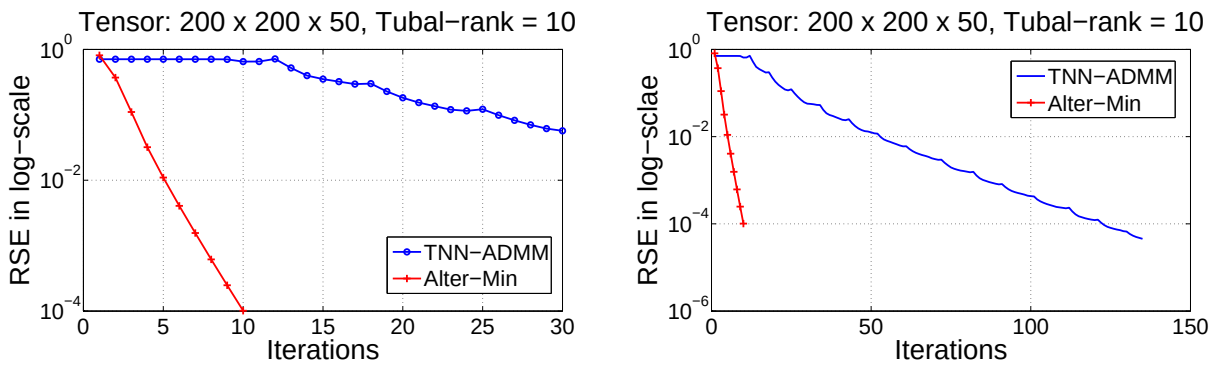

Fig. 4. Convergence speed for the two algorithms on synthetic data.

sampling rates $5\% \sim 95\%$, respectively. Alter-Min achieves lower orders of error magnitude for sampling rates higher than 50%, while the RSE does not decrease much for TNN-ADMM. For sampling rates $15\% \sim 45\%$, the RSE of Tubal-Alt-Min is approximately half of or one order lower than TNN-ADMM except an abnormal case at sampling rate 25%. However, for sampling rates $5\% \sim 10\%$, Alter-Min behaves badly since with insufficient elements, Alter-Min switches between the factors tensors that are both not well determined. The possible reason TNN-ADMM cannot achieve relatively lower orders RSE as Alter-Min does is that TNN-ADMM lacks the exact tubal-rank value while it is an input in Alter-Min.

For convergence speed, the input is a low-tubal-rank tensor of size $200 \times 200 \times 50$ and tubal-rank 10. We set the maximum iteration number to be 10 for Tubal-Alt-Min and 500 for TNN-ADMM. We fix the sampling rate to be 50% and record the RSE in each iteration. TNN-ADMM terminates at the 134-th iteration because the algorithm detected that the decrease of RSE is lower than a preset threshold.

Fig. 4 shows the decreasing RSE across the iterations for Tubal-Alt-Min and TNN-ADMM. For a clear comparison, we draw two plots at iteration number 30 and 140 for TNN-ADMM, respectively. Clearly, Tubal-Alt-Min decrease much faster than TNN-ADMM. Unexpectedly, TNN-ADMM behaves very badly during the first 11 iteration.

Fig. 5 shows the fitting results for the convergence rates for Tubal-Alt-Min and TNN-ADMM. We use linear functions to fit our observed RSE (in $\log$ form) across the iterations. For Tubal-Alt-Min, the fitted function is $y = -0.4423x + 0.339$ and the estimated convergence rate is $10^{-0.4423} = 0.3612$. For TNN-ADMM, the fitted function is $y = -0.0322x - 0.1618$ and the estimated convergence rate is $10^{-0.0322} = 0.9285$. Therefore, Tubal-Alt-Min convergence faster than TNN-ADMM.

For running time, we measure the CPU time in seconds. We vary the tensor size from $25 \times 25 \times 20$

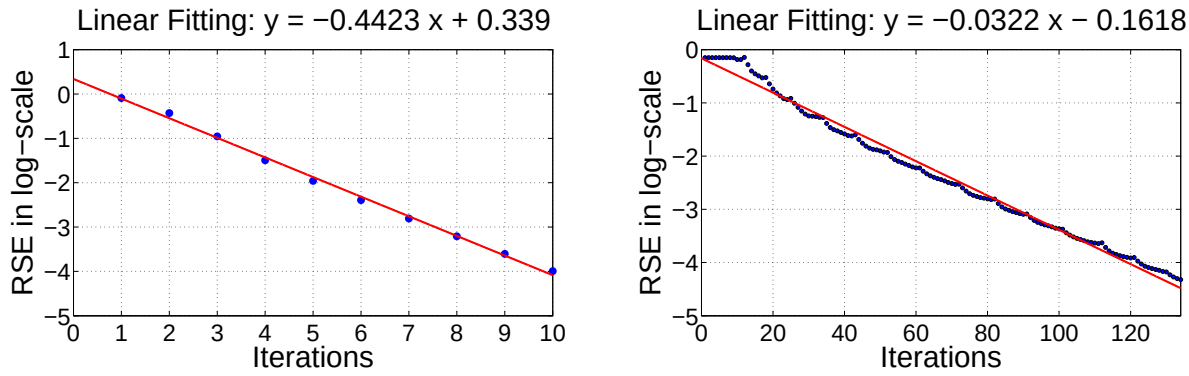

Fig. 5. Using linear fitting to estimate the convergence rate, Left: Tubal-Tubal-Alt-Min, Right: TNN-ADMM.

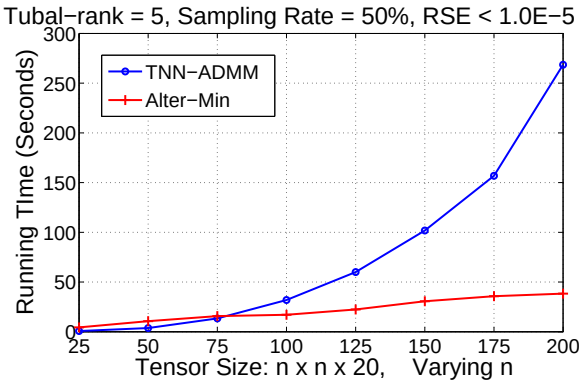

Fig. 6. Comparison of running time for varying tensor size.

to $200 \times 200 \times 20$. Note that the tubal-rank tensor model views a tensor as a matrix of vectors, the third-demention k acts as a linear scaling factor, therefore, to test large tensor cases we set $k = 20$ smaller than n_1 and n_2 to avoid the ERR: out of memory. The tubal-rank is set to be 5 for all cases, the sampling rate is 50%, while the target RSE less than 10^{-5}.

Fig. 6 shows the running time comparison for Tubal-Alt-Min and TNN-ADMM. For tensors larger than $75 \times 75 \times 20$, our Tubal-Alt-Min algorithm beats TNN-ADMM. The acceleration ratio is 2 times for tensors of size $125 \times 125 \times 20$ and 5 times for tensors of size $200 \times 200 \times 20$. However, since our implementation of Tubal-Alt-Min in Section III-D introduces intermediate matrices of size $nk \times nk \times n$, we encounter the ERR: out of memory and thus we are not able to test larger tensor sizes.

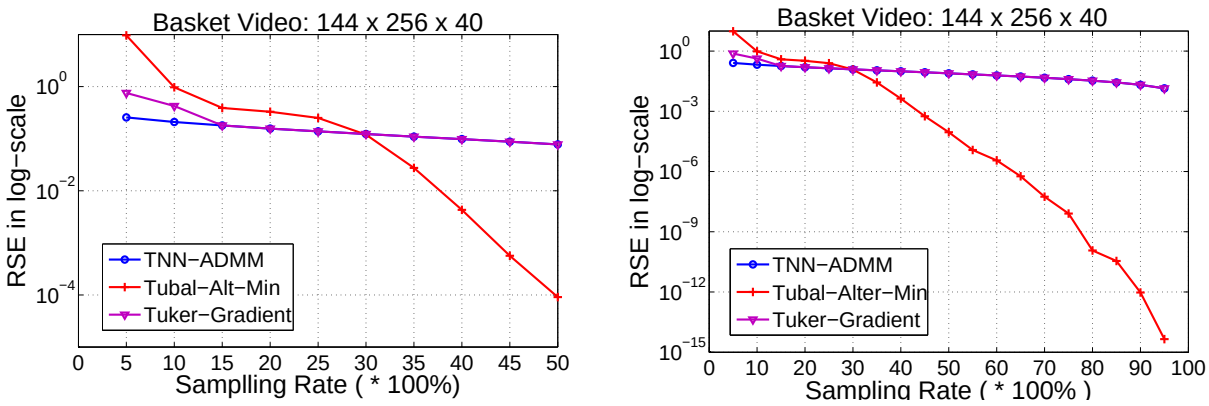

Fig. 7. Recovery error RSE in log scale for different sampling rates.

C. Real Data

For the basket video, we only compare the recovery error. Since we are interested in fast algorithms and we think it is not fair to compare the running time and convergence speed of our Tubal-Alt-Min algorithm with the Tuker-Gradient [38][6] while it has much higher recovery error.

Fig. 7 shows the recovery error performance (RSE in log-scale) of Alter-Min, TNN-ADMM and Tuker-Gradient for varying sampling rates. For a clear comparison, we draw two plots for sampling rate $\leq 50\%$ and sampling rates $5\% \sim 95\%$, respectively. Alter-Min achieves relative lower orders of error magnitude for sampling rates higher than 35%, while the RSE does not decrease much for TNN-ADMM, with a much clearer superior performance than that in Fig. 3. Here, Alter-Min behaves worse than that in Fig. 3, for sampling rates $5\% \sim 30\%$ and much better for sampling rates $35\% \sim 95\%$. The reasons are: 1) the basket video data has larger tubal-rank (approximately 30) than the synthetic input tensor with tubal-rank 10, and 2) Tubal-Alt-Min is more capable of dealing with noise in the real-world video data. Note that in log-scale, the performance of TNN-ADMM and Tuker-Gradient are in the same order and thus indistinguishable.

VI. Conclusions

Alternating minimization provides an empirically appealing and popular approach to solve low-rank matrix and tensor completion problems. We provide the first theoretical guarantees on the global optimality

[6]The Tuker-Gradient [38] implementation tested tensor of size $10,000 \times 10,000 \times 10,000$, which is not possible in our implementation due to the ERR: out of memory for our implementation in Section III-D.

for the low-tubal-rank tensor completion problem. Based on extensive evaluations, we show that the proposed algorithm is fast and accurate.

This paper proposes a fast iterative algorithm, called *Tubal-Alt-Min*, that is based on the alternating minimization approach for low-tubal-rank tensor completion. The unknown low-tubal-rank tensor is parameterized as the product of two much smaller tensors with the low-tubal-rank property being automatically incorporated, and Tubal-Alt-Min alternates between estimating those two tensors using tensor least squares minimization. We derive theoretical performance guarantees on the success of the proposed algorithm under the tensor incoherency conditions as derived in [24]. The performance of the proposed algorithm is demonstrated for both synthetic and real-world data, compared with both the corresponding convex approach and other tensor factorization approaches.

APPENDIX

A. Tensor Least Squares Minimization

Alg. 5 describes a tensor least squares minimization update step, specialized to the case of a symmetric square tensor. Our goal in this section is to express the tensor least squares minimization update step as $\mathcal{Y} = \mathcal{T} * \mathcal{X} + \mathcal{G}$, then we will be able to apply our convergence analysis of the noisy tensor-column subspace iteration in Appendix B. This syntactic transformation is given in Appendix A1 that is followed by a bound on the norm of the noise term $\mathcal{G}$ in Appendix A2. With this, we prove in Appendix A3 that the element-wise median process (in Alg. 4) on $t = O(\log n)$ tensor least squares minimizations will result in a much tighter bound of the noise term $\underline{\mathcal{G}}$ that is the average of those $t = O(\log n)$ copies of $\mathcal{G}$.

1) From Alternating Least Squares to Noisy Tensor-Column Subspace Iteration: We first show that the tensor completion can be analyzed in its circular form in Lemma 5, then give an optimality condition in the circular form that the optimizer $\mathcal{Y}$ satisfies a set of linear equations in Lemma 6. With these constraints, we express the tensor least squares minimization update step as $\mathcal{Y} = \mathcal{T} * \mathcal{X} + \mathcal{G}$ in Lemma 7.

Lemma 5. *The function $f(\mathcal{Y}) = ||\mathcal{P}_\Omega(\mathcal{T} - \mathcal{X} * \mathcal{Y}^\dagger)||_F^2$ has an injective mapping to the circular form $f(Y^c) = \frac{1}{\sqrt{k}}||\mathcal{P}_{\Omega'}(T^c - X^c Y^{c\dagger})||_F^2$. (Note that in Section IV-A we show that those two objective functions are different.)*

Proof: For the least squares update $\mathcal{Y} = \arg\min_{\mathcal{Y} \in \mathbb{R}^{n \times r \times k}} ||\mathcal{P}_\Omega(\mathcal{T} - \mathcal{X} * \mathcal{Y}^\dagger)||_F^2$, it is equivalent to the following optimization problem:

$$\min_{\mathcal{Y}} \quad ||\mathcal{G}||_F^2, \qquad s.t. \;\; \mathcal{P}_\Omega(\mathcal{T}) = \mathcal{P}_\Omega(\mathcal{X} * \mathcal{Y}^\dagger + \mathcal{G}), \tag{35}$$

where $\mathcal{G}$ is a noise term. The circular form of (35) is:

$$\min_{Y} \quad \frac{1}{k}||G^c||_F^2, \qquad s.t. \;\; P_{\Omega'}T^c = P_{\Omega'}X^cY^\dagger + P_{\Omega'}G^c. \tag{36}$$

Since the addition/subtraction and inverse operations are closed in the circulant algebra [32], then $Y^\dagger = (P^{\Omega'}X^c)^{-1}(P_{\Omega'}T^c - P_{\Omega'}G^c)$ is also circular.

We can see that $f(\mathcal{Y})$ implies $f(Y^c)$ while the opposite direction does not hold as shown in Section IV-A. Therefore, this mapping is injective. ■

Lemma 6. *(Optimality Condition). Let $P_i : \mathbb{R}^{nk} \to \mathbb{R}^{nk}$ be the linear projection onto the coordinates in $\Omega'_i = \{j : (i,j) \in \Omega'\}$ scaled by $p^{-1} = (nk)^2/(\mathbb{E}|\Omega'|)$, i.e., $P_i = p^{-1}\sum_{j\in\Omega'_i} e_j e_j^\dagger$ where e_i, e_j are the standard vector bases and $e_j^\dagger$ is the corresponding row basis. Further, define the matrix $B_i \in \mathbb{R}^{rk\times rk}$ as $B_i = X^{c\dagger}P_iX^c$ (note that B_i is invertible as shown in [29]). Then, for every $i \in [nk]$, the i-th row of Y^c satisfies $e_i^\dagger Y^c = e_i^\dagger T^c P_i X^c B_i^{-1}$.*

Proof: By Lemma 5, we consider the circular objective function $f(Y^c) = \frac{1}{\sqrt{k}}||\mathcal{P}^{\Omega'}(T^c - X^cY^{c\dagger})||_F^2$. For every $i \in [nk]$, $j \in [rk]$, we have $\frac{\partial f}{\partial Y^c_{ij}} = -\frac{2}{\sqrt{k}}\sum_{s\in\Omega'_i} T^c_{is}X^c_{sj} + \frac{2}{\sqrt{k}}\sum_{t=1}^{rk} Y^c_{it}\sum_{s\in\Omega'_i} X^c_{sj}X^c_{st}$. Therefore, we know that the optimal $\mathcal{Y}$ must satisfy $e_i^\dagger T^c P_i X^c = e_i^\dagger Y^c X^{c\dagger} P_i X^c = e_i^\dagger Y^c B_i$, hence, $e_i^\dagger Y^c = e_i^\dagger T^c P_i X^c B_i^{-1}$. ■

Lemma 7. *Let $E^c = (I^c - X^cX^{c\dagger})U^c$, and assume that $\mathcal{T}$ is a noisy tensor (approximately r-tubal-rank) that is the superposition of an exact r-tubal-rank tensor $\mathcal{M}$ and a noisy tensor $\mathcal{N}$, i.e., $\mathcal{T} = \mathcal{M} + \mathcal{N}$. We express the least squares update as $\mathcal{Y} = \mathcal{T} * \mathcal{X} + \mathcal{G}$ where $\mathcal{G} = \mathcal{G}_\mathcal{M} + \mathcal{G}_\mathcal{N}$ and their circular matrices $G^c_\mathcal{M}$ and $G^c_\mathcal{N}$ satisfy that each row $i \in [nk]$, we have the following expressions:*

$$\begin{aligned} e_i^\dagger G^c_\mathcal{M} &= e_i^\dagger U^c \Lambda^c_{U^c} E^{c\dagger} P_i X^c B_i^{-1} \\ e_i^\dagger G^c_\mathcal{N} &= e_i^\dagger (N^c P_i X^c B_i^{-1} - N^c X^c). \end{aligned} \tag{37}$$

Proof: Since $\mathcal{T} = \mathcal{M} + \mathcal{N}$, we have $\mathcal{M} = \mathcal{U} * \Theta * \mathcal{U}^\dagger$ and $\mathcal{N} = (\mathcal{I} - \mathcal{U} * \mathcal{U}^\dagger) * \mathcal{T}$. By Lemma 6, $e_i^\dagger Y^c = e_i^\dagger T^c P_i X^c B_i^{-1} = e_i^\dagger M^c P_i X^c B_i^{-1} + e_i^\dagger N^c P_i X^c B_i^{-1}$ since $Y^c = M^c + N^c$. Let $C_i = U^{c\dagger}P_iX^c$ and $D = U^{c\dagger}X^c$. We have:

$$\begin{aligned} e_i^\dagger M^c P_i X^c B_i^{-1} = -e_i^\dagger U^c \Lambda_{U^c} C_i B_i^{-1} &= e_i^\dagger (U^c \Lambda_{U^c} D - U^c \Lambda_{U^c}(DB_i - C_i)B_i^{-1}) \\ &= e_i^\dagger M^c X^c - e_i^\dagger U^c \Lambda_{U^c}(DB_i - C_i)B_i^{-1}, \\ C_i = U^{c\dagger}P_iX^c = (X^cX^{c\dagger}U^c + E^c)P_iX^c &= (U^{c\dagger}X^c)X^{c\dagger}P_iX^c + E^{c\dagger}P_iX^c \\ &= DB_i + E^{c\dagger}P_iX^c. \end{aligned} \tag{38}$$

Then, we have $e_i^\dagger M^c P_i X^c B_i^{-1} = e_i^\dagger M^c X^c - e_i^\dagger U^c \Lambda_{U^c} E^{c\dagger} P_i X^c B_i^{-1}$. From (37) we know that $e_i^\dagger N^c P_i X^c B_i^{-1} = e_i^\dagger N^c X^c + e_i^\dagger G_{\mathcal{N}}^c$. Putting all together, we have that $Y^c = M^c X^c + G_{\mathcal{M}}^c + N^c X^c + G_{\mathcal{N}}^c = T^c X^c + G_{\mathcal{M}}^c + G_{\mathcal{N}}^c$, therefore, transforming it to the tensor form we have $\mathcal{Y} = \mathcal{T} * \mathcal{X} + \mathcal{G}$. ■

2) Bound the Noisy Term $\mathcal{G}$: We bound the spectral norm of each horizontal slice of $\mathcal{G}$. An intriguing fact is that the matrix E^c appearing in the expression for the error terms (37) satisfies $||E^c|| = ||V^{c\dagger} X^c||$, i.e., $||\mathcal{E}|| = ||\mathcal{V}^\dagger * \mathcal{X}||$ where $\mathcal{V}$ is defined in the t-SVD (Definition 7). This allows us to obtain a bound through the quantity $||V^{c\dagger} X^c||$ that equals to $||\mathcal{V}^\dagger * \mathcal{X}||$ according to Lemma 2.

Lemma 8. *Let $\delta \in (0, 1)$. Assume that each entry is included in Ω independently with probability*

$$p_+ \geq \frac{r\mu(\mathcal{X}) \log nk}{\delta^2 n}, \tag{39}$$

then, for $\forall i \in [n]$, $\mathbb{P}\left\{||\dot{e}_i^\dagger * \mathcal{G}|| > \delta\left(||\dot{e}_i^\dagger * \mathcal{M}|| \cdot ||\mathcal{V}^\dagger * \mathcal{X}|| + ||\dot{e}_i^\dagger * \mathcal{N}||\right)\right\} \leq \frac{1}{5}$.

Proof: The set Ω' is exactly k replicas of Ω, and the probability p_+ of Ω corresponds to one replica of those k replicas. According to Lemma 4.3 [29], if the probability $p_+ \geq \frac{r\mu(X^c) \log nk}{\delta^2 nk}$ for each replica in Ω', then we have $\mathbb{P}\{||e_i^\dagger G^c|| > \delta(||e_i^\dagger M^c|| \cdot ||V^{c\dagger} X^c|| + ||e_i^\dagger N^c||)\} \leq \frac{1}{5}$, corresponding to $\mathbb{P}\left\{||\dot{e}_i^\dagger * \mathcal{G}|| > \delta\left(||\dot{e}_i^\dagger * \mathcal{M}|| \cdot ||\mathcal{V}^\dagger * \mathcal{X}|| + ||\dot{e}_i^\dagger * \mathcal{N}||\right)\right\} \leq \frac{1}{5}$ (Lemma 2). Note that $\mu(X^c) = k\mu(\mathcal{X})$, we have $p_+ \geq \frac{r\mu(\mathcal{X}) \log nk}{\delta^2 n}$. ■

3) Median Tensor Least Squares Minimization: Here, we further analyze the element-wise median process in Alg. 4. Given the previous error bound of $\mathcal{G}$ in Lemma 8, we can further derive a stronger concentration bound by taking the element-wise median of multiple independent samples of the error term.

Lemma 9. *Let Ω be a sample set in which each element is included independently with probability p_+. Let $\mathcal{G}_1,\ \mathcal{G}_2, ..., \mathcal{G}_t$ be i.i.d. copies of $\mathcal{G}$, and $\underline{\mathcal{G}} = median(\mathcal{G}_1,\ \mathcal{G}_2, ..., \mathcal{G}_t)$ be the element-wise median, and assume p_+ satisfy (39). Then, for every $i \in [n]$,*

$$\mathbb{P}\{||\dot{e}_i^\dagger * \underline{\mathcal{G}}|| > \delta\left(||\dot{e}_i^\dagger * \mathcal{M}|| \cdot ||\mathcal{V}^\dagger * \mathcal{X}|| + ||\dot{e}_i^\dagger * \mathcal{N}||\right)\} \leq \exp(-\Omega(t)), \tag{40}$$

where $\Omega(t)$ denotes some polynomial of t.

Proof: For each $i \in [n]$, let $g_1, g_2, ..., g_t \in \mathbb{R}^{1\times r\times k}$ denote the i-th horizontal slice of $\mathcal{G}_1, \mathcal{G}_2, ..., \mathcal{G}_t$. Let $S = \{j \in [t] : ||g_j|| \leq B\}$ where $B = \frac{\delta}{4}\left(||\dot{e}_i^\dagger * \mathcal{M}|| \cdot ||\mathcal{V}^\dagger * \mathcal{X}|| + ||\dot{e}_i^\dagger * \mathcal{N}||\right)$. Applying Lemma 8 with error parameter $\frac{\delta}{4}$, we have $\mathbb{E}|S| \geq 4t/5$ with g_j being drawn independently. Then we apply a Chernoff bound to argue that $\mathbb{P}(|S| > 2t/3) \geq 1 - \exp(-\Omega(t))$.

Fixing a coordinate $s \in [r]$. By the median property we have $|\{j||g_j(1,s,:)||_F^2 \geq ||\underline{g}_j||_F^2\}| \geq t/2$. Since $|S| > 2t/3$, we know that at least $t/3$ horizontal slices with $j \in S$ have $||g_j(1,s,:)||_F^2 \geq ||\underline{g}_j||_F^2$. Therefore, the average value of $||g_j(1,s,:)||_F^2$ over all $j \in [S]$ much be at least $\frac{t||\underline{g}_j||_F^2}{3|S|} \geq ||\underline{g}_j||_F^2/3$. This means that the average of $||g_j||_F^2$ over all $j \in [S]$ much be at least $||\underline{g}||_F^2/3$. On the other hand, we also know that the average squared Frobenius norm in S is at most B^2 by the definition of S. Then, the lemma is proved. ■

We then provide a strong concentration bound for the median of multiple independent solutions to the tensor least squares minimization step.

Lemma 10. *Let Ω be a sample set in which each element is included independently with probability $p_+ \geq \frac{r\mu(\mathcal{X})\log nk}{\delta^2 n}$. Let $\mathcal{Y} \leftarrow$ MedianLS-Y($\mathcal{P}^{\Omega_\ell}(\mathcal{T}), \Omega, \mathcal{X}, r$). Then, we have with probability $1-1/n^3$ that $\underline{\mathcal{Y}} = \mathcal{T} * \mathcal{X} + \underline{\mathcal{G}}$ with $\underline{\mathcal{G}}$ satisfying: $||\dot{e}_i^\dagger * \underline{\mathcal{G}}|| \leq \delta \left(||\dot{e}_i^\dagger * \mathcal{M}|| \cdot ||\mathcal{V}^\dagger * \mathcal{X}|| + ||\dot{e}_i^\dagger * \mathcal{N}|| \right)$.*

Proof: Using the Split(Ω, t) process, we know that the sample sets $\Omega_1, ...\Omega_j, ..., \Omega_t$ are independent and each set Ω_j includes each element with probability at least p_+/t. The output satisfies $\underline{\mathcal{Y}} =$ median$(\mathcal{Y}_1, ..., \mathcal{Y}_j, ..., \mathcal{Y}_t)$, where $\mathcal{Y}_j = \mathcal{T} * \mathcal{X} + \mathcal{G}_j$. Then median$(\mathcal{Y}_1, ..., \mathcal{Y}_j, ..., \mathcal{Y}_t) = \mathcal{T} * \mathcal{X} + \underline{\mathcal{G}}$.

Therefore, apply Lemma 9 combining the fact $t = O(\log n)$, we take a union bound over all n horizontal slices of $\mathcal{G}$ to conclude this lemma. ■

B. Convergence of Noisy Tensor-Column Subspace Iteration

Algorithm 7 Noisy Tensor-Column Subspace Iteration

Input: Tensor $\mathcal{T} \in \mathbb{R}^{n\times n\times k}$, number of iterations L, target dimension r

Let $\mathcal{X}_0 \in \mathbb{R}^{n\times r\times k}$ be an orthonormal tensor.

For $\ell = 1$ to L

Let $\mathcal{G}_\ell \in \mathbb{R}^{n\times r\times k}$ be an arbitrary perturbation.

$\mathcal{Z}_\ell \leftarrow \mathcal{T} * \mathcal{X}_{\ell-1} + \mathcal{G}_\ell$.

$\mathcal{X}_\ell \leftarrow GS(\mathcal{Z}_\ell)$.

Output: Tensor $\mathcal{X}_\ell \in \mathbb{R}^{n\times r\times k}$.

Alg. 7 describes our noisy tensor-column subspace iteration, where $GS(\mathcal{Y}_l)$ denotes the Gram-Schimidt process which orthonormalizes the lateral slices of the tensor $\mathcal{Z}_\ell$. The detailed steps of the Gram-Schimidt process for third-order tubal-rank tensor is given in [22]). Note that Alg. 7 is different from the recently

proposed power method [32]: 1) we simultaneously compute multiple top-k eigenslices while the power method considered only the top-1 eigenslice, 2) in each iteration ℓ, the computation in Alg. 7 is perturbed by a tensor $\mathcal{G}_\ell$ which can be adversarially and adaptively chosen, and 3) the Gram-Schimidt process for third-order tubal-rank tensor is introduced to manipulate $\mathcal{G}_\ell$.

For the matrix case, an important observation of [27–29] is that the least squares minimization can be analyzed as a noisy update step of the well known *subspace iteration* (or power method). Therefore, the convergence of the alternating minimization iteration is equivalent to the convergence of the noisy subspace iteration. The corresponding convergence analysis exploits the tangent function of the largest principle angle between the subspace U spanned by the first r singular vectors of the input matrix and the r-dimensional space spanned by the columns of the iterate X_ℓ.

To show the convergence results of noisy tensor column subspace iteration, we use the largest principal angle between two tensor-column subspaces as the potential function. Borrowing idea from [32], we show that the noisy tensor-column subspace iteration can be transformed to k *parallel* noisy subspace iterations in the frequency domain.

Lemma 11. *The noisy tensor-columns subspace iteration in Alg. 7 converges at a geometric rate*[7].

Proof: The key iterative operations in Alg. 7 are

$$\begin{aligned} \mathcal{Z}_\ell &\leftarrow \mathcal{T} * \mathcal{X}_{\ell-1} + \mathcal{G}_\ell, \\ \mathcal{X}_\ell &\leftarrow \text{GS}(\mathcal{Z}_\ell). \end{aligned} \tag{41}$$

Introducing the $\text{cft}(\cdot)$ opertion in (14), we know that (41) be represented as follows:

$$\text{cft}(\mathcal{Y}_\ell) \leftarrow \text{cft}(\mathcal{X})\text{cft}(\mathcal{X}_{\ell-1}) + \text{cft}(\mathcal{G}_\ell). \tag{42}$$

This implies that (41) equals to k parallel standard noisy subspace iteration in the frequency domain. Therefore, combining the convergence results of [29] that noisy subspace iteration converges at a geometric rate, our noisy tensor-columns subspace iteration in Alg. 7 will also converge at a geometric rate.

■

In the following, we first provide the definitions of principal angles and corresponding inequalities for the matrix case [29]. Then, we need to establish explicit inequalities along the iteration process, so that we will be able to bound the recovery error of Alg. 2.

[7]We do not explicitly state the convergence rate because the one in [29] depends on the condition number, while we encounter a constant convergence rate from our experiments.

Definition 17. ***Largest principal angle***. *Let $X, Y \in \mathbb{R}^{n\times r}$ be orthonormal bases for subspaces $\mathcal{S}_X, \mathcal{S}_Y$, respectively. Then, the sine of the* largest principal angle *between $\mathcal{S}_X$ and $\mathcal{S}_Y$ is defined as* $\sin\theta(\mathcal{S}_X, \mathcal{S}_Y) \doteq ||(I - XX^\dagger)Y||$.

Lemma 12. *(Matrix Local Convergence) Let $0 \le \epsilon \le 1/4$, $\Delta = \max_{1\le\ell\le L} ||G_\ell||$, and $\gamma_r = 1-\sigma_{r+1}/\sigma_r$. Assume that $||V^\dagger X_0|| \le 1/4$ and $\sigma_r \ge 8\Delta/\gamma_r\epsilon$. Then*

$$||V^\dagger X_L|| \le \max\{\epsilon, 2||V^\dagger X_0|| \exp(\gamma_r L/2)\} \tag{43}$$

Similarly we can prove the following lemma for our tensor case.

Lemma 13. *(Tensor Local Convergence) Let $0 \le \epsilon \le 1/4$, $\Delta = \max_{1\le\ell\le L} ||\mathcal{G}_\ell||$, and $\gamma_{rk} = 1 - \overline{\sigma}_{rk+1}/\overline{\sigma}_{rk}$. Assume that $||\mathcal{V}^\dagger * \mathcal{X}_0|| \le 1/4$ and $\overline{\sigma}_{rk} \ge 8\Delta/\gamma_{rk}\epsilon$. Then*

$$||\mathcal{V}^\dagger * \mathcal{X}_L|| \le \max\{\epsilon, 2||\mathcal{V}^\dagger * \mathcal{X}_0|| \cdot \exp(\gamma_{rk} L/2)\} \tag{44}$$

Proof: According to Lemma 2, we have

$$||\mathcal{V}^\dagger * \mathcal{X}_L|| = ||\overline{\mathcal{V}^\dagger * \mathcal{X}_L}|| = ||\overline{\mathcal{V}^\dagger}\ \overline{\mathcal{X}_L}||. \tag{45}$$

This means that the largest principle angle between $\mathcal{V}^\dagger$ and $\mathcal{X}_L$ equals to that of these two tensor-column subspaces in the frequency domain.

Note that $||\mathcal{V}^\dagger * \mathcal{X}_0|| \le 1/4$ will be provided in Lemma 5, thus $||\overline{\mathcal{V}^\dagger}\ \overline{\mathcal{X}_0}|| \le 1/4$. Let $\Delta = \max_{1\le\ell\le L} ||\mathcal{G}_\ell||$, $\gamma_{rk} = 1 - \overline{\sigma}_{rk+1}/\overline{\sigma}_{rk}$, and $\overline{\sigma}_{rk} \ge 8\Delta/\gamma_{rk}\epsilon$. From Definition 5 and 7, we know that a tensor $\mathcal{T}$ with tubal-rank r has a corresponding block diagonal matrix $\overline{\mathcal{T}}$ with rank rk. Therefore, applying Lemma 12 we get

$$\begin{aligned} ||\overline{\mathcal{V}^\dagger}\ \overline{\mathcal{X}_L}|| &\le \max\{\epsilon, 2||\overline{\mathcal{V}^\dagger}\ \overline{\mathcal{X}_0}|| \cdot \exp(\gamma_{rk} L/2)\} \\ &= \max\{\epsilon, 2||\mathcal{V}^\dagger * \mathcal{X}_0|| \cdot \exp(\gamma_{rk} L/2)\}. \end{aligned} \tag{46}$$

Combining with (45), the lemma is proof. ■

To prove the convergence of the noisy tensor-column subspace iteration, we show that the error term $||\mathcal{G}_\ell||$ decrease as ℓ increases and Alg. 7 starts to converge. We define the following condition as a convergence bound for this type of shrinking error.

Definition 18. *(Tensor Admissible). Let $\gamma_{rk} = 1 - \overline{\sigma}_{rk+1}/\overline{\sigma}_{rk}$. We say that the pair of tensors $(\mathcal{X}_{\ell-1}, \mathcal{G}_\ell)$ is ϵ-admissible for noisy tensor-column subspace iteration if*

$$||\mathcal{G}_\ell|| \le \frac{1}{32}\gamma_{rk}\overline{\sigma}_{rk}||\mathcal{V}^\dagger * \mathcal{X}_{\ell-1}|| + \frac{\epsilon}{32}\gamma_{rk}\overline{\sigma_{rk}}. \tag{47}$$

One can say that a sequence of tensors $\{(\mathcal{X}_{\ell-1}, \mathcal{G}_\ell)\}$ is ϵ-admissible for noisy tensor-column subspace iteration if each element of this sequence is ϵ-admissible. In the following we will use the notation $\{(\mathcal{G}_\ell$ as a shorthand for $\{(\mathcal{X}_{\ell-1}, \mathcal{G}_\ell)\}_{\ell=1}^{L}$.

With Lemma 13 and Definition 18, we are able to get the following convergence guarantee for admissible noise tensors.

Theorem 4. *Let $\gamma_{rk} = 1 - \overline{\sigma}_{rk+1}/\overline{\sigma}_{rk}$, and $\epsilon \leq 1/2$. Assume that the sequence of noisy tensors $\{\mathcal{G}_\ell\}$ is $(\epsilon/2)$-admissible for the noisy tensor-columns subspace iteration and that $||\mathcal{V}^\dagger * \mathcal{X}_0|| \leq 1/4$. Then, we have $||\mathcal{V}^\dagger * \mathcal{X}_L|| \leq \epsilon$ for any $L \geq 4\gamma_{rk}^{-1} \log(1/\epsilon)$.*

Proof: We prove by induction that for every integer $t \geq 0$ after $L_t = 4t\gamma_{rk}^{-1}$ steps, we have $||\mathcal{V}^\dagger * \mathcal{X}_{L_t}|| \leq \max\{2^{-(t+1)}, \epsilon\}$. For the base case $t = 0$, the lemma holds because of the assumption that $||\mathcal{V}^\dagger * \mathcal{X}_0|| \leq 1/4$. For $t \geq 1$, we assume that $||\mathcal{V}^\dagger * \mathcal{X}_{L_t}|| \leq \max\{2^{-(t+1)}, \epsilon\}$. Apply Lemma 13 with $\mathcal{X}_0 = \mathcal{X}_{L_t}$, error parameter $\max\{2^{-t+2}, \epsilon\}$ and $L = L_{t+1} - L_t = 4/\gamma_{rk}$. The conditions of the lemma are satisfied due to the assumption that $\{\mathcal{G}_\ell\}$ is $\epsilon/2$-admissible. Therefore, we get

$$||\mathcal{V}^\dagger * \mathcal{X}_{L_{t+1}}|| \leq \max\{\epsilon, 2\max\{2^{-(t+1)}, \epsilon\} \cdot \exp(\gamma_{rk}(L_{t+1} - L_t)/2)\} \leq \max\{\epsilon, 2^{-(t+2)}\} \tag{48}$$

■

C. Initialization

Alg. 3 computes the top-r eigenslices of $\mathcal{P}_\Omega(\mathcal{T})$, and truncates them in a scaling manner to ensure incoherence. This initialization procedure servers as an acceleration of our main algorithm Alg. 2. We analyze Alg. 3 and derive the required sampling probability p_0 to get a good starting point in Theorem 5. The corresponding proofs relies mainly on the matrix Bernstein inequality in Lemma 14 and the Davis-Kahan $\sin\theta$-theorem [39] in Lemma 16.

Theorem 5. *Let $\mathcal{T} = \mathcal{U} * \Theta * \mathcal{V}^\dagger \in \mathbb{R}^{n\times n\times k}$ be a symmetric square tensor with tubal-rank r. Assume that each element is included in Ω independently with probability*

$$p_0 \geq \frac{6144r^2\mu(\mathcal{U})(||\mathcal{T}||_F/\overline{\gamma_{rk}\sigma_{rk}})^2 \log n}{n} + \frac{64r^{3/2}\mu(\mathcal{U})(||\mathcal{T}||_F/\overline{\gamma_{rk}\sigma_{rk}}) \log n}{n} \tag{49}$$

where $\overline{\sigma_{rk}}$ denotes the rk-th singular value of the block diagonal matrix $\overline{\mathcal{T}}$, and $\overline{\gamma_{rk}} = 1 - \overline{\sigma_{rk+1}}/\overline{\sigma_{rk}}$. Then, Alg.3 returns an orthonormal tensor $\mathcal{X} \in \mathbb{R}^{n\times r\times k}$ such that with probability at least $1 - 1/n^2$, we have

$$||\mathcal{V}^\dagger * \mathcal{X}||_F \leq 1/4, \quad \textit{and} \quad \mu(\mathcal{X}) \leq 32\mu(\mathcal{U}) \log n. \tag{50}$$

Proof: The proof follows directly from Lemma 15, Lemma 16, Lemma 17, and Lemma 18. ■

Lemma 14. *(Matrix Bernstein Inequality) Consider a finite sequence $\{Z_i\}$ of independent random matrices with dimensions $d_1 \times d_2$. Assume that each random matrix satisfies $\mathbb{E}Z_i = 0$ and $||Z_i|| \leq R$ almost surely. Define $\zeta^2 \doteq \max\{||\sum_i \mathbb{E}Z_i Z_i^\dagger||, ||\sum_i \mathbb{E}Z_i^\dagger Z_i||\}$. Then, for all $u \geq 0$,*

$$\mathbb{P}\left\{\left\|\sum_i Z_i\right\| \geq u\right\} \leq (d_1+d_2)\exp\left\{\frac{-u^2/2}{\zeta^2 + Ru/3}\right\}. \tag{51}$$

Lemma 15. *Suppose that $\mathcal{T} \in \mathbb{R}^{n\times n\times k}$ and let $\Omega \in [n]\times[n]\times[k]$ be a random subset with each entry being included independently with probability p_0. Then*

$$\mathbb{P}\left\{||\mathcal{P}_\Omega(\mathcal{T}) - \mathcal{T}|| > u\right\} \leq n\exp\left\{\frac{-u^2/2}{\zeta^2 + \frac{u}{3p_0}||\mathcal{T}||_\infty}\right\} \tag{52}$$

where $\zeta^2 = 1/p_0 \max\left\{||\mathcal{T}||^2_{\infty,2^*}, ||\mathcal{T}||_\infty\right\}$.

Proof: Define a random variable $\xi_{ij\ell} = 1_{(i,j,\ell)\in\Omega}$ where $1_{(\cdot)}$ is the indicator function. Consider the sum of independent random tensors $\mathcal{P}_\Omega(\mathcal{T}) - \mathcal{T} = \sum_{i,j,\ell}(\frac{\xi_{ij\ell}}{p_0} - 1)\mathcal{T}_{ij\ell}\dot{e}_i * e_\ell * \dot{e}_j^\dagger$, where $\dot{e}_i \in \mathbb{R}^{n\times 1\times k}$ is the column basis with $\dot{e}_i(i,1,1) = 1$, and $e_\ell \in \mathbb{R}^{1\times 1\times k}$ is the tube basis with $e_\ell(1,1,\ell) = 1$.

Let $\mathcal{E}_{ij\ell} = (\frac{\xi_{ij\ell}}{p_0} - 1)\mathcal{T}_{ij\ell}\dot{e}_i * e_\ell * \dot{e}_j^\dagger$, and $\overline{\mathcal{E}_{ij\ell}} = (\frac{\xi_{ij\ell}}{p_0} - 1)\mathcal{T}_{ij\ell}\bar{\dot{e}}_i\bar{e}_\ell\bar{\dot{e}}_j^\dagger$. Notice that $\mathbb{E}[\overline{\mathcal{E}}_{ij\ell}] = 0$ and $||\overline{\mathcal{E}}_{ij\ell}|| \leq \frac{1}{p_0}||\mathcal{T}||_\infty$.

$$\begin{aligned}
\left\|\mathbb{E}\left[\sum_{i,j,\ell}\overline{\mathcal{E}}^\dagger_{ij\ell}\overline{\mathcal{E}}_{ij\ell}\right]\right\| &= \left\|\mathbb{E}\left[\sum_{i,j,\ell}\mathcal{E}^\dagger_{ij\ell} * \mathcal{E}_{ij\ell}\right]\right\| \\
&= \left\|\sum_{ij\ell}\mathcal{T}^2_{ij\ell}\dot{e}_j * \dot{e}_j^\dagger \mathbb{E}\left(\frac{1}{p_0}\xi_{ij\ell} - 1\right)^2\right\| \\
&= \left\|\frac{1-p_0}{p_0}\sum_{ij\ell}\mathcal{T}^2_{ij\ell}\dot{e}_j * \dot{e}_j^\dagger\right\|
\end{aligned} \tag{53}$$

since $\dot{e}_j * \dot{e}_j^\dagger$ will return a zero tensor except for $(j,j,1)$-th entry equaling 1, we have

$$\left\|\mathbb{E}\left[\sum_{i,j,\ell}\overline{\mathcal{E}}^\dagger_{ij\ell}\overline{\mathcal{E}}_{ij\ell}\right]\right\| = \frac{1-p_0}{p_0}\max_j\left|\sum_{i,\ell}\mathcal{T}^2_{ij\ell}\right| \leq \frac{1}{p_0}||\mathcal{T}||^2_{\infty,2^*}, \tag{54}$$

And similarly, $\left\|\mathbb{E}\left[\sum_{i,j,\ell}\overline{\mathcal{E}}_{ij\ell}\overline{\mathcal{E}}^\dagger_{ij\ell}\right]\right\| \leq \frac{1}{p_0}||\mathcal{T}||_\infty$ which is bounded. Then using Lemma 14 concludes the proof.

■

Lemma 16. *To assure that* $||\mathcal{P}_\Omega(\mathcal{T})-\mathcal{T}|| \leq \frac{\overline{\gamma_{rk}\sigma_{rk}}}{32\sqrt{r}}$ *holds with high probability at least* $1-\frac{1}{n^2}$*, it requires that*

$$p_0 \geq \frac{6144r^2\mu(\mathcal{U})(||\mathcal{T}||_F/\overline{\gamma_{rk}\sigma_{rk}})^2\log n}{n} + \frac{64r^{3/2}\mu(\mathcal{U})(||\mathcal{T}||_F/\overline{\gamma_{rk}\sigma_{rk}})\log n}{n} \tag{55}$$

where $\overline{\sigma_{rk}}$ *denotes the* rk*-th singular value of the block diagonal matrix* $\overline{\mathcal{T}}$*, and* $\overline{\gamma_{rk}} = 1-\overline{\sigma_{rk+1}}/\overline{\sigma_{rk}}$*. Then, we have*

$$\mathbb{P}\left\{||\mathcal{V}^\dagger * \mathcal{W}|| \leq \frac{1}{16\sqrt{r}}\right\} > 1-1/n^2. \tag{56}$$

Proof: We have that

$$\begin{aligned} ||\mathcal{T}||^2_{\infty,2^*} &= \frac{r\mu(\mathcal{U})}{n}||\mathcal{T}||^2_F, \\ ||\mathcal{T}||_\infty &= \frac{r\mu(\mathcal{U})}{n}||\mathcal{T}||_F. \end{aligned} \tag{57}$$

Then ζ^2 in Lemma 15 becomes

$$\zeta^2 = \frac{1}{p_0}\max\left\{\frac{r\mu(\mathcal{U})}{n}||\mathcal{T}||^2_F, \frac{r\mu(\mathcal{U})}{n}||\mathcal{T}||_F\right\} = \frac{r\mu(\mathcal{U})}{p_0 n}||\mathcal{T}||^2_F. \tag{58}$$

Set the right hand side of (52) to be $\leq n^{-(C-1)}$, then taking log-function we have:

$$\begin{aligned} &-\frac{u^2}{2} \leq -C\left(\zeta^2 + \frac{u}{3p_0}||\mathcal{T}||_\infty\right)\log n = -C\left(\frac{r\mu(\mathcal{U})}{p_0 n}||\mathcal{T}||^2_F + \frac{u}{3p_0}\frac{r\mu(\mathcal{U})}{n}||\mathcal{T}||_F\right)\log n, \\ &u^2 - 2C\log n\frac{r\mu(\mathcal{U})}{3p_0 n}||\mathcal{T}||_F\ u - 2C\log n\frac{r\mu(\mathcal{U})}{p_0 n}||\mathcal{T}||^2_F \geq 0. \end{aligned} \tag{59}$$

This can be re-arranged to get

$$p_0 \geq \frac{2Cr\mu(\mathcal{U})\log n}{n}\ \frac{||\mathcal{T}||^2_F}{u^2} + \frac{2Cr\mu(\mathcal{U})\log n}{n}\ \frac{||\mathcal{T}||_F}{u}. \tag{60}$$

Set $u = \frac{\gamma^c_{rk}\sigma^c_{rk}}{32\sqrt{r}}$ (γ^c_{rk} and σ^c_{rk} are introduced for reasons to be clear in (62)), leading to the condition that $p_0 \geq \frac{2048Cr^2\mu(\mathcal{U})(||\mathcal{T}||_F/\gamma^c_{rk}\sigma^c_{rk})^2\log n}{n} + \frac{64Cr^{3/2}\mu(\mathcal{U})(||\mathcal{T}||_F/\gamma^c_{rk}\sigma^c_{rk})\log n}{3n}$. Let $C=3$, plug in the above parameters into Lemma 15, we get:

$$\mathbb{P}\left\{||\mathcal{P}_\Omega(\mathcal{T})-\mathcal{T}|| > \frac{\gamma^c_{rk}\sigma^c_{rk}}{32\sqrt{r}}\right\} \leq \frac{1}{n^2}. \tag{61}$$

Let $\mathcal{W}$ be the top r eigenslices of $\mathcal{P}_\Omega(\mathcal{T})$, σ^c_{rk} denote the rk-th singular value of the circular matrix T^c and define $\gamma^c_{rk} = 1-\sigma^c_{rk+1}/\sigma^c_{rk}$. Now let us assume that $||\mathcal{T}-\mathcal{P}_\Omega(\mathcal{T})|| \leq u$, then $||T^c - P_{\Omega'}(T^c)|| \leq u$, thus $\sigma^c_{rk}(P_{\Omega'}(T^c)) > \sigma^c_{rk}(T^c) - u > \sigma^c_{rk} - \gamma^c_{rk}\sigma^c_{rk}/2$, and $\sigma^c_{rk}(P_{\Omega'}(T^c)) - \sigma^c_{rk+1}(T^c) = \sigma^c_{rk}(P_{\Omega'}(T^c)) - \sigma^c_{rk} + \gamma^c_{rk}\sigma^c_{rk} > \gamma^c_{rk}\sigma^c_{rk}/2$.

By Davis-Kahan $\sin\theta$-theorem [39] and combining Definition 12, we have that

$$\begin{aligned} ||\mathcal{V}^\dagger * \mathcal{W}|| = ||V^{c\dagger}W^c|| = \sin\theta_{rk}(U^c, W^c) &\leq \frac{||T^c - P_{\Omega'}(T^c)||}{\sigma^c_{rk}(P_{\Omega'}(T^c)) - \sigma^c_{rk+1}(T^c)} \\ &\leq \frac{u}{\sigma^c_{rk}(P_{\Omega'}(T^c)) - \sigma^c_{rk+1}(T^c)} \leq \frac{2u}{\gamma^c_{rk}\sigma^c_{rk}} = \frac{1}{16\sqrt{r}}. \end{aligned} \tag{62}$$

Note that the rk-th singular value σ^c_{rk} of the circular matrix T^c equals to that (i.e., $\overline{\sigma_{rk}}$) of the block diagonal matrix $\overline{\mathcal{T}}$. Then the probability formula becomes $p_0 \geq \frac{6144r^2\mu(\mathcal{U})(||\mathcal{T}||_F/\overline{\gamma_{rk}\sigma_{rk}})^2\log n}{n} + \frac{64r^{3/2}\mu(\mathcal{U})(||\mathcal{T}||_F/\overline{\gamma_{rk}\sigma_{rk}})\log n}{n}$. ■

Lemma 17. *Let $\mathcal{W} \in \mathbb{R}^{n\times r\times k}$ be any orthonormal basis with $\mathcal{W} \leq \mu$. Then, for a random orthonormal tensor $\mathcal{O} \in \mathbb{R}^{r\times r\times k}$, we have $\mathbb{P}\left\{max_{i,j}||[\mathcal{W}*\mathcal{O}](i,j,:)||_F > \sqrt{8\mu\log n/n}\right\} \leq \frac{1}{n^2}$.*

Proof: Consider a tube $Z = [\mathcal{W}*\mathcal{O}](i,j,:)$. Note that $||Z||_F$ is distributed like a coordinate of a random vector $\mathbb{R}^r$ of norm at most $\sqrt{\mu r/n}$. By measure concentration, we obtain:

$$\mathbb{P}\left\{||Z||_F > \epsilon\sqrt{\mu r/n}\right\} \leq 4\exp\{-\epsilon^2 r/2\}. \tag{63}$$

This follows from Levy's Lemma [40], as in [29]. Set $\epsilon = \sqrt{8\log n/r}$, then

$$\mathbb{P}\left\{||Z||_F > \sqrt{8\mu\log n/n}\right\} \leq 4\exp\{-4\log n\} = 4n^{-4}. \tag{64}$$

Taking a union bound over all $nr \leq n^2/4$ tubes of $\mathcal{W}*\mathcal{O}$, we have that with probability $1-1/n^2$,

$$\max_{i,j}||[\mathcal{W}*\mathcal{O}](i,j,:)||_F \leq \sqrt{8\mu\log n/n}. \tag{65}$$

■

Lemma 18. *Assume that $||\mathcal{V}^\dagger * \mathcal{W}|| \leq \frac{1}{16\sqrt{r}}$. Then with probability $1-1/n^2$, we have $||\mathcal{V}^\dagger * \mathcal{X}||_F \leq 1/4$ and $\mu(\mathcal{X}) \leq 32\mu(\mathcal{U})\log n$.*

Proof: Assume that $||\mathcal{V}^\dagger * \mathcal{W}|| \leq \frac{1}{16\sqrt{r}}$, then there exists an orthonormal transformation $\mathcal{Q} \in \mathbb{R}^{r\times r\times k}$ such that $||\mathcal{U}*\mathcal{Q}-\mathcal{W}||_F \leq 1/16$. Because of the following three facts: $||\mathcal{U}*\mathcal{Q}-\mathcal{W}||_F \leq \left(\sum_{i=1}^r ||\sigma_i||_F^2\right)^{1/2}$ where σ_i denotes the i-th eigentube of $(\mathcal{U}*\mathcal{Q}-\mathcal{W})$, $||\sigma_1||_F \geq ||\sigma_2||_F \geq ... \geq ||\sigma_r||_F$, and $||\sigma_1||_F = ||\mathcal{V}^\dagger * \mathcal{W}|| \leq \frac{1}{16\sqrt{r}}$.

Since $\mu(\mathcal{U}*\mathcal{Q}) = \mu(\mathcal{U}) \leq \mu$ (the orthonormal transformation $\mathcal{Q}$ does not change the incoherence of $\mathcal{U}$). Therefore, $\mathcal{W}$ is close in Frobenius norm to an orthonormal basis of small coherence. However, it is possible that some tubes of $\mu(\mathcal{U}*\mathcal{Q})$ have Frobenius norm as large as $\sqrt{\mu r/n}$. Rotating $\mu(\mathcal{U}*\mathcal{Q})$ by a random rotation $\mathcal{O}$, Lemma 17 asserts that with probability $1-1/n^2$, $||[\mathcal{U}*\mathcal{Q}*\mathcal{O}](i,j,:)||_F \leq \mu' = \sqrt{8\mu\log n/n}$, for all i,j. Moreover, because a rotation does not increase Frobenius norm, then we have $||\mathcal{U}*\mathcal{Q}*\mathcal{O} - \mathcal{W}*\mathcal{O}||_F \leq 1/16$. Truncating the tubes of $\mathcal{W}*\mathcal{O}$ that has Frobenius norm larger than μ' to μ' can therefore only decrease the distance in Frobenius norm to $\mathcal{U}*\mathcal{Q}*\mathcal{O}$, hence, $||\mathcal{U}*\mathcal{Q}*\mathcal{O} - \mathcal{Z}'||_F \leq 1/16$.

Since truncation is a projection onto the set $\{\mathcal{B} : ||\mathcal{B}(i,j,:)||_F \leq \mu'\}$ with respect to Frobenius norm, we have:

$$||\mathcal{W} * \mathcal{O} - \mathcal{Z}'||_F \leq ||\mathcal{U} * \mathcal{Q} * \mathcal{O} - \mathcal{Z}'||_F \leq \frac{1}{16}. \tag{66}$$

We can write $\mathcal{X} = \mathcal{Z}' * \mathcal{R}^{-1}$ where $\mathcal{R}$ is an invertible linear transformation with the same eigentubes as $\mathcal{Z}'$ and thus satisfies

$$||\mathcal{R}^{-1}|| = \frac{1}{||\sigma_1(\mathcal{Z}')||_F} \leq \frac{1}{||\sigma_1(\mathcal{W} * \mathcal{O})||_F - ||\sigma_1(\mathcal{W} * \mathcal{O} - \mathcal{Z}')||_F} \leq \frac{1}{1 - 1/16} \leq 2. \tag{67}$$

Therefore,

$$||\dot{e}_i^\dagger * \mathcal{X}|| = ||\dot{e}_i^\dagger * \mathcal{T} * \mathcal{R}^{-1}|| \leq ||\dot{e}_i^\dagger * \mathcal{T}||||\mathcal{R}^{-1}|| \leq 2||\dot{e}_i^\dagger * \mathcal{T}|| \leq 2\sqrt{8r\mu(\mathcal{U}) \log n / n}. \tag{68}$$

Hence,

$$\mu(\mathcal{X}) \leq \frac{n}{r} \frac{32r\mu(\mathcal{U}) \log n}{n} \leq 32\mu(\mathcal{U}) \log n. \tag{69}$$

$$\begin{aligned} ||\mathcal{V}^\dagger * \mathcal{X}||_F = ||\mathcal{V}^\dagger * \mathcal{T} * \mathcal{R}^{-1}||_F &\leq ||\mathcal{V}^\dagger * \mathcal{T}||_F ||\mathcal{R}^{-1}|| \leq 2||\mathcal{V}^\dagger * \mathcal{T}||_F \\ &\leq 2||\mathcal{V}^\dagger * \mathcal{W} * \mathcal{O}||_F + 2||\mathcal{W} * \mathcal{O} - \mathcal{T}||_F \leq 2||\mathcal{V}^\dagger * \mathcal{W}||_F + \frac{1}{8} \leq \frac{1}{4}. \end{aligned} \tag{70}$$

■

D. Incoherence via the SmoothQR Procedure

As a requirement for our proof in Appendix E, we need to show that each intermediate solution $\mathcal{Y}_\ell$ (accordingly $\mathcal{X}_\ell$) has small coherence. Lemma 20 states that applying the SmoothQR factorization (in Alg. 6) on $\mathcal{Y}_\ell$ will return a tensor $\mathcal{Z}$ satisfying this coherence requirement. Note that before orthonormalizing $\mathcal{Y}_\ell$, a small Gaussian perturbation $\mathcal{H}$ is added to $\mathcal{Y}_\ell$. There exists such noisy term that will cause little effect as long as its norm is bounded by that of $\mathcal{G}_\ell$.

Lemma 19. *Let $\mathcal{G} \in \mathbb{R}^{n \times r \times k}$ be any tensor with $||\mathcal{G}|| \leq 1$, $\mathcal{V} \in \mathbb{R}^{n \times (n-r) \times k}$ be a $(n-r)$ dimensional tensor-column subspace with orthogonal projection $\mathcal{P}_\mathcal{V}$, and $\mathcal{H} \in \mathbb{R}^{n \times r \times k} \sim \mathcal{N}(0, \tau^2/n)$ be a random Gaussian tensor. Assume that $r = o(n/\log n)$ where $o(n)$ denotes an order that is lower than n. Then, with probability $1 - \exp(-\Omega(n))$, we have $\sigma_{rk}(\overline{\mathcal{P}_\mathcal{V}(\mathcal{G} + \mathcal{H})}) \geq \Omega(\tau)$.*

Proof: Consider a tensor $\mathcal{X} \in \mathbb{R}^{r \times 1 \times k}$ with $||\mathcal{X}||_F = 1$, we have

$$||\mathcal{P}_\mathcal{V}(\mathcal{G} + \mathcal{H}) * \mathcal{X}||^2 > ||\mathcal{P}_\mathcal{V} * \mathcal{H} * \mathcal{X}||^2 - |\langle \mathcal{P}_\mathcal{V} * \mathcal{G} * \mathcal{X}, \mathcal{P}_\mathcal{V} * \mathcal{H} * \mathcal{X} \rangle|. \tag{71}$$

Note that $g = \mathcal{H} * \mathcal{X} \in \mathbb{R}^{n \times 1 \times k}$ follows the distribution $N(0, \tau^2/n)^{n \times 1 \times k}$, $y = \mathcal{P}_\mathcal{V} * \mathcal{G} * \mathcal{X}$ has spectral norm at most 1, and $\mathcal{V}$ is a $n - k$ dimensional tensor-column subspace, and $h = \mathcal{P}_\mathcal{V} * \mathcal{H} * \mathcal{X}$ follows the

distribution $N(0,\tau^2/n)^{n\times 1\times k}$. Then, we need to lower bound $||h||^2 - |\langle y,h\rangle|$. Since $\mathbb{E}||h||^2 > \tau^2/2$, by standard concentration bounds for the norm of a Gaussian variable, we get

$$\mathbb{P}\{||h||^2 \leq \tau^2/4\} \leq \exp\Omega(n). \tag{72}$$

On the other hand, $\langle y,h\rangle$ is distributed like a one-dimensional Gaussian variable of variance at most τ^2/n. By Gaussian tail bounds, $\mathbb{P}\{\langle y,h\rangle^2 > \tau^2/8\} \leq \exp -\Omega(n)$. Therefore, with probability $1-\exp\Omega(n)$, we have $||\mathcal{P}_{\mathcal{V}}(\mathcal{G}+\mathcal{H})*\mathcal{X}|| > \Omega(\tau)$.

Taking a union bound over a set of the unit sphere in $\mathbb{R}^{r\times 1\times k}$ of size $\exp O(r\log r)$, we have that with probability $1-\exp(O(r\log r))\exp(-\Omega(n))$, $||\mathcal{P}_{\mathcal{V}}(\mathcal{G}+\mathcal{H})*\mathcal{X}|| > \Omega(\tau)$ for all unit tensors $\mathcal{X}\in\mathbb{R}^{r\times 1\times k}$, i.e., $\sigma_{rk}(\overline{\mathcal{P}_{\mathcal{V}}(\mathcal{G}+\mathcal{H})}) > \Omega(\tau)$.

Note that $\exp O(r\log r) = \exp(o(n))$, hence this event occurs with probability $1-\exp\Omega(n)$.

■

We introduce a variant of μ-coherence, i.e., ρ-coherence, that applies to tensors rather than tensor-column subspaces. The next lemma (Lemma 20) show that adding a Gaussian noise term leads to a bound on the coherence after applying the QR-factorization.

Definition 19. *(ρ-coherence). Given a tensor $\mathcal{G}\in\mathbb{R}^{n\times r\times k}$ we let $\rho(\mathcal{G}) \doteq \frac{n}{r}||\dot{e}_i^\dagger * \mathcal{G}||^2$.*

Lemma 20. *Let $r = \Omega(n/\log n)$ and $\tau\in(0,1)$. Let $\mathcal{U}\in\mathbb{R}^{n\times r\times k}$ be an orthonormal tensor, and $\mathcal{G}\in\mathbb{R}^{n\times r\times k}$ be a tensor such that $||\mathcal{G}||\leq 1$. Let $\mathcal{H}\sim\mathcal{N}(0,\tau^2/n)^{n\times r\times k}$ be a random Gaussian tensor. Then, with probability $1-\exp(-\Omega(n))-n^{-5}$, there exists an orthonormal tensor $\mathcal{Q}\in\mathbb{R}^{n\times 2r\times k}$ such that*

- *$\mathcal{R}(\mathcal{Q}) = \mathcal{R}([\mathcal{U}\mid\mathcal{G}+\mathcal{H}])$ where $\mathcal{R}(\mathcal{Q})$ denotes the range of $\mathcal{Q}$;*
- *$\mu(\mathcal{Q}) \leq O(\frac{1}{\tau}(\rho(\mathcal{Q})+\mu(\mathcal{U})+\log n))$.*

Proof: First, $\mathcal{R}([\mathcal{U}\mid\mathcal{G}+\mathcal{H}]) = \mathcal{R}([\mathcal{U}\mid(\mathcal{I}-\mathcal{U}*\mathcal{U}^\dagger)*(\mathcal{G}+\mathcal{H})])$. Let $\mathcal{B} = (\mathcal{I}-\mathcal{U}*\mathcal{U}^\dagger)*(\mathcal{G}+\mathcal{H})])$. Applying the QR-factorization to $[\mathcal{U}\mid\mathcal{B}]$, we can find two orthonormal tensors $\mathcal{Q}_1,\mathcal{Q}_2\in\mathbb{R}^{n\times r\times k}$ such that $[\mathcal{Q}_1\mid\mathcal{Q}_2] = [\mathcal{U}\mid\mathcal{B}*\mathcal{R}^{-1}]$ where $\mathcal{R}\in\mathbb{R}^{r\times r\times k}$. Since $\mathcal{U}$ is already orthonormal, we can have $\mathcal{Q}_1 = \mathcal{U}$. Furthermore, the lateral slices of $\mathcal{B}$ are orthogonal to $\mathcal{U}$ and thus we apply the QR-factorization to $\mathcal{U}$ and $\mathcal{B}$ independently.

Applying Lemma 19 to the $(n-r)$-dimensional tensor-column subspace $\mathcal{U}^\perp$ and the tensor $\mathcal{G}+\mathcal{H}$, we get that with probability $1-\exp(-\Omega(n))$, $\sigma_{rk}(\mathcal{B})\geq\Omega(\tau)$. Assume that this hold in the following.

We verify the second condition. We have

$$\frac{n}{r}||\dot{e}_i^\dagger * \mathcal{Q}||^2 = \frac{n}{r}||\dot{e}_i^\dagger * \mathcal{U}||^2 + \frac{n}{r}||\dot{e}_i^\dagger * \mathcal{B} * \mathcal{R}^{-1}||^2 = \mu(\mathcal{U}) + \frac{n}{r}||\dot{e}_i^\dagger * \mathcal{B} * \mathcal{R}^{-1}||^2. \tag{73}$$

On the other hand, we also have

$$\frac{n}{r}||\dot{e}_i^\dagger * \mathcal{B} * \mathcal{R}^{-1}||^2 \leq \frac{n}{r}||\dot{e}_i^\dagger * \mathcal{B}||^2||\mathcal{R}^{-1}||^2 \leq O\left(\frac{n}{r\tau^2}||\dot{e}_i^\dagger * \mathcal{B}||^2\right), \tag{74}$$

where we used the fact that $||\mathcal{R}^{-1}|| = 1/\sigma_{rk}(\mathcal{R}) = O(1/\tau)$.

Moreover, we have

$$\begin{aligned}\frac{n}{r}||\dot{e}_i^\dagger * \mathcal{B} * \mathcal{R}^{-1}||^2 &\leq 2\frac{n}{r}||\dot{e}_i^\dagger * (\mathcal{I} - \mathcal{U} * \mathcal{U}^\dagger)||^2 + 2\rho((\mathcal{I} - \mathcal{U} * \mathcal{U}^\dagger) * \mathcal{H}) \\ &\leq 2\rho(\mathcal{G}) + 2\rho(\mathcal{U} * \mathcal{U}^\dagger * \mathcal{G}) + 2\rho((\mathcal{I} - \mathcal{U} * \mathcal{U}^\dagger) * \mathcal{H}).\end{aligned} \tag{75}$$

Note that $\rho(\mathcal{U} * \mathcal{U}^\dagger * \mathcal{G}) \leq \mu(\mathcal{U})||\sqcap^\dagger * \mathcal{G}||^2 \leq \mu(\mathcal{G})$.

Combining the Lemma 21 (in the following), we have $(\mathcal{I} - \mathcal{U} * \mathcal{U}^\dagger) * \mathcal{H} \leq O(\log n)$ with probability $1 - n^{-5}$. Summing up the probability concludes the lemma. ■

Lemma 21. *Let $\mathcal{P}$ be the projection onto an $n - r$-dimensional tensor-column subspace. Let $\mathcal{H} \sim N(0, 1/n)^{n\times r\times k}$. Then, $\rho(\mathcal{P} * \mathcal{H}) \leq O(\log n)$ with probability $1 - 1/n^5$.*

Proof: Let $\mathcal{P} = (\mathcal{I} - \mathcal{U} * \mathcal{U}^\dagger)$ for some r-dimensional tensor-column basis $\mathcal{U}$. Then,

$$\rho(\mathcal{P} * \mathcal{H}) \leq \rho(\mathcal{H}) + \rho(\mathcal{U} * \mathcal{U}^\dagger * \mathcal{H}). \tag{76}$$

Using concentration bounds for the norm of each horizontal slice of $\mathcal{H}$ and a union bound over all horizontal slices, it follows that

$$\begin{aligned}\rho(\mathcal{H}) &\leq O(\log n) \quad \text{with probability} \quad 1 - \frac{1}{2}n^{-5}, \\ \rho(\mathcal{U} * \mathcal{U}^\dagger * \mathcal{H}) &\leq \rho(\mathcal{U})||\mathcal{U}^\dagger * \mathcal{H}||^2.\end{aligned} \tag{77}$$

Note that $\mathcal{U}^\dagger * \mathcal{H}$ is a Gaussian tensor following the distribution $N(0, 1/n)^{r\times r\times k}$, and its largest singular value satisfies $||\mathcal{U}^\dagger * \mathcal{H}||^2 \leq O(r \log n/n)$ with probability $1 - \frac{1}{2}n^{-5}$. Summing up the probability concludes the lemma. ■

The next theorem states that when SmoothQR is called on an input of the form $\mathcal{T} * \mathcal{X} + \mathcal{G}$ with suitable parameters, the algorithm outputs a tensor of the form $\mathcal{X}' = \text{QR}(\mathcal{T} * \mathcal{X} + \mathcal{G} + \mathcal{H})$ whose coherence is bounded in terms of $\mathcal{G}$ and $\rho(\mathcal{G})$, and $\mathcal{H}$ satisfies a bound on its norm.

Theorem 6. *Let $\tau > 0$, $r = \Omega(n/\log n)$, $\mathcal{G} \in \mathcal{R}^{n\times r\times k}$, and $\mathcal{X} \in \mathbb{R}^{n\times r\times k}$ be an orthonormal tensor such that $\upsilon \geq \max\{||\mathcal{G}||, ||\mathcal{N} * \mathcal{X}||\}$. There exist a constant $C > 0$, assume that*

$$\mu \geq \frac{C}{\tau^2}\left(\mu(\mathcal{U}) + \frac{\rho(\mathcal{G}) + \rho(\mathcal{N} * \mathcal{X})}{\upsilon^2} + \log n\right), \tag{78}$$

then, for every $\epsilon \leq \tau\upsilon$ satisfying $\log(n/\epsilon) \leq n$ and every $\mu \leq n$ we have with probability $1 - O(n^{-4})$, the algorithm SmoothQR (Alg. 6) terminates in $\log(n/\epsilon)$ steps and outputs $(\mathcal{X}', \mathcal{H})$ such that $\mu(\mathcal{X}') \leq \mu$ and $||\mathcal{H}|| \leq \tau\upsilon$.

Proof: If the algorithm SmoothQR (Alg. 6) terminates in an iteration where $\varsigma \leq \tau^2\upsilon^2/4$ (proved in Lemma 22), we claim that in this case, with probability $1 - \exp(-\Omega(n))$ we must have that $||\mathcal{H}|| \leq \tau\upsilon$. Assume that the algorithm SmoothQR (Alg. 6) terminates in an iteration where $\varsigma \leq \tau^2\upsilon^2/r$, then the algorithm takes at most $t = O(\log(n/\epsilon)) \leq O(n)$ steps.

Let $\mathcal{H}_1, ..., \mathcal{H}_t$ denote the random Gaussian tensors generated in each step. We claim that each of them satisfies $\mathcal{H} \leq \tau\upsilon$. Note that for all t we have $\mathbb{E}||\mathcal{H}_t||^2 \leq \tau^2\upsilon^2/4$. The claim therefore follows directly from tail bounds for the Frobenius norm of Gaussian random tensors and holds with probabilities $1 - \exp(-\Omega(n))$.

■

Lemma 22. *With probability $1 - O(n^{-4})$, Alg. 6 terminates in an iteration where $\varsigma \leq \tau^2\upsilon^2/4$.*

Proof: Consider the first iteration in which $\varsigma \leq \tau^2\upsilon^2/8$. Let us define $\mathcal{G}' = (\mathcal{N} * \mathcal{X} + \mathcal{G})/2\upsilon$. Apply Lemma 20 to the tensor $\mathcal{G}'$ which satisfies the required assumption that $||\mathcal{G}'|| \leq 1$. Lemma 20 states that with probability $1 - O(n^{-4})$, there is an orthonormal $n \times 2r \times k$ tensor $\mathcal{Q}$ such that

$$\begin{aligned} &\mathcal{R}(\mathcal{Q}) = \mathcal{R}([\mathcal{U} \mid \mathcal{G}' + \mathcal{H}]) = \mathcal{R}([\mathcal{U} \mid \mathcal{G} + \mathcal{N} * \mathcal{X} + \mathcal{H}]), \\ &\mu(\mathcal{Q}) \leq O(\frac{1}{\tau^2}(\rho(\mathcal{G} + \mu(\mathcal{U} + \log n)))). \end{aligned} \tag{79}$$

On one hand, we have

$$\mathcal{R}(\mathcal{X}') = \mathcal{R}(\mathcal{T} * \mathcal{X} + \mathcal{G} + \mathcal{H}) = \mathcal{R}(\mathcal{M} * \mathcal{X} + \mathcal{N} * \mathcal{X} + \mathcal{G} + \mathcal{H}) \subset \mathcal{R}([\mathcal{U} \mid \mathcal{N} * \mathcal{X} + \mathcal{G} + \mathcal{H}]) = \mathcal{R}((V)), \tag{80}$$

where we use the fact that $\mathcal{U}$ is an orthonormal basis for the range of $\mathcal{M} * \mathcal{X} = \mathcal{U} * \Theta_{\mathcal{U}} * \mathcal{U}^\dagger * \mathcal{X}$. On the other hand, $\rho(\mathcal{G}') = O(\rho(\mathcal{G}/\upsilon) + \rho(\mathcal{N} * \mathcal{X}/\upsilon'))$.

Therefore, combining Lemma 23 and the fact that $\dim(\mathcal{Q}) \leq 2\dim(\mathcal{X}')$ where $\dim(\cdot)$ denotes the dimension, we have $\mu(\mathcal{X}') \leq 2\mu(\mathcal{Q}) \leq \mu$. This lemma is proved as long as C is large enough.

■

Lemma 23. *Let $\mathcal{X}, \mathcal{Y}$ be r and r' dimensional tensor-column subspaces, respectively, such that $\mathcal{X} \subset \mathcal{Y}$. Then, we have $\mathcal{X} \leq \frac{r'}{r}\mu(\mathcal{Y})$.*

Proof: We know that $\mu(\mathcal{Y})$ is rotationally invariant. Therefore, without loss of generality we assume that $\mathcal{Y} = [\mathcal{X} \mid \mathcal{X}']$ for some orthonormal tensor $\mathcal{X}'$. Here, we identify $\mathcal{X}$ and $\mathcal{Y}$ with orthonormal bases. Therefore,

$$\mu(\mathcal{X}) = \frac{n}{r}\max_{i\in[n]}||\dot{e}_i^\dagger * \mathcal{X}||^2 \leq \frac{n}{r}\max_{i\in[n]}\left(||\dot{e}_i^\dagger * \mathcal{X}||^2 + ||\dot{e}_i^\dagger * \mathcal{X}'||^2\right) = \frac{n}{r}\max_{i\in[n]}||\cdot e_i^\dagger * \mathcal{Y}|| = \frac{r'}{r}\mu(\mathcal{Y}). \quad (81)$$

■

E. Proof of Theorem 2

To guarantee the global optima of the iterative algorithm Alg. 2, we follow the standard two-stage approach. In the first stage, we prove that Alg. 3 (the first step of Alg. 2) starts from a good initial point that is relatively close to the global optima, while in the second stage, Alg. 4 (the second step of Alg. 2) converges locally.

Theorem 5 in Appendix C states that $\mathcal{X}_0$ is relatively close to the tensor-columns subspace $\mathcal{U}$ well such that $||\mathcal{V}^\dagger * \mathcal{X}||_F \leq 1/4$ with high probability, if the elements are included in Ω_0 with at least probability $p_0 \geq \frac{6144r^2\mu(\mathcal{U})(||\mathcal{T}||_F/\overline{\gamma_{rk}\sigma_{rk}})^2\log n}{n} + \frac{64r^{3/2}\mu(\mathcal{U})(||\mathcal{T}||_F/\overline{\gamma_{rk}\sigma_{rk}})\log n}{n}$.

One the one hand, the convergence of Alg. 2 relies on the fact that alternating Alg. 4 will converge. First, we express each iteration as a noisy tensor-column subspace iteration in Lemma 7. Second, we show that it converges at a geometric rate (whose value is unknown) in Lemma 11 if the error term $\mathcal{G}$ is bounded in Lemma 9. Third, to guarantee $\mathcal{G}$ is bounded in Frobenius norm, we obtain a requirement that the elements are included in Ω_+ with at least probability $p_+ = O\left(\frac{r\mu(\mathcal{X})\log nk}{\delta^2 n}\right)$ in Lemma 8.

Therefore, the sampling complexity is the sum of two terms: the samples required by the initialization process, and the samples required to bound the error term in the tensor least squares minimization steps, i.e., $p = p_0 + p_+$. The following theorem shows that if the sampling complexity is larger than the sum of these two terms, Alg. 2 converges to the true unknown tensor rapidly.

Theorem 7. *Given a sample set Ω of size $O(pn^2k)$, each element randomly drawn with probability $p = p_0 + p_+$ from an unknown $n \times n \times k$ (symmetric square) tensor $\mathcal{T} = \mathcal{M} + \mathcal{N}$ where $\mathcal{M} = \mathcal{U} * \Theta * \mathcal{U}^\dagger$ of tubal-rank r and $\mathcal{N}$ satisfies condition (34). Let $\gamma_{rk} = 1 - \sigma_{rk+1}/\sigma_{rk}$ where σ_{rk} is the smallest singular value of $\overline{\mathcal{M}}$ and σ_{rk+1} is the largest singular value of $\overline{\mathcal{N}}$ ($\sigma_{rk+1} = 0, \gamma_{rk} = 1$ for the exact tensor completion problem).*

Then, there exist parameters $\mu = \Theta(\gamma_{rk}^{-2} r(\mu^* + \log n))$ *and* $L = \Theta(\gamma_{rk}^{-2} \log(n/\epsilon))$ *such that Alg. 2 will output* $(\widehat{\mathcal{X}}, \widehat{\mathcal{Y}})$ *such that* $||(\mathcal{I} - \mathcal{U} * \mathcal{U}^\dagger) * \widehat{\mathcal{X}}|| \leq \epsilon$.

Before the proof, we state the following remark on the reconstruction error in the Frobenius norm.

Remark 5. *Under the assumptions of Theorem 7, the output* $(\widehat{\mathcal{X}}, \widehat{\mathcal{Y}})$ *of Alg. 2 satisfies* $\mathcal{T} - \mathcal{X} * \mathcal{Y}^\dagger \leq \epsilon ||\mathcal{T}||_F$.

Proof: First, Theorem 5 concludes that with probability at least $1 - 1/n^2$, the initial tensor $\mathcal{X}_0$ satisfies that $||\mathcal{V}^\dagger * \mathcal{X}_0|| \leq 1/4$ and $\mu(\mathcal{X}_0) \leq 32\mu(\mathcal{U}) \log n$. Assume that this condition holds, then our goal is to apply Theorem 4, leading to our final bound of recovery error.

Consider the sequence of tensors $\{(\mathcal{X}_{\ell-1}, \widetilde{\mathcal{G}}_\ell)\}_{\ell=1}^L$ obtained along the iterations of Alg. 2. Let $\widetilde{\mathcal{G}}_\ell = \mathcal{G}_\ell + \mathcal{H}_\ell$ where $\mathcal{G}_\ell$ is the error term corresponding to the ℓ-step of MedianLS, and $\mathcal{H}_\ell$ is the error term induced by the SmoothQR algorithm in Alg. 6 at step ℓ. To apply Theorem 4, we need to show that this sequence of tensors is $\epsilon/2$-admissible for Noisy Tensor-Column Subspace Iteration. Then, this theorem directly indicates that $\mathcal{V}^\dagger * \mathcal{X}_L \leq \epsilon$ and this would conclude our proof.

(The $\epsilon/2$-admissible requirement). Let $\tau = \frac{\gamma_{rk}}{128}$, $\hat{\mu} = \frac{C}{\tau^2}(20\mu^* + \log n)$, and μ be any number satisfying $\mu \geq \mu^*$. Since $\mu^* = \theta(\gamma_{rk}^{-2} r(\mu^* + \log n))$, it satisfies the requirement of Theorem 4. We prove that with probability $1 - 1/n^2$, the following three claims hold:

- $\{(\mathcal{X}_{\ell-1}, \mathcal{G}_\ell)\}_{\ell=1}^L$ is $\epsilon/4$-admissible,
- $\{(\mathcal{X}_{\ell-1}, \mathcal{H}_\ell)\}_{\ell=1}^L$ is $\epsilon/4$-admissible,
- for $\ell \in [L]$, we have $\mu(\mathcal{X}_\ell) \leq \mu$.

If the above three claims hold, then it implies that $\widetilde{\mathcal{G}}_\ell$ is $\epsilon/2$-admissible, using a triangle inequality as $\widetilde{\mathcal{G}}_\ell = \mathcal{G}_\ell + \mathcal{H}_\ell$.

To prove these three claims, we apply a mutual induction approach. For $\ell = 0$, it only requires to check the third claim which follows from Theorem 5 that $\mathcal{X}_0$ satisfies the incoherence bound. Now let us assume that all three claims hold at step $\ell - 1$, our goal is to argue that with probability $1 - 1/n^2$, all three claims will hold at step ℓ.

The first claim will hold from Lemma 9 using the induction hypothesis of the third claim that $\mu(\mathcal{X}_\ell) \leq \mu$. Specifically, the parameters should be set properly as follows. Let $\delta = \frac{c\gamma_{rk}\overline{\sigma}_{rk}}{||\mathcal{T}||_F}$ for sufficiently small constant $c > 0$. The lemma requires the lower bound $p_0 \geq \frac{k\mu^* \log^2 n}{\delta^2 n}$ where the $\log n$ factor comes from $L = \theta(\gamma_{rk}^{-1} \log(n/\epsilon))$ round (each round has new samples as input). Therefore, Lemma 9 states that with probability $1 - 1/n^3$, the upper bound $||\mathcal{G}_\ell||_F \leq \frac{1}{4}\left(\frac{1}{32}\gamma_{rk}\overline{\sigma}_{rk}||\mathcal{V}^\dagger * \mathcal{X}_{\ell-1}|| + \frac{\epsilon}{32}\gamma_{rk}\overline{\sigma_{rk}}\right)$, satisfying the

$\epsilon/4$-admissible condition.

The remaining two lemmas follow from Theorem 6. We will apply this theorem to $\mathcal{A} * \mathcal{X}_\ell + \mathcal{G}_\ell$ with $\upsilon = \overline{\sigma}_{rk}(||\mathcal{V}^\dagger * \mathcal{X}_{\ell-1}|| + \epsilon)$ and τ as above. (noisy tensor) Since $||\mathcal{N} * \mathcal{X}_{\ell-1}|| \leq \overline{\sigma}_{rk}||\mathcal{V}^\dagger * \mathcal{X}_\ell||$, it holds that $\upsilon \geq \max\{||\mathcal{G}_\ell||, ||\mathcal{N} * \mathcal{X}_{\ell-1}||\}$ as required by Theorem 6. and it also requires a lower bound μ. To satisfy the lower bound, we combing Lemma 24 states that with probability $1 - 1/n^2$, we have $\frac{1}{\upsilon}(\rho(\mathcal{G}) + \rho(\mathcal{N} * \mathcal{X})) \leq 10\mu^*$. The SmoothQR process produces with probability at least $1 - 1/n^4$ a tensor $\mathcal{H}_\ell$ such that $||\mathcal{H}_\ell|| \leq \tau\upsilon \leq \frac{\gamma_{rk}\upsilon}{128} \leq \frac{1}{4}\left(\frac{1}{32}\gamma_{rk}\overline{\sigma}_{rk}||\mathcal{V}^\dagger * \mathcal{X}_{\ell-1}|| + \frac{\epsilon}{32}\gamma_{rk}\overline{\sigma_{rk}}\right)$, satisfying the $\epsilon/4$-admissible condition. Therefore, the second and third claim hold.

Note that all error probabilities that incurred were less than $1/n$, i.e., $\Omega(1/n)$, then we sum up the error probabilities over all $L \leq n$ steps to conclude the proof.

■

Lemma 24. *Under the assumptions of Theorem 7, we have for every $\ell \in [L]$ and $\upsilon = \frac{\overline{\sigma}_{rk}}{32}\left(||\mathcal{V}^\dagger * \mathcal{X}_{\ell-1}|| + \epsilon\right)$ with probability $1 - 1/n^2$, $\frac{1}{\upsilon}(\rho(\mathcal{G} + \rho(\mathcal{N} * \mathcal{X}_\ell))) \leq \mu^*$.*

Proof: Given the lower bound on p in Theorem 7, we apply Lemma 9 to conclude that $||\dot{e}_i^\dagger * \mathcal{G}_\ell^{\mathcal{M}}|| \leq \sqrt{r\mu(\mathcal{U})/n}\dot{\upsilon}$ and $||\dot{e}_i^\dagger * \mathcal{G}_\ell^{\mathcal{N}}|| \leq \sqrt{\mu^*/n}\dot{\upsilon}$. Thus, $\rho(\mathcal{G}_\ell)/\upsilon^2 \leq \mu^*$.

Further, we claim that $||\dot{e}_i^\dagger * \mathcal{N} * \mathcal{X}||^2 \leq (\mu^*/n)\overline{\sigma}_{rk}||\mathcal{V} * \mathcal{U}||$ for all $i \in [n]$, since

$$||\dot{e}_i^\dagger * \mathcal{N} * \mathcal{X}|| \leq ||\dot{e}_i^\dagger * \mathcal{V} * \Theta_{\mathcal{V}}||\dot{||}\mathcal{V}^\dagger * \mathcal{X}_\ell|| = ||\dot{e}_i^\dagger \mathcal{N}||\dot{||}\mathcal{V}^\dagger * \mathcal{X}_\ell||, \tag{82}$$

where we used the following fact that

$$||\dot{e}_i^\dagger \mathcal{N}||^2 = ||\dot{e}_i^\dagger \mathcal{N} * \mathcal{V}||^2 + ||\dot{e}_i^\dagger \mathcal{N} * \mathcal{U}||^2 = ||\dot{e}_i^\dagger \mathcal{N} * \mathcal{V}||^2 = ||\dot{e}_i^\dagger * \mathcal{V} * \Theta_{\mathcal{V}}||\dot{||}\mathcal{V}^\dagger * \mathcal{X}_\ell||^2. \tag{83}$$

Combining (34), this shows that $\rho(\mathcal{N} * \mathcal{X}_{\ell-1})/\upsilon^2 \leq \mu^*$, and conclude the proof.

■

REFERENCES

[1] R. Baraniuk, "More is less: Signal processing and the data deluge," *Science*, 331(6018), pp. 717-719, 2011.

[2] A. Cichocki, C. Mandic, AH. Phan, C. Caiafa, G. Zhou, Q. Zhao, and L. Lathauwer, "Tensor decompositions for signal processing applications: From two-way to multiway component analysis," *IEEE Signal Processing Magazine*, 32(2), pp. 145-163, 2015.

[3] E.J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on Information Theory*, 52(2): 489-509, 2006.

[4] R. Little and D.B. Rubin, "Statistical analysis with missing data," *John Wiley & Sons*, 2014.

[5] L. Kong, L. He, X.-Y. Liu, Y. Gu, M.-Y. Wu, and X. Liu, "Privacy-preserving compressive sensing for crowdsensing based trajectory recovery," *IEEE ICDCS*, 2015.

[6] X.-Y. Liu, S. Aeron, V. Aggarwal, X. Wang, and M. Wu, "Adaptive sampling of RF fingerprints for fine-grained indoor localization," *IEEE Transactions on Mobile Computing*, 2015

[7] Y. Zhang, M. Roughan, W. Willinger, and L. Qiu, "Spatio-temporal compressive sensing and internet traffic matrices," *ACM SIGCOMM*, 39(4): 267-278, 2009.

[8] S. Rallapalli, L. Qiu, Y. Zhang, Y.-C. Chen, "Exploiting temporal stability and low-rank structure for localization in mobile networks," *ACM MobiCom*, 2010.

[9] Y.-C. Chen, L. Qiu, Y. Zhang, Z. Hu, and G. Xue, "Robust network compressive sensing," *ACM MobiCom*, 2014.

[10] L. Kong, M. Xia, X.-Y. Liu, X. Liu, and M.-Y. Wu, "Data loss and reconstruction in sensor networks," *IEEE INFOCOM*, pp. 1654-1662, 2013.

[11] X.-Y. Liu, Y. Zhu, L. Kong, C. Liu, Y. Gu, and M.-Y. Wu, "CDC: Compressive data collection for wireless sensor networks," *IEEE Transactions on Parallel and Distributed Systems*, pp. 2188-2197, 2015.

[12] M. Signoretto, Q.T. Dinh, L.D. Lathauwer, and J. Suykens, "Learning with tensors: a framework based on convex optimization and spectral regularization," *Springer Machine Learning*, 94(3): 303-351, 2014.

[13] P. Chen and D. Suter, "Recovering the missing components in a large noisy low-rank matrix: Application to SFM," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 26(8): 1051-1063, 2004.

[14] J. Liu, P. Musialski, P. Wonka, and J. Ye, "Tensor completion for estimating missing values in visual data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1): 208-220, 2013.

[15] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *IEEE Computer*, vol. 42, no. 8, 2009.

[16] Z. Liu and L. Vandenberghe, "Interior-point method for nuclear norm approximation with application to system identification," *SIAM Journal on Matrix Analysis and Applications*, 31(3): 1235-1256, 2009.

[17] E. Candès and T. Tao, "Near optimal signal recovery from random projections: Universal encoding strategies?" *IEEE Trans. on Information Theory*, 52(12), pp. 5406-5425, 2006.

[18] E. Candès and B. Recht, "Exact matrix completion via convex optimization," *Springer Foundations of Computational Mathematics*, 9(6), pp. 717-772, 2009.

[19] E. Candès and T. Tao, "The power of convex relaxation: Near-optimal matrix completion," *IEEE Transactions on Information Theory*, 56(5), pp. 2053-2080, 2010.

[20] T.G. Kolda and B.W. Bader, "Tensor decompositions and applications," *SIAM review*, 51(3), pp. 455-500, 2009.

[21] M.E. Kilmer and C.D. Martin, "Factorization strategies for third-order tensors," *Linear Algebra and its Applications*, 435(3), pp. 641-658, 2011.

[22] M.E. Kilmer, K. Braman, N. Hao, and R.C. Hoover, "Third-order tensors as operators on matrices: A

theoretical and computational framework with applications in imaging," *SIAM Journal on Matrix Analysis and Applications*, 34(1), pp. 148-172, 2013.

[23] M Fazel, "Matrix rank minimization with applications," *Diss. PhD thesis, Stanford University*, 2002.

[24] Z. Zhang and S. Aeron, "Exact tensor completion using t-SVD," *arXiv: http://arxiv.org/abs/1502.04689*, Mar., 2015.

[25] Z. Zhang, G. Ely, S. Aeron, N. Hao, and M. Kilmer, "Novel methods for multilinear data completion and de-noising based on tensor-SVD," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3842-3849, 2014.

[26] H. Zou, T. Hastie, and R. Tibshirani, "Sparse principal component analysis," *JCGS*, 15(2), pp. 262-286, 2006.

[27] P. Jain, P. Netrapalli, and S. Sanghavi, "Low-rank matrix completion using alternating minimization," *ACM Symposium on Theory of Computing (STOC)*, pp. 665-674, 2013.

[28] M. Hardt and Mary Wootters, "Fast matrix completion without the condition number," *Springer, COLT*, 2014.

[29] M. Hardt. "Understanding alternating minimization for matrix completion," *IEEE 55th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 651-660, 2014.

[30] E. Kernfeld, S. Aeron, and M. Kilmer, "Clustering multi-way data: A novel algebraic approach," 2014.

[31] G. Ely, S. Aeron, N. Hao, and M.E. Kilmer, "5D seismic data completion and denoising using a novel class of tensor decompositions, *Geophysics*, 2015.

[32] D.F. Gleich, C. Greif, and J.M. Varah, "The power and Arnoldi methods in an algebra of circulants," *Numerical Linear Algebra with Applications*, 20(5): 809-831, 2013.

[33] M. Hardt and E. Price, "The noisy power method: A meta algorithm with applications," *Advances in Neural Information Processing Systems (NIPS)*, pp. 2861-2869, 2014.

[34] K. Braman, "Third-order tensors as linear operators on a space of matrices," *Linear Algebra and its Applications*, 433(7): 1241-1253, 2010.

[35] R. Sun and Z.Q. Luo, "Guaranteed matrix completion via non-convex factorization," *IEEE 56th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 270-289, 2015.

[36] V. Aggarwal, H. Ma, A. Mahimkar, W. Willinger, Z. Zhang, and S. Aeron, "FlashFill: inferring fine-grained service quality in cellular networks," submitted.

[37] P. Jain and S. Oh, "Provable tensor factorization with missing data," in Proc. *ACM NIPS,* 2014

[38] H. Kasai and B. Mishra, "Riemannian preconditioning for tensor completion," *arXiv preprint*, arXiv:1506.02159, 2015.

[39] C. Davis and W.M. Kahan, "The rotation of eigenvectors by a perturbation," *SIAM J. Numer. Anal.*, 7:1-46, 1970.

[40] J. Matousek, "Lectures on discrete geometry," *Springer-Verlag New York, Inc.*, 2001.