Power Iteration for Tensor PCA

Jiaoyang Huang*[1], Daniel Z. Huang†[2], Qing Yang‡[3], and Guang Cheng§[4]

[1] *New York University, New York, NY*
[2] *California Institute of Technology, Pasadena, CA*
[3] *University of Science and Technology of China, China*
[4] *Purdue University, West Lafayette, IN*

Abstract

In this paper, we study the power iteration algorithm for the spiked tensor model, as introduced in [44]. We give necessary and sufficient conditions for the convergence of the power iteration algorithm. When the power iteration algorithm converges, for the rank one spiked tensor model, we show the estimators for the spike strength and linear functionals of the signal are asymptotically Gaussian; for the multi-rank spiked tensor model, we show the estimators are asymptotically mixtures of Gaussian. This new phenomenon is different from the spiked matrix model. Using these asymptotic results of our estimators, we construct valid and efficient confidence intervals for spike strengths and linear functionals of the signals.

1 Introduction

Modern real scientific data call for more advanced structures than matrices. High order arrays, or tensors have been actively considered in neuroimaging analysis, topic modeling, signal processing and recommendation system [16, 17, 21, 22, 31, 43, 45, 46, 52]. Setting the stage, imagine that the signal is in the form of a large symmetric low-rank k-th order tensor

$$\boldsymbol{X}^* = \sum_{j=1}^{r} \beta_j \boldsymbol{v}_j^{\otimes k} \in \otimes^k \mathbb{R}^n, \tag{1}$$

where r ($r \ll n$) represents the rank and β_j are the strength of the signals. Such low-rank tensor components appear in various applications, e.g. community detection [2], moments estimation for latent variable models [3, 26] and hypergraph matching [19]. Suppose that we do not have access to perfect measurements about the entries of this signal tensor. The observations $\boldsymbol{X} = \boldsymbol{X}^* + \boldsymbol{Z}$ are contaminated by a substantial amount of random noise (reflected by the random tensor $\boldsymbol{Z}$ which has i.i.d. Gaussian entries with mean 0 and variance $1/n$.). The aim is to perform reliable estimation and inference on the unseen signal tensor $\boldsymbol{X}^*$. In literature, this is the spiked tensor model, introduced in [44].

In the special case, when $k = 2$, the above model reduces to the well-known "spiked matrix model" [28]. In this setting it is known that there is an order 1 critical signal-to-noise ratio β_c,

*jh4427@nyu.edu
†dzhuang@caltech.edu
‡yangq@ustc.edu.cn
§chengg@purdue.edu

such that below β_c, it is information-theoretical impossible to detect the spikes, and above β_c, it is possible to detect the spikes by Principal Component Analysis (PCA). A body of work has quantified the behavior of PCA in this setting [5–11, 18, 20, 28, 29, 34, 38, 40, 48]. We refer readers to the review articles [30] for more discussion and references to this and related lines of work

Tensor problems are far more than an extension of matrices. Not only the more involved structures and high-dimensionality, many concepts are not well defined [33], e.g. eigenvalues and eigenvectors, and most tensor problems are NP-hard [23]. Despite a large body of work tackling the spiked tensor model, there are several fundamental yet unaddressed challenges that deserve further attention.

Computational Hardness. The same as the spiked matrix model, for spiked tensor model, there is an order 1 critical signal-to-noise ratio β_k (depending on the order k), such that below β_k, it is information-theoretical impossible to detect the spikes, and above β_k, the maximum likelihood estimator is a distinguishing statistics [12, 13, 27, 35, 42]. In the matrix setting the maximum likelihood estimator is the top eigenvector, which can be computed in polynomial time by, e.g., power iteration. However, for order $k \geqslant 3$ tensor, computing the maximum likelihood estimator is NP-hard in generic setting. In this setting, it is widely believed that there is a regime of signal-to-noise ratios for which it is information theoretically possible to recover the signal but there is no known algorithm to efficiently approximate it. In the pioneer work [44], the algorithmic aspects of this model has been studied under the special setting when the rank $r = 1$. They showed that tensor power iteration with random initialization recovers the signal provided $\beta \gg n^{(k-1)/2}$, and tensor unfolding recovers the signal provided $\beta \gg n^{(\lceil k/2 \rceil - 1)/2}$. Based on heuristic arguments, they predicted that the necessary and sufficient condition for power iteration to succeed is $\beta \gg n^{(k-2)/2}$, and for tensor unfolding is $\beta \gg n^{(k-2)/4}$. Langevin dynamics and gradient descent were studied in [4], and shown to recover the signal provided $\beta \gg n^{(k-2)/2}$. Later the sharp threshold $\beta \gg n^{(k-2)/4}$ is achieved using Sum-of-Squares algorithms [24, 25, 32] and sophisticated iteration algorithms [36, 50]. The necessary part of this threshold still remains open, and its relation with hypergraphic planted clique problem was discussed in [37].

Statistical inferences. In many applications, it is often the case that the ultimate goal is not to characterize the L_2 or "bulk" behavior (e.g. the mean squared estimation error) of the signals, but rather to reason about the signals along a few preconceived yet important directions. In the example of community detecting for hypergraphs, the entries of the vector $\boldsymbol{v}$ can represent different community memberships. The testing of whether any two nodes belong to the same community is reduced to the hypothesis testing problem of whether the corresponding entries of $\boldsymbol{v}$ are equal. These problems can be formulated as estimation and inference for linear functionals of a signal, namely, quantities of the form $\langle \boldsymbol{a}, \boldsymbol{v}_j \rangle$, $1 \leqslant j \leqslant r$ with a prescribed vector $\boldsymbol{a}$. A natural starting point is to plug in an estimator $\widehat{\boldsymbol{v}}_j$ of $\boldsymbol{v}_j$, i.e. the estimator $\langle \boldsymbol{a}, \widehat{\boldsymbol{v}}_j \rangle$. However, a most prior works [24, 25, 32, 36, 44, 50] on spiked tensor models focuses on the L_2 risk analysis, which is often too coarse to give tight uncertainty bound for the plug-in estimator. To further complicate matters, there is often a bias issue surrounding the plug-in estimator. Addressing these issues calls for refined risk analysis of the algorithms.

1.1 Our Contributions

We consider the power iteration algorithm given by the following recursion

$$\boldsymbol{u}_0 = \boldsymbol{u}, \quad \boldsymbol{u}_{t+1} = \frac{\boldsymbol{X}[\boldsymbol{u}_t^{\otimes(k-1)}]}{\|\boldsymbol{X}[\boldsymbol{u}_t^{\otimes(k-1)}]\|_2} \tag{2}$$

where $\boldsymbol{u} \in \mathbb{R}^n$ with $\|\boldsymbol{u}\|_2 = 1$ is the initial vector, and $\boldsymbol{X}[\boldsymbol{v}^{\otimes(k-1)}] \in \mathbb{R}^n$ is the vector with i-th entry given by $\langle \boldsymbol{X}, \boldsymbol{e}_i \otimes \boldsymbol{v}^{\otimes(k-1)} \rangle$. The estimators are given by

$$\widehat{\boldsymbol{v}} = \boldsymbol{u}_T, \quad \widehat{\beta} = \langle \boldsymbol{X}, \widehat{\boldsymbol{v}}^{\otimes k} \rangle. \tag{3}$$

for some large T. Although in a worst case scenario, i.e. with random initialization, power iteration algorithm underperforms tensor unfolding. However, if extra information about the signals $\boldsymbol{v}_j$ is available, power iteration algorithm with a warm start can be used to obtain a much better estimator. In fact this approach is commonly used to obtain refined estimators. In this paper, we study the convergence and statistical inference aspects of the power iteration algorithm. The main contributions of this paper are summarized below,

Convergence criterion. We give necessary and sufficient conditions for the convergence of the power iteration algorithm. In the rank one case $r = 1$, we show that the power iteration algorithm converges to the true signal $\boldsymbol{v}$, provided $|\beta \langle \boldsymbol{u}, \boldsymbol{v} \rangle^{k-2}| \gg 1$ where $\boldsymbol{u}$ is the initialization vector. In the complementary setting, if $|\beta \langle \boldsymbol{u}, \boldsymbol{v} \rangle^{k-2}| \ll 1$, the output of the power iteration algorithm behaves like random Gaussian vectors, and has no correlations with the signal. With random initialization, i.e. $\boldsymbol{u}$ is a uniformly random vector on the unit sphere, our results assert that the power iteration algorithm converges in finite time, if and only if $\beta \gg n^{(k-2)/2}$, which verifies the prediction in [44]. This is analogous to the PCA of spiked matrix model, where power iteration recovers the top eigenvalue. However, the multi-rank spiked tensor model, i.e. $r \geqslant 2$, is different from multi-rank spiked matrix. The power iteration algorithm for multi-rank spiked tensor model is more sensitive to the initialization, i.e. the power iteration algorithm converges if $\max_j |\beta_j \langle \boldsymbol{u}, \boldsymbol{v}_j \rangle^{k-2}| \gg 1$. In this case, it converges to $\boldsymbol{v}_{j_*}$ with $j_* = \operatorname{argmax}_j |\beta_j \langle \boldsymbol{u}, \boldsymbol{v}_j \rangle^{k-2}|$.

Statistical inference We consider the statistical inference problem for the spiked tensor model. We develop the limiting distributions of the above power iteration estimators. In the rank one case, above the threshold $|\beta \langle \boldsymbol{u}, \boldsymbol{v} \rangle^{k-2}| \gg 1$, we show that our estimator $\langle \boldsymbol{a}, \widehat{\boldsymbol{v}} \rangle$ (modulo some global sign) admits the following first order approximation

$$\langle \boldsymbol{a}, \widehat{\boldsymbol{v}} \rangle \approx \left(1 - \frac{1}{2\beta^2}\right) \langle \boldsymbol{a}, \boldsymbol{v} \rangle + \frac{\langle \boldsymbol{a}^\perp, \boldsymbol{\xi} \rangle}{\beta},$$

where $\boldsymbol{a}^\perp = \boldsymbol{a} - \langle \boldsymbol{a}, \boldsymbol{v} \rangle \boldsymbol{v}$, and $\boldsymbol{\xi} = \boldsymbol{Z}[\boldsymbol{v}^{\otimes(k-1)}]$, is an n-dim vector, with each entry i.i.d. $\mathcal{N}(0, 1/n)$ Gaussian random variables. For multi-rank spiked tensor model, the output of power iteration algorithm depends on the angle between the initialization $\boldsymbol{u}$ and the signals $\boldsymbol{v}_j$. We consider the case the initialization $\boldsymbol{u}$ is a uniformly random vector on the unit sphere. For such initialization, very interestingly, our estimator $\langle \boldsymbol{a}, \widehat{\boldsymbol{v}} \rangle$ is asymptotically a mixture of Gaussian, with modes at $\langle \boldsymbol{a}, \boldsymbol{v}_j \rangle$ and mixture weights depending on the signal strength β_j. Using these asymptotic results of our estimators, we construct valid and efficient confidence intervals for the linear functionals $\langle \boldsymbol{a}, \boldsymbol{v}_j \rangle$.

1.2 Notations:

For a vector $\boldsymbol{v} \in \mathbb{R}^n$, we denote its i-th coordinate as $\boldsymbol{v}(i)$. We equate k-th order tensors in $\otimes^k \mathbb{R}^n$ with vectors of dimension n^k, i.e. $\boldsymbol{\tau} = (\boldsymbol{\tau}_{i_1 i_2 \cdots i_k})_{1 \leqslant i_1, i_2, \cdots, i_k \leqslant n}$. For any two k-th order tensors $\boldsymbol{\tau}, \boldsymbol{\eta} \in \otimes^k \mathbb{R}^n$, we denote their inner product as $\langle \boldsymbol{\tau}, \boldsymbol{\eta} \rangle := \sum_{1 \leqslant i_1, i_2, \cdots, i_k \leqslant n} \boldsymbol{\tau}_{i_1 i_2 \cdots i_k} \boldsymbol{\eta}_{i_1 i_2 \cdots i_k}$. A k-th order tensor can act on a $(k-1)$-th order tensor, and return a vector: $\boldsymbol{\tau} \in \otimes^k \mathbb{R}^n$ and $\boldsymbol{\eta} \in \otimes^{k-1} \mathbb{R}^n$

$$\boldsymbol{\tau}[\boldsymbol{\eta}] \in \mathbb{R}^n, \quad \boldsymbol{\tau}[\boldsymbol{\eta}](i) = \langle \boldsymbol{\tau}, \boldsymbol{e}_i \otimes \boldsymbol{\eta} \rangle = \sum_{1 \leqslant i_1, \cdots, i_{k-1} \leqslant n} \boldsymbol{\tau}_{i i_1 i_2 \cdots i_{k-1}} \boldsymbol{\eta}_{i_1 i_2 \cdots i_{k-1}}. \tag{4}$$

We denote the L_2 norm of a vector $\boldsymbol{v}$ as $\|\boldsymbol{v}\|$. We use $\stackrel{\mathrm{d}}{=}$ for the equality in law, and $\stackrel{\mathrm{d}}{\to}$ for the convergence in law. We denote the index sets $[\![a,b]\!] = \{a, a+1, a+2, \cdots, b\}$ and $[\![n]\!] = \{1, 2, 3, \cdots, n\}$. We use C to represent large universal constant, and c a small universal constant, which may be different from line by line. We write that $X = \mathrm{O}(Y)$ if there exists some universal constant such that $|X| \leqslant CY$. We write $X = \mathrm{o}(Y)$ if the ratio $|X|/Y \to \infty$ as n goes to infinity. We write $X \asymp Y$ if there exist universal constants such that $cY \leqslant |X| \leqslant CY$. We say an event holds with high probability, if for there exists $c > 0$, and n large enough, the event holds with probability at least $1 - n^{-c \log n}$.

An outline of the paper is given as follows. In Section 2.1, we state our main results for the rank-one spiked tensor model. In particular, with general initialization a distributional result for the power iteration algorithm is developed. Section 2.2 investigates the general rank-r spiked tensor model. A similar distributional result is established with general initialization as in Section 2.1. While with uniformly distributed initialization over the unit sphere, we obtain a multinoimal distribution which yields a mixture Gaussian. Numerical simulations are presented in Section 3. All proofs and technical details are deferred to the appendix.

Acknowledgement. The research collaboration was initiated when both G.C. and J.H. were warmly hosted by IAS in the special year of Deep Learning Theory. The research of J.H. is supported by the Simons Foundation as a Junior Fellow at the Simons Society of Fellows.

2 Main Results

2.1 Rank one spiked tensor model

In this section, we state our main results for the rank-one spiked tensor model (corresponding to $r = 1$ in (1)):

$$\boldsymbol{X} = \beta \boldsymbol{v}^{\otimes k} + \boldsymbol{Z}, \tag{5}$$

where

- $\boldsymbol{X} \in \otimes^k \mathbb{R}^n$ is the k-th order tensor observation.
- $\boldsymbol{Z} \in \otimes^k \mathbb{R}^n$ is a noise tensor. The entries of $\boldsymbol{Z}$ are i.i.d. standard $\mathcal{N}(0, 1/n)$ Gaussian random variables.
- $\beta \in \mathbb{R}$ is the signal size.
- $\boldsymbol{v} \in \mathbb{R}^n$ is an unknown unit vector to be recovered.

We obtain a distributional result for the power iteration algorithm (2) with general initialization $\boldsymbol{u}$: when $|\beta|$ is above certain threshold, $\boldsymbol{u}_t$ converges to $\boldsymbol{v}$, and the error is asymptotically Gaussian; when $|\beta|$ is below the same threshold, the algorithm does not converge.

Theorem 2.1. *Fix the initialization $\boldsymbol{u} \in \mathbb{R}^n$ with $\|\boldsymbol{u}\|_2 = 1$ and $\langle \boldsymbol{u}, \boldsymbol{v} \rangle \gtrsim 1/\sqrt{n}$. If $|\beta \langle \boldsymbol{u}, \boldsymbol{v} \rangle^{k-2}| \geqslant n^{\varepsilon}$ with arbitrarily small $\varepsilon > 0$, the behavior of the power iteration algorithm depends on the parity of k and the sign of β in the following sense:*

1. *If k is odd, and $\beta > 0$ then $(\boldsymbol{X}[\boldsymbol{u}_t^{\otimes k}], \boldsymbol{u}_t)$ converges to $(\beta, \boldsymbol{v})$;*
2. *If k is odd, and $\beta < 0$ then $(\boldsymbol{X}[\boldsymbol{u}_t^{\otimes k}], \boldsymbol{u}_t)$ converges to $(-\beta, -\boldsymbol{v})$;*

3. *If k is even, and $\beta > 0$, then $(\boldsymbol{X}[\boldsymbol{u}_t^{\otimes k}], \boldsymbol{u}_t)$ converges to $(\beta, \operatorname{sgn}(\langle \boldsymbol{u}, \boldsymbol{v}\rangle)\boldsymbol{v})$ depending on the initialization $\boldsymbol{u}$;*

4. *If k is even, and $\beta < 0$, then $(\boldsymbol{X}[\boldsymbol{u}_t^{\otimes k}], \boldsymbol{u}_t)$ does not converge, but instead alternates between $(\beta, \operatorname{sgn}(\langle \boldsymbol{u}, \boldsymbol{v}\rangle)\boldsymbol{v})$ and $(\beta, -\operatorname{sgn}(\langle \boldsymbol{u}, \boldsymbol{v}\rangle)\boldsymbol{v})$.*

In Case 1, for any fixed unit vector $\boldsymbol{a} \in \mathbb{R}^n$, and

$$T \geqslant 1 + \frac{1}{\varepsilon}\left(\frac{1}{2} + \frac{2\log|\beta|}{\log n}\right), \tag{6}$$

with probability $1 - \mathrm{O}(n^{-c(\log n)^2})$, the estimators $\widehat{\boldsymbol{v}} = \boldsymbol{u}_T$, and $\widehat{\beta} = \boldsymbol{X}[\widehat{\boldsymbol{v}}^{\otimes k}]$ satisfies

$$\begin{aligned}\langle \boldsymbol{a}, \widehat{\boldsymbol{v}}\rangle = \langle \boldsymbol{a}, \boldsymbol{u}_T\rangle = \left(1 - \frac{1}{2\beta^2}\right)\langle \boldsymbol{a}, \boldsymbol{v}\rangle + \frac{\langle \boldsymbol{a}, \boldsymbol{\xi}\rangle - \langle \boldsymbol{a}, \boldsymbol{v}\rangle\langle \boldsymbol{v}, \boldsymbol{\xi}\rangle}{\beta} \\ + \mathrm{O}\left(\frac{\log n}{\beta^2\sqrt{n}} + \frac{(\log n)^{3/2}}{\beta^{3/2}n^{3/4}} + \frac{|\langle \boldsymbol{a}, \boldsymbol{v}\rangle|}{\beta^4}\right),\end{aligned} \tag{7}$$

where $\boldsymbol{\xi} = \boldsymbol{Z}[\boldsymbol{v}^{\otimes(k-1)}]$, is an n-dim vector, with each entry i.i.d. $\mathcal{N}(0, 1/n)$ Gaussian random variable. And

$$\widehat{\beta} = \boldsymbol{X}[\boldsymbol{u}_T^{\otimes k}] = \beta + \langle \boldsymbol{\xi}, \boldsymbol{v}\rangle - \frac{k/2 - 1}{\beta} + \mathrm{O}\left(\frac{\log n}{|\beta|\sqrt{n}} + \frac{(\log n)^{3/2}}{|\beta|^{1/2}n^{3/4}} + \frac{1}{|\beta|^3}\right). \tag{8}$$

Under the same assumption, we have similar results for Cases 2, 3, 4, by simply changing $(\beta, \boldsymbol{v})$ in the righthand side of (7) and (8) to the corresponding limit.

Theorem 2.2. *Fix the initialization $\boldsymbol{u} \in \mathbb{R}^n$ with $\|\boldsymbol{u}\|_2 = 1$. If $|\beta| \geqslant n^{\varepsilon}$ and $|\beta\langle \boldsymbol{u}, \boldsymbol{v}\rangle^{k-2}| \leqslant n^{-\varepsilon}$ with arbitrarily small $\varepsilon > 0$, then $\boldsymbol{u}_t$ does not converge to $\pm\boldsymbol{v}$, and $\boldsymbol{u}_t$ behaves like a random Gaussian vector. For*

$$T \geqslant 1 + \frac{1}{\varepsilon}\left(\frac{1}{2} - \frac{\log|\beta|}{(k-2)\log n}\right) \tag{9}$$

with probability $1 - \mathrm{O}(n^{-c(\log n)^2})$, it holds

$$\widehat{\boldsymbol{v}} = \boldsymbol{u}_T = \frac{\tilde{\boldsymbol{\xi}}}{\|\tilde{\boldsymbol{\xi}}\|_2} + \mathrm{O}\left(|\beta|\left(\frac{\log n}{\sqrt{n}}\right)^{k-1}\right), \tag{10}$$

where $\tilde{\boldsymbol{\xi}}$ is the standard Gaussian vector in $\mathbb{R}^n$, the error term is a vector of length bounded by $|\beta|(\log n/\sqrt{n})^{k-1}$.

In Theorem 2.1, we assume that $\langle \boldsymbol{u}, \boldsymbol{v}\rangle \gtrsim 1/\sqrt{n}$, which is generic and is true for a random $\boldsymbol{u}$. Moreover, if the initial vector $\boldsymbol{u}$ is random, then $|\langle \boldsymbol{u}, \boldsymbol{v}\rangle| \asymp n^{-1/2}$. Notably, Theorems 2.1 and 2.2 together state that power iteration recovers $\boldsymbol{v}$ if $|\beta| \gg n^{(k-2)/2}$ and fails if $|\beta| \ll n^{(k-2)/2}$. This gives a rigorous proof of the prediction in [44] that the necessary and sufficient condition for the convergence is given by $|\beta| \gtrsim n^{(k-2)/2}$. In practice, it may be possible to use domain knowledge to choose better initialization points. For example, in the classical topic modeling applications [3], the unknown vectors $\boldsymbol{v}$ are related to the topic word distributions, and many documents may be primarily composed of words from just single topic. Therefore, good initialization points can be derived from these single-topic documents.

The special case for $k = 2$, i.e. the spiked matrix model, has been intensively studied since the pioneer work of Johnstone [28]. In this setting it is known [30] that there is an order O(1) critical signal-to-noise ratio, such that below the threshold, it is information-theoretically impossible to recover $\boldsymbol{v}$, and above the threshold, the PCA (partially) recovers the unseen eigenvector $\boldsymbol{v}$ [1,14,15,39,41,47,49,51]. The special case of our results Theorem 2.1 recovers some abovementioned results.

As a consequence of Theorem 2.1, we have the following central limit theorem for our estimators.

Corollary 2.3. *(Central Limit Theorem) Fix the initialization $\boldsymbol{u} \in \mathbb{R}^n$ with $\|\boldsymbol{u}\|_2 = 1$ and $|\langle \boldsymbol{u}, \boldsymbol{v}\rangle| \gtrsim 1/\sqrt{n}$. If $|\beta\langle \boldsymbol{u}, \boldsymbol{v}\rangle^{k-2}| \geqslant n^{\varepsilon}$ with arbitrarily small $\varepsilon > 0$, in Case 1 of Theorem 2.1, for any fixed unit vector $\boldsymbol{a} \in \mathbb{R}^n$ obeying*

$$|\langle \boldsymbol{a}, \boldsymbol{v}\rangle| = \mathrm{o}\left(\frac{\beta^3}{\sqrt{n}}\right), \tag{11}$$

and time

$$T \geqslant 1 + \frac{1}{\varepsilon}\left(\frac{1}{2} + \frac{2\log|\beta|}{\log n}\right). \tag{12}$$

the estimators $\widehat{\boldsymbol{v}} = \boldsymbol{u}_T$, and $\widehat{\beta} = \boldsymbol{X}[\widehat{\boldsymbol{v}}^{\otimes k}]$ satisfies

$$\frac{\sqrt{n}\widehat{\beta}}{\sqrt{\langle \boldsymbol{a}, (\boldsymbol{I}_n - \widehat{\boldsymbol{v}}\widehat{\boldsymbol{v}}^\top)\boldsymbol{a}\rangle}}\left[\big(1 - \frac{1}{2\widehat{\beta}^2}\big)^{-1}\langle \boldsymbol{a}, \widehat{\boldsymbol{v}}\rangle - \langle \boldsymbol{a}, \boldsymbol{v}\rangle\right] \xrightarrow{d} \mathcal{N}(0, 1), \tag{13}$$

as n tends to infinity. We have similar results for Cases 2, 3, 4, by simply changing $(\beta, \boldsymbol{v})$ in (13) to the corresponding limit.

We remark that in Corollary 2.3, we assume that $|\langle \boldsymbol{a}, \boldsymbol{v}\rangle| = \mathrm{o}\big(\beta^3/\sqrt{n}\big)$, which is generic. For example, if $\boldsymbol{v}$ is delocalized, and $\boldsymbol{a}$ is supported on finitely many entries, we will have that $|\langle \boldsymbol{a}, \boldsymbol{v}\rangle| \lesssim 1/\sqrt{n}$, and (11) is satisfied.

With the central limit theorem for our estimators in Corollary 2.3, we can easily write down the confidence interval for our estimators.

Corollary 2.4. *(Prediction Interval) Given the asymptotic significance level α, and let $z_\alpha = \Phi(1 - \alpha/2)$ where $\Phi(\cdot)$ is the CDF of a standard Gaussian. If $|\beta\langle \boldsymbol{u}, \boldsymbol{v}\rangle^{k-2}| \geqslant n^{\varepsilon}$ with arbitrarily small $\varepsilon > 0$, in Case 1 of Theorem 2.1, for any fixed unit vector $\boldsymbol{a} \in \mathbb{R}^n$ obeying*

$$|\langle \boldsymbol{a}, \boldsymbol{v}\rangle| = \mathrm{o}\left(\frac{\beta^3}{\sqrt{n}}\right), \tag{14}$$

and time

$$T \geqslant 1 + \frac{1}{\varepsilon}\left(\frac{1}{2} + \frac{2\log|\beta|}{\log n}\right), \tag{15}$$

let $\widehat{\boldsymbol{v}} = \boldsymbol{u}_T$, and $\widehat{\beta} = \boldsymbol{X}[\widehat{\boldsymbol{v}}^{\otimes k}]$. The asymptotic confidence interval of $\langle \boldsymbol{a}, \boldsymbol{v}\rangle$ is given by

$$\frac{1}{1 - 1/(2\widehat{\beta}^2)}\left[\langle \boldsymbol{a}, \widehat{\boldsymbol{v}}\rangle - z_\alpha\frac{\sqrt{\langle \boldsymbol{a}, (\boldsymbol{I}_n - \widehat{\boldsymbol{v}}\widehat{\boldsymbol{v}}^\top)\boldsymbol{a}\rangle}}{\sqrt{n}\widehat{\beta}},\ \langle \boldsymbol{a}, \widehat{\boldsymbol{v}}\rangle + z_\alpha\frac{\sqrt{\langle \boldsymbol{a}, (\boldsymbol{I}_n - \widehat{\boldsymbol{v}}\widehat{\boldsymbol{v}}^\top)\boldsymbol{a}\rangle}}{\sqrt{n}\widehat{\beta}}\right]. \tag{16}$$

We have similar results for Cases 2, 3, 4, by simply changing $(\beta, \boldsymbol{v})$ in (16) to the corresponding limit.

2.2 General Results: rank-r spiked tensor model

In this section, we state our main results for the general case, the rank-r spiked tensor model (1). Before stating our main results, we need to introduce some more notations and assumptions.

Assumption 2.5. *We assume that the initialization does not distinguish $\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_r$, such that there exists some large constant $\kappa > 0$*

$$1/\kappa \leqslant \left| \frac{\langle \boldsymbol{u}, \boldsymbol{v}_i \rangle}{\langle \boldsymbol{u}, \boldsymbol{v}_j \rangle} \right| \leqslant \kappa, \tag{17}$$

for all $1 \leqslant i, j \leqslant r$.

If we take the uniform initialization, i.e. $\boldsymbol{u}_0 = \boldsymbol{u}$ is a uniformly distributed vector in $\mathbb{S}^{n-1}$. Then with probability $1 - \mathrm{O}(r/\sqrt{\kappa})$ we will have $1/\sqrt{\kappa n} \leqslant |\langle \boldsymbol{u}, \boldsymbol{v}_i \rangle| \leqslant \sqrt{\kappa/n}$ for $1 \leqslant i \leqslant r$, and Assumption 2.5 holds.

The same as in the rank-1 case, the quantities $|\beta_j \langle \boldsymbol{u}, \boldsymbol{v}_j \rangle^{k-2}|$ play a crucial role in our power iteration algorithm. We need to make the following technical assumption:

Assumption 2.6. *Let $j_* = \mathrm{argmax}_j \, |\beta_j \langle \boldsymbol{u}, \boldsymbol{v}_j \rangle^{k-2}|$. We assume that there exists some large constant $\kappa > 0$*

$$(1 - 1/\kappa) |\beta_{j_*} \langle \boldsymbol{u}, \boldsymbol{v}_{j_*} \rangle^{k-2}| \geqslant |\beta_j \langle \boldsymbol{u}, \boldsymbol{v}_j \rangle^{k-2}|, \tag{18}$$

for all $1 \leqslant j \leqslant r$ and $j \neq j_*$.

It turns out under Assumptions 2.5 and 2.6, the power iteration converges to $\boldsymbol{v}_{j_*}$. Moreover, if we simply take the uniform initialization, i.e. $\boldsymbol{u}_0 = \boldsymbol{u}$ is a uniformly distributed vector in $\mathbb{S}^{n-1}$. Assumption 2.6 holds for some $1 \leqslant j_* \leqslant r$ with probability $1 - \mathrm{O}(1/\kappa)$.

Theorem 2.7. *Fix the initialization $\boldsymbol{u} \in \mathbb{R}^n$ with $\|\boldsymbol{u}\|_2 = 1$ and $|\langle \boldsymbol{u}, \boldsymbol{v}_j \rangle| \gtrsim 1/\sqrt{n}$, for $1 \leqslant j \leqslant r$. Let $j_* = \mathrm{argmax}_j \, |\beta_j \langle \boldsymbol{u}, \boldsymbol{v}_j \rangle^{k-2}|$. Under Assumptions 2.5 and 2.6, if $|\beta_{j_*} \langle \boldsymbol{u}, \boldsymbol{v}_{j_*} \rangle^{k-2}| \geqslant n^{\varepsilon}$ with arbitrarily small $\varepsilon > 0$, the behavior of the power iteration algorithm depends on the parity of k and the sign of β_{j_*}:*

1. *If k is odd, and $\beta_{j_*} > 0$ then $(\boldsymbol{X}[\boldsymbol{u}_t^{\otimes k}], \boldsymbol{u}_t)$ converges to $(\beta_{j_*}, \boldsymbol{v}_{j_*})$;*
2. *If k is odd, and $\beta_{j_*} < 0$ then $(\boldsymbol{X}[\boldsymbol{u}_t^{\otimes k}], \boldsymbol{u}_t)$ converges to $(-\beta_{j_*}, -\boldsymbol{v}_{j_*})$;*
3. *If k is even, and $\beta_{j_*} > 0$, then $(\boldsymbol{X}[\boldsymbol{u}_t^{\otimes k}], \boldsymbol{u}_t)$ converges to $(\beta_{j_*}, \mathrm{sgn}(\langle \boldsymbol{u}, \boldsymbol{v}_{j_*} \rangle) \boldsymbol{v}_{j_*})$ depending on the initialization $\boldsymbol{u}$;*
4. *If k is even, and $\beta_{j_*} < 0$, then $(\boldsymbol{X}[\boldsymbol{u}_t^{\otimes k}], \boldsymbol{u}_t)$ does not converge, but instead alternating between $(\beta_{j_*}, \mathrm{sgn}(\langle \boldsymbol{u}, \boldsymbol{v}_{j_*} \rangle) \boldsymbol{v}_{j_*})$ and $(\beta_{j_*}, -\mathrm{sgn}(\langle \boldsymbol{u}, \boldsymbol{v}_{j_*} \rangle) \boldsymbol{v}_{j_*})$.*

In Case 1, for any fixed unit vector $\boldsymbol{a} \in \mathbb{R}^n$, and

$$T \geqslant 1 + \frac{1}{\varepsilon} \left(\frac{1}{2} + \frac{2 \log |\beta_1|}{\log n} \right) + \frac{\log \log (\sqrt{n} |\beta_1|)}{\log (k-1)}, \tag{19}$$

the estimators $\widehat{\boldsymbol{v}} = \boldsymbol{u}_T$, and $\widehat{\beta} = \boldsymbol{X}[\widehat{\boldsymbol{v}}^{\otimes k}]$ satisfies

$$\begin{aligned} \langle \boldsymbol{a}, \widehat{\boldsymbol{v}} \rangle = \langle \boldsymbol{a}, \boldsymbol{u}_T \rangle = & \left(1 - \frac{1}{2\beta_{j_*}^2} \right) \langle \boldsymbol{a}, \boldsymbol{v}_{j_*} \rangle + \frac{\langle \boldsymbol{a}, \boldsymbol{\xi} \rangle - \langle \boldsymbol{a}, \boldsymbol{v}_{j_*} \rangle \langle \boldsymbol{v}_{j_*}, \boldsymbol{\xi} \rangle}{\beta_{j_*}} \\ & + \mathrm{O}_{\mathbb{P}} \left(\frac{\log n}{\sqrt{n}} \left(\frac{\log n}{\sqrt{n} |\beta_1|} \right)^{k-1} + \frac{\log n}{|\beta_1|^2 \sqrt{n}} + \frac{(\log n)^{3/2}}{|\beta_1|^{3/2} n^{3/4}} + \frac{1}{|\beta_1|^4} \right), \end{aligned} \tag{20}$$

where $\boldsymbol{\xi} = \boldsymbol{Z}[\boldsymbol{v}_{j_*}^{\otimes(k-1)}]$*, is an* n*-dim vector, with each entry i.i.d.* $\mathcal{N}(0, 1/n)$ *Gaussian random variable. And*

$$\begin{aligned}\widehat{\beta} = \boldsymbol{X}[\boldsymbol{u}_T^{\otimes k}] = \beta_{j_*} + \langle \boldsymbol{\xi}, \boldsymbol{v}_{j_*}\rangle - \frac{k/2-1}{\beta_{j_*}} \\ + \mathrm{O}_{\mathbb{P}}\left(\frac{\log n}{\sqrt{n}}\left(\frac{\log n}{\sqrt{n}|\beta_1|}\right)^{k-1} + \frac{\log n}{|\beta_1|\sqrt{n}} + \frac{(\log n)^{3/2}}{|\beta_1|^{1/2}n^{3/4}} + \frac{1}{|\beta_1|^3}\right).\end{aligned} \tag{21}$$

Under the same assumption, we have similar results for Cases 2, 3, 4, by simply changing $(\beta_{j_*}, \boldsymbol{v}_{j_*})$ *in the righthand side of* (20) *and* (21) *to the corresponding limit.*

In Theorem 2.7, we assume that $|\langle \boldsymbol{u}, \boldsymbol{v}_j\rangle| \gtrsim 1/\sqrt{n}$ for $1 \leqslant j \leqslant r$. This is generic and is true for a random initialization $\boldsymbol{u}$.

We want to remark that for multi-rank spiked tensor model, the senarios for $k = 2$, i.e. the spiked matrix model, and $k \geqslant 3$ are very different. For the spiked matrix model, in Theorem 2.7, we always have that $j_* = \mathrm{argmax}_j\, |\beta_j| = 1$, and power iteration algorithm always converges to the eigenvector corresponding to the largest eigenvalue. However, for rank $k \geqslant 3$, the power iteration algorithm may converge to any vector $\boldsymbol{v}_j$ provided that the initialization $\boldsymbol{u}$ is sufficiently close to $\boldsymbol{v}_j$. As a consequence of Theorem 2.7, we have the following central limit theorem for our estimators.

Corollary 2.8. *Fix the initialization* $\boldsymbol{u} \in \mathbb{R}^n$ *with* $\|\boldsymbol{u}\|_2 = 1$ *and* $|\langle \boldsymbol{u}, \boldsymbol{v}_j\rangle| \gtrsim 1/\sqrt{n}$ *for* $1 \leqslant j \leqslant r$*. We assume* $|\beta\langle \boldsymbol{u}, \boldsymbol{v}_{j_*}\rangle^{k-2}| \geqslant n^{\varepsilon}$ *with arbitrarily small* $\varepsilon > 0$*, and Assumptions 2.5 and 2.6. In Case 1 of Theorem 2.7, for any fixed unit vector* $\boldsymbol{a} \in \mathbb{R}^n$*, for any fixed unit vector* $\boldsymbol{a} \in \mathbb{R}^n$ *obeying*

$$|\langle \boldsymbol{a}, \boldsymbol{v}_{j_*}\rangle| = \mathrm{o}\left(\frac{|\beta_1|^3}{\sqrt{n}}\right), \tag{22}$$

and time

$$T \geqslant 1 + \frac{1}{\varepsilon}\left(\frac{1}{2} + \frac{2\log|\beta_1|}{\log n}\right), \tag{23}$$

the estimators $\widehat{\boldsymbol{v}} = \boldsymbol{u}_T$*, and* $\widehat{\beta} = \boldsymbol{X}[\boldsymbol{u}_T^{\otimes k}]$ *satisfy*

$$\frac{\sqrt{n}\widehat{\beta}_{j_*}}{\sqrt{\langle \boldsymbol{a}, (\boldsymbol{I}_n - \widehat{\boldsymbol{v}}\widehat{\boldsymbol{v}}^\top)\boldsymbol{a}\rangle}}\left[\big(1 - \frac{1}{2\widehat{\beta}_{j_*}^2}\big)^{-1}\langle \boldsymbol{a}, \widehat{\boldsymbol{v}}\rangle - \langle \boldsymbol{a}, \boldsymbol{v}_{j_*}\rangle\right] \xrightarrow{d} \mathcal{N}(0, 1). \tag{24}$$

We have similar results for Cases 2, 3, 4, by simply changing $(\beta_{j_*}, \boldsymbol{v}_{j_*})$ *in* (24) *to the corresponding limit.*

In the following we take $\boldsymbol{u}$ to be a random vector uniformly distributed over the unit sphere. The power iteration algorithm can be easily understood in this setting, thanks to Theorem 2.7. More precisely if $j_* = \mathrm{argmax}_j\, |\beta_j\langle \boldsymbol{u}, \boldsymbol{v}_j\rangle^{k-2}|$ and the initialization $\boldsymbol{u}$ satisfies Assumptions 2.5 and 2.6, then the power iteration estimator $(\widehat{\boldsymbol{v}}, \widehat{\beta})$ recovers $(\boldsymbol{v}_{j_*}, \beta_{j_*})$. From the discussions below, for a random vector $\boldsymbol{u}$ uniformly distributed over the unit sphere, Assumptiosn 2.5 and 2.6 holds with probability $1 - \mathrm{O}(1/\sqrt{\kappa})$. We can compute explicitly the probability that index i achieves $\mathrm{argmax}_j\, |\beta_j\langle \boldsymbol{u}, \boldsymbol{v}_j\rangle^{k-2}|$:

$$\begin{aligned} p_i &:= \mathbb{P}(i = \mathrm{argmax}_j\, |\beta_j\langle \boldsymbol{u}, \boldsymbol{v}_j\rangle^{k-2}|) \\ &= \int_0^\infty \sqrt{\frac{2}{\pi}} e^{-x^2/2}\left(\prod_{\ell \neq i}\int_0^{\left(\frac{|\beta_i|}{|\beta_\ell|}\right)^{\frac{1}{k-2}}x} \sqrt{\frac{2}{\pi}} e^{-y^2/2}\mathrm{d}y\right)\mathrm{d}x, \end{aligned} \tag{25}$$

for any $1 \leqslant i \leqslant r$. For spiked matrix model, i.e. $k = 2$, we always have $1 = \mathrm{argmax}_j \, |\beta_j \langle \boldsymbol{u}, \boldsymbol{v}_j \rangle^{k-2}|$, and $p_1 = 1, p_2 = p_3 = \cdots = 0$. For spiked tensor models with $k \geqslant 3$, all those p_i are nonnegative and $p_1 \geqslant p_2 \geqslant p_3 \geqslant \cdots > 0$.

Theorem 2.9. *Fix large $\kappa > 0$ and recall p_i as defined* (25)*. If $\boldsymbol{u}$ is uniformly distributed over the unit sphere, and $|\beta_1| \geqslant n^{(k-2)/2+\varepsilon}$ with arbitrarily small $\varepsilon > 0$, then for any $1 \leqslant i \leqslant r$:*

1. *If k is odd, and $\beta_i > 0$ then with probability $p_i + \mathrm{O}(1/\sqrt{\kappa})$, $(\boldsymbol{X}[\boldsymbol{u}_t^{\otimes k}], \boldsymbol{u}_t)$ converges to $(\beta_i, \boldsymbol{v}_i)$;*
2. *If k is odd, and $\beta_i < 0$ then with probability $p_i + \mathrm{O}(1/\sqrt{\kappa})$, $(\boldsymbol{X}[\boldsymbol{u}_t^{\otimes k}], \boldsymbol{u}_t)$ converges to $(-\beta_i, -\boldsymbol{v}_i)$;*
3. *If k is even, and $\beta_i > 0$, then with probability $p_i/2 + \mathrm{O}(1/\sqrt{\kappa})$, $(\boldsymbol{X}[\boldsymbol{u}_t^{\otimes k}], \boldsymbol{u}_t)$ converges to $(\beta_i, +\boldsymbol{v}_i)$, and with probability $p_i/2 + \mathrm{O}(1/\sqrt{\kappa})$, $(\boldsymbol{X}[\boldsymbol{u}_t^{\otimes k}], \boldsymbol{u}_t)$ converges to $(\beta_i, -\boldsymbol{v}_i)$.*
4. *If k is even, and $\beta_i < 0$, then with probability $p_i + \mathrm{O}(1/\sqrt{\kappa})$, $(\boldsymbol{X}[\boldsymbol{u}_t^{\otimes k}], \boldsymbol{u}_t)$ alternates between $(\beta_i, \boldsymbol{v}_i)$ and $(\beta_i, -\boldsymbol{v}_i)$.*

In Case 1, for any fixed unit vector $\boldsymbol{a} \in \mathbb{R}^n$, and

$$T \geqslant 1 + \frac{1}{\varepsilon}\left(\frac{1}{2} + \frac{2\log|\beta_1|}{\log n}\right) + \frac{\log\log(\sqrt{n}|\beta_1|)}{\log(k-1)}, \tag{26}$$

with probability $p_i + \mathrm{O}(1/\sqrt{\kappa})$, the estimators $\widehat{\boldsymbol{v}} = \boldsymbol{u}_T$, and $\widehat{\beta} = \boldsymbol{X}[\boldsymbol{u}_T^{\otimes k}]$ satisfy

$$\begin{aligned} \langle \boldsymbol{a}, \widehat{\boldsymbol{v}} \rangle = \langle \boldsymbol{a}, \boldsymbol{u}_T \rangle = \left(1 - \frac{1}{2\beta_i^2}\right)\langle \boldsymbol{a}, \boldsymbol{v}_i \rangle + \frac{\langle \boldsymbol{a}, \boldsymbol{\xi} \rangle - \langle \boldsymbol{a}, \boldsymbol{v}_i \rangle \langle \boldsymbol{v}_i, \boldsymbol{\xi} \rangle}{\beta_i} \\ + \mathrm{O}_{\mathbb{P}}\left(\frac{\log n}{\sqrt{n}}\left(\frac{\log n}{\sqrt{n}|\beta_1|}\right)^{k-1} + \frac{\log n}{|\beta_1|^2\sqrt{n}} + \frac{(\log n)^{3/2}}{|\beta_1|^{3/2} n^{3/4}} + \frac{1}{|\beta_1|^4}\right), \end{aligned} \tag{27}$$

where $\boldsymbol{\xi} = \boldsymbol{Z}[\boldsymbol{v}_i^{\otimes(k-1)}]$, is an n-dim vector, with each entry i.i.d. $\mathcal{N}(0, 1/n)$ Gaussian random variable. And

$$\begin{aligned} \widehat{\beta} = \boldsymbol{X}[\boldsymbol{u}_T^{\otimes k}] = \beta_i + \langle \boldsymbol{\xi}, \boldsymbol{v}_i \rangle - \frac{k/2 - 1}{\beta_i} \\ + \mathrm{O}_{\mathbb{P}}\left(\frac{\log n}{\sqrt{n}}\left(\frac{\log n}{\sqrt{n}|\beta_1|}\right)^{k-1} + \frac{\log n}{|\beta_1|\sqrt{n}} + \frac{(\log n)^{3/2}}{|\beta_1|^{1/2} n^{3/4}} + \frac{1}{|\beta_1|^3}\right). \end{aligned} \tag{28}$$

Under the same assumption, we have similar results for Cases 2, 3, 4, by simply changing $(\beta_i, \boldsymbol{v}_i)$ in the righthand side of (27) *and* (28) *to the corresponding limit.*

We want to emphasize here that the senarios for $k = 2$, i.e. the spiked matrix model, and $k \geqslant 3$ are very different. For spiked matrix model, i.e. $k = 2$, we always have that $p_1 = 0, p_2 = p_3 = \cdots = 0$. The power iteration algorithm always converges to the eigenvector corresponding to the largest eigenvalue. We can only recover $(\beta_1, \boldsymbol{v}_1)$ no matter how many times we repeat the algorithm. However, for spiked tensor models with $k \geqslant 3$, all those p_i are nonnegative, $p_1 \geqslant p_2 \geqslant p_3 \geqslant \cdots > 0$. By repeating the power iteration algorithm for sufficiently many times, it recovers $(\beta_i, \boldsymbol{v}_i)$ with probability roughly p_i.

Similar to the rank one case in Section 2.1, we are also able to establish the asymptotic distribution and confidence interval for multi-rank spiked tensor model with uniformly distributed initialization $\boldsymbol{u}$.

Corollary 2.10. *Fix $k \geqslant 3$, assume $\boldsymbol{u}$ to be a random vector uniformly distributed over the unit sphere and $|\beta_1| \geqslant n^{(k-2)/2+\varepsilon}$ with arbitrarily small $\varepsilon > 0$. In Case 1 of Theorem 2.9, for any fixed unit vector $\boldsymbol{a} \in \mathbb{R}^n$, and time*

$$T \geqslant 1 + \frac{1}{\varepsilon}\left(\frac{1}{2} + \frac{2\log|\beta_1|}{\log n}\right) + \frac{\log\log(\sqrt{n}|\beta_1|)}{\log(k-1)},$$

for any $1 \leqslant i \leqslant r$, with probability $p_i + \mathrm{O}(1/\sqrt{\kappa})$, the estimators $\widehat{\boldsymbol{v}} = \boldsymbol{u}_T$ and $\widehat{\beta} = \boldsymbol{X}[\boldsymbol{u}_T^{\otimes k}]$ satisfy

$$\frac{\sqrt{n}\widehat{\beta}}{\sqrt{\langle \boldsymbol{a}, (\boldsymbol{I}_n - \widehat{\boldsymbol{v}}\widehat{\boldsymbol{v}}^\top)\boldsymbol{a}\rangle}}\left[\langle \boldsymbol{a}, \widehat{\boldsymbol{v}}\rangle - \big(1 - \frac{1}{2\widehat{\beta}^2}\big)\langle \boldsymbol{a}, \boldsymbol{v}_i\rangle\right] \xrightarrow{d} \mathcal{N}(0,1). \tag{29}$$

And

$$\sqrt{n}\left(\beta_i - \widehat{\beta} - \frac{k/2-1}{\widehat{\beta}}\right) \xrightarrow{d} \mathcal{N}(0,1). \tag{30}$$

We have similar results for Cases 2, 3, 4, by simply changing $(\beta_i, \boldsymbol{v}_i)$ above to the corresponding limit.

We want to emphasize the difference between Corollary 2.3 and Corollary 2.10. In the rank one case, the estimators $\widehat{\beta}$ and $\langle \boldsymbol{a}, \widehat{\boldsymbol{v}}\rangle$ are asymptotically Gaussian. In the multi-rank spiked tensor model with $k \geqslant 3$, those estimators $\widehat{\beta}$ and $\langle \boldsymbol{a}, \widehat{\boldsymbol{v}}\rangle$ are no longer Gaussian. Instead, they are asymptotically a mixture Gaussian with mixture weights $p_1 \geqslant p_2 \geqslant p_3 \geqslant \cdots$.

Corollary 2.11. *Given the asymptotic significance level α, and let $z_\alpha = \Phi(1-\alpha/2)$ where $\Phi(\cdot)$ is the CDF of a standard Gaussian. Under the conditions in Corollary 2.10, in Case 1 of Theorem 2.9, we can find the asymptotic confidence interval of $\langle \boldsymbol{a}, \boldsymbol{v}_i\rangle$ as*

$$\frac{1}{1 - 1/(2\widehat{\beta}^2)}\left[\langle \boldsymbol{a}, \widehat{\boldsymbol{v}}\rangle - z_\alpha\frac{\sqrt{\langle \boldsymbol{a}, (\boldsymbol{I}_n - \widehat{\boldsymbol{v}}\widehat{\boldsymbol{v}}^\top)\boldsymbol{a}\rangle}}{\sqrt{n}\widehat{\beta}},\ \langle \boldsymbol{a}, \widehat{\boldsymbol{v}}\rangle + z_\alpha\frac{\sqrt{\langle \boldsymbol{a}, (\boldsymbol{I}_n - \widehat{\boldsymbol{v}}\widehat{\boldsymbol{v}}^\top)\boldsymbol{a}\rangle}}{\sqrt{n}\widehat{\beta}}\right]$$

and the asymptotic confidence interval of β_i as

$$\left[\widehat{\beta} + \frac{k/2-1}{\widehat{\beta}} - \frac{z_\alpha}{\sqrt{n}},\quad \widehat{\beta} + \frac{k/2-1}{\widehat{\beta}} + \frac{z_\alpha}{\sqrt{n}}\right].$$

We have similar results for Cases 2, 3, 4, by changing $(\beta_i, \boldsymbol{v}_i)$ above to the corresponding limit.

3 Numerical Study

In this section, we conduct numerical experiments on synthetic data to demonstrate our distributional results provided in Sections 2.1 and 2.2. We fix the dimension $n = 600$ and rank $k = 3$.

3.1 Rank one spiked tensor model

We begin with numerical experiments on rank one case. This section is devoted to numerically studying the efficiency of our estimators for the strength of signals and linear functionals of the signals. We take the signal $\boldsymbol{v}$ a random vector sampled from the unit sphere in $\mathbb{R}^n$, and the vector

$$\boldsymbol{a} = \frac{1}{\sqrt{3}}(\boldsymbol{e}_{n/3} + \boldsymbol{e}_{2n/3} + \boldsymbol{e}_n) \tag{31}$$

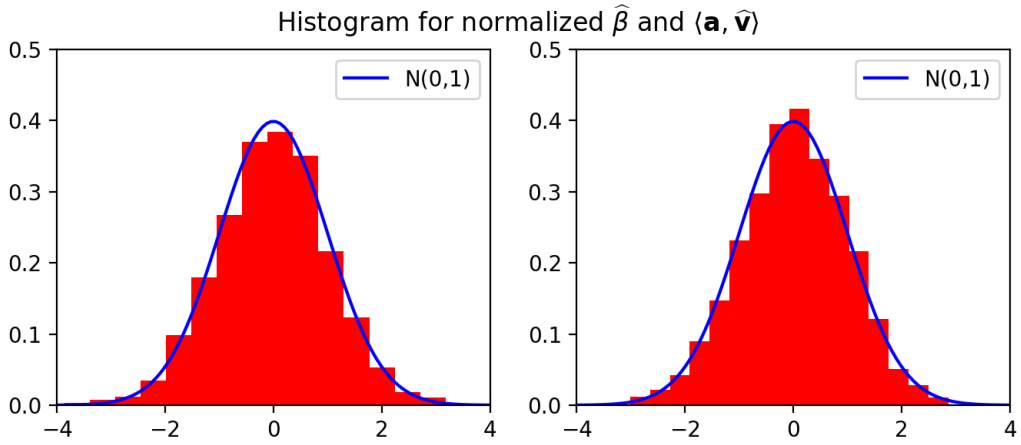

Figure 1: The empirical density of normalized $\widehat{\beta}$ as in (32) (left panel), and normalized $\langle \boldsymbol{a}, \widehat{\boldsymbol{v}} \rangle$ as in (33). The results are reported over 2000 independent trials where the initialization of our power iteration algorithm $\boldsymbol{u}$ a random vector sampled from the unit sphere in $\mathbb{R}^n$, and the strength of signal $\beta = n^{(k-2)/2} \approx 24.495$.

For the setting without prior information of the signal, we take the initialization of our power iteration algorithm $\boldsymbol{u}$ a random vector sampled from the unit sphere in $\mathbb{R}^n$, and the strength of signal $\beta = n^{(k-2)/2} \approx 24.495$. We plot in Figure 1 our estimators for the strength of signals after normalization

$$\widehat{\beta} + \frac{k/2-1}{\widehat{\beta}} - \beta \tag{32}$$

and our estimators for the linear functionals of the signals

$$\frac{\sqrt{n}\widehat{\beta}}{\sqrt{\langle \boldsymbol{a}, (\boldsymbol{I}_n - \widehat{\boldsymbol{v}}\widehat{\boldsymbol{v}}^\top)\boldsymbol{a} \rangle}} \left[\big(1 - \frac{1}{2\widehat{\beta}^2}\big)^{-1} \langle \boldsymbol{a}, \widehat{\boldsymbol{v}} \rangle - \langle \boldsymbol{a}, \boldsymbol{v} \rangle \right] \tag{33}$$

as in Corollary 2.3.

For the setting that there is prior information of the signal, we take the initilization of our power iteration algorithm $\boldsymbol{u} = (\boldsymbol{v} + \boldsymbol{w})/\|\boldsymbol{v} + \boldsymbol{w}\|_2$, where $\boldsymbol{v}$ is a random vector sampled from the unit sphere in $\mathbb{R}^n$. We plot our estimators for the strength of signals after normalization (32) and our estimators for the linear functionals of the signals (33) for $\beta = 5$ in Figure 2, and for $\beta = 10$ in Figure 3. Although our Theorem 2.1 and Corollary 2.3 requires $|\beta \langle \boldsymbol{u}, \boldsymbol{v} \rangle^{k-2}| \geqslant n^{\varepsilon} \gg 1$, Figures 2 and 3 indicate that our estimators $\widehat{\beta}$ and $\langle \boldsymbol{a}, \widehat{\boldsymbol{v}} \rangle$ are asymptotically Gaussian even with small β, i.e. $\beta = 5, 10$. Theorem 2.1 also indicates that error term in Corollary (2.3), i.e. the error term in (13), is of order $1/|\beta|$. This matches with our simulation. In Figures 2 and 3, the the difference between the Gaussian fit of our empirical density and the density of $\mathcal{N}(0,1)$ decreases as β increases from 5 to 10.

In Figure 5, we test the threshold signal-to-noise ratio for the power iteration algorithm. Our Theorems 2.1 and 2.2 state that for $|\beta \langle \boldsymbol{u}_0, \boldsymbol{v} \rangle^{k-2}| \gg 1$ tensor power iteration recovers the signal $\boldsymbol{v}$, and fails when $|\beta \langle \boldsymbol{u}_0, \boldsymbol{v} \rangle^{k-2}| \ll 1$. Especially for random initialization, we have that $|\langle \boldsymbol{u}_0, \boldsymbol{v} \rangle| \asymp 1/\sqrt{n}$. Our Theorems state that for $|\beta| \gg n^{(k-2)/2}$ tensor power iteration recovers the signal $\boldsymbol{v}$, and fails when $|\beta| \ll n^{(k-2)/2}$. Take $k = 3$. In the left panel of Figure 5, we test tensor power iteration

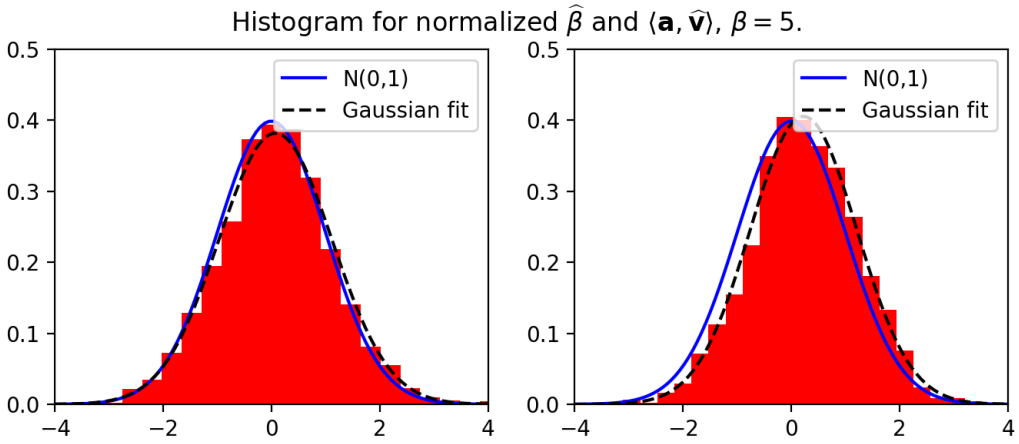

Figure 2: The empirical density of normalized $\widehat{\beta}$ as in (32) (left panel), and normalized $\langle \boldsymbol{a}, \widehat{\boldsymbol{v}} \rangle$ as in (33). The results are reported over 2000 independent trials where the initialization of our power iteration algorithm $\boldsymbol{u}$ a random vector sampled from the unit sphere in $\mathbb{R}^n$, and the strength of signal $\beta = 5$.

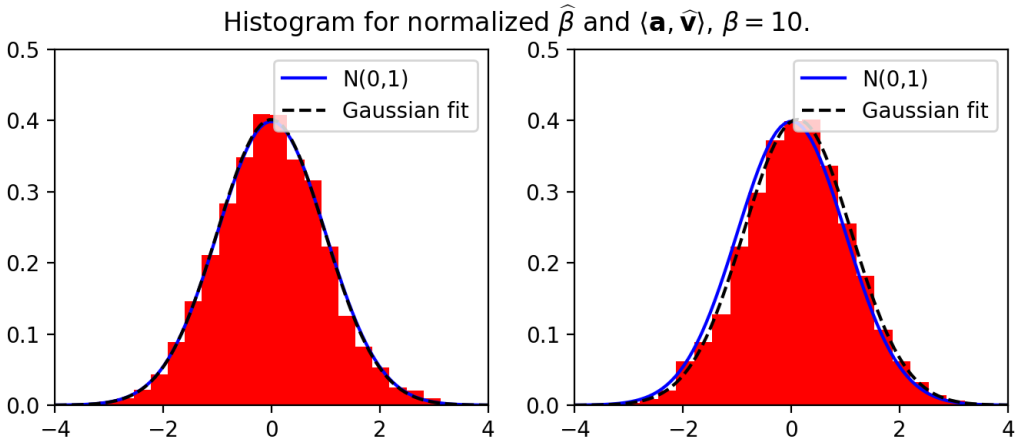

Figure 3: The empirical density of normalized $\widehat{\beta}$ as in (32) (left panel), and normalized $\langle \boldsymbol{a}, \widehat{\boldsymbol{v}} \rangle$ as in (33). The results are reported over 2000 independent trials where the initialization of our power iteration algorithm $\boldsymbol{u}$ a random vector sampled from the unit sphere in $\mathbb{R}^n$, and the strength of signal $\beta = 10$.

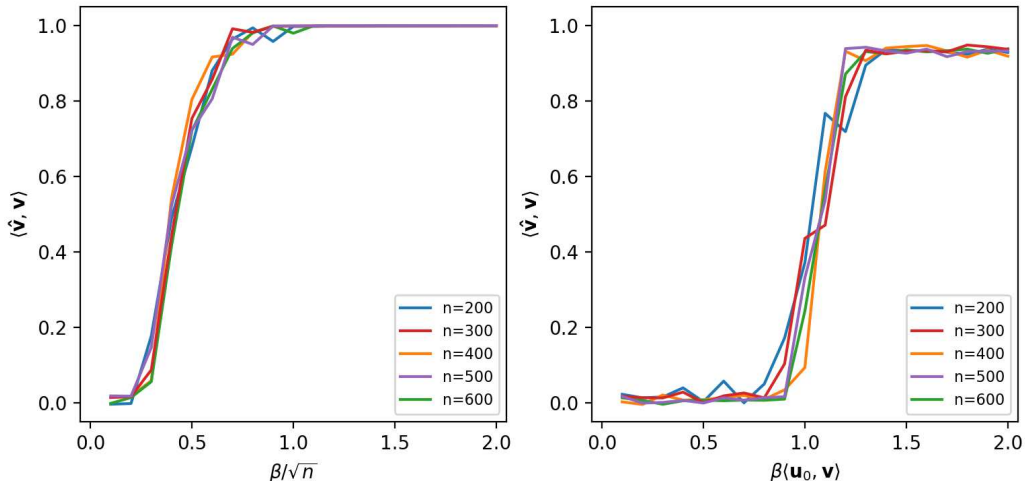

Figure 4: Output of tensor power iteration with random initialization for various signal strength $\beta/\sqrt{n} \in (0, 2]$ (left panel), and tensor power iteration with fixed small $\beta = 3$ and informative initialization $\beta\langle \boldsymbol{u}_0, \boldsymbol{v}\rangle \in (0, 2]$.

with random initialization for various dimensions $n \in \{200, 300, 400, 500, 600\}$ and signal strength $\beta/\sqrt{n} \in (0, 2]$. In the right panel of Figure 5, we test tensor power iteration with fixed small $\beta = 3$ and informative initialization $\beta\langle \boldsymbol{u}_0, \boldsymbol{v}\rangle \in (0, 2]$ for various dimensions $n \in \{200, 300, 400, 500, 600\}$. The outputs $\langle \widehat{\boldsymbol{v}}, \boldsymbol{v}\rangle$ are averaged over 60 independent trials.

3.2 Rank-r spiked tensor model

In this section, we conduct numerical experiments to demonstrate our distributional results for the multi-rank spiked tensor model. We consider the simplest case that there are two spikes with signals $\boldsymbol{v}_1, \boldsymbol{v}_2$, such that they are uniformly sampled from the unit sphere in $\mathbb{R}^n$ and orthogonal to each other $\langle \boldsymbol{v}_1, \boldsymbol{v}_2\rangle = 0$, and the vector

$$\boldsymbol{a} = \frac{1}{\sqrt{3}}(\boldsymbol{e}_{n/3} + \boldsymbol{e}_{2n/3} + \boldsymbol{e}_n). \tag{34}$$

We test the setting that there is no prior information of the signal. We take the strength of signals $\beta_1 = 1.2 \times n^{(k-2)/2} \approx 29.394$ and $\beta_2 = n^{(k-2)/2} \approx 24.495$ and the initialization of our power iteration algorithm $\boldsymbol{u}$ a random vector sampled from the unit sphere in $\mathbb{R}^n$. We scatter plot in Figure 5 our estimator $\widehat{\beta}$ for the strength of signals, and our estimator $\langle \boldsymbol{a}, \widehat{\boldsymbol{v}}\rangle$ for the linear functionals of the signals over 5000 independent trials. As seen in the first panel of Figure 5, our estimators $(\widehat{\beta}, \langle \boldsymbol{a}, \widehat{\boldsymbol{v}}\rangle)$ form two clusters, centered around $(\beta_1, \langle \boldsymbol{a}, \widehat{\boldsymbol{v}}_1\rangle) \approx (29.394, 0.000)$ and $(\beta_2, \langle \boldsymbol{a}, \widehat{\boldsymbol{v}}_2\rangle) \approx (24.495, 0.039)$. In the second and third panels, we zoom in, and scatter plot for the cluster corresponding to $(\beta_1, \langle \boldsymbol{a}, \widehat{\boldsymbol{v}}_1\rangle) \approx (29.394, 0.000)$

$$\widehat{\beta} + \frac{k/2 - 1}{\widehat{\beta}} - \beta, \quad \frac{\sqrt{n}\widehat{\beta}_1}{\sqrt{\langle \boldsymbol{a}, (\boldsymbol{I}_n - \widehat{\boldsymbol{v}}\widehat{\boldsymbol{v}}^\top)\boldsymbol{a}\rangle}} \left[\left(1 - \frac{1}{2\widehat{\beta}^2}\right)^{-1} \langle \boldsymbol{a}, \widehat{\boldsymbol{v}}\rangle - \langle \boldsymbol{a}, \boldsymbol{v}_1\rangle \right], \tag{35}$$

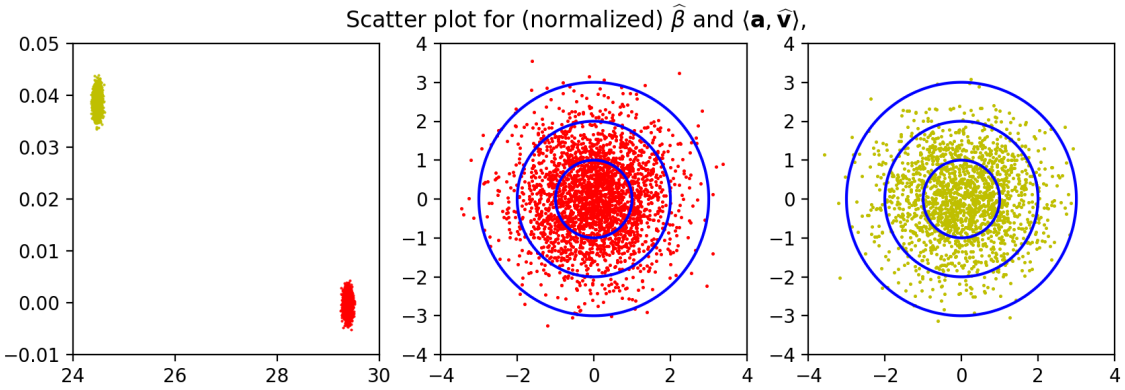

Figure 5: Scatter plot of $(\widehat{\beta}, \langle \boldsymbol{a}, \widehat{\boldsymbol{v}} \rangle)$ (first panel), the normalized $(\widehat{\beta}, \langle \boldsymbol{a}, \widehat{\boldsymbol{v}} \rangle)$ as in (35) for the cluster corresponding to $(\beta_1, \langle \boldsymbol{a}, \boldsymbol{v}_1 \rangle)$ (second panel), the normalized $(\widehat{\beta}, \langle \boldsymbol{a}, \widehat{\boldsymbol{v}} \rangle)$ as in (36) for the cluster corresponding to $(\beta_2, \langle \boldsymbol{a}, \boldsymbol{v}_2 \rangle)$. The contour plot is a standard 2-dim Gaussian distribution, at $1, 2, 3$ standard deviation. The results are reported over 5000 independent trials where the initialization of our power iteration algorithm $\boldsymbol{u}$ a random vector sampled from the unit sphere in $\mathbb{R}^n$.

and scatter plot for the cluster corresponding to $(\beta_2, \langle \boldsymbol{a}, \widehat{\boldsymbol{v}}_2 \rangle) \approx (24.495, 0.039)$

$$\widehat{\beta} + \frac{k/2-1}{\widehat{\beta}} - \beta_2, \quad \frac{\sqrt{n}\widehat{\beta}}{\sqrt{\langle \boldsymbol{a}, (\boldsymbol{I}_n - \widehat{\boldsymbol{v}}\widehat{\boldsymbol{v}}^\top)\boldsymbol{a} \rangle}} \left[\big(1 - \frac{1}{2\widehat{\beta}^2}\big)^{-1} \langle \boldsymbol{a}, \widehat{\boldsymbol{v}} \rangle - \langle \boldsymbol{a}, \boldsymbol{v}_2 \rangle \right]. \tag{36}$$

As predicted by our Theorem 2.9, both clusters are asymptotically Gaussian, and the normalized estimators matches pretty well with the contour plot of standard 2-dim Gaussian distribution, at $1, 2, 3$ standard deviation.

We plot in Figure 6 our estimators for the strength of signals and the linear functionals of the signals after normalization, for the first cluster (35), and for the second cluster (36).

In Table (1), for each $n \in \{50, 100, 200, 400, 600, 800\}$ and $k = 3$, we take the strength of signals $\beta_1 = n^{(k-2)/2}$ and $\beta_2 = 1.2 \times n^{(k-2)/2}$. Over 1000 independent trials for power iteration with random initialization for each n, we estimate the percentage $\widehat{p}_1$ of estimators converging to β_1, and the percentage $\widehat{p}_2$ of estimators converging to β_2. Our theoretical values are

$$p_1 = \mathbb{P}(|\beta_1 \langle \boldsymbol{u}, \boldsymbol{v}_1 \rangle| > |\beta_2 \langle \boldsymbol{u}, \boldsymbol{v}_2 \rangle|) \approx 0.44,$$
$$p_2 = \mathbb{P}(|\beta_1 \langle \boldsymbol{u}, \boldsymbol{v}_1 \rangle| < |\beta_2 \langle \boldsymbol{u}, \boldsymbol{v}_2 \rangle|) \approx 0.56.$$

We also exam the numerical coverage rates for our 95% confidence intervals over 1000 independent trials.

4 Proof of main theorems

4.1 Proof of Theorems 2.1 and 2.2

The following lemma on the conditioning of Gaussian tensors will be repeatedly use in the remaining of this section.

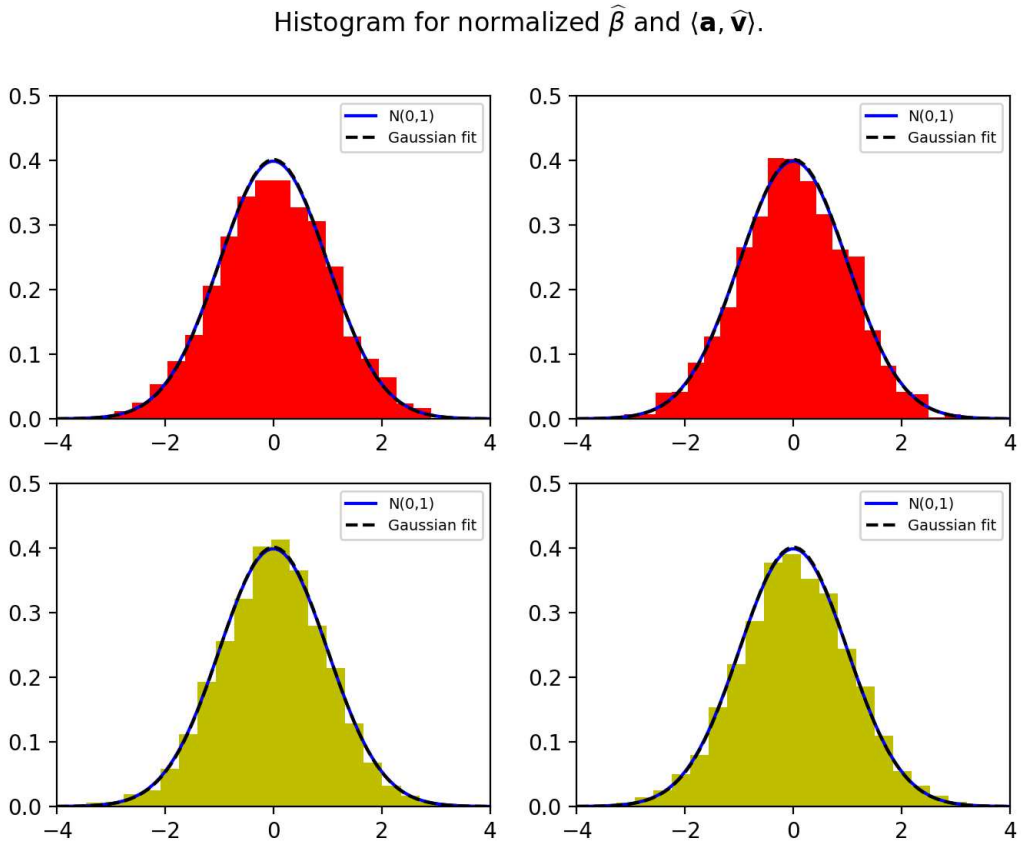

Figure 6: The empirical density of the normalized $(\widehat{\beta}, \langle \boldsymbol{a}, \widehat{\boldsymbol{v}} \rangle)$ as in (35) for the cluster corresponding to $(\beta_1, \langle \boldsymbol{a}, \boldsymbol{v}_1 \rangle)$ (second panel), the normalized $(\widehat{\beta}, \langle \boldsymbol{a}, \widehat{\boldsymbol{v}} \rangle)$ as in (36) for the cluster corresponding to $(\beta_2, \langle \boldsymbol{a}, \boldsymbol{v}_2 \rangle)$. The results are reported over 5000 independent trials where the initialization of our power iteration algorithm $\boldsymbol{u}$ a random vector sampled from the unit sphere in $\mathbb{R}^n$.

	$n = 50$	$n = 100$	$n = 200$	$n = 400$	$n = 600$	$n = 800$
$\widehat{p}_1$	0.405	0.399	0.421	0.381	0.422	0.401
$\widehat{p}_2$	0.595	0.579	0.601	0.619	0.578	0.599
signal β_1	0.9136	0.9223	0.9596	0.9291	0.9313	0.9551
linear form $\langle \boldsymbol{a}, \boldsymbol{v}_1 \rangle$	0.9680	0.9499	0.9572	0.9580	0.9668	0.9526
signal β_2	0.9462	0.9334	0.9430	0.9612	0.9602	0.9599
linear form $\langle \boldsymbol{a}, \boldsymbol{v}_2 \rangle$	0.9445	0.9434	0.94819	0.9677	0.9533	0.9549

Table 1: Estimated $\widehat{p}_1, \widehat{p}_2$ over 1000 independent trials for dimension $n \in \{50, 100, 200, 400, 600, 800\}$ (top two rows), and numerical coverage rates for our 95% confidence intervals over 1000 independent trials for dimension $n \in \{50, 100, 200, 400, 600, 800\}$ (last four rows).

Lemma 4.1. *Let $\boldsymbol{Z} \in \otimes^k \mathbb{R}^n$ be a random Gaussian tensor. The entries of $\boldsymbol{Z}$ are i.i.d. standard $\mathcal{N}(0, 1/n)$ Gaussian random variables. Fix $\boldsymbol{\tau}_1, \boldsymbol{\tau}_2, \cdots, \boldsymbol{\tau}_t \in \otimes^{k-1}\mathbb{R}^n$ orthonormal $(k-1)$-th order tensors, i.e. $\langle \boldsymbol{\tau}_i, \boldsymbol{\tau}_j \rangle = \delta_{ij}$, and vectors $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \cdots, \boldsymbol{\xi}_t \in \mathbb{R}^n$. Then the distribution of $\boldsymbol{Z}[\boldsymbol{\tau}]$ conditioned on $\boldsymbol{Z}[\boldsymbol{\tau}_s] = \boldsymbol{\xi}_s$ for $1 \leqslant s \leqslant t$ is*

$$\boldsymbol{Z}[\boldsymbol{\tau}] \stackrel{d}{=} \sum_{s=1}^{t} \langle \boldsymbol{\tau}_s, \boldsymbol{\tau} \rangle \boldsymbol{\xi}_s + \tilde{\boldsymbol{Z}} \left[\boldsymbol{\tau} - \sum_{s=1}^{t} \langle \boldsymbol{\tau}_s, \boldsymbol{\tau} \rangle \boldsymbol{\tau}_s \right],$$

where $\tilde{\boldsymbol{Z}}$ is an independent copy of $\boldsymbol{Z}$.

Proof of Lemma 4.1. For any $(k-1)$-th order tensor $\boldsymbol{\tau}$, viewed as a vector in $\mathbb{R}^{n^{k-1}}$, we can decompose it as the projection on the span of $\boldsymbol{\tau}_1, \boldsymbol{\tau}_2, \cdots, \boldsymbol{\tau}_t$ and the orthogonal part

$$\boldsymbol{\tau} = \sum_{s=1}^{t} \langle \boldsymbol{\tau}_s, \boldsymbol{\tau} \rangle \boldsymbol{\tau}_s + \left(\boldsymbol{\tau} - \sum_{s=1}^{t} \langle \boldsymbol{\tau}_s, \boldsymbol{\tau} \rangle \boldsymbol{\tau}_s \right). \tag{37}$$

Using the above decomposition and $\boldsymbol{Z}[\boldsymbol{\tau}_s] = \boldsymbol{\xi}_s$, we can write $\boldsymbol{Z}[\boldsymbol{\tau}]$ as

$$\boldsymbol{Z}[\boldsymbol{\tau}] \stackrel{\mathrm{d}}{=} \sum_{s=1}^{t} \langle \boldsymbol{\tau}_s, \boldsymbol{\tau} \rangle \boldsymbol{\xi}_s + \boldsymbol{Z} \left[\boldsymbol{\tau} - \sum_{s=1}^{t} \langle \boldsymbol{\tau}_s, \boldsymbol{\tau} \rangle \boldsymbol{\tau}_s \right], \tag{38}$$

and the first sum and the second term on the righthand side of (38) are independent. The claim (37) follows. □

Proof of Theorem 2.1. We define an auxiliary iteration, $\boldsymbol{y}_0 = \boldsymbol{u}$ and

$$\boldsymbol{y}_{t+1} = \boldsymbol{X}[\boldsymbol{y}_t^{\otimes(k-1)}]. \tag{39}$$

Then with $\boldsymbol{y}_t$, our original power iteration (2) is given by $\boldsymbol{u}_t = \boldsymbol{y}_t / \|\boldsymbol{y}_t\|_2$.

Let $\boldsymbol{\xi} = \boldsymbol{Z}[\boldsymbol{v}^{\otimes(k-1)}] \in \mathbb{R}^n$. Then the entries of $\boldsymbol{\xi}$ are given by

$$\boldsymbol{\xi}(i) = \boldsymbol{Z}[\boldsymbol{v}^{\otimes(k-1)}](i) = \langle \boldsymbol{Z}, \boldsymbol{e}_i \otimes \boldsymbol{v}^{\otimes(k-1)} \rangle = \sum_{i_1, i_2, \cdots, i_{k-1} \in [\![1,n]\!]} \boldsymbol{Z}_{i i_1 i_2 \cdots i_{k-1}} \boldsymbol{v}(i_1) \boldsymbol{v}(i_2) \cdots \boldsymbol{v}(i_{k-1}). \tag{40}$$

From the expression, $\boldsymbol{\xi}(i)$ is a linear combination of Gaussian random variables, itself is also a Gaussian. Moreover, these entries $\boldsymbol{\xi}(i)$ are i.i.d. Gaussian variables with mean zero and variance $1/n$:

$$\mathbb{E}[\boldsymbol{\xi}(i)^2] = \sum_{i_1, i_2, \cdots, i_{k-1} \in [\![1,n]\!]} \mathbb{E}[\boldsymbol{Z}_{i i_1 i_2 \cdots i_{k-1}}^2] \boldsymbol{v}(i_1)^2 \boldsymbol{v}(i_2)^2 \cdots \boldsymbol{v}(i_{k-1})^2 = \frac{1}{n}. \tag{41}$$

We can compute $\boldsymbol{y}_t$ iteratively: $\boldsymbol{y}_1$ is given by

$$\boldsymbol{y}_1 = \boldsymbol{X}[\boldsymbol{y}_0^{\otimes(k-1)}] = \beta \langle \boldsymbol{y}_0, \boldsymbol{v} \rangle^{k-1} \boldsymbol{v} + \boldsymbol{Z}[\boldsymbol{y}_0^{\otimes(k-1)}]. \tag{42}$$

For the last term on the righthand side of (42), we can decompose $\boldsymbol{y}_0^{\otimes(k-1)}$ as a projection on $\boldsymbol{v}^{\otimes(k-1)}$ and its orthogonal part:

$$\boldsymbol{y}_0^{\otimes(k-1)} = \langle \boldsymbol{y}_0, \boldsymbol{v} \rangle^{k-1} \boldsymbol{v}^{\otimes(k-1)} + \sqrt{1 - \langle \boldsymbol{y}_0, \boldsymbol{v} \rangle^{2(k-1)}} \boldsymbol{\tau}_0, \tag{43}$$

where $\boldsymbol{\tau}_0 \in \otimes^{(k-1)}\mathbb{R}^n$ and $\langle \boldsymbol{v}^{\otimes(k-1)}, \boldsymbol{\tau}_0\rangle = 0$, $\langle \boldsymbol{\tau}_0, \boldsymbol{\tau}_0\rangle = 1$. Thanks to Lemma 4.1, conditioning on $\boldsymbol{\xi} = \boldsymbol{Z}[\boldsymbol{v}^{\otimes(k-1)}]$, $\boldsymbol{\xi}_1 = \boldsymbol{Z}[\boldsymbol{\tau}_0]$ has the same law as $\tilde{\boldsymbol{Z}}[\boldsymbol{\tau}_0]$, where $\tilde{\boldsymbol{Z}}$ is an independent copy of $\boldsymbol{Z}$. Since $\langle \boldsymbol{\tau}_0, \boldsymbol{\tau}_0\rangle = 1$, $\boldsymbol{\xi}_1$ is a Gaussian vector with each entry $\mathcal{N}(0, 1/n)$. With those notations we can rewrite the expression (42) of $\boldsymbol{y}_1$ as

$$\boldsymbol{y}_1 = \beta\langle \boldsymbol{y}_0, \boldsymbol{v}\rangle^{k-1}\boldsymbol{v} + \langle \boldsymbol{y}_0, \boldsymbol{v}\rangle^{k-1}\boldsymbol{\xi} + \sqrt{1 - \langle \boldsymbol{y}_0, \boldsymbol{v}\rangle^{2(k-1)}}\boldsymbol{\xi}_1. \tag{44}$$

In the following we show that:

Claim 4.2. *We can compute $\boldsymbol{y}_1, \boldsymbol{y}_2, \boldsymbol{y}_3, \cdots, \boldsymbol{y}_t$ inductively. The Gram-Schmidt orthonormalization procedure gives an orthogonal base of $\boldsymbol{v}^{\otimes(k-1)}, \boldsymbol{y}_0^{\otimes(k-1)}, \boldsymbol{y}_1^{\otimes(k-1)}, \cdots, \boldsymbol{y}_{t-1}^{\otimes(k-1)}$ as:*

$$\boldsymbol{v}^{\otimes(k-1)}, \boldsymbol{\tau}_0, \boldsymbol{\tau}_1, \cdots, \boldsymbol{\tau}_{t-1}. \tag{45}$$

Let $\boldsymbol{\xi}_{s+1} = \boldsymbol{Z}[\boldsymbol{\tau}_s]$ for $0 \leqslant s \leqslant t-1$. Conditioning on $\boldsymbol{\xi} = \boldsymbol{Z}[\boldsymbol{v}^{\otimes(k-1)}]$ and $\boldsymbol{\xi}_{s+1} = \boldsymbol{Z}[\boldsymbol{\tau}_s]$ for $0 \leqslant s \leqslant t-2$, $\boldsymbol{\xi}_t = \boldsymbol{Z}[\boldsymbol{\tau}_{t-1}]$ is an independent Gaussian vector, with each entry $\mathcal{N}(0, 1/n)$. Then $\boldsymbol{y}_t$ is in the following form

$$\boldsymbol{y}_t = a_t\boldsymbol{v} + b_t\boldsymbol{w}_t + c_t\boldsymbol{\xi}_t, \quad b_t\boldsymbol{w}_t = b_{t0}\boldsymbol{\xi} + b_{t1}\boldsymbol{\xi}_1 + \cdots + b_{tt-1}\boldsymbol{\xi}_{t-1}, \tag{46}$$

where $\|\boldsymbol{w}_t\|_2 = 1$.

Proof of Claim 4.2. The Claim 4.2 for $t = 1$ follows from (44). In the following, assuming Claim 4.2 holds for t, we prove it for $t+1$.

Let $\boldsymbol{v}^{\otimes(k-1)}, \boldsymbol{\tau}_0, \boldsymbol{\tau}_1, \cdots, \boldsymbol{\tau}_t$ be an orthogonal base for $\boldsymbol{v}^{\otimes(k-1)}, \boldsymbol{y}_0^{\otimes(k-1)}, \boldsymbol{y}_1^{\otimes(k-1)}, \cdots, \boldsymbol{y}_t^{\otimes(k-1)}$, obtained by the Gram-Schmidt orthonormalization procedure. More precisely, given those tensors $\boldsymbol{v}^{\otimes(k-1)}, \boldsymbol{\tau}_0, \boldsymbol{\tau}_1, \cdots, \boldsymbol{\tau}_{t-1}$, we denote

$$\begin{aligned} &b_{(t+1)0} = \langle \boldsymbol{y}_t^{\otimes(k-1)}, \boldsymbol{v}^{\otimes(k-1)}\rangle, \quad c_{t+1} = \langle \boldsymbol{y}_t^{\otimes(k-1)}, \boldsymbol{\tau}_t\rangle, \\ &b_{(t+1)(s+1)} = \langle \boldsymbol{y}_t^{\otimes(k-1)}, \boldsymbol{\tau}_s\rangle, \quad 0 \leqslant s \leqslant t-1. \end{aligned} \tag{47}$$

then $b_{(t+1)0}\boldsymbol{v}^{\otimes(k-1)} + b_{(t+1)1}\boldsymbol{\tau}_0 + b_{(t+1)2}\boldsymbol{\tau}_1 + \cdots b_{(t+1)t}\boldsymbol{\tau}_{t-1}$ is the projection of $\boldsymbol{y}_t^{\otimes(k-1)}$ on the span of $\boldsymbol{v}^{\otimes(k-1)}, \boldsymbol{y}_0^{\otimes(k-1)}, \boldsymbol{y}_1^{\otimes(k-1)}, \cdots, \boldsymbol{y}_{t-1}^{\otimes(k-1)}$. With those notations, we can write $\boldsymbol{y}_t^{\otimes(k-1)}$ as

$$\boldsymbol{y}_t^{\otimes(k-1)} = b_{(t+1)0}\boldsymbol{v}^{\otimes(k-1)} + b_{(t+1)1}\boldsymbol{\tau}_0 + b_{(t+1)2}\boldsymbol{\tau}_1 + \cdots b_{(t+1)t}\boldsymbol{\tau}_{t-1} + c_{t+1}\boldsymbol{\tau}_t, \tag{48}$$

Using (46) and (48), we notice that

$$\langle \beta\boldsymbol{v}^{\otimes k-1}, \boldsymbol{y}_t^{\otimes(k-1)}\rangle = \beta(a_t + b_t\langle \boldsymbol{w}_t, \boldsymbol{v}\rangle + c_t\langle \boldsymbol{\xi}_t, \boldsymbol{v}\rangle)^{k-1}\boldsymbol{v}, \tag{49}$$

and the iteration (39) implies that

$$\boldsymbol{y}_{t+1} = \beta(a_t + b_t\langle \boldsymbol{w}_t, \boldsymbol{v}\rangle + c_t\langle \boldsymbol{\xi}_t, \boldsymbol{v}\rangle)^{k-1}\boldsymbol{v} + b_{t+1}\boldsymbol{w}_{t+1} + c_{t+1}\boldsymbol{Z}[\boldsymbol{\tau}_t], \tag{50}$$

where

$$\begin{aligned} b_{t+1}\boldsymbol{w}_{t+1} &= \boldsymbol{Z}[b_{(t+1)0}\boldsymbol{v}^{\otimes(k-1)} + b_{(t+1)1}\boldsymbol{\tau}_0 + b_{(t+1)2}\boldsymbol{\tau}_1 + \cdots b_{(t+1)t}\boldsymbol{\tau}_{t-1}] \\ &= b_{(t+1)0}\boldsymbol{\xi} + b_{(t+1)1}\boldsymbol{\xi}_1 + b_{(t+1)2}\boldsymbol{\xi}_2 + \cdots b_{(t+1)t}\boldsymbol{\xi}_t. \end{aligned} \tag{51}$$

Since $\boldsymbol{\tau}_t$ is orthogonal to $\boldsymbol{v}^{\otimes(k-1)}, \boldsymbol{\tau}_0, \boldsymbol{\tau}_1, \cdots, \boldsymbol{\tau}_{t-1}$, Lemma 4.1 implies that conditioning on $\boldsymbol{\xi} = \boldsymbol{Z}[\boldsymbol{v}^{\otimes(k-1)}]$ and $\boldsymbol{\xi}_{s+1} = \boldsymbol{Z}[\boldsymbol{\tau}_s]$ for $0 \leqslant s \leqslant t-1$, $\boldsymbol{\xi}_{t+1} = \boldsymbol{Z}[\boldsymbol{\tau}_t]$ is an independent Gaussian vector, with each entry $\mathcal{N}(0, 1/n)$. The above discussion gives us that

$$\boldsymbol{y}_{t+1} = a_{t+1}\boldsymbol{v} + b_{t+1}\boldsymbol{w}_{t+1} + c_{t+1}\boldsymbol{\xi}_{t+1}, \quad a_{t+1} = \beta(a_t + b_t\langle \boldsymbol{w}_t, \boldsymbol{v}\rangle + c_t\langle \boldsymbol{\xi}_t, \boldsymbol{v}\rangle)^{k-1}. \tag{52}$$

In this way, for any $t \geqslant 0$, $\boldsymbol{y}_t$ is given in the form (46). □

In the following, We study the case that $\langle \boldsymbol{u}, \boldsymbol{v}\rangle > 0$. The case $\langle \boldsymbol{u}, \boldsymbol{v}\rangle < 0$ can be proven in exactly the same way, by simply changing $(\beta, \boldsymbol{v})$ with $((-1)^k\beta, -\boldsymbol{v})$. We prove by induction

Claim 4.3. *For any fixed time t, with probability at least $1 - \mathrm{O}(e^{-c(\log N)^2})$ the following holds: for any $s \leqslant t$,*

$$\begin{aligned} &|a_s| \gtrsim |\beta|(|b_{s0}| + |b_{s1}| + \cdots + |b_{s(s-1)}|), \\ &|a_s| \gtrsim n^{\varepsilon} \max\{\mathbf{1}(k \geqslant 3)|c_s/\beta^{1/(k-2)}|, |c_s/\sqrt{n}|\}. \end{aligned} \tag{53}$$

and

$$\begin{aligned} &\|\boldsymbol{\xi}\|, \|\boldsymbol{\xi}_s\|_2 = 1 + \mathrm{O}(\log n/\sqrt{n}), \quad |\langle \boldsymbol{v}, \boldsymbol{\xi}\rangle|, |\langle \boldsymbol{a}, \boldsymbol{\xi}\rangle|, |\langle \boldsymbol{a}, \boldsymbol{\xi}_s\rangle|, \\ &\|\mathrm{Proj}_{\mathrm{Span}\{\boldsymbol{v}, \boldsymbol{\xi}, \boldsymbol{\xi}_1, \cdots, \cdots, \boldsymbol{\xi}_{s-1}\}}(\boldsymbol{\xi}_s)\|_2 \lesssim \log n/\sqrt{n}. \end{aligned} \tag{54}$$

Proof of Claim 4.3. From (44), $\boldsymbol{y}_1 = \beta\langle \boldsymbol{u}, \boldsymbol{v}\rangle^{k-1}\boldsymbol{v} + \langle \boldsymbol{u}, \boldsymbol{v}\rangle^{k-1}\boldsymbol{\xi} + \sqrt{1 - \langle \boldsymbol{u}, \boldsymbol{v}\rangle^{2(k-1)}}\boldsymbol{\xi}_1$. We have $a_1 = \beta\langle \boldsymbol{u}, \boldsymbol{v}\rangle^{k-1}$, $b_{10} = \langle \boldsymbol{u}, \boldsymbol{v}\rangle^{k-1}$, $b_1\boldsymbol{w}_1 = \langle \boldsymbol{u}, \boldsymbol{v}\rangle^{k-1}\boldsymbol{\xi}$ and $c_1 = \sqrt{1 - \langle \boldsymbol{u}, \boldsymbol{v}\rangle^{2(k-1)}}$. Since $\boldsymbol{\xi}$ is a Gaussian vector with each entry mean zero and variance $1/n$, the concentration for chi-square distribution implies that

$$\|\boldsymbol{\xi}\|_2 = \sqrt{\sum_{i=1}^{n} \boldsymbol{\xi}(i)^2} = 1 + \mathrm{O}(\log n/\sqrt{n}) \tag{55}$$

with probability $1 - e^{c(\log n)^2}$. We can check that $|a_1| = |\beta b_{10}|$, $|\beta^{1/(k-2)}a_1| = |\beta\langle \boldsymbol{u}, \boldsymbol{v}\rangle^{k-2}|^{(k-1)/(k-2)} \gtrsim n^{(k-1)\varepsilon/(k-2)} \geqslant n^{\varepsilon}|c_1|$, and $|\sqrt{n}a_1| = |\beta\langle \boldsymbol{u}, \boldsymbol{v}\rangle^{k-2}||\sqrt{n}\langle \boldsymbol{u}, \boldsymbol{v}\rangle| \gtrsim n^{\varepsilon} \geqslant n^{\varepsilon}|c_1|$. Moreover, conditioning on $\boldsymbol{Z}[\boldsymbol{v}^{\otimes(k-1)}] = \boldsymbol{\xi}$, Lemma 4.1 implies that $\boldsymbol{\xi}_1 = \boldsymbol{Z}[\boldsymbol{\tau}_0]$ is an independent Gaussian random vector with each entry $\mathcal{N}(0, 1/n)$. By the standard concentration inequality, it holds that with probability $1 - e^{c(\log n)^2}$, $\|\boldsymbol{\xi}_1\|_2 = 1 + \mathrm{O}(\log n/\sqrt{n})$, $|\langle \boldsymbol{a}, \boldsymbol{\xi}_1\rangle|$ and the projection of $\boldsymbol{\xi}_1$ on the span of $\{\boldsymbol{v}, \boldsymbol{\xi}\}$ is bounded by $\log n/\sqrt{n}$. So far we have proved that (53) and (54) for $t = 1$.

In the following, we assume that (53) holds for t, and prove it for $t + 1$. We recall from (46) and (52) that

$$a_{t+1} = \beta(a_t + b_t\langle \boldsymbol{w}_t, \boldsymbol{v}\rangle + c_t\langle \boldsymbol{\xi}_t, \boldsymbol{v}\rangle)^{k-1}, \quad b_t\boldsymbol{w}_t = b_{t0}\boldsymbol{\xi} + b_{t1}\boldsymbol{\xi}_1 + \cdots + b_{tt-1}\boldsymbol{\xi}_{t-1} \tag{56}$$

By our induction hypothesis, we have that

$$|b_t\langle \boldsymbol{w}_t, \boldsymbol{v}\rangle| \lesssim |b_{t0}\langle \boldsymbol{\xi}, \boldsymbol{v}\rangle| + |b_{t1}\langle \boldsymbol{\xi}_1, \boldsymbol{v}\rangle| + \cdots + |b_{t(t-1)}\langle \boldsymbol{\xi}_{t-1}, \boldsymbol{v}\rangle| \lesssim (\log n/\sqrt{n})|a_t|/|\beta|, \tag{57}$$

and

$$|c_t\langle \boldsymbol{\xi}_t, \boldsymbol{v}\rangle| \lesssim (\log n/\sqrt{n})|c_t| \lesssim (\log n)|a_t|/n^{\varepsilon}. \tag{58}$$

It follows from plugging (57) and (58) into (56), we get

$$a_{t+1} = \beta(a_t + \mathrm{O}(\log n|a_t|/n^{\varepsilon}))^{k-1} = (1 + \mathrm{O}(\log n/n^{\varepsilon}))\beta a_t^{k-1}. \tag{59}$$

We recall from (48), the coefficients $b_{(t+1)0}, b_{(t+1)1}, \cdots, b_{(t+1)t}$ are determined from the projection of $\boldsymbol{y}_t^{\otimes(k-1)}$ on $\boldsymbol{v}^{\otimes(k-1)}, \boldsymbol{\tau}_0, \boldsymbol{\tau}_1, \cdots, \boldsymbol{\tau}_{t-1}$

$$\boldsymbol{y}_t^{\otimes(k-1)} = b_{(t+1)0}\boldsymbol{v}^{\otimes(k-1)} + b_{(t+1)1}\boldsymbol{\tau}_0 + b_{(t+1)2}\boldsymbol{\tau}_1 + \cdots b_{(t+1)t}\boldsymbol{\tau}_{t-1} + c_{t+1}\boldsymbol{\tau}_t. \tag{60}$$

We also recall that $\boldsymbol{v}^{\otimes(k-1)}, \boldsymbol{\tau}_0, \boldsymbol{\tau}_1, \cdots, \boldsymbol{\tau}_{t-1}$ are obtained from $\boldsymbol{v}^{\otimes(k-1)}, \boldsymbol{y}_0^{\otimes(k-1)}, \boldsymbol{y}_1^{\otimes(k-1)}, \cdots, \boldsymbol{y}_{t-1}^{\otimes(k-1)}$ by the Gram-Schmidt orthonormalization procedure. So we have that the span of vectors (viewed as vectors) $\boldsymbol{v}^{\otimes(k-1)}, \boldsymbol{\tau}_0, \boldsymbol{\tau}_1, \cdots, \boldsymbol{\tau}_{t-1}$ is the same as the span of tensors $\boldsymbol{v}^{\otimes(k-1)}, \boldsymbol{y}_0^{\otimes(k-1)}, \boldsymbol{y}_1^{\otimes(k-1)}, \cdots, \boldsymbol{y}_{t-1}^{\otimes(k-1)}$, which is contained in the span of $\{\boldsymbol{v}, \boldsymbol{w}_t, \boldsymbol{y}_0, \cdots, \boldsymbol{y}_{t-1}\}^{\otimes(k-1)}$. Moreover from the relation (46), one can see that the span of $\{\boldsymbol{v}, \boldsymbol{w}_t, \boldsymbol{y}_0, \cdots, \boldsymbol{y}_{t-1}\}$ is the same as the span of $\{\boldsymbol{v}, \boldsymbol{\xi}, \boldsymbol{\xi}_1, \cdots, \boldsymbol{\xi}_{t-1}\}$. It follows that

$$\begin{aligned}
&\sqrt{b_{(t+1)0}^2 + b_{(t+1)1}^2 + b_{(t+1)2}^2 + \cdots + b_{(t+1)t}^2} \\
&= \|\mathrm{Proj}_{\mathrm{Span}\{\boldsymbol{v}^{\otimes(k-1)}, \boldsymbol{\tau}_0, \boldsymbol{\tau}_1, \cdots, \boldsymbol{\tau}_{t-1}\}}(a_t\boldsymbol{v} + b_t\boldsymbol{w}_t + c_t\boldsymbol{\xi}_t)^{\otimes(k-1)}\|_2 \\
&\leqslant \|\mathrm{Proj}_{\mathrm{Span}\{\boldsymbol{v}, \boldsymbol{w}_t, \boldsymbol{y}_0, \cdots, \boldsymbol{y}_{t-1}\}^{\otimes(k-1)}}(a_t\boldsymbol{v} + b_t\boldsymbol{w}_t + c_t\boldsymbol{\xi}_t)^{\otimes(k-1)}\|_2 \\
&\leqslant \|\mathrm{Proj}_{\mathrm{Span}\{\boldsymbol{v}, \boldsymbol{w}_t, \boldsymbol{y}_0, \cdots, \boldsymbol{y}_{t-1}\}}(a_t\boldsymbol{v} + b_t\boldsymbol{w}_t + c_t\boldsymbol{\xi}_t)\|_2^{k-1} \\
&= \|a_t\boldsymbol{v} + b_t\boldsymbol{w}_t + c_t\mathrm{Proj}_{\mathrm{Span}\{\boldsymbol{v}, \boldsymbol{\xi}, \boldsymbol{\xi}_1, \cdots, \boldsymbol{\xi}_{t-1}\}}(\boldsymbol{\xi}_t)\|_2^{k-1} \\
&\lesssim \left(|a_t| + |b_t| + \frac{\log n|c_t|}{\sqrt{n}}\right)^{k-1} \lesssim |a_t|^{k-1} \lesssim |a_{t+1}|/|\beta|,
\end{aligned} \tag{61}$$

where in the last line we used our induction hypothesis that $\|\mathrm{Proj}_{\mathrm{Span}\{\boldsymbol{v}, \boldsymbol{\xi}, \boldsymbol{\xi}_1, \cdots, \boldsymbol{\xi}_{t-1}\}}(\boldsymbol{\xi}_t)\|_2 \lesssim \log n/\sqrt{n}$.

Finally we estimate c_{t+1}. We recall from (48), the coefficient c_{t+1} is the remainder of $\boldsymbol{y}_t^{\otimes(k-1)}$ after projecting on $\boldsymbol{v}^{\otimes(k-1)}, \boldsymbol{\tau}_0, \boldsymbol{\tau}_1, \cdots, \boldsymbol{\tau}_{t-1}$. It is bounded by the remainder of $\boldsymbol{y}_t^{\otimes(k-1)}$ after projecting on $\boldsymbol{v}^{\otimes(k-1)}$,

$$|c_{t+1}| \leqslant \|\boldsymbol{y}_t^{\otimes(k-1)} - a_t^{k-1}\boldsymbol{v}^{\otimes(k-1)}\|_2 = \|(a_t\boldsymbol{v} + b_t\boldsymbol{w}_t + c_t\boldsymbol{\xi}_t)^{\otimes(k-1)} - a_t^{k-1}\boldsymbol{v}^{\otimes(k-1)}\|_2. \tag{62}$$

The difference $(a_t\boldsymbol{v} + b_t\boldsymbol{w}_t + c_t\boldsymbol{\xi}_t)^{\otimes(k-1)} - a_t^{k-1}\boldsymbol{v}^{\otimes(k-1)}$ is a sum of terms in the following form,

$$\boldsymbol{\eta}_1 \otimes \boldsymbol{\eta}_2 \otimes \cdots \otimes \boldsymbol{\eta}_{k-1}, \tag{63}$$

where vectors $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \cdots, \boldsymbol{\eta}_{k-1} \in \{a_t\boldsymbol{v}, b_t\boldsymbol{w}_t + c_t\boldsymbol{\xi}_t\}$, and at least one of them is $b_t\boldsymbol{w}_t + c_t\boldsymbol{\xi}_t$. We notice that by our induction hypothesis, $\|b_t\boldsymbol{w}_t + c_t\boldsymbol{\xi}_t\|_2 \lesssim |b_t|\|\boldsymbol{w}_t\|_2 + |c_t|\|\boldsymbol{\xi}_t\|_2 \lesssim |b_t| + |c_t|$. For the L_2 norm of (63), each copy of $a_t\boldsymbol{v}$ contributes a_t and each copy of $b_t\boldsymbol{w}_t + c_t\boldsymbol{\xi}_t$ contributes a factor $|b_t| + |c_t|$. We conclude that

$$|c_{t+1}| \leqslant \|(a_t\boldsymbol{v} + b_t\boldsymbol{w}_t + c_t\boldsymbol{\xi}_t)^{\otimes(k-1)} - a_t^{k-1}\boldsymbol{v}^{\otimes(k-1)}\|_2 \lesssim \sum_{r=1}^{k-1} |a_t|^{k-1-r}(|b_t| + |c_t|)^r. \tag{64}$$

Combining the above estimate with (59) that $|a_{t+1}| \asymp |\beta||a_t|^{k-1}$, we divide both sides of (64) by $|\beta||a_t|^{k-1}$,

$$\frac{|c_{t+1}|}{|a_{t+1}|} \lesssim \frac{1}{|\beta|}\sum_{r=1}^{k-1}\left(\frac{|b_t|}{|a_t|} + \frac{|c_t|}{|a_t|}\right)^r \lesssim \frac{1}{|\beta|}\sum_{r=1}^{k-1}\left(\frac{1}{|\beta|} + \frac{|c_t|}{|a_t|}\right)^r, \tag{65}$$

where we used our induction hypothesis that $|a_t| \gtrsim |\beta||b_t|$. There are three cases:

1. If $|c_t|/|a_t| \geqslant 1$, then

$$\frac{|c_{t+1}|}{|a_{t+1}|} \lesssim \frac{1}{|\beta|}\sum_{r=1}^{k-1}\left(\frac{1}{|\beta|} + \frac{|c_t|}{|a_t|}\right)^r \lesssim \frac{1}{|\beta|}\left(\frac{|c_t|}{|a_t|}\right)^{k-1}. \tag{66}$$

   If $k = 2$, then our assumption $|\beta\langle\boldsymbol{u}, \boldsymbol{v}\rangle^{k-2}| = |\beta| \geqslant n^\varepsilon$, implies that $|c_{t+1}|/|a_{t+1}| \lesssim (|c_t|/|a_t|)/n^\varepsilon$. If $k \geqslant 2$, by our induction hypothesis $|c_t|/|a_t| \lesssim \beta^{1/(k-2)}/n^\varepsilon$. This implies $(|c_t|/|a_t|)^{k-2}/|\beta| \lesssim 1/n^\varepsilon$, and we still get that $|c_{t+1}|/|a_{t+1}| \lesssim (|c_t|/|a_t|)/n^\varepsilon$.

2. If $1/|\beta| \lesssim |c_t|/|a_t| \leqslant 1$, then

$$\frac{|c_{t+1}|}{|a_{t+1}|} \lesssim \frac{1}{|\beta|}\sum_{r=1}^{k-1}\left(\frac{1}{|\beta|}+\frac{|c_t|}{|a_t|}\right)^r \lesssim \frac{1}{|\beta|}\left(\frac{|c_t|}{|a_t|}\right) \lesssim \frac{1}{n^\varepsilon}\left(\frac{|c_t|}{|a_t|}\right), \tag{67}$$

where we used that $|\beta| \geqslant |\beta\langle \boldsymbol{u}, \boldsymbol{v}\rangle^{k-2}| \geqslant n^\varepsilon$.

3. Finally for $|c_t|/|a_t| \lesssim 1/|\beta|$, we will have

$$\frac{|c_{t+1}|}{|a_{t+1}|} \lesssim \frac{1}{|\beta|}\sum_{r=1}^{k-1}\left(\frac{1}{|\beta|}+\frac{|c_t|}{|a_t|}\right)^r \lesssim \frac{1}{|\beta|}\left(\frac{1}{|\beta|}\right) \lesssim \frac{1}{|\beta|^2}. \tag{68}$$

In all these cases if $|c_t|/|a_t| \lesssim \min\{\sqrt{n}, \mathbf{1}(k \geqslant 3)|\beta|^{1/(k-2)}\}/n^\varepsilon$, we have $|c_{t+1}|/|a_{t+1}| \lesssim \min\{\sqrt{n}, \mathbf{1}(k \geqslant 3)|\beta|^{1/(k-2)}\}/n^\varepsilon$. This finishes the proof of the induction (53).

For (54), since $\boldsymbol{\tau}_t$ is orthogonal to $\boldsymbol{v}^{\otimes(k-1)}, \boldsymbol{\tau}_0, \boldsymbol{\tau}_1, \cdots, \boldsymbol{\tau}_{t-1}$, Lemma 4.1 implies that conditioning on $\boldsymbol{\xi} = \boldsymbol{Z}[\boldsymbol{v}^{\otimes(k-1)}]$ and $\boldsymbol{\xi}_{s+1} = \boldsymbol{Z}[\boldsymbol{\tau}_s]$ for $0 \leqslant s \leqslant t-1$, $\boldsymbol{\xi}_{t+1} = \boldsymbol{Z}[\boldsymbol{\tau}_t]$ is an independent Gaussian vector, with each entry $\mathcal{N}(0, 1/n)$. By the standard concentration inequality, it holds that with probability $1 - e^{c(\log n)^2}$, $\|\boldsymbol{\xi}_{t+1}\|_2 = 1 + \mathrm{O}(\log n/\sqrt{n})$, $|\langle \boldsymbol{a}, \boldsymbol{\xi}_{t+1}\rangle|$ and the projection of $\boldsymbol{\xi}_{t+1}$ on the span of $\{\boldsymbol{v}, \boldsymbol{\xi}, \boldsymbol{\xi}_1, \cdots, \boldsymbol{\xi}_t\}$ is bounded by $\log n/\sqrt{n}$. This finishes the proof of the induction (54). □

Next, using (53) and (54) in Claim 4.3 as input, we prove that for

$$t \geqslant 1 + \frac{1}{\varepsilon}\left(\frac{1}{2} + \frac{2\log|\beta|}{\log n}\right), \tag{69}$$

with probability $1 - e^{c(\log n)^2}$ we have

$$\boldsymbol{y}_t = a_t\boldsymbol{v} + b_{t0}\boldsymbol{\xi} + b_{t1}\boldsymbol{\xi}_1 + \cdots + b_{tt-1}\boldsymbol{\xi}_{t-1} + c_t\boldsymbol{\xi}_t, \tag{70}$$

such that

$$b_{t0} = \frac{a_t}{\beta} + \mathrm{O}\left(\frac{\log n|a_t|}{|\beta|^2\sqrt{n}}\right) \quad |b_{t1}|, |b_{t2}|, \cdots, |b_{t(t-1)}| \lesssim \frac{(\log n)^{1/2}|a_t|}{|\beta|^{3/2}n^{1/4}}, \quad |c_t| \lesssim |a_t|/\beta^2. \tag{71}$$

Let $x_t = |c_t/a_t| \ll |\beta|^{1/(k-2)}$, then (65) implies

$$x_{t+1} \lesssim \frac{1}{|\beta|}\sum_{r=1}^{k-1}\left(\frac{1}{|\beta|} + x_t\right)^r, \tag{72}$$

from the discussion after (65), we have that either $x_{t+1} \lesssim 1/\beta^2$, or $x_{t+1} \lesssim x_t/n^\varepsilon$. Since $x_1 = |c_1/a_1| \lesssim n^{1/2-\varepsilon}$, we conclude that it holds

$$x_t = |c_t/a_t| \lesssim 1/\beta^2, \quad \text{when } t \geqslant \frac{1}{\varepsilon}\left(\frac{1}{2} + \frac{2\log|\beta|}{\log n}\right). \tag{73}$$

To derive the upper bound of $b_{t1}, b_{t2}, \cdots, b_{t(t-1)}$, we use (61).

$$\begin{aligned}
&b_{(t+1)0}^2 + b_{(t+1)1}^2 + b_{(t+1)2}^2 + \cdots + b_{(t+1)t}^2 \\
&\leqslant \|a_t\boldsymbol{v} + b_t\boldsymbol{w}_t + c_t\mathrm{Proj}_{\mathrm{Span}\{\boldsymbol{v},\boldsymbol{\xi},\boldsymbol{\xi}_1,\cdots,\boldsymbol{\xi}_{t-1}\}}(\boldsymbol{\xi}_t)\|_2^{2(k-1)} \\
&= \left(a_t^2 + \mathrm{O}\left(|a_t|\left(|b_t| + |c_t|\right)\frac{\log n}{\sqrt{n}} + \left(|b_t| + |c_t|\frac{\log n}{\sqrt{n}}\right)^2\right)\right)^{k-1},
\end{aligned} \tag{74}$$

where we used our induction (54) that $|\langle \boldsymbol{\xi}, \boldsymbol{v}\rangle|, |\langle \boldsymbol{\xi}_1, \boldsymbol{v}\rangle|, \cdots, |\langle \boldsymbol{\xi}_t, \boldsymbol{v}\rangle| \lesssim \log n/\sqrt{n}$ and the projection $\|\mathrm{Proj}_{\mathrm{Span}\{\boldsymbol{v},\boldsymbol{\xi},\boldsymbol{\xi}_1,\cdots,\boldsymbol{\xi}_{t-1}\}}(\boldsymbol{\xi}_t)\|_2 \lesssim \log n/\sqrt{n}$. Moreover, the first term $b_{(t+1)0}$ is the projection of $\boldsymbol{y}_t^{\otimes(k-1)}$ on $\boldsymbol{v}^{\otimes(k-1)}$,

$$b_{(t+1)0} = \langle a_t\boldsymbol{v} + b_t\boldsymbol{w}_t + c_t\boldsymbol{\xi}_t, \boldsymbol{v}\rangle^{k-1} = \left(a_t + \mathrm{O}\left(\frac{\log n(|b_t|+|c_t|)}{\sqrt{n}}\right)\right)^{k-1}, \tag{75}$$

where we used (54) that $|\langle \boldsymbol{\xi}, \boldsymbol{v}\rangle|, |\langle \boldsymbol{\xi}_1, \boldsymbol{v}\rangle|, \cdots, |\langle \boldsymbol{\xi}_t, \boldsymbol{v}\rangle| \lesssim \log n/\sqrt{n}$. Now we can take difference of (74) and (75), and use that $|b_t| \lesssim |a_t|/|\beta|$ from (53) and $|c_t| \lesssim |a_t|/|\beta|$ from (73),

$$b_{(t+1)0} = a_t^{k-1} + \mathrm{O}\left(|a_t|^{k-1}\frac{\log n}{|\beta|\sqrt{n}}\right), \quad b_{(t+1)1}^2 + b_{(t+1)2}^2 + \cdots + b_{(t+1)t}^2 \lesssim a_t^{2(k-1)}\frac{\log n}{|\beta|\sqrt{n}}. \tag{76}$$

From (56) and (59), we have that

$$a_{t+1} = \beta b_{(t+1)0} \asymp \beta a_t^{k-1}. \tag{77}$$

Using the above relation, we can simplify (76) as

$$b_{(t+1)0} = \frac{a_{t+1}}{\beta} + \mathrm{O}\left(\frac{\log n|a_{t+1}|}{|\beta|^2\sqrt{n}}\right), \quad |b_{(t+1)1}|, |b_{(t+1)2}|, \cdots |b_{(t+1)t}| \lesssim \frac{(\log n)^{1/2}|a_{t+1}|}{|\beta|^{3/2}n^{1/4}}. \tag{78}$$

This finishes the proof of (71).

With the expression (71), we can process to prove our main results (7) and (8). Thanks to (54), for t satisfies (69), we have that with probability at least $1 - \mathrm{O}(e^{-c(\log N)^2})$

$$\|\boldsymbol{y}_t\|_2^2 = a_t^2\left(1 + \frac{1}{\beta^2} + \frac{2\langle \boldsymbol{v}, \boldsymbol{\xi}\rangle}{\beta} + \mathrm{O}\left(\frac{\log n}{\beta^2\sqrt{n}} + \frac{(\log n)^{3/2}}{|\beta|^{3/2}n^{3/4}} + \frac{1}{\beta^4}\right)\right). \tag{79}$$

By rearranging it we get

$$1/\|\boldsymbol{y}_t\|_2 = \frac{1}{|a_t|}\left(1 - \frac{1}{2\beta^2} - \frac{\langle \boldsymbol{v}, \boldsymbol{\xi}\rangle}{\beta} + \mathrm{O}\left(\frac{\log n}{\beta^2\sqrt{n}} + \frac{(\log n)^{3/2}}{|\beta|^{3/2}n^{3/4}} + \frac{1}{\beta^4}\right)\right). \tag{80}$$

We can take the inner product $\langle \boldsymbol{a}, \boldsymbol{y}_t\rangle$ using (70) and (71), and multiply (80)

$$\begin{aligned}\langle \boldsymbol{a}, \boldsymbol{u}_t\rangle = \frac{\langle \boldsymbol{a}, \boldsymbol{y}_t\rangle}{\|\boldsymbol{y}_t\|_2} = \mathrm{sgn}(a_t)\left(\left(1 - \frac{1}{2\beta^2}\right)\langle \boldsymbol{a}, \boldsymbol{v}\rangle + \frac{\langle \boldsymbol{a}, \boldsymbol{\xi}\rangle - \langle \boldsymbol{a}, \boldsymbol{v}\rangle\langle \boldsymbol{v}, \boldsymbol{\xi}\rangle}{\beta}\right) \\ + \mathrm{O}_{\mathbb{P}}\left(\frac{\log n}{\beta^2\sqrt{n}} + \frac{(\log n)^{3/2}}{|\beta|^{3/2}n^{3/4}} + \frac{|\langle \boldsymbol{a}, \boldsymbol{v}\rangle|}{\beta^4}\right),\end{aligned} \tag{81}$$

where we used (54) that with high probability $|\langle \boldsymbol{a}, \boldsymbol{\xi}\rangle|$, $|\langle \boldsymbol{a}, \boldsymbol{\xi}_s\rangle|$ for $1 \leqslant s \leqslant t$ are bounded by $\log n/\sqrt{n}$. This finishes the proof of (7). For $\widehat{\beta}$ in (8), we have

$$\boldsymbol{X}[\boldsymbol{u}_t^{\otimes k}] = \frac{\boldsymbol{X}[\boldsymbol{y}_t^{\otimes k}]}{\|\boldsymbol{y}_t\|_2^k} = \frac{\langle \boldsymbol{y}_t, \boldsymbol{X}[\boldsymbol{y}_t^{\otimes(k-1)}]\rangle}{\|\boldsymbol{y}_t\|_2^k} = \frac{\langle \boldsymbol{y}_t, \boldsymbol{y}_{t+1}\rangle}{\|\boldsymbol{y}_t\|_2^k}. \tag{82}$$

Thanks to (73), (52) and (54), for t satisfies (69), with probability at least $1 - \mathrm{O}(e^{-c(\log N)^2})$, we have

$$\boldsymbol{y}_{t+1} = a_{t+1}\boldsymbol{v} + b_{(t+1)0}\boldsymbol{\xi} + b_{(t+1)1}\boldsymbol{\xi}_1 + \cdots + b_{(t+1)t}\boldsymbol{\xi}_t + c_{t+1}\boldsymbol{\xi}_{t+1}, \tag{83}$$

where $|c_{t+1}| \lesssim |a_t|^{k-1}/\beta^2$,

$$\begin{aligned} a_{t+1} &= \beta(a_t + b_{t0}\langle \boldsymbol{\xi}, \boldsymbol{v}\rangle + b_{t1}\langle \boldsymbol{\xi}_1, \boldsymbol{v}\rangle + \cdots + b_{t(t-1)}\langle \boldsymbol{\xi}_{t-1}, \boldsymbol{v}\rangle + c_t\langle \boldsymbol{\xi}_t, \boldsymbol{v}\rangle)^{k-1} \\ &= \beta a_t^{k-1}\left(1 + \frac{\langle \boldsymbol{\xi}, \boldsymbol{v}\rangle}{\beta} + \mathrm{O}\left(\frac{\log n}{\beta^2\sqrt{n}} + \frac{(\log n)^{3/2}}{|\beta|^{3/2}n^{3/4}}\right)\right)^{k-1}, \end{aligned} \tag{84}$$

and

$$b_{(t+1)0} = a_t^{k-1}\left(1 + \mathrm{O}\left(\frac{\log n}{|\beta|\sqrt{n}}\right)\right), \quad |b_{(t+1)1}|, |b_{(t+1)2}| + \cdots + |b_{(t+1)t}| \lesssim a_t^{k-1}\frac{(\log n)^{1/2}}{|\beta|^{1/2}n^{1/4}}. \tag{85}$$

From the discussion above, combining with (70) and (71) with straightforward computation, we have

$$\langle \boldsymbol{y}_t, \boldsymbol{y}_{t+1}\rangle = \beta a_t^k\left(1 + \frac{1}{\beta^2} + \frac{(k+1)\langle \boldsymbol{\xi}, \boldsymbol{v}\rangle}{\beta} + \mathrm{O}\left(\frac{\log n}{\beta^2\sqrt{n}} + \frac{(\log n)^{3/2}}{|\beta|^{3/2}n^{3/4}}\right)\right). \tag{86}$$

By plugging (80) and (86) into (82), we get

$$\boldsymbol{X}[\boldsymbol{u}_t^{\otimes k}] = \mathrm{sgn}(a_t)^k\left(\beta + \langle \boldsymbol{\xi}, \boldsymbol{v}\rangle - \frac{k/2-1}{\beta}\right) + \mathrm{O}\left(\frac{\log n}{|\beta|\sqrt{n}} + \frac{(\log n)^{3/2}}{|\beta|^{1/2}n^{3/4}} + \frac{1}{|\beta|^3}\right) \tag{87}$$

Since by our assumption, in Case 1 we have that $\beta > 0$. Thanks to (84) $a_{t+1} = \beta a_t^{k-1}(1 + \mathrm{o}(1))$, especially a_{t+1} and a_t are of the same sign. In the case $\langle \boldsymbol{u}, \boldsymbol{v}\rangle > 0$, we have $a_1 = \beta\langle \boldsymbol{u}, \boldsymbol{v}\rangle^{k-1} > 0$. We conclude that $a_t > 0$. Therefore $\mathrm{sgn}(\boldsymbol{X}[\boldsymbol{u}_t^{\otimes k}]) = \mathrm{sgn}(a_t)^k = +$, and it follows that

$$\boldsymbol{X}[\boldsymbol{u}_t^{\otimes k}] = \beta + \langle \boldsymbol{\xi}, \boldsymbol{v}\rangle - \frac{k/2-1}{\beta} + \mathrm{O}\left(\frac{\log n}{|\beta|\sqrt{n}} + \frac{(\log n)^{3/2}}{|\beta|^{1/2}n^{3/4}} + \frac{1}{|\beta|^3}\right) \tag{88}$$

This finishes the proof of (8). The Cases 2, 3, 4 follow by simply changing $(\beta, \boldsymbol{v})$ in the righthand side of (7) and (8) to the corresponding limit. □

Proof of Theorem 2.2. We use the same notations as in the proof of Theorem 2.1. If $|\beta| \geqslant n^{\varepsilon}$ and $|\beta\langle \boldsymbol{u}, \boldsymbol{v}\rangle^{k-2}| \leqslant n^{-\varepsilon}$, then we first prove by induction that for any fixed time t, with probability at least $1 - \mathrm{O}(e^{-c(\log N)^2})$ the following holds: for any $s \leqslant t$,

$$\begin{aligned} &|b_{s0}|, |b_{s1}|, \cdots, |b_{s(s-1)}| \lesssim \max\{|c_s|/|\beta|^{(k-1)/(k-2)}, (\log n)^{k-1}|c_s|/n^{(k-1)/2}\}, \\ &|c_s| \geqslant n^{\varepsilon}\beta^{1/(k-2)}|a_s|, \end{aligned} \tag{89}$$

and

$$\begin{aligned} &\|\boldsymbol{\xi}\|, \|\boldsymbol{\xi}_s\|_2 = 1 + \mathrm{O}(\log n/\sqrt{n}), \\ &|\langle \boldsymbol{v}, \boldsymbol{\xi}\rangle|, \|\mathrm{Proj}_{\mathrm{Span}\{\boldsymbol{v}, \boldsymbol{\xi}, \boldsymbol{\xi}_1, \cdots, \cdots, \boldsymbol{\xi}_{s-1}\}}(\boldsymbol{\xi}_s)\|_2 \lesssim \log n/\sqrt{n}. \end{aligned} \tag{90}$$

From (44), $a_1 = \beta\langle \boldsymbol{u}, \boldsymbol{v}\rangle^{k-1}$, $b_{10} = \langle \boldsymbol{u}, \boldsymbol{v}\rangle^{k-1}$ and $c_1 = \sqrt{1 - \langle \boldsymbol{u}, \boldsymbol{v}\rangle^{2(k-1)}}$. Since $|\beta| \geqslant n^{\varepsilon}$ and $|\beta\langle \boldsymbol{u}, \boldsymbol{v}\rangle^{k-2}| \leqslant n^{-\varepsilon}$, we have that $|\langle \boldsymbol{u}, \boldsymbol{v}\rangle| \leqslant n^{-2\varepsilon/(k-2)} \ll 1$ and therefore $|c_1| \asymp 1$. We can check that $|\beta^{1/(k-2)}a_1| = |\beta\langle \boldsymbol{u}, \boldsymbol{v}\rangle^{k-2}|^{(k-1)/(k-2)} \leqslant n^{-\varepsilon} \lesssim n^{-\varepsilon}|c_1|$ and $|b_{10}| = |a_1/\beta| \lesssim n^{-\varepsilon}|c_1/\beta^{(k-1)/(k-2)}|$. Moreover, conditioning on $\boldsymbol{Z}[\boldsymbol{v}^{\otimes(k-1)}] = \boldsymbol{\xi}$, Lemma 4.1 implies that $\boldsymbol{\xi}_1 = \boldsymbol{Z}[\boldsymbol{\tau}_0]$ is an independent Gaussian random vector with each entry $\mathcal{N}(0, 1/n)$. By the standard concentration inequality, it holds that with probability $1 - e^{c(\log n)^2}$, $\|\boldsymbol{\xi}_1\|_2 = 1 + \mathrm{O}(\log n/\sqrt{n})$, and the

projection of $\boldsymbol{\xi}_1$ on the span of $\{\boldsymbol{v}, \boldsymbol{\xi}\}$ is bounded by $\log n/\sqrt{n}$. So far we have proved (89) and (90) for $t = 1$.

In the following, assuming the statements (89) and (90) hold for t, we prove them for $t + 1$. From (56), using (56) and (57), we have

$$\begin{aligned}
|a_{t+1}| &= \left|\beta(a_t + b_t\langle \boldsymbol{w}_t, \boldsymbol{v}\rangle + c_t\langle \boldsymbol{\xi}_t, \boldsymbol{v}\rangle)^{k-1}\right| \\
&\lesssim |\beta| \left(|a_t| + \frac{\log n(|b_{t0}| + |b_{t1}| + \cdots + |b_{t(t-1)}|)}{\sqrt{n}} + \frac{\log n|c_t|}{\sqrt{n}}\right)^{k-1} \\
&\lesssim |\beta| \left(|a_t| + \frac{\log n|c_t|}{\sqrt{n}}\right)^{k-1} \lesssim |\beta| \left(\frac{|c_t|}{n^{\varepsilon}|\beta|^{1/(k-2)}} + \frac{\log n|c_t|}{\sqrt{n}}\right)^{k-1} \\
&\lesssim |\beta||c_t|^{k-1} \left(\frac{1}{n^{\varepsilon}|\beta|^{1/(k-2)}} + \frac{\log n}{\sqrt{n}}\right)^{k-1} \lesssim \frac{|c_t|^{k-1}}{n^{\varepsilon}|\beta|^{1/(k-2)}},
\end{aligned} \tag{91}$$

where in the third line we used our induction hypothesis that $|b_{t0}| + |b_{t1}| + \cdots + |b_{t(t-1)}| \lesssim |c_t|$, and $n^{-\varepsilon} \geqslant |\beta||\langle \boldsymbol{u}, \boldsymbol{v}\rangle|^{k-2} \gtrsim |\beta|/n^{(k-2)/2}$.

For $b_{(t+1)0}, b_{(t+1)1}, \cdots, b_{(t+1)t}$, from (61) we have

$$\begin{aligned}
&\sqrt{b_{(t+1)0}^2 + b_{(t+1)1}^2 + b_{(t+1)2}^2 + \cdots + b_{(t+1)t}^2} \lesssim \left(|a_t| + |b_t| + \frac{\log n|c_t|}{\sqrt{n}}\right)^{k-1} \\
&\lesssim \left(|a_t| + \frac{\log n|c_t|}{\sqrt{n}}\right)^{k-1} \lesssim \left(\frac{|c_t|}{n^{\varepsilon}|\beta|^{1/(k-2)}} + \frac{\log n|c_t|}{\sqrt{n}}\right)^{k-1} \\
&\lesssim |c_t|^{k-1} \left(\frac{1}{n^{\varepsilon}|\beta|^{1/(k-2)}} + \frac{\log n}{\sqrt{n}}\right)^{k-1}.
\end{aligned} \tag{92}$$

Finally we estimate c_{t+1}. We recall from (48), the coefficient c_{t+1} is the remainder of $\boldsymbol{y}_t^{\otimes(k-1)}$ after projecting on $\boldsymbol{v}^{\otimes(k-1)}, \boldsymbol{\tau}_0, \boldsymbol{\tau}_1, \cdots, \boldsymbol{\tau}_{t-1}$. We have the following lower bound for c_{t+1}

$$\begin{aligned}
|c_{t+1}|^2 &= \|(a_t\boldsymbol{v} + b_t\boldsymbol{w}_t + c_t\boldsymbol{\xi}_t)^{\otimes(k-1)}\|_2^2 - (b_{(t+1)0}^2 + b_{(t+1)1}^2 + b_{(t+1)2}^2 + \cdots + b_{(t+1)t}^2) \\
&\geqslant \|a_t\boldsymbol{v} + b_t\boldsymbol{w}_t + c_t\boldsymbol{\xi}_t\|_2^{2(k-1)} - \mathrm{O}\left(|c_t|^{2(k-1)} \left(\frac{1}{n^{\varepsilon}|\beta|^{1/(k-2)}} + \frac{\log n}{\sqrt{n}}\right)^{2(k-1)}\right).
\end{aligned} \tag{93}$$

For the first term on the righthand side of (93), using our induction hypothesis (89) and (90) that $|a_t| \lesssim |c_t|$, we have

$$\begin{aligned}
\|a_t\boldsymbol{v} + b_t\boldsymbol{w}_t + c_t\boldsymbol{\xi}_t\|_2^2 &= a_t^2 + b_t^2 + c_t^2\|\boldsymbol{\xi}_t\|_2^2 + 2a_tb_t\langle \boldsymbol{v}, \boldsymbol{w}_t\rangle + 2a_tc_t\langle \boldsymbol{v}, \boldsymbol{\xi}_t\rangle + 2b_tc_t\langle \boldsymbol{w}_t, \boldsymbol{\xi}_t\rangle \\
&= \left(1 + \mathrm{O}\left(\frac{\log n}{\sqrt{n}} + \frac{1}{n^{2\varepsilon}\beta^{2/(k-2)}}\right)\right) c_t^2.
\end{aligned} \tag{94}$$

We get the following lower for c_{t+1} by plugging (94) into (93), and rearranging

$$|c_{t+1}| \geqslant \left(1 + \mathrm{O}\left(\frac{\log n}{\sqrt{n}} + \frac{1}{n^{2\varepsilon}\beta^{2/(k-2)}}\right)\right) |c_t|^{k-1} \tag{95}$$

The claim that $|b_{(t+1)0}|, |b_{(t+1)1}|, \cdots, |b_{(t+1)t}| \lesssim \max\{|c_{t+1}|/|\beta|^{(k-1)/(k-2)}, (\log n)^{k-1}|c_{t+1}|/n^{(k-1)/2}\}$ follows from combining (92) and (95). The claim that $|c_{t+1}| \geqslant n^{\varepsilon}\beta^{1/(k-2)}|a_{t+1}|$ follows from combining (91) and (95).

For (90), since $\boldsymbol{\tau}_t$ is orthogonal to $\boldsymbol{v}^{\otimes(k-1)}, \boldsymbol{\tau}_0, \boldsymbol{\tau}_1, \cdots, \boldsymbol{\tau}_{t-1}$, Lemma 4.1 implies that conditioning on $\boldsymbol{\xi} = \boldsymbol{Z}[\boldsymbol{v}^{\otimes(k-1)}]$ and $\boldsymbol{\xi}_{s+1} = \boldsymbol{Z}[\boldsymbol{\tau}_s]$ for $0 \leqslant s \leqslant t-1$, $\boldsymbol{\xi}_{t+1} = \boldsymbol{Z}[\boldsymbol{\tau}_t]$ is an independent Gaussian vector, with each entry $\mathcal{N}(0, 1/n)$. By the standard concentration inequality, it holds that with probability $1 - e^{c(\log n)^2}$, $\|\boldsymbol{\xi}_{t+1}\|_2 = 1 + \mathrm{O}(\log n/\sqrt{n})$, and the projection of $\boldsymbol{\xi}_{t+1}$ on the span of $\{\boldsymbol{v}, \boldsymbol{\xi}, \boldsymbol{\xi}_1, \cdots, \boldsymbol{\xi}_t\}$ is bounded by $\log n/\sqrt{n}$. This finishes the proof of the induction (90).

Next, using (53) and (54) as input, we prove that for

$$t \geqslant 1 + \frac{1}{\varepsilon}\left(\frac{1}{2} - \frac{\log|\beta|}{(k-2)\log n}\right), \tag{96}$$

we have

$$\boldsymbol{y}_t = a_t \boldsymbol{v} + b_{t0}\boldsymbol{\xi} + b_{t1}\boldsymbol{\xi}_1 + \cdots + b_{t(t-1)}\boldsymbol{\xi}_{t-1} + c_t \boldsymbol{\xi}_t, \tag{97}$$

such that

$$|a_t|, |b_{t0}|, |b_{t1}|, \cdots, |b_{t(t-1)}| \lesssim |c_t||\beta|\left(\frac{\log n}{\sqrt{n}}\right)^{k-1}. \tag{98}$$

Let $x_t = |a_t/c_t|$, then (89) implies that $x_t \leqslant 1/(n^{\varepsilon}|\beta|^{1/(k-2)})$. By taking the ratio of (91) and (95), we get

$$x_{t+1} \lesssim |\beta|\left(\frac{\log n}{\sqrt{n}} + x_t\right)^{k-1}. \tag{99}$$

there are two cases,

1. if $\log n/\sqrt{n} \lesssim x_t \leqslant 1/(n^{\varepsilon}|\beta|^{1/(k-2)})$, then
$$x_{t+1} \lesssim |\beta| x_t^{k-1} = x_t(|\beta|^{1/(k-2)}x_t)^{k-2} \leqslant x_t/n^{\varepsilon}; \tag{100}$$
2. If $x_t \lesssim \log n/\sqrt{n}$, then $|x_{t+1}| \lesssim |\beta|(\log n/\sqrt{n})^{k-1}$.

Since $x_1 = |a_1/c_1| \lesssim 1/(n^{\varepsilon}|\beta|^{1/(k-2)})$, we conclude that

$$x_t = |a_t/c_t| \lesssim |\beta|(\log n/\sqrt{n})^{k-1}, \quad \text{when } t \geqslant \frac{1}{\varepsilon}\left(\frac{1}{2} - \frac{\log|\beta|}{(k-2)\log n}\right). \tag{101}$$

In this regime, (92) implies that

$$\begin{aligned} &|b_{(t+1)0}|, |b_{(t+1)1}|, |b_{(t+1)2}|, \cdots, |b_{(t+1)t}| \lesssim |\beta||c_t|^{k-1}\left(\frac{|a_t|}{|c_t|} + \frac{\log n}{\sqrt{n}}\right)^{k-1} \\ \lesssim\ &|\beta||c_t|^{k-1}\left(\frac{\log n}{\sqrt{n}}\right)^{k-1} \lesssim |c_{t+1}||\beta|\left(\frac{\log n}{\sqrt{n}}\right)^{k-1}, \end{aligned} \tag{102}$$

where we used (95) in the last inequality. This finishes the proof of (98). Using (98), we can compute $\boldsymbol{u}_t$,

$$\boldsymbol{u}_t = \frac{\boldsymbol{y}_t}{\|\boldsymbol{y}_t\|} = \frac{\boldsymbol{\xi}_t}{\|\boldsymbol{\xi}_t\|_2} + \mathrm{O}_{\mathbb{P}}\left(|\beta|\left(\frac{\log n}{\sqrt{n}}\right)^{k-1}\right), \tag{103}$$

where the error term is a vector of length bounded by $|\beta|(\log n/\sqrt{n})^{k-1}$. This finishes the proof of Theorem 2.1.

□

4.2 Proof of Corollarys 2.3 and 2.4

Proof of Corollary 2.3. According to the definition of $\boldsymbol{\xi}$ in (7) of Theorem 2.1, i.e. $\boldsymbol{\xi} = \boldsymbol{Z}[\boldsymbol{v}^{\otimes(k-1)}]$, is an n-dim vector, with each entry i.i.d. $\mathcal{N}(0, 1/n)$ Gaussian random variable. We see that

$$\langle \boldsymbol{\xi}, \boldsymbol{v} \rangle \overset{d}{=} \mathcal{N}(0, 1/n).$$

Especially with high probability we will have that $|\langle \boldsymbol{\xi}, \boldsymbol{v} \rangle| \lesssim \log n/\sqrt{n}$. Then we conclude from (8), with high probability it holds

$$\widehat{\beta} = \beta + \mathrm{O}\left(\frac{1}{\beta} + \frac{\log n}{\sqrt{n}}\right). \tag{104}$$

With the bound (180), we can replace $\langle \boldsymbol{a}, \boldsymbol{v} \rangle/(2\beta^2)$ on the righthand side of (7) by $\langle \boldsymbol{a}, \boldsymbol{v} \rangle/(2\widehat{\beta}^2)$, which gives an error

$$\left| \frac{\langle \boldsymbol{a}, \boldsymbol{v} \rangle}{2\beta^2} - \frac{\langle \boldsymbol{a}, \boldsymbol{v} \rangle}{2\widehat{\beta}^2} \right| = \mathrm{O}\left(|\langle \boldsymbol{a}, \boldsymbol{v} \rangle| \left(\frac{1}{|\beta|^4} + \frac{\log n}{|\beta|^3 \sqrt{n}} \right) \right). \tag{105}$$

Combining the above discussion together, we can rewrite (7) as

$$\langle \boldsymbol{a}, \widehat{\boldsymbol{v}} \rangle - \left(1 - \frac{1}{2\widehat{\beta}^2}\right) \langle \boldsymbol{a}, \boldsymbol{v} \rangle = \frac{\langle \boldsymbol{a}, \boldsymbol{\xi} \rangle - \langle \boldsymbol{a}, \boldsymbol{v} \rangle \langle \boldsymbol{v}, \boldsymbol{\xi} \rangle}{\beta} + \mathrm{O}\left(\frac{\log n}{\beta^2 \sqrt{n}} + \frac{(\log n)^{3/2}}{\beta^{3/2} n^{3/4}} + \frac{|\langle \boldsymbol{a}, \boldsymbol{v} \rangle|}{\beta^4} \right) \tag{106}$$

with high probability.

Again thanks to the definition of $\boldsymbol{\xi}$ in (7) of Theorem 2.1, i.e. $\boldsymbol{\xi} = \boldsymbol{Z}[\boldsymbol{v}^{\otimes(k-1)}]$, is an n-dim vector, with each entry i.i.d. $\mathcal{N}(0, 1/n)$ Gaussian random variable, we see that

$$\langle \boldsymbol{a}, \boldsymbol{\xi} \rangle - \langle \boldsymbol{a}, \boldsymbol{v} \rangle \langle \boldsymbol{v}, \boldsymbol{\xi} \rangle = \langle \boldsymbol{a} - \langle \boldsymbol{a}, \boldsymbol{v} \rangle \boldsymbol{v}, \boldsymbol{\xi} \rangle, \tag{107}$$

is a Gaussian random variable, with mean zero and variance

$$\mathbb{E}[\langle \boldsymbol{a} - \langle \boldsymbol{a}, \boldsymbol{v} \rangle \boldsymbol{v}, \boldsymbol{\xi} \rangle^2] = \frac{1}{n} \| \boldsymbol{a} - \langle \boldsymbol{a}, \boldsymbol{v} \rangle \boldsymbol{v} \|_2^2 = \frac{1}{n} \langle \boldsymbol{a}, (\boldsymbol{I}_n - \boldsymbol{v}\boldsymbol{v}^\top) \boldsymbol{a} \rangle = \frac{1 + \mathrm{o}(1)}{n} \langle \boldsymbol{a}, (\boldsymbol{I}_n - \widehat{\boldsymbol{v}}\widehat{\boldsymbol{v}}^\top) \boldsymbol{a} \rangle. \tag{108}$$

This together with (180), (182) as well as our assumption (11)

$$\frac{\sqrt{n}\widehat{\beta}}{\sqrt{\langle \boldsymbol{a}, (\boldsymbol{I}_n - \widehat{\boldsymbol{v}}\widehat{\boldsymbol{v}}^\top) \boldsymbol{a} \rangle}} \left[\big(1 - \frac{1}{2\widehat{\beta}^2}\big)^{-1} \langle \boldsymbol{a}, \widehat{\boldsymbol{v}} \rangle - \langle \boldsymbol{a}, \boldsymbol{v} \rangle \right] \xrightarrow{d} \mathcal{N}(0, 1). \tag{109}$$

Under the same assumption, we have similar results for Cases 2, 3, 4, by simply changing $(\beta, \boldsymbol{v})$ in the righthand side of (7) and (8) to the corresponding expression. □

Proof of Corollary 2.4. Given the significance level α, the asymptotic confidence intervals in Corollary 2.4 can be calculated from Corollary 2.3 by bounding the absolute values of the left hand sides of (13) at z_α. □

4.3 Proof of Theorem 2.7

Proof of Theorem 2.7. We define an auxiliary iteration, $\boldsymbol{y}_0 = \boldsymbol{u}$ and

$$\boldsymbol{y}_{t+1} = \boldsymbol{X}[\boldsymbol{y}_t^{\otimes(k-1)}]. \tag{110}$$

Then we have that $\boldsymbol{u}_t = \boldsymbol{y}_t/\|\boldsymbol{y}_t\|_2$.

For index $\boldsymbol{j} = (j_1, j_2, \cdots, j_{k-1}) \in [\![1, r]\!]^{k-1}$. Let $\boldsymbol{\xi_j} = \boldsymbol{Z}[\boldsymbol{v}_{j_1} \otimes \boldsymbol{v}_{j_2} \otimes \cdots \otimes \boldsymbol{v}_{j_{k-1}}]$. Its entries

$$\boldsymbol{\xi_j}(i) = \sum_{i_1, i_2, \cdots, i_{k-1} \in [\![1,n]\!]} \boldsymbol{Z}_{i i_1 i_2 \cdots i_{k-1}} \boldsymbol{v}_{j_1}(i_1) \boldsymbol{v}_{j_2}(i_2) \cdots \boldsymbol{v}_{j_{k-1}}(i_{k-1}), \tag{111}$$

are linear combination of Gaussian random variables, which is also Gaussian. These entries are i.i.d. Gaussian variables with mean zero and variance $1/n$,

$$\mathbb{E}[\boldsymbol{\xi_j}(i)^2] = \sum_{i_1, i_2, \cdots, i_{k-1} \in [\![1,n]\!]} \mathbb{E}[\boldsymbol{Z}^2_{i i_1 i_2 \cdots i_{k-1}}] \boldsymbol{v}_{j_1}(i_1)^2 \boldsymbol{v}_{j_2}(i_2)^2 \cdots \boldsymbol{v}_{j_{k-1}}(i_{k-1})^2 = \frac{1}{n}. \tag{112}$$

We can compute $\boldsymbol{y}_t$ iteratively:

$$\boldsymbol{y}_1 = \boldsymbol{X}[\boldsymbol{y}_0^{\otimes(k-1)}] = \sum_{j=1}^{r} \beta_j \langle \boldsymbol{y}_0, \boldsymbol{v}_j \rangle^{k-1} \boldsymbol{v}_j + \boldsymbol{Z}[\boldsymbol{y}_0^{\otimes(k-1)}]. \tag{113}$$

For the last term on the righthand side of (113), we can decompose $\boldsymbol{y}_0^{\otimes(k-1)}$ as a projection on $\boldsymbol{v}_{j_1} \otimes \boldsymbol{v}_{j_2} \cdots \otimes \boldsymbol{v}_{j_{k-1}}$ for $\boldsymbol{j} \in [\![1, r]\!]^{k-1}$, and its orthogonal part:

$$\boldsymbol{y}_0^{\otimes(k-1)} = \sum_{\boldsymbol{j}} \prod_{s=1}^{k-1} \langle \boldsymbol{y}_0, \boldsymbol{v}_{j_s} \rangle \boldsymbol{v}_{j_1} \otimes \boldsymbol{v}_{j_2} \otimes \cdots \otimes \boldsymbol{v}_{j_{k-1}} + \sqrt{1 - \left(\sum_{j=1}^{r} \langle \boldsymbol{y}_0, \boldsymbol{v}_j \rangle^2 \right)^{(k-1)}} \boldsymbol{\tau}_0, \tag{114}$$

where the sum is over $\boldsymbol{j} \in [\![1, r]\!]^{k-1}$, $\boldsymbol{\tau}_0 \in \otimes^k \mathbb{R}^n$ and $\|\boldsymbol{\tau}_0\|_2 = 1$. Let $\boldsymbol{\xi}_1 = \boldsymbol{Z}[\boldsymbol{\tau}_0]$. By our construction $\boldsymbol{v}_{j_1} \otimes \boldsymbol{v}_{j_2} \otimes \cdots \otimes \boldsymbol{v}_{j_{k-1}}$ for any $\boldsymbol{j} \in [\![1, r]\!]^{k-1}$ and $\boldsymbol{\tau}_0$ are othorgonal to each other. Thanks to Lemma 4.1, conditioning on $\boldsymbol{\xi_j} := \boldsymbol{Z}[\boldsymbol{v}_{j_1} \otimes \boldsymbol{v}_{j_2} \otimes \cdots \otimes \boldsymbol{v}_{j_{k-1}}]$ for index $\boldsymbol{j} = (j_1, j_2, \cdots, j_{k-1}) \in [\![1, r]\!]^{k-1}$, $\boldsymbol{\xi}_1 = \boldsymbol{Z}[\boldsymbol{\tau}_0]$ has the same law as $\tilde{\boldsymbol{Z}}[\boldsymbol{\tau}_0]$, where $\tilde{\boldsymbol{Z}}$ is an independent copy of $\boldsymbol{Z}$. Since $\langle \boldsymbol{\tau}_0, \boldsymbol{\tau}_0 \rangle = 1$, $\boldsymbol{\xi}_1$ is a Gaussian vector with each entry $\mathcal{N}(0, 1/n)$. With those notations we can rewrite $\boldsymbol{y}_1$ as

$$\boldsymbol{y}_1 = \sum_{j=1}^{r} \beta_j \langle \boldsymbol{y}_0, \boldsymbol{v}_j \rangle^{k-1} \boldsymbol{v}_j + \sum_{\boldsymbol{j}} \prod_{s=1}^{k-1} \langle \boldsymbol{y}_0, \boldsymbol{v}_{j_s} \rangle \boldsymbol{\xi_j} + \sqrt{1 - \left(\sum_{j=1}^{r} \langle \boldsymbol{y}_0, \boldsymbol{v}_j \rangle^2 \right)^{(k-1)}} \boldsymbol{\xi}_1. \tag{115}$$

In the following we show that:

Claim 4.4. *We can compute $\boldsymbol{y}_2, \boldsymbol{y}_3, \cdots, \boldsymbol{y}_t$ inductively. The Gram-Schmidt orthonormalization procedure gives an orthogonal base of $\boldsymbol{v}_{j_1} \otimes \boldsymbol{v}_{j_2} \otimes \cdots \otimes \boldsymbol{v}_{j_{k-1}}$ for $\boldsymbol{j} \in [\![1, r]\!]^{k-1}$ and $\boldsymbol{y}_0^{\otimes(k-1)}, \boldsymbol{y}_1^{\otimes(k-1)}, \cdots, \boldsymbol{y}_{t-1}^{\otimes(k-1)}$ as:*

$$\{\boldsymbol{v}_{j_1} \otimes \boldsymbol{v}_{j_2} \otimes \cdots \otimes \boldsymbol{v}_{j_{k-1}}\}_{\boldsymbol{j} \in [\![1,r]\!]^{k-1}}, \boldsymbol{\tau}_0, \boldsymbol{\tau}_1, \cdots, \boldsymbol{\tau}_{t-1}. \tag{116}$$

Let $\boldsymbol{\xi_j} = \boldsymbol{Z}[\boldsymbol{v}_{j_1} \otimes \boldsymbol{v}_{j_2} \otimes \cdots \otimes \boldsymbol{v}_{j_{k-1}}]$ for $\boldsymbol{j} = (j_1, j_2, \cdots, j_{k-1}) \in [\![1, r]\!]^{k-1}$, and $\boldsymbol{\xi}_{s+1} = \boldsymbol{Z}[\boldsymbol{\tau}_s]$ for $0 \leqslant s \leqslant t-1$. Conditioning on $\boldsymbol{\xi_j} = \boldsymbol{Z}[\boldsymbol{v}_{j_1} \otimes \boldsymbol{v}_{j_2} \otimes \cdots \otimes \boldsymbol{v}_{j_{k-1}}]$ for $\boldsymbol{j} = (j_1, j_2, \cdots, j_{k-1}) \in [\![1, r]\!]^{k-1}$

and $\boldsymbol{\xi}_{s+1} = \boldsymbol{Z}[\boldsymbol{\tau}_s]$ *for* $0 \leqslant s \leqslant t-2$, $\boldsymbol{\xi}_t = \boldsymbol{Z}[\boldsymbol{\tau}_{t-1}]$ *is an independent Gaussian vector, with each entry* $\mathcal{N}(0, 1/n)$. *Then* $\boldsymbol{y}_t$ *is in the following form*

$$\boldsymbol{y}_t = a_t \boldsymbol{v}_t + b_t \boldsymbol{w}_t + c_t \boldsymbol{\xi}_t, \tag{117}$$

where

$$a_t \boldsymbol{v}_t = a_{t1}\boldsymbol{v}_1 + a_{t2}\boldsymbol{v}_2 + \cdots + a_{tr}\boldsymbol{v}_r, \quad b_t \boldsymbol{w}_t = \sum_{\boldsymbol{j}} b_{t\boldsymbol{j}} \boldsymbol{\xi}_{\boldsymbol{j}} + b_{t1}\boldsymbol{\xi}_1 + \cdots + b_{tt-1}\boldsymbol{\xi}_{t-1}, \tag{118}$$

and $\|\boldsymbol{v}_1\|_2, \|\boldsymbol{v}_2\|_2, \cdots, \|\boldsymbol{v}_r\|_2, \|\boldsymbol{w}_t\|_2 = 1$.

Proof of Claim 4.4. The Claim 4.4 for $t = 1$ follows from (115). In the following, assuming Claim 4.4 holds for t, we prove it for $t+1$.

Conditioning on $\boldsymbol{\xi}_{\boldsymbol{j}} = \boldsymbol{Z}[\boldsymbol{v}_{j_1} \otimes \boldsymbol{v}_{j_2} \otimes \cdots \otimes \boldsymbol{v}_{j_{k-1}}]$ for index $\boldsymbol{j} = (j_1, j_2, \cdots, j_{k-1}) \in [\![1, r]\!]^{k-1}$ and $\boldsymbol{Z}[\boldsymbol{\tau}_s] = \boldsymbol{\xi}_{s+1}$ for $0 \leqslant s \leqslant t-2$, Lemma 4.1 implies that $\boldsymbol{\xi}_t = \boldsymbol{Z}[\boldsymbol{\tau}_{t-1}]$ has the same law as $\tilde{\boldsymbol{Z}}[\boldsymbol{\tau}_{t-1}]$, where $\tilde{\boldsymbol{Z}}$ is an independent copy of $\boldsymbol{Z}$. Since $\boldsymbol{\tau}_{t-1}$ is orthogonal to $\boldsymbol{v}_{j_1} \otimes \boldsymbol{v}_{j_2} \otimes \cdots \otimes \boldsymbol{v}_{j_{k-1}}$ for index $\boldsymbol{j} = (j_1, j_2, \cdots, j_{k-1}) \in [\![1, r]\!]^{k-1}$ and $\boldsymbol{Z}[\boldsymbol{\tau}_s] = \boldsymbol{\xi}_{s+1}$ for $0 \leqslant s \leqslant t-2$, $\boldsymbol{\xi}_t$ is an independent Gaussian random vector with each entry $\mathcal{N}(0, 1/n)$.

Let $\{\boldsymbol{v}_{j_1} \otimes \boldsymbol{v}_{j_2} \otimes \cdots \otimes \boldsymbol{v}_{j_{k-1}}\}_{\boldsymbol{j} \in [\![1,r]\!]^{k-1}}, \boldsymbol{\tau}_0, \boldsymbol{\tau}_1, \cdots, \boldsymbol{\tau}_t$ be an orthogonal base for $\boldsymbol{v}_{j_1} \otimes \boldsymbol{v}_{j_2} \otimes \cdots \otimes \boldsymbol{v}_{j_{k-1}}$ for $\boldsymbol{j} \in [\![1, r]\!]^{k-1}$ and $\boldsymbol{y}_0^{\otimes(k-1)}, \boldsymbol{y}_1^{\otimes(k-1)}, \cdots, \boldsymbol{y}_t^{\otimes(k-1)}$, obtained by the Gram-Schmidt orthonormalization procedure. More precisely, given those tensors $\{\boldsymbol{v}_{j_1} \otimes \boldsymbol{v}_{j_2} \otimes \cdots \otimes \boldsymbol{v}_{j_{k-1}}\}_{\boldsymbol{j} \in [\![1,r]\!]^{k-1}}, \boldsymbol{\tau}_0, \boldsymbol{\tau}_1, \cdots, \boldsymbol{\tau}_{t-1}$, we denote

$$\begin{aligned} &b_{(t+1)\boldsymbol{j}} = \langle \boldsymbol{y}_t^{\otimes(k-1)}, \boldsymbol{v}_{j_1} \otimes \boldsymbol{v}_{j_2} \otimes \cdots \otimes \boldsymbol{v}_{j_{k-1}} \rangle, \quad \boldsymbol{j} = (j_1, j_2, \cdots, j_{k-1}) \in [\![1, r]\!]^{k-1}, \\ &b_{(t+1)s} = \langle \boldsymbol{y}_s^{\otimes(k-1)}, \boldsymbol{\tau}_{s-1} \rangle, \quad 1 \leqslant s \leqslant t, \quad c_{t+1} = \langle \boldsymbol{y}_t^{\otimes(k-1)}, \boldsymbol{\tau}_t \rangle \end{aligned} \tag{119}$$

and $\sum_{\boldsymbol{j}} b_{(t+1)\boldsymbol{j}} \boldsymbol{v}_{j_1} \otimes \boldsymbol{v}_{j_2} \otimes \cdots \otimes \boldsymbol{v}_{j_{k-1}} + b_{(t+1)1}\boldsymbol{\tau}_0 + b_{(t+1)2}\boldsymbol{\tau}_1 + \cdots b_{(t+1)t}\boldsymbol{\tau}_{t-1}$ is the projection of $\boldsymbol{y}_t^{\otimes(k-1)}$ on the span of $\{\boldsymbol{v}_{j_1} \otimes \boldsymbol{v}_{j_2} \otimes \cdots \otimes \boldsymbol{v}_{j_{k-1}}\}_{\boldsymbol{j} \in [\![1,r]\!]^{k-1}}, \boldsymbol{y}_0^{\otimes(k-1)}, \boldsymbol{y}_1^{\otimes(k-1)}, \cdots, \boldsymbol{y}_{t-1}^{\otimes(k-1)}$. Then we can write $\boldsymbol{y}_t^{\otimes(k-1)}$ in terms of the base (116)

$$\boldsymbol{y}_t^{\otimes(k-1)} = \sum_{\boldsymbol{j}} b_{(t+1)\boldsymbol{j}} \boldsymbol{v}_{j_1} \otimes \boldsymbol{v}_{j_2} \otimes \cdots \otimes \boldsymbol{v}_{j_{k-1}} + b_{(t+1)1}\boldsymbol{\tau}_0 + b_{(t+1)2}\boldsymbol{\tau}_1 + \cdots b_{(t+1)t}\boldsymbol{\tau}_{t-1} + c_{t+1}\boldsymbol{\tau}_t. \tag{120}$$

The recursion (110) implies that

$$\boldsymbol{y}_{t+1} = \sum_{j=1}^{r} \beta_j (a_{tj} + b_t \langle \boldsymbol{w}_t, \boldsymbol{v}_j \rangle + c_t \langle \boldsymbol{\xi}_t, \boldsymbol{v}_j \rangle)^{k-1} \boldsymbol{v}_j + b_{t+1}\boldsymbol{w}_{t+1} + c_{t+1}\boldsymbol{Z}[\boldsymbol{\tau}_t] \tag{121}$$

where

$$\begin{aligned} b_{t+1}\boldsymbol{w}_{t+1} &= \boldsymbol{Z}[\sum_{\boldsymbol{j}} b_{(t+1)\boldsymbol{j}} \boldsymbol{v}_{j_1} \otimes \boldsymbol{v}_{j_2} \otimes \cdots \otimes \boldsymbol{v}_{j_{k-1}} + b_{(t+1)1}\boldsymbol{\tau}_0 + b_{(t+1)2}\boldsymbol{\tau}_1 + \cdots b_{(t+1)t}\boldsymbol{\tau}_{t-1}] \\ &= \sum_{\boldsymbol{j}} b_{(t+1)\boldsymbol{j}} \boldsymbol{\xi}_{\boldsymbol{j}} + b_{(t+1)1}\boldsymbol{\xi}_1 + b_{(t+1)2}\boldsymbol{\xi}_2 + \cdots b_{(t+1)t}\boldsymbol{\xi}_t. \end{aligned} \tag{122}$$

Since $\boldsymbol{\tau}_t$ is orthogonal to $\{\boldsymbol{v}_{j_1} \otimes \boldsymbol{v}_{j_2} \otimes \cdots \otimes \boldsymbol{v}_{j_{k-1}}\}_{\boldsymbol{j} \in [\![1,r]\!]^{k-1}}, \boldsymbol{\tau}_0, \boldsymbol{\tau}_1, \cdots, \boldsymbol{\tau}_{t-1}$, Lemma 4.1 implies that conditioning on $\boldsymbol{\xi}_{\boldsymbol{j}} = \boldsymbol{Z}[\boldsymbol{v}_{j_1} \otimes \boldsymbol{v}_{j_2} \otimes \cdots \otimes \boldsymbol{v}_{j_{k-1}}]$ for $\boldsymbol{j} = (j_1, j_2, \cdots, j_{k-1}) \in [\![1, r]\!]^{k-1}$ and

$\boldsymbol{\xi}_{s+1} = \boldsymbol{Z}[\boldsymbol{\tau}_s]$ for $0 \leqslant s \leqslant t-1$, $\boldsymbol{\xi}_{t+1} = \boldsymbol{Z}[\boldsymbol{\tau}_t]$ is an independent Gaussian vector, with each entry $\mathcal{N}(0, 1/n)$. The above discussion gives us that

$$\boldsymbol{y}_{t+1} = a_{t+1}\boldsymbol{v}_{t+1} + b_{t+1}\boldsymbol{w}_{t+1} + c_{t+1}\boldsymbol{\xi}_{t+1}, \quad a_{t+1} = \sqrt{a_{(t+1)1}^2 + a_{(t+1)2}^2 + \cdots + a_{(t+1)r}^2}. \tag{123}$$

and

$$a_{(t+1)j} = \beta_j(a_{tj} + b_t\langle \boldsymbol{w}_t, \boldsymbol{v}_j\rangle + c_t\langle \boldsymbol{\xi}_t, \boldsymbol{v}_j\rangle)^{k-1}, \quad 1 \leqslant j \leqslant r. \tag{124}$$

□

We recall that by our Assumption 2.5, that

$$1/\kappa \leqslant \left|\frac{\langle \boldsymbol{u}, \boldsymbol{v}_i\rangle}{\langle \boldsymbol{u}, \boldsymbol{v}_j\rangle}\right| \leqslant \kappa, \tag{125}$$

for all $1 \leqslant i, j \leqslant r$. If $j_* = \mathrm{argmax}_j\, \beta_j\langle \boldsymbol{u}, \boldsymbol{v}_j\rangle^{k-2}$, it is necessary that $\beta_{j_*} \gtrsim \beta_1$, where the implicit constant depends on κ.

In the following, we study the case that $\langle \boldsymbol{u}, \boldsymbol{v}_{j_*}\rangle > 0$. The case $\langle \boldsymbol{u}, \boldsymbol{v}_{j_*}\rangle < 0$ can be proven in exactly the same way, by simply changing $(\beta, \boldsymbol{v}_{j_*})$ with $((-1)^k\beta, -\boldsymbol{v}_{j_*})$. We prove by induction

Claim 4.5. *For any fixed time t, with probability at least $1 - \mathrm{O}(e^{-c(\log n)^2})$ the following holds: for any $s \leqslant t$,*

$$\begin{aligned} &|a_{sj_*}| \geqslant |a_{sj}|, \quad |a_s| \gtrsim |\beta_1|(\sum_{\boldsymbol{j}} |b_{s\boldsymbol{j}}| + |b_{s1}| + \cdots + |b_{s(s-1)}|), \\ &|a_s| \gtrsim n^{\varepsilon}\max\{\mathbf{1}(k \geqslant 3)|c_s/\beta_1^{1/(k-2)}|, |c_s/\sqrt{n}|\}, \end{aligned} \tag{126}$$

and for $\boldsymbol{j} = (j_1, j_2, \cdots, j_{k-1}) \in [\![1, r]\!]^{k-1}$

$$\begin{aligned} &\|\boldsymbol{\xi}_{\boldsymbol{j}}\|, \|\boldsymbol{\xi}_s\|_2 = 1 + \mathrm{O}(\log n/\sqrt{n}), \quad |\langle \boldsymbol{v}_j, \boldsymbol{\xi}_{\boldsymbol{j}}\rangle|, |\langle \boldsymbol{a}, \boldsymbol{\xi}_{\boldsymbol{j}}\rangle|, |\langle \boldsymbol{a}, \boldsymbol{\xi}_s\rangle| \lesssim \log n/\sqrt{n}. \\ &\|\mathrm{Proj}_{\mathrm{Span}\{\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_r, \{\boldsymbol{\xi}_{\boldsymbol{j}}\}_{\boldsymbol{j}\in[\![1,r]\!]^{k-1}}, \boldsymbol{\xi}_1, \cdots, \cdots, \boldsymbol{\xi}_{s-1}\}}(\boldsymbol{\xi}_s)\|_2 \lesssim \log n/\sqrt{n}. \end{aligned} \tag{127}$$

Proof of Claim 4.5. From (115), we have

$$\boldsymbol{y}_1 = \sum_{j=1}^r \beta_j\langle \boldsymbol{y}_0, \boldsymbol{v}_j\rangle^{k-1}\boldsymbol{v}_j + \sum_{\boldsymbol{j}}\prod_{s=1}^{k-1}\langle \boldsymbol{y}_0, \boldsymbol{v}_{j_s}\rangle\boldsymbol{\xi}_{\boldsymbol{j}} + \sqrt{1 - \left(\sum_{j=1}^r \langle \boldsymbol{y}_0, \boldsymbol{v}_j\rangle^2\right)^{(k-1)}}\,\boldsymbol{\xi}_1 \tag{128}$$

$$= \sum_{j=1}^r a_{1j}\boldsymbol{v}_j + \sum_{\boldsymbol{j}} b_{1\boldsymbol{j}}\boldsymbol{\xi}_{\boldsymbol{j}} + c_1\boldsymbol{\xi}_1, \tag{129}$$

where $a_{1j} = \beta_j\langle \boldsymbol{u}, \boldsymbol{v}_j\rangle^{k-1}$ for $1 \leqslant j \leqslant r$, $b_{1\boldsymbol{j}} = \prod_{s=1}^{k-1}\langle \boldsymbol{u}, \boldsymbol{v}_{j_s}\rangle$ for any index $\boldsymbol{j} = (j_1, j_2, \cdots, j_{k-1})$ and $c_1 = \sqrt{1 - \left(\sum_{j=1}^r \langle \boldsymbol{u}, \boldsymbol{v}_j\rangle^2\right)^{(k-1)}}$. Since $\boldsymbol{\xi}_{\boldsymbol{j}}$ are independent Gaussian vectors with each entry mean zero and variance $1/n$, the concentration for chi-square distribution implies that $\|\boldsymbol{\xi}_{\boldsymbol{j}}\|_2 = 1 + \mathrm{O}(\log n/\sqrt{n})$ with probability $1 - e^{c(\log n)^2}$. Since $j_* = \mathrm{argmax}_j\, |\beta_j\langle \boldsymbol{u}, \boldsymbol{v}_j\rangle^{k-2}|$, combining with

our Assumption 2.5, it gives that $|a_{1j_*}| \geqslant |a_{1j}|/\kappa$. As a consequence, we also have that $|a_1| = \sqrt{a_{11}^2 + a_{12}^2 + \cdots + a_{1r}^2} \asymp |a_{1j_*}|$. Again using our Assumption 2.5

$$\sum_{\boldsymbol{j}} |b_{1\boldsymbol{j}}| \lesssim \left(\sum_{j=1}^{r} |\langle \boldsymbol{u}, \boldsymbol{v}_j\rangle|\right)^{k-1} \lesssim \sum_{j=1}^{r} |\langle \boldsymbol{u}, \boldsymbol{v}_j\rangle|^{k-1} \lesssim |a_{1j_*}|/\beta_{j_*} \lesssim |a_1|/\beta_1. \tag{130}$$

We can check that $|\beta_1^{1/(k-2)} a_1| \asymp |\beta_{j_*}^{1/(k-2)} a_{1j_*}| = |\beta_{j_*}\langle \boldsymbol{u}, \boldsymbol{v}_{j_*}\rangle^{k-2}|^{(k-1)/(k-2)} \geqslant n^{\varepsilon} \geqslant n^{\varepsilon}|c_1|$, and $|\sqrt{n}a_1| \asymp |\sqrt{n}a_{1j_*}| = |\beta_{j_*}\langle \boldsymbol{u}, \boldsymbol{v}_{j_*}\rangle^{k-2}||\sqrt{n}\langle \boldsymbol{u}, \boldsymbol{v}_{j_*}\rangle| \gtrsim n^{\varepsilon} \geqslant n^{\varepsilon}|c_1|$. Moreover, conditioning on $\boldsymbol{\xi}_{\boldsymbol{j}} = \boldsymbol{Z}[\boldsymbol{v}_{j_1} \otimes \boldsymbol{v}_{j_2} \otimes \cdots \otimes \boldsymbol{v}_{j_{k-1}}]$ for $\boldsymbol{j} = (j_1, j_2, \cdots, j_{k-1}) \in [\![1, r]\!]^{k-1}$, Lemma 4.1 implies that $\boldsymbol{\xi}_1 = \boldsymbol{Z}[\boldsymbol{\tau}_0]$ is an independent Gaussian random vector with each entry $\mathcal{N}(0, 1/n)$. By the standard concentration inequality, it holds that with probability $1 - e^{c(\log n)^2}$, $\|\boldsymbol{\xi}_1\|_2 = 1 + \mathrm{O}(\log n/\sqrt{n})$, $|\langle \boldsymbol{a}, \boldsymbol{\xi}_1\rangle|$ and the projection of $\boldsymbol{\xi}_1$ on the span of $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_r, \{\boldsymbol{\xi}_{\boldsymbol{j}}\}_{\boldsymbol{j}\in[\![1,r]\!]^{k-1}}\}$ is bounded by $\log n/\sqrt{n}$. So far we have proved that (53) and (54) hold for $t = 1$.

In the following, we assume that (126) and (127) hold for t, and prove it for $t + 1$. We recall from (118) and (124) that

$$a_{(t+1)j} = \beta_j(a_{tj} + b_t\langle \boldsymbol{w}_t, \boldsymbol{v}_j\rangle + c_t\langle \boldsymbol{\xi}_t, \boldsymbol{v}_j\rangle)^{k-1}, \quad b_t\boldsymbol{w}_t = \sum_{\boldsymbol{j}} b_{t\boldsymbol{j}}\boldsymbol{\xi}_{\boldsymbol{j}} + b_{t1}\boldsymbol{\xi}_1 + \cdots + b_{tt-1}\boldsymbol{\xi}_{t-1}. \tag{131}$$

By our induction hypothesis, we have

$$|b_t\langle \boldsymbol{w}_t, \boldsymbol{v}_j\rangle| \lesssim \sum_{\boldsymbol{j}} |b_{t\boldsymbol{j}}\langle \boldsymbol{\xi}_{\boldsymbol{j}}, \boldsymbol{v}_j\rangle| + |b_{t1}\langle \boldsymbol{\xi}_1, \boldsymbol{v}_j\rangle| + \cdots + |b_{t(t-1)}\langle \boldsymbol{\xi}_{t-1}, \boldsymbol{v}_j\rangle| \lesssim (\log n/\sqrt{n})|a_t|/|\beta_1|, \tag{132}$$

and

$$|c_t\langle \boldsymbol{\xi}_t, \boldsymbol{v}_j\rangle| \lesssim (\log n/\sqrt{n})|c_t| \lesssim (\log n)|a_t|/n^{\varepsilon}. \tag{133}$$

It follows from plugging (132) and (133) into (131), we get

$$a_{(t+1)j} = \beta_j(a_{tj} + b_t\langle \boldsymbol{w}_t, \boldsymbol{v}_j\rangle + c_t\langle \boldsymbol{\xi}_t, \boldsymbol{v}_j\rangle)^{k-1} = \beta_j(a_{tj} + \mathrm{O}(\log n|a_t|/n^{\varepsilon}))^k \lesssim \beta_j|a_t|^{k-1}, \tag{134}$$

and especially

$$a_{(t+1)j_*} = \beta_{j_*}(a_{tj_*} + \mathrm{O}(\log n|a_{tj_*}|/n^{\varepsilon}))^k = (1 + \mathrm{O}(\log n/n^{\varepsilon}))\beta_{j_*}a_{tj_*}^{k-1}. \tag{135}$$

Therefore, we conclude that

$$|a_{t+1j_*}| \asymp |\beta_{j_*}a_{tj_*}^{k-1}| \asymp |\beta a_t^{k-1}|, \tag{136}$$

and

$$|a_{(t+1)j}| \lesssim \beta_j|a_t|^{k-1} \lesssim \beta_{j_*}|a_{tj_*}|^{k-1} \lesssim |a_{t+1j_*}|. \tag{137}$$

We recall from (119), $\sum_{\boldsymbol{j}} b_{(t+1)\boldsymbol{j}}\boldsymbol{v}_{j_1} \otimes \boldsymbol{v}_{j_2} \otimes \cdots \otimes \boldsymbol{v}_{j_{k-1}} + b_{(t+1)1}\boldsymbol{\tau}_0 + b_{(t+1)2}\boldsymbol{\tau}_1 + \cdots b_{(t+1)t}\boldsymbol{\tau}_{t-1}$ is the projection of $\boldsymbol{y}_t^{\otimes(k-1)}$ on the span of $\{\boldsymbol{v}_{j_1} \otimes \boldsymbol{v}_{j_2} \otimes \cdots \otimes \boldsymbol{v}_{j_{k-1}}\}_{\boldsymbol{j}\in[\![1,r]\!]^{k-1}}, \boldsymbol{y}_0^{\otimes(k-1)}, \boldsymbol{y}_1^{\otimes(k-1)}, \cdots, \boldsymbol{y}_{t-1}^{\otimes(k-1)}$. We also recall that $\{\boldsymbol{v}_{j_1} \otimes \boldsymbol{v}_{j_2} \otimes \cdots \otimes \boldsymbol{v}_{j_{k-1}}\}_{\boldsymbol{j}\in[\![1,r]\!]^{k-1}}, \boldsymbol{\tau}_0, \boldsymbol{\tau}_1, \cdots, \boldsymbol{\tau}_{t-1}$ are obtained from $\{\boldsymbol{v}_{j_1} \otimes \boldsymbol{v}_{j_2} \otimes \cdots \otimes \boldsymbol{v}_{j_{k-1}}\}_{\boldsymbol{j}\in[\![1,r]\!]^{k-1}}, , \boldsymbol{y}_0^{\otimes(k-1)}, \boldsymbol{y}_1^{\otimes(k-1)}, \cdots, \boldsymbol{y}_{t-1}^{\otimes(k-1)}$ by the Gram-Schmidt orthonormalization procedure. So we have that the span of vectors $\{\boldsymbol{v}_{j_1} \otimes \boldsymbol{v}_{j_2} \otimes \cdots \otimes \boldsymbol{v}_{j_{k-1}}\}_{\boldsymbol{j}\in[\![1,r]\!]^{k-1}}, \boldsymbol{\tau}_0, \boldsymbol{\tau}_1, \cdots, \boldsymbol{\tau}_{t-1}$ is the same as the span of vectors $\{\boldsymbol{v}_{j_1} \otimes \boldsymbol{v}_{j_2} \otimes \cdots \otimes \boldsymbol{v}_{j_{k-1}}\}_{\boldsymbol{j}\in[\![1,r]\!]^{k-1}}, , \boldsymbol{y}_0^{\otimes(k-1)}, \boldsymbol{y}_1^{\otimes(k-1)}, \cdots, \boldsymbol{y}_{t-1}^{\otimes(k-1)}$, which is

contained in the span of $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_r, \boldsymbol{w}_t, \boldsymbol{y}_0, \cdots, \boldsymbol{y}_{t-1}\}^{\otimes(k-1)}$. Moreover from the relation (117) and (118), one can see that the span of $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_r, \boldsymbol{w}_t, \boldsymbol{y}_0, \cdots, \boldsymbol{y}_{t-1}\}$ is the same as the span of $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_r, \{\boldsymbol{\xi}_{\boldsymbol{j}}\}_{\boldsymbol{j}\in[\![1,r]\!]^{k-1}}, \boldsymbol{\xi}_1, \cdots, \boldsymbol{\xi}_{t-1}\}$. It follows that

$$\begin{aligned}
|b_{t+1}| &\lesssim \sqrt{\sum_{\boldsymbol{j}} b_{(t+1)\boldsymbol{j}}^2 + b_{(t+1)1}^2 + b_{(t+1)2}^2 + \cdots + b_{(t+1)t}^2} \\
&= \|\mathrm{Proj}_{\mathrm{Span}\{\{\boldsymbol{v}_{j_1}\otimes\boldsymbol{v}_{j_2}\otimes\cdots\otimes\boldsymbol{v}_{j_{k-1}}\}_{\boldsymbol{j}\in[\![1,r]\!]^{k-1}}, \boldsymbol{\tau}_0, \boldsymbol{\tau}_1, \cdots, \boldsymbol{\tau}_{t-1}\}}(a_t\boldsymbol{v}_t + b_t\boldsymbol{w}_t + c_t\boldsymbol{\xi}_t)^{\otimes(k-1)}\|_2 \\
&\leqslant \|\mathrm{Proj}_{\mathrm{Span}\{\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_r, \boldsymbol{w}_t, \boldsymbol{y}_0, \cdots, \boldsymbol{y}_{t-1}\}^{\otimes(k-1)}}(a_t\boldsymbol{v}_t + b_t\boldsymbol{w}_t + c_t\boldsymbol{\xi}_t)^{\otimes(k-1)}\|_2 \\
&\leqslant \|\mathrm{Proj}_{\mathrm{Span}\{\boldsymbol{v}, \boldsymbol{w}_t, \boldsymbol{y}_0, \cdots, \boldsymbol{y}_{t-1}\}}(a_t\boldsymbol{v}_t + b_t\boldsymbol{w}_t + c_t\boldsymbol{\xi}_t)\|_2^{k-1} \\
&= \|a_t\boldsymbol{v}_t + b_t\boldsymbol{w}_t + c_t\mathrm{Proj}_{\mathrm{Span}\{\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_r, \{\boldsymbol{\xi}_{\boldsymbol{j}}\}_{\boldsymbol{j}\in[\![1,r]\!]^{k-1}}, \boldsymbol{\xi}_1, \cdots, \boldsymbol{\xi}_{t-1}\}}(\boldsymbol{\xi}_t)\|_2^{k-1} \\
&\lesssim \left(|a_t| + |b_t| + \frac{\log n|c_t|}{\sqrt{n}}\right)^{k-1} \lesssim |a_t|^{k-1} \lesssim |a_{t+1}|/\beta_1,
\end{aligned} \tag{138}$$

where in the first line we used (122), and in the last line of (138) we used our induction hypothesis that $\|\mathrm{Proj}_{\mathrm{Span}\{\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_r, \{\boldsymbol{\xi}_{\boldsymbol{j}}\}_{\boldsymbol{j}\in[\![1,r]\!]^{k-1}}, \boldsymbol{\xi}_1, \cdots, \boldsymbol{\xi}_{t-1}\}}(\boldsymbol{\xi}_t)\|_2 \lesssim \log n/\sqrt{n}$.

Finally we estimate c_{t+1}. We recall from (119), the coefficient c_{t+1} is the remainder of $\boldsymbol{y}_t^{\otimes(k-1)}$ after projecting on $\{\boldsymbol{v}_{j_1}\otimes\boldsymbol{v}_{j_2}\otimes\cdots\otimes\boldsymbol{v}_{j_{k-1}}\}_{\boldsymbol{j}\in[\![1,r]\!]^{k-1}}, \boldsymbol{\tau}_0, \boldsymbol{\tau}_1, \cdots, \boldsymbol{\tau}_{t-1}$. It is bounded by the remainder of $\boldsymbol{y}_t^{\otimes(k-1)}$ after projecting on $\{\boldsymbol{v}_{j_1}\otimes\boldsymbol{v}_{j_2}\otimes\cdots\otimes\boldsymbol{v}_{j_{k-1}}\}_{\boldsymbol{j}\in[\![1,r]\!]^{k-1}}$,

$$|c_{t+1}| \leqslant \|\boldsymbol{y}_t^{\otimes(k-1)} - a_t^{k-1}\boldsymbol{v}_t^{\otimes(k-1)}\|_2 = \|(a_t\boldsymbol{v}_t + b_t\boldsymbol{w}_t + c_t\boldsymbol{\xi}_t)^{\otimes(k-1)} - a_t^{k-1}\boldsymbol{v}_t^{\otimes(k-1)}\|_2. \tag{139}$$

The difference $(a_t\boldsymbol{v}_t + b_t\boldsymbol{w}_t + c_t\boldsymbol{\xi}_t)^{\otimes(k-1)} - a_t^{k-1}\boldsymbol{v}_t^{\otimes(k-1)}$ is a sum of terms in the following form,

$$\boldsymbol{\eta}_1\otimes\boldsymbol{\eta}_2\otimes\cdots\otimes\boldsymbol{\eta}_{k-1}, \tag{140}$$

where $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \cdots, \boldsymbol{\eta}_{k-1} \in \{a_t\boldsymbol{v}_t, b_t\boldsymbol{w}_t + c_t\boldsymbol{\xi}_t\}$, and at least one of them is $b_t\boldsymbol{w}_t + c_t\boldsymbol{\xi}_t$. We notice that by our induction hypothesis, $\|b_t\boldsymbol{w}_t + c_t\boldsymbol{\xi}_t\|_2 \lesssim |b_t|\|\boldsymbol{w}_t\|_2 + |c_t|\|\boldsymbol{\xi}_t\|_2 \lesssim |b_t| + |c_t|$. For the L_2 norm of (140), each copy of $a_t\boldsymbol{v}_t$ contributes a_t and each copy of $b_t\boldsymbol{w}_t + c_t\boldsymbol{\xi}_t$ contributes a factor $|b_t| + |c_t|$. We conclude that

$$|c_{t+1}| \leqslant \|(a_t\boldsymbol{v}_t + b_t\boldsymbol{w}_t + c_t\boldsymbol{\xi}_t)^{\otimes(k-1)} - a_t^{k-1}\boldsymbol{v}_t^{\otimes(k-1)}\|_2 \lesssim \sum_{r=1}^{k-1}|a_t|^{k-1-r}(|b_t| + |c_t|)^r. \tag{141}$$

Combining with (136) that $|a_{t+1}| \asymp |\beta_1||a_t|^{k-1}$, we divide both sides of (141) by $|\beta_1||a_t|^{k-1}$,

$$\frac{|c_{t+1}|}{|a_{t+1}|} \lesssim \frac{1}{|\beta_1|}\sum_{r=1}^{k-1}\left(\frac{|b_t|}{|a_t|} + \frac{|c_t|}{|a_t|}\right)^r \lesssim \frac{1}{|\beta_1|}\sum_{r=1}^{k-1}\left(\frac{1}{|\beta_1|} + \frac{|c_t|}{|a_t|}\right)^r \tag{142}$$

There are three cases:

1. If $|c_t|/|a_t| \geqslant 1$, then

$$\frac{|c_{t+1}|}{|a_{t+1}|} \lesssim \frac{1}{|\beta_1|}\sum_{r=1}^{k-1}\left(\frac{1}{|\beta_1|} + \frac{|c_t|}{|a_t|}\right)^r \lesssim \frac{1}{|\beta_1|}\left(\frac{|c_t|}{|a_t|}\right)^{k-1}. \tag{143}$$

If $k = 2$, then $|c_{t+1}|/|a_{t+1}| \lesssim (|c_t|/|a_t|)/n^\varepsilon$. If $k \geqslant 2$, by our induction hypothesis $|c_t|/|a_t| \lesssim \beta_1^{1/(k-2)}/n^\varepsilon$. Especially, $(|c_t|/|a_t|)^{k-2}/|\beta_1| \lesssim 1/n^\varepsilon$. We still get that $|c_{t+1}|/|a_{t+1}| \lesssim (|c_t|/|a_t|)/n^\varepsilon$.

2. If $1/|\beta_1| \lesssim |c_t|/|a_t| \leqslant 1$, then

$$\frac{|c_{t+1}|}{|a_{t+1}|} \lesssim \frac{1}{|\beta_1|}\sum_{r=1}^{k-1}\left(\frac{1}{|\beta_1|}+\frac{|c_t|}{|a_t|}\right)^r \lesssim \frac{1}{|\beta_1|}\left(\frac{|c_t|}{|a_t|}\right) \lesssim \frac{1}{n^{\varepsilon}}\left(\frac{|c_t|}{|a_t|}\right). \tag{144}$$

3. Finally for $|c_t|/|a_t| \lesssim 1/|\beta_1|$, we will have

$$\frac{|c_{t+1}|}{|a_{t+1}|} \lesssim \frac{1}{|\beta_1|}\sum_{r=1}^{k-1}\left(\frac{1}{|\beta_1|}+\frac{|c_t|}{|a_t|}\right)^r \lesssim \frac{1}{|\beta_1|}\left(\frac{1}{|\beta_1|}\right) \lesssim \frac{1}{|\beta_1|^2}. \tag{145}$$

In all these cases we have $|c_{t+1}|/|a_{t+1}| \lesssim \min\{\sqrt{n}, \mathbf{1}(k\geqslant 3)|\beta_1|^{1/(k-2)}\}/n^{\varepsilon}$. This finishes the proof of the induction (126).

For (127), since $\boldsymbol{\tau}_t$ is orthogonal to $\{\boldsymbol{v}_{j_1}\otimes\boldsymbol{v}_{j_2}\otimes\cdots\otimes\boldsymbol{v}_{j_{k-1}}\}_{\boldsymbol{j}\in[\![1,r]\!]^{k-1}}, \boldsymbol{\tau}_0, \boldsymbol{\tau}_1, \cdots, \boldsymbol{\tau}_{t-1}$, Lemma 4.1 implies that conditioning on $\boldsymbol{\xi}_{\boldsymbol{j}} = \boldsymbol{Z}[\boldsymbol{v}_{j_1}\otimes\boldsymbol{v}_{j_2}\otimes\cdots\otimes\boldsymbol{v}_{j_{k-1}}]$ for index $\boldsymbol{j}=(j_1,j_2,\cdots,j_{k-1})\in[\![1,r]\!]^{k-1}$ and $\boldsymbol{\xi}_{s+1} = \boldsymbol{Z}[\boldsymbol{\tau}_s]$ for $0\leqslant s\leqslant t-1$, $\boldsymbol{\xi}_{t+1} = \boldsymbol{Z}[\boldsymbol{\tau}_t]$ is an independent Gaussian vector, with each entry $\mathcal{N}(0,1/n)$. By the standard concentration inequality, it holds that with probability $1-e^{c(\log n)^2}$, $\|\boldsymbol{\xi}_{t+1}\|_2 = 1+\mathrm{O}(\log n/\sqrt{n})$, $|\langle \boldsymbol{a}, \boldsymbol{\xi}_{t+1}\rangle|$ and the projection of $\boldsymbol{\xi}_{t+1}$ on the span of $\{\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_r,\{\boldsymbol{\xi}_{\boldsymbol{j}}\}_{\boldsymbol{j}\in[\![1,r]\!]^{k-1}},\boldsymbol{\xi}_1,\cdots,\boldsymbol{\xi}_{t-1}\}$ is bounded by $\log n/\sqrt{n}$. This finishes the proof of the induction (127). □

Next, using (126) and (127) as input, we prove that for

$$t \geqslant 1+\frac{1}{\varepsilon}\left(\frac{1}{2}+\frac{2\log|\beta_1|}{\log n}\right)+\frac{\log\log(\sqrt{n}|\beta_1|)}{\log(k-1)} \tag{146}$$

we have

$$\boldsymbol{y}_t = \sum_{j=1}^{r} a_{tj}\boldsymbol{v}_j + \sum_{\boldsymbol{j}} b_{t\boldsymbol{j}}\boldsymbol{\xi}_{\boldsymbol{j}} + b_{t1}\boldsymbol{\xi}_1+\cdots+b_{tt-1}\boldsymbol{\xi}_{t-1}+c_t\boldsymbol{\xi}_t, \tag{147}$$

such that

$$\begin{aligned}
&|a_{tj}| \lesssim \left(\frac{\log n}{\sqrt{n}}\frac{1}{|\beta_1|}\right)^{k-1}|a_{tj_*}|, \quad j\neq j_*,\\
&b_{t(j_*,j_*,\cdots,j_*)} = \frac{a_{tj_*}}{\beta_{j_*}} + \mathrm{O}\left(\frac{\log n|a_t|}{|\beta_1|^2\sqrt{n}}\right), |b_{(t+1)\boldsymbol{j}_*}| \lesssim \frac{\log n}{\sqrt{n}|\beta_1|^2}|a_{tj_*}|, \quad \boldsymbol{j}_* = (j_*,j_*,\cdots,j_*),\\
&|b_{t1}|,|b_{t2}|,\cdots,|b_{t(t-1)}| \lesssim \frac{(\log n)^{1/2}|a_t|}{|\beta_1|^{3/2}n^{1/4}}, \quad |c_t| \lesssim |a_t|/\beta_1^2
\end{aligned} \tag{148}$$

Let $x_t = |c_t/a_t| \leqslant n^{-\varepsilon}|\beta|^{1/(k-2)}$, and $r_t = \max_{j\neq j_*}(\beta_j^{1/(k-2)}a_{tj})/(\beta_{j_*}^{1/(k-2)}a_{tj_*})$. For $t=1$, our Assumption 2.6 implies that

$$\begin{aligned}
\beta_j^{1/(k-2)}a_{1j} &\leqslant (\beta_j\langle\boldsymbol{u},\boldsymbol{v}_j\rangle^{k-2})^{(k-1)/(k-2)}\\
&\leqslant ((1-1/\kappa)\beta_{j_*}\langle\boldsymbol{u},\boldsymbol{v}_{j_*}\rangle^{k-2})^{(k-1)/(k-2)} \leqslant (1-1/\kappa)\beta_{j_*}^{1/(k-2)}a_{1j_*}.
\end{aligned} \tag{149}$$

Thus we have that $r_1\leqslant(1-1/\kappa)$. We recall from (131)

$$\begin{aligned}
\beta_j^{1/(k-2)}a_{(t+1)j} &= \left(\beta_j^{1/k-2}(a_{tj}+b_t\langle\boldsymbol{w}_t,\boldsymbol{v}_j\rangle+c_t\langle\boldsymbol{\xi}_t,\boldsymbol{v}_j\rangle)\right)^{k-1}\\
&= \left(\beta_j^{1/k-2}(a_{tj}+\mathrm{O}\left(|a_t|\frac{\log n(1/|\beta_1|+x_t)}{\sqrt{n}}\right)\right)^{k-1},
\end{aligned} \tag{150}$$

where we used (126) and (127). Thus it follows that

$$\begin{aligned} r_{t+1} &= \max_{j\neq j_*} \left(\frac{\beta_j^{1/k-2}(a_{tj} + \mathrm{O}\left(|a_t|\log n(1/|\beta_1| + x_t)/\sqrt{n}\right)}{\beta_{j_*}^{1/k-2}(a_{tj_*} + \mathrm{O}\left(|a_t|\log n(1/|\beta_1| + x_t)/\sqrt{n}\right)} \right)^{k-1} \\ &\leqslant \left(\frac{r_t + \mathrm{O}\left(\log n(1/|\beta_1| + x_t)/\sqrt{n}\right)}{1 + \mathrm{O}\left(\log n(1/|\beta_1| + x_t)/\sqrt{n}\right)} \right)^{k-1} \end{aligned} \tag{151}$$

For x_t, (142) implies

$$x_{t+1} \lesssim \frac{1}{|\beta_1|} \sum_{r=1}^{k-1} \left(\frac{1}{|\beta_1|} + x_t \right)^r, \tag{152}$$

from the discussion after (142), we have that either $x_{t+1} \lesssim 1/|\beta_1|^2$, or $x_{t+1} \lesssim x_t/n^{\varepsilon}$. Since $x_1 = |c_1/a_1| \lesssim n^{1/2-\varepsilon}$, and $r_1 \leqslant (1 - 1/\kappa)$ we conclude from (151) and (152) that

$$x_t = |c_t/a_t| \lesssim 1/\beta_1^2, \quad r_t \lesssim (\log n/(|\beta_1|\sqrt{n}))^{k-1}, \tag{153}$$

when

$$t \geqslant \frac{1}{\varepsilon}\left(\frac{1}{2} + \frac{2\log|\beta_1|}{\log n}\right) + \frac{\log\log(\sqrt{n}|\beta_1|)}{\log(k-1)}. \tag{154}$$

To derive the upper bound of $b_{t1}, b_{t2}, \cdots, b_{t(t-1)}$, we use (138).

$$\begin{aligned} &\sum_{\boldsymbol{j}} b_{(t+1)\boldsymbol{j}}^2 + b_{(t+1)1}^2 + b_{(t+1)2}^2 + \cdots + b_{(t+1)t}^2 \\ &\leqslant \|a_t\boldsymbol{v}_t + b_t\boldsymbol{w}_t + c_t \mathrm{Proj}_{\mathrm{Span}\{\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_r,\{\boldsymbol{\xi}_{\boldsymbol{j}}\}_{\boldsymbol{j}\in[\![1,r]\!]^{k-1}},\boldsymbol{\xi}_1,\cdots,\boldsymbol{\xi}_{t-1}\}}(\boldsymbol{\xi}_t)\|_2^{2(k-1)} \\ &= \left(a_t^2 + \mathrm{O}\left(|a_t|\left(|b_t| + |c_t|\right)\frac{\log n}{\sqrt{n}} + \left(|b_t| + |c_t|\frac{\log n}{\sqrt{n}}\right)^2 \right) \right)^{k-1}, \end{aligned} \tag{155}$$

where we used (127). The first term $b_{(t+1)\boldsymbol{j}}$ is the projection of $\boldsymbol{y}_t^{\otimes(k-1)}$ on $\boldsymbol{v}_{j_1} \otimes \boldsymbol{v}_{j_2} \otimes \cdots \otimes \boldsymbol{v}_{j_{k-1}}$,

$$b_{(t+1)\boldsymbol{j}} = \prod_{s=1}^{k-1} \langle a_t\boldsymbol{v}_t + b_t\boldsymbol{w}_t + c_t\boldsymbol{\xi}_t, \boldsymbol{v}_{j_s}\rangle = \prod_{s=1}^{k-1} \left(a_{tj_s} + \mathrm{O}\left(\frac{\log n(|b_t| + |c_t|)}{\sqrt{n}} \right) \right), \tag{156}$$

and

$$\begin{aligned} \sum_{\boldsymbol{j}} b_{(t+1)\boldsymbol{j}}^2 &= \left(\sum_{s=1}^{k-1} |\langle a_t\boldsymbol{v}_t + b_t\boldsymbol{w}_t + c_t\boldsymbol{\xi}_t, \boldsymbol{v}_{j_s}\rangle|^2 \right)^k \\ &= \left(a_t^2 + \mathrm{O}\left(|a_t|\left(|b_t| + |c_t|\right)\frac{\log n}{\sqrt{n}} + \left(|b_t| + |c_t|\frac{\log n}{\sqrt{n}}\right)^2 \right) \right)^{k-1}, \end{aligned} \tag{157}$$

where we used (54) that $|\langle\boldsymbol{\xi}_{\boldsymbol{j}}, \boldsymbol{v}_j\rangle|, |\langle\boldsymbol{\xi}_1, \boldsymbol{v}_j\rangle|, \cdots, |\langle\boldsymbol{\xi}_t, \boldsymbol{v}_j\rangle| \lesssim \log n/\sqrt{n}$. Now we can take difference of (155) and (157), and use that $|b_t| \lesssim |a_t|/|\beta_1|$ from (126) and $|c_t| \lesssim |a_t|/|\beta_1|^2$ from (153),

$$b_{(t+1)1}^2 + b_{(t+1)2}^2 + \cdots + b_{(t+1)t}^2 \lesssim a_t^{2(k-1)} \frac{\log n}{|\beta|\sqrt{n}}. \tag{158}$$

Using (156) and (153), we get that

$$\begin{aligned} b_{(t+1)\boldsymbol{j}_*} &= a_{tj_*}^{k-1}\left(1+\mathrm{O}\left(\frac{\log n}{\sqrt{n}|\beta_1|}\right)\right), \quad \boldsymbol{j}_* = (j_*, j_*, \cdots, j_*) \\ |b_{(t+1)\boldsymbol{j}}| &\lesssim \frac{\log n}{\sqrt{n}|\beta_1|}|a_{tj_*}|^{k-1}, \quad \boldsymbol{j} \neq \boldsymbol{j}_*. \end{aligned} \tag{159}$$

From (131), (136) and (153), we have that

$$\begin{aligned} a_{(t+1)j_*} &= \beta_{j_*} b_{(t+1)\boldsymbol{j}_*} = \beta_{j_*} a_{tj_*}^{k-1}\left(1+\mathrm{O}\left(\frac{\log n}{\sqrt{n}|\beta_1|}\right)\right), \\ |a_{(t+1)j}| &\lesssim \left(\frac{\log n}{\sqrt{n}|\beta_1|}\right)^{k-1}|a_{(t+1)j_*}|, \quad j \neq j_*. \end{aligned} \tag{160}$$

Using the above relation, we can simplify (158) and (159) as

$$|b_{(t+1)1}|, |b_{(t+1)2}|, \cdots |b_{(t+1)t}| \lesssim \frac{(\log n)^{1/2}|a_{t+1}|}{|\beta_1|^{3/2}n^{1/4}}. \tag{161}$$

and

$$\begin{aligned} b_{(t+1)\boldsymbol{j}_*} &= \frac{a_{(t+1)j_*}}{\beta_{j_*}}\left(1+\mathrm{O}\left(\frac{\log n}{\sqrt{n}|\beta_1|}\right)\right), \\ |b_{(t+1)\boldsymbol{j}}| &\lesssim \frac{\log n}{\sqrt{n}|\beta_1|^2}|a_{(t+1)j_*}|, \quad \boldsymbol{j} \neq \boldsymbol{j}_*. \end{aligned} \tag{162}$$

This finishes the proof of (148).

With the expression (148), we can process to prove our main results (20) and (21). Thanks to (127) and (147), for t satisfies (146), we have that with probability at least $1-\mathrm{O}(e^{-c(\log n)^2})$

$$\|\boldsymbol{y}_t\|_2^2 = a_{tj_*}^2\left(1+\frac{1}{\beta_{j_*}^2}+\frac{2\langle \boldsymbol{v}_{j_*}, \boldsymbol{\xi}_{\boldsymbol{j}_*}\rangle}{\beta_{j_*}}+\mathrm{O}\left(\frac{\log n}{\sqrt{n}}\left(\frac{\log n}{\sqrt{n}|\beta_1|}\right)^{k-1}+\frac{\log n}{\beta_1^2\sqrt{n}}+\frac{(\log n)^{3/2}}{|\beta_1|^{3/2}n^{3/4}}+\frac{1}{\beta_1^4}\right)\right) \tag{163}$$

where $\boldsymbol{j}_* = (j_*, j_*, \cdots, j_*)$. By rearranging it we get

$$1/\|\boldsymbol{y}_t\|_2 = a_{tj_*}^2\left(1-\frac{1}{2\beta_{j_*}^2}-\frac{2\langle \boldsymbol{v}_{j_*}, \boldsymbol{\xi}_{\boldsymbol{j}_*}\rangle}{\beta_{j_*}}+\mathrm{O}\left(\frac{\log n}{\sqrt{n}}\left(\frac{\log n}{\sqrt{n}|\beta_1|}\right)^{k-1}+\frac{\log n}{\beta_1^2\sqrt{n}}+\frac{(\log n)^{3/2}}{|\beta_1|^{3/2}n^{3/4}}+\frac{1}{\beta_1^4}\right)\right) \tag{164}$$

We can take the inner product $\langle \boldsymbol{a}, \boldsymbol{y}_t\rangle$, and multiply (164)

$$\begin{aligned} \langle \boldsymbol{a}, \boldsymbol{u}_t\rangle = \frac{\langle \boldsymbol{a}, \boldsymbol{y}_t\rangle}{\|\boldsymbol{y}_t\|_2} &= \mathrm{sgn}(a_{tj_*})\left(\left(1-\frac{1}{2\beta_{j_*}^2}\right)\langle \boldsymbol{a}, \boldsymbol{v}_{j_*}\rangle + \frac{\langle \boldsymbol{a}, \boldsymbol{\xi}_{\boldsymbol{j}_*}\rangle - \langle \boldsymbol{a}, \boldsymbol{v}_{j_*}\rangle\langle \boldsymbol{v}_{j_*}, \boldsymbol{\xi}_{\boldsymbol{j}_*}\rangle}{\beta}\right) \\ &\quad + \mathrm{O}_{\mathbb{P}}\left(\frac{\log n}{\sqrt{n}}\left(\frac{\log n}{\sqrt{n}|\beta_1|}\right)^{k-1}+\frac{\log n}{\beta_1^2\sqrt{n}}+\frac{(\log n)^{3/2}}{|\beta_1|^{3/2}n^{3/4}}+\frac{1}{\beta_1^4}\right), \end{aligned} \tag{165}$$

where we used (54) that with high probability $|\langle \boldsymbol{a}, \boldsymbol{\xi}_{\boldsymbol{j}}\rangle|$, $|\langle \boldsymbol{a}, \boldsymbol{\xi}_s\rangle|$ for $1 \leqslant s \leqslant t$ are bounded by $\log n/\sqrt{n}$. This finishes the proof of (20). For $\widehat{\beta}$ in (21), we have that

$$\boldsymbol{X}[\boldsymbol{u}_t^{\otimes k}] = \frac{\boldsymbol{X}[\boldsymbol{y}_t^{\otimes k}]}{\|\boldsymbol{y}_t\|_2^k} = \frac{\langle \boldsymbol{y}_t, \boldsymbol{X}[\boldsymbol{y}_t^{\otimes (k-1)}]\rangle}{\|\boldsymbol{y}_t\|_2^k} = \frac{\langle \boldsymbol{y}_t, \boldsymbol{y}_{t+1}\rangle}{\|\boldsymbol{y}_t\|_2^k}. \tag{166}$$

Thanks to (153), (160) and (127), for t satisfies (146), with probability at least $1-\mathrm{O}(e^{-c(\log n)^2})$, we can write the first term on the righthand side of (82), we have

$$\boldsymbol{y}_{t+1} = \sum_j a_{(t+1)j}\boldsymbol{v}_j + \sum_{\boldsymbol{j}} b_{(t+1)\boldsymbol{j}}\boldsymbol{\xi}_{\boldsymbol{j}} + b_{(t+1)1}\boldsymbol{\xi}_1 + \cdots + b_{(t+1)t}\boldsymbol{\xi}_t + c_{t+1}\boldsymbol{\xi}_{t+1}, \tag{167}$$

where $|c_{t+1}| \lesssim |a_t|^{k-1}/\beta^2$,

$$\begin{aligned} a_{(t+1)j_*} &= \beta_{j_*} b_{(t+1)\boldsymbol{j}_*} = \beta_{j_*} a_{tj_*}^{k-1}\left(1+\mathrm{O}\left(\frac{\log n}{\sqrt{n}|\beta_1|}\right)\right), \\ |a_{(t+1)j}| &\lesssim \left(\frac{\log n}{\sqrt{n}|\beta_1|}\right)^{k-1} |a_{(t+1)j_*}|, \quad j \neq j_* \end{aligned} \tag{168}$$

and

$$\begin{aligned} b_{(t+1)\boldsymbol{j}_*} &= \frac{a_{(t+1)j_*}}{\beta_{j_*}}\left(1+\mathrm{O}\left(\frac{\log n}{\sqrt{n}|\beta_1|}\right)\right), \\ |b_{(t+1)\boldsymbol{j}}| &\lesssim \frac{\log n}{\sqrt{n}|\beta_1|^2}|a_{(t+1)j_*}|, \quad \boldsymbol{j} \neq \boldsymbol{j}_*, \\ |b_{(t+1)1}|, |b_{(t+1)2}| + \cdots + |b_{(t+1)t}| &\lesssim a_t^{k-1}\frac{(\log n)^{1/2}}{|\beta|^{1/2}n^{1/4}}. \end{aligned} \tag{169}$$

From the discussion above, combining with (147) and (148) with straightforward computation, we have

$$\langle \boldsymbol{y}_t, \boldsymbol{y}_{t+1}\rangle = \beta_{j_*}a_{tj_*}^k\left(1+\frac{1}{\beta_{j_*}^2}+\frac{(k+1)\langle \boldsymbol{\xi}_{\boldsymbol{j}_*}, \boldsymbol{v}_{j_*}\rangle}{\beta_{j_*}} + \mathrm{O}\left(\frac{\log n}{\sqrt{n}}\left(\frac{\log n}{\sqrt{n}|\beta_1|}\right)^{k-1} + \frac{\log n}{\beta_1^2\sqrt{n}} + \frac{(\log n)^{3/2}}{|\beta_1|^{3/2}n^{3/4}}\right)\right). \tag{170}$$

By plugging (164) and (170) into (166), we get

$$\boldsymbol{X}[\boldsymbol{u}_t^{\otimes k}] = \mathrm{sgn}(a_{tj_*}^k)\left(\beta_{j_*} + \langle \boldsymbol{\xi}_{\boldsymbol{j}_*}, \boldsymbol{v}_{j_*}\rangle - \frac{k/2-1}{\beta_{j_*}}\right) \tag{171}$$

$$+\mathrm{O}\left(\frac{\log n}{\sqrt{n}}\left(\frac{\log n}{\sqrt{n}|\beta_1|}\right)^{k-1} + \frac{\log n}{|\beta_1|\sqrt{n}} + \frac{(\log n)^{3/2}}{|\beta_1|^{1/2}n^{3/4}} + \frac{1}{|\beta_1|^3}\right) \tag{172}$$

Since by our assumption, in Case 1 we have that $\beta_{j^*} > 0$. Thanks to (160) $a_{t+1j_*} = \beta a_{tj_*}^{k-1}(1+\mathrm{o}(1))$, especially a_{t+1j_*} and a_{tj_*} are of the same sign. In the case $\langle \boldsymbol{u}, \boldsymbol{v}_{j_*}\rangle > 0$, we have $a_{1j_*} = \beta\langle \boldsymbol{u}, \boldsymbol{v}_{j_*}\rangle^{k-1} > 0$. We conclude that $a_{tj_*} > 0$. Therefore $\mathrm{sgn}(\boldsymbol{X}[\boldsymbol{u}_t^{\otimes k}]) = \mathrm{sgn}(a_{tj_*})^k = +$, and it follows that

$$\boldsymbol{X}[\boldsymbol{u}_t^{\otimes k}] = \beta_{j_*} + \langle \boldsymbol{\xi}_{\boldsymbol{j}_*}, \boldsymbol{v}_{j_*}\rangle - \frac{k/2-1}{\beta_{j_*}} \tag{173}$$

$$+\mathrm{O}\left(\frac{\log n}{\sqrt{n}}\left(\frac{\log n}{\sqrt{n}|\beta_1|}\right)^{k-1} + \frac{\log n}{|\beta_1|\sqrt{n}} + \frac{(\log n)^{3/2}}{|\beta_1|^{1/2}n^{3/4}} + \frac{1}{|\beta_1|^3}\right) \tag{174}$$

This finishes the proof of (21). The Cases 2, 3, 4, by simply changing $(\beta_{j_*}, \boldsymbol{v}_{j_*})$ in the righthand side of (20) and (21) to the corresponding limit.

□

4.4 Proof of Theorem 2.9

Proof of Theorem 2.9. We first prove (25). If $\boldsymbol{u}$ is uniformly distributed over the unit sphere, then it has the same law as $\boldsymbol{\eta}/\|\boldsymbol{\eta}\|_2$, where $\boldsymbol{\eta}$ is an n-dim standard Gaussian vector, with each entry $\mathcal{N}(0,1)$. With this notation

$$|\beta_j\langle \boldsymbol{u},\boldsymbol{v}_j\rangle^{k-2}| = |\beta_j\langle \boldsymbol{\eta},\boldsymbol{v}_j\rangle^{k-2}|/\|\boldsymbol{\eta}\|_2^{k-2}, \tag{175}$$

and we can rewrite $\mathbb{P}(i=\mathrm{argmax}_j\,|\beta_j\langle \boldsymbol{u},\boldsymbol{v}_j\rangle^{k-2}|)$ as

$$\mathbb{P}(i=\mathrm{argmax}_j\,|\beta_j\langle \boldsymbol{u},\boldsymbol{v}_j\rangle^{k-2}|) = \mathbb{P}(i=\mathrm{argmax}_j\,|\beta_j\langle \boldsymbol{\eta},\boldsymbol{v}_j\rangle^{k-2}|). \tag{176}$$

Since $\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_r$ are orthonormal vectors, $\langle \boldsymbol{v}_1,\boldsymbol{\eta}\rangle,\langle \boldsymbol{v}_2,\boldsymbol{\eta}\rangle,\cdots,\langle \boldsymbol{v}_r,\boldsymbol{\eta}\rangle$ are independent standard Gaussian random variables. Then we have

$$\begin{aligned} p_i &= \mathbb{P}(i=\mathrm{argmax}_j\,|\beta_j\langle \boldsymbol{\eta},\boldsymbol{v}_j\rangle^{k-2}|) = \mathbb{P}(|\beta_i/\beta_\ell|^{1/k-2}\langle \boldsymbol{\eta},\boldsymbol{v}_i\rangle| \geqslant |\langle \boldsymbol{\eta},\boldsymbol{v}_\ell\rangle|,\ \text{for all } i\neq\ell) \\ &= \int_0^\infty \sqrt{\frac{2}{\pi}}e^{-x^2/2}\left(\prod_{\ell\neq i}\int_0^{\left(\frac{|\beta_i|}{|\beta_\ell|}\right)^{\frac{1}{k-2}}x}\sqrt{\frac{2}{\pi}}e^{-y^2/2}\mathrm{d}y\right)\mathrm{d}x. \end{aligned} \tag{177}$$

This gives (25). Using the fact we can rewrite $\boldsymbol{u}$ as $\boldsymbol{\eta}/\|\boldsymbol{\eta}\|_2$, we have that with probability $1-\mathrm{O}(1/\sqrt{\kappa})$,

$$1/\sqrt{\kappa n} \leqslant |\langle \boldsymbol{u},\boldsymbol{v}_i\rangle| \leqslant \sqrt{\kappa/n}, \tag{178}$$

for all $1\leqslant i\leqslant r$. Thus Assumption (2.5) holds, and especially,

$$\max_j |\beta_j\langle \boldsymbol{u},\boldsymbol{v}_j\rangle^{k-2}| \geqslant |\beta_1\langle \boldsymbol{u},\boldsymbol{v}_1\rangle^{k-2}| \geqslant |\beta_1(1/\sqrt{\kappa n})^{k-2}| \gtrsim n^{\varepsilon}. \tag{179}$$

Theorem 2.9 then follows directly from Theorem 2.7. □

4.5 Proof of Corollaries 2.8, 2.10 and 2.11

Proof of Corollary 2.8. According to the definition of $\boldsymbol{\xi}$ in (20) of Theorem 2.7, i.e. $\boldsymbol{\xi}=\boldsymbol{Z}[\boldsymbol{v}_{j_*}^{\otimes(k-1)}]$, is an n-dim vector, with each entry i.i.d. $\mathcal{N}(0,1/n)$ Gaussian random variable. We see that

$$\langle \boldsymbol{\xi},\boldsymbol{v}\rangle \overset{d}{=} \mathcal{N}\left(0,1/n\right).$$

Especially with high probability we will have that $|\langle \boldsymbol{\xi},\boldsymbol{v}\rangle| \lesssim \log n/\sqrt{n}$. Then we conclude from (21), with high probability it holds

$$\widehat{\beta} = \beta_{j_*} + \mathrm{O}\left(\frac{1}{\beta_{j_*}} + \frac{\log n}{\sqrt{n}}\right). \tag{180}$$

With the bound (180), we can replace $\langle \boldsymbol{a},\boldsymbol{v}\rangle/(2\beta^2)$ on the righthand side of (20) by $\langle \boldsymbol{a},\boldsymbol{v}\rangle/(2\widehat{\beta}^2)$, which gives an error

$$\left|\frac{\langle \boldsymbol{a},\boldsymbol{v}\rangle}{2\beta_{j_*}^2} - \frac{\langle \boldsymbol{a},\boldsymbol{v}\rangle}{2\widehat{\beta}^2}\right| = \mathrm{O}\left(|\langle \boldsymbol{a},\boldsymbol{v}\rangle|\left(\frac{1}{|\beta_{j_*}|^4}+\frac{\log n}{|\beta_{j_*}|^3\sqrt{n}}\right)\right). \tag{181}$$

Combining the above discussion together, we can rewrite (20) as

$$\langle \boldsymbol{a}, \widehat{\boldsymbol{v}} \rangle - \left(1 - \frac{1}{2\widehat{\beta}^2}\right) \langle \boldsymbol{a}, \boldsymbol{v} \rangle = \frac{\langle \boldsymbol{a}, \boldsymbol{\xi} \rangle - \langle \boldsymbol{a}, \boldsymbol{v}_{j_*} \rangle \langle \boldsymbol{v}_{j_*}, \boldsymbol{\xi} \rangle}{\beta_{j_*}} + \mathrm{O}_{\mathbb{P}}\left(\frac{\log n}{\sqrt{n}} \left(\frac{\log n}{\sqrt{n}|\beta_1|} \right)^{k-1} + \frac{\log n}{|\beta_1|^2 \sqrt{n}} + \frac{(\log n)^{3/2}}{|\beta_1|^{3/2} n^{3/4}} + \frac{1}{|\beta_1|^4} \right), \tag{182}$$

with high probability, where we used that $|\beta_{j_*}| \gtrsim |\beta_1|$.

Again thanks to the definition of $\boldsymbol{\xi}$ in (20) of Theorem 2.1, i.e. $\boldsymbol{\xi} = \boldsymbol{Z}[\boldsymbol{v}^{\otimes(k-1)}]$, is an n-dim vector, with each entry i.i.d. $\mathcal{N}(0, 1/n)$ Gaussian random variable, we see that

$$\langle \boldsymbol{a}, \boldsymbol{\xi} \rangle - \langle \boldsymbol{a}, \boldsymbol{v}_{j_*} \rangle \langle \boldsymbol{v}_{j_*}, \boldsymbol{\xi} \rangle = \langle \boldsymbol{a} - \langle \boldsymbol{a}, \boldsymbol{v}_{j_*} \rangle \boldsymbol{v}_{j_*}, \boldsymbol{\xi} \rangle, \tag{183}$$

is a Gaussian random variable, with mean zero and variance

$$\mathbb{E}[\langle \boldsymbol{a} - \langle \boldsymbol{a}, \boldsymbol{v}_{j_*} \rangle \boldsymbol{v}_{j_*}, \boldsymbol{\xi} \rangle^2] = \frac{1}{n} \| \boldsymbol{a} - \langle \boldsymbol{a}, \boldsymbol{v}_{j_*} \rangle \boldsymbol{v}_{j_*} \|_2^2 = \frac{1}{n} \langle \boldsymbol{a}, (\boldsymbol{I}_n - \boldsymbol{v}_{j_*} \boldsymbol{v}_{j_*}^\top) \boldsymbol{a} \rangle \tag{184}$$

$$= \frac{1 + \mathrm{o}(1)}{n} \langle \boldsymbol{a}, (\boldsymbol{I}_n - \widehat{\boldsymbol{v}}_{j_*} \widehat{\boldsymbol{v}}_{j_*}^\top) \boldsymbol{a} \rangle. \tag{185}$$

This together with (180), (182) as well as our assumption (22)

$$\frac{\sqrt{n}\widehat{\beta}}{\sqrt{\langle \boldsymbol{a}, (\boldsymbol{I}_n - \widehat{\boldsymbol{v}}\widehat{\boldsymbol{v}}^\top) \boldsymbol{a} \rangle}} \left[(1 - \frac{1}{2\widehat{\beta}^2})^{-1} \langle \boldsymbol{a}, \widehat{\boldsymbol{v}} \rangle - \langle \boldsymbol{a}, \boldsymbol{v}_{j_*} \rangle \right] \xrightarrow{d} \mathcal{N}(0, 1). \tag{186}$$

Under the same assumption, we have similar results for Cases 2, 3, 4, by simply changing $(\beta_{j_*}, \boldsymbol{v}_{j_*})$ in the righthand side of (7) and (8) to the corresponding expression. □

Proof of Corollary 2.10. For $k \geqslant 3$ and $|\beta_1| \geqslant n^{(k-2)/2+\varepsilon}$, the assumption 22 holds trivially. The claim (29) follows from (24). For (30), we recall that in (28), $\boldsymbol{\xi} = \boldsymbol{Z}[\boldsymbol{v}_i^{\otimes(k-1)}]$, is an n-dim vector, with each entry i.i.d. $\mathcal{N}(0, 1/n)$ Gaussian random variable. We see that

$$\langle \boldsymbol{\xi}, \boldsymbol{v}_i \rangle \overset{d}{=} \mathcal{N}(0, 1/n).$$

Especially with high probability we will have that $|\langle \boldsymbol{\xi}, \boldsymbol{v} \rangle| \lesssim \log n / \sqrt{n}$. Then we conclude from (28), with high probability it holds

$$\widehat{\beta} = \beta_i + \mathrm{O}\left(\frac{1}{\beta_i} + \frac{\log n}{\sqrt{n}} \right). \tag{187}$$

With the bound (187), we can replace $(k/2 - 1)/\beta_i$ on the righthand side of (28) by $(k/2 - 1)/\widehat{\beta}$, which gives an error

$$\left| \frac{k/2 - 1}{\beta_i} - \frac{k/2 - 1}{\widehat{\beta}} \right| = \mathrm{O}\left(\frac{1}{|\beta_1|^2} + \frac{\log n}{|\beta_1| \sqrt{n}} \right), \tag{188}$$

where we used that $|\beta_i| \gtrsim |\beta_1|$. Combining the above discussion together, we can rewrite (28) as

$$\beta_i = \widehat{\beta} + \frac{k/2 - 1}{\widehat{\beta}} - \langle \boldsymbol{\xi}, \boldsymbol{v}_i \rangle + \mathrm{O}_{\mathbb{P}}\left(\frac{\log n}{\sqrt{n}} \left(\frac{\log n}{\sqrt{n}|\beta_1|} \right)^{k-1} + \frac{\log n}{|\beta_1| \sqrt{n}} + \frac{(\log n)^{3/2}}{|\beta_1|^{1/2} n^{3/4}} + \frac{1}{|\beta_1|^2} \right). \tag{189}$$

Since $\langle \boldsymbol{\xi}, \boldsymbol{v}_i \rangle \stackrel{d}{=} \mathcal{N}(0, 1/n)$, and the error term in (189) is much smaller than $1/\sqrt{n}$. We conclude from (189)

$$\sqrt{n}\left(\beta_i - \widehat{\beta} + \frac{k/2-1}{\widehat{\beta}}\right) \stackrel{d}{\to} \mathcal{N}(0,1). \tag{190}$$

This finishes the proof of (30).

□

Proof of Corollary 2.11. Given the significance level α, the asymptotic confidence intervals in Corollary 2.11 can be calculated from Corollary 2.10 by bounding the absolute values of the left hand sides of (29) and (30) at z_α. □

References

[1] E. Abbe, J. Fan, K. Wang, Y. Zhong, et al. Entrywise eigenvector analysis of random matrices with low expected rank. *Annals of Statistics*, 48(3):1452–1474, 2020.

[2] A. Anandkumar, R. Ge, D. Hsu, and S. M. Kakade. A tensor approach to learning mixed membership community models. *The Journal of Machine Learning Research*, 15(1):2239–2312, 2014.

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

[4] G. B. Arous, R. Gheissari, A. Jagannath, et al. Algorithmic thresholds for tensor pca. *Annals of Probability*, 48(4):2052–2087, 2020.

[5] Z. Bai and J. Yao. On sample eigenvalues in a generalized spiked population model. *Journal of Multivariate Analysis*, 106:167–177, 2012.

[6] J. Baik, G. B. Arous, S. Péché, et al. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, 33(5):1643–1697, 2005.

[7] J. Baik and J. W. Silverstein. Eigenvalues of large sample covariance matrices of spiked population models. *Journal of multivariate analysis*, 97(6):1382–1408, 2006.

[8] F. Benaych-Georges and R. R. Nadakuditi. The singular values and vectors of low rank perturbations of large rectangular random matrices. *Journal of Multivariate Analysis*, 111:120–135, 2012.

[9] A. Birnbaum, I. M. Johnstone, B. Nadler, and D. Paul. Minimax bounds for sparse pca with noisy high-dimensional data. *Annals of statistics*, 41(3):1055, 2013.

[10] T. Cai, Z. Ma, and Y. Wu. Optimal estimation and rank detection for sparse spiked covariance matrices. *Probability theory and related fields*, 161(3-4):781–815, 2015.

[11] T. T. Cai, Z. Ma, Y. Wu, et al. Sparse pca: Optimal rates and adaptive estimation. *The Annals of Statistics*, 41(6):3074–3110, 2013.

[12] W.-K. Chen et al. Phase transition in the spiked random tensor with rademacher prior. *The Annals of Statistics*, 47(5):2734–2756, 2019.

[13] W.-K. Chen, M. Handschy, and G. Lerman. Phase transition in random tensors with multiple spikes. *arXiv preprint arXiv:1809.06790*, 2018.

[14] Y. Chen, C. Cheng, and J. Fan. Asymmetry helps: Eigenvalue and eigenvector analyses of asymmetrically perturbed low-rank matrices. *arXiv preprint arXiv:1811.12804*, 2018.

[15] C. Cheng, Y. Wei, and Y. Chen. Inference for linear forms of eigenvectors under minimal eigenvalue separation: Asymmetry and heteroscedasticity. *arXiv preprint arXiv:2001.04620*, 2020.

[16] A. Cichocki, D. Mandic, L. De Lathauwer, G. Zhou, Q. Zhao, C. Caiafa, and H. A. Phan. Tensor decompositions for signal processing applications: From two-way to multiway component analysis. *IEEE signal processing magazine*, 32(2):145–163, 2015.

[17] P. Comon. Tensors: a brief introduction. *IEEE Signal Processing Magazine*, 31(3):44–53, 2014.

[18] D. L. Donoho, M. Gavish, and I. M. Johnstone. Optimal shrinkage of eigenvalues in the spiked covariance model. *Annals of statistics*, 46(4):1742, 2018.

[19] O. Duchenne, F. Bach, I.-S. Kweon, and J. Ponce. A tensor-based algorithm for high-order graph matching. *IEEE transactions on pattern analysis and machine intelligence*, 33(12):2383–2395, 2011.

[20] N. El Karoui et al. Spectrum estimation for large dimensional covariance matrices using random matrix theory. *The Annals of Statistics*, 36(6):2757–2790, 2008.

[21] E. Frolov and I. Oseledets. Tensor methods and recommender systems. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 7(3):e1201, 2017.

[22] W. Hackbusch. *Tensor spaces and numerical tensor calculus*, volume 42. Springer, 2012.

[23] C. J. Hillar and L.-H. Lim. Most tensor problems are np-hard. *Journal of the ACM (JACM)*, 60(6):1–39, 2013.

[24] S. B. Hopkins, T. Schramm, J. Shi, and D. Steurer. Fast spectral algorithms from sum-of-squares proofs: tensor decomposition and planted sparse vectors. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 178–191, 2016.

[25] S. B. Hopkins, J. Shi, and D. Steurer. Tensor principal component analysis via sum-of-square proofs. In *Conference on Learning Theory*, pages 956–1006, 2015.

[26] D. Hsu and S. M. Kakade. Learning mixtures of spherical gaussians: moment methods and spectral decompositions. In *Proceedings of the 4th conference on Innovations in Theoretical Computer Science*, pages 11–20, 2013.

[27] A. Jagannath, P. Lopatto, and L. Miolane. Statistical thresholds for tensor pca. *arXiv preprint arXiv:1812.03403*, 2018.

[28] I. M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *Annals of statistics*, pages 295–327, 2001.

[29] I. M. Johnstone and A. Y. Lu. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486):682–693, 2009.

[30] I. M. Johnstone and D. Paul. PCA in high dimensions: An orientation. *Proceedings of the IEEE*, 106(8):1277–1292, 2018.

[31] A. Karatzoglou, X. Amatriain, L. Baltrunas, and N. Oliver. Multiverse recommendation: n-dimensional tensor factorization for context-aware collaborative filtering. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 79–86, 2010.

[32] C. Kim, A. S. Bandeira, and M. X. Goemans. Community detection in hypergraphs, spiked tensor models, and sum-of-squares. In *2017 International Conference on Sampling Theory and Applications (SampTA)*, pages 124–128. IEEE, 2017.

[33] T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009.

[34] O. Ledoit, M. Wolf, et al. Nonlinear shrinkage estimation of large-dimensional covariance matrices. *The Annals of Statistics*, 40(2):1024–1060, 2012.

[35] T. Lesieur, L. Miolane, M. Lelarge, F. Krzakala, and L. Zdeborová. Statistical and computational phase transitions in spiked tensor estimation. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 511–515. IEEE, 2017.

[36] Y. Luo, G. Raskutti, M. Yuan, and A. R. Zhang. A sharp blockwise tensor perturbation bound for orthogonal iteration. *arXiv preprint arXiv:2008.02437*, 2020.

[37] Y. Luo and A. R. Zhang. Open problem: Average-case hardness of hypergraphic planted clique detection. In *Conference on Learning Theory*, pages 3852–3856. PMLR, 2020.

[38] Z. Ma et al. Sparse principal component analysis and iterative thresholding. *The Annals of Statistics*, 41(2):772–801, 2013.

[39] S. O'Rourke, V. Vu, and K. Wang. Random perturbation of low rank matrices: Improving classical bounds. *Linear Algebra and its Applications*, 540:26–59, 2018.

[40] D. Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, pages 1617–1642, 2007.

[41] S. Péché. The largest eigenvalue of small rank perturbations of hermitian random matrices. *Probability Theory and Related Fields*, 134(1):127–173, 2006.

[42] A. Perry, A. S. Wein, A. S. Bandeira, et al. Statistical limits of spiked tensor models. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 56, pages 230–264. Institut Henri Poincaré, 2020.

[43] S. Rendle and L. Schmidt-Thieme. Pairwise interaction tensor factorization for personalized tag recommendation. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 81–90, 2010.

[44] E. Richard and A. Montanari. A statistical model for tensor pca. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2897–2905. Curran Associates, Inc., 2014.

[45] N. D. Sidiropoulos, L. De Lathauwer, X. Fu, K. Huang, E. E. Papalexakis, and C. Faloutsos. Tensor decomposition for signal processing and machine learning. *IEEE Transactions on Signal Processing*, 65(13):3551–3582, 2017.

[46] E. Simony, C. J. Honey, J. Chen, O. Lositsky, Y. Yeshurun, A. Wiesel, and U. Hasson. Dynamic reconfiguration of the default mode network during narrative comprehension. *Nature communications*, 7:12141, 2016.

[47] V. Vu. Singular vectors under random perturbation. *Random Structures & Algorithms*, 39(4):526–538, 2011.

[48] V. Q. Vu, J. Lei, et al. Minimax sparse principal subspace estimation in high dimensions. *The Annals of Statistics*, 41(6):2905–2947, 2013.

[49] A. Zhang, T. T. Cai, and Y. Wu. Heteroskedastic pca: Algorithm, optimality, and applications. *arXiv preprint arXiv:1810.08316*, 2018.

[50] A. Zhang and D. Xia. Tensor svd: Statistical and computational limits. *IEEE Transactions on Information Theory*, 64(11):7311–7338, 2018.

[51] Y. Zhong. Eigenvector under random perturbation: A nonasymptotic rayleigh-schr\"{o} dinger theory. *arXiv preprint arXiv:1702.00139*, 2017.

[52] H. Zhou, L. Li, and H. Zhu. Tensor regression with applications in neuroimaging data analysis. *Journal of the American Statistical Association*, 108(502):540–552, 2013.