Online and Differentially-Private Tensor Decomposition

Yining Wang[1] and Animashree Anandkumar[2]

[1]Machine Learning Department, Carnegie Mellon University
[2]Department of EECS, University of California, Irvine

June 20, 2016

Abstract

Tensor decomposition is positioned to be a pervasive tool in the era of big data. In this paper, we resolve many of the key algorithmic questions regarding robustness, memory efficiency, and differential privacy of tensor decomposition. We propose simple variants of the tensor power method which enjoy these strong properties. We present the first guarantees for online tensor power method which has a linear memory requirement. Moreover, we present a noise calibrated tensor power method with efficient privacy guarantees. At the heart of all these guarantees lies a careful perturbation analysis derived in this paper which improves up on the existing results significantly.

Keywords: Tensor decomposition, tensor power method, online methods, streaming, differential privacy, perturbation analysis.

1 Introduction

In recent years, tensor decomposition has emerged as a powerful tool to solve many challenging problems in unsupervised [1], supervised [18] and reinforcement learning [4]. Tensors are higher order extensions of matrices which can reveal far greater information compared to matrices, while retaining most of the efficiencies of matrix operations [1].

A central task in tensor analysis is the process of decomposing the tensor into its rank-1 components, which is usually referred to as *CP (Candecomp/Parafac) decomposition* in the literature. While decomposition of arbitrary tensors is NP-hard [13], it becomes tractable for the class of tensors with linearly independent components. Through a simple *whitening* procedure, such tensors can be converted to orthogonally decomposable tensors. Tensor power method is a popular method for computing the decomposition of an orthogonal tensor. It is simple and efficient to implement, and a natural extension of the matrix power method.

In the absence of noise, the tensor power method correctly recovers the components under a random initialization followed by deflation. On the other hand, perturbation analysis of tensor power method is much more delicate compared to the matrix case. This is because the problem of tensor decomposition is NP-hard, and if a large amount of arbitrary noise is added to an orthogonal tensor, the decomposition can again become intractable. In [1], guaranteed recovery of components was proven under bounded noise, and the bound was improved in [2]. In this paper, we significantly improve upon the noise requirements, i.e. the extent of noise that can be withstood by the tensor power method.

In order for tensor methods to be deployed in large-scale systems, we require fast, parallelizable and scalable algorithms. To achieve this, we need to avoid the exponential increase in computation and memory requirements with the order of the tensor; i.e. a naive implementation on a 3rd-order d-dimensional tensor would require $O(d^3)$ computation and memory. Instead, we analyze the online tensor power method that requires only linear (in d) memory and does not form the entire tensor. This is achieved in settings, where the tensor is an empirical higher order moment, computed from the stream of data samples. We can avoid explicit construction of the tensor by running online tensor power method directly on i.i.d. data samples. We show that this algorithm correctly recovers tensor components in time[1] $\tilde{O}(nk^2d)$ and $\tilde{O}(dk)$ memory for a rank-k tensor and n number of data samples. Additionally, we provide efficient sample complexity analysis.

As spectral methods become increasingly popular with recommendation system and health analytics applications [29, 17], data privacy is particularly relevant in the context of preserving sensitive private information. We propose the first differentially private tensor decomposition algorithm with both privacy and utility guarantees via noise calibrated power iterations. We show that under the natural assumption of tensor incoherence, privacy parameters have no (polynomial) dependence on tensor dimension d. On the other hand, straightforward input perturbation type methods lead to far worse bounds and do not yield guaranteed recovery for all values of privacy parameters.

1.1 Related work

Online tensor SGD Stochastic gradient descent (SGD) is an intuitive approach for online tensor decomposition and has been successful in practical large-scale tensor decomposition problems [16]. Despite its simplicity, theoretical properties are particularly hard to establish. [11] considered a variant of the SGD objective and proved its correctness. However, the approach in [11] only works for even-order tensors and its sample complexity dependency upon tensor dimension d is poor.

Tensor PCA In the *statistical tensor PCA* [24] model a $d \times d \times d$ tensor $\mathbf{T} = \boldsymbol{v}^{\otimes 3} + \mathbf{E}$ is observed and one wishes to recover component $\boldsymbol{v}$ under the presence of Gaussian random noise $\mathbf{E}$. [24] shows that $\|\mathbf{E}\|_{\mathrm{op}} = O(d^{-1/2})$ is sufficient to guarantee approximate recovery of $\boldsymbol{v}$ and [14] further improves the noise condition to $\|\mathbf{E}\|_{\mathrm{op}} = O(d^{-1/4})$ via a 4th-order sum-of-squares relaxation. Techniques in both [24, 14] are rather complicated and could be difficult to adapt to memory or privacy constraints. Furthermore, in [24, 14] only one component is considered. On the other hand, [25] shows that $\|\mathbf{E}\|_{\mathrm{op}} = O(d^{-1/2})$ is sufficient for recovering multiple components from noisy tensors. However, [25] assumes exact computation of rank-1 tensor approximation, which is NP-hard in general.

Noisy matrix power methods Our relaxed noise condition analysis for tensor power method is inspired by recent analysis of noisy matrix power methods [12, 6]. Unlike the matrix case, tensor decomposition no longer requires *spectral gap* among eigenvalues and eigenvectors are usually recovered one at a time [1, 2]. This poses new challenges and requires non-trivial extensions of matrix power method analysis to the tensor case.

1.2 Notation and Preliminaries

We use $[n]$ to denote the set $\{1, 2, \cdots, n\}$. We use bold characters $\mathbf{A}, \mathbf{T}, \boldsymbol{v}$ for matrices, tensors, vectors and normal characters λ, μ for scalars. A pth order tensor $\mathbf{T}$ of dimensions $d_1, \cdots, d_p$ has $d_1 \times \cdots \times d_p$ elements, each indexed by a p-tuple $(i_1, \cdots, i_p) \in [d_1] \times \cdots \times [d_p]$. A tensor $\mathbf{T}$ of dimensions $d \times \cdots \times d$ is *super-symmetric* or simply *symmetric* if $\mathbf{T}_{i_1,\cdots,i_p} = \mathbf{T}_{\sigma(i_1),\cdots,\sigma(i_p)}$ for all permutations $\sigma : [p] \to [p]$. For a tensor $\mathbf{T} \in \mathbb{R}^{d_1 \times \cdots \times d_p}$ and matrices $\mathbf{A}_1 \in \mathbb{R}^{m_1 \times d_1}, \cdots, \mathbf{A}_p \in \mathbb{R}^{m_p \times d_p}$, the *multi-linear form* $\mathbf{T}(\mathbf{A}_1, \cdots, \mathbf{A}_p)$ is a $m_1 \times \cdots \times m_p$ tensor defined as

$$[\mathbf{T}(\mathbf{A}_1, \cdots, \mathbf{A}_p)]_{i_1,\cdots,i_p} = \sum_{j_1 \in [d_1]} \cdots \sum_{j_p \in [d_p]} \mathbf{T}_{j_1,\cdots,j_p} [\mathbf{A}_1]_{j_1,i_1} \cdots [\mathbf{A}_p]_{j_p,i_p}.$$

We use $\|\boldsymbol{v}\|_2 = \sqrt{\sum_i \boldsymbol{v}_i^2}$ for vector 2-norm and $\|\boldsymbol{v}\|_\infty = \max_i |\boldsymbol{v}_i|$ for vector infinity norm. We use $\|\mathbf{T}\|_{\mathrm{op}}$ to denote the *operator norm* or *spectral norm* of a tensor $\mathbf{T}$, which is defined as $\|\mathbf{T}\|_{\mathrm{op}} = \sup_{\|\boldsymbol{u}_1\|_2 = \cdots \|\boldsymbol{u}_p\|_2 = 1} \mathbf{T}(\boldsymbol{u}_1, \cdots, \boldsymbol{u}_p)$. An event $\mathcal{A}$ is said to occur *with overwhelming probability* if $\Pr[\mathcal{A}] \geq 1 - d^{-10}$.

We limit ourselves to symmetric 3rd-order tensors ($p = 3$) in this paper. The results can be directly extended to asymmetric tensors since they can first be symmetrized using simple matrix operations (see [1]). Extension to higher-order tensors is also straightforward. A symmetric 3rd-order tensor $\mathbf{T}$ is rank-1 if it can be written in the form of

$$\mathbf{T} = \lambda \cdot \boldsymbol{v} \otimes \boldsymbol{v} \otimes \boldsymbol{v} = \lambda \boldsymbol{v}^{\otimes 3} \quad \Longleftrightarrow \quad \mathbf{T}_{i,j,\ell} = \lambda \cdot \boldsymbol{v}(i) \cdot \boldsymbol{v}(j) \cdot \boldsymbol{v}(\ell), \tag{1}$$

[1] $\tilde{O}$ hides poly-logarithmic factors.

Algorithm 1 Robust tensor power method [1]

```
1: Input: symmetric $d \times d \times d$ tensor $\widetilde{\mathbf{T}}$, number of components $k \leq d$, number of iterations $L$, $R$.
2: for $i = 1$ to $k$ do
3:     Initialization: Draw $\boldsymbol{u}_0$ uniformly at random from the unit sphere in $\mathbb{R}^d$.
4:     Power iteration: Compute $\boldsymbol{u}_t = \widetilde{\mathbf{T}}(\mathbf{I}, \boldsymbol{u}_{t-1}, \boldsymbol{u}_{t-1}) / \|\widetilde{\mathbf{T}}(\mathbf{I}, \boldsymbol{u}_{t-1}, \boldsymbol{u}_{t-1})\|_2$ for $t = 1, \cdots, R$.
5:     Boosting: Repeat Steps 3 and 4 for $L$ times and obtain $\boldsymbol{u}_R^{(1)}, \cdots, \boldsymbol{u}_R^{(L)}$. Let $\tau^* = \text{argmax}_{\tau=1}^{L} \widetilde{\mathbf{T}}(\boldsymbol{u}_R^{(\tau)}, \boldsymbol{u}_R^{(\tau)}, \boldsymbol{u}_R^{(\tau)})$. Set $\hat{\boldsymbol{v}}_i = \boldsymbol{u}_R^{(\tau)}$ and $\hat{\lambda}_i = \widetilde{\mathbf{T}}(\boldsymbol{u}_R^{(\tau)}, \boldsymbol{u}_R^{(\tau)}, \boldsymbol{u}_R^{(\tau)})$.
6:     Deflation: $\widetilde{\mathbf{T}} \leftarrow \widetilde{\mathbf{T}} - \hat{\lambda}_i \hat{\boldsymbol{v}}_i^{\otimes 3}$.
7: end for
8: Output: Estimated eigenvalue/Eigenvector pairs $\{\hat{\lambda}_i, \hat{\boldsymbol{v}}_i\}_{i=1}^k$.
```

where $\otimes$ represents the *outer product*, and $\boldsymbol{v} \in \mathbb{R}^d$ is a unit vector (i.e., $\|\boldsymbol{v}\|_2 = 1$) and $\lambda \in \mathbb{R}^+$. [2] A tensor $\mathbf{T} \in \mathbb{R}^{d \times d \times d}$ is said to have a CP (Candecomp/Parafac) *rank* k if it can be (minimally) written as the sum of k rank-1 tensors:

$$\mathbf{T} = \sum_{i \in [k]} \lambda_i \boldsymbol{v}_i \otimes \boldsymbol{v}_i \otimes \boldsymbol{v}_i, \quad \lambda_i \in \mathbb{R}^+, \;\; \boldsymbol{v}_i \in \mathbb{R}^d. \tag{2}$$

A tensor is said to be orthogonally decomposable if in the above decomposition $\langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle = 0$ for $i \neq j$. Any tensor can be converted to an orthogonal tensor through an invertible whitening transform, provided that $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_k$ are linearly independent [1]. We thus limit our analysis to orthogonal tensors in this paper since it can be extended to this more general class in a straightforward manner.

Tensor Power Method: A popular algorithm for finding the tensor decomposition in (2) is through the tensor power method. The full algorithm is given in Algorithm 1. We propose simple variants of the power method in this paper to enable memory efficient streaming tensor decomposition as well as differentially private decomposition. Moreover, we provide an improved noise analysis for the "vanilla" robust power method in Algorithm 1 in Sec. 4.

2 Memory-Efficient Streaming Tensor Decomposition

Tensor power method in Algorithm 1 requires significant storage to be deployed: $\Omega(d^3)$ memory is required to store a dense $d \times d \times d$ tensor, which is prohibitively large in many real-world applications as tensor dimension d could be really high. We show in this section how to compute tensor decomposition in a memory efficient manner, with storage scaling *linearly* in d. In particular, we consider the case when tensor $\mathbf{T}$ to be decomposed is a *population moment* $\mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}}[\boldsymbol{x}^{\otimes 3}]$ with respect to some unknown underlying data distribution $\mathcal{D}$, and data points $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots$ i.i.d. sampled from $\mathcal{D}$ are fed into a tensor decomposition algorithm in a streaming fashion. One classical example is topic modeling, where each $\boldsymbol{x}_i$ represents documents that come in streams and consistent estimation of topics can be achieved by decomposing variants of the population moment [1, 3].

Algorithm 2 displays memory-efficient tensor decomposition procedure on streaming data points. The main idea is to replace the power iteration step $\mathbf{T}(\mathbf{I}, \boldsymbol{u}, \boldsymbol{u})$ in Algorithm 1 with a "data association" step that exploits the empirical-moment structure of the tensor $\mathbf{T}$ to be decomposed and evaluates approximate power iterations from stochastic data samples. This procedure is highly efficient, in that both time and space complexity scale linearly with tensor dimension d:

Proposition 2.1. *Algorithm 2 runs in $O(nkdLR)$ time and $O(d(k+L))$ memory, with $O(nkR)$ sample complexity (number of data point gone through).*

In the remainder of this section we show Algorithm 2 recovers eigenvectors of the population moment $\mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}}[\boldsymbol{x}^{\otimes 3}]$ with high probability and we derive corresponding sample complexity bounds. To facilitate our theoretical analysis we need several assumptions on the data distribution $\mathcal{D}$. The first natural assumption is the low-rankness of the population moment $\mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}}[\boldsymbol{x}^{\otimes 3}]$ to be decomposed:

[2] One can always assume without loss of generality that $\lambda \geq 0$ by replacing $\boldsymbol{v}$ with $-\boldsymbol{v}$ instead.

Algorithm 2 Online robust tensor power method

```
Input: data stream $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots \in \mathbb{R}^d$, no. of components $k$, parameters $L, R, n$.
for $i = 1$ to $k$ do
    Draw $\boldsymbol{u}_0^{(1)}, \cdots, \boldsymbol{u}_0^{(L)}$ i.i.d. uniformly at random from the unit sphere $\mathcal{S}^{d-1}$.
    for $t = 0$ to $R - 1$ do
        Initialization: Set accumulators $\tilde{\boldsymbol{u}}_{t+1}^{(1)}, \cdots, \tilde{\boldsymbol{u}}_{t+1}^{(L)}$ and $\tilde{\lambda}^{(1)}, \cdots, \tilde{\lambda}^{(L)}$ to 0.
        Data association: Read the next $n$ data points; update $\tilde{\boldsymbol{u}}_{t+1}^{(\tau)} \leftarrow \tilde{\boldsymbol{u}}_{t+1}^{(\tau)} + \frac{1}{n}(\boldsymbol{x}_i^\top \boldsymbol{u}_t^{(\tau)})^2 \boldsymbol{x}_i$ and $\tilde{\lambda}^{(\tau)} \leftarrow \tilde{\lambda}^{(\tau)} + \frac{1}{n}(\boldsymbol{x}_i^\top \boldsymbol{u}_t^{(\tau)})^3$ for each $i \in [n]$ and $\tau \in [L]$.
        Deflation: For each $\tau \in [L]$, update $\tilde{\boldsymbol{u}}_{t+1}^{(\tau)} \leftarrow \tilde{\boldsymbol{u}}_{t+1}^{(\tau)} - \sum_{j=1}^{i-1} \hat{\lambda}_j \xi_{j,\tau}^2 \boldsymbol{v}_j$ and $\tilde{\lambda}^{(\tau)} \leftarrow \tilde{\lambda}^{(\tau)} - \sum_{j=1}^{i-1} \hat{\lambda}_j \xi_{j,\tau}^3$, where $\xi_{j,\tau} = \hat{\boldsymbol{v}}_j^\top \boldsymbol{u}_t^{(\tau)}$.
        Normalization: $\boldsymbol{u}_{t+1}^{(\tau)} = \tilde{\boldsymbol{u}}_{t+1}^{(\tau)} / \|\tilde{\boldsymbol{u}}_{t+1}^{(\tau)}\|_2$, for each $\tau \in [L]$.
    end for
    Find $\tau^* = \mathrm{argmax}_{\tau \in [L]} \tilde{\lambda}^{(\tau)}$ and store $\hat{\lambda}_i = \tilde{\lambda}^{(\tau^*)}$, $\hat{\boldsymbol{v}}_i = \boldsymbol{u}_R^{(\tau^*)}$.
end for
Output: approximate eigenvalue and eigenvector pairs $\{\hat{\lambda}_i, \hat{\boldsymbol{v}}_i\}_{i=1}^k$ of $\hat{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}}[\boldsymbol{x}^{\otimes 3}]$.
```

Assumption 2.1 (Low-rank moment)**.** *The mean tensor* $\mathbf{T} = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}[\boldsymbol{x}^{\otimes 3}]$ *admits a low-rank representation* $\mathbf{T} = \sum_{i=1}^k \lambda_i \boldsymbol{v}_i^{\otimes 3}$ *for* $\lambda_1, \cdots, \lambda_k > 0$ *and orthonormal* $\{\boldsymbol{v}_1, \cdots, \boldsymbol{v}_k\} \subseteq \mathbb{R}^d$.

We also place restrictions on the "noise model", which imply that the population moment $\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}[\boldsymbol{x}^{\otimes 3}]$ can be well approximated by a reasonable number of samples with high probability. In particular, we consider sub-Gaussian noise as formulated in Definition 2.1 and Assumption 2.2:

Definition 2.1 (Multivariate sub-Gaussian distribution, [15])**.** *A D-dimensional random variable* $\boldsymbol{x}$ *belongs to the sub-Gaussian distribution family* $\mathcal{SG}_D(\sigma)$ *with parameter* $\sigma > 0$ *if it has zero mean and* $\mathbb{E}\left[\exp(\boldsymbol{a}^\top \boldsymbol{x})\right] \leq \exp\left\{\|\boldsymbol{a}\|_2^2 \sigma^2/2\right\}$ *for all* $\boldsymbol{a} \in \mathbb{R}^D$.

Assumption 2.2 (Sub-Gaussian noise)**.** *There exists* $\sigma > 0$ *such that the mean-centered vectorized random variable* $\mathrm{vec}(\boldsymbol{x}^{\otimes 3} - \mathbb{E}[\boldsymbol{x}^{\otimes 3}])$ *belongs to* $\mathcal{SG}_{d^3}(\sigma)$ *as defined in Definition 2.1.*

We remark that Assumption 2.2 includes a wide family of distributions that are of practical importance, for example noise that have compact support. Assumption 2.2 also resembles (B,p)*-round noise* considered in [12] that imposes spherical symmetry constraints onto the noise distribution.

We are now ready to present the main theorem that bounds the recovery (approximation) error of eigenvalues and eigenvectors of the streaming robust tensor power method in Algorithm 2:

Theorem 2.1 (Analysis of streaming robust tensor power method)**.** *Let Assumptions 2.1, 2.2 hold true and suppose* $\epsilon < C_1 \lambda_{\min}/\sqrt{k}$ *for some sufficiently small absolute constant* $C_1 > 0$*. If*

$$n = \widetilde{\Omega}\left(\min\left\{\frac{\sigma^2 d}{\epsilon^2}, \frac{\sigma^2 d^2}{\lambda_{\min}^2}\right\}\right), \quad R = \Omega(\log(\lambda_{\max} d/\epsilon)), \quad L = \Omega(k \log k),$$

then with probability at least 0.9 there exists permutation $\pi : [k] \to [k]$ *such that*

$$|\lambda_i - \hat{\lambda}_{\pi(i)}| \leq C_2 \epsilon, \quad \|\boldsymbol{v}_i - \hat{\boldsymbol{v}}_{\pi(i)}\|_2 \leq C_3 \epsilon/\lambda_i, \quad \forall i = 1, \cdots, k$$

for some universal constants $C_2, C_3 > 0$.

Corollary 2.1 is then an immediate consequence of Theorem 2.1, which simplifies the bounds and highlights asymptotic dependencies over important model parameters d, k and σ:

Corollary 2.1. *Under Assumptions 2.1, 2.2, Algorithm 2 correctly learns* $\{\lambda_i, \boldsymbol{v}_i\}_{i=1}^k$ *up to* $O(1/\sqrt{d})$ *additive error with* $\tilde{O}(\sigma^2 k d^2)$ *samples and* $\tilde{O}(dk)$ *memory.*

Algorithm 3 Differentially private robust tensor power method

```
Input: tensor $\mathbf{T}$, no. of components $k$, number of iterations $L, R$, privacy parameters $\varepsilon, \delta$.
Initialization: $\mathbf{D} = \mathbf{0}$, $\nu = \frac{6\sqrt{2\ln(1.25/\delta')}}{\varepsilon'}$, $\delta' = \frac{\delta}{2K}$, $\varepsilon' = \frac{\varepsilon}{\sqrt{K(4+\ln(2/\delta))}}$, $K = kL(R+1)$.
for $i = 1$ to $k$ do
    Initialization: Draw $\boldsymbol{u}_0^{(1)}, \cdots, \boldsymbol{u}_0^{(\tau)}$ uniformly at random from the unit sphere in $\mathbb{R}^d$.
    for $t = 0$ to $R - 1$ do
        Power iteration: compute $\tilde{\boldsymbol{u}}_{t+1}^{(\tau)} = (\mathbf{T} - \mathbf{D})(\mathbf{I}, \boldsymbol{u}_t^{(\tau)}, \boldsymbol{u}_t^{(\tau)})$.
        Noise calibration: release $\bar{\boldsymbol{u}}_{t+1}^{(\tau)} = \tilde{\boldsymbol{u}}_{t+1}^{(\tau)} + \nu\|\boldsymbol{u}_t^{(\tau)}\|_\infty^2 \cdot \boldsymbol{z}_t^{(\tau)}$, where $\boldsymbol{z}_t^{(\tau)} \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$.
        Normalization: $\boldsymbol{u}_{t+1}^{(\tau)} = \bar{\boldsymbol{u}}_{t+1}^{(\tau)} / \|\bar{\boldsymbol{u}}_{t+1}^{(\tau)}\|_2$.
    end for
    Compute $\tilde{\lambda}^{(\tau)} = (\mathbf{T} - \mathbf{D})(\boldsymbol{u}_R^{(\tau)}, \boldsymbol{u}_R^{(\tau)}, \boldsymbol{u}_R^{(\tau)}) + \nu\|\boldsymbol{u}_R^{(\tau)}\|_\infty^3 \cdot z_\tau$ and let $\tau^* = \mathrm{argmax}_\tau \tilde{\lambda}^{(\tau)}$.
    Deflation: $\hat{\lambda}_i = \tilde{\lambda}^{(\tau^*)}$, $\hat{\boldsymbol{v}}_i = \boldsymbol{u}_R^{(\tau^*)}$, $\mathbf{D} \leftarrow \mathbf{D} + \hat{\lambda}_i \hat{\boldsymbol{v}}_i^{\otimes 3}$.
end for
Output: eigenvalue/eigenvector pairs $\{\hat{\lambda}_i, \hat{\boldsymbol{v}}_i\}_{i=1}^k$.
```

Proofs of Theorem 2.1 and Corollary 2.1 are both deferred to Appendix C. Compared to streaming noisy matrix PCA considered in [12], the bound is weaker with an additional $1/k$ factor in the term involving ϵ and $1/d$ factor in the term that does not involve ϵ. We conjecture this to be a fundamental difficulty of the tensor decomposition problem, as we show in Sec. 5 that our noise conditions cannot generally be improved. On the other hand, our bounds resulting from the analysis in Sec. 4 have a $O(1/d)$ improvement compared to applying existing analysis in [1] directly (cf. Sec. 4.1).

Remark on comparison with SGD: Our proposed streaming tensor power method is nothing but the projected stochastic gradient descent (SGD) procedure on the objective of maximizing the tensor norm on the sphere. The optimal solution of this coincides with the objective of finding the best rank-1 approximation of the tensor. Here, we can estimate all the components of the tensor through deflation. An alternative method is to run SGD based a combined objective function to obtain all the components of the tensor simultaneously, as considered in [16, 11]. However, the analysis in [11] only works for even-order tensors and has worse dependency (at least d^9) on tensor dimension d.

3 Differentially private tensor decomposition

The objective of private data processing is to release data summaries such that any particular entry of the original data cannot be reliably inferred from the released results. Formally speaking, we adopt the popular (ε, δ)-differential privacy criterion proposed in [9]:

Definition 3.1 ((ε, δ)-differential privacy [9])**.** *Let $\mathcal{M}$ denote all symmetric d-dimensional real third order tensors and $\mathcal{O}$ be an arbitrary output set. A randomized algorithm $A : \mathcal{M} \to \mathcal{O}$ is* (ε, δ)-differentially private *if for all neighboring tensors $\mathbf{T}, \mathbf{T}'$ and measurable set $O \subseteq \mathcal{O}$ we have*

$$\Pr[A(\mathbf{T}) \in O] \leq e^\varepsilon \Pr[A(\mathbf{T}') \in O] + \delta,$$

where $\varepsilon > 0$, $\delta \in [0, 1)$ are privacy parameters and probabilities are taken over randomness in A.

Since our tensor decomposition analysis concerns symmetric tensors primarily, we adopt a "symmetric" definition of neighboring tensors in Definition 3.1, as shown below:

Definition 3.2 (Neighboring tensors)**.** *Two $d \times d \times d$ symmetric tensors $\mathbf{T}, \mathbf{T}'$ are* neighboring tensors *if there exists $i, j, k \in [d]$ such that*

$$\mathbf{T}' - \mathbf{T} = \pm\text{symmetrize}(\boldsymbol{e}_i \otimes \boldsymbol{e}_j \otimes \boldsymbol{e}_k) = \pm\left(\boldsymbol{e}_i \otimes \boldsymbol{e}_j \otimes \boldsymbol{e}_k + \boldsymbol{e}_i \otimes \boldsymbol{e}_k \otimes \boldsymbol{e}_j + \cdots + \boldsymbol{e}_k \otimes \boldsymbol{e}_j \otimes \boldsymbol{e}_i\right).$$

As noted earlier, the above notions can be similarly extended to asymmetric tensors as well as the guarantees for tensor power method on asymmetric tensors.

In a nutshell, Definitions 3.1, 3.2 state that an algorithm A is differentially private if, conditioned on any set of possible outputs of A, one cannot distinguish with high probability between two "neighboring" tensors $\mathbf{T}, \mathbf{T}'$ that differ only in a single entry (up to symmetrization), thus protecting the privacy of any particular element in the original tensor $\mathbf{T}$. Here ε, δ are parameters controlling the level of privacy, with smaller ε, δ values implying stronger privacy guarantee as $\Pr[A(\mathbf{T}) \in O]$ and $\Pr[A(\mathbf{T}') \in O]$ are closer to each other.

Algorithm 3 describes the procedure of privately releasing eigenvectors of a low-rank input tensor $\mathbf{T}$. The main idea for privacy preservation is the following *noise calibration* step

$$\bar{\boldsymbol{u}}_{t+1} = \tilde{\boldsymbol{u}}_{t+1} + \nu\|\boldsymbol{u}_t\|_\infty^2 \cdot \boldsymbol{z}_t,$$

where $\boldsymbol{z}_t$ is a d-dimensional standard Normal random variable and $\nu\|\boldsymbol{u}_t\|_\infty^2$ is a carefully designed noise magnitude in order to achieved desired privacy level (ε, δ). One key aspect is that the noise calibration step occurs at *every* power iteration, which adds to the robustness of the algorithm and achieves sharper bounds. We discuss at the end of this section.

Lemma 3.1 shows that Algorithm 3 is (ε, δ)-differentially private:

Lemma 3.1 (Privacy guarantee)**.** *Algorithm 3 satisfies (ε, δ)-differential privacy.*

The proof is via advanced composition of Gaussian mechanism [9] and is deferred to Appendix D.

The rest of the section is devoted to discussing the "utility" of Algorithm 3; i.e., to show that the algorithm is still capable of producing approximate eigenvectors, despite the privacy constraints. Similar to [12], we adopt the following incoherence assumptions on the eigenspace of $\mathbf{T}$:

Assumption 3.1 (Incoherent basis)**.** *Suppose $\mathbf{V} \in \mathbb{R}^{d\times k}$ is the stacked matrix of orthonormal component vectors $\{\boldsymbol{v}_i\}_{i=1}^k$. There exists constant $\mu_0 > 0$ such that*

$$\frac{d}{k} \max_{1\le i\le d} \|\mathbf{V}^\top \boldsymbol{e}_i\|_2^2 \le \mu_0. \tag{3}$$

Note that by definition, μ_0 is always in the range of $[1, n/k]$. Intuitively, Assumption 3.1 with small constant μ_0 implies a relatively "flat" distribution of element magnitudes in $\mathbf{T}$. The incoherence level μ_0 plays an important role in the utility guarantee of Algorithm 3, as we show below:

Theorem 3.1 (Guaranteed recovery of eigenvector under privacy requirements)**.** *Suppose $\mathbf{T} = \sum_{i=1}^k \lambda_i \boldsymbol{v}_i^{\otimes 3}$ for $\lambda_1 > \lambda_2 \ge \lambda_3 \ge \cdots \ge \lambda_k > 0$ with orthonormal $\boldsymbol{v}_1, \cdots, \boldsymbol{v}_k \in \mathbb{R}^d$, and suppose Assumption 3.1 holds with μ_0. Assume $\lambda_1 - \lambda_2 \ge c/\sqrt{d}$ for some sufficiently small universal constant $c > 0$. If $R = \Theta(\log(\lambda_{\max} d))$, $L = \Theta(k \log k)$ and ε, δ satisfy*

$$\varepsilon = \Omega\left(\frac{\mu_0 k^2 \log(\lambda_{\max} d/\delta)}{\lambda_{\min}}\right), \tag{4}$$

then with probability at least 0.9 the first eigen pair $(\hat{\lambda}_1, \hat{\boldsymbol{v}}_1)$ returned by Algorithm 3 satisfies

$$\left|\lambda_1 - \hat{\lambda_1}\right| = O(1/\sqrt{d}), \qquad \|\boldsymbol{v}_1 - \hat{\boldsymbol{v}}_1\|_2 = O(1/(\lambda_1\sqrt{d})).$$

At a high level, Theorem 3.1 states that when the privacy parameter ε is not too small (i.e., privacy requirements are not too stringent), Algorithm 3 approximately recovers the largest eigenvalue and eigenvector with high probability. Furthermore, when μ_0 is a constant, the lower bound condition on the privacy parameter ε does *not* depend polynomially upon tensor dimension d, which is a much desired property for high-dimensional data analysis. On the other hand, similar results cannot be achieved via simpler methods like input perturbation, as we discuss below:

Comparison with input perturbation Input perturbation is perhaps the simplest method for differentially private data analysis and has been successful in numerous scenarios, e.g. private matrix PCA [10]. In our context, this would entail appending a random Gaussian tensor $\mathbf{E}$ directly onto the input tensor $\mathbf{T}$ *before* tensor power iterations. By Gaussian mechanism, the standard deviation σ of each element in $\mathbf{E}$ scales as $\sigma = \Omega(\varepsilon^{-1}\sqrt{\log(1/\delta)})$. On the other hand, noise analysis for tensor decomposition derived in [24, 2] and in the subsequent section of this paper requires $\sigma = O(1/d)$ or $\|\mathbf{E}\|_{\mathrm{op}} = O(1/\sqrt{d})$, which implies $\varepsilon = \tilde{\Omega}(d)$ (cf. Lemma E.9). That is, the privacy parameter ε must scale *linearly* with tensor dimension d to successfully recover even the first principle eigenvector, which renders the privacy guarantee of the input perturbation procedure useless for high-dimensional tensors. Thus, we require a non-trivial new approach for differentially private tensor decomposition.

Finally, we remark that a more desired utility analysis would bound the approximation error $\|\boldsymbol{v}_i - \hat{\boldsymbol{v}}_i\|_2$ for every component $\boldsymbol{v}_1, \cdots, \boldsymbol{v}_k$, and not just the top eigenvector. Unfortunately, our current analysis cannot handle deflation effectively as the deflated vector $\hat{\boldsymbol{v}}_i - \boldsymbol{v}_i$ may not be incoherent. Extension to deflated tensor decomposition remains an interesting open question.

4 Improved Noise Analysis for Tensor Power Method

When the tensor $\mathbf{T}$ has an exact orthogonal decomposition, the power method provably recovers all the components with random initialization and deflation. However, the analysis is more subtle under noise. While matrix perturbation bounds are well understood, it is an open problem in the case of tensors. This is because the problem of tensor decomposition is NP-hard, and becomes tractable only under special conditions such as orthogonality (and more generally linear independence). If a large amount of arbitrary noise is added, the decomposition can again become intractable. In [1], guaranteed recovery of components was proven under bounded noise and we recap the result below.

Theorem 4.1 ([1] Theorem 5.1, simplified version)**.** *Suppose $\widetilde{\mathbf{T}} = \mathbf{T} + \boldsymbol{\Delta}_T$, where $\mathbf{T} = \sum_{i=1}^k \lambda_i \boldsymbol{v}_i^{\otimes 3}$ with $\lambda_i > 0$ and orthonormal basis vectors$\{\boldsymbol{v}_1, \cdots, \boldsymbol{v}_k\} \subseteq \mathbb{R}^d$, $d \geq k$, and noise $\boldsymbol{\Delta}_T$ satisfies $\|\boldsymbol{\Delta}_T\|_{\mathrm{op}} \leq \epsilon$. Let $\lambda_{\max}, \lambda_{\min}$ be the largest and smallest values in $\{\lambda_i\}_{i=1}^k$ and $\{\hat{\lambda}_i, \hat{\boldsymbol{v}}_i\}_{i=1}^k$ be outputs of Algorithm 1. There exist absolute constants $K_0, C_1, C_2, C_3 > 0$ such that if $k \geq K_0$ and*

$$\epsilon \leq C_1 \cdot \lambda_{\min}/d, \quad R = \Omega(\log d + \log\log(\lambda_{\max}/\epsilon)), \quad L = \Omega(k \log k), \tag{5}$$

then with probability at least 0.9, there exists a permutation $\pi : [k] \to [k]$ such that

$$|\lambda_i - \hat{\lambda}_{\pi(i)}| \leq C_2\epsilon, \quad \|\boldsymbol{v}_i - \hat{\boldsymbol{v}}_{\pi(i)}\|_2 \leq C_3\epsilon/\lambda_i, \quad \forall i = 1, \cdots, k.$$

Theorem 4.1 is the first provably correct result on robust tensor decomposition under general noise conditions. In particular, the noise term $\boldsymbol{\Delta}_T$ can be deterministic or even adversarial. However, one important drawback of Theorem 4.1 is that $\|\boldsymbol{\Delta}_T\|_{\mathrm{op}}$ must be upper bounded by $O(\lambda_{\min}/d)$, which is a strong assumption for many practical applications [28]. On the other hand, [2, 24] show that by using smart initializations the robust tensor power method is capable of tolerating $O(\lambda_{\min}/\sqrt{d})$ magnitude of noise, and [25] suggests that such noise magnitude cannot be improved if deflation (i.e., successive rank-one approximation) is to be performed.

In this paper, we show that the relaxed noise bound $O(\lambda_{\min}/\sqrt{d})$ holds even if the initialization of robust TPM is as simple as a vector uniformly sampled from the d-dimensional sphere (Algorithm 1). Our claim is formalized below:

Theorem 4.2 (Improved noise tolerance analysis for robust TPM)**.** *Assume the same notation as in Theorem 4.1. Let $\epsilon \in (0, 1/2)$ be an error tolerance parameter. There exist absolute constants $K_0, C_0, C_1, C_2, C_3 > 0$ such that if $k \geq K_0$ and $\boldsymbol{\Delta}_T$ satisfies*

$$\|\boldsymbol{\Delta}_T(\mathbf{I}, \boldsymbol{u}_t^{(\tau)}, \boldsymbol{u}_t^{(\tau)})\|_2 \leq \epsilon, \quad |\boldsymbol{\Delta}_T(\boldsymbol{v}_i, \boldsymbol{u}_t^{(\tau)}, \boldsymbol{u}_t^{(\tau)})| \leq \min\{\epsilon/\sqrt{k}, C_0\lambda_{\min}/d\} \tag{6}$$

for all $i \in [k]$, $t \in [T], \tau \in [L]$ and furthermore

$$\epsilon \le C_1 \cdot \lambda_{\min}/\sqrt{k}, \quad R = \Omega(\log(\lambda_{\max} d/\epsilon)), \quad L = \Omega(k \log k) \tag{7}$$

then with probability at least 0.9*, there exists a permutation $\pi : [k] \to [k]$ such that*

$$|\lambda_i - \hat{\lambda}_{\pi(i)}| \le C_2 \epsilon, \quad \|\boldsymbol{v}_i - \hat{\boldsymbol{v}}_{\pi(i)}\|_2 \le C_3 \epsilon/\lambda_i, \quad \forall i = 1, \cdots, k.$$

Due to space constraints, proof of Theorem 4.2 is placed in Appendix A. We next make several remarks on our results. In particular, we consider three scenarios with increasing assumptions imposed on the noise tensor $\boldsymbol{\Delta}_T$ and compare the noise conditions in Theorem 4.2 with existing results on orthogonal tensor decomposition:

1. $\boldsymbol{\Delta}_T$ *does not have any special structure*: in this case, we only have $|\boldsymbol{\Delta}_T(\boldsymbol{v}_i, \boldsymbol{u}_t, \boldsymbol{u}_t)| \le \|\boldsymbol{\Delta}_T\|_{\mathrm{op}}$ and our noise conditions reduces to the classical one: $\|\boldsymbol{\Delta}_T\|_{\mathrm{op}} = O(\lambda_{\min}/d)$.
2. $\boldsymbol{\Delta}_T$ *is "round"* in the sense that $|\boldsymbol{\Delta}_T(\boldsymbol{v}_i, \boldsymbol{u}_t, \boldsymbol{u}_t)| \le O(1/\sqrt{d}) \cdot \|\boldsymbol{\Delta}_T(\mathbf{I}, \boldsymbol{u}_t, \boldsymbol{u}_t)\|_2$: this is the typical setting when the noise $\boldsymbol{\Delta}_T$ follows Gaussian or sub-Gaussian distributions, as we explain in Sec. 2 and 3. Our noise condition in this case is $\|\boldsymbol{\Delta}_T\|_{\mathrm{op}} = O(\lambda_{\min}/\sqrt{d})$, strictly improving Theorem 4.1 on robust tensor power method with random initializations and matching the bound for more advanced SVD initialization techniques in [2].
3. $\boldsymbol{\Delta}_T$ *is weakly correlated with signal* in the sense that $\|\boldsymbol{\Delta}_T(\boldsymbol{v}_i, \mathbf{I}, \mathbf{I})\|_2 = O(\lambda_{\min}/d)$ for all $i \le k$: in this case our noise condition reduces to $\|\boldsymbol{\Delta}_T\|_{\mathrm{op}} = O(\lambda_{\min}/\sqrt{k})$, strictly improving over SVD initialization [2] in the "undercomplete" regime $k = o(d)$. Note that the whitening trick [3, 1] does not attain our bound, as we explain in Sec. 4.2.

Finally, we remark that the $\log\log(1/\epsilon)$ quadratic convergence rate in Eq. (5) is worsened to $\log(1/\epsilon)$ linear rate in Eq. (7). We are not sure whether this is an artifact of our analysis, because similar analysis for the matrix noisy power method [12] also reveals a linear convergence rate.

4.1 Implications on memory-efficient and differentially private tensor power method

Our bounds in Theorem 4.2 results in sharper analysis of both memory-efficient and differentially private tensor power methods in Sec. 2, 3. Using the original analysis (Theorem 4.1) for the two applications, the memory-efficient tensor power method would have sample complexity *cubic* in the dimension d and for differentially private tensor decomposition the privacy level ε needs to scale as $\widetilde{\Omega}(\sqrt{d})$ as d increases, which is particularly bad as the quality of privacy protection e^{ε} degrades exponentially with tensor dimension d. On the other hand, our improved noise condition in Theorem 4.2 greatly sharpens the bounds in both applications: for memory efficient decomposition, we now require only quadratic sample complexity and for differentially private decomposition, the privacy level ε has no polynomial dependence on d. This makes our results far more practical for high-dimensional tensor decomposition applications.

4.2 Comparison with whitening and matrix SVD decompositions

Another popular thread of tensor decomposition techniques involve whitening and reducing the problem to a matrix SVD decomposition, which is very effective at reducing the dimensionality of the problem in the $k = o(d)$ undercomplete settings [1, 21, 28]. When *only* the 3rd-order tensor $\mathbf{T}$ is available, one common approach is to randomly "marginalized out" one view of $\widetilde{\mathbf{T}}$:

$$\mathbf{M}(\boldsymbol{\theta}) := \widetilde{\mathbf{T}}(\mathbf{I}, \mathbf{I}, \boldsymbol{\theta}), \qquad \boldsymbol{\theta} \text{ randomly drawn on the unit } d\text{-dimensional sphere;}$$

and then evaluate top-k eigen-decomposition of $\mathbf{M}(\boldsymbol{\theta})$. Let $\mathcal{W} = \mathrm{span}(\boldsymbol{v}_1, \cdots, \boldsymbol{v}_k)$ be the span of the true components of $\mathbf{T}$ and $\hat{\mathcal{W}}$ be the top-k eigenspace of matrix $\mathbf{M}(\boldsymbol{\theta})$ obtained by collapsing one view of $\widetilde{\mathbf{T}}$. We then have the following proposition that bounds the perturbation between $\mathcal{W}$ and $\hat{\mathcal{W}}$:

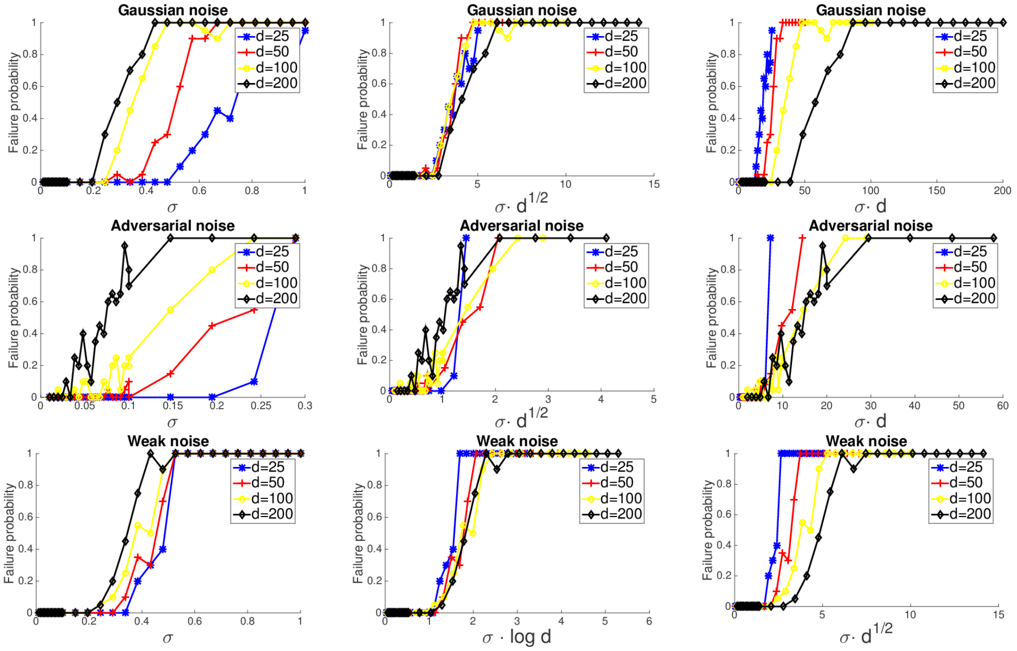

Figure 1: Failure probability against scaled noise magnitude on synthetic tensors. From top to bottom rows: random Gaussian noise, adversarial noise and noise that weakly correlate with the signals (details in main text). From left to right: scaled noise magnitudes: σ, $\sigma\sqrt{d}$, σd and $\sigma\ln(d)$, labeled on each plot below the X axis.

Proposition 4.1. *Suppose $\widetilde{\mathbf{T}} = \mathbf{T} + \boldsymbol{\Delta}_T$ as in Theorems 4.1, 4.2 and let $\boldsymbol{\Pi}_{\mathcal{W}}, \boldsymbol{\Pi}_{\hat{\mathcal{W}}}$ be the projection operators of $\mathcal{W}$ and $\hat{\mathcal{W}}$, respectively. Then with probability at least 0.9 over the random draw of $\boldsymbol{\theta}$,*

$$\left\|\boldsymbol{\Pi}_{\mathcal{W}} - \boldsymbol{\Pi}_{\hat{\mathcal{W}}}\right\|_2 \leq \widetilde{O}\left(\frac{\sqrt{d}\|\boldsymbol{\Delta}_T\|_{\mathrm{op}}}{\lambda_{\min}}\right), \qquad \textit{if} \;\; \|\boldsymbol{\Delta}_T\|_{\mathrm{op}} = \widetilde{O}\left(\frac{\lambda_{\min}}{\sqrt{d}}\right),$$

Proof of Proposition 4.1 is based on matrix SVD perturbation bounds and is deferred to Appendix B. This simple result shows that the whitening trick does not trivially lead to matching noise conditions in Theorem 4.2 under $k = o(d)$ settings.

5 Simulation results

We verify our main theoretical results in Theorem 4.2 on synthetic tensors. $\mathbf{T}$ is taken to be a rank-3 tensor $\mathbf{T} = \boldsymbol{v}_1^{\otimes 3} + 0.75\boldsymbol{v}_2^{\otimes 3} + 0.5\boldsymbol{v}_3^{\otimes 3}$, where $\boldsymbol{v}_1 = (1, 0, 0, 0, \cdots,)$, $\boldsymbol{v}_2 = (0, 1, 0, 0, \cdots)$ and $\boldsymbol{v}_3 = (0, 0, 1, 0, \cdots)$. The noise tensor $\mathbf{E}$ is synthesized according to the following three regimes:

1. **Random Gaussian noise**: First generate $\mathbf{E}_{ijk} \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$ and then super-symmetrize $\mathbf{E}$.

2. **Adversarial Gaussian noise**: $\mathbf{E} = \sum_{i=1}^d \boldsymbol{v}_2 \otimes \boldsymbol{e}_i \otimes \boldsymbol{e}_i + \boldsymbol{e}_i \otimes \boldsymbol{v}_2 \otimes \boldsymbol{e}_i + \boldsymbol{e}_i \otimes \boldsymbol{e}_i \otimes \boldsymbol{v}_2$, where $\boldsymbol{e}_i = (0, \cdots, 0, 1, 0, \cdots, 0)$ has all zero entries except for the ith one.

3. **Weakly correlated noise**: Let $\{\boldsymbol{v}_4, \cdots, \boldsymbol{v}_d\}$ be an orthonormal basis of the orthogonal complement of span$\{\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3\}$. Set $\mathbf{E} = \sum_{i=4}^d \boldsymbol{v}_i \otimes \boldsymbol{v}_i \otimes \boldsymbol{v}_i$.

In Fig. 1 we plot the "failure probability" (measured via 20 independent trials per setting) of the robust tensor power method with random initialization against controlled noise magnitude $\|\mathbf{E}\|_{\mathrm{op}} = \sigma$. A trial is "successful" if for all $i \in \{1, 2, 3\}$ the recovered eigenvector $\hat{\boldsymbol{v}}_i$ satisfies $\hat{\boldsymbol{v}}_i^\top \boldsymbol{v}_i \geq 1/4$. To control $\|\mathbf{E}\|_{\mathrm{op}}$, we first compute the operator norm of the generated raw noise tensor by invoking the `eig_sshopm` routine in Matlab tensor toolbox [5] (algorithm based on [20]) and then re-scale the entries. By inspecting the noise levels at which phase transition of failure probabilities occurs for different tensor dimensions d, ranging from 25 to 200. It is quite clear from Fig. 1 that the phase transitions occur at $\sigma = O(1/\sqrt{d})$ for random Gaussian noise, $\sigma = O(1/d)$ for adversarial noise and $\sigma = O(1/\log d)$ for weakly correlated noises, which matches our theoretical findings in Sec. 4 up to logarithmic terms. Our simulation results and explicit construction of an "adversarial" noise matrix also suggests that our analysis for robust tensor power method with random initializations under random Gaussian noise and existing analysis for worst-case noise in [1] are tight.

6 Conclusion

We consider memory-efficient and differentially private tensor decomposition problems in this paper and derive efficient algorithms for both online and private tensor decomposition based on the popular tensor power method framework. Through an improved noise condition analysis of robust tensor power method, we obtain sharper dimension-dependent sample complexity bounds for online tensor decomposition and wider range of privacy parameters values for private tensor decomposition while still retaining utility. Simulation results verify the tightness of our noise conditions in principle.

One important direction of future research is to extend our online and/or private tensor decomposition algorithms and analysis to practical applications such as topic modeling and community detection, where tensor decomposition acts as one critical step for data analysis. An end-to-end analysis of online/private methods for these applications would be theoretically interesting and could also greatly benefit practical machine learning of important models.

References

[1] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15(1):2773–2832, 2014.

[2] A. Anandkumar, R. Ge, and M. Janzamin. Learning overcomplete latent variable models through tensor methods. In *Proc. of COLT*, 2015.

[3] A. Anandkumar, Y.-k. Liu, D. J. Hsu, D. P. Foster, and S. M. Kakade. A spectral algorithm for latent dirichlet allocation. In *NIPS*, 2012.

[4] K. Azizzadenesheli, A. Lazaric, and A. Anandkumar. Reinforcement learning of POMDP's using spectral methods. In *COLT*, 2016.

[5] B. W. Bader and T. G. Kolda. Algorithm 862: Matlab tensor classes for fast algorithm prototyping. *ACM Transactions on Mathematical Software*, 32(4):635–653, 2006.

[6] M.-F. Balcan, S. Du, Y. Wang, and A. W. Yu. An improved gap-dependency analysis of the noisy power method. In *COLT*, 2016.

[7] L. Birgé. An alternative point of view on lepski's method. *Lecture Notes-Monograph Series*, pages 113–133, 2001.

[8] B. Cirel'soN, I. Ibragimov, and V. Sudakov. Norms of gaussian sample functions. *Lecture Notes in Mathematics*, 550:20–41, 1976.

[9] C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

[10] C. Dwork, K. Talwar, A. Thakurta, and L. Zhang. Analyze gauss: optimal bounds for privacy-preserving principal component analysis. In *STOC*, 2014.

[11] R. Ge, F. Huang, C. Jin, and Y. Yuan. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *COLT*, 2015.

[12] M. Hardt and E. Price. The noisy power method: A meta algorithm with applications. In *NIPS*, 2014.

[13] C. J. Hillar and L.-H. Lim. Most tensor problems are np-hard. *Journal of the ACM (JACM)*, 60(6):45, 2013.

[14] S. B. Hopkins, J. Shi, and D. Steurer. Tensor principal component analysis via sum-of-squares proofs. In *COLT*, 2015.

[15] D. Hsu, S. M. Kakade, and T. Zhang. A tail inequality for quadratic forms of subgaussian random vectors. *Electron. Commun. Probab*, 17(52):1–6, 2012.

[16] F. Huang, U. Niranjan, M. U. Hakeem, and A. Anandkumar. Online tensor methods for learning latent variable models. *Journal of Machine Learning Research*, 16:2797–2835, 2015.

[17] F. Huang, I. Perros, R. Chen, J. Sun, A. Anandkumar, et al. Scalable latent tree model and its application to health analytics. *arXiv preprint arXiv:1406.4566*, 2014.

[18] M. Janzamin, H. Sedghi, and A. Anandkumar. Beating the perils of non-convexity: Guaranteed training of neural networks using tensor methods. *arXiv preprint arXiv:1506.08473*, 2015.

[19] G. Kamath. Bounds on the expectation of the maximum of samples from a gaussian. [Online; accessed April, 2016].

[20] T. G. Kolda and J. R. Mayo. Shifted power method for computing tensor eigenpairs. *SIAM Journal on Matrix Analysis and Applications*, 32(4):1095–1124, 2011.

[21] V. Kuleshov, A. T. Chaganty, and P. Liang. Tensor factorization via matrix factorization. In *AISTATS*, 2015.

[22] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338, 2000.

[23] P. Massart. *Concentration inequalities and model selection*, volume 6. Springer, 2007.

[24] A. Montanari and E. Richard. A statistical model for tensor PCA. In *NIPS*, 2014.

[25] C. Mu, D. Hsu, and D. Goldfarb. Successive rank-one approximations for nearly orthogonally decomposable symmetric tensors. *SIAM Journal on Matrix Analysis and Applications*, 36(4):1638–1659, 2015.

[26] G. W. Stewart, J.-g. Sun, and H. B. Jovanovich. *Matrix perturbation theory*. Academic press New York, 1990.

[27] R. Tomioka and T. Suzuki. Spectral norm of random tensors. *arXiv:1407.1870*, 2014.

[28] Y. Wang, H.-Y. Tung, A. J. Smola, and A. Anandkumar. Fast and guaranteed tensor decomposition via sketching. In *NIPS*, 2015.

[29] Y. Wang and J. Zhu. Spectral methods for supervised topic models. In *NIPS*, 2014.

A Proof of Theorem 4.2

A.1 Proof sketch of Theorem 4.2

In this section we sketch the proof of Theorem 4.2. Our proof is mostly built upon the analysis in [1] for robust tensor power method. However, we borrow new ideas from [12] to substantially revise the per-iteration analysis (Lemma A.2), which subsequently results in desired relaxation of noise conditions. Some results and arguments in [1], especially those involved with absolute constants, are simplified for accessibility purposes.

We start with Lemma A.1 that analyzes random initializations against eigenvectors.

Lemma A.1. *Fix $j^* \in \{1, \cdots, k\}$ and $\eta \in (0, 1/2)$. Suppose L satisfies $L = \Omega(k/\eta)$. Then with probability at least $1 - \eta$ there exists a initialization $\boldsymbol{u}_0$ such that*

$$\max_{1 \leq j \leq k, j \neq j^*} |\boldsymbol{v}_j^\top \boldsymbol{u}_0| \leq 0.5 |\boldsymbol{v}_{j^*}^\top \boldsymbol{u}_0| \quad \textit{and} \quad |\boldsymbol{v}_{j^*}^\top \boldsymbol{u}_0| \geq 1/\sqrt{d}. \tag{8}$$

Roughly speaking, Lemma A.1 shows that with $L = \Omega(d \log d)$ initializations the initial vector $\boldsymbol{u}_0$ will slightly bias towards one of the directions j^* with overwhelming probability. The lemma is a slight generalization of Lemma B.1 in [1] to the $k \leq d$ case and their proofs are similar. For completeness purposes we include its proof in Appendix A.2. Applying a standard boosting argument we have the following corollary, which guarantees exponentially decaying failure probabilities:

Corollary A.1. *For any $\tilde{\eta} \in (0, 1/2)$, with $L = \Omega(k \log(1/\tilde{\eta}))$ initializations Eq. (8) holds for at least one initialization with probability at least $1 - \tilde{\eta}$.*

The following lemma is the key lemma that characterizes the recovery of *single* eigenvectors of the robust tensor power method.

Lemma A.2. *Suppose $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k \geq 0$ and assume without loss of generality that the conditions in Lemma A.1 hold with respect to $j^* = 1$. Assume in addition that*

$$\|\boldsymbol{\Delta}_T(\mathbf{I}, \boldsymbol{u}_t, \boldsymbol{u}_t)\|_2 \leq \min\left\{\tilde{\epsilon}_t, \frac{\lambda_1}{40\sqrt{d}}\right\}, \quad |\boldsymbol{\Delta}_T(\boldsymbol{v}_j, \boldsymbol{u}_t, \boldsymbol{u}_t)| \leq \min\left\{\frac{\tilde{\epsilon}_t}{\sqrt{k}}, \frac{\lambda_1}{8d}\right\}, \quad \tilde{\epsilon}_t \leq \frac{\lambda_1}{200}$$

for all $t \in [T]$ and j such that $\lambda_j > 0$. We then have that [3]

$$\max_{j \neq 1} \lambda_j |\boldsymbol{v}_j^\top \boldsymbol{u}_t| \leq 0.5 \lambda_1 |\boldsymbol{v}_1^\top \boldsymbol{u}_t|, \qquad \tan\theta(\boldsymbol{v}_1, \boldsymbol{u}_t) \leq 0.8 \tan\theta(\boldsymbol{v}_1, \boldsymbol{u}_{t-1}) + 8\tilde{\epsilon}_t/\lambda_1. \tag{9}$$

In addition, if $\theta(\boldsymbol{v}_1, \boldsymbol{u}_t) \leq \pi/3$ we have further that

$$\frac{|\boldsymbol{v}_j^\top \boldsymbol{u}_{t+1}|}{|\boldsymbol{v}_1^\top \boldsymbol{u}_{t+1}|} \leq 0.8 \frac{|\boldsymbol{v}_j^\top \boldsymbol{u}_t|}{|\boldsymbol{v}_1^\top \boldsymbol{u}_t|} + \frac{8\tilde{\epsilon}_t}{\lambda_1 \sqrt{k}}, \qquad \forall j > 1 \textit{ and } \lambda_j > 0. \tag{10}$$

Compared to existing analysis in (Propositions B.1, B.2, Lemmas B.2, B.3, B.4 in [1]), our proof in Appendix A.2 analyzes the two-phase behavior of robust tensor power method in a unified framework and is thus much cleaner. Furthermore, we borrow ideas from [12] to prove shrinkage of the tangent angle between $\boldsymbol{v}_1$ and $\boldsymbol{u}_t$, which subsequently leads to relaxed noise conditions. We also prove additional bounds regarding $|\boldsymbol{v}_j^\top \boldsymbol{u}_t|$ for $j > 1$ to facilitate later deflation analysis. This result is used for relaxing noise conditions only and is hence not proved in previous work [1].

Finally, we present the following lemma that analyzes the deflation step in the robust noisy power method, in which both “element-wise” and “full-vector” conditions on the deflated tensor are proved.

[3] For notational simplicity, let $\tan\theta(\boldsymbol{v}_1, \boldsymbol{u}_{-1}) = \infty$.

Lemma A.3. *Let $\{\hat{\lambda}_i, \hat{\boldsymbol{v}}_i\}_{i=1}^k$ be eigenvalue and (orthonormal) eigenvector pairs that approximates $\{\lambda_i, \boldsymbol{v}_i\}_{i=1}^k$ with $\lambda_1 \geq \cdots \geq \lambda_k > 0$ such that for all $i \in [k]$,*

$$|\hat{\lambda}_i - \lambda_i| \leq C\epsilon, \quad \tan\theta(\boldsymbol{v}_i, \boldsymbol{v}_i) \leq \min\{\sqrt{2}, C\epsilon/\lambda_i\} \quad |\hat{\boldsymbol{v}}_i^\top \boldsymbol{v}_j| \leq C\epsilon/(\lambda_i\sqrt{k}), \;\; \forall j > i \tag{11}$$

for some absolute constant $C > 0$ and error tolerance parameter $\epsilon > 0$. Denote $\mathbf{E}_i = \hat{\lambda}_i \hat{\boldsymbol{v}}_i^{\otimes 3} - \lambda_i \boldsymbol{v}_i^{\otimes 3}$ as the ith reconstruction error tensor. Let $\delta \in (0,1)$ be an arbitrary small constant. There exist universal constants $C > 0$ such that if $\epsilon \leq C'\lambda_{\min}/\sqrt{k}$ then the following holds for all $t \in [k]$ and $\|\boldsymbol{u}\|_2 = 1$:

$$\left\| \sum_{i=1}^t \mathbf{E}_i(\mathbf{I}, \boldsymbol{u}, \boldsymbol{u}) \right\|_2 \leq \kappa_t(\boldsymbol{u})\epsilon \quad \text{and} \quad \left| \sum_{i=1}^t \mathbf{E}_i(\boldsymbol{v}_j, \boldsymbol{u}, \boldsymbol{u}) \right| \leq \kappa_t(\boldsymbol{u})^2 \frac{\epsilon}{\sqrt{k}}, \;\; \forall j > t, \tag{12}$$

where $\kappa_t(\boldsymbol{u}) = \sqrt{\delta + C'' \sum_{i=1}^t |\boldsymbol{v}_i^\top \boldsymbol{u}|^2}$ and $C'' > 0$ is a universal constant.

We are now ready to prove the main theorem.

Proof of Theorem 4.2. We use induction to prove the theorem. For $i = 1$ all conditions in Lemma A.2 are satisfied with $\tilde{\epsilon}_t = 2\epsilon$ when $\epsilon \leq C_1\lambda_{\min}/\sqrt{k}$ for some sufficiently small constant $C_1 > 0$. Lemma A.2 then asserts that, with $L = \Omega(d \log d)$ initializations and $R = \Omega(\log(\lambda_1 k/\epsilon))$ iterations, $\|\hat{\boldsymbol{v}}_1 - \boldsymbol{v}_1\|_2 \leq \tan\theta(\hat{\boldsymbol{v}}_1, \boldsymbol{v}_1) \leq C_2\epsilon/\lambda_1$ for some universal constant $C_2 > 0$. Furthermore,

$$\begin{aligned}
|\hat{\lambda}_1 - \lambda_1| &= \left|\widetilde{\mathbf{T}}(\hat{\boldsymbol{v}}_1, \hat{\boldsymbol{v}}_1, \hat{\boldsymbol{v}}_1) - \lambda_1\right| \leq \left|\boldsymbol{\Delta}_T(\hat{\boldsymbol{v}}_1, \hat{\boldsymbol{v}}_1, \hat{\boldsymbol{v}}_1)\right| + \left|\mathbf{T}(\hat{\boldsymbol{v}}_1, \hat{\boldsymbol{v}}_1, \hat{\boldsymbol{v}}_1) - \lambda_1\right| \\
&\leq O\left(\frac{\epsilon}{\sqrt{k}}\right) + \left|\lambda_1 |\boldsymbol{v}_1^\top \hat{\boldsymbol{v}}_1|^3 - \lambda_1 + \sum_{j>1} \lambda_j |\boldsymbol{v}_j^\top \hat{\boldsymbol{v}}_1|^3\right| \\
&\leq O\left(\frac{\epsilon}{\sqrt{k}}\right) + \left|\lambda_1 \left[1 + O\left(\frac{\epsilon}{\lambda_1}\right)\right] - \lambda_1 + \sum_{j>1} \lambda_j O\left(\frac{\epsilon^3}{\lambda_j^3 k^{1.5}}\right)\right| \\
&\leq O(\epsilon), \quad \text{if } \epsilon \leq C_1\lambda_{\min}/\sqrt{k} \text{ for some sufficiently small constant } C_1.
\end{aligned}$$

We next prove the theorem for the case of $i+1$ assuming by induction that the theorem holds for all $\{\lambda_j, \boldsymbol{v}_j\}_{j=1}^i$. In this case, the "new" noise tensor $\widetilde{\boldsymbol{\Delta}}_T$ comes from both the original noise and also noise introduced by deflations; that is, $\widetilde{\boldsymbol{\Delta}}_T = \widetilde{\mathbf{T}} + \sum_{j=1}^i \mathbf{E}_i$. Invoking Lemma A.3 we have that $\widetilde{\boldsymbol{\Delta}}_T$ satisfies conditions in Lemma A.2 with

$$\tilde{\epsilon}_t = \epsilon\left(1 + \max\{\kappa_i(\boldsymbol{u}_t), \kappa_i(\boldsymbol{u}_t)^2\}\right),$$

where $\kappa_i(\boldsymbol{u}) = \sqrt{\delta + C'' \sum_{j=1}^i |\boldsymbol{u}^\top \boldsymbol{v}_j|^2}$ as defined in Lemma A.3, provided that $\epsilon \leq C_1\lambda_{\min}/\sqrt{k}$ for some sufficiently small constant C_1. Furthermore, note that for arbitrary $\delta \in (0,1)$, we can again pick $C_1' > 0$ to be a sufficiently small constant (possibly depending on δ) such that $\epsilon \leq C_1'\lambda_{\min}/\sqrt{k}$ would imply $\tilde{\epsilon}_t \leq \min\{\lambda_1/200, 0.01\lambda_{\min}\sqrt{\delta/(C''k)}\}$. Subsequently, by Eq. (9) we know that after $\Omega(\log(\lambda_{\max}k/\epsilon))$ iterations we have that $\tan\theta(\boldsymbol{u}_t, \boldsymbol{v}_{i+1}) \leq 0.1\sqrt{\delta/(C''k)}$ and hence for any $j \leq i$, $|\boldsymbol{u}_t^\top \boldsymbol{v}_j| = \cos\theta(\boldsymbol{u}_t, \boldsymbol{v}_j) = \sin\theta(\boldsymbol{u}_t, \boldsymbol{v}_{i+1}) \leq \tan\theta(\boldsymbol{u}_t, \boldsymbol{v}_{i+1}) \leq 0.1\sqrt{\delta/(C''k)}$. Consequently, $C'' \sum_{j=1}^i |\boldsymbol{u}_t^\top \boldsymbol{v}_j|^2 \leq 0.01\delta$ and therefore $\kappa_i(\boldsymbol{u}_t) \leq \sqrt{1.01\delta} \leq 1$. We then have that $\tilde{\epsilon}_t \leq 2\epsilon$ and hence the resuling bounds on $|\hat{\lambda}_{i+1} - \lambda_{i+1}|$ and $\tan\theta(\boldsymbol{u}_t, \boldsymbol{v}_{i+1})$ hold with the same constant C as all previous iterations before i. Finally, applying Lemma A.1 and taking a union bound over all k iterations we complete the proof. □

A.2 Proof of technical lemmas

Proof of Lemma A.1. Let $\tilde{\boldsymbol{u}}_0^{(\tau)} \overset{i.i.d.}{\sim} \mathcal{N}_d(0, \mathbf{I}_{d\times d})$ for $\tau \in [L]$ and define $Z_{j,\tau} = \boldsymbol{v}_j^\top \tilde{\boldsymbol{u}}_0^{(\tau)}$ for $j \in [d]$ and $\tau \in [L]$. Without loss of generality, assume $j^* = 1$. Consider the following sets of events:

$$\mathcal{E}_1 \;\; := \;\; \left\{ Z : \max_{\tau\in[L]} |Z_{1,\tau}| \geq 0.5\sqrt{\ln L} - \sqrt{2\ln(6/\eta)} \right\}, \tag{13}$$

$$\mathcal{E}_{2,\tau} := \left\{Z_{\cdot,\tau}: \max_{1<j\leq k}|Z_{j,\tau}| \leq \sqrt{2\ln k} + \sqrt{2\ln(3/\eta)}\right\}, \tag{14}$$

$$\mathcal{E}_{3,\tau} := \left\{Z_{\cdot,\tau}: \sum_{j=k+1}^{d}|Z_{j,\tau}|^2 \leq 3\ln(3/\eta)\cdot d + 2\ln(3/\eta)\right\}. \tag{15}$$

Suppose $\mathcal{E}_1$ holds with $\tau^* = \mathrm{argmax}_\tau|Z_{1,\tau}|$ and suppose in addition that $\mathcal{E}_{2,\tau^*}$ and $\mathcal{E}_{3,\tau^*}$ hold. To derive Eq. (8) we need to show the following inequalities:

$$0.5\sqrt{\ln L} - \sqrt{2\ln(6/\eta)} \quad \geq \quad 0.5\left(\sqrt{2\ln k} + \sqrt{2\ln(3/\eta)}\right);$$

$$\frac{(0.6\sqrt{\ln L} - \sqrt{2\ln(6/\eta)})^2}{k\cdot(0.6\sqrt{\ln L} - \sqrt{2\ln(6/\eta)})^2 + 3\ln(3/\eta)d + 2\ln(3/\eta)} \quad \geq \quad \frac{1}{d}.$$

It can be easily verified that $L = \Omega(k/\eta)$ satisfies the above inequalities and hence imply Eq. (8) under $\mathcal{E}_1 \cap \mathcal{E}_{2,\tau^*} \cap \mathcal{E}_{3,\tau^*}$.

The rest of the proof is to lower bound the probabilities of events $\mathcal{E}_1, \mathcal{E}_{2,\tau^*}$ and $\mathcal{E}_{3,\tau^*}$. We first consider $\mathcal{E}_1$. Because $Z_{1,1}, \cdots, Z_{1,L} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ and $f(Z_{1,1}, \cdots, Z_{1,L}) = \max_\tau |Z_{1,\tau}|$ is a 1-Lipschitz function, applying Lemma E.1 we have that

$$\Pr\left[\max_\tau |Z_{1,\tau}| < \mu - t\right] \leq 2e^{-t^2/2}, \tag{16}$$

where $\mu = \mathbb{E}[\max_\tau |Z_{1,\tau}|]$. By Lemma E.2, $\mu \geq \mathbb{E}[\max_\tau Z_{1,\tau}] \geq \sqrt{\ln L}/\sqrt{\pi\ln 2} \geq 0.5\sqrt{\ln L}$. Setting $t = \sqrt{2\ln(6/\eta)}$ in Eq. (16) we have that $\Pr[\mathcal{E}_1] \geq 1 - \eta/3$.

Next, suppose $\mathcal{E}_1$ holds with $\tau^* = \mathrm{argmax}_\tau|Z_{1,\tau}|$. Note that $\mathcal{E}_{2,\tau^*}$ and $\mathcal{E}_{3,\tau^*}$ are independent regardless of the choice of τ^*, because $Z_{1,\tau^*}, \cdots, Z_{d,\tau^*}$ are independent Gaussian random variables. We can then lower bound the probabilities of $\mathcal{E}_{2,\tau^*}$ and $\mathcal{E}_{3,\tau^*}$ separately. We consider $\mathcal{E}_{2,\tau^*}$ first. Because $Z_{2,\tau^*}, \cdots, Z_{k,\tau^*}$ are i.i.d. standard Normal random variables, applying Lemma E.3 we obtain

$$\Pr\left[\max_{2\leq j\leq k}|Z_{j,\tau^*}| > \sqrt{2\ln k} + \sqrt{2t}\right] \leq e^{-t}. \tag{17}$$

Putting $t = \ln(3/\eta)$ in Eq. (17) we have that $\Pr[\mathcal{E}_{2,\tau^*}|\mathcal{E}_1] \geq 1 - \eta/3$. For $\mathcal{E}_{3,\tau^*}$, it is obvious by definition that $\sum_{j=k+1}^d |Z_{j,\tau^*}|^2$ is a χ^2_{d-k}-distributed random variable and is independent of $\mathcal{E}_1$ and $\mathcal{E}_{2,\tau^*}$. Applying Lemma E.4 the following holds:

$$\Pr\left[\sum_{j=k+1}^d |Z_{j,\tau^*}|^2 > d + 2\sqrt{dt} + 2t\right] \leq e^{-t}. \tag{18}$$

Putting $t = \ln(3/\eta)$ in Eq. (18) and noting that $\sqrt{d} \leq d$, $t \geq 1$, we conclude that $\Pr[\mathcal{E}_{3,\tau^*}|\mathcal{E}_1] \geq 1 - \eta/3$. Finally, applying union bound we have that $\Pr[\mathcal{E}_1 \cap \mathcal{E}_{2,\tau^*} \cap \mathcal{E}_{3,\tau^*}] \geq 1 - \eta$. □

Proof of Lemma A.2. First, as a consequence of Corollary A.1, we know that $|\boldsymbol{v}_1^\top \boldsymbol{u}_0| \geq 1/\sqrt{d}$. The conditions in Lemma A.2 then imply $|\boldsymbol{\Delta}_T(\boldsymbol{v}_j, \boldsymbol{u}_t, \boldsymbol{u}_t)| \leq \lambda_1|\boldsymbol{v}_1^\top \boldsymbol{u}_0|^2/8$. We now use induction to prove Eq. (9). When $t = 0$ Eq. (9) trivially holds due to Lemma A.1 and the condition that $j^* = 1$ corresponds to the largest eigenvalue λ_1. The objective is then to prove Eq. (9) for the case of $t+1$, assuming it holds for all iterations up to t.

We first consider the second part of Eq. (9) concerning $\tan\theta(\boldsymbol{v}_1, \boldsymbol{u}_t)$. Let $\mathbf{V} \in \mathbb{R}^{d\times(k-1)}$ be an orthonormal basis of the complement subspace $\mathcal{V}_\perp = \mathrm{span}\{\boldsymbol{v}_2, \cdots, \boldsymbol{v}_k\}$. Further let $\boldsymbol{\varepsilon}_t = \boldsymbol{\Delta}_T(\mathbf{I}, \boldsymbol{u}_t, \boldsymbol{u}_t)$. We then have that

$$\tan\theta(\boldsymbol{v}_1, \boldsymbol{u}_{t+1}) = \tan\theta(\boldsymbol{v}_1, \mathbf{T}(\mathbf{I}, \boldsymbol{u}_t, \boldsymbol{u}_t) + \boldsymbol{\varepsilon}_t) \leq \frac{\|\mathbf{V}^\top\mathbf{T}(\mathbf{I}, \boldsymbol{u}_t, \boldsymbol{u}_t)\|_2 + \|\mathbf{V}^\top\boldsymbol{\varepsilon}_t\|_2}{|\boldsymbol{v}_1^\top\mathbf{T}(\mathbf{I}, \boldsymbol{u}, \boldsymbol{u})| - |\boldsymbol{v}_1^\top\boldsymbol{\varepsilon}_t|},$$

In addition, note that

$$\|\mathbf{V}^\top\mathbf{T}(\mathbf{I}, \boldsymbol{u}_t, \boldsymbol{u}_t)\|_2 = \sqrt{\sum_{j=2}^k \lambda_j^2|\boldsymbol{v}_j^\top\boldsymbol{u}_t|^4} \leq \max_{j\neq 1}\lambda_j|\boldsymbol{v}_j^\top\boldsymbol{u}_t|\cdot\sqrt{\sum_{j=2}^k|\boldsymbol{v}_j^\top\boldsymbol{u}_t|^2},$$

where the first equality is due to the orthogonality of $\{\boldsymbol{v}_2, \cdots, \boldsymbol{v}_k\}$ and in the last inequality we apply H´older's inequality. Because $\sqrt{\sum_{j=2}^k |\boldsymbol{v}_j^\top \boldsymbol{u}_t|^2} = \|\mathbf{V}^\top \boldsymbol{u}_t\|_2$, we have that

$$\begin{aligned}
\tan\theta(\boldsymbol{v}_1, \boldsymbol{u}_{t+1}) &\leq \frac{\|\mathbf{V}^\top \boldsymbol{u}_t\|_2 \cdot \max_{j\neq 1} \lambda_j |\boldsymbol{v}_j^\top \boldsymbol{u}_t| + \|\mathbf{V}^\top \boldsymbol{\varepsilon}_t\|_2}{|\boldsymbol{v}_1^\top \boldsymbol{u}_t| \cdot \lambda_1 |\boldsymbol{v}_1^\top \boldsymbol{u}_t| - |\boldsymbol{v}_1^\top \boldsymbol{\varepsilon}_t|} \\
&= \tan\theta(\boldsymbol{v}_1, \boldsymbol{u}_t) \left[\frac{\max_{j\neq 1} \lambda_j |\boldsymbol{v}_j^\top \boldsymbol{u}_t| + \|\mathbf{V}^\top \boldsymbol{\varepsilon}_t\|_2 / \|\mathbf{V}^\top \boldsymbol{u}_t\|_2}{\lambda_1 |\boldsymbol{v}_1^\top \boldsymbol{u}_t| - |\boldsymbol{v}_1^\top \boldsymbol{\varepsilon}_t| / |\boldsymbol{v}_1^\top \boldsymbol{u}_t|}\right] \\
&\leq \tan\theta(\boldsymbol{v}_1, \boldsymbol{u}_t) \left[\frac{0.5\lambda_1 |\boldsymbol{v}_1^\top \boldsymbol{u}_t| + \|\mathbf{V}^\top \boldsymbol{\varepsilon}_t\|_2 / \|\mathbf{V}^\top \boldsymbol{u}_t\|_2}{\lambda_1 |\boldsymbol{v}_1^\top \boldsymbol{u}_t| - |\boldsymbol{v}_1^\top \boldsymbol{\varepsilon}_t| / |\boldsymbol{v}_1^\top \boldsymbol{u}_t|}\right] \\
&= \tan\theta(\boldsymbol{v}_1, \boldsymbol{u}_t) \underbrace{\left[\frac{1}{2} \frac{1}{1 - |\boldsymbol{v}_1^\top \boldsymbol{\varepsilon}_t| / (\lambda_1 |\boldsymbol{v}_1^\top \boldsymbol{u}_t|^2)}\right]}_{\alpha} + \underbrace{\frac{1}{1 - |\boldsymbol{v}_1^\top \boldsymbol{\varepsilon}_t| / (\lambda_1 |\boldsymbol{v}_1^\top \boldsymbol{u}_t|^2)}}_{2\alpha} \cdot \underbrace{\frac{\|\mathbf{V}^\top \boldsymbol{\varepsilon}_t\|_2}{\lambda_1 |\boldsymbol{v}_1^\top \boldsymbol{u}_t|^2}}_{\beta}.
\end{aligned} \tag{19}$$

Here in Line 19 we apply the induction hypothesis that $\max_{j\neq 1} \lambda_j |\boldsymbol{v}_j^\top \boldsymbol{u}_t| \leq 0.5\lambda_1 |\boldsymbol{v}_1^\top \boldsymbol{u}_t|$. Before proceeding the analysis we first show that $|\boldsymbol{v}_1^\top \boldsymbol{u}_0| \leq |\boldsymbol{v}_1^\top \boldsymbol{u}_t|$. Applying the induction hypothesis, we have that

$$\tan\theta(\boldsymbol{v}_1, \boldsymbol{u}_t) \leq 0.8^t \tan\theta(\boldsymbol{v}_1, \boldsymbol{u}_0) + 40\tilde{\epsilon}_t/\lambda_1 \leq 0.8 \tan\theta(\boldsymbol{v}_1, \boldsymbol{u}_0) + 40\tilde{\epsilon}_t/\lambda_1 \leq \tan\theta(\boldsymbol{v}_1, \boldsymbol{u}_0),$$

where the last inequality is due to $\tilde{\epsilon}_t \leq \lambda_1/200$. Subsequently, $\theta(\boldsymbol{v}_1, \boldsymbol{u}_t) \leq \theta(\boldsymbol{v}_1, \boldsymbol{u}_0)$ and hence $|\boldsymbol{v}_1^\top \boldsymbol{u}_t| = \cos\theta(\boldsymbol{v}_1, \boldsymbol{u}_t) \geq \cos\theta(\boldsymbol{v}_1, \boldsymbol{u}_0) = |\boldsymbol{v}_1^\top \boldsymbol{u}_0|$. Now applying $|\boldsymbol{v}_1^\top \boldsymbol{\varepsilon}_t| \leq |\boldsymbol{v}_1^\top \boldsymbol{u}_0|^2/4$ we obtain

$$|\boldsymbol{v}_1^\top \boldsymbol{\varepsilon}_t| \leq \frac{\lambda_1 |\boldsymbol{v}_1^\top \boldsymbol{u}_0|^2}{4} \leq \frac{\lambda_1 |\boldsymbol{v}_1^\top \boldsymbol{u}_t|^2}{4} \implies \frac{1}{1 - |\boldsymbol{v}_1^\top \boldsymbol{\varepsilon}_t| / (\lambda_1 |\boldsymbol{v}_1^\top \boldsymbol{u}_t|^2)} \leq \frac{3}{2} \implies \alpha \leq \frac{3}{4}. \tag{20}$$

Next we bound β by considering two cases. In the first case of $|\boldsymbol{v}_1^\top \boldsymbol{u}_t| \leq 0.5$, we have that

$$\beta = \tan\theta(\boldsymbol{v}_1, \boldsymbol{u}_t) \cdot \frac{\|\mathbf{V}^\top \boldsymbol{\varepsilon}_t\|_2}{\lambda_1 |\boldsymbol{v}_1^\top \boldsymbol{u}_t| \sqrt{1 - |\boldsymbol{v}_1^\top \boldsymbol{u}_t|^2}} \leq \frac{2\|\boldsymbol{\varepsilon}_t\|_2}{\lambda_1 |\boldsymbol{v}_1^\top \boldsymbol{u}_t|} \cdot \tan\theta(\boldsymbol{v}_1, \boldsymbol{u}_t) \leq 0.05 \tan\theta(\boldsymbol{v}_1, \boldsymbol{u}_t). \tag{21}$$

where the last inequality is due to the condition that $\|\boldsymbol{\varepsilon}_t\|_2 \leq \frac{\lambda_1 |\boldsymbol{v}_1^\top \boldsymbol{u}_0|}{40} \leq \frac{\lambda_1 |\boldsymbol{v}_1^\top \boldsymbol{u}_t|}{40}$. On the other hand, if $|\boldsymbol{v}_1^\top \boldsymbol{u}_t| > 0.5$ the following holds:

$$\beta = \frac{\|\mathbf{V}^\top \boldsymbol{\varepsilon}_t\|_2}{\lambda_1 |\boldsymbol{v}_1^\top \boldsymbol{u}_t|^2} \leq \frac{4\|\boldsymbol{\varepsilon}_t\|_2}{\lambda_1} \leq \frac{4\tilde{\epsilon}_t}{\lambda_1}. \tag{22}$$

Combining Eq. (20,21,22) we obtain $\tan\theta(\boldsymbol{v}_1, \boldsymbol{u}_{t+1}) \leq 0.8 \tan\theta(\boldsymbol{v}_1, \boldsymbol{u}_t) + 8\tilde{\epsilon}_t/\lambda_1$.

We next prove the first part of Eq. (9), namely that $\max_{j\neq 1} \lambda_j |\boldsymbol{v}_j^\top \boldsymbol{u}_{t+1}| \leq 0.5\lambda_1 |\boldsymbol{v}_1^\top \boldsymbol{u}_{t+1}|$. For those j with $\lambda_j = 0$ the bound trivially holds. So we consider only $j > 1$ with $\lambda_j > 0$. We then have that

$$\frac{\lambda_1 |\boldsymbol{v}_1^\top \boldsymbol{u}_{t+1}|}{\lambda_j |\boldsymbol{v}_j^\top \boldsymbol{u}_{t+1}|} = \frac{\lambda_1 |\boldsymbol{v}_1^\top [\mathbf{T}(\mathbf{I}, \boldsymbol{u}_t, \boldsymbol{u}_t) + \boldsymbol{\varepsilon}_t]|}{\lambda_j |\boldsymbol{v}_j^\top [\mathbf{T}(\mathbf{I}, \boldsymbol{u}_t, \boldsymbol{u}_t) + \boldsymbol{\varepsilon}_t]|} \geq \underbrace{\left(\frac{\lambda_1 |\boldsymbol{v}_1^\top \boldsymbol{u}_t|}{\lambda_j |\boldsymbol{v}_j^\top \boldsymbol{u}_t|}\right)^2}_{\alpha'} \cdot \underbrace{\frac{1 - |\boldsymbol{v}_1^\top \boldsymbol{\varepsilon}_t| / (\lambda_1 |\boldsymbol{v}_1^\top \boldsymbol{u}_t^2|)}{1 + |\boldsymbol{v}_j^\top \boldsymbol{\varepsilon}_t| / (\lambda_j |\boldsymbol{v}_j^\top \boldsymbol{\varepsilon}_t|^2)}}_{\beta'}.$$

By induction hypothesis $\alpha' \geq 4$. Applying conditions on $|\boldsymbol{v}_1^\top \boldsymbol{\varepsilon}_t|$ we get $|\boldsymbol{v}_1^\top \boldsymbol{\varepsilon}_t| \leq \frac{\lambda_1 |\boldsymbol{v}_1^\top \boldsymbol{u}_0|^2}{4} \leq \frac{\lambda_1 |\boldsymbol{v}_1^\top \boldsymbol{u}_t|^2}{4}$ and hence $|\boldsymbol{v}_1^\top \boldsymbol{\varepsilon}_t| / (\lambda_1 |\boldsymbol{v}_1^\top \boldsymbol{u}_t|^2) \leq 1/4$. On the other hand,

$$\left(\frac{\lambda_1 |\boldsymbol{v}_1^\top \boldsymbol{u}_t|}{\lambda_j |\boldsymbol{v}_j^\top \boldsymbol{u}_t|}\right)^2 \left[1 + \frac{|\boldsymbol{v}_j^\top \boldsymbol{\varepsilon}|}{\lambda_j |\boldsymbol{v}_j^\top \boldsymbol{u}_t|^2}\right]^{-1} = \left[\left(\frac{\lambda_j |\boldsymbol{v}_j^\top \boldsymbol{u}_t|}{\lambda_1 |\boldsymbol{v}_1^\top \boldsymbol{u}_t|}\right)^2 + \frac{\lambda_j |\boldsymbol{v}_j^\top \boldsymbol{\varepsilon}_t|}{\lambda_1^2 |\boldsymbol{v}_1^\top \boldsymbol{u}_t|^2}\right]^{-1} \geq \left[\frac{1}{4} + \frac{|\boldsymbol{v}_j^\top \boldsymbol{\varepsilon}_t|}{\lambda_1 |\boldsymbol{v}_1^\top \boldsymbol{u}_t|^2}\right]^{-1}.$$

Because $|\boldsymbol{v}_j^\top \boldsymbol{\varepsilon}_t| \leq \frac{\lambda_1 |\boldsymbol{v}_1^\top \boldsymbol{u}_0|^2}{8} \leq \frac{\lambda_1 |\boldsymbol{v}_1^\top \boldsymbol{u}_t|^2}{8}$, the right-hand side of the above equation is lower bounded by 8/3. Therefore, $\alpha' \beta' \geq \frac{8}{3}(1 - \frac{1}{4}) \geq 2$.

The last part of this proof is devoted to showing Eq. (10). Under the condition that $\theta(\boldsymbol{v}_1, \boldsymbol{u}_t) \le \pi/3$ we have that $\cos\theta(\boldsymbol{v}_1, \boldsymbol{u}_t) = |\boldsymbol{v}_1^\top \boldsymbol{u}_t| \ge 1/2$. Subsequently, for arbitrary $j > 1$ with $\lambda_j > 0$ the following holds:

$$\frac{|\boldsymbol{v}_j^\top \boldsymbol{u}_{t+1}|}{|\boldsymbol{v}_1^\top \boldsymbol{u}_{t+1}|} \le \frac{\lambda_j |\boldsymbol{v}_j^\top \boldsymbol{u}_t|^2 + |\boldsymbol{v}_j^\top \boldsymbol{\varepsilon}_t|}{\lambda_1 |\boldsymbol{v}_1^\top \boldsymbol{u}_t|^2 - |\boldsymbol{v}_1^\top \boldsymbol{\varepsilon}_t|} \le \frac{|\boldsymbol{v}_j^\top \boldsymbol{u}_t|}{|\boldsymbol{v}_1^\top \boldsymbol{u}_t|} \cdot \underbrace{\frac{1}{2}\frac{1}{1 - |\boldsymbol{v}_1^\top \boldsymbol{\varepsilon}_t|/(\lambda_1 |\boldsymbol{v}_1^\top \boldsymbol{u}_t|^2)}}_{\alpha} + \underbrace{\frac{|\boldsymbol{v}_j^\top \boldsymbol{\varepsilon}_t|}{\lambda_1 |\boldsymbol{v}_1^\top \boldsymbol{u}_t|^2 - |\boldsymbol{v}_1^\top \boldsymbol{\varepsilon}_t|}}_{\gamma}.$$

Because $|\boldsymbol{v}_1^\top \boldsymbol{u}_t| \ge 1/2$ and $|\boldsymbol{v}_1^\top \boldsymbol{\varepsilon}_t| \le \frac{1}{2}\lambda_1 |\boldsymbol{v}_1^\top \boldsymbol{u}_0|^2 \le \frac{1}{2}\lambda_1 |\boldsymbol{v}_1^\top \boldsymbol{u}_t|^2$, we have $\gamma \le 8|\boldsymbol{v}_j^\top \boldsymbol{\varepsilon}_t|/\lambda_1$ and hence

$$|\boldsymbol{v}_j^\top \boldsymbol{u}_{t+1}| \le 0.8|\boldsymbol{v}_j^\top \boldsymbol{u}_t| + \frac{8|\boldsymbol{v}_j^\top \boldsymbol{\varepsilon}_t|}{\lambda_1} \le 0.8|\boldsymbol{v}_j^\top \boldsymbol{u}_t| + \frac{8\tilde{\epsilon}_t}{\lambda_1\sqrt{k}}.$$

□

Proof of Lemma A.3. The first part of Eq. (12) is a simplified result of Lemma B.5 [4] in [1] because $\|\hat{\boldsymbol{v}}_i - \boldsymbol{v}_i\|_2 \le \tan\theta(\hat{\boldsymbol{v}}_i, \boldsymbol{v}_i)$ when $\|\hat{\boldsymbol{v}}_i\|_2 = \|\boldsymbol{v}_i\|_2 = 1$ and $\theta < \pi/2$. The proofs are almost identical. So we focus on proving the second part of Eq. (12) here. Recall that $\boldsymbol{v}_j^\top \boldsymbol{v}_i = 0$ for all $j > i$. Subsequently,

$$\left|\sum_{i=1}^t \mathbf{E}_i(\boldsymbol{v}_j, \boldsymbol{u}, \boldsymbol{u})\right| \le \sum_{i=1}^t \hat{\lambda}_i |\boldsymbol{u}^\top \hat{\boldsymbol{v}}_i|^2 |\boldsymbol{v}_j^\top \hat{\boldsymbol{v}}_i| \le \frac{C\epsilon}{\sqrt{k}} \sum_{i=1}^t \frac{\hat{\lambda}_i}{\lambda_i} |\boldsymbol{u}^\top \hat{\boldsymbol{v}}_i|^2.$$

Define $\hat{\boldsymbol{v}}_i^\perp = \hat{\boldsymbol{v}}_i - (\hat{\boldsymbol{v}}_i^\top \boldsymbol{v}_i)\boldsymbol{v}_i$ as the difference between $\hat{\boldsymbol{v}}_i$ and its projection on $\boldsymbol{v}_i$. It is then by definition that $\|\hat{\boldsymbol{v}}_i^\perp\|_2 = \|\hat{\boldsymbol{v}}_i\|_2 \sin\theta(\hat{\boldsymbol{v}}_i, \boldsymbol{v}_i) \le \tan\theta(\hat{\boldsymbol{v}}_i, \boldsymbol{v}_i)$. Subsequently,

$$\begin{aligned}\sum_{i=1}^t \frac{\hat{\lambda}_i}{\lambda_i} |\boldsymbol{u}^\top \hat{\boldsymbol{v}}_i|^2 &\le \sum_{i=1}^t \left(1 + \frac{|\hat{\lambda}_i - \lambda_i|}{\lambda_i}\right) |\boldsymbol{u}^\top \hat{\boldsymbol{v}}_i|^2 \le \frac{C\epsilon}{\lambda_{\min}} + \sum_{i=1}^t \left(|\boldsymbol{u}^\top \boldsymbol{v}_i|^2 + |\boldsymbol{u}^\top \hat{\boldsymbol{v}}_i^\perp|^2\right)\\ &\le \frac{C\epsilon}{\lambda_{\min}} + k\|\hat{\boldsymbol{v}}_i^\perp\|_2^2 + \sum_{i=1}^t |\boldsymbol{u}^\top \boldsymbol{v}_i|^2 \le \underbrace{\frac{C\epsilon}{\lambda_{\min}} + \frac{C^2 k \epsilon^2}{\lambda_{\min}^2}}_{a} + \sum_{i=1}^t |\boldsymbol{u}^\top \boldsymbol{v}_i|^2.\end{aligned}$$

For arbitrary $\delta \in (0,1)$, set $C' \le \min\{\frac{\delta}{2C^2}, \frac{\delta}{2C^3}\}$ and $\epsilon \le C'\lambda_{\min}/\sqrt{k}$ we have that $a \le \delta/C$ and hence the second part of Eq. (12) holds with $C'' = C$. □

B Proof of Proposition 4.1

Proof of Proposition 4.1. First, we decompose $\mathbf{M}(\boldsymbol{\theta})$ into two terms:

$$\mathbf{M}(\boldsymbol{\theta}) = \sum_{i=1}^k \lambda_i(\boldsymbol{v}_i^\top \boldsymbol{\theta}) \cdot \boldsymbol{v}_i \boldsymbol{v}_i^\top + \boldsymbol{\Delta}_T(\mathbf{I}, \mathbf{I}, \boldsymbol{\theta}).$$

Define $\bar{\lambda}_i = \lambda_i(\boldsymbol{v}_i^\top \boldsymbol{\theta})$ and $\bar{\mathbf{E}} = \boldsymbol{\Delta}_T(\mathbf{I}, \mathbf{I}, \boldsymbol{\theta})$. We then have that

$$\mathbf{M}(\boldsymbol{\theta}) = \mathbf{M}_0 + \bar{\mathbf{E}},$$

where $\mathbf{M}_0$ is a $d \times d$ rank-k matrix with eigenvalues $(\bar{\lambda}_1, \cdots, \bar{\lambda}_k)$ and eigenvectors $(\boldsymbol{v}_1, \cdots, \boldsymbol{v}_k)$, and $\bar{\mathbf{E}}$ satisfies $\|\bar{\mathbf{E}}\|_2 \le \|\boldsymbol{\Delta}_T\|_{\mathrm{op}}$. Since $\boldsymbol{\theta}$ is uniformly sampled from the d-dimensional unit sphere, by standard concentration arguments we have that $|\boldsymbol{v}_j^\top \boldsymbol{\theta}| = \widetilde{\Omega}(1/\sqrt{d})$ with overwhelming probability for all $j = 1, \cdots, k$ and hence

$$\sigma_k(\mathbf{M}_0) = \widetilde{\Omega}(\lambda_{\min}/\sqrt{d}),$$

[4] Except that we operate under a $k < d$ regime, which adds no difficulty to the proof.

where $\sigma_k(\cdot)$ denotes the kth largest singular value of a matrix. Applying Weyl's theorem (Lemma E.7) we have that

$$\sigma_k(\mathbf{M}(\boldsymbol{\theta})) \geq \sigma_k(\mathbf{M}_0) - \|\bar{\mathbf{E}}\|_2 = \widetilde{\Omega}(\lambda_{\min}/\sqrt{d}),$$

where the last inequality is due to the condition imposed on noise magnitude $\|\boldsymbol{\Delta}_T\|_{\mathrm{op}}$ and the fact that $\|\bar{\mathbf{E}}\|_2 \leq \|\boldsymbol{\Delta}_T\|_{\mathrm{op}}$. Applying Wedin's theorem (Lemma E.8) with $\alpha = 0$ and $\delta = \sigma_k(\mathbf{M}(\boldsymbol{\theta})) = \widetilde{\Omega}(\lambda_{\min}/\sqrt{d})$ we arrive at

$$\|\boldsymbol{\Pi}_{\mathcal{W}} - \boldsymbol{\Pi}_{\hat{\mathcal{W}}}\|_2 \leq \frac{\|\bar{\mathbf{E}}\|_2}{\sigma_k(\mathbf{M}(\boldsymbol{\theta}))} \leq \widetilde{O}\left(\frac{\sqrt{d}\|\boldsymbol{\Delta}_T\|_{\mathrm{op}}}{\lambda_{\min}}\right).$$

□

C Proof of results for streaming robust tensor power method

Proof of Theorem 2.1. First, note that if $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n \overset{i.i.d.}{\sim} P$, $P \in \mathcal{SG}_D(\sigma)$ then the distribution of the sample mean $\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i$ belongs to $\mathcal{SG}_D(\sigma/\sqrt{n})$. To see this, fix any $\boldsymbol{a} \in \mathbb{R}^D$ and one can show that

$$\mathbb{E}\left[\exp(\boldsymbol{a}^\top \bar{\boldsymbol{x}})\right] = \prod_{i=1}^n \mathbb{E}\left[\exp(\boldsymbol{a}^\top \boldsymbol{x}_i/n)\right] \leq \prod_{i=1}^n \exp(\|\boldsymbol{a}\|_2^2\sigma^2/n^2) = \exp(\|\boldsymbol{a}\|_2^2\sigma^2/n),$$

where the second inequality is due to the fact that $\boldsymbol{x}_i \in \mathcal{SG}_D(\sigma)$ and $\|\boldsymbol{a}/n\|_2^2 = \|\boldsymbol{a}\|_2^2/n^2$.

Under Assumptions 2.1, 2.2 and using the the above arguments, we know that

$$\mathrm{vec}(\boldsymbol{\Delta}_T) = \mathrm{vec}\left[\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i^{\otimes 3} - \mathbf{T}\right] \in \mathcal{SG}_{d^3}(\sigma/n)$$

Now fix $\boldsymbol{v}_i, \boldsymbol{u}_t \in \mathbb{R}^d$ with unit L_2 norms. Applying Lemma E.6 with respect to $\boldsymbol{\Sigma} = \mathrm{vec}(\boldsymbol{v}_i \otimes \boldsymbol{u}_t \otimes \boldsymbol{u}_t)\mathrm{vec}(\boldsymbol{v}_i \otimes \boldsymbol{u}_t \otimes \boldsymbol{u}_t)^\top$ we obtain that

$$\Pr\left[\left|\boldsymbol{\Delta}_T(\boldsymbol{v}_i, \boldsymbol{u}_t, \boldsymbol{u}_t)\right|^2 > (1 + 2\sqrt{t} + t)\frac{\sigma^2}{n}\right] \leq e^{-t}, \qquad \forall t > 0. \tag{23}$$

Subsequently, with overwhelming probability (e.g., $\geq 1 - n^{-10}$) we have that

$$\|\boldsymbol{\Delta}_T(\mathbf{I}, \boldsymbol{u}_t, \boldsymbol{u}_t)\|_2 = \widetilde{O}\left(\sigma\sqrt{\frac{d}{n}}\right), \qquad \left|\boldsymbol{\Delta}_T(\boldsymbol{v}_i, \boldsymbol{u}_t, \boldsymbol{u}_t)\right| = \widetilde{O}\left(\sigma\sqrt{\frac{1}{n}}\right).$$

Finally, with

$$n = \widetilde{\Omega}\left(\min\left\{\frac{\sigma^2 d}{\epsilon^2}, \frac{\sigma^2 d^2}{\lambda_{\min}^2}\right\}\right)$$

the conditions in Eq. (6) are satisfied with overwhelming probability and hence the error bounds on $|\lambda_i - \hat{\lambda}_{\pi(i)}|$ and $\|\boldsymbol{v}_i - \hat{\boldsymbol{v}}_{\pi(i)}\|_2$. □

D Proofs of results for differentially private tensor decomposition

Proof of Proposition 3.1. The procedure described in Algorithm 3 can be thought of as a SuLQ framework with $K = kL(R+1)$ queries, each time querying $\mathbf{T}(\mathbf{I}, \boldsymbol{u}, \boldsymbol{u})$ and/or $\mathbf{T}(\boldsymbol{u}, \boldsymbol{u}, \boldsymbol{u})$ for some specific unit d-dimensonal vector $\boldsymbol{u}$. Suppose $\mathbf{T}' = \mathbf{T} + \mathrm{symmetrize}(\boldsymbol{e}_i \otimes \boldsymbol{e}_j \otimes \boldsymbol{e}_k)$. The ℓ_2-sensitivity of both queries can be separately bounded as

$$\begin{aligned}\Delta_2 f_1 &= \sup_{\mathbf{T}'} \|\mathbf{T}(\mathbf{I}, \boldsymbol{u}, \boldsymbol{u}) - \mathbf{T}'(\mathbf{I}, \boldsymbol{u}, \boldsymbol{u})\|_2 \leq \sup_{i,j,k} 2(|\boldsymbol{u}_i\boldsymbol{u}_j| + |\boldsymbol{u}_i\boldsymbol{u}_k| + |\boldsymbol{u}_j\boldsymbol{u}_k|) \leq 6\|\boldsymbol{u}\|_\infty^2;\\ \Delta_2 f_2 &= \sup_{\mathbf{T}'} \left|\mathbf{T}(\boldsymbol{u}, \boldsymbol{u}, \boldsymbol{u}) - \mathbf{T}'(\boldsymbol{u}, \boldsymbol{u}, \boldsymbol{u})\right| = \sup_{i,j,k} 6\left|\boldsymbol{u}_i\boldsymbol{u}_j\boldsymbol{u}_k\right| \leq 6\|\boldsymbol{u}\|_\infty^3.\end{aligned}$$

Thus, applying the Gaussian mechanism [9] we can (ε, δ)-privately release *one output* of either $f_1(\boldsymbol{u})$ or $f_2(\boldsymbol{u})$ by releasing

$$f_\ell(\boldsymbol{u}) + \frac{\Delta_2 f_\ell \cdot \sqrt{2\ln(1.25/\delta)}}{\varepsilon} \cdot \boldsymbol{w},$$

where $\ell = 1, 2$ and $\boldsymbol{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ are i.i.d. standard Normal random variables. Finally, applying *advanced composition* [9] across $K = kL(R+1)$ private releases we complete the proof of this proposition. Note that both normalization and deflation steps do not affect the differential privacy of Algorithm 3 due to the *closeness under post-processing* property of DP.

□

Before proving Theorem 3.1, we first present a lemma that upper bounds $\|\boldsymbol{u}_t\|_\infty$ when the components $\mathbf{V} \in \mathbb{R}^{d\times k}$ is incoherent (Assumption 3.1) and Gaussian noise across power updates is added.

Lemma D.1. *Suppose $\mathbf{T} = \sum_{i=1}^k \lambda_i \boldsymbol{v}_i^{\otimes} 3$ and $\mathbf{V} = (\boldsymbol{v}_1, \cdots, \boldsymbol{v}_k)$ satisfies Assumption 3.1 with coherence level μ_0. Fix $\boldsymbol{u} \in \mathbb{R}^d$ with $\|\boldsymbol{u}\|_2 = 1$ and let $\bar{\boldsymbol{u}} = \mathbf{T}(\mathbf{I}, \boldsymbol{u}, \boldsymbol{u}) + \sigma \cdot \boldsymbol{z}$, where $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d\times d})$ are zero-mean independently distributed Gaussian random variables. We then have that*

$$\frac{\|\bar{\boldsymbol{u}}\|_\infty}{\|\bar{\boldsymbol{u}}\|_2} = O\left(\sqrt{\frac{\mu_0 k \log d}{d}}\right).$$

with overwhelming probability.

Proof. We prove this lemma by showing an upper bound for $\|\bar{\boldsymbol{u}}\|_\infty$ and a lower bound on $\|\bar{\boldsymbol{u}}\|_2$, both with overwhelming probabilities. For the infinity-norm upper bound, we consider the following decomposition via triangle inequality:

$$\|\bar{\boldsymbol{u}}\|_\infty \le \|\tilde{\boldsymbol{u}}\|_\infty + \sigma\|\boldsymbol{z}\|_\infty,$$

where $\tilde{\boldsymbol{u}} = \mathbf{T}(\mathbf{I}, \boldsymbol{u}, \boldsymbol{u})$ and $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d\times d})$. By definition,

$$\|\tilde{\boldsymbol{u}}\|_\infty = \left\|\sum_{i=1}^k \lambda_i |\boldsymbol{u}^\top \boldsymbol{v}_i|^2 \boldsymbol{v}_i\right\|_\infty = \max_{1\le j\le d}\left|\boldsymbol{\lambda}^\top(\mathbf{V}^\top \boldsymbol{e}_j)\right|,$$

where $\boldsymbol{\lambda}$ is a k-dimensional vector defined as $\boldsymbol{\lambda} = (\lambda_1|\boldsymbol{u}^\top\boldsymbol{v}_1|^2, \cdots, \lambda_k|\boldsymbol{u}^\top\boldsymbol{v}_k|^2)$. By Cauchy-Schwarts inequality, we have that

$$\|\tilde{\boldsymbol{u}}\|_\infty = \max_{1\le j\le d}\left|\boldsymbol{\lambda}^\top(\mathbf{V}^\top \boldsymbol{e}_j)\right| \le \|\boldsymbol{\lambda}\|_2 \cdot \max_{1\le j\le d}\|\mathbf{V}^\top\boldsymbol{e}_j\|_2 \le \sqrt{\frac{\mu_0 k}{d}\left(\sum_{i=1}^k \lambda_i^2|\boldsymbol{u}^\top\boldsymbol{v}_i|^4\right)},$$

where the last inequality is due to the condition that $\mathbf{V}$ is incoherent with coherence level μ_0. In addition, $\|\boldsymbol{z}\|_\infty = O(\sqrt{\log d})$ with overwhelming probability, by applying Lemma E.3. As a result,

$$\|\bar{\boldsymbol{u}}\|_\infty \le \sqrt{\frac{2k\mu_0}{d}\left(\sum_{i=1}^k \lambda_i^2|\boldsymbol{u}^\top\boldsymbol{v}_i|^4\right)} + O(\sigma\sqrt{\log d}). \tag{24}$$

We next lower bound the denominator term $\|\bar{\boldsymbol{u}}\|_2$. By definition, $\bar{\boldsymbol{u}}$ follows a multi-variate Gaussian distribution with mean $\tilde{\boldsymbol{u}}$ and co-variance $\sigma^2\mathbf{I}_{d\times d}$. Applying Lemma E.5 with $\mu = \|\tilde{\boldsymbol{u}}\|_2^2/\sigma^2$ and $t = O(\log d)$ we have that $\|\bar{\boldsymbol{u}}\|_2^2 = \Omega(\|\tilde{\boldsymbol{u}}\|_2^2 + \sigma^2 d)$ with overwhelming probability. Note also that

$$\|\tilde{\boldsymbol{u}}\|_2^2 = \left\|\sum_{i=1}^k \lambda_i|\boldsymbol{u}^\top\boldsymbol{v}_i|^2\boldsymbol{v}_i\right\|_2^2 = \sum_{i=1}^k \lambda_i^2|\boldsymbol{u}^\top\boldsymbol{v}_i|^4$$

because $\{\boldsymbol{v}_i\}_{i=1}^k$ are orthonormal vectors. Consequently,

$$\|\bar{\boldsymbol{u}}\|_2^2 = \Omega\left(\sqrt{\sigma^2 d + \sum_{i=1}^k \lambda_i^2|\boldsymbol{u}^\top\boldsymbol{v}_i|^4}\right). \tag{25}$$

Combining Eqs. (24,25) we obtain

$$\frac{\|\bar{\boldsymbol{u}}\|_\infty}{\|\bar{\boldsymbol{u}}\|_2} \leq \frac{\sqrt{\frac{2\mu_0 k}{d}\sum_{i=1}^k \lambda_i^2|\boldsymbol{u}^\top\boldsymbol{v}_i|^4} + O(\sigma\sqrt{\log d})}{\Omega\left(\sqrt{\sigma^2 d + \sum_{i=1}^k \lambda_i^2|\boldsymbol{u}^\top\boldsymbol{v}_i|^4}\right)} \leq O\left(\sqrt{\frac{\mu_0 k}{d}}\right) + O\left(\sqrt{\frac{\log d}{d}}\right) = O\left(\sqrt{\frac{\mu_0 k\log d}{d}}\right).$$

□

We are now ready to prove Theorem 3.1.

Proof of Theorem 3.1. Applying Lemma D.1 we can with overwhelming probability upper bound the per-coordinate standard deviation of Gaussian noise calibrated in Algorithm 3:

$$\max_{\substack{0\leq t\leq T\\ 1\leq\tau\leq L}}\left\{\nu\|\boldsymbol{u}_t^{(\tau)}\|_\infty^2, \nu\|\boldsymbol{u}_t^{(\tau)}\|_\infty^3\right\} \leq O\left(\frac{\sqrt{K}\cdot\log(1/\delta)}{\varepsilon}\cdot\frac{\mu_0 k\log d}{d}\right),$$

where $K = kL(T+1) = \widetilde{O}(k^2\log(\lambda_{\max}d))$. Let $\boldsymbol{\epsilon}_t^{(\tau)} = \mathbf{E}(\mathbf{I}, \boldsymbol{u}_t^{(\tau)}, \boldsymbol{u}_t^{(\tau)}) = \sigma_t^{(\tau)}\cdot\boldsymbol{z}$ be the noise vector calibrated, where $\sigma_t^{(\tau)} = \nu\|\boldsymbol{u}_t^{(\tau)}\|_\infty^2$. We then have that with overwhelming probability,

$$\|\boldsymbol{\epsilon}_t^{(\tau)}\|_2 = O\left(\frac{\mu_0 k^2\log(\lambda_{\max}d/\delta)}{\varepsilon\sqrt{d}}\right) \qquad \text{and} \qquad \left|\boldsymbol{v}_1^\top\boldsymbol{\epsilon}_t^{(\tau)}\right| = O\left(\frac{\mu_0 k^2\log(\lambda_{\max}d/\delta)}{\varepsilon d}\right).$$

Equating the upper bound for $|\boldsymbol{v}_1^\top\boldsymbol{\epsilon}_t^{(\tau)}|$ with $O(\lambda_{\min}/d)$ we obtain the desired privacy level condition:

$$\varepsilon = \Omega\left(\frac{\mu_0 k^2\log(\lambda_{\max}d/\delta)}{\lambda_{\min}}\right).$$

It can also be easily verified that all noise conditions in Theorem 4.2 are satisfied with above lower bound condition on ε. □

E Technical lemmas

E.1 Tail inequalities

Lemma E.1 (Tail bound of Lipschitz function of Gaussian random variables, [8])**.** *Suppose $\boldsymbol{x}\sim\mathcal{N}_d(0,\sigma^2\mathbf{I}_{d\times d})$ are d-dimensional independent Gaussian random variables and let $f:\mathbb{R}^d\to\mathbb{R}$ be an L-Lipschitz function; that is, $|f(\boldsymbol{x})-f(\boldsymbol{y})|\leq L\|\boldsymbol{x}-\boldsymbol{y}\|_2$ for all $\boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^d$. Suppose $\mu = \mathbb{E}_{\boldsymbol{x}}[f(\boldsymbol{x})]$. Then for all $t>0$, we have that*

$$\Pr\left[\left|f(\boldsymbol{x})-\mu\right|\geq t\right]\leq 2e^{-t^2/(2\sigma^2L^2)}.$$

Lemma E.2 (Bounds on maximum of Gaussian random variables, [19])**.** *Suppose $X_1,\cdots,X_n \overset{i.i.d.}{\sim}\mathcal{N}(0,\sigma^2)$ and let $Y=\max_{1\leq i\leq n}X_i$. We then have that*

$$\frac{\sigma}{\sqrt{\pi\ln 2}}\sqrt{\ln n}\leq\mathbb{E}[Y]\leq\sigma\sqrt{2}\sqrt{\ln n}.$$

Lemma E.3 (Bounds on maximum absolute values of Gaussian random variables; Theorem 3.12, [23])**.** *Suppose $X_1,\cdots,X_n\overset{i.i.d.}{\sim}\mathcal{N}(0,\sigma^2)$ and let $Y=\max_{1\leq i\leq n}|X_i|$. We then have that*

$$\Pr\left[Y\geq\sigma\sqrt{2\ln n}+\sigma\sqrt{2t}\right]\leq e^{-t},\quad\forall t>0.$$

Lemma E.4 (Bounds on Chi-square random variables, [22])**.** *Suppose $X \sim \chi_k^2$; that is, $X = \sum_{j=1}^k Y_j^2$ for i.i.d. standard Normal random variables $Y_1, \cdots, Y_k$. We then have that $\forall t > 0$,*

$$\Pr\left[X \geq k + 2\sqrt{kt} + 2t\right] \leq e^{-t}, \qquad \Pr\left[X \leq k - 2\sqrt{kt}\right] \leq e^{-t}.$$

Lemma E.5 (Bounds on non-central Chi-square random variables, [7])**.** *Suppose $X \sim \chi_k^2(\mu)$; that is, $X = \sum_{j=1}^k Y_k^2$ for independent Normal random variables $Y_1, \cdots, Y_k$ distributed as $Y_j \sim \mathcal{N}(\mu_j, 1)$, $\sum_j \mu_j = \mu$. We then have that*

$$\begin{aligned} \Pr\left[X \geq (k+\mu) + 2\sqrt{(k+2\mu)t} + 2t\right] &\leq e^{-t}, \\ \Pr\left[X \leq (k+\mu) - 2\sqrt{(k+2\mu)t}\right] &\leq e^{-t}. \end{aligned}$$

Lemma E.6 (Bounds on quadratic forms of sub-Gaussian random variables, [15])**.** *Suppose $X \sim \mathcal{SG}_D(\sigma)$ and let $\Sigma \in \mathbb{R}^{D \times D}$ be a positive semidefinite matrix. Then for all $t > 0$ we have that*

$$\Pr\left[X^\top \Sigma X > \sigma^2 \left(\mathrm{tr}(\Sigma) + 2\sqrt{\mathrm{tr}(\Sigma^2)t} + 2\|\Sigma\|t\right)\right] \leq e^{-t}.$$

E.2 Matrix perturbation lemmas

Lemma E.7 (Weyl's theorem; Theorem 4.11, p. 204 in [26])**.** *Let $\mathbf{A}, \mathbf{E}$ be given $m \times n$ matrices with $m \geq n$. Then*

$$\max_{i \in [n]} \left|\sigma_i(\mathbf{A} + \mathbf{E}) - \sigma_i(\mathbf{A})\right| \leq \|\mathbf{E}\|_2.$$

Lemma E.8 (Wedin's theorem; Theorem 4.4, pp. 262 in [26])**.** *Let $\mathbf{A}, \mathbf{E} \in \mathbb{R}^{m \times n}$ be given matrices with $m \geq n$. Let $\mathbf{A}$ have the following singular value decomposition*

$$\begin{bmatrix} \mathbf{U}_1^\top \\ \mathbf{U}_2^\top \\ \mathbf{U}_3^\top \end{bmatrix} \mathbf{A} \begin{bmatrix} \mathbf{V}_1 & \mathbf{V}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{\Sigma}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_2 \\ \mathbf{0} & \mathbf{0} \end{bmatrix},$$

where $\mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3, \mathbf{V}_1, \mathbf{V}_2$ have orthonormal columns and $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ are diagonal matrices. Let $\widetilde{\mathbf{A}} = \mathbf{A} + \mathbf{E}$ be a perturbed version of $\mathbf{A}$ and $(\widetilde{\mathbf{U}}_1, \widetilde{\mathbf{U}}_2, \widetilde{\mathbf{U}}_3, \widetilde{\mathbf{V}}_1, \widetilde{\mathbf{V}}_2, \widetilde{\mathbf{\Sigma}}_1, \widetilde{\mathbf{\Sigma}}_2)$ be analogous singular value decomposition of $\widetilde{\mathbf{A}}$. Let $\mathbf{\Phi}$ be the matrix of canonical angles between Range$(\mathbf{U}_1)$ *and* Range$(\widetilde{\mathbf{U}}_1)$ *and $\mathbf{\Theta}$ be the matrix of canonical angles between* Range$(\mathbf{V}_1)$ *and* Range$(\widetilde{\mathbf{V}}_1)$. *If there exists $\alpha, \delta > 0$ such that*

$$\min_i \sigma_i(\widetilde{\mathbf{\Sigma}}_1) \geq \alpha + \delta \quad \text{and} \quad \max_i \sigma_i(\mathbf{\Sigma}_2) \leq \alpha,$$

then

$$\max\{\|\mathbf{U}_1\mathbf{U}_1^\top - \widetilde{\mathbf{U}}_1\widetilde{\mathbf{U}}_1^\top\|_2, \|\mathbf{U}_1\mathbf{U}_1^\top - \widetilde{\mathbf{V}}_1\widetilde{\mathbf{V}}_1^\top\|_2\} = \max\{\|\sin\mathbf{\Phi}\|_2, \|\sin\mathbf{\Theta}\|_2\} \leq \frac{\|\mathbf{E}\|_2}{\delta}.$$

E.3 Lemmas on random tensors

Lemma E.9 (Spectral norm bound of random tensors, [27])**.** *Suppose $\mathbf{X}$ is a pth order tensor with dimensions $d_1, \cdots, d_p$ and each element of $\mathbf{X}$ is sampled i.i.d. from Gaussian distribution $\mathcal{N}(0, \sigma^2)$. Then the following upper bound on $\|\mathbf{X}\|_{\mathrm{op}}$ holds with probability at least $(1 - \delta)$:*

$$\|\mathbf{X}\|_{\mathrm{op}} \leq \sqrt{8\sigma^2 \left(\left(\sum_{k=1}^p d_p\right) \ln(2K/K_0) + \ln(2/\delta)\right)},$$

where $K_0 = \ln(3/2)$.