# Tensor Decomposition via Simultaneous Power Iteration (Supplementary Material)

**Po-An Wang** [1] **Chi-Jen Lu** [1]

## A. Technical Lemmas

For a matrix $A$, let $\sigma_{\max}(A)$ and $\sigma_{\min}(A)$ denote its largest and smallest singular values, respectively. Then we will need the following lemma relating such singular values of a matrix and its sub-matrix.

**Lemma A.1.** *(Corollary 3.1.3 in (Hom & Johnson, 1991)) Let $A$ and $B$ be matrices such that $B$ is derived from $A$ by deleting some of its rows and/or columns. Then $\sigma_{\max}(A) \geq \sigma_{\max}(B)$ and $\sigma_{\min}(A) \leq \sigma_{\min}(B)$.*

For a matrix $Z$, let $Z^{\odot 2} = Z \odot Z$ denote the Hadamard (entry-wise) product of $Z$ with itself. Then we will need the following lemma relating the singular values of matrices $Z$ and $Z^{\odot 2}$.

**Lemma A.2.** *For any matrix $Z$, $\sigma_{\min}(Z^{\odot 2}) \geq (\sigma_{\min}(Z))^2$ and $\sigma_{\max}(Z^{\odot 2}) \leq (\sigma_{\max}(Z))^2$.*

*Proof.* One can relate the singular values of the Hadamard product $Z^{\odot 2}$ to those of the Kronecker product $Z \otimes Z$. In particular, as $Z \odot Z$ can be obtain from $Z \otimes Z$ by deleting some rows and columns, Lemma A.1 tells us that $\sigma_{\min}(Z \odot Z) \geq \sigma_{\min}(Z \otimes Z)$ and $\sigma_{\max}(Z \odot Z) \leq \sigma_{\max}(Z \otimes Z)$. Then the lemma follows as the Kronecker product $Z \otimes Z$ is known to have the property that $\sigma_{\min}(Z \otimes Z) = (\sigma_{\min}(Z))^2$ and $\sigma_{\max}(Z \otimes Z) = (\sigma_{\max}(Z))^2$.[1] □

We will need the following two tail bounds. The first is for the sum of the squares of independent standard normal random variables, known as the $\chi$-square distribution, which follows from the bound in (Laurent & Massart, 2000).

**Lemma A.3.** *Let $z_1, \dots, z_L$ be a sequence of i.i.d. random variables, each from the distribution $\mathcal{N}(0,1)$. Then for any $\delta \in (0,1)$, we have*

$$\Pr\left[\left|\frac{1}{L}\sum_{i\in[L]} z_i^2 - 1\right| \geq \delta\right] \leq 2^{-\Omega(\delta^2 L)}.$$

The second is the following matrix version of the Bernstein inequality (see e.g. Theorem 1.6 in (Tropp, 2012)).

**Lemma A.4.** *Consider a finite sequence $Z_1, \dots, Z_n$ of independent, random, matrices in $\mathbb{R}^{d\times k}$. Assume that each random matrix satisfies $\mathbb{E}[Z_i] = 0$ and $\|Z_i\| \leq R$ almost surely. Define the variance parameter*

$$\sigma^2 = \max\left\{\left\|\sum_{i=1}^{n}\mathbb{E}[Z_i Z_i^\top]\right\|, \left\|\sum_{i=1}^{n}\mathbb{E}[Z_i^\top Z_i]\right\|\right\}.$$

*Then, for all $t \geq 0$,*

$$\Pr\left[\left\|\sum_{i=1}^{n} Z_i\right\| \geq t\right] \leq (d+k)\cdot 2^{\frac{-t^2}{\sigma^2 + Rt/3}}.$$

We will also need the following two matrix perturbation bounds.

**Lemma A.5.** *(Theorem 2.5 in (Stewart & Sun, 1990)) Let $A, E \in \mathbb{R}^{k\times k}$ be given. If $A$ is invertible, and $\|A^{-1}E\| < 1$ then $\bar{A} := A + E$ is invertible, and*

$$\|\bar{A}^{-1} - A^{-1}\| \leq \frac{\|E\|\,\|A^{-1}\|^2}{1 - \|A^{-1}E\|}.$$

**Lemma A.6.** *(Lemma 2.2 in (Schmitt, 1992)) Given any $A, \bar{A} \in \mathbb{R}^{k\times k}$ with smallest singular values $\sigma > 0$ and $\bar{\sigma} > 0$, respectively, we have*

$$\left\|\bar{A}^{\frac{1}{2}} - A^{\frac{1}{2}}\right\| \leq \frac{\|\bar{A} - A\|}{\bar{\sigma}^{\frac{1}{2}} + \sigma^{\frac{1}{2}}}.$$

## B. Proofs in Section 3

### B.1. Proof of Lemma 1

Recall that $Q^{(t)}$ is derived from $Y^{(t)}$ by the QR decomposition $Y^{(t)} = Q^{(t)} \cdot R^{(t)}$ via the Gram-Schmidt process,

[1] Academia Sinica, Taiwan. Correspondence to: Po-An Wang <poanwang@iis.sinica.edu.tw>.

*Proceedings of the 34th International Conference on Machine Learning*, Sydney, Australia, PMLR 70, 2017. Copyright 2017 by the author(s).

[1] See e.g. Theorem 4.2.12 in (Hom & Johnson, 1991) for the case of square matrices; the extension to rectangular matrices is straightforward.

which has the same effect as performing $k$ copies of the QR decomposition on the $k$ sub-matrices $Y_{[m]}^{(t)}$, for $m \in [k]$, to obtain the $k$ sub-matrices $Q_{[m]}^{(t)}$, for $m \in [k]$.

Let us fix any $m \in [k]$ and $t \geq 0$. To simply our notation, we will drop the indices of $m$ and $t$ in the following. We will write $Q$ for $Q_{[m]}^{(t-1)}$, $Q'$ for $Q_{[m]}^{(t)}$, $Y$ for $Y_{[m]}^{(t)}$, and $\hat{\Phi}$ for $\hat{\Phi}_{[m]}^{(t)}$. We will write $U$ for $U_{[m]}$, with the vector $u_1, \ldots, u_m$ as its columns, while we will use $V$ to denote the $d \times (d-m)$ matrix having the vectors $u_{m+1}, \ldots, u_d$ as its columns. We will write $\tan, \cos, \sin$ for $\tan_m, \cos_m, \sin_m$, respectively. Furthermore, for a matrix $A$, let $\sigma_{\min}(A)$ and $\sigma_{\max}(A)$ denote its smallest and largest singular values, respectively.

Recall that our goal is to bound $\tan(Q')$ in terms of $\tan(Q)$. As discussed before, $Q'$ is derived from $Y$ by a QR decomposition, with $Y = Q'R$ for some matrix $R$. To achieve our goal, we will first show that $R$ is invertible so that $Q' = YR^{-1}$, and then relate $\tan(Q')$ to the singular values $\sigma_{\max}(V^\top Y)$ and $\sigma_{\min}(U^\top Y)$, followed by bounding these two singular values.

First, from the condition (3), we have $\cos Q > 0$ which implies that $Q$ has full rank and consists of orthonormal columns. Our key lemma is the following, which we will prove later in Subsection B.1.1.

**Lemma B.1.** *The following two bounds hold:*

- $\sigma_{\min}\left(U^\top Y\right) \geq \lambda_m \cos^2(Q) - \|\hat{\Phi}\|$, *and*
- $\sigma_{\max}\left(V^\top Y\right) \leq \lambda_{m+1} \sin^2(Q) + \|\hat{\Phi}\|$.

Using this lemma and the assumption $\|\hat{\Phi}\| \leq \triangle \cos^2(Q)$ in (3), we have

$$\sigma_{\min}\left(U^\top Y\right) \geq \lambda_m \cos^2(Q) - \triangle \cos^2(Q) > 0,$$

as $\Delta \leq \frac{\lambda_m}{2}$. This implies that $Y$ has full rank, $R^{-1}$ exists, $Q' = YR^{-1}$ has orthonormal columns, and $(U^\top Q')^{-1}$ exists. Then from standard properties of principal angles (see e.g. (Zhu & Knyazev, 2012)), we know that

$$\tan(Q') = \frac{\sigma_{\max}(V^\top Q')}{\sigma_{\min}(U^\top Q')} = \left\|(V^\top Q')(U^\top Q')^{-1}\right\|,$$

which equals

$$\begin{aligned} \left\|(V^\top Y R^{-1})(U^\top Y R^{-1})^{-1}\right\| &= \left\|(V^\top Y)(U^\top Y)^{-1}\right\| \\ &\leq \frac{\sigma_{\max}(V^\top Y)}{\sigma_{\min}(U^\top Y)}. \end{aligned}$$

Then by Lemma B.1, together with the assumption from (3) that $\|\hat{\Phi}\| \leq \triangle\beta = \triangle\beta \sin^2(Q) + \triangle\beta \cos^2(Q)$, we can bound the denominator by

$$\sigma_{\max}(V^\top Y) \leq (\lambda_{m+1} + \triangle\beta) \sin^2(Q) + \triangle\beta \cos^2(Q),$$

while using the assumption $\|\hat{\Phi}\| \leq \triangle \cos^2(Q)$, we can bound the enumerator by

$$\sigma_{\min}(U^\top Y) \geq \lambda_m \cos^2(Q) - \triangle \cos^2(Q).$$

Combining these bounds together, we obtain

$$\tan(Q') \leq \frac{\lambda_{m+1} + \triangle\beta}{\lambda_m - \triangle} \tan^2(Q) + \frac{\triangle\beta}{\lambda_m - \triangle}.$$

Then the rest of the analysis is identical to that of Hardt & Price (2014) (for the proof of their Lemma 2.2). Specifically, we can rewrite the righthand side above as the weighted average of two terms

$$(1-\alpha) \cdot \frac{\lambda_{m+1} + \triangle\beta}{\lambda_{m+1} + 2\triangle} \tan^2(Q) + \alpha \cdot \beta,$$

with $\alpha = \frac{\triangle}{\lambda_{m+1} + 3\triangle}$, which can be upper-bounded by

$$\max\left\{\frac{\lambda_{m+1} + \triangle\beta}{\lambda_{m+1} + 2\triangle} \tan^2(Q), \beta\right\},$$

and similarly, we can also have

$$\frac{\lambda_{m+1} + \triangle\beta}{\lambda_{m+1} + 2\triangle} \leq \max\left\{\frac{\lambda_{m+1}}{\lambda_{m+1} + \triangle}, \beta\right\}.$$

Since $\frac{\lambda_{m+1}}{\lambda_{m+1} + \triangle} \leq (\frac{\lambda_{m+1}}{\lambda_{m+1} + 4\triangle})^{1/4} = (\frac{\lambda_{m+1}}{\lambda_m})^{1/4} = \rho$, we thus have the desired bound

$$\tan(Q') \leq \max\left\{\max\left\{\rho, \beta\right\} \tan^2(Q), \beta\right\}.$$

To finish the proof, it remains to prove Lemma B.1, which we do next.

### B.1.1. Proof of Lemma B.1

Recall from (2) that for any column $Y_j$ of $Y$ and for any target vector $u_i$,

$$u_i^\top Y_j = \lambda_i \left(u_i^\top Q_j\right)^2 + u_i^\top \hat{\Phi}_j.$$

These equations can be summarized as

$$\begin{aligned} U^\top Y &= \Lambda \left(U^\top Q\right)^{\odot 2} + U^\top \hat{\Phi}, \text{ and} \\ V^\top Y &= \bar{\Lambda} \left(V^\top Q\right)^{\odot 2} + V^\top \hat{\Phi}, \end{aligned}$$

using the notation $\Lambda$ for the $m \times m$ diagonal matrix with $\lambda_1, \ldots, \lambda_m$ at its diagonal, $\bar{\Lambda}$ for the $(d-m) \times (d-m)$ diagonal matrix with $\lambda_{m+1}, \ldots, \lambda_d$ at its diagonal, and $A^{\odot 2} = A \odot A$ for the Hadamard (entry-wise) product of matrix $A$ with itself. From this, we have

$$\begin{aligned} \sigma_{\min}\left(U^\top Y\right) &= \sigma_{\min}\left(\Lambda \left(U^\top Q\right)^{\odot 2} + U^\top \hat{\Phi}\right) \\ &\geq \sigma_{\min}\left(\Lambda \left(U^\top Q\right)^{\odot 2}\right) - \left\|U^\top \hat{\Phi}\right\| \\ &\geq \sigma_{\min}\left(\Lambda\right) \sigma_{\min}\left(\left(U^\top Q\right)^{\odot 2}\right) - \left\|\hat{\Phi}\right\|, \end{aligned}$$

as well as

$$\begin{aligned}\sigma_{\max}\left(V^\top Y\right) &= \sigma_{\max}\left(\bar{\Lambda}\left(V^\top Q\right)^{\odot 2} + V^\top \hat{\Phi}\right)\\ &\le \sigma_{\max}\left(\bar{\Lambda}\left(V^\top Q\right)^{\odot 2}\right) + \left\|V^\top \hat{\Phi}\right\|\\ &\le \sigma_{\max}\left(\bar{\Lambda}\right)\sigma_{\max}\left(\left(V^\top Q\right)^{\odot 2}\right) + \left\|\hat{\Phi}\right\|.\end{aligned}$$

From Lemma A.2, we have

$$\sigma_{\min}\left(\left(U^\top Q\right)^{\odot 2}\right) \ge \left(\sigma_{\min}\left(U^\top Q\right)\right)^2 = \cos^2(Q),$$

since $Q$ has orthonormal columns, and moreover

$$\sigma_{\max}\left(\left(V^\top Q\right)^{\odot 2}\right) \le \left(\sigma_{\max}\left(V^\top Q\right)\right)^2 = \sin^2(Q).$$

As $\sigma_{\min}\left(\bar{\Lambda}\right) = \lambda_m$ and $\sigma_{\max}\left(\Lambda\right) = \lambda_{m+1}$, Lemma B.1 follows.

### B.2. Proof of Theorem 2

Suppose we have $Q^{(0)}$ such that for every $m \in [k]$, $\tan_m(Q^{(0)}) \le 1$ and hence $\cos_m(Q^{(0)}) \ge \frac{1}{\sqrt{2}}$. We would like to apply Lemma 1 repeatedly with $\beta = \frac{\varepsilon}{2}$ to achieve $\tan_m(Q^{(t)}) \le \frac{\varepsilon}{2}$ for every $m$. To be able to do this, we need to verify that for every $t$, the condition (3) in Lemma 1 is satisfied. For this, we first claim that $\|\hat{\Phi}^{(t)}_{[m]}\| \le \frac{\Delta\varepsilon}{2}$. This holds since for any vector $x = (x_1, \dots, x_m)$ of unit length,

$$\left\|\hat{\Phi}^{(t)}_{[m]} x\right\| \le \sum_{j\in[m]} \left\|x_j \Phi(I_d, Q^{(t)}_j, Q^{(t)}_j)\right\| \le \sum_{j\in[m]} |x_j| \cdot \|\Phi\|$$

which by the Cauchy-Schwarz inequality is at most

$$\sqrt{m}\|x\| \cdot \|\Phi\| = \sqrt{m} \cdot \|\Phi\| \le \frac{\Delta\varepsilon}{2}.$$

Moreover, we can assume without loss of generality that $\max\{\frac{\varepsilon}{2}, \rho\} = \rho$ because otherwise, we immediately have $\tan_m(Q^{(1)}) \le \frac{\varepsilon}{2}$ for every $m$, as the condition (3) is satisfied for $t = 1$. Then a simple induction shows that for every $t \ge 1$, the condition (3) in Lemma 1 holds and

$$\begin{aligned}\tan_m(Q^{(t)}) &\le \max\left\{\frac{\varepsilon}{2}, \rho \tan^2_m(Q^{(t-1)})\right\}\\ &\le \max\left\{\frac{\varepsilon}{2}, \rho^{2^t-1}\right\}\end{aligned}$$

for every $m$. Thus, we have $\rho^{2^t-1} \le \frac{\varepsilon}{2}$ and $\tan_m(Q^{(t)}) \le \frac{\varepsilon}{2}$ whenever $t \ge N - 1$, for some

$$N = \mathcal{O}\left(\log \frac{\log \frac{1}{\varepsilon}}{\log \frac{1}{\rho}}\right) = \mathcal{O}\left(\log\left(\frac{1}{\gamma} \log \frac{1}{\varepsilon}\right)\right),$$

by noting that $\log \frac{1}{\rho} \ge \Omega(\log \frac{1}{1-\gamma}) \ge \Omega(\gamma)$.

Next, we show that for any $t \ge N$, each $Q^{(t)}_i$ is close enough to $u_i$. For this, we rely on the following.

**Proposition B.1.** *For any $t \ge N$, we have $u_i^\top Q^{(t)}_i \ge \sqrt{1 - \frac{\varepsilon^2}{2}}$ for every $i \in [k]$.*

*Proof.* Let us fix any $i \in [k]$. In the following, we first show that $(u_i^\top Q^{(t)}_i)^2 \ge 1 - \frac{\varepsilon^2}{2}$ for any $t \ge N - 1$, and then we show that $u_i^\top Q^{(t)}_i \ge 0$ for any $t \ge N$, which together prove the proposition.

First, consider any $t \ge N - 1$, which from the discussion above has $\tan_i(Q^{(t)}) \le \frac{\varepsilon}{2}$. Note that $(u_i^\top Q^{(t)}_i)^2 \ge \cos^2_i(Q^{(t)}) - \sin^2_{i-1}(Q^{(t)})$, because

$$\begin{aligned}\cos^2_i(Q^{(t)}) &\le \left\|u_i^\top Q^{(t)}_{[i]}\right\|^2\\ &= \left\|u_i^\top Q^{(t)}_{[i-1]}\right\|^2 + \left(u_i^\top Q^{(t)}_i\right)^2\\ &\le \sin^2_{i-1}(Q^{(t)}) + \left(u_i^\top Q^{(t)}_i\right)^2.\end{aligned}$$

From this and the fact that $\cos^2_i(Q^{(t)}) = \frac{1}{1+\tan^2_i(Q^{(t)})}$ and $\sin^2_{i-1}(Q^{(t)}) \le \tan^2_{i-1}(Q^{(t)})$, we get

$$\left(u_i^\top Q^{(t)}_i\right)^2 \ge \frac{1}{1+\frac{\varepsilon^2}{4}} - \frac{\varepsilon^2}{4} \ge 1 - \frac{\varepsilon^2}{2}.$$

Next, consider any $t \ge N$ and our goal is to show that $u_i^\top Q^{(t)}_i > 0$. Recall that $Q^{(t)}_i$ is derived from

$$Y^{(t)}_i = \sum_j \lambda_j \left(u_j^\top Q^{(t-1)}_i\right)^2 \cdot u_j + \hat{\Phi}^{(t-1)}_i$$

by subtracting from it its projection to some unit vector $z$ in the column space of $Q^{(t)}_{[i-1]}$ and then scaling it to unit length. Thus, the sign of $u_i^\top Q^{(t)}_i$ is the same as that of

$$\begin{aligned}&u_i^\top\left(Y^{(t)}_i - \left(z^\top Y^{(t)}_i\right) z\right)\\ &= u_i^\top Y^{(t)}_i - \left(z^\top Y^{(t)}_i\right)\left(u_i^\top z\right)\\ &\ge \lambda_i \left(u_i^\top Q^{(t-1)}_i\right)^2 - \left\|\hat{\Phi}^{(t-1)}_i\right\| - \sin_{i-1}(Q^{(t)})\\ &\ge \lambda_i\left(1 - \frac{\varepsilon^2}{2}\right) - \frac{5\varepsilon}{8},\end{aligned}$$

since $(u_i^\top Q^{(t-1)}_i)^2 \ge 1 - \frac{\varepsilon^2}{2}$ for $t-1 \ge N-1$, $\|\hat{\Phi}^{(t-1)}_i\| \le \frac{\Delta\varepsilon}{2} \le \frac{\varepsilon}{8}$, and $\sin_{i-1}(Q^{(t)}) \le \tan_{i-1}(Q^{(t)}) \le \frac{\varepsilon}{2}$ for $t \ge N$. Finally, as $\lambda_i \ge \lambda_k \ge 2\varepsilon$, the last line above is positive, which implies that $u_i^\top Q^{(t)}_i > 0$. □

As $\|Q^{(t)}_i\| = \|u_i\| = 1$, this proposition immediately implies that for any $t \ge N$ and $i \in [k]$,

$$\left\|Q^{(t)}_i - u_i\right\| = \sqrt{2 - 2u_i^\top Q^{(t)}_i} \le \varepsilon,$$

as $u_i^\top Q_i^{(t)} \geq (u_i^\top Q_i^{(t)})^2 \geq 1 - \frac{\varepsilon^2}{2}$.

Finally, let us show that for any $t \geq N$, each $\lambda_i$ can be approximated well by $\hat{\lambda}_i = \bar{T}(Q_i^{(t)}, Q_i^{(t)}, Q_i^{(t)})$. Fix any $t \geq N$ and $i \in [k]$. Note that $|\lambda_i - \hat{\lambda}_i|$ is at most

$$\left|\lambda_i - T\left(Q_i^{(t)}, Q_i^{(t)}, Q_i^{(t)}\right)\right| + \left|\Phi\left(Q_i^{(t)}, Q_i^{(t)}, Q_i^{(t)}\right)\right|.$$

The second term above is at most $\|\Phi\| \|Q_i^{(t)}\|^3 \leq \frac{\varepsilon}{4}$, and the first term above is at most

$$\begin{aligned}
&\left|\lambda_i - \lambda_i \left(u_i^\top Q_i^{(t)}\right)^3\right| + \sum_{j \neq i} \left|\lambda_j \left(u_j^\top Q_i^{(t)}\right)^3\right| \\
\leq \quad & \lambda_i \left(1 - \left(u_i^\top Q_i^{(t)}\right)^3\right) + \sum_{j \neq i} \lambda_j \left(1 - \left(u_i^\top Q_i^{(t)}\right)^2\right) \\
\leq \quad & \lambda_i \frac{3\varepsilon^2}{4} + \sum_{j \neq i} \lambda_j \frac{\varepsilon^2}{2},
\end{aligned}$$

where the first inequality uses the fact that for $j \neq i$, $(u_j^\top Q_i^{(t)})^3 \leq (u_j^\top Q_i^{(t)})^2 \leq \|Q_i^{(t)}\|^2 - (u_i^\top Q_i^{(t)})^2 = 1 - (u_i^\top Q_i^{(t)})^2$, while the second inequality uses the bound $(u_i^\top Q_i^{(t)})^3 \geq (1 - \frac{\varepsilon^2}{2})^{3/2} \geq 1 - \frac{3\varepsilon^2}{4}$. As a result, we have

$$\left|\lambda_i - \hat{\lambda}_i\right| \leq \sum_j \lambda_j \frac{3\varepsilon^2}{4} + \frac{\varepsilon}{4} \leq \varepsilon,$$

using the assumption $\sum_j \lambda_j \leq 1$ from (1). This completes the proof of the theorem.

### B.3. Proof of Lemma 2

From the definition of $\bar{w}$, we can express $u_i^\top \bar{w}$ as

$$\frac{1}{L} \sum_{j \in [L]} T(u_i, w_j, w_j) + \frac{1}{L} \sum_{j \in [L]} \Phi(u_i, w_j, w_j). \tag{1}$$

The first term above equals

$$\frac{1}{L} \sum_{j \in [L]} \lambda_i \left(u_i^\top w_j\right)^2 = \lambda_i \cdot \frac{1}{L} \sum_{j \in [L]} \left(u_i^\top w_j\right)^2,$$

and note that the sum has a $\chi$-square distribution because each $u_i^\top w_j$ is an independent random variable with the standard normal distribution $\mathcal{N}(0,1)$.[2] Then from Lemma A.3, we know that for $\delta = \frac{\gamma}{4}$, there exists some $L \leq \mathcal{O}(\frac{1}{\gamma^2} \log d)$ such that the first term in (1) differs from $\lambda_i$ by at most $\frac{\lambda_i \gamma}{4}$ with probability at least $1 - \frac{1}{200d^2}$.

The second term in (1) can be bounded in a similar way as follows. Since $\|u_i\| = 1$ and $\Phi$ is symmetric, we know that the matrix $\Phi(u_i, I_d, I_d)$ has norm $\|\Phi(u_i, I_d, I_d)\| \leq \|\Phi\| \leq \frac{\Delta}{3d}$ and can be decomposed as $\sum_{r \in [d]} \tilde{\lambda}_r \cdot \tilde{u}_r \otimes \tilde{u}_r$, for some orthonormal vectors $\tilde{u}_r$'s as well as some values $\tilde{\lambda}_r$'s, each with $|\tilde{\lambda}_r| \leq \frac{\Delta}{3d}$. Then by a similar analysis as above, together with a union bound, we can have with probability at least $1 - d \cdot \frac{1}{200d^2} = 1 - \frac{1}{200d}$ that

$$\begin{aligned}
\left|\frac{1}{L} \sum_{j \in [L]} \Phi(u_i, w_j, w_j)\right| \quad & \leq \quad \sum_{r \in [d]} \left|\tilde{\lambda}_r\right| \cdot \frac{1}{L} \sum_{j \in [L]} \left(\tilde{u}_r^\top w_j\right)^2 \\
& \leq \quad \sum_{r \in [d]} \frac{\Delta}{3d} \cdot \left(1 + \frac{\gamma}{4}\right) \\
& \leq \quad \frac{\Delta}{2}.
\end{aligned}$$

By combining the two bounds above, we can conclude that for any $i \in [d]$, the sum in (1) differs from $\lambda_i$ by at most $\frac{1}{4}(\lambda_i \gamma + 2\Delta)$ with probability at least $1 - \frac{1}{100d}$. Then the lemma immediately follows by a union bound.

### B.4. Proof of Lemma 3

Consider any $\bar{w}$ satisfying the condition (5) in Lemma 2. By definition, $\bar{M} = \bar{T}(I_d, I_d, \bar{w})$ can be decomposed as

$$T(I_d, I_d, \bar{w}) + \Phi(I_d, I_d, \bar{w}).$$

The first matrix can be expressed as

$$T(I_d, I_d, \bar{w}) = \sum_{i \in [d]} \lambda_i \left(u_i^\top \bar{w}\right) \cdot u_i \otimes u_i,$$

with $\bar{\lambda}_i = \lambda_i (u_i^\top \bar{w})$ and $u_i$ as its $i$'th eigenvalue and eigenvector, respectively. Note that as $\Delta \leq \frac{\lambda_i - \lambda_{i+1}}{4}$, we have

$$\begin{aligned}
\bar{\lambda}_i \quad & \geq \quad \lambda_i^2 - \frac{1}{4}\left(\lambda_i^2 \gamma + 2\lambda_i \Delta\right) \\
& \geq \quad \lambda_i^2 - \frac{\lambda_i^2 - \lambda_{i+1}^2}{4} - \frac{\lambda_i^2 - \lambda_i \lambda_{i+1}}{8} \\
& \geq \quad \lambda_i^2 - \frac{3\left(\lambda_i^2 - \lambda_{i+1}^2\right)}{8},
\end{aligned}$$

as well as

$$\begin{aligned}
\bar{\lambda}_{i+1} \quad & \leq \quad \lambda_{i+1}^2 + \frac{1}{4}\left(\lambda_{i+1}^2 \gamma + 2\lambda_{i+1} \Delta\right) \\
& \leq \quad \lambda_{i+1}^2 + \lambda_{i+1}^2 \frac{\lambda_i^2 - \lambda_{i+1}^2}{4\lambda_i^2} + \frac{\lambda_{i+1}\lambda_i - \lambda_{i+1}^2}{8} \\
& \leq \quad \lambda_{i+1}^2 + \frac{3\left(\lambda_i^2 - \lambda_{i+1}^2\right)}{8},
\end{aligned}$$

which together imply that

$$\bar{\lambda}_i - \bar{\lambda}_{i+1} \geq \frac{\lambda_i^2 - \lambda_{i+1}^2}{4} \geq \Delta^2.$$

[2] This is because each component of $w_j$ has the distribution $\mathcal{N}(0,1)$, and the distribution of $u_i^\top w_j$ has mean $\sum_r u_{i,r} \cdot 0 = 0$ and variance $\sum_r u_{i,r}^2 \cdot 1 = 1$, where $u_{i,r}$ denotes the $r$'th component of $u_i$.

It remains to bound the norm of $\Phi(I_d, I_d, \bar{w})$, which is

$$\|\Phi(I_d, I_d, \bar{w})\| \le \|\Phi\| \cdot \|\bar{w}\|,$$

where $\|\bar{w}\|^2 = \sum_{i\in[d]} \left(u_i^\top \bar{w}\right)^2$ is at most

$$\sum_{i\in[d]} \left(\lambda_i + \frac{1}{4}\left(\lambda_i \gamma + 2\Delta\right)\right)^2 \le \sum_{i\in[d]} \left(2\lambda_i\right)^2 \le 4.$$

This implies that $\|\Phi(I_d, I_d, \bar{w})\| \le 2\|\Phi\|$, which completes the proof of the lemma.

### B.5. Proof of Lemma 4

Suppose we have a matrix $\bar{M} = M + \bar{\Phi}$, where

$$M = \sum_{i\in[d]} \bar{\lambda}_i \cdot u_i \otimes u_i$$

with $\bar{\lambda}_i - \bar{\lambda}_{i+1} \ge \Delta^2$ for every $i \in [k]$ and $\|\bar{\Phi}\| \le \frac{\alpha_1 \Delta^2}{\sqrt{dk}}$ for a small enough constant $\alpha_1$. The key observation is that although we run one copy of the matrix power method of Hardt & Price (2014) to update the whole $d\times k$ matrix $Z^{(s)}$, we can actually see our algorithm as running $k$ copies of the matrix power method on $k$ sub-matrices $Z^{(s)}_{[1]}, \dots, Z^{(s)}_{[k]}$ simultaneously. This allows us to apply their analysis immediately.

More precisely, although our QR decomposition at each step $s$ is applied to the whole $d \times k$ matrix $Y^{(s)}$ to obtain our $d \times k$ matrix $Z^{(s)}$, the Gram-Schmidt process we use has the effect that each $d \times m$ sub-matrix $Z^{(s)}_{[m]}$ can also be seen as obtained from the $d \times m$ matrix $Y^{(s)}_{[m]}$ by a QR decomposition. Thus, our algorithm can be seen as running $k$ copies of the algorithm of (Hardt & Price, 2014) simultaneously, and we can apply the following lemma of theirs[3] simultaneously for every $m \in [k]$ with $X^{(s)}$ being our $d\times k$ matrix $Z^{(s)}_{[m]}$.

**Lemma B.2.** *Fix any $m \in [k]$. Suppose that the initial $X^{(0)}$ and the noise $G^{(s)}_m = \bar{\Phi} \cdot X^{(s)}$ at each step $s$ is such that*

$$\begin{aligned} 5\left\|U_{[m]}^\top G^{(s)}_m\right\| &\le \left(\bar{\lambda}_m - \bar{\lambda}_{m+1}\right) \cos_m(X^{(0)}) \\ 5\left\|G^{(s)}_m\right\| &\le \left(\bar{\lambda}_m - \bar{\lambda}_{m+1}\right) \epsilon \end{aligned}$$

*for some $\epsilon < \frac{1}{2}$. Then for $\gamma_m = 1 - \frac{\bar{\lambda}_{m+1}}{\bar{\lambda}_m}$, there exists some $S = \mathcal{O}(\frac{1}{\gamma_m} \log \frac{\tan_m(X^{(0)})}{\epsilon})$ such that for any $t \ge S$ we have $\tan_m(X^{(t)}) \le \epsilon$.*

It remains to show that we can have an initial $Z^{(0)}$, such that for each $m \in [k]$, $Z^{(0)}_{[m]}$ satisfies the condition required by Lemma B.2. For this, we need the following bound from (Mitliagkas et al., 2013).

**Proposition B.2.** *For any $\delta$, we have*

$$\Pr\left[\cos_k(Z^{(0)}) \le \frac{\delta}{\sqrt{dk}}\right] \le \mathcal{O}(\delta) + 2^{-\Omega(d)}.$$

By applying this proposition, with $\delta = 10\alpha_0$ for a small enough constant $\alpha_0$, we can have $\cos_k(Z^{(0)}) \ge \frac{10\alpha_0}{\sqrt{dk}}$ with high probability. From Lemma A.1, we know that for any $m \in [k]$, $\cos_m(Z^{(0)}) = \sigma_{\min}(U^\top_{[m]} Z^{(0)}_{[m]}) \ge \sigma_{\min}(U^\top Z^{(0)}) = \cos_k(Z^{(0)})$. Thus, with high probability we in fact have $\cos_m(Z^{(0)}) \ge \frac{10\alpha_0}{\sqrt{dk}}$ for every $m \in [k]$. Given such an initial $Z^{(0)}$, we can have for every $s$ and $m$ that

$$5\|G^{(s)}_m\| \le 5\|\bar{\Phi}\| \le \frac{10\alpha_0 \Delta^2}{\sqrt{dk}}$$

which satisfies the two conditions needed by Lemma B.2, with $\epsilon = \frac{1}{3}$. Then we can repeatedly apply Lemma B.2, simultaneously for every $m \in [k]$, and a simple induction shows that for some $S = \mathcal{O}(\frac{1}{\gamma} \log d)$, we have $\tan_m(Z^{(s)}) \le \epsilon < 1$ for any $m \in [k]$ and $s \ge S$. This completes the proof of our Lemma 4

## C. Proofs in Section 5

### C.1. Proof of Lemma 5

First, we claim that $\tan_k(Q) < 1$ with high probability. To show this, note that by Proposition B.2, we have with high probability that $\cos_k(Z) > \frac{4\alpha_0}{\sqrt{dk}}$ for a small enough constant $\alpha_0$. In the following, let us assume that we indeed have such a matrix $Z$, and note that it has $\tan_k(Z) < \frac{\sqrt{dk}}{4\alpha_0}$. Then we need the following.

**Lemma C.1.** *(Lemma 2.2 in (Hardt & Price, 2014)) Let $Z, G \in \mathbb{R}^{d\times k}$ satisfy*

$$\begin{aligned} 4\left\|U^\top G\right\| &\le \left(\lambda_k - \lambda_{k+1}\right) \cos_k(Z) \\ 4\left\|G\right\| &\le \left(\lambda_k - \lambda_{k+1}\right) \beta \end{aligned}$$

*for some $\beta < 1$. Then for $\rho = \left(\frac{\lambda_{k+1}}{\lambda_k}\right)^{1/4}$, we have $\tan_k(MZ + G) \le \max\{\beta, \max\{\beta, \rho\} \tan_k(Z)\}$.*

Recall that we assume $\lambda_{k+1} = 0$, and to apply the lemma, let $\beta = \frac{4\alpha_0}{\sqrt{dk}}$ and $G = \bar{\Phi} Z$. Note that

$$\|G\| \le \left\|\bar{\Phi}\right\| \le \frac{\lambda_k \beta}{4},$$

which satisfies both requirements of the lemma, and thus with $\bar{Y} = MZ + G$, we have

$$\tan_k(Q) = \tan_k(\bar{Y}) \le \beta \tan_k(Z) < 1.$$

[3] It corresponds to Theorem 2.3 in (Hardt & Price, 2014). Although it is stated there for $m = k$, it in fact works for any value of $k$ and hence $m$.

Next, let us bound $\sigma_{\min}(P)$ and $\sigma_{\max}(P)$. Recall that

$$P = Q^\top M Q = Q^\top U \Lambda U^\top Q,$$

which implies that

$$\begin{aligned} \sigma_{\min}(P) &\geq \sigma_{\min}\left(Q^\top U\right)\sigma_{\min}\left(\Lambda\right)\sigma_{\min}\left(U^\top Q\right) \text{ and} \\ \sigma_{\max}(P) &\leq \sigma_{\max}\left(Q^\top U\right)\sigma_{\max}\left(\Lambda\right)\sigma_{\max}\left(U^\top Q\right). \end{aligned}$$

Since the matrix $Q$ has orthonormal columns, we have

$$(\sigma_{\min}\left(U^\top Q\right))^2 = (\cos_k(Q))^2 = \frac{1}{1+\tan_k^2(Q)} \geq \frac{1}{2},$$

as well as

$$\sigma_{\max}\left(U^\top Q\right) \leq \|U\|\,\|Q\| = 1.$$

Finally, as $\sigma_{\max}\left(\Lambda\right) = \lambda_1$ and $\sigma_{\min}\left(\Lambda\right) = \lambda_k$, we have

$$\sigma_{\min}(P) \geq \frac{\lambda_k}{2} \text{ and } \sigma_{\max}(P) \leq \lambda_1.$$

### C.2. Proof of Lemma 6

First, from the definition, we have

$$\left\|\bar{P} - P\right\| = \left\|Q^\top \bar{M} Q - Q^\top M Q\right\| \leq \|Q\|^2 \left\|\bar{\Phi}\right\| \leq \epsilon$$

as $Q$ has orthonormal columns so that $\|Q\|^2 \leq 1$. Therefore, given the assumption that $0 < \epsilon \leq \frac{\sigma_{\min}(P)}{2}$, we have

$$\sigma_{\min}(\bar{P}) \geq \sigma_{\min}(P) - \left\|\bar{P} - P\right\| > 0,$$

which implies that $\bar{P}$ is invertible.

Then according to Lemma A.5, we have

$$\begin{aligned} \left\|\bar{P}^{-1} - P^{-1}\right\| &\leq \frac{\left\|\bar{P} - P\right\| \left\|P^{-1}\right\|^2}{1 - \left\|\bar{P} - P\right\| \left\|P^{-1}\right\|} \\ &\leq 2\epsilon(\sigma_{\min}(P))^{-2}, \end{aligned}$$

as $\left\|P^{-1}\right\| = (\sigma_{\min}(P))^{-1}$ and $\left\|\bar{P} - P\right\| \leq \epsilon \leq \frac{\sigma_{\min}(P)}{2}$. Combining this with Lemma A.6, we obtain

$$\begin{aligned} \left\|\bar{P}^{-\frac{1}{2}} - P^{-\frac{1}{2}}\right\| &\leq \frac{\left\|\bar{P}^{-1} - P^{-1}\right\|}{(\sigma_{\min}(\bar{P}^{-1}))^{\frac{1}{2}} + (\sigma_{\min}(P^{-1}))^{\frac{1}{2}}} \\ &\leq 2\epsilon(\sigma_{\min}(P))^{-2}(\sigma_{\max}(P))^{\frac{1}{2}}, \end{aligned}$$

since $\sigma_{\min}(\bar{P}^{-1}) \geq 0$ and $\sigma_{\min}(P^{-1}) = (\sigma_{\max}(P))^{-1}$.

### C.3. Proof of Theorem 3

First, given $\varepsilon \in (0, \frac{1}{2})$ and $\|\bar{\Phi}\| \leq \alpha_0 \varepsilon \min\{\frac{\lambda_k}{\sqrt{dk}}, \frac{\lambda_k^3}{\sqrt{\lambda_1}}\}$, for a small enough constant $\alpha_0$, we know from Lemma 5 and Lemma 6 that with high probability,

- $\sigma_{\max}(P) \leq \lambda_1$,
- $\sigma_{\min}(P) \geq \frac{\lambda_k}{2}$, and
- $\|\bar{P}^{-\frac{1}{2}} - P^{-\frac{1}{2}}\| \leq \frac{\lambda_k \varepsilon}{64}$.

Assume from now on that the above three conditions hold. Next, observe that

$$\begin{aligned} &\left\|\bar{T}\left(\bar{W}, \bar{W}, \bar{W}\right) - T\left(W, W, W\right)\right\| \\ \leq\ &\left\|\bar{T}\left(\bar{W}, \bar{W}, \bar{W}\right) - T\left(\bar{W}, \bar{W}, \bar{W}\right)\right\| + \quad (2) \\ &\left\|T\left(\bar{W}, \bar{W}, \bar{W}\right) - T\left(W, \bar{W}, \bar{W}\right)\right\| + \quad (3) \\ &\left\|T\left(W, \bar{W}, \bar{W}\right) - T\left(W, W, \bar{W}\right)\right\| + \quad (4) \\ &\left\|T\left(W, W, \bar{W}\right) - T\left(W, W, W\right)\right\|. \quad (5) \end{aligned}$$

The term in (2) is at most

$$\left\|\Phi\left(\bar{W}, \bar{W}, \bar{W}\right)\right\| \leq \|\Phi\| \left\|\bar{W}\right\|^3 \leq \frac{\varepsilon}{4}$$

as $\|\Phi\| \leq \alpha_0 \lambda_k^{\frac{3}{2}} \varepsilon$ for a small enough constant $\alpha_0$ and

$$\left\|\bar{W}\right\| \leq \|Q\| \left\|\bar{P}^{-\frac{1}{2}}\right\| \leq \left\|\bar{P}^{-\frac{1}{2}}\right\|,$$

which can be upper-bounded by

$$\left\|P^{-\frac{1}{2}}\right\| + \left\|\bar{P}^{-\frac{1}{2}} - P^{-\frac{1}{2}}\right\| \leq 4\lambda_k^{-\frac{1}{2}}.$$

The term in (3) is at most

$$\begin{aligned} \left\|T\left(\bar{W} - W, \bar{W}, \bar{W}\right)\right\| &\leq \|T\| \left\|\bar{W} - W\right\| \left\|\bar{W}\right\|^2 \\ &\leq \left\|\bar{P}^{-\frac{1}{2}} - P^{-\frac{1}{2}}\right\| 16\lambda_k^{-1} \\ &\leq \frac{\varepsilon}{4}. \end{aligned}$$

Similarly, the term in (4) can be upper-bounded by

$$\|T\| \left\|\bar{W} - W\right\| \|W\| \left\|\bar{W}\right\| \leq \frac{\varepsilon}{4}$$

and the term in (5) can be upper-bounded by

$$\|T\| \left\|\bar{W} - W\right\| \|W\|^2 \leq \frac{\varepsilon}{4}.$$

As a result, we can conclude that

$$\left\|\bar{T}\left(\bar{W}, \bar{W}, \bar{W}\right) - T\left(W, W, W\right)\right\| \leq \varepsilon$$

with high probability, which proves the theorem.

## D. Proofs in Section 6

Our streaming algorithm for orthogonal tensors with $g$ of the form $g(x) = x \otimes x \otimes x$ is summarized in Algorithm 2. We will use the parameters

$$L = \frac{c_0 \log k}{\Delta^2}, S = \frac{c_0 \log k}{\gamma}, N = c_0 \log\left(\frac{1}{\gamma} \log \frac{1}{\varepsilon}\right)$$

**Algorithm 2** Streaming robust tensor power method

```
Input: a stream of data $\{x_1, x_2, \dots, \}$, parameters $L, S, N$, index sets $\{B_s\}_{s=1}^{S}, \{J_t\}_{t=1}^{N}$.
Initialization Phase
Let $\bar{w} = \mathbf{0} \in \mathbb{R}^d$.
for $\tau = 1$ to $L$ do
    Update $\bar{w} = \bar{w} + \frac{1}{L} x_\tau$.
end for
Sample $Y_1^{(0)}, \dots, Y_k^{(0)} \sim \mathcal{N}^d(0, 1)$.
Factorize $Y^{(0)}$ as $Z^{(0)} R^{(0)}$ by QR decomposition.
for $s = 1$ to $S$ do
    for $\tau \in B_s$ do
        Update $Y^{(s)} = Y^{(s)} + \frac{1}{|B_s|} \left( x_\tau^\top \bar{w} \right) x_\tau x_\tau^\top Z^{(s-1)}$.
    end for
    Factorize $Y^{(s)}$ as $Z^{(s)} R^{(s)}$ by QR decomposition.
end for
Tensor power phase
Let $Q^{(0)} = Z^{(S)}$.
for $t = 1$ to $N$ do
    for $\tau \in J_t$ do
        Update $Y_i^{(t)} = Y_i^{(t)} + \frac{1}{|J_t|} x_\tau \left( x_\tau^\top Q_i^{(t)} \right)^2, \forall i \in [k]$.
        Update $\lambda_i^{(t)} = \lambda_i^{(t)} + \frac{1}{|J_t|} \left( x_\tau^\top Q_i^{(t)} \right)^3, \forall i \in [k]$.
    end for
    Factorize $Y^{(t)}$ as $Q^{(t)} R^{(t)}$ by QR decomposition.
end for
Output: $\hat{u}_i = Q_i^{(N)}$ and $\hat{\lambda}_i = \lambda_i^{(N)}, \forall i \in [k]$
```

for a large enough constant $c_0$. Moreover, we partition the time steps into consecutive blocks: with the first block $[L]$ for finding the vector $\bar{w}$, the next $S$ blocks $B_1, \dots, B_S$ for the matrix power method in the initialization phase, followed by $N$ blocks $J_1, \dots, J_N$ for the tensor power phase, with their sizes $|B_s|$ and $|J_t|$ given in (6) and (6) respectively. The proofs of related lemmas in Section 6 are given next.

### D.1. Proof of Lemma 7

First, from the assumption that $T = \mathbb{E}_x[x \otimes x \otimes x]$, where $\mathbb{E}_x[\cdot]$ denotes the expectation over the distribution of $x$, we have the following.

**Proposition D.1.** $\mathbb{E}_x[\|x\|^2 x] = \sum_{i \in [d]} \lambda_i u_i$.

*Proof.* Recall that if we sample $w$ according to the distribution $\mathcal{N}^d(0, 1)$, then for any $u \in \mathbb{R}^d$, we have $\mathbb{E}_w[(u^\top w)^2] = \|u\|^2$, where $\mathbb{E}_w[\cdot]$ denotes the expectation over $w$. Then we have

$$\mathbb{E}_w[T(I_d, w, w)] = \sum_{i \in [d]} \lambda_i \mathbb{E}_w \left[ \left( u_i^\top w \right)^2 \right] u_i = \sum_{i \in [d]} \lambda_i u_i,$$

as $\|u_i\|^2 = 1$. On the other hand, from the assumption that $T = \mathbb{E}_x[x \otimes x \otimes x]$, we also have

$$\begin{aligned} \mathbb{E}_w[T(I_d, w, w)] &= \mathbb{E}_w \left[ \mathbb{E}_x \left[ \left( x^\top w \right)^2 x \right] \right] \\ &= \mathbb{E}_x \left[ \mathbb{E}_w \left[ \left( x^\top w \right)^2 x \right] \right] \\ &= \mathbb{E}_x \left[ \|x\|^2 x \right]. \end{aligned}$$

The proposition follows by combining these two equalities. □

This suggests us to take $\bar{w} = \frac{1}{L} \sum_{\tau=1}^{L} (\|x_\tau\|^2 x_\tau)$, for some $L$ to be determined next. This is because for any $i \in [k]$, the random variable $z_\tau = u_i^\top (\|x_\tau\|^2 x_\tau)$ falls in $[-1, 1]$ and has expected value

$$\mathbb{E}[z_\tau] = u_i^\top \sum_{i \in [d]} \lambda_i u_i = \lambda_i$$

for each $\tau$, so that for $\delta = \frac{1}{4}(\lambda_i \gamma + 2\Delta)$,

$$\Pr\left[ \left| u_i^\top \bar{w} - \lambda_i \right| > \delta \right] = \Pr\left[ \left| \frac{1}{L} \sum_{\tau=1}^{L} z_\tau - \lambda_i \right| > \delta \right]$$

which by Hoeffding inequality is at most

$$2^{-\Omega(\delta^2 L)} \leq \frac{1}{100k}$$

for some $L = \mathcal{O}(\frac{1}{\delta^2}) = \mathcal{O}(\frac{1}{\Delta^2} \log k)$. Then by a union bound, with probability 0.99 we have $\bar{w}$ satisfying $|u_i^\top \bar{w} - \lambda_i| \leq \delta$ for every $i \in [k]$. As $\bar{w}$ can clearly be computed in $\mathcal{O}(d)$ space, the lemma follows.

### D.2. Proof of Lemma 8

The streaming algorithm for this lemma can be found in the initialization phase of our Algorithm 2, which is based on that of Li et al. (2016).

Recall that Li et al. (2016) considered the matrix case, in which each vector $x_\tau$ in the stream has the expectation $\mathbb{E}[x_\tau \otimes x_\tau] = M$ for some $d \times d$ matrix $M$ to be decomposed. To apply their result, let us make the connection by seeing $T(I_d, I_d, \bar{w})$ as their matrix $M$ and $M_\tau = g(x_\tau)(I_d, I_d, \bar{w}) = (x_\tau^\top \bar{w}) \cdot x_\tau \otimes x_\tau$ as their estimator $x_\tau \otimes x_\tau$, by noting that

$$\mathbb{E}[M_\tau] = \mathbb{E}[g(x_\tau)(I_d, I_d, \bar{w})] = \mathbb{E}[g(x_\tau)](I_d, I_d, \bar{w}) = M.$$

Since $\|\bar{w}\| \leq 1$, $\|M_\tau\| \leq \|\bar{w}\| \|x_\tau\|^3 \leq 1$, and $\|M\| \leq \|T\| \|\bar{w}\| \leq 1$, we have

$$\|M_\tau - M\| \leq \|M_\tau\| + \|M\| \leq 2.$$

Thus, we have from the matrix Bernstein inequality (Lemma A.4) that

$$\Pr\left[ \left\| \frac{1}{|B|} \sum_{\tau \in B} M_\tau - M \right\| \geq \delta \right] \leq 2d 2^{-\Omega(\delta^2 |B|)},$$

for any block $B$ of time steps, and this allows us to apply the analysis of (Li et al., 2016).

Following (Li et al., 2016), we use the parameters

$$\varepsilon_s = \varepsilon_0 \rho^s \text{ and } \beta_s = \min\left\{\rho/\sqrt{1+\varepsilon_{s-1}^2}, \rho\varepsilon_{s-1}\right\},$$

with $\varepsilon_0 = \frac{\sqrt{dk}}{\alpha_0}$ for a small enough constant $\alpha_0$, and divide the time steps into $S = \mathcal{O}(\frac{1}{\gamma}\log d)$ blocks, with the $s$'th block $B_s$ having size

$$|B_s| \le \frac{c_0 \log(ds)}{\Delta^4 \beta_s^2}, \tag{6}$$

for a large enough constant $c_0$. Then according to the analysis in (Li et al., 2016) together with that in our proof of Lemma 4, one can show that $\tan_m(Z^{(s)}) \le \varepsilon_s$ for every $s \le S$, so that we can have $\tan_m(Z^{(S)}) \le 1$, for every $m \in [k]$. Moreover, from the analysis in (Li et al., 2016), we know that the number of samples needed can be bounded by

$$\sum_{s=1}^{S} |B_s| \le \mathcal{O}\left(\frac{\varepsilon_0^2 \log(dS)}{\Delta^4 \gamma}\right) = \mathcal{O}\left(\frac{dk \log(dS)}{\Delta^4 \gamma}\right).$$

Finally, note that for each update, the matrix product $M_\tau Z^{(s-1)}$ equals $(x_\tau^\top \bar{w}) x_\tau x_\tau^\top Z^{(s-1)}$, which can be computed in $\mathcal{O}(kd)$ space. Thus, the algorithm works in $\mathcal{O}(kd)$ space, and the lemma follows.

### D.3. Proof of Theorem 4

According to Lemma 7 and Lemma 8, let us assume that we have obtained some $Z \in \mathbb{R}^{d\times k}$ such that $\tan_m(Z) < 1$ for every $m \in [k]$. Now let us focus on the tensor power phase.

Consider a fixed iteration $t$. We would like to show that $\tan_m(Q^{(t)}) \le \beta_t$ with high probability, using Lemma 1. For this, we need to show that the condition (3) there is satisfied with high probability. For $j \in [k]$, let $q_j$ denote $Q_j^{(t-1)}$, and recall that $\hat{\Phi}_j^{(t)} = \Phi(I_d, q_j, q_j)$, which now equals

$$\frac{1}{|J_t|} \sum_{\tau \in J_t} \left(x_\tau^\top q_j\right)^2 x_\tau - T\left(I_d, q_j, q_j\right).$$

Let $\hat{\Phi}^{(t)}$ be the $d \times k$ matrix with $\hat{\Phi}_j^{(t)}$ as its $j$'th column. Then we have the following.

**Lemma D.1.** *$\|\hat{\Phi}^{(t)}\| > \frac{\Delta\beta_t}{2}$ with probability at most $\frac{1}{200t^2}$.*

*Proof.* Let us see $\hat{\Phi}^{(t)}$ as the average of $|J_t|$ i.i.d. random matrices, so that we can apply the matrix Bernstein inequality (Lemma A.4). More precisely, for $\tau \in J_t$, let $A_\tau$ denote the $d \times k$ matrix with

$$\left(x_\tau^\top q_j\right)^2 x_\tau - T\left(I_d, q_j, q_j\right)$$

as its $j$'th column, so that $\hat{\Phi}^{(t)} = \frac{1}{|J_t|}\sum_{\tau \in J_t} A_\tau$. Note that we have $\mathbb{E}[A_\tau] = 0$ for each $\tau$, because

$$\begin{aligned} \mathbb{E}\left[\left(x_\tau^\top q_j\right)^2 x_\tau\right] &= \mathbb{E}\left[\left(x_\tau \otimes x_\tau \otimes x_\tau\right)\left(I_d, q_j, q_j\right)\right] \\ &= \left(\mathbb{E}\left[x_\tau \otimes x_\tau \otimes x_\tau\right]\right)\left(I_d, q_j, q_j\right) \\ &= T\left(I_d, q_j, q_j\right). \end{aligned}$$

Moreover, we claim that $\|A_\tau\| \le 2$. This is because for any $v = (v_1, \ldots, v_k) \in \mathbb{R}^k$ with $\|v\| = 1$, $\|A_\tau v\|$ is at most

$$\left\|\sum_{j\in[k]} v_j \left(x_\tau^\top q_j\right)^2 x_\tau\right\| + \left\|\sum_{j\in[k]} v_j T\left(I_d, q_j, q_j\right)\right\|$$

where the first term above is at most

$$\sum_{j\in[k]} |v_j| \left(x_\tau^\top q_j\right)^2 \|x_\tau\| \le \sum_{j\in[k]} \left(x_\tau^\top q_j\right)^2 \le \|x_\tau\| \le 1$$

as the $q_j$'s are orthonormal, while the second term above is

$$\left\|\sum_{j\in[k]} v_j \sum_{i\in[d]} \lambda_i \left(u_i^\top q_j\right)^2 u_i\right\|$$

which can be upper-bounded by

$$\sum_{i\in[d]} \lambda_i \left\|\sum_{j\in[k]} v_j \left(u_i^\top q_j\right)^2 u_i\right\| \le \sum_{i\in[d]} \lambda_i \le 1$$

using a similar argument and the assumption that $\sum_{i\in[d]} \lambda_i \le 1$. Then we can apply Lemma A.4, and conclude that

$$\Pr\left[\left\|\frac{1}{|J_t|}\sum_{\tau\in J_t} A_\tau\right\| > \frac{\Delta\beta_t}{2}\right] \le \frac{1}{200t^2}$$

for our choice of $|J_t|$. □

Note that $\|\hat{\Phi}_{[m]}^{(t)}\| \le \|\hat{\Phi}^{(t)}\|$ for any $m \in [k]$, and recall that we start with $\cos_m^2(Q^{(0)}) = \frac{1}{\tan_m^2(Q^{(0)})+1} \ge \frac{1}{2}$. Therefore, given this lemma, we can then apply Lemma 1 repeatedly and a simple induction shows that the probability that $\tan_m(Q^{(t)}) > \beta_t$ for some $m$ and $t$ is at most $\sum_t \frac{1}{200t^2} \le 0.01$. Thus, with high probability we have $\tan_m(Q^{(t)}) \le \beta_t$ for every $m$ and $t$. Let $N$ be the number such that $\beta_t > \frac{\varepsilon}{2}$ for $t \le N-2$ and $\beta = \frac{\varepsilon}{2}$ for $t \ge N-1$. Note that

$$N = \mathcal{O}\left(\log \frac{\log\frac{1}{\varepsilon}}{\log\frac{1}{\rho}}\right) = \mathcal{O}\left(\log\left(\frac{1}{\gamma}\log\frac{1}{\varepsilon}\right)\right),$$

and recall from the proof of Theorem 2 that from $Q^{(N)}$ we can obtain the required $\hat{u}_i$'s and $\hat{\lambda}_i$'s.

It remains to bound the number of samples needed in this phase, which is

$$\sum_{t=1}^{N} |J_t| \leq \mathcal{O}\left(\frac{\log(dN)}{\Delta^2}\right) \sum_{t=1}^{N} \frac{1}{\beta_t^2}.$$

As $\beta_t = \max\{\rho^{2^t-1}, \frac{\varepsilon}{2}\}$, we have $\beta_t = \frac{\varepsilon}{2}$ for $t \geq N-1$ and $\beta_t = \beta_{N-2}\rho^{2^t-2^{N-2}} \geq \frac{\varepsilon}{2}\rho^{2^t-2^{N-2}}$ for $t \leq N-2$. Therefore,

$$\sum_{t=1}^{N} \frac{1}{\beta_t^2} \leq \frac{8}{\varepsilon^2} + \frac{4}{\varepsilon^2} \sum_{t=1}^{N-2} \rho^{2(2^{N-2}-2^t)} \leq \mathcal{O}\left(\frac{1}{\varepsilon^2(1-\rho^4)}\right),$$

where $\frac{1}{1-\rho^4} = \frac{1+\rho^4}{1-\rho^8} \leq \frac{2}{1-\rho^8}$ and $1-\rho^8 = 1 - \max_{i\in[k]} \frac{\lambda_{i+1}^2}{\lambda_i^2} = \min_{i\in[k]} \frac{\lambda_i^2-\lambda_{i+1}^2}{\lambda_i^2} = \gamma$. As a result, we have

$$\sum_{t=1}^{N} |J_t| \leq \mathcal{O}\left(\frac{\log(dN)}{\Delta^2\gamma\varepsilon^2}\right).$$

Combining this with the number of samples for the initialization phase, including that for finding $\bar{w}$, we have the stated sample complexity bound of the theorem.

## References

Hardt, Moritz and Price, Eric. The noisy power method: A meta algorithm with applications. In *Advances in Neural Information Processing Systems*, pp. 2861–2869, 2014.

Hom, Roger A and Johnson, Charles R. Topics in matrix analysis. *Cambridge University Press, New York*, 1991.

Laurent, B. and Massart, P. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, 28(5):1302–1338, 2000.

Li, Chun-Liang, Lin, Hsuan-Tien, and Lu, Chi-Jen. Rivalry of two families of algorithms for memory-restricted streaming PCA. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 473–481, 2016.

Mitliagkas, Ioannis, Caramanis, Constantine, and Jain, Prateek. Memory limited, streaming PCA. In *Advances in Neural Information Processing Systems*, pp. 2886–2894, 2013.

Schmitt, Bernhard A. Perturbation bounds for matrix square roots and pythagorean sums. *Linear algebra and its applications*, 174:215–227, 1992.

Stewart, Gilbert W. and Sun, Ji-Guang. *Matrix Perturbation Theory*. Academic Press, 1990.

Tropp, Joel A. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

Zhu, Peizhen. and Knyazev, Andrew V. Angles between subspaces and their tangents. *Arxiv preprint at arXiv:1209.0523*, 2012.