Journal of Machine Learning Research 25 (2024) 1-42 Submitted 1/24; Revised 5/24; Published 6/24

Sharp Analysis of Power Iteration for Tensor PCA

Yuchen Wu WUYC14@WHARTON.UPENN.EDU
Department of Statistics and Data Science
University of Pennsylvania
Philadelphia, PA 19104-6303, USA

Kangjie Zhou KANGJIE@STANFORD.EDU
Department of Statistics
Stanford University
Stanford, CA 94305-2004, USA

Editor: Daniel Hsu

Abstract

We investigate the power iteration algorithm for the tensor PCA model introduced in Richard and Montanari (2014). Previous work studying the properties of tensor power iteration is either limited to a constant number of iterations, or requires a non-trivial data-independent initialization. In this paper, we move beyond these limitations and analyze the dynamics of randomly initialized tensor power iteration up to polynomially many steps. Our contributions are threefold: First, we establish sharp bounds on the number of iterations required for power method to converge to the planted signal, for a broad range of the signal-to-noise ratios. Second, our analysis reveals that the actual algorithmic threshold for power iteration is smaller than the one conjectured in the literature by a polylog(n) factor, where n is the ambient dimension. Finally, we propose a simple and effective stopping criterion for power iteration, which provably outputs a solution that is highly correlated with the true signal. Extensive numerical experiments verify our theoretical results.

Keywords: Spiked model, tensor PCA, power iteration, approximate message passing, non-convex optimization

1. Introduction

Tensors are multi-dimensional arrays that have found wide applications across various domains, including neuroscience (Wozniak et al., 2007; Zhou et al., 2013), recommendation systems (Rendle and Schmidt-Thieme, 2010; Shah and Yu, 2019), image processing (Liu et al., 2012; Sidiropoulos et al., 2017), community detection (Nickel et al., 2011; Jing et al., 2021), and genomics (Hore et al., 2016; Wang et al., 2019). In these applications, oftentimes the tensor exhibits a *low-rank* structure, meaning that the data admits the form of a low-dimensional signal corrupted by random noise. Efficient recovery of this intrinsic low-rank signal not only facilitates various important machine learning tasks, e.g., clustering (Zhou et al., 2019; Luo and Zhang, 2022; Zhou and Chen, 2023), but also spurs the development of important methodologies in the field of scientific computing (Khoromskij and Schwab, 2011; Grasedyck et al., 2013).

In this paper, we study the problem of recovering a low-rank tensor from noisy observations of its entries. Such a problem is also known as Tensor Principal Component Analysis

(Tensor PCA) in the literature. To set the stage, we consider the single-spike model proposed by Richard and Montanari (Richard and Montanari, 2014). Under this model, we observe a k-th order rank-one tensor corrupted by random noise:

$$\mathbf{T} = \lambda_n \boldsymbol{v}^{\otimes k} + \boldsymbol{W}. \tag{1}$$

Here, $\lambda_n > 0$ is the signal-to-noise ratio that depends on the ambient dimension n, $\boldsymbol{v} \in \mathbb{S}^{n-1}$ is a planted signal that lies on the n-dimensional unit sphere, and $\boldsymbol{W} \in (\mathbb{R}^n)^{\otimes k}$ stands for the random noise that has i.i.d. standard Gaussian entries and is independent of the signal. We denote by $k \in \mathbb{N}_+$ the order of the tensor. The special case $k = 2$ has been well studied by statisticians, where model (1) reduces to the spiked Wigner model (Johnstone, 2001). In particular, detection and estimation problems under the spiked matrix model have been extensively investigated under various contexts (Johnstone, 2001; Baik et al., 2005; Benaych-Georges and Nadakuditi, 2012; Lelarge and Miolane, 2019; Montanari and Wu, 2022), and numerous computationally efficient algorithms have been proposed to recover the signal (Journée et al., 2010; Ma, 2013; Deshpande et al., 2016; Montanari and Venkataramanan, 2021). The main focus of the present paper will be tensors of order $k \geq 3$.

Comparing to its matrix counterpart, tensor problems with order $k \geq 3$ are, in many scenarios, much more challenging. For instance, the spectral decomposition of a matrix or matrix PCA can be performed efficiently using polynomial-time algorithms, while tensor decomposition and tensor PCA are known to be NP-hard (Hillar and Lim, 2013). Despite the NP-hardness of tensor PCA, researchers have designed scalable algorithms that are capable of recovering the planted signal, if one additionally assumes that the data are random and follow natural distributional assumptions. This random data perspective not only simplifies analysis in many scenarios, but also offers valuable insights on the computational complexity and estimation accuracy of the proposed methodologies from an average-case point of view. Exemplary algorithms that come with average-case theoretical guarantees include iterative algorithms (Ma, 2013; Anandkumar et al., 2017a; Zhang and Xia, 2018; Arous et al., 2020; Han et al., 2022; Huang et al., 2022), sum-of-squares (SOS) algorithms (Hopkins et al., 2015, 2016; Potechin and Steurer, 2017; Kim et al., 2017), and spectral algorithms (Montanari and Sun, 2018; Xia and Yuan, 2019; Cai et al., 2019).

Another striking feature of the tensor PCA model (1) is that it exhibits the so-called *computational-to-statistical gap*, meaning that there exists regime of signal-to-noise ratios within which it is information theoretically possible to recover the planted signal, while no polynomial-time algorithms are known (Berthet and Rigollet, 2013; Arous et al., 2019; Gamarnik, 2021; Dudeja and Hsu, 2021). A sequence of works have established that the information-theoretic threshold of tensor PCA (1) is of order $\Theta(\sqrt{n})$ (Richard and Montanari, 2014; Lesieur et al., 2017; Chen, 2019). On the other hand, the algorithmic threshold—the minimal signal-to-noise ratio above which recovery is efficiently achievable—is conjectured to be $\Theta(n^{k/4})$ (Richard and Montanari, 2014).[1] This critical threshold, in fact, has been achieved by various algorithms which originate from different ideas (Hopkins et al., 2015; Anandkumar et al., 2017a; Biroli et al., 2020).

In spite of the strong theoretical guarantees achieved by strategically crafted algorithms, in practice, it is often preferable to resort to simple iterative algorithms. Among them,

1. Under the scaling of Richard and Montanari (2014), these two thresholds are $\Theta(1)$ (constant order) and $\Theta(n^{(k-2)/4})$, respectively.

tensor power iteration has been extensively applied to solving a large number of tensor problems, e.g., tensor decomposition and tensor PCA (Kolda and Mayo, 2011; Richard and Montanari, 2014; Wang and Anandkumar, 2016; Anandkumar et al., 2017b; Huang et al., 2022; Wu and Zhou, 2023). For tensor PCA, doing power iteration is equivalent to running projected gradient ascent on a non-convex polynomial objective function with infinite step size. Towards understanding the dynamics of this algorithm, in Richard and Montanari (2014) the authors proved that tensor power iteration with random initialization converges rapidly to the true signal provided that $\lambda_n \gtrsim n^{k/2}$. They also employed a heuristic argument to suggest that the necessary and sufficient condition for convergence is actually $\lambda_n \gtrsim n^{(k-1)/2}$.[2] Later, a more refined analysis was carried out by Huang et al. (2022). The authors showed that tensor power iteration with a constant number of iterates succeeds when $\lambda_n \gtrsim n^{(k-1)/2+\varepsilon}$, and fails when $\lambda_n \lesssim n^{(k-1)/2-\varepsilon}$ for an arbitrarily small positive constant ε, thus partially confirming the $n^{(k-1)/2}$ threshold conjectured by Richard and Montanari (2014). However, their analysis is restricted to a fixed number of power iterates, and therefore fails to capture the dynamics of tensor power iteration when the number of iterations grows with the input dimension. Further, they did not characterize the asymptotic behavior of tensor power iteration when $\lambda_n \asymp n^{(k-1)/2}$. Therefore, a complete picture is still lacking. As a side note, past works have also considered tensor power iteration with a warm start depending on some extra side information (Richard and Montanari, 2014; Huang et al., 2022). However, how to obtain such initialization in practice remains elusive. In this paper, we will only consider tensor power iteration with random initialization independent of the data.

1.1 Our contribution

This paper is devoted to establishing a more comprehensive picture of tensor power iteration starting from a random initialization. Our contributions are summarized below.

Algorithmic threshold. First, we give a partial answer to the open problem in Richard and Montanari (2014). To be concrete, our results imply that tensor power iteration with a random initialization provably converges to the planted signal in polynomially many steps, requiring only $\lambda_n \gtrsim n^{(k-1)/2}(\log n)^{-C}$ for some positive constant C that depends only on k. Recall that the conjectured threshold in Richard and Montanari (2014) is $\Theta(n^{(k-1)/2})$, our conclusion actually shows that the true phase transition for power iteration occurs at a slightly lower signal-to-noise ratio than the one conjectured in Richard and Montanari (2014). In order to establish such a result, we introduce the concept of *alignment* to measure the correlation between the iterates obtained from power iteration and the planted signal, and show that the evolution of this alignment can be well approximated by a low-dimensional *polynomial recurrence process*. We then conduct a precise analysis on the dynamics of this process to establish the convergence of power iteration for tensor PCA.

Number of iterations required for convergence. Second, we present a sharp characterization of the number of iterations required for convergence, for λ_n ranging from $\ll n^{(k-1)/2}$ to $\gg n^{(k-1)/2}$. To be precise, when $\gamma_n := n^{-(k-1)/2}\lambda_n \in [c, n^{o(1)}]$ for $c > 0$ an arbitrary constant that does not depend on n, we show that $(1+o_n(1)) \log_{k-1}(\log_{k-1} n / \log_{k-1} \gamma_n)$

2. Again, under their scaling, these two conditions should be $\lambda_n \gtrsim n^{(k-1)/2}$ and $\lambda_n \gtrsim n^{(k-2)/2}$, respectively.

iterations are both necessary and sufficient for convergence to occur. In a different weak signal regime where $(\log n)^{-(k-2)/2} \ll \gamma_n \ll 1$, we establish upper and lower bounds on the number of iterations that are of the same order of magnitude on a logarithmic scale (see Theorem 2.1 for a formal statement). Our analysis is similar to that of Huang et al. (2022). However, the technical treatment in Huang et al. (2022) is only able to analyze the dynamics of tensor power iteration up to a constant number of steps, which is not sufficient if the initialization is random. To overcome this difficulty, in this work we develop a novel Gaussian conditioning lemma (Lemma 3.1), which allows us to accurately analyze the dynamics of tensor power iteration up to polynomially many steps. To the best of our knowledge, this is the first result that studies the dynamics of tensor power iteration beyond a constant number of iterations under the setting of model (1).

Stopping criterion. We also propose a stopping criterion that allows us to decide when to terminate the iteration in practice. Our proposal is simple, effective, and comes with rigorous theoretical guarantee. To summarize, the proposed stopping criterion finds an iterate that with high probability correlates well with the hidden spike. Besides, if we implement the proposed stopping rule, then the actual number of power iteration we implement matches well with the upper and lower bounds we have established, emphasizing both accuracy and efficiency of our proposal.

Gaussian conditioning beyond constant steps. The tool that we employ to establish the above results is based on the Gaussian conditioning technique, which has been widely applied to analyze the Approximate Message Passing (AMP) algorithm (Bayati and Montanari, 2011) as well as many other iterative algorithms. Prior art along this line of research mostly studies only a constant number of iterations. Encouragingly, recent years have witnessed significant progress towards generalizing such Gaussian conditioning type analysis to accommodate settings that allow the number of iterations to grow simultaneously with the input dimension (Rush and Venkataramanan, 2018; Li and Wei, 2022; Li et al., 2023; Wu and Zhou, 2023). Our work contributes to this active field of research by establishing the first result of this kind under the tensor PCA model (1). From a technical perspective, we believe our results not only push forward the development of AMP theory, but also enrich the toolbox to analyze general iterative algorithms.

Future directions. One interesting future direction is to generalize our convergence analysis to sub-Gaussian tensors, which requires developing non-asymptotic AMP-type analysis for sub-Gaussian random ensembles beyond a constant number of iterations, and is highly non-trivial. Progress in this direction is made only recently by Jones and Pesenti (2024), and their main results are based on some complicated combinatorial arguments.

Another fascinating future direction would be to extend our main results to the multi-rank case, namely, the observed tensor $\mathbf{T}$ is a rank-r tensor perturbed by random Gaussian noise for some $r > 1$. Under this setting, we believe that the same Gaussian conditioning technique can be employed to analyze the (properly defined) alignment between the power iterates and the planted rank-r signal, and we leave that for future work.

1.2 Organization

The rest of this paper is organized as follows. Section 2 formulates the framework and gives our main result. We present the proof of the main theorem in Section 3, while deferring the proofs of several auxiliary lemmas to appendices. In Section 4 we report numerical experiments that support our theorems.

1.3 Notation

Throughout the proof, with a slight abuse of notation, we use letters c, C to represent various constants (which can only depend on the tensor order k), whose values might not necessarily be the same in each occurrence. For a matrix $\mathbf{S} \in \mathbb{R}^{d\times m}$, we denote by $\Pi_{\mathbf{S}} \in \mathbb{R}^{d\times d}$ the projection matrix onto the column space of $\mathbf{S}$, and let $\Pi_{\mathbf{S}}^{\perp} := \boldsymbol{I}_d - \Pi_{\mathbf{S}}$. For $n \in \mathbb{N}_+$, we define $[n] = \{1, 2, \cdots, n\}$. For two sequences of positive numbers $\{a_n\}_{n\in\mathbb{N}_+}$ and $\{b_n\}_{n\in\mathbb{N}_+}$, we say $a_n \gtrsim b_n$ if there exists a positive constant c, such that $a_n \geq c\, b_n$, we say $a_n = o_n(b_n)$ if $a_n/b_n \to 0$ as $n \to \infty$, and we say $a_n \gg b_n$ if $a_n/b_n \to \infty$ as $n \to \infty$.

2. Main results

We summarize in this section our main results. We first give a formal definition of tensor power iteration from a random initialization. Then, we present our main theorem in which we determine the regime of convergence and characterize the number of iterations required. We also give a stopping criterion that determines when to terminate the power iteration.

2.1 Tensor power iteration

We denote by $\boldsymbol{v}^0 = \tilde{\boldsymbol{v}}^0 \sim \mathrm{Unif}(\mathbb{S}^{n-1})$ the random initialization that is independent of $\mathbf{T}$. Tensor power iteration initialized at $\boldsymbol{v}^0$ is defined recursively as follows:

$$\begin{aligned} \boldsymbol{v}^{t+1} &= \mathbf{T}[(\tilde{\boldsymbol{v}}^t)^{\otimes(k-1)}] = \lambda_n \langle \boldsymbol{v}, \tilde{\boldsymbol{v}}^t\rangle^{k-1}\boldsymbol{v} + \boldsymbol{W}[(\tilde{\boldsymbol{v}}^t)^{\otimes(k-1)}], \\ \tilde{\boldsymbol{v}}^{t+1} &= \frac{\boldsymbol{v}^{t+1}}{\|\boldsymbol{v}^{t+1}\|_2}, \end{aligned} \tag{2}$$

where $\boldsymbol{W}[(\tilde{\boldsymbol{v}}^t)^{\otimes(k-1)}]$ is an n-dimensional vector whose i-th entry is $\langle \boldsymbol{W}, \boldsymbol{e}_i \otimes (\tilde{\boldsymbol{v}}^t)^{\otimes(k-1)}\rangle$. Here, $\boldsymbol{e}_i$ is an n-dimensional vector that has the i-th entry being one and all the rest being zero.

As a side remark, iteration (2) can be regarded as projected gradient ascent with infinite step size for the following constrained optimization problem:

$$\text{maximize } \langle \mathbf{T}, \boldsymbol{\sigma}^{\otimes k}\rangle, \quad \text{subject to } \boldsymbol{\sigma} \in \mathbb{S}^{n-1}.$$

2.2 Convergence analysis

Next, we study the number of iterations required for algorithm (2) to converge. To this end, we first define the convergence criterion. For any fixed positive constant δ, let

$$T_\delta^{\mathrm{conv}} := \min\left\{t \in \mathbb{N}_+ : |\langle \tilde{\boldsymbol{v}}^t, \boldsymbol{v}\rangle| \geq 1 - \delta\right\}. \tag{3}$$

Our main result provides upper and lower bounds on T_δ^{conv}, in a signal-to-noise ratio regime when we simultaneously have $\gamma_n \gg (\log n)^{-(k-2)/2}$ and $\gamma_n = n^{o(1)}$.

Theorem 2.1 *Recall that $\gamma_n = n^{-(k-1)/2}\lambda_n$. Assume $\gamma_n \gg (\log n)^{-(k-2)/2}$ and $\gamma_n = n^{o(1)}$. Then for any fixed $\delta, \eta > 0$, with probability $1 - o_n(1)$ we have*

$$
\begin{aligned}
T_\delta^{\mathrm{conv}} &\geq \max\left\{\exp\left(\frac{1-\eta}{2}\left(\frac{C_k}{\gamma_n}\right)^{2/(k-2)}\right),\ (1-\eta)\log_{k-1}\frac{\log_{k-1} n}{\max\{\log_{k-1}\gamma_n, 1\}}\right\}, \\
T_\delta^{\mathrm{conv}} &\leq \exp\left(\frac{1+\eta}{2}\left(\frac{1}{\gamma_n}\right)^{2/(k-2)}\right) + (1+\eta)\log_{k-1}\frac{\log_{k-1} n}{\max\{\log_{k-1}\gamma_n, 1\}},
\end{aligned}
\tag{4}
$$

where $C_k = (k-2)^{k-2}/(k-1)^{k-1}$.

Remark 2.1 *When $\gamma_n \gtrsim 1$, Theorem 2.1 implies that*

$$
\frac{T_\delta^{\mathrm{conv}}}{\log_{k-1}\frac{\log_{k-1} n}{\max\{\log_{k-1}\gamma_n, 1\}}} \xrightarrow{\mathbb{P}} 1, \tag{5}
$$

which gives a sharp characeterization of the number of steps required for convergence. On the other hand, as γ_n drops below the constant level, T_δ^{conv} grows drastically, but is still polynomial in n provided that $\gamma_n \gg (\log n)^{-(k-2)/2}$. In addition, based on the upper and lower bounds presented in the theorem, we conjecture that the time complexity of tensor power iteration is super-polynomial when $\gamma_n \ll (\log n)^{-(k-2)/2}$. However, due to the use of the Gaussian conditioning scheme (Lemma 3.1), our current approach can only accurately analyze the dynamics of tensor power iteration up to polynomially many steps. Therefore, justifying this conjecture would require the development of new theoretical tools, which we leave for future work. Further, our Theorem 2.1 assumes $\gamma_n = n^{o(1)}$, since Huang et al. (2022) already proved that a constant number of power iterations is sufficient to recover the true signal when $\gamma_n \gtrsim n^c$, for any constant $c > 0$ that does not depend on n.

We present the proof of Theorem 2.1 in Section 3, with proofs of auxiliary lemmas and propositions deferred to the appendices.

2.3 Stopping criterion

Theorem 2.1 gives lower and upper bounds on the number of iterations required for convergence. However, the theorem falls short of providing practical guidance regarding when should we terminate the power iteration, as we do not assume we know any prior information about the signal-to-noise ratio λ_n, or equivalently γ_n.

To tackle this issue, we propose in this section a simple while effective stopping criterion that with high probability finds an iterate that aligns well with the hidden spike. In addition, we give upper and lower bounds on the actual number of power iterations we implement if we follow the proposed stopping criterion, which matches that introduced in Theorem 2.1.

To give a high level description, we propose to terminate the algorithm if we find any two consecutive iterates being moderately correlated with each other. To be precise, we define

$$
T_{\mathsf{stop}} := \inf\left\{t \in \mathbb{N}_+ : \left|\langle \tilde{\boldsymbol{v}}^{t-2}, \tilde{\boldsymbol{v}}^{t-3}\rangle\right| \geq 1/2\right\}. \tag{6}
$$

We shall output $\tilde{\boldsymbol{v}}^{T_{\mathsf{stop}}}$ as an estimate of $\boldsymbol{v}$. We give theoretical guarantee for our approach in the theorem below, which will be proved in Appendix F.

Theorem 2.2 *We assume the conditions of Theorem 2.1. Then, for any positive constant δ, with probability $1 - o_n(1)$ we have $|\langle \tilde{\boldsymbol{v}}^{T_{\mathsf{stop}}}, \boldsymbol{v}\rangle| \geq 1 - \delta$. In addition, with high probability we still get the upper and lower bounds as indicated in Eq. (4) if we replace T_δ^{conv} with T_{stop}.*

3. Proof of Theorem 2.1

We present in this section the proof of Theorem 2.1. The idea is to track the alignment between the iterates obtained from tensor power iteration and the planted signal. Equipped with the Gaussian conditioning technique, we are able to control the difference between this alignment and a scalar polynomial recurrence process that we define below, and prove that they are with high probability close to each other. This allows us to simplify our analysis by resorting to a reduction, and the remaining convergence analysis is conducted directly on this polynomial recurrence process.

3.1 Reduction to the polynomial recurrence process

In this section we define the alignment with the true signal and show that it can be captured by a polynomial recurrence process.

For $t \in \mathbb{N}_+$, we define

$$\alpha_t := \lambda_n \langle \boldsymbol{v}, \tilde{\boldsymbol{v}}^{t-1}\rangle^{k-1} = \gamma_n \left(\sqrt{n}\langle \boldsymbol{v}, \tilde{\boldsymbol{v}}^{t-1}\rangle\right)^{k-1}.$$

The magnitude of α_t measures the level of alignment between the obtained iterates and the hidden spike. We also define $\alpha_0 = 0$ for convenience. In the first iteration, the initial alignment is of the same order as γ_n, since by taking a random initialization roughly speaking we have $|\langle \boldsymbol{v}, \tilde{\boldsymbol{v}}^0\rangle| \asymp n^{-1/2}$. Throughout the paper, α_t will be the key quantity that characterizes the evolution of iteration (2).

As we have mentioned, the main goal of this section is to establish that $\{\alpha_t\}_{t\in\mathbb{N}_+}$ can be closely tracked by a one-dimensional discrete Markov process $\{X_t\}_{t\in\mathbb{N}}$, given by the following recurrence equation:

$$X_0 = 0, \quad \text{and } X_{t+1} = \gamma_n (X_t + Z_t)^{k-1} \text{ for } t \geq 0, \tag{7}$$

where $\{Z_t\}_{t\in\mathbb{N}_+}$ is a sequence of i.i.d. standard Gaussian random variables. To this end, we first develop the recurrence equation for the alignment sequence $\{\alpha_t\}_{t\in\mathbb{N}_+}$. Our derivation is based on the Gaussian conditioning technique, which has been widely applied to study the AMP algorithm (Bayati and Montanari, 2011).

Decomposing tensor power iterates

Next, we give a useful decomposition of the tensor power iterates. By definition of tensor power iteration, we have

$$\begin{aligned}\boldsymbol{v}^{t+1} &= \lambda_n \langle \boldsymbol{v}, \tilde{\boldsymbol{v}}^t\rangle^{k-1}\boldsymbol{v} + \boldsymbol{W}\left[(\tilde{\boldsymbol{v}}^t)^{\otimes(k-1)}\right] \\ &= \alpha_{t+1}\boldsymbol{v} + \boldsymbol{W}\left[(\tilde{\boldsymbol{v}}^t)^{\otimes(k-1)}\right],\end{aligned}$$

where we recall that $\alpha_{t+1} = \lambda_n \langle \boldsymbol{v}, \tilde{\boldsymbol{v}}^t\rangle^{k-1}$.

Before proceeding, we shall first introduce several concepts that are useful for our analysis. For $t \in \mathbb{N}$, we let $\boldsymbol{V}_t \in \mathbb{R}^{n\times(t+1)}$ be a matrix whose i-th column is $\boldsymbol{v}^{i-1}$. Based on the column space of $\boldsymbol{V}_t$, we can decompose the vector $\boldsymbol{v}^t$ as $\boldsymbol{v}^t = \boldsymbol{v}^t_\perp + \boldsymbol{v}^t_\parallel$, where $\boldsymbol{v}^t_\perp := \Pi^\perp_{\boldsymbol{V}_{t-1}}\boldsymbol{v}^t$ and $\boldsymbol{v}^t_\parallel := \Pi_{\boldsymbol{V}_{t-1}}\boldsymbol{v}^t$. Analogously, we set $\tilde{\boldsymbol{v}}^t_\perp = \boldsymbol{v}^t_\perp/\|\boldsymbol{v}^t\|_2$ and $\tilde{\boldsymbol{v}}^t_\parallel = \boldsymbol{v}^t_\parallel/\|\boldsymbol{v}^t\|_2$ as their normalized versions. When the original vector is an all-zero one, we simply set its normalized version to be itself.

We immediately see that the vectors $\{\tilde{\boldsymbol{v}}^i_\perp/\|\tilde{\boldsymbol{v}}^i_\perp\|_2 : 0 \le i \le t\}$ form an orthonormal basis of the linear space spanned by $\{\tilde{\boldsymbol{v}}^i : 0 \le i \le t\}$. As a consequence, $\tilde{\boldsymbol{v}}^t$ admits the following decomposition:

$$\tilde{\boldsymbol{v}}^t = \sum_{i=0}^{t} \frac{\langle \tilde{\boldsymbol{v}}^t, \tilde{\boldsymbol{v}}^i_\perp\rangle}{\|\tilde{\boldsymbol{v}}^i_\perp\|_2} \cdot \frac{\tilde{\boldsymbol{v}}^i_\perp}{\|\tilde{\boldsymbol{v}}^i_\perp\|_2}.$$

For $(i_1, \cdots, i_{k-1}) \in \{0, 1, \cdots, t\}^{k-1}$, we define

$$\begin{aligned}
\beta^{(t)}_{i_1,i_2,\cdots,i_{k-1}} &:= \prod_{j=1}^{k-1} \frac{\langle \tilde{\boldsymbol{v}}^{i_j}_\perp, \tilde{\boldsymbol{v}}^t\rangle}{\|\tilde{\boldsymbol{v}}^{i_j}_\perp\|_2} \in \mathbb{R}, \\
\boldsymbol{w}_{i_1,i_2,\cdots,i_{k-1}} &:= \prod_{j=1}^{k-1} \|\tilde{\boldsymbol{v}}^{i_j}_\perp\|_2^{-1} \boldsymbol{W}\left[\tilde{\boldsymbol{v}}^{i_1}_\perp \otimes \tilde{\boldsymbol{v}}^{i_2}_\perp \otimes \cdots \otimes \tilde{\boldsymbol{v}}^{i_{k-1}}_\perp\right] \in \mathbb{R}^n.
\end{aligned}$$

As will become clear soon, with probability 1 over the randomness of the data generation process, it holds that $\|\tilde{\boldsymbol{v}}^t_\perp\|_2 \neq 0$ for all $t = O(n^{1/(2k-2)})$. Therefore, $\boldsymbol{w}_{i_1,i_2,\cdots,i_{k-1}}$ and $\beta^{(t)}_{i_1,i_2,\cdots,i_{k-1}}$ are almost surely well-defined. With these definitions, we see that $\boldsymbol{v}^{t+1}$ can be decomposed as the sum of the following terms:

$$\begin{aligned}
\boldsymbol{v}^{t+1} &= \alpha_{t+1}\boldsymbol{v} + \boldsymbol{W}\left[(\tilde{\boldsymbol{v}}^t)^{\otimes(k-1)}\right] \\
&= \alpha_{t+1}\boldsymbol{v} + \sum_{(i_1,\cdots,i_{k-1})\in\boldsymbol{H}_t} \beta^{(t)}_{i_1,i_2,\cdots,i_{k-1}} \boldsymbol{w}_{i_1,i_2,\cdots,i_{k-1}} \\
&= \alpha_{t+1}\boldsymbol{v} + \sum_{(i_1,\cdots,i_{k-1})\in\boldsymbol{H}_{t-1}} \beta^{(t)}_{i_1,i_2,\cdots,i_{k-1}} \boldsymbol{w}_{i_1,i_2,\cdots,i_{k-1}} + \sqrt{1 - \|\tilde{\boldsymbol{v}}^t_\parallel\|_2^{2k-2}}\,\mathbf{g}_{t+1},
\end{aligned} \tag{8}$$

where $\boldsymbol{H}_t = \{0, 1, \cdots, t\}^{k-1}$, and

$$\mathbf{g}_{t+1} = \frac{1}{\sqrt{1 - \|\tilde{\boldsymbol{v}}^t_\parallel\|_2^{2k-2}}} \sum_{(i_1,i_2,\cdots,i_{k-1})\in\boldsymbol{H}_t\setminus\boldsymbol{H}_{t-1}} \beta^{(t)}_{i_1,i_2,\cdots,i_{k-1}} \boldsymbol{w}_{i_1,i_2,\cdots,i_{k-1}}. \tag{9}$$

Here, we make the convention that $\boldsymbol{H}_{-1} = \emptyset$. In what follows, we will characterize the joint distribution of the vectors $\boldsymbol{w}_{i_1,i_2,\cdots,i_{k-1}}$ for all $(i_1, i_2, \cdots, i_{k-1}) \in \boldsymbol{H}_t$ via a Gaussian conditioning lemma, and derive the relationship between α_{t+1} and α_t.

The Gaussian conditioning lemma

As an important ingredient of our conditioning analysis, we introduce the sigma-algebra $\mathcal{F}_t$, which roughly speaking, is generated by the vectors associated with the first t iterations. To be precise, we define

$$\mathcal{F}_t := \sigma\left(\left\{\boldsymbol{w}_{i_1,i_2,\cdots,i_{k-1}} : (i_1, i_2, \cdots, i_{k-1}) \in \boldsymbol{H}_{t-1}\right\} \cup \left\{\mathbf{g}, \boldsymbol{v}^0, \boldsymbol{v}^1, \cdots, \boldsymbol{v}^t, \boldsymbol{v}\right\}\right). \tag{10}$$

Notice that $\mathcal{F}_t \subseteq \mathcal{F}_{t+1}$. With these notations, we state our Gaussian conditioning lemma as follows.

Lemma 3.1 *For all $t < n$ and all $(i_1, i_2, \cdots, i_{k-1}) \in \boldsymbol{H}_t \backslash \boldsymbol{H}_{t-1}$, it holds that $\boldsymbol{w}_{i_1,i_2,\cdots,i_{k-1}} \perp \mathcal{F}_t$. Furthermore, $\boldsymbol{w}_{i_1,i_2,\cdots,i_{k-1}} \overset{\text{i.i.d.}}{\sim} \mathsf{N}(\mathbf{0}, \boldsymbol{I}_n)$.*

The proof of Lemma 3.1 is deferred to Appendix B. With the aid of this lemma, we know that $\mathbf{g}_{t+1}$ is independent of the $\boldsymbol{w}_{i_1,i_2,\cdots,i_{k-1}}$'s for $(i_1, \cdots, i_{k-1}) \in \boldsymbol{H}_{t-1}$. Further, since

$$
\begin{aligned}
\sum_{(i_1,i_2,\cdots,i_{k-1}) \in \boldsymbol{H}_t \backslash \boldsymbol{H}_{t-1}} \left(\beta^{(t)}_{i_1,i_2,\cdots,i_{k-1}}\right)^2 &= 1 - \sum_{(i_1,i_2,\cdots,i_{k-1}) \in \boldsymbol{H}_{t-1}} \left(\beta^{(t)}_{i_1,i_2,\cdots,i_{k-1}}\right)^2 && (11) \\
&= 1 - \prod_{j=1}^{k-1} \sum_{i_j=0}^{t-1} \frac{\langle \tilde{\boldsymbol{v}}^{i_j}_{\perp}, \tilde{\boldsymbol{v}}^t \rangle^2}{\|\tilde{\boldsymbol{v}}^{i_j}_{\perp}\|_2^2} && (12) \\
&= 1 - \left(\sum_{i=0}^{t-1} \frac{\langle \tilde{\boldsymbol{v}}^{i}_{\perp}, \tilde{\boldsymbol{v}}^t \rangle^2}{\|\tilde{\boldsymbol{v}}^{i}_{\perp}\|_2^2}\right)^{k-1} && (13) \\
&= 1 - \|\tilde{\boldsymbol{v}}^t_{\|}\|_2^{2k-2}, && (14)
\end{aligned}
$$

it follows that $\mathbf{g}_{t+1} \sim \mathsf{N}(\mathbf{0}, \boldsymbol{I}_n)$.

Recurrence equation for the alignment

With the aid of Lemma 3.1 and decomposition (8), we are ready to establish the recurrence equation for the alignment. For notational convenience, we let

$$
\boldsymbol{h}_{t+1} := \sum_{(i_1,\cdots,i_{k-1}) \in \boldsymbol{H}_{t-1}} \beta^{(t)}_{i_1,i_2,\cdots,i_{k-1}} \boldsymbol{w}_{i_1,i_2,\cdots,i_{k-1}},
$$

where we make the convention that $\boldsymbol{h}_0 = \mathbf{0}$. It then follows that

$$
\begin{aligned}
\boldsymbol{v}^{t+1} &= \alpha_{t+1}\boldsymbol{v} + \boldsymbol{h}_{t+1} + \sqrt{1 - \|\tilde{\boldsymbol{v}}^t_{\|}\|_2^{2k-2}}\,\mathbf{g}_{t+1}, \\
\Longrightarrow \langle \boldsymbol{v}^{t+1}, \boldsymbol{v} \rangle &= \alpha_{t+1} + \langle \boldsymbol{h}_{t+1}, \boldsymbol{v} \rangle + \sqrt{1 - \|\tilde{\boldsymbol{v}}^t_{\|}\|_2^{2k-2}}\ \langle \mathbf{g}_{t+1}, \boldsymbol{v} \rangle,
\end{aligned}
$$

which further implies

$$
\begin{aligned}
\alpha_{t+2} &= \lambda_n \langle \tilde{\boldsymbol{v}}^{t+1}, \boldsymbol{v} \rangle^{k-1} = \gamma_n \cdot \left(\frac{\sqrt{n}}{\|\boldsymbol{v}^{t+1}\|_2}\right)^{k-1} \cdot \langle \boldsymbol{v}^{t+1}, \boldsymbol{v} \rangle^{k-1} \\
&= \gamma_n \left(\frac{\sqrt{n}}{\|\boldsymbol{v}^{t+1}\|_2}\right)^{k-1} \left(\alpha_{t+1} + \langle \boldsymbol{h}_{t+1}, \boldsymbol{v} \rangle + \sqrt{1 - \|\tilde{\boldsymbol{v}}^t_{\|}\|_2^{2k-2}} \cdot \langle \mathbf{g}_{t+1}, \boldsymbol{v} \rangle\right)^{k-1}.
\end{aligned}
$$

Decrementing the index by one, we get the following recurrence equation for the sequence $\{\alpha_t\}_{t \in \mathbb{N}}$:

$$
\alpha_{t+1} = \gamma_n \left(\frac{\sqrt{n}}{\|\boldsymbol{v}^t\|_2}\right)^{k-1} \left(\alpha_t + \langle \boldsymbol{h}_t, \boldsymbol{v} \rangle + \sqrt{1 - \|\tilde{\boldsymbol{v}}^{t-1}_{\|}\|_2^{2k-2}} \cdot \langle \mathbf{g}_t, \boldsymbol{v} \rangle\right)^{k-1}. \tag{15}
$$

The remaining parts of this section will be devoted to the analysis of $\{\alpha_t\}_{t \in \mathbb{N}}$ based on the above equation.

Controlling the error terms

We then show that the recurrence equation (15) can be viewed as a perturbed version of the polynomial process (7) with small errors. We start with defining some key quantities in Eq. (15):

$$\zeta_t := \left(\frac{\sqrt{n}}{\|\boldsymbol{v}^t\|_2} \right)^{k-1}, \quad b_t := \langle \boldsymbol{h}_t, \boldsymbol{v} \rangle, \quad c_t := \sqrt{1 - \|\tilde{\boldsymbol{v}}_{\parallel}^{t-1}\|_2^{2k-2}}. \tag{16}$$

In the above display, we make the convention that $\tilde{\boldsymbol{v}}_{\parallel}^{-1} = \mathbf{0}$. At initialization, since $\zeta_0 = 1$, $b_0 = 0$, and $c_0 = 1$, we know that the first iteration of Eq. (15) is equivalent to

$$\alpha_1 = \gamma_n \zeta_0 \left(\alpha_0 + b_0 + c_0 Z_0\right)^{k-1} = \gamma_n \left(\alpha_0 + Z_0\right)^{k-1},$$

where $Z_0 = \langle \mathbf{g}_0, \boldsymbol{v} \rangle \sim \mathsf{N}(0,1)$ is by Lemma 3.1 (recall that $\mathbf{g}_0$ is defined in Eq. (9)). Similarly, by the law of large numbers, we know that $\zeta_1 = 1 + o_{n,\mathbb{P}}(1)$, $b_1 = o_{n,\mathbb{P}}(1)$ and $c_1 = 1 + o_{n,\mathbb{P}}(1)$, hence the next iteration has the following approximation:

$$\alpha_2 = \gamma_n \zeta_1 \left(\alpha_1 + b_1 + c_1 Z_1\right)^{k-1} \approx \gamma_n \left(\alpha_1 + Z_1\right)^{k-1},$$

where $Z_1 = \langle \mathbf{g}_1, \boldsymbol{v} \rangle \sim \mathsf{N}(0,1)$ is independent of α_1. Indeed, we will show that the above approximation is valid up to polynomially many steps along the power iteration path until the alignment α_t reaches a certain threshold. To be precise, we establish the following lemma:

Lemma 3.2 *For any fixed $\varepsilon \in (1/4, 1/2)$, define the stopping time*

$$T_\varepsilon := \min\left\{t \in \mathbb{N}_+ : |\alpha_t| \geq n^\varepsilon\right\}. \tag{17}$$

Then, there exists an absolute constant $C > 0$, such that with probability no less than $1 - \exp(-C\sqrt{n})$, the following happens: For all $t < \min(T_\varepsilon, n^{1/2(k-1)})$,

$$\zeta_t \in [1 - n^{-1/6}, 1 + n^{-1/6}], \ |b_t| \leq C n^{1/4 + (k-1)(\varepsilon - 1/2)}, \ |c_t - 1| \leq C n^{2(k-1)(\varepsilon - 1/2)}. \tag{18}$$

We defer the proof of Lemma 3.2 to Appendix C. As a direct corollary of Lemma 3.2, we immediately obtain the following proposition:

Proposition 3.1 *Under the same setting as in Lemma 3.2, and let $\varepsilon = \varepsilon_k = (6k - 11)/12(k-1)$, which satisfies $\varepsilon_k \in (1/4, 1/2)$ for all $k \geq 3$. Then, we have*

$$\alpha_{t+1} = \gamma_n \zeta_t (\alpha_t + b_t + c_t Z_t)^{k-1}, \ \alpha_0 = 0. \tag{19}$$

where $Z_t \sim \mathsf{N}(0,1)$ is independent of $(\zeta_t, \alpha_t, b_t, c_t)$. Further, with probability at least $1 - \exp(-C\sqrt{n})$, the following happens: For all $t < \min(T_\varepsilon, n^{1/2(k-1)})$,

$$\zeta_t \in [1 - n^{-1/6}, 1 + n^{-1/6}], \ |b_t| \leq C n^{-1/6}, \ |c_t - 1| \leq C n^{-5/6}. \tag{20}$$

The above proposition establishes that up to $\min(T_\varepsilon, n^{1/2(k-1)})$ steps, the iteration of the alignment is closely tracked by that of the one-dimensional stochastic process defined in Eq. (7). In what follows, we show that the convergence of power iteration for tensor PCA can be precisely characterized by the stopping time T_ε. Before proceeding, we establish high probability upper and lower bounds on T_ε, detailed by the following lemma.

Lemma 3.3 *Under the assumptions of Theorem 2.1, and let* $\varepsilon = \varepsilon_k = (6k-11)/12(k-1)$ *as in the statement of Proposition 3.1. Then, for any sufficiently large* $n \in \mathbb{N}$ *and any* $\eta \in (0,1)$*, with probability* $1 - o_n(1)$ *one has*

$$T_\varepsilon \geq \max\left\{ \exp\left(\frac{1-\eta}{2} \left(\frac{C_k}{\gamma_n} \right)^{2/(k-2)} \right),\ (1-\eta) \log_{k-1} \frac{\log_{k-1} n}{\max\{\log_{k-1} \gamma_n, 1\}} \right\}, \tag{21}$$

$$T_\varepsilon \leq \exp\left(\frac{1+\eta}{2} \left(\frac{1}{\gamma_n} \right)^{2/(k-2)} \right) + (1+\eta) \log_{k-1} \frac{\log_{k-1} n}{\max\{\log_{k-1} \gamma_n, 1\}}, \tag{22}$$

where $C_k = (k-2)^{k-2}/(k-1)^{k-1}$.

The proof of Lemma 3.3 is based on Proposition 3.1. For the compactness of presentation, we delay the proof of Lemma 3.3 to Appendix D.

3.2 Convergence of tensor power iteration

Recall that T_δ^{conv} is defined in Eq. (3) and T_ε is defined in Eq. (17). For fixed positive constants δ and $\varepsilon \in (1/4, 1/2)$, we see that for n large enough we have $T_\delta^{\mathrm{conv}} \geq T_\varepsilon$. In this section, we also show that with high probability $T_\delta^{\mathrm{conv}} \leq T_\varepsilon + 1$. Putting together these results, we conclude that if we can establish bounds on T_ε, then this automatically gives bounds on T_δ^{conv} as well.

Now let $t = T_\varepsilon$. Naively we have $t - 1 < T_\varepsilon$ and $|\alpha_t| \geq n^\varepsilon$. According to the power iteration equation, we obtain that

$$\boldsymbol{v}^t = \alpha_t \boldsymbol{v} + \boldsymbol{h}_t + \sqrt{1 - \|\tilde{\boldsymbol{v}}_{\|}^{t-1}\|_2^{2k-2}}\, \mathbf{g}_t.$$

Invoking Lemma 3.3, we see that for a large enough n, with probability $1 - o_n(1)$ it holds that $T_\varepsilon \leq n^{1/2(k-1)}$. In this case we have $t - 1 < \min(T_\varepsilon, n^{1/2(k-1)})$. Re-examining the proof of concentration of c_t in the proof of Lemma 3.2, we find that (note $\varepsilon > 1/4$)

$$\mathbb{P}\left(\|\tilde{\boldsymbol{v}}_{\|}^{t-1}\|_2 \leq C n^{\varepsilon - 1/2} \right) \geq 1 - \exp(-C n^{2\varepsilon}) \geq 1 - \exp(-C\sqrt{n}), \tag{23}$$

where C is a positive constant, and consequently

$$\|\boldsymbol{h}_t\|_2 \leq 2\sqrt{n} \cdot \|\tilde{\boldsymbol{v}}_{\|}^{t-1}\|_2^{k-1} \leq C n^{1/12}$$

if we choose $\varepsilon = \varepsilon_k = (6k-11)/12(k-1)$ as per Proposition 3.1. Note that

$$\begin{aligned} \left\|\boldsymbol{v}^t\right\|_2 &= \left\| \alpha_t \boldsymbol{v} + \boldsymbol{h}_t + \sqrt{1 - \|\tilde{\boldsymbol{v}}_{\|}^{t-1}\|_2^{2k-2}}\, \mathbf{g}_t \right\|_2 \\ &\leq |\alpha_t| + \|\mathbf{g}_t\|_2 + C n^{1/12} \leq |\alpha_t| + C\sqrt{n} \end{aligned}$$

with probability at least $1 - \exp(-C\sqrt{n})$. Therefore,

$$|\langle \tilde{\boldsymbol{v}}^t, \boldsymbol{v} \rangle| = \frac{|\langle \boldsymbol{v}^t, \boldsymbol{v} \rangle|}{\|\boldsymbol{v}^t\|_2} = \frac{|\alpha_t + b_t + c_t Z_t|}{\|\boldsymbol{v}^t\|_2} \geq \frac{|\alpha_t| - |b_t| - |c_t Z_t|}{|\alpha_t| + C\sqrt{n}}. \tag{24}$$

Again according to the proof of Lemma 3.2 and the choice of ε in Proposition 3.1, we know that $|b_t| \leq Cn^{-1/6}$ and $|c_t| \leq 1 + Cn^{-5/6}$ with probability at least $1 - \exp(C\sqrt{n})$. Further since $Z_t \sim \mathsf{N}(0, 1)$, it finally follows that with high probability $|\langle \tilde{\boldsymbol{v}}^t, \boldsymbol{v} \rangle| \geq Cn^{\varepsilon - 1/2}$, which leads to

$$|\alpha_{t+1}| = \gamma_n \left|\sqrt{n}\langle \boldsymbol{v}, \tilde{\boldsymbol{v}}^t \rangle\right|^{k-1} > Cn^{1/2+1/24} \tag{25}$$

for sufficiently large n, provided that $\gamma_n \gg (\log n)^{-(k-2)/2}$. Consider the next iteration:

$$\boldsymbol{v}^{t+1} = \alpha_{t+1}\boldsymbol{v} + \boldsymbol{h}_{t+1} + \sqrt{1 - \|\tilde{\boldsymbol{v}}_{\|}^t\|_2^{2k-2}}\mathbf{g}_{t+1}.$$

Using standard concentration arguments, we know that

$$\left\|\boldsymbol{h}_{t+1} + \sqrt{1 - \|\tilde{\boldsymbol{v}}_{\|}^t\|_2^{2k-2}}\mathbf{g}_{t+1}\right\|_2 \leq Cn^{1/2} \leq Cn^{-1/24}|\alpha_{t+1}|$$

with probability at least $1 - \exp(-Cn)$, which immediately implies that

$$|\langle \tilde{\boldsymbol{v}}^{t+1}, \boldsymbol{v} \rangle| \geq 1 - Cn^{-1/24}. \tag{26}$$

Therefore, $T_\delta^{\text{conv}} \leq t + 1 = T_\varepsilon + 1$ with probability at least $1 - \exp(-C\sqrt{n})$. We summarize the main result of this section in the following lemma:

Lemma 3.4 *Assume $\gamma_n \gg (\log n)^{-(k-2)/2}$ and $\gamma_n = n^{o(1)}$. Then, with probability at least $1 - \exp(-C\sqrt{n})$, we have*

$$|\langle \tilde{\boldsymbol{v}}^{T_\varepsilon+1}, \boldsymbol{v} \rangle| \geq 1 - Cn^{-1/24}. \tag{27}$$

Namely, tensor power iteration converges in one step after α_t reaches the level n^ε.

Combining the conclusions of Lemma 3.3 and Lemma 3.4 completes the proof of Theorem 2.1.

4. Numerical experiments

We present in this section simulations that support our theories. For the simplicity of presentation, in the main text we only present experiments for several representative settings. We refer interested readers to Appendix G for simulation outcomes under more settings.

4.1 Comparing alignment and the polynomial recurrence process

As demonstrated in Section 3, a key ingredient of our proof is to connect the tensor alignments $\{\alpha_t\}_{t\geq 0}$ with the polynomial recurrence process $\{X_t\}_{t\geq 0}$ defined in Eq. (7). Theoretical result that suggests their closeness has already been established in Proposition 3.1. We complement to this result in this section by providing empirical evidence.

To set the stage, we choose $n = 200$, $k = 3$, $\lambda_n = n^{(k-1)/2}$, and generate the tensor data according to Eq. (1). We then run tensor power iteration with random initialization and compare the marginal distributions of α_t and X_t, for all $t \in \{1, 2, 3, 4\}$. We repeat this procedure 1000 times independently, and collect the realized values of α_t to form the corresponding empirical distributions. We also simulate the polynomial recurrence process $\{X_t\}_{t\geq 0}$ and obtain 1000 independent samples. We display the simulation outcomes in Figure 1, which suggests that the marginal distributions already match well with a moderately large n.

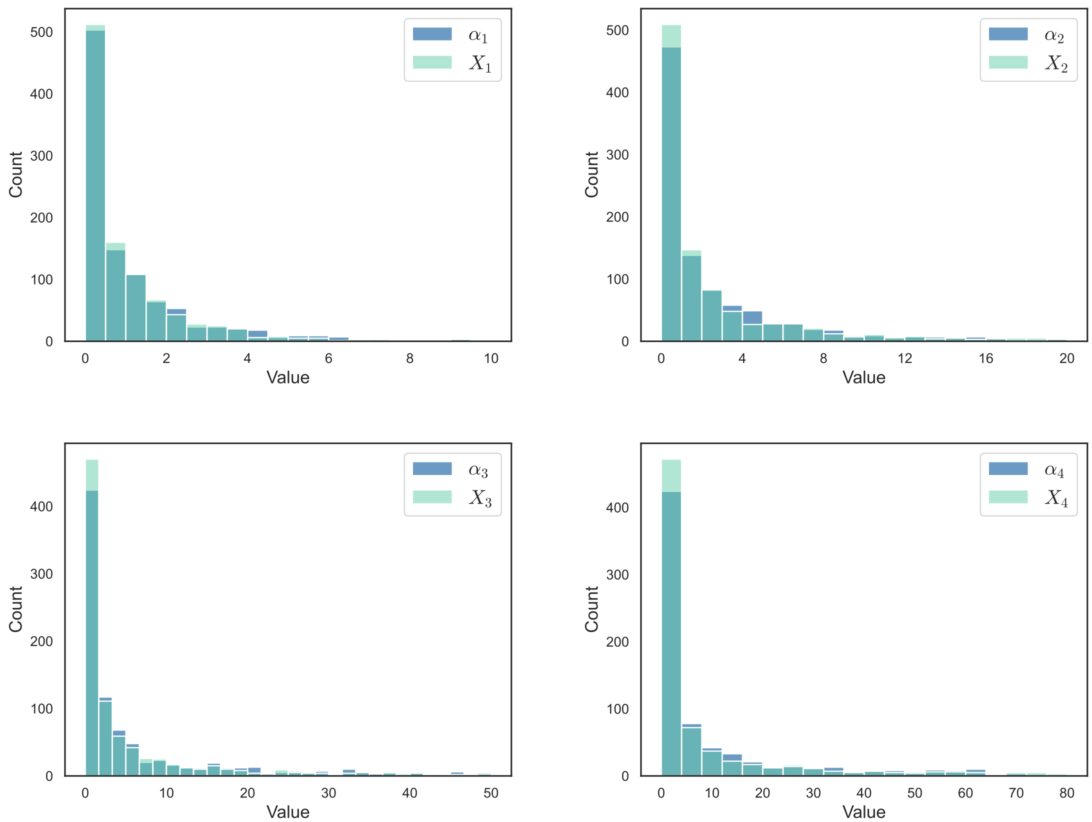

Figure 1: Comparison of the marginal distributions between α_t and X_t, for $t \in \{1, 2, 3, 4\}$. Here, we set $n = 200$, $k = 3$, $\lambda_n = n^{(k-1)/2}$, and run tensor power iteration from random initialization on independent datasets for 1000 times. Note that in this figure, the histograms for α_t and X_t overlap a lot with each other (their overlapping regions are indicated by the third color), meaning that the marginal distributions of α_t and X_t are indeed very close.

4.2 Evolution of correlation

Theorem 2.1 implies that as long as $\lambda_n \gtrsim 1$, tensor power iteration with random initialization will converge to the planted spike within $O(\log \log n)$ iterations. In this experiment, we provide numerical evidence that supports this claim. Throughout the experiment, we set $\lambda_n = n^{(k-1)/2}$. In Figure 2, we plot the evolution of correlation $|\langle \tilde{\boldsymbol{v}}^t, \boldsymbol{v} \rangle|$ as a function of the number of iterations t. From the figure, we see that the correlation rapidly increases from 0 to 1 as t increases. Furthermore, the number of iterations required for convergence is nearly independent of the input dimension, suggesting the correctness of our claim.

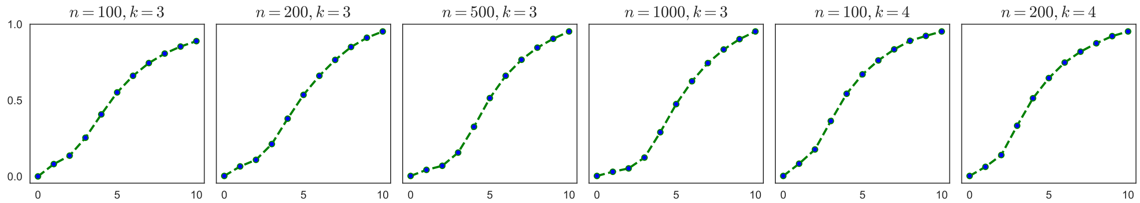

Figure 2: Evolution of correlation $|\langle \tilde{\boldsymbol{v}}^t, \boldsymbol{v} \rangle|$ as a function of the number of iterations t. Here, the x axis represents the number of iterations ranging from 0 to 10, and the y axis gives the level of correlation. We repeat the experiment independently for 1000 times for every combination of (n, k), and compute the average correlation.

4.3 Convergence probability

Next, we investigate the probability of tensor power iteration with a random initialization converging to the planted spike. For this part we let $\lambda_n = n^{(k-1)/2}$, $k = 3$, and use different values of n. For each tensor realization, we run tensor power iteration from a random initialization for a sufficiently large number of iterations and check the convergence. For each $n \in \{25, 50, 100, 200, 400, 800\}$, we repeat this procedure independently for 1000 times and compute the empirical convergence probability. Here, we say an iterate $\tilde{\boldsymbol{v}}^t$ converges to the true spike if and only if $|\langle \tilde{\boldsymbol{v}}^t, \boldsymbol{v} \rangle| > 0.99$. We plot such empirical probabilities in Figure 3. Inspecting the figure, we see that the γ-threshold above which power iteration with a random start achieves near probability one convergence decreases and approaches 0 as $n \to \infty$, once again suggesting the correctness of our main theorem.

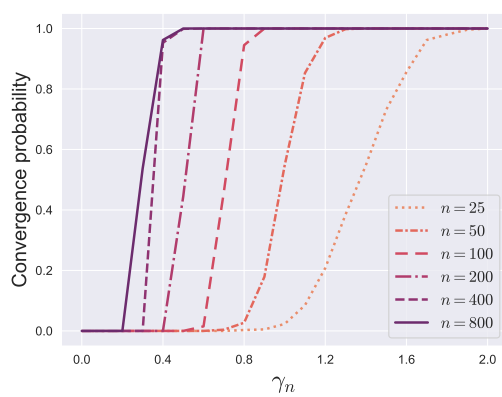

Figure 3: Probability of tensor power iteration with random initialization converging to the hidden spike. The x axis stands for γ_n and the y axis gives the empirical convergence probability averaged over 1000 independent experiments.

4.4 Stopping rule

Finally, we evaluate the performance of the stopping rule proposed in Section 2.3. In this experiment, we fix $\lambda_n = n^{(k-1)/2}$ and consider different combinations of (n, k). For each configuration, we independently generate five tensors following model (1), and implement tensor power iteration from a random initialization. We then compute the correlation between the true spike and the iterates measured by $|\langle \tilde{\boldsymbol{v}}^t, \boldsymbol{v} \rangle|$, and plot the evolution of this correlation over the first 100 iterations. We also calculate T_{stop} using Eq. (6). We present the simulation results in Figure 4. From the figure, we see that the proposed stopping rule effectively terminates the algorithm at an early stage while simultaneously maintains high estimation accuracy.

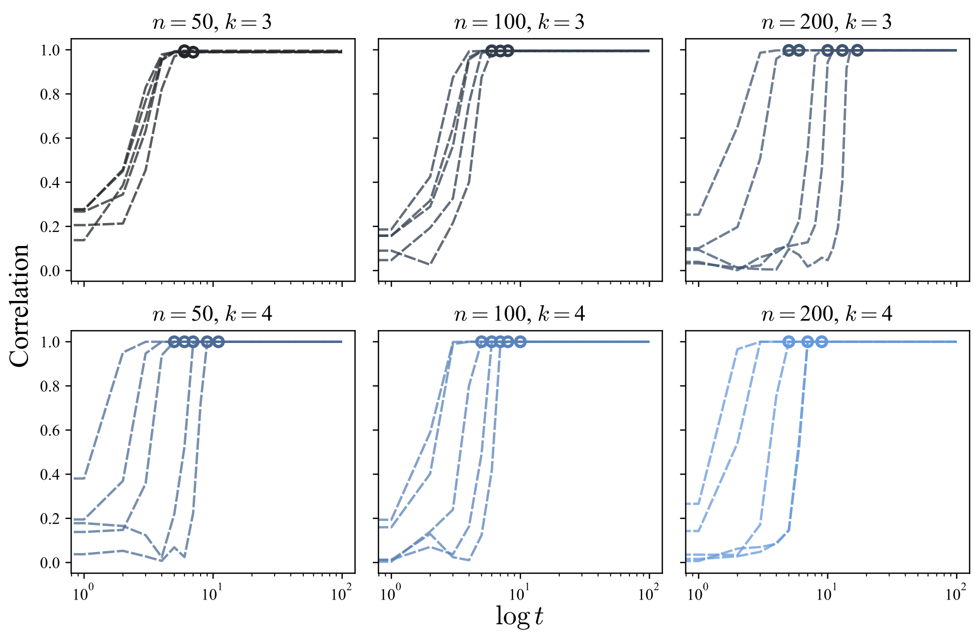

Figure 4: Illustration of the effectiveness of the stopping rule. The x axis here is the logarithmic of the number of iterations, and the y axis shows the correlation. We independently repeat the experiment 5 times for each setting, and record the correlation along the power iteration trajectory. Here, T_{stop} is computed using the power iteration iterates and is marked with a circle in the figure.

In Appendix G.2, we alter the value of the stopping threshold (which is 1/2 in the current experiment) and still observe good performance, suggesting the proposed stopping rule is not sensitive to the choice of the stopping threshold.

Acknowledgments

The authors would like to thank Tselil Schramm for suggesting this topic, as well as for many helpful conversations. The authors also thank Jiaoyang Huang for an insightful discussion. K.Z. was supported by the NSF through award DMS-2031883 and the Simons Foundation through Award 814639 for the Collaboration on the Theoretical Foundations of Deep Learning and by the NSF grant CCF-2006489.

References

Animashree Anandkumar, Yuan Deng, Rong Ge, and Hossein Mobahi. Homotopy analysis for tensor pca. In *Conference on Learning Theory*, pages 79–104. PMLR, 2017a.

Animashree Anandkumar, Rong Ge, Majid Janzamin, et al. Analyzing tensor power method dynamics in overcomplete regime. *Journal of Machine Learning Research*, 18(22):1–40, 2017b.

Gerard Ben Arous, Song Mei, Andrea Montanari, and Mihai Nica. The landscape of the spiked tensor model. *Communications on Pure and Applied Mathematics*, 72(11):2282–2330, 2019.

Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. Algorithmic thresholds for tensor pca. *The Annals of Probability*, 48(4):2052–2087, 2020.

Jinho Baik, Gérard Ben Arous, and Sandrine Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, 33(5): 1643 – 1697, 2005.

Mohsen Bayati and Andrea Montanari. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory*, 57 (2):764–785, 2011.

Florent Benaych-Georges and Raj Rao Nadakuditi. The singular values and vectors of low rank perturbations of large rectangular random matrices. *Journal of Multivariate Analysis*, 111:120–135, 2012.

Quentin Berthet and Philippe Rigollet. Optimal detection of sparse principal components in high dimension. *The Annals of Statistics*, 41(4):1780–1815, 2013.

Giulio Biroli, Chiara Cammarota, and Federico Ricci-Tersenghi. How to iron out rough landscapes and get optimal performances: averaged gradient descent and its application to tensor pca. *Journal of Physics A: Mathematical and Theoretical*, 53(17):174003, 2020.

Changxiao Cai, Gen Li, H Vincent Poor, and Yuxin Chen. Nonconvex low-rank tensor completion from noisy data. *Advances in neural information processing systems*, 32, 2019.

Wei-Kuo Chen. Phase transition in the spiked random tensor with rademacher prior. *The Annals of Statistics*, 47(5):2734–2756, 2019.

Yash Deshpande, Andrea Montanari, et al. Sparse pca via covariance thresholding. *Journal of Machine Learning Research*, 17(141):1–41, 2016.

Rishabh Dudeja and Daniel Hsu. Statistical query lower bounds for tensor pca. *The Journal of Machine Learning Research*, 22(1):3729–3779, 2021.

David Gamarnik. The overlap gap property: A topological barrier to optimizing over random structures. *Proceedings of the National Academy of Sciences*, 118(41):e2108492118, 2021.

Lars Grasedyck, Daniel Kressner, and Christine Tobler. A literature survey of low-rank tensor approximation techniques. *GAMM-Mitteilungen*, 36(1):53–78, 2013.

Rungang Han, Rebecca Willett, and Anru R Zhang. An optimal statistical and computational framework for generalized tensor estimation. *The Annals of Statistics*, 50(1):1–29, 2022.

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

Samuel B Hopkins, Jonathan Shi, and David Steurer. Tensor principal component analysis via sum-of-square proofs. In *Conference on Learning Theory*, pages 956–1006. PMLR, 2015.

Samuel B Hopkins, Tselil Schramm, Jonathan Shi, and David 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.

Victoria Hore, Ana Vinuela, Alfonso Buil, Julian Knight, Mark I McCarthy, Kerrin Small, and Jonathan Marchini. Tensor decomposition for multiple-tissue gene expression experiments. *Nature genetics*, 48(9):1094–1100, 2016.

Jiaoyang Huang, Daniel Z Huang, Qing Yang, and Guang Cheng. Power iteration for tensor pca. *Journal of Machine Learning Research*, 23(128):1–47, 2022.

Bing-Yi Jing, Ting Li, Zhongyuan Lyu, and Dong Xia. Community detection on mixture multilayer networks via regularized tensor decomposition. *The Annals of Statistics*, 49 (6):3181–3205, 2021.

Iain M Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *The Annals of statistics*, 29(2):295–327, 2001.

Chris Jones and Lucas Pesenti. Diagram analysis of iterative algorithms. *arXiv preprint arXiv:2404.07881*, 2024.

Michel Journée, Yurii Nesterov, Peter Richtárik, and Rodolphe Sepulchre. Generalized power method for sparse principal component analysis. *Journal of Machine Learning Research*, 11(2), 2010.

Boris N Khoromskij and Christoph Schwab. Tensor-structured galerkin approximation of parametric and stochastic elliptic pdes. *SIAM journal on scientific computing*, 33(1): 364–385, 2011.

Chiheon Kim, Afonso S Bandeira, and Michel 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.

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

Marc Lelarge and Léo Miolane. Fundamental limits of symmetric low-rank matrix estimation. *Probability Theory and Related Fields*, 173:859–929, 2019.

Thibault Lesieur, Léo Miolane, Marc Lelarge, Florent Krzakala, and Lenka Zdeborová. Statistical and computational phase transitions in spiked tensor estimation. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 511–515. IEEE, 2017.

Gen Li and Yuting Wei. A non-asymptotic framework for approximate message passing in spiked models. *arXiv preprint arXiv:2208.03313*, 2022.

Gen Li, Wei Fan, and Yuting Wei. Approximate message passing from random initialization with applications to $\mathbb{Z}_2$ synchronization. *arXiv preprint arXiv:2302.03682*, 2023.

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

Yuetian Luo and Anru R Zhang. Tensor clustering with planted structures: Statistical optimality and computational limits. *The Annals of Statistics*, 50(1):584–613, 2022.

Zongming Ma. Sparse principal component analysis and iterative thresholding. *Annals of Statistics*, 41(296):772–801, 2013.

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

Andrea Montanari and Ramji Venkataramanan. Estimation of low-rank matrices via approximate message passing. *The Annals of Statistics*, 49(1), 2021.

Andrea Montanari and Yuchen Wu. Fundamental limits of low-rank matrix estimation with diverging aspect ratios. *arXiv preprint arXiv:2211.00488*, 2022.

Maximilian Nickel, Volker Tresp, Hans-Peter Kriegel, et al. A three-way model for collective learning on multi-relational data. In *Icml*, volume 11, pages 3104482–3104584, 2011.

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

Steffen Rendle and Lars 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.

Emile Richard and Andrea Montanari. A statistical model for tensor pca. *Advances in neural information processing systems*, 27, 2014.

Cynthia Rush and Ramji Venkataramanan. Finite sample analysis of approximate message passing algorithms. *IEEE Transactions on Information Theory*, 64(11):7264–7286, 2018.

Devavrat Shah and Christina Lee Yu. Iterative collaborative filtering for sparse noisy tensor estimation. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pages 41–45. IEEE, 2019.

Nicholas D Sidiropoulos, Lieven De Lathauwer, Xiao Fu, Kejun Huang, Evangelos E Papalexakis, and Christos Faloutsos. Tensor decomposition for signal processing and machine learning. *IEEE Transactions on signal processing*, 65(13):3551–3582, 2017.

Miaoyan Wang, Jonathan Fischer, and Yun S Song. Three-way clustering of multi-tissue multi-individual gene expression data using semi-nonnegative tensor decomposition. *The annals of applied statistics*, 13(2):1103, 2019.

Yining Wang and Anima Anandkumar. Online and differentially-private tensor decomposition. *Advances in Neural Information Processing Systems*, 29, 2016.

Jeffrey R Wozniak, Linda Krach, Erin Ward, Bryon A Mueller, Ryan Muetzel, Sarah Schnoebelen, Andrew Kiragu, and Kelvin O Lim. Neurocognitive and neuroimaging correlates of pediatric traumatic brain injury: a diffusion tensor imaging (dti) study. *Archives of Clinical Neuropsychology*, 22(5):555–568, 2007.

Yuchen Wu and Kangjie Zhou. Lower bounds for the convergence of tensor power iteration on random overcomplete models. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 3783–3820. PMLR, 2023.

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

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

Hua Zhou, Lexin Li, and Hongtu Zhu. Tensor regression with applications in neuroimaging data analysis. *Journal of the American Statistical Association*, 108(502):540–552, 2013.

Pan Zhou, Canyi Lu, Jiashi Feng, Zhouchen Lin, and Shuicheng Yan. Tensor low-rank representation for data recovery and clustering. *IEEE transactions on pattern analysis and machine intelligence*, 43(5):1718–1732, 2019.

Yuchen Zhou and Yuxin Chen. Heteroskedastic tensor clustering. *arXiv preprint arXiv:2311.02306*, 2023.

Appendix A. Technical lemmas

Lemma A.1 (Tails of the normal distribution) *Let $g \sim \mathsf{N}(0,1)$. Then for all $t > 0$, we have*

$$\left(\frac{1}{t} - \frac{1}{t^3}\right) \cdot \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \leq \mathbb{P}\left(g \geq t\right) \leq \frac{1}{t} \cdot \frac{1}{\sqrt{2\pi}} e^{-t^2/2}.$$

Lemma A.2 (Bernstein's inequality) *Let $X_1, \cdots, X_N$ be independent, mean zero, sub-exponential random variables. Then, for every $t \geq 0$, we have*

$$\mathbb{P}\left(\left|\sum_{i=1}^{N} X_i\right| \geq t\right) \leq 2\exp\left[-c\min\left(\frac{t^2}{\sum_{i=1}^{N}\|X_i\|_{\Psi_1}^2}, \frac{t}{\max_i \|X_i\|_{\Psi_1}}\right)\right].$$

Lemma A.3 *Let $\boldsymbol{x}_i \overset{\text{i.i.d.}}{\sim} \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_n)$ for $i = 1, \cdots, m$, where $m < \sqrt{n}$. Then, we have*

$$\mathbb{P}\left(\sup_{\boldsymbol{\alpha}\in\mathbb{S}^{m-1}} \left|\left\|\sum_{i=1}^{m} \alpha_i \boldsymbol{x}_i\right\|_2 - \sqrt{n}\right| \geq \varepsilon\sqrt{n}\right) \leq \exp(-C\varepsilon^2 n).$$

for some absolute constant $C > 0$.

Proof We use a covering argument. For $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{S}^{m-1}$ with $\|\boldsymbol{\alpha} - \boldsymbol{\beta}\|_2 \leq \varepsilon$, we have

$$\left|\left\|\sum_{i=1}^{m} \alpha_i \boldsymbol{x}_i\right\|_2 - \left\|\sum_{i=1}^{m} \beta_i \boldsymbol{x}_i\right\|_2\right| \leq \|\boldsymbol{X}\|_{\text{op}} \|\boldsymbol{\alpha} - \boldsymbol{\beta}\|_2 \leq C\sqrt{n}\varepsilon$$

with probability at least $1 - \exp(-Cn)$, where $\boldsymbol{X}$ is the matrix whose i-th column is $\boldsymbol{x}_i$. For any fixed $\boldsymbol{\alpha} \in \mathbb{S}^{m-1}$, we have $\sum_{i=1}^{m} \alpha_i \boldsymbol{x}_i \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_n)$. Therefore, for any $t \in (0, \sqrt{n})$, we have

$$\mathbb{P}\left(\left|\left\|\sum_{i=1}^{m} \alpha_i \boldsymbol{x}_i\right\|_2 - \sqrt{n}\right| \geq t\right) \leq 2\exp\left(-\frac{t^2}{8}\right),$$

where the last inequality follows from concentration of sub-exponential random variables. Now, let N_ε^m be an ε-net of $\mathbb{S}^{m-1}$, we thus obtain that

$$\begin{aligned}
&\mathbb{P}\left(\sup_{\boldsymbol{\alpha}\in\mathbb{S}^{m-1}} \left|\left\|\sum_{i=1}^{m} \alpha_i \boldsymbol{x}_i\right\|_2 - \sqrt{n}\right| \geq 2C\sqrt{n}\varepsilon\right) \\
&\leq \exp(-Cn) + \mathbb{P}\left(\sup_{\boldsymbol{\alpha}\in N_\varepsilon^m} \left|\left\|\sum_{i=1}^{m} \alpha_i \boldsymbol{x}_i\right\|_2 - \sqrt{n}\right| \geq C\sqrt{n}\varepsilon\right) \\
&\leq \exp(-Cn) + \left(\frac{C}{\varepsilon}\right)^m \times 2\exp\left(-\frac{C^2 n\varepsilon^2}{8}\right) \leq \exp(-C\varepsilon^2 n).
\end{aligned}$$

Replacing ε by $C\varepsilon$ completes the proof. ■

Appendix B. Proof of Lemma 3.1

Recall that $\mathcal{F}_t$ is defined in Eq. (10). We first show that

$$\mathcal{F}_t = \sigma\left(\left\{\boldsymbol{w}_{i_1,i_2,\cdots,i_{k-1}} : (i_1, i_2, \cdots, i_{k-1}) \in \boldsymbol{H}_{t-1}\right\} \cup \{\mathbf{g}, \boldsymbol{v}\}\right),$$

which is equivalent to proving that $\boldsymbol{v}^0, \boldsymbol{v}^1, \cdots, \boldsymbol{v}^t$ are measurable with respect to the σ-algebra on the right hand side of the above equation. Since $\boldsymbol{v}^0 = \mathbf{g}/\left\|\mathbf{g}\right\|_2$, we know that $\boldsymbol{v}^0$ is measurable. Using decomposition (8) with $t = 0$, we know that $\boldsymbol{v}^1$ is measurable as well. Repeating this argument yields that $\boldsymbol{v}^2, \cdots, \boldsymbol{v}^t$ are all measurable. This proves our claim.

Next, we are in position to prove the lemma. To avoid heavy notation, we denote

$$\boldsymbol{u}_{i_1,\cdots,i_{k-1}} = \prod_{j=1}^{k-1} \left\|\tilde{\boldsymbol{v}}_{\perp}^{i_j}\right\|_2^{-1} \cdot \left(\tilde{\boldsymbol{v}}_{\perp}^{i_1} \otimes \cdots \otimes \tilde{\boldsymbol{v}}_{\perp}^{i_{k-1}}\right).$$

We further define for $(i_1, \cdots, i_{k-1}) \in \mathbb{N}^{k-1}$ and $\boldsymbol{W} \in (\mathbb{R}^n)^{\otimes k}$ the rank-one tensor:

$$\mathrm{P}_{(i_1,\cdots,i_{k-1})}\boldsymbol{W} = \boldsymbol{W}\left[\boldsymbol{u}_{i_1,\cdots,i_{k-1}}\right] \otimes \boldsymbol{u}_{i_1,\cdots,i_{k-1}}. \tag{28}$$

Straightforward calculation reveals that for $(i_1, \cdots, i_{k-1}),\ (j_1, \cdots, j_{k-1}) \in \mathbb{N}^{k-1}$,

$$\begin{aligned}
&\mathrm{P}_{(j_1,\cdots,j_{k-1})}\mathrm{P}_{(i_1,\cdots,i_{k-1})}\boldsymbol{W} = \mathrm{P}_{(i_1,\cdots,i_{k-1})}\boldsymbol{W}\left[\boldsymbol{u}_{j_1,\cdots,j_{k-1}}\right] \otimes \boldsymbol{u}_{j_1,\cdots,j_{k-1}} \\
=& \left\langle \boldsymbol{u}_{i_1,\cdots,i_{k-1}}, \boldsymbol{u}_{j_1,\cdots,j_{k-1}} \right\rangle \cdot \boldsymbol{W}\left[\boldsymbol{u}_{i_1,\cdots,i_{k-1}}\right] \otimes \boldsymbol{u}_{j_1,\cdots,j_{k-1}} \\
\overset{(i)}{=}& \delta_{i_1 j_1} \cdots \delta_{i_{k-1} j_{k-1}} \cdot \boldsymbol{W}\left[\boldsymbol{u}_{i_1,\cdots,i_{k-1}}\right] \otimes \boldsymbol{u}_{j_1,\cdots,j_{k-1}} \\
=& \delta_{i_1 j_1} \cdots \delta_{i_{k-1} j_{k-1}} \cdot \mathrm{P}_{(i_1,\cdots,i_{k-1})}\boldsymbol{W},
\end{aligned}$$

where (i) follows from the fact that the $\boldsymbol{u}_{i_1,\cdots,i_{k-1}}$'s are mutually orthogonal, and δ represents the Kronecker delta: $\delta_{ij} = \mathbf{1}\{i = j\}$. Moreover, for any subset $S \subset \mathbb{N}^{k-1}$, define

$$\mathrm{P}_S\boldsymbol{W} = \sum_{(i_1,\cdots,i_{k-1}) \in S} \mathrm{P}_{(i_1,\cdots,i_{k-1})}\boldsymbol{W}, \quad \mathrm{P}_S^{\perp}\boldsymbol{W} = \boldsymbol{W} - \mathrm{P}_S\boldsymbol{W}. \tag{29}$$

One can prove using the previous calculation that for $S,\ T \subset \mathbb{N}^{k-1}$,

$$\mathrm{P}_S\mathrm{P}_T = \mathrm{P}_T\mathrm{P}_S = \mathrm{P}_{S\cap T}. \tag{30}$$

We will show that for all $t \in \mathbb{N}$,

$$\boldsymbol{W} = \mathrm{P}_{\boldsymbol{H}_{t-1}}\boldsymbol{W} + \mathrm{P}_{\boldsymbol{H}_{t-1}}^{\perp}\tilde{\boldsymbol{W}}_t, \tag{31}$$

where $\tilde{\boldsymbol{W}}_t \overset{d}{=} \boldsymbol{W}$ and is independent of $\mathcal{F}_t$. We prove Eq. (31) by induction. For $t = 0$, it is obvious that we can simply choose $\tilde{\boldsymbol{W}}_0 = \boldsymbol{W}$, since $\boldsymbol{W}$ is independent of $\mathcal{F}_0 = \sigma(\{\mathbf{g}, \boldsymbol{v}\})$. Now assume (31) holds for $t = s$, namely we have

$$\boldsymbol{W} = \mathrm{P}_{\boldsymbol{H}_{s-1}}\boldsymbol{W} + \mathrm{P}_{\boldsymbol{H}_{s-1}}^{\perp}\tilde{\boldsymbol{W}}_s, \quad \tilde{\boldsymbol{W}}_s \overset{d}{=} \boldsymbol{W},\ \tilde{\boldsymbol{W}}_s \perp \mathcal{F}_s. \tag{32}$$

Then, for $t = s + 1$, let us define

$$\tilde{\boldsymbol{W}}_{s+1} = \mathrm{P}^{\perp}_{\boldsymbol{H}_s} \tilde{\boldsymbol{W}}_s + \mathrm{P}_{\boldsymbol{H}_s} \boldsymbol{W}'_s, \tag{33}$$

where $\boldsymbol{W}'_s \overset{d}{=} \boldsymbol{W}$ and is independent of everything else. We show that $\tilde{\boldsymbol{W}}_{s+1}$ defined as above satisfies our requirement. To this end, first note that since $\boldsymbol{H}_{s-1} \subset \boldsymbol{H}_s$,

$$\begin{aligned} \mathrm{P}^{\perp}_{\boldsymbol{H}_s} \boldsymbol{W} &= \mathrm{P}^{\perp}_{\boldsymbol{H}_s} \mathrm{P}_{\boldsymbol{H}_{s-1}} \boldsymbol{W} + \mathrm{P}^{\perp}_{\boldsymbol{H}_s} \mathrm{P}^{\perp}_{\boldsymbol{H}_{s-1}} \tilde{\boldsymbol{W}}_s = \mathrm{P}^{\perp}_{\boldsymbol{H}_s} \tilde{\boldsymbol{W}}_s, \\ \mathrm{P}^{\perp}_{\boldsymbol{H}_s} \tilde{\boldsymbol{W}}_{s+1} &= \mathrm{P}^{\perp}_{\boldsymbol{H}_s} \mathrm{P}^{\perp}_{\boldsymbol{H}_s} \tilde{\boldsymbol{W}}_s + \mathrm{P}^{\perp}_{\boldsymbol{H}_s} \mathrm{P}_{\boldsymbol{H}_s} \boldsymbol{W}'_s = \mathrm{P}^{\perp}_{\boldsymbol{H}_s} \tilde{\boldsymbol{W}}_s, \end{aligned}$$

which further implies that $\mathrm{P}^{\perp}_{\boldsymbol{H}_s} \boldsymbol{W} = \mathrm{P}^{\perp}_{\boldsymbol{H}_s} \tilde{\boldsymbol{W}}_{s+1}$. Hence, we deduce that

$$\boldsymbol{W} = \mathrm{P}_{\boldsymbol{H}_s} \boldsymbol{W} + \mathrm{P}^{\perp}_{\boldsymbol{H}_s} \boldsymbol{W} = \mathrm{P}_{\boldsymbol{H}_s} \boldsymbol{W} + \mathrm{P}^{\perp}_{\boldsymbol{H}_s} \tilde{\boldsymbol{W}}_{s+1}, \tag{34}$$

i.e., Eq. (31) holds for $t = s + 1$. Next, it suffices to show that $\tilde{\boldsymbol{W}}_{s+1} \overset{d}{=} \boldsymbol{W}$ and that $\tilde{\boldsymbol{W}}_{s+1}$ is independent of $\mathcal{F}_{s+1}$. Recall that we already proved

$$\mathcal{F}_{s+1} = \sigma \left(\left\{ \boldsymbol{w}_{i_1, i_2, \cdots, i_{k-1}} : (i_1, i_2, \cdots, i_{k-1}) \in \boldsymbol{H}_s \right\} \cup \{ \mathbf{g}, \boldsymbol{v} \} \right). \tag{35}$$

According to Eq. (32) and direct calculation, we know that for $(i_1, \cdots, i_{k-1}) \in \boldsymbol{H}_s \backslash \boldsymbol{H}_{s-1}$,

$$\boldsymbol{w}_{i_1, \cdots, i_{k-1}} = \boldsymbol{W} \left[\boldsymbol{u}_{i_1, \cdots, i_{k-1}} \right] = \tilde{\boldsymbol{W}}_s \left[\boldsymbol{u}_{i_1, \cdots, i_{k-1}} \right],$$

and that $\boldsymbol{u}_{i_1, \cdots, i_{k-1}} \in \mathcal{F}_s$. Therefore,

$$\mathcal{F}_{s+1} = \sigma \left(\mathcal{F}_s \cup \sigma \left\{ \tilde{\boldsymbol{W}}_s \left[\boldsymbol{u}_{i_1, \cdots, i_{k-1}} \right] : (i_1, \cdots, i_{k-1}) \in \boldsymbol{H}_s \backslash \boldsymbol{H}_{s-1} \right\} \right). \tag{36}$$

Next we compute the conditional distribution of $\mathrm{P}^{\perp}_{\boldsymbol{H}_s} \tilde{\boldsymbol{W}}_s$ given $\mathcal{F}_{s+1}$, which is equivalent to the law of $\mathrm{P}^{\perp}_{\boldsymbol{H}_s} \tilde{\boldsymbol{W}}_s$ conditioning on $\mathcal{F}_s$ and the random variables $\tilde{\boldsymbol{W}}_s \left[\boldsymbol{u}_{i_1, \cdots, i_{k-1}} \right]$ for $(i_1, \cdots, i_{k-1}) \in \boldsymbol{H}_s \backslash \boldsymbol{H}_{s-1}$. Here, we can view the $(k-1)$-tensors $\boldsymbol{u}_{i_1, \cdots, i_{k-1}}$ as fixed since they are measurable with respect to $\mathcal{F}_s$. By definition, these $\boldsymbol{u}_{i_1, \cdots, i_{k-1}}$'s are mutually orthogonal and belong to $\boldsymbol{H}_s$, and we know from induction hypothesis that $\tilde{\boldsymbol{W}}_s | \mathcal{F}_s \overset{d}{=} \boldsymbol{W}$. Applying Lemma 4.1 in Huang et al. (2022) yields that the conditional distribution of $\mathrm{P}^{\perp}_{\boldsymbol{H}_s} \tilde{\boldsymbol{W}}_s$ is equal to the law of $\mathrm{P}^{\perp}_{\boldsymbol{H}_s} \boldsymbol{W}''_s$, where $\boldsymbol{W}''_s$ is an independent copy of $\tilde{\boldsymbol{W}}_s$ which is further independent of $\mathcal{F}_{s+1}$. As a consequence, it follows that

$$\tilde{\boldsymbol{W}}_{s+1} | \mathcal{F}_{s+1} \overset{d}{=} \mathrm{P}^{\perp}_{\boldsymbol{H}_s} \boldsymbol{W}''_s + \mathrm{P}_{\boldsymbol{H}_s} \boldsymbol{W}'_s | \mathcal{F}_{s+1} \overset{d}{=} \boldsymbol{W},$$

i.e., $\tilde{\boldsymbol{W}}_{s+1} \overset{d}{=} \boldsymbol{W}$ is independent of $\mathcal{F}_{s+1}$. This completes the induction. Now, using Eq. (31), we know that for $(i_1, \cdots, i_{k-1}) \in \boldsymbol{H}_t \backslash \boldsymbol{H}_{t-1}$, $\boldsymbol{w}_{i_1, \cdots, i_{k-1}} = \tilde{\boldsymbol{W}}_t [\boldsymbol{u}_{i_1, \cdots, i_{k-1}}]$, where $\tilde{\boldsymbol{W}}_t \perp \mathcal{F}_t$, $\{ \boldsymbol{u}_{i_1, \cdots, i_{k-1}} \}$ is an orthonormal set that is measurable with respect to $\mathcal{F}_t$. It then follows immediately that $\boldsymbol{w}_{i_1, \cdots, i_{k-1}} \overset{\text{i.i.d.}}{\sim} \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_n)$ are independent of $\mathcal{F}_t$ for $(i_1, \cdots, i_{k-1}) \in \boldsymbol{H}_t \backslash \boldsymbol{H}_{t-1}$. This completes the proof of Lemma 3.1.

Appendix C. Proof of Lemma 3.2

The entire argument is divided into three parts:

Concentration of ζ_t. To this end, we need to estimate $\left\|\boldsymbol{v}^t\right\|_2$. From Eq. (8), we know that

$$\boldsymbol{v}^t = \alpha_t \boldsymbol{v} + \sum_{(i_1,\cdots,i_{k-1})\in \boldsymbol{H}_{t-1}} \beta^{(t-1)}_{i_1,i_2,\cdots,i_{k-1}} \boldsymbol{w}_{i_1,i_2,\cdots,i_{k-1}}. \tag{37}$$

Since $\boldsymbol{w}_{i_1,i_2,\cdots,i_{k-1}} \overset{\text{i.i.d.}}{\sim} \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_n)$, and $\sum_{(i_1,\cdots,i_{k-1})\in \boldsymbol{H}_{t-1}} (\beta^{(t-1)}_{i_1,i_2,\cdots,i_{k-1}})^2 = 1$, we deduce from Lemma A.3 that with probability at least $1-\exp(-C\eta^2 n)$, for all $t < n^{1/2(k-1)}$,

$$(1-\eta)\sqrt{n} \le \left\| \sum_{(i_1,\cdots,i_{k-1})\in \boldsymbol{H}_{t-1}} \beta^{(t-1)}_{i_1,i_2,\cdots,i_{k-1}} \boldsymbol{w}_{i_1,i_2,\cdots,i_{k-1}} \right\|_2 \le (1+\eta)\sqrt{n}, \tag{38}$$

where $\eta > 0$ is a small constant. Further, as long as $t < T_\varepsilon$, by definition we have

$$\left\|\alpha_t \boldsymbol{v}\right\|_2 = |\alpha_t| < n^\varepsilon,$$

which leads to

$$(1-2\eta)\sqrt{n} \le (1-\eta)\sqrt{n} - n^\varepsilon \le \left\|\boldsymbol{v}^t\right\|_2 \le (1+\eta)\sqrt{n} + n^\varepsilon \le (1+2\eta)\sqrt{n}. \tag{39}$$

To summarize, we conclude that with probability at least $1-\exp(-C\eta^2 n)$, the following happens:

$$\text{For all } t < \min(T_\varepsilon, n^{1/2(k-1)}),\ \left\|\boldsymbol{v}^t\right\|_2 \in [(1-\eta)\sqrt{n}, (1+\eta)\sqrt{n}]. \tag{40}$$

Recall that $\zeta_t = (\sqrt{n}/\left\|\boldsymbol{v}^t\right\|_2)^{k-1}$, the above bound also implies that $\zeta_t \in [1-2(k-1)\eta, 1+2(k-1)\eta]$ for sufficiently small $\eta > 0$. In fact, from the proof of Lemma A.3 we know that η can be chosen as $n^{-1/6}/2(k-1)$, so that

$$\mathbb{P}\left(\zeta_t \in [1-n^{-1/6}, 1+n^{-1/6}] \text{ for all } t < \min(T_\varepsilon, n^{1/2(k-1)})\right) \ge 1-\exp(-Cn^{2/3}). \tag{41}$$

Concentration of c_t. To show that c_t is close to 1 with high probability, we need to control $\|\tilde{\boldsymbol{v}}^{t-1}_{\parallel}\|_2$. By definition, we have

$$\left\|\tilde{\boldsymbol{v}}^{t-1}_{\parallel}\right\|_2 = \frac{\left\|\boldsymbol{v}^{t-1}_{\parallel}\right\|_2}{\left\|\boldsymbol{v}^{t-1}\right\|_2} \le \frac{1}{1-\eta} \frac{\left\|\boldsymbol{v}^{t-1}_{\parallel}\right\|_2}{\sqrt{n}}$$

for all $t < \min(T_\varepsilon, n^{1/2(k-1)}) + 1$ with probability at least $1-\exp(-C\eta^2 n)$. It then suffices to establish an upper bound for $\left\|\boldsymbol{v}^t_{\parallel}\right\|_2$ with $t < \min(T_\varepsilon, n^{1/2(k-1)})$. Note that

$$\begin{aligned}
\boldsymbol{v}^t_{\parallel} &= \Pi_{\boldsymbol{V}^{t-1}} \boldsymbol{v}^t = \Pi_{\boldsymbol{V}^{t-1}} \left(\alpha_t \boldsymbol{v} + \boldsymbol{h}_t + \sqrt{1-\|\tilde{\boldsymbol{v}}^{t-1}_{\parallel}\|_2^{2k-2}}\mathbf{g}_t\right) \\
&= \alpha_t \cdot \Pi_{\boldsymbol{V}^{t-1}} \boldsymbol{v} + \Pi_{\boldsymbol{V}^{t-1}} \boldsymbol{h}_t + \sqrt{1-\|\tilde{\boldsymbol{v}}^{t-1}_{\parallel}\|_2^{2k-2}} \cdot \Pi_{\boldsymbol{V}^{t-1}} \mathbf{g}_t,
\end{aligned}$$

which further implies that

$$\left\|\boldsymbol{v}^t_{\parallel}\right\|_2 \le \alpha_t + \left\|\boldsymbol{h}_t\right\|_2 + \sqrt{1-\|\tilde{\boldsymbol{v}}^{t-1}_{\parallel}\|_2^{2k-2}} \cdot \left\|\Pi_{\boldsymbol{V}^{t-1}} \mathbf{g}_t\right\|_2. \tag{42}$$

By definition of $\boldsymbol{h}_t$, we know that $\boldsymbol{h}_t$ can be written as

$$\boldsymbol{h}_t = \|\tilde{\boldsymbol{v}}_{\|}^{t-1}\|_2^{k-1} \cdot \sum_{j=1}^{(t-1)^{k-1}} \beta_j \boldsymbol{w}_j,$$

where $\sum_{j=1}^{(t-1)^{k-1}} \beta_j^2 = 1$ and $\boldsymbol{w}_j \sim_{\text{i.i.d.}} \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_n)$. Applying again Lemma A.3, it follows that

$$\|\boldsymbol{h}_t\|_2 \le 2\sqrt{n} \cdot \|\tilde{\boldsymbol{v}}_{\|}^{t-1}\|_2^{k-1} \text{ for all } t < n^{1/2(k-1)} \tag{43}$$

with probability at least $1 - \exp(-Cn)$. Further, since $\mathbf{g}_t$ is independent of $\boldsymbol{V}^{t-1}$, we have $\|\Pi_{\boldsymbol{V}^{t-1}} \mathbf{g}_t\|_2^2 \sim \chi^2(t)$. As a consequence,

$$\mathbb{P}\left(\|\Pi_{\boldsymbol{V}^{t-1}} \mathbf{g}_t\|_2 \le Cn^{\varepsilon} \text{ for all } t < n^{1/2(k-1)}\right) \ge 1 - \exp\left(-Cn^{2\varepsilon}\right). \tag{44}$$

To summarize, with probability $1 - \exp(-Cn^{2\varepsilon})$, the following estimate holds for all $t < \min(T_\varepsilon, n^{1/2(k-1)})$:

$$\begin{aligned}
\left\|\boldsymbol{v}_{\|}^{t}\right\|_2 &\le \alpha_t + \|\boldsymbol{h}_t\|_2 + \sqrt{1 - \|\tilde{\boldsymbol{v}}_{\|}^{t-1}\|_2^{2k-2}} \cdot \|\Pi_{\boldsymbol{V}^{t-1}} \mathbf{g}_t\|_2 \\
&\le n^{\varepsilon} + 2\sqrt{n} \cdot \|\tilde{\boldsymbol{v}}_{\|}^{t-1}\|_2^{k-1} + Cn^{\varepsilon}\sqrt{1 - \|\tilde{\boldsymbol{v}}_{\|}^{t-1}\|_2^{2k-2}} \\
&\le C\left(n^{\varepsilon} + \sqrt{n} \cdot \|\tilde{\boldsymbol{v}}_{\|}^{t-1}\|_2^{k-1}\right),
\end{aligned}$$

which further implies that

$$\left\|\tilde{\boldsymbol{v}}_{\|}^{t}\right\|_2 \le C\left(n^{\varepsilon - 1/2} + \left\|\tilde{\boldsymbol{v}}_{\|}^{t-1}\right\|_2^{k-1}\right).$$

At initialization, we have $\left\|\tilde{\boldsymbol{v}}_{\|}^{0}\right\|_2 = 0$. We will use induction to show that $\left\|\tilde{\boldsymbol{v}}_{\|}^{t}\right\|_2 \le (C + 1)n^{\varepsilon - 1/2}$ as long as the above inequality holds. Assume this is true for $t-1$, then we have

$$\begin{aligned}
\left\|\tilde{\boldsymbol{v}}_{\|}^{t}\right\|_2 &\le C\left(n^{\varepsilon - 1/2} + (C+1)^{k-1} n^{(k-1)(\varepsilon - 1/2)}\right) \\
&\le n^{\varepsilon - 1/2}\left(C + C(C+1)^{k-1} n^{(k-2)(\varepsilon - 1/2)}\right) \\
&\le (C+1) n^{\varepsilon - 1/2}
\end{aligned}$$

for sufficiently large n. We thus conclude that

$$\mathbb{P}\left(|c_t - 1| \le Cn^{2(k-1)(\varepsilon - 1/2)} \text{ for all } t < \min(T_\varepsilon, n^{1/2(k-1)})\right) \ge 1 - \exp(-Cn^{2\varepsilon}). \tag{45}$$

Concentration of b_t. Recall the definition of b_t:

$$\begin{aligned}
b_t = \langle \boldsymbol{h}_t, \boldsymbol{v} \rangle &= \sum_{(i_1, \cdots, i_{k-1}) \in \boldsymbol{H}_{t-2}} \beta^{(t-1)}_{i_1, i_2, \cdots, i_{k-1}} \cdot \left\langle \boldsymbol{w}_{i_1, i_2, \cdots, i_{k-1}}, \boldsymbol{v} \right\rangle \\
&= \|\tilde{\boldsymbol{v}}_{\|}^{t-1}\|_2^{k-1} \cdot \sum_{j=1}^{(t-1)^{k-1}} \beta_j \langle \boldsymbol{w}_j, \boldsymbol{v} \rangle,
\end{aligned}$$

where we have

$$\sum_{j=1}^{(t-1)^{k-1}} \beta_j^2 = 1, \text{ and } \boldsymbol{w}_j \overset{\text{i.i.d.}}{\sim} \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_n). \tag{46}$$

It then follows that

$$|b_t| \leq \|\tilde{\boldsymbol{v}}_{\|}^{t-1}\|_2^{k-1} \cdot \sup_{\boldsymbol{\beta}\in\mathbb{S}^{(t-1)^{k-1}-1}} \left| \sum_{j=1}^{(t-1)^{k-1}} \beta_j \langle \boldsymbol{w}_j, \boldsymbol{v}\rangle \right| \leq \|\tilde{\boldsymbol{v}}_{\|}^{t-1}\|_2^{k-1} \cdot \|\boldsymbol{W}\boldsymbol{v}\|_2 .$$

Since $\boldsymbol{W}\boldsymbol{v} \in \mathbb{R}^{(t-1)^{k-1}}$ has i.i.d. standard normal entries, we know that $\|\boldsymbol{W}\boldsymbol{v}\|_2 \leq Cn^{1/4}$ with probability no less than $1 - \exp(-C\sqrt{n})$. Therefore,

$$\mathbb{P}\left(|b_t| \leq Cn^{1/4} \|\tilde{\boldsymbol{v}}_{\|}^{t-1}\|_2^{k-1} \text{ for all } t < n^{1/2(k-1)} \right) \geq 1 - \exp(-C\sqrt{n}). \tag{47}$$

Using the upper bound on $\|\tilde{\boldsymbol{v}}_{\|}^{t-1}\|_2$ we obtained in the previous paragraph, it follows that

$$|b_t| \leq Cn^{1/4} \|\tilde{\boldsymbol{v}}_{\|}^{t-1}\|_2^{k-1} \leq Cn^{1/4+(k-1)(\varepsilon-1/2)} \tag{48}$$

for all $t < \min(T_\varepsilon, n^{1/2(k-1)})$ with probability at least $1 - \exp(-C\sqrt{n}) - \exp(-Cn^{2\varepsilon}) \geq 1 - \exp(-C\sqrt{n})$. This proves the concentration bound on b_t.

Appendix D. Proof of Lemma 3.3

Note that Proposition 3.1 implies that the following event occurs with probability at least $1 - \exp(-C\sqrt{n})$:

$$E = \left\{ \zeta_t \in [1 - n^{-1/6}, 1 + n^{-1/6}],\ |b_t| \leq Cn^{-1/6},\ |c_t - 1| \leq Cn^{-5/6} \text{ for all } t < \min(T_\varepsilon, n^{1/2(k-1)}) \right\}. \tag{49}$$

To facilitate our analysis, we define an auxiliary stochastic process $\{\bar{\alpha}_t\}_{t\in\mathbb{N}}$ as follows:

1. $\{\bar{\alpha}_t\}$ and $\{\alpha_t\}$ have the same initialization, i.e., $\bar{\alpha}_0 = \alpha_0 = 0$.
2. For all $t \in \mathbb{N}$, we have $\bar{\alpha}_{t+1} = \gamma_n \bar{\zeta}_t (\bar{\alpha}_t + \bar{b}_t + \bar{c}_t Z_t)^{k-1}$, where $(\bar{\zeta}_t, \bar{b}_t, \bar{c}_t) = (\zeta_t, b_t, c_t)$ if $t < \min(T_\varepsilon, n^{1/2(k-1)})$, and $(1, 0, 1)$ otherwise.

By definition, we know that $\alpha_t = \bar{\alpha}_t$ for all $t \in \{0, 1, \cdots, \min(T_\varepsilon, n^{1/2(k-1)})\}$. Further, setting $\delta = (C+1)n^{-1/6}$, Proposition 3.1 then implies that

$$\max\left\{ \left|\bar{\zeta}_t - 1\right|, \left|\bar{b}_t\right|, \left|\bar{c}_t - 1\right| \right\} \leq \delta, \quad \forall t \in \mathbb{N}. \tag{50}$$

Below we will establish lower and upper bounds on the first hitting time of $\{\bar{\alpha}_t\}_{t\in\mathbb{N}}$ to certain level sets, which is defined as follows:

$$\bar{T}_\varepsilon = \min\left\{ t \in \mathbb{N}_+ : |\bar{\alpha}_t| \geq n^\varepsilon \right\}. \tag{51}$$

Then, we will show that $\bar{T}_\varepsilon = T_\varepsilon$ with high probability, and consequently obtain the same lower and upper bounds on T_ε. To begin with, we state a helper lemma that is useful for establishing upper and lower bounds on $\bar{T}_\varepsilon$:

Lemma D.1 *Let $\delta := (C+1)n^{-1/6}$. Fix $\Delta \in [\delta, 1]$. Define the deterministic sequences $\{\bar{b}_{t,\Delta}\}_{t\geq 0}$ and $\{\underline{b}_{t,\Delta}\}_{t\geq 0}$ recursively as follows:*

$$\bar{b}_{0,\Delta} \geq 0, \quad \bar{b}_{t+1,\Delta} = \gamma_n(1+\Delta)^k \cdot \bar{b}_{t,\Delta}^{k-1}, \tag{52}$$

$$\underline{b}_{0,\Delta} \geq 0, \quad \underline{b}_{t+1,\Delta} = \gamma_n(1-\Delta)^k \cdot \underline{b}_{t,\Delta}^{k-1}. \tag{53}$$

Then, we have

$$\begin{aligned}\bar{b}_{t,\Delta} &= \Big(\gamma_n(1+\Delta)^k\Big)^{((k-1)^t-1)/(k-2)} \cdot \bar{b}_{0,\Delta}^{(k-1)^t}, \\ \underline{b}_{t,\Delta} &= \Big(\gamma_n(1-\Delta)^k\Big)^{((k-1)^t-1)/(k-2)} \cdot \underline{b}_{0,\Delta}^{(k-1)^t}.\end{aligned} \tag{54}$$

Furthermore, for any $t \in \mathbb{N}$,

$$\begin{aligned}&\mathbb{P}\left(|\bar{\alpha}_{t+s}| \leq \bar{b}_{s,\Delta},\ \forall s \in \mathbb{N} \;\middle|\; |\bar{\alpha}_t| \leq \bar{b}_{0,\Delta}\right) \\ &\geq \mathbb{P}\left(|Z_s| \leq \frac{\Delta\bar{b}_{s,\Delta} - \delta}{1+\delta}, \forall s \in \mathbb{N}\right) \geq 1 - 2\sum_{s=0}^{\infty}\Phi\left(-\frac{\Delta\bar{b}_{s,\Delta} - \delta}{1+\delta}\right),\end{aligned} \tag{55}$$

$$\begin{aligned}&\mathbb{P}\left(|\bar{\alpha}_{t+s}| \geq \underline{b}_{s,\Delta},\ \forall s \in \mathbb{N} \;\middle|\; |\bar{\alpha}_t| \geq \underline{b}_{0,\Delta}\right) \\ &\geq \mathbb{P}\left(|Z_s| \leq \frac{\Delta\underline{b}_{s,\Delta} - \delta}{1+\delta}, \forall s \in \mathbb{N}\right) \geq 1 - 2\sum_{s=0}^{\infty}\Phi\left(-\frac{\Delta\underline{b}_{s,\Delta} - \delta}{1+\delta}\right).\end{aligned} \tag{56}$$

We prove Lemma D.1 in Appendix E.1.

D.1 Lower bound on $\bar{T}_\varepsilon$

We start with two useful propositions, whose proofs are deferred to Appendices E.2 and E.3, respectively.

Proposition D.1 *Let $C_k = (k-2)^{k-2}/(k-1)^{k-1}$ and $\delta = (C+1)n^{-1/6}$. Define*

$$M(k,\gamma_n,\delta) := \frac{1}{k-2}\left(\frac{C_k}{\gamma_n(1+\delta)}\right)^{1/(k-2)},\ N(k,\gamma_n,\delta) := \frac{1}{1+\delta}\left(\frac{C_k}{\gamma_n(1+\delta)}\right)^{1/(k-2)} - \frac{\delta}{1+\delta}. \tag{57}$$

Then, for any $T \in \mathbb{N}$, with probability at least $1 - 2T\Phi(-N(k,\gamma_n,\delta))$, we have

$$\max_{0\leq t\leq T} |\bar{\alpha}_t| \leq M(k,\gamma_n,\delta). \tag{58}$$

Proposition D.2 *Assume $\gamma_n = n^{o(1)}$ and $\gamma_n \gg (\log n)^{-(k-2)/2}$. Let $\varepsilon = \varepsilon_k = (6k-11)/12(k-1)$. Then, for any fixed $\eta \in (0,1)$, and $t_n = \lfloor (1-\eta)\log_{k-1}\frac{\log_{k-1} n}{\max\{\log_{k-1}\gamma_n, 1\}}\rfloor$, we have*

$$\mathbb{P}\left(\max_{0\leq t\leq t_n} |\bar{\alpha}_t| \leq \frac{n^\varepsilon}{2}\right) \geq 1 - n^{-C}, \tag{59}$$

where $C > 0$ is an absolute constant.

With the aid of Proposition D.1 and Proposition D.2, we prove the follwing lower bound on $\bar{T}_\varepsilon$: For any fixed $\eta \in (0,1)$ and large enough n, with probability at least $1 - n^{-C}$ we have

$$\bar{T}_\varepsilon \geq \max\left\{\exp\left(\frac{1-\eta}{2}\left(\frac{C_k}{\gamma_n}\right)^{2/(k-2)}\right),\ (1-\eta)\log_{k-1}\frac{\log_{k-1} n}{\max\{\log_{k-1}\gamma_n, 1\}}\right\}. \tag{60}$$

To show Eq. (60), note that by definition of $\bar{T}_\varepsilon$ and Proposition D.2, it immediately follows that

$$\mathbb{P}\left(\bar{T}_\varepsilon \geq (1-\eta)\log_{k-1}\frac{\log_{k-1} n}{\max\{\log_{k-1}\gamma_n, 1\}}\right) \geq 1 - n^{-C}.$$

It then suffices to consider the case $(\log n)^{-(k-2)/2} \ll \gamma_n \ll 1$, otherwise the lower bound on the right hand side of Eq. (60) is just $(1-\eta)\log_{k-1}(\log_{k-1} n/\max\{\log_{k-1}\gamma_n, 1\})$ for a large enough n. Recall that $M(k, \gamma_n, \delta)$ and $N(k, \gamma_n, \delta)$ are defined in Proposition D.1. For a large enough n, we have

$$M(k,\gamma_n,\delta) \leq \frac{1}{k-2}\left(\frac{C_k}{\gamma_n}\right)^{1/(k-2)} \ll n^\varepsilon, \quad N(k,\gamma_n,\delta) \geq \frac{1}{1+10\delta}\left(\frac{C_k}{\gamma_n}\right)^{1/(k-2)} \gg 1,$$

and

$$1 - 2T\Phi\left(-N(k,\gamma_n,\delta)\right) \geq 1 - \frac{2T}{N(k,\gamma_n,\delta)}\exp\left(-\frac{1}{2}N(k,\gamma_n,\delta)^2\right).$$

As a consequence, as long as $T \leq \exp(N(k,\gamma_n,\delta)^2/2)$, with high probability we have

$$\max_{0\leq t\leq T}|\bar{\alpha}_t| \leq M(k,\gamma_n,\delta) \ll n^\varepsilon,$$

which further implies that

$$\bar{T}_\varepsilon \geq \exp\left(\frac{1}{2(1+10\delta)^2}\left(\frac{C_k}{\gamma_n}\right)^{2/(k-2)}\right) \geq \exp\left(\frac{1-\eta}{2}\left(\frac{C_k}{\gamma_n}\right)^{2/(k-2)}\right),$$

since $\delta \to 0$ as $n \to \infty$. This completes the proof of the lower bound given in Eq. (60).

D.2 Upper bound on $\bar{T}_\varepsilon$

Next, we establish an upper bound for $\bar{T}_\varepsilon$. Our proof consists of two steps. First, we show that in a moderately many steps, the alignment $\bar{\alpha}_t$ will reach a sufficiently large magnitude. Then, we use Lemma D.1 to prove that, after reaching this magnitude, it takes at most $(1+\eta)\log_{k-1}(\log_{k-1} n/\max\{\log_{k-1}\gamma_n, 1\})$ steps for $|\bar{\alpha}_t|$ to reach n^ε, where $\eta \in (0,1)$ is an arbitrarily small fixed positive constant. To get started, we establish the following proposition, the proof of which can be found in Appendix E.4.

Proposition D.3 *For any $m > 0$ and $T \in \mathbb{N}$, we have*

$$\mathbb{P}\left(\max_{0\leq t\leq T}|\bar{\alpha}_t| \geq m\right) \geq 1 - \exp\left(-T\Phi\left(-\left(\frac{m}{\gamma_n(1-\delta)^k}\right)^{1/(k-1)}\right)\right),$$

where we recall that $\delta = (C+1)n^{-1/6}$.

For $\Delta \in (\delta, 1)$ and $L > 0$, we define

$$m(\Delta, L) := \max\left\{ \left(\frac{1+\Delta}{\gamma_n (1-\Delta)^k} \right)^{1/(k-2)}, L \right\}, \tag{61}$$

we know that (using $\Delta \geq \delta$)

$$\left(\frac{m(\Delta, L)}{\gamma_n (1-\delta)^k} \right)^{1/(k-1)} \leq \max\left\{ \left(\frac{(1+\Delta)^{1/(k-1)}}{\gamma_n (1-\Delta)^k} \right)^{1/(k-2)}, \left(\frac{L(1+\Delta)^{1/(k-1)}}{\gamma_n (1-\Delta)^k} \right)^{1/(k-1)} \right\}.$$

Define $A_{k,\Delta} := (1+\Delta)^{1/(k-1)}/(1-\Delta)^k$ and $B_{k,\Delta} := (1+\Delta)/(1-\Delta)^k$. Invoking Proposition D.3 and taking $m = m(\Delta, L)$, we know that with probability at least

$$1 - \max\left\{ \exp\left(-T\Phi\left(-\left(\frac{A_{k,\Delta}}{\gamma_n} \right)^{1/(k-2)} \right) \right), \exp\left(-T\Phi\left(-\left(\frac{A_{k,\Delta} L}{\gamma_n} \right)^{1/(k-1)} \right) \right) \right\},$$

there exists $t \leq T$ satisfying

$$|\bar{\alpha}_t| \geq m(\Delta, L) = \max\left\{ \left(\frac{B_{k,\Delta}}{\gamma_n} \right)^{1/(k-2)}, L \right\}.$$

In what follows, we will show that for a sufficiently large L, starting from such $\bar{\alpha}_t$, it takes at most $(1+\eta)\log_{k-1} \frac{\log_{k-1} n}{\max\{\log_{k-1}\gamma_n, 1\}}$ steps for $\bar{\alpha}_t$ reach order n^{ε}. Such a result is established as Proposition D.4 below. We delay the proof of this proposition to Appendix E.5.

Proposition D.4 *Assume $\gamma_n = n^{o(1)}$ and $\gamma_n \gg (\log n)^{-(k-2)/2}$. Let $\varepsilon = \varepsilon_k = (6k-11)/12(k-1)$. Then, for any fixed $\eta \in (0, \infty)$ and $t_n = \lfloor (1+\eta)\log_{k-1} \frac{\log_{k-1} n}{\max\{\log_{k-1}\gamma_n, 1\}} \rfloor$, for L, n sufficiently large we have*

$$\mathbb{P}\left(|\bar{\alpha}_{t+t_n}| \geq n^{\varepsilon} \;\middle|\; |\bar{\alpha}_t| \geq m(\Delta, L) \right) \geq 1 - C\exp\left(-\frac{\Delta^2 m(\Delta, L)^2}{8} \right), \tag{62}$$

where C is an absolute constant.

Putting together Proposition D.3 and Proposition D.4, we obtain the following theorem:

Lemma D.2 (Upper bound on $\bar{T}_{\varepsilon}$) *Assume $\gamma_n = n^{o(1)}$, $\gamma_n \gg (\log n)^{-(k-2)/2}$, and $\varepsilon = \varepsilon_k = (6k-11)/12(k-1)$. For any fixed $\eta > 0$ and sufficiently large $n \in \mathbb{N}$, with probability $1 - o_n(1)$ we have*

$$\bar{T}_{\varepsilon} \leq \exp\left(\frac{1+\eta}{2} \left(\frac{1}{\gamma_n} \right)^{2/(k-2)} \right) + (1+\eta)\log_{k-1} \frac{\log_{k-1} n}{\max\{\log_{k-1}\gamma_n, 1\}}. \tag{63}$$

Proof [Proof of Lemma D.2] Let $L = L_n$ be such that $L \to \infty$ as $n \to \infty$ (specific choice of L will be discussed later). Then, from the discussion following Proposition D.3, we know that as long as

$$T \gg \max\left\{ \exp\left(\frac{1+\eta}{2} \left(\frac{A_{k,\Delta}}{\gamma_n} \right)^{2/(k-2)} \right), \exp\left(\frac{1+\eta}{2} \left(\frac{A_{k,\Delta} L}{\gamma_n} \right)^{2/(k-1)} \right) \right\}, \tag{64}$$

then with high probability there exists $t \le T$ satisfying

$$|\bar{\alpha}_t| \ge m(\Delta, L) = \max\left\{ \left(\frac{B_{k,\Delta}}{\gamma_n}\right)^{1/(k-2)}, L \right\}.$$

Applying Proposition D.4 yields that

$$|\bar{\alpha}_{t+t_n}| \ge n^{\varepsilon}, \quad t_n = \lfloor (1+\eta) \log_{k-1} \frac{\log_{k-1} n}{\max\{\log_{k-1} \gamma_n, 1\}} \rfloor. \tag{65}$$

The above calculation implies that

$$\bar{T}_{\varepsilon} \le T + (1+\eta) \log_{k-1} \frac{\log_{k-1} n}{\max\{\log_{k-1} \gamma_n, 1\}}. \tag{66}$$

Next we discuss the choice of L, which will eventually lead to an upper bound on T. Let L be the solution to the equation below

$$\exp\left(\frac{1+\eta}{2} \left(\frac{A_{k,\Delta} L}{\gamma_n}\right)^{2/(k-1)} \right) = \exp\left(\frac{1+\eta}{2} \left(\frac{A_{k,\Delta}}{\gamma_n}\right)^{2/(k-2)} \right) + \left(\log_{k-1} \frac{\log_{k-1} n}{\max\{\log_{k-1} \gamma_n, 1\}} \right)^{1-\eta},$$

then one can show that $L \to \infty$ as $n \to \infty$. In fact, the above equation implies

$$\begin{aligned} L &\ge C_{k,\Delta,\eta} \cdot \max\left\{ \gamma_n^{-1/(k-2)},\ \gamma_n \cdot \left((1-\eta) \cdot \log\log_{k-1} \frac{\log_{k-1} n}{\max\{\log_{k-1} \gamma_n, 1\}} \right)^{(k-1)/2} \right\} \\ &\ge C_{k,\Delta,\eta} \cdot \left((1-\eta) \cdot \log\log_{k-1} \frac{\log_{k-1} n}{\max\{\log_{k-1} \gamma_n, 1\}} \right)^{1/2} \to \infty, \end{aligned}$$

where $C_{k,\Delta,\eta}$ is a constant only depending on k, Δ, η. Under this choice of L, setting

$$T = \exp\left(\frac{1+2\eta}{2} \left(\frac{A_{k,\Delta}}{\gamma_n}\right)^{2/(k-2)} \right) + \eta \log_{k-1} \frac{\log_{k-1} n}{\max\{\log_{k-1} \gamma_n, 1\}} \tag{67}$$

verifies Eq. (64). We finally deduce that

$$\bar{T}_{\varepsilon} \le \exp\left(\frac{1+2\eta}{2} \left(\frac{A_{k,\Delta}}{\gamma_n}\right)^{2/(k-2)} \right) + (1+2\eta) \log_{k-1} \frac{\log_{k-1} n}{\max\{\log_{k-1} \gamma_n, 1\}}. \tag{68}$$

Since η and Δ can be arbitrarily small, the above upper bound is equivalent to

$$\bar{T}_{\varepsilon} \le \exp\left(\frac{1+\eta}{2} \left(\frac{1}{\gamma_n}\right)^{2/(k-2)} \right) + (1+\eta) \log_{k-1} \frac{\log_{k-1} n}{\max\{\log_{k-1} \gamma_n, 1\}}. \tag{69}$$

This concludes the proof. ■

D.3 Proof of the lemma

Finally, we put together results in the previous two sections and prove Lemma 3.3. Denote the lower and upper bounds in the statement of the theorem as L_ε and U_ε, respectively. Then we know that $L_\varepsilon \le \bar{T}_\varepsilon \le U_\varepsilon \le n^{1/2(k-1)}$ with high probability. Note that $\mathbb{P}(E_n) \to 1$ as $n \to \infty$, where

$$E_n := \left\{ \alpha_t = \bar{\alpha}_t \text{ for all } t < \min(T_\varepsilon, n^{1/2(k-1)}) + 1 \text{ and } L_\varepsilon \le \bar{T}_\varepsilon \le U_\varepsilon \right\}. \tag{70}$$

We will show that $T_\varepsilon = \bar{T}_\varepsilon$ on E_n. Assume this is not true, then there are two possibilities:

(a) $T_\varepsilon < \bar{T}_\varepsilon$. Since $\bar{T}_\varepsilon \le n^{1/2(k-1)}$, we know that $\bar{\alpha}_t = \alpha_t$ for all $t \le T_\varepsilon$, which further implies that $|\bar{\alpha}_{T_\varepsilon}| \ge n^\varepsilon$. As a consequence, $T_\varepsilon \ge \bar{T}_\varepsilon$, a contradiction.

(b) $T_\varepsilon > \bar{T}_\varepsilon$. In this case, we know that $\bar{\alpha}_t = \alpha_t$ for all $t < \min(\bar{T}_\varepsilon + 1, n^{1/2(k-1)}) = \bar{T}_\varepsilon + 1$. Therefore, $|\alpha_{\bar{T}_\varepsilon}| \ge n^\varepsilon$, and we know that $T_\varepsilon \le \bar{T}_\varepsilon$, a contradiction.

We thus conclude that $T_\varepsilon = \bar{T}_\varepsilon$ on E_n. Hence, with high probability $L_\varepsilon \le T_\varepsilon \le U_\varepsilon$ as well. This completes the proof of Lemma 3.3.

Appendix E. Proofs of auxiliary lemmas in Appendix D

E.1 Proof of Lemma D.1

We first prove Eq. (54). Note that for any sequence $\{b_t\}_{t\ge 0}$ satisfying $b_{t+1} = q b_t^{k-1}$ for some $q \in \mathbb{R}$, one has

$$q^{1/(k-2)} b_{t+1} = \left(q^{1/(k-2)} b_t \right)^{k-1},$$

thus leading to

$$q^{1/(k-2)} b_t = \left(q^{1/(k-2)} b_0 \right)^{(k-1)^t} \implies b_t = q^{((k-1)^t - 1)/(k-2)} b_0^{(k-1)^t}.$$

Specializing the above equation to $(b_0, q) = (\bar{b}_{0,\Delta}, \gamma_n(1+\Delta)^k)$ and $(b_0, q) = (\underline{b}_{0,\Delta}, \gamma_n(1-\Delta)^k)$ proves Eq. (54).

Next, we prove Eq. (55). The proof of Eq. (56) follows similarly and we omit it for the sake of simplicity. Without loss of generality we may assume $t = 0$. The proof applies without change to positive t. To prove Eq. (55), it suffices to prove the following claim:

$$|\bar{\alpha}_0| \le \bar{b}_{0,\Delta}, \text{ and } |Z_s| \le \frac{\Delta \bar{b}_{s,\Delta} - \delta}{1+\delta}, \ \forall s \in \mathbb{N} \implies |\bar{\alpha}_s| \le \bar{b}_{s,\Delta}, \ \forall s \in \mathbb{N}. \tag{71}$$

We establish the above relationship via induction. For $s = 0$, it holds trivially. Now assume $|\bar{\alpha}_s| \le \bar{b}_{s,\Delta}$, then we know that

$$\begin{aligned} |\bar{\alpha}_{s+1}| &= \gamma_n \bar{\zeta}_s \cdot \left| \bar{\alpha}_s + \bar{b}_s + \bar{c}_s Z_s \right|^{k-1} \le \gamma_n (1+\delta) \cdot (|\bar{\alpha}_s| + \delta + (1+\delta)|Z_s|)^{k-1} \\ &\le \gamma_n (1+\Delta) \cdot \left(\bar{b}_{s,\Delta} + \Delta \bar{b}_{s,\Delta} \right)^{k-1} = \gamma_n (1+\Delta)^k \cdot \bar{b}_{s,\Delta}^{k-1} = \bar{b}_{s+1,\Delta}. \end{aligned}$$

This completes the induction. As a consequence, we obtain that

$$\begin{aligned}
&\mathbb{P}\left(|\bar{\alpha}_{t+s}| \le \bar{b}_{s,\Delta},\ \forall s \in \mathbb{N} \;\middle|\; |\bar{\alpha}_t| \le \bar{b}_{0,\Delta}\right) \\
\ge &\,\mathbb{P}\left(|Z_s| \le \frac{\Delta\bar{b}_{s,\Delta} - \delta}{1+\delta}, \forall s \in \mathbb{N}\right) = \prod_{s=0}^{\infty} \mathbb{P}\left(|Z_s| \le \frac{\Delta\bar{b}_{s,\Delta} - \delta}{1+\delta}\right) \\
= &\prod_{s=0}^{\infty}\left(1 - 2\Phi\left(-\frac{\Delta\bar{b}_{s,\Delta} - \delta}{1+\delta}\right)\right) \ge 1 - 2\sum_{s=0}^{\infty} \Phi\left(-\frac{\Delta\bar{b}_{s,\Delta} - \delta}{1+\delta}\right),
\end{aligned}$$

where the last line follows from the inequality $\prod_{s=0}^{\infty}(1-x_s) \ge 1 - \sum x_s$ for $x_s \in [0,1]$. This completes the proof of the lemma.

E.2 Proof of Proposition D.1

Note that by definition of $\{\bar{\alpha}_t\}_{t\ge 0}$, we have

$$|\bar{\alpha}_{t+1}| \le \gamma_n(1+\delta)\left(|\bar{\alpha}_t| + \delta + (1+\delta)\,|Z_t|\right)^{k-1}.$$

Recall that $C_k = (k-2)^{k-2}/(k-1)^{k-1}$, we next show that as long as

$$|Z_t| \le \frac{1}{1+\delta}\left(\frac{C_k}{\gamma_n(1+\delta)}\right)^{1/(k-2)} - \frac{\delta}{1+\delta} \text{ for all } 0 \le t \le T-1,$$

then we also have

$$|\bar{\alpha}_t| \le \frac{1}{k-2}\left(\frac{C_k}{\gamma_n(1+\delta)}\right)^{1/(k-2)} \text{ for all } 0 \le t \le T.$$

To this end, we use induction. For $t = 0$, we already have $\alpha_0 = 0$, so the above inequality holds automatically. Assume that it is true for all $t \le T-1$, then one has

$$\begin{aligned}
|\bar{\alpha}_{t+1}| &\le \gamma_n(1+\delta)\left(|\bar{\alpha}_t| + \delta + (1+\delta)\,|Z_t|\right)^{k-1} \\
&\le \gamma_n(1+\delta)\left(|\bar{\alpha}_t| + \left(\frac{C_k}{\gamma_n(1+\delta)}\right)^{1/(k-2)}\right)^{k-1} \\
&\le \gamma_n(1+\delta) \times \frac{(k-1)^{k-1}}{(k-2)^{k-1}} \times \left(\frac{C_k}{\gamma_n(1+\delta)}\right)^{(k-1)/(k-2)} \\
&= \frac{1}{k-2}\left(\frac{C_k}{\gamma_n(1+\delta)}\right)^{1/(k-2)} = M(k, \gamma_n, \delta).
\end{aligned}$$

This completes the induction. As a consequence, we deduce that

$$
\begin{aligned}
&\mathbb{P}\left(|\bar{\alpha}_t| \le \frac{1}{k-2}\left(\frac{C_k}{\gamma_n(1+\delta)}\right)^{1/(k-2)} \text{ for all } 0 \le t \le T\right) \\
\ge\, &\mathbb{P}\left(|Z_t| \le \frac{1}{1+\delta}\left(\frac{C_k}{\gamma_n(1+\delta)}\right)^{1/(k-2)} - \frac{\delta}{1+\delta} \text{ for all } 0 \le t \le T-1\right) \\
=\, &\left(1 - 2\Phi\left(\frac{\delta}{1+\delta} - \frac{1}{1+\delta}\left(\frac{C_k}{\gamma_n(1+\delta)}\right)^{1/(k-2)}\right)\right)^T \\
\ge\, &1 - 2T\Phi\left(\frac{\delta}{1+\delta} - \frac{1}{1+\delta}\left(\frac{C_k}{\gamma_n(1+\delta)}\right)^{1/(k-2)}\right) = 1 - 2T\Phi\left(-N(k,\gamma_n,\delta)\right).
\end{aligned}
$$

This completes the proof.

E.3 Proof of Proposition D.2

Take $\Delta = 1$. Consider the sequence $\{\bar{b}_{t,\Delta}\}_{t\ge 0}$ defined via $\bar{b}_{0,\Delta} = M_{k,n} = \log_{k-1} n > 0$, and

$$\bar{b}_{t+1,\Delta} = \gamma_n 2^k \cdot \bar{b}_{t,\Delta}^{k-1}.$$

Then, according to Lemma D.1 we know that

$$\bar{b}_{t,\Delta} = \left(2^k\gamma_n\right)^{((k-1)^t-1)/(k-2)} \cdot M_{k,n}^{(k-1)^t}, \tag{72}$$

and that

$$\mathbb{P}\left(|\bar{\alpha}_t| \le \bar{b}_{t,\Delta},\ \forall t \in \mathbb{N}\right) \overset{(i)}{=} \mathbb{P}\left(|\bar{\alpha}_t| \le \bar{b}_{t,\Delta},\ \forall t \in \mathbb{N} \;\big|\; |\bar{\alpha}_0| \le M_{k,n}\right) \ge\ 1 - 2\sum_{t=0}^{\infty}\Phi\left(-\frac{\bar{b}_{t,\Delta}-\delta}{1+\delta}\right),$$

where (i) is because $\bar{\alpha}_0 = 0$. Since $\gamma_n \gg (\log n)^{-(k-2)/2}$, we then see that $M_{k,n} \gg \max\{\gamma_n^{-1/(k-2)}, 1\}$, hence for large enough n:

$$\frac{\bar{b}_{t,\Delta}-\delta}{1+\delta} \ge \frac{\bar{b}_{t,\Delta}}{2} \text{ for all } t \ge 0.$$

Therefore, for sufficiently large n it holds that

$$
\begin{aligned}
1 - 2\sum_{t=0}^{\infty}\Phi\left(-\frac{\bar{b}_{t,\Delta}-\delta}{1+\delta}\right) &\ge 1 - 2\sum_{t=0}^{\infty}\Phi\left(-\frac{\bar{b}_{t,\Delta}}{2}\right) \ge\ 1 - 2\sum_{t=0}^{\infty}\Phi\left(-\left(2^{k-1}\gamma_n M_{k,n}^{k-2}\right)^t \cdot M_{k,n}\right) \\
&\ge 1 - C\sum_{t=0}^{\infty}\exp\left(-\frac{M_{k,n}^2}{2}\cdot\left(2^{k-1}\gamma_n M_{k,n}^{k-2}\right)^{2t}\right) \\
&\ge 1 - C\sum_{t=0}^{\infty}\exp\left(-\frac{(t+1)M_{k,n}^2}{2}\right) \ge 1 - C\exp\left(-\frac{M_{k,n}^2}{2}\right),
\end{aligned}
$$

where C is a positive numerical constant. Furthermore, we have

$$M_{k,n} = \log_{k-1} n \implies 1 - C\exp\left(-\frac{M_{k,n}^2}{2}\right) \geq 1 - n^{-C}.$$

It then follows that for large enough n, with probability at least $1 - n^{-C}$,

$$\begin{aligned}
&\max_{0\leq t\leq t_n} |\bar{\alpha}_t| \leq \max_{0\leq t\leq t_n} \bar{b}_{t,\Delta} \leq \max_{0\leq t\leq t_n} \Big(\max(2^k\gamma_n, 1)\cdot M_{k,n}\Big)^{(k-1)^t} \\
&\leq \Big(\max(2^k\gamma_n, 1)\cdot M_{k,n}\Big)^{(k-1)^{t_n}} \leq \Big(\max(2^k\gamma_n, 1)\cdot M_{k,n}\Big)^{\left(\frac{\log_{k-1} n}{\max\{\log_{k-1}\gamma_n, 1\}}\right)^{1-\eta}} \\
&= \exp\left(\left(\frac{\log_{k-1} n}{\max\{\log_{k-1}\gamma_n, 1\}}\right)^{1-\eta}\cdot\Big(\log M_{k,n} + \log\Big(\max(2^k\gamma_n, 1)\Big)\Big)\right) \\
&\leq \exp\Big(C_0\Big((\log n)^{1-\eta/2} + (\log n)^{1-\eta}(\log(\max\{\gamma_n, 1\}))^{\eta}\Big)\Big) \overset{(i)}{=} n^{o(1)} \leq \frac{n^{\varepsilon}}{2}
\end{aligned}$$

for sufficiently large n, where (i) is due to our assumption: $\gamma_n = n^{o(1)}$. In the above display, C_0 is another positive numerical constant. This completes the proof of the proposition.

E.4 Proof of Proposition D.3

Note that for any $0 \leq t \leq T-1$, Z_t is independent of $\bar{\alpha}_t + \bar{b}_t$. Therefore, for any $x \geq 0$, with probability $\Phi(-x)$, we have $|Z_t| \geq x$ and $\text{sign}(Z_t) = \text{sign}(\bar{\alpha}_t + \bar{b}_t)$. When this event occurs, we have

$$\left|\bar{\alpha}_t + \bar{b}_t + \bar{c}_t Z_t\right| \geq \bar{c}_t\,|Z_t| \geq (1-\delta)x \implies |\bar{\alpha}_{t+1}| \geq \gamma_n(1-\delta)^k x^{k-1}.$$

Define $E_t := \{|Z_t| < x \text{ or } \text{sign}(Z_t) \neq \text{sign}(\bar{\alpha}_t + \bar{b}_t)\}$, then $\mathbb{P}(E_t) = \mathbb{P}\left(E_t | \cap_{s=0}^{t-1} E_s\right) = 1 - \Phi(-x)$ for all $0 \leq t \leq T-1$. Note that on $(\cap_{t=0}^{T-1} E_t)^c$, we have

$$\max_{0\leq t\leq T} |\bar{\alpha}_t| \geq \gamma_n(1-\delta)^k x^{k-1}. \tag{73}$$

Since $\{\bar{\alpha}_t\}_{t\geq 0}$ is a Markov chain, we obtain that

$$\begin{aligned}
\mathbb{P}\left((\cap_{t=0}^{T-1} E_t)^c\right) &= 1 - \mathbb{P}\left(\cap_{t=0}^{T-1} E_t\right) = 1 - \prod_{t=0}^{T-1}\mathbb{P}\left(E_t | \cap_{s=0}^{t-1} E_s\right) \\
&= 1 - (1-\Phi(-x))^T \geq 1 - \exp(-T\Phi(-x)).
\end{aligned}$$

Now, choosing $x = (m/\gamma_n(1-\delta)^k)^{1/(k-1)}$, it follows that

$$\mathbb{P}\left(\max_{0\leq t\leq T} |\bar{\alpha}_t| \geq m\right) \geq 1 - \exp\left(-T\Phi\left(-\left(\frac{m}{\gamma_n(1-\delta)^k}\right)^{1/(k-1)}\right)\right). \tag{74}$$

This completes the proof.

E.5 Proof of Proposition D.4

Consider the sequence $\{\underline{b}_{t,\Delta}\}_{t\geq 0}$ defined as per Eq. (53) with $\underline{b}_{0,\Delta} = m(\Delta, L)$. Applying Lemma D.1 to $\{\bar{\alpha}_t\}$ and $\{\underline{b}_{t,\Delta}\}$ yields

$$\underline{b}_{t,\Delta} = \Big(\gamma_n(1-\Delta)^k\Big)^{((k-1)^t-1)/(k-2)} \cdot m(\Delta, L)^{(k-1)^t}, \tag{75}$$

and consequently,

$$\begin{aligned} &\mathbb{P}\left(|\bar{\alpha}_{t+t_n}| \geq \underline{b}_{t_n,\Delta} \;\middle|\; |\bar{\alpha}_t| \geq m(\Delta, L)\right) \\ \geq &\mathbb{P}\left(|\bar{\alpha}_{t+s}| \geq \underline{b}_{s,\Delta},\ \forall s \in \mathbb{N} \;\middle|\; |\bar{\alpha}_t| \geq m(\Delta, L)\right) \\ \geq &1 - 2\sum_{s=0}^{\infty} \Phi\left(-\frac{\Delta \underline{b}_{s,\Delta} - \delta}{1+\delta}\right). \end{aligned} \tag{76}$$

Note that by definition $\gamma_n(1-\Delta)^k m(\Delta, L)^{k-2} \geq 1 + \Delta$. Also note that

$$\underline{b}_{t,\Delta} = \Big(\gamma_n(1-\Delta)^k\Big)^{-1/(k-2)} \cdot \Big(m(\Delta, L) \cdot \gamma_n^{1/(k-2)}(1-\Delta)^{k/(k-2)}\Big)^{(k-1)^t},$$

hence $\{\underline{b}_{t,\Delta}\}_{t\geq 0}$ is an increasing sequence, lower bounded by $m(\Delta, L)$. Therefore, for sufficiently large n the last line in Eq. (76) if further no smaller than $1 - 2\sum_{s=0}^{\infty} \Phi(-\Delta \underline{b}_{s,\Delta}/2)$, which by Eq. (75) is equal to

$$\begin{aligned} &1 - 2\sum_{s=0}^{\infty} \Phi\left(-\frac{\Delta}{2}\Big(\gamma_n(1-\Delta)^k\Big)^{((k-1)^s-1)/(k-2)} \cdot m(\Delta, L)^{(k-1)^s}\right) \\ =&1 - 2\sum_{s=0}^{\infty} \Phi\left(-\frac{\Delta}{2}\Big(\gamma_n(1-\Delta)^k m(\Delta, L)^{k-2}\Big)^{((k-1)^s-1)/(k-2)} \cdot m(\Delta, L)\right) \\ \overset{(i)}{\geq}&1 - 2\sum_{s=0}^{\infty} \Phi\left(-\frac{\Delta}{2}(1+\Delta)^{((k-1)^s-1)/(k-2)} \cdot m(\Delta, L)\right) \overset{(ii)}{\geq} 1 - 2\sum_{s=0}^{\infty} \Phi\left(-\frac{\Delta}{2}(1+\Delta)^s m(\Delta, L)\right) \\ \overset{(iii)}{\geq}&1 - C\sum_{s=0}^{\infty} \exp\left(-\frac{\Delta^2}{8}(1+\Delta)^{2s} m(\Delta, L)^2\right) \geq 1 - C\sum_{s=0}^{\infty} \exp\left(-\frac{1+2s\Delta}{8}\Delta^2 m(\Delta, L)^2\right) \\ =&1 - \frac{C\exp(-\Delta^2 m(\Delta, L)^2/8)}{1 - \exp(-\Delta^3 m(\Delta, L)^2/4)} \geq 1 - C\exp\left(-\frac{\Delta^2 m(\Delta, L)^2}{8}\right), \end{aligned}$$

where (i) follows from our choice of $m(\Delta, L)$: $\gamma_n(1-\Delta)^k m(\Delta, L)^{k-2} \geq 1 + \Delta$, (ii) is due to the inequality $(k-1)^s \geq 1 + (k-2)s$, and (iii) follows from the well-known fact regarding Gaussian tail bound: $\Phi(-x) \leq C\exp(-x^2/2)$ for $x \leq 0$, where $C > 0$ is a numerical constant.

It then suffices to show that $\underline{b}_{t_n,\Delta} \geq n^{\varepsilon}$. By direct calculation, we obtain that for a sufficiently large L and n, it holds that

$$\begin{aligned}
\underline{b}_{t_n,\Delta} &= \Big(\gamma_n(1-\Delta)^k\Big)^{((k-1)^{t_n}-1)/(k-2)} \cdot m(\Delta,L)^{(k-1)^{t_n}} \\
&\geq \Big(\gamma_n(1-\Delta)^k m(\Delta,L)^{k-2}\Big)^{((k-1)^{t_n}-1)/(k-2)} \\
&\geq (\gamma_n(1-\Delta)^k m(\Delta,L)^{k-2})^{(k-1)^{t_n-1}} \\
&\geq (\gamma_n(1-\Delta)^k m(\Delta,L)^{k-2})^{(k-1)^{(1+\eta/2)\log_{k-1}(\log_{k-1} n/\max\{\log_{k-1}\gamma_n,1\})}} \\
&\geq (\gamma_n(1-\Delta)^k m(\Delta,L)^{k-2})^{(\log_{k-1} n/\max\{\log_{k-1}\gamma_n,1\})^{(1+\eta/2)}}. \qquad (77)
\end{aligned}$$

By definition of $m(\Delta, L)$, we see that $\gamma_n(1-\Delta)^k m(\Delta,L)^{k-2} \geq C_{\Delta,k,L}\max\{\gamma_n, e\}$, where $C_{\Delta,k,L} > 1$ is a constant that depends only on (Δ, k, L) (this is true if we choose L large enough). For simplicity, define $\omega = \max\{\gamma_n, e\}$. To show that the last line of Eq. (77) is no smaller than n^{ε}, it suffices to prove the following:

$$\begin{aligned}
&\left(\frac{\log_{k-1} n}{\log_{k-1}\omega}\right)^{1+\eta/2} \cdot \log_{k-1}(C_{\Delta,k,L}\,\omega) \geq \varepsilon \log_{k-1} n \\
\Longleftrightarrow\; &\big(\log_{k-1} n\big)^{\eta/2} \cdot \log_{k-1}(C_{\Delta,k,L}\,\omega) \geq \varepsilon \big(\log_{k-1}\omega\big)^{1+\eta/2}.
\end{aligned}$$

This is true for large enough n since $\omega \ll n$ and $C_{\Delta,k,L} > 1$. The proof is complete.

Appendix F. Proof of Theorem 2.2

In this section, we follow the definitions and notations introduced in the proof of Theorem 2.1. Recall from Eq. (17) that $T_\varepsilon = \min\{t \in \mathbb{N}_+ : |\alpha_t| \geq n^{\varepsilon}\}$, where

$$\alpha_t = \lambda_n \langle \boldsymbol{v}, \tilde{\boldsymbol{v}}^{t-1}\rangle^{k-1} = \gamma_n \left(\sqrt{n}\langle \boldsymbol{v}, \tilde{\boldsymbol{v}}^{t-1}\rangle\right)^{k-1}.$$

Part I: before T_ε

We first prove that with probability $1 - o_n(1)$,

$$\|\tilde{\boldsymbol{v}}^{t} - \tilde{\boldsymbol{v}}^{t-1}\|_2 < 1/2$$

simultaneously for all $t \in \{1, 2, \cdots, T_\varepsilon - 1\}$. Invoking Lemma 3.3, we see that for $\varepsilon = (6k-11)/12(k-1)$, with probability $1 - o_n(1)$ we have $T_\varepsilon + 10 < n^{1/10(k-1)}$. Recall that $\boldsymbol{w}_{i_1,i_2,\cdots,i_{k-1}} \overset{d}{=} \mathbf{g}_t \overset{d}{=} \mathsf{N}(\mathbf{0}, \boldsymbol{I}_n)$. Employing standard Gaussian concentration inequalities, we see that with probability at least $1 - o_n(1)$, it holds that

$$\big\|\boldsymbol{w}_{i_1,i_2,\cdots,i_{k-1}}\big\|_2 \leq \sqrt{n\log n}, \qquad \big|\|\mathbf{g}_t\|_2 - \sqrt{n}\big| \leq \log n, \qquad \big\|\mathbf{g}_t - \mathbf{g}_{t+1}\big\|_2 \geq \sqrt{n} \qquad (78)$$

for all $t \in \{0, 1, \cdots, \lceil n^{1/10(k-1)}\rceil\}$ and $(i_1, \cdots, i_{k-1}) \in \boldsymbol{H}_t$. In addition, Lemma 3.2 implies that with probability $1 - o_n(1)$ we have

$$\left|\sqrt{1 - \|\tilde{\boldsymbol{v}}_{\|}^{t-1}\|_2^{2k-2}} - 1\right| = |c_t - 1| \leq C n^{-5/6}$$

for an absolute constant $C > 0$ and all $t \in \{0, 1, \cdots, T_\varepsilon - 1\}$. Observe that for $t \in \{1, \cdots, \lceil n^{1/10(k-1)} \rceil\}$,

$$\sum_{(i_1,\cdots,i_{k-1}) \in \boldsymbol{H}_{t-2}} |\beta_{i_1,i_2,\cdots,i_{k-1}}|^2 + c_t^2 = 1, \qquad \sum_{(i_1,\cdots,i_{k-1}) \in \boldsymbol{H}_{t-3}} |\beta_{i_1,i_2,\cdots,i_{k-1}}|^2 + c_{t-1}^2 = 1,$$

hence using Cauchy-Schwartz inequality we get

$$\sum_{(i_1,\cdots,i_{k-1}) \in \boldsymbol{H}_{t-2}} |\beta_{i_1,i_2,\cdots,i_{k-1}}| \le C n^{-5/12+1/20}, \qquad \sum_{(i_1,\cdots,i_{k-1}) \in \boldsymbol{H}_{t-3}} |\beta_{i_1,i_2,\cdots,i_{k-1}}| \le C n^{-5/12+1/20}. \tag{79}$$

Plugging these bounds into Eq. (8) and applying the triangle inequality, we see that with probability $1 - o_n(1)$

$$\begin{aligned}
&\|\boldsymbol{v}^t\|_2 \le n^{1/2} + \log n + n^{(6k-11)/12(k-1)} + C n^{-11/30} \sqrt{n \log n} + C n^{-5/6} \cdot (n^{1/2} + \log n), \\
&\|\boldsymbol{v}^t\|_2 \ge n^{1/2} - \log n - n^{(6k-11)/12(k-1)} - C n^{-11/30} \sqrt{n \log n} - C n^{-5/6} \cdot (n^{1/2} + \log n).
\end{aligned}$$

for all $t \in \{0, 1, \cdots, T_\varepsilon - 1\}$. From the above upper bound, we see that with high probability

$$\left| \|\boldsymbol{v}^t\|_2 - n^{1/2} \right| \le 2 n^{(6k-11)/12(k-1)}$$

for all $t \in \{0, 1, \cdots, T_\varepsilon - 1\}$.

We can also employ Eq. (78) and Eq. (79) to lower bound $\|\boldsymbol{v}^t - \boldsymbol{v}^{t-1}\|_2$. Invoking the triangle inequality, we get:

$$\begin{aligned}
&\left\|\boldsymbol{v}^t - \boldsymbol{v}^{t-1}\right\|_2 \\
\ge & \|\mathbf{g}_t - \mathbf{g}_{t-1}\|_2 - |\alpha_t| - |\alpha_{t-1}| - |1 - c_t| \cdot \|\mathbf{g}_t\|_2 - |1 - c_{t-1}| \cdot \|\mathbf{g}_{t-1}\|_2 - \\
& \sum_{(i_1,\cdots,i_{k-1}) \in \boldsymbol{H}_{t-2}} |\beta_{i_1,i_2,\cdots,i_{k-1}}| \cdot \|\boldsymbol{w}_{i_1,i_2,\cdots,i_{k-1}}\|_2 - \sum_{(i_1,\cdots,i_{k-1}) \in \boldsymbol{H}_{t-3}} |\beta_{i_1,i_2,\cdots,i_{k-1}}| \cdot \|\boldsymbol{w}_{i_1,i_2,\cdots,i_{k-1}}\|_2.
\end{aligned}$$

By definition of T_ε we have $\max\{|\alpha_t|, |\alpha_{t-1}|\} \le n^\varepsilon = n^{(6k-11)/12(k-1)}$. By Lemma 3.2 we have $\max\{|c_t - 1|, |c_{t-1} - 1|\} \le C n^{-5/6}$. Putting together the above equations, we get

$$\begin{aligned}
&\left\|\boldsymbol{v}^t - \boldsymbol{v}^{t-1}\right\|_2 \\
\ge & \|\mathbf{g}_t - \mathbf{g}_{t-1}\|_2 - 2n^{(6k-11)/12(k-1)} - C n^{-5/6} \cdot \left(n^{1/2} + \log n\right) - C n^{-11/30} \sqrt{n \log n} \\
\ge & n^{1/2} - 3n^{(6k-11)/12(k-1)},
\end{aligned}$$

which holds with high probability for all $t \in \{0, 1, \cdots, T_\varepsilon - 1\}$ for a sufficiently large n. We can then use this to give a high-probability lower bound for $\|\tilde{\boldsymbol{v}}^t - \tilde{\boldsymbol{v}}^{t-1}\|_2$. Leveraging the triangle inequality, we obtain that for a large enough n, with probability $1 - o_n(1)$

$$\begin{aligned}
\left\|\tilde{\boldsymbol{v}}^t - \tilde{\boldsymbol{v}}^{t-1}\right\|_2 &= \left\| \frac{\boldsymbol{v}^t}{\|\boldsymbol{v}^t\|_2} - \frac{\boldsymbol{v}^{t-1}}{\|\boldsymbol{v}^{t-1}\|_2} \right\|_2 \\
&\ge \left\| \frac{\boldsymbol{v}^t}{\sqrt{n}} - \frac{\boldsymbol{v}^{t-1}}{\sqrt{n}} \right\|_2 - \left| \frac{1}{\|\boldsymbol{v}^t\|_2} - \frac{1}{\sqrt{n}} \right| \cdot \|\boldsymbol{v}^t\|_2 - \left| \frac{1}{\|\boldsymbol{v}^{t-1}\|_2} - \frac{1}{\sqrt{n}} \right| \cdot \|\boldsymbol{v}^{t-1}\|_2 \\
&\ge 1 - 20 n^{-5/12(k-1)}
\end{aligned}$$

for all $t \in \{0, 1, \cdots, T_\varepsilon - 1\}$. Therefore, with probability $1 - o_n(1)$ we have $\|\tilde{\boldsymbol{v}}^t - \tilde{\boldsymbol{v}}^{t-1}\|_2 < 1/2$ for all $t \in \{1, 2, \cdots, T_\varepsilon - 1\}$. This completes the proof for Part I. The takeaway message is that with high probability $T_{\mathsf{stop}} \geq T_\varepsilon + 1$.

Part II: after T_ε

In the second part of the proof, we show with high probability $|\langle \tilde{\boldsymbol{v}}^{T_\varepsilon+1}, \tilde{\boldsymbol{v}}^{T_\varepsilon+2}\rangle| \geq 1/2$, hence $T_{\mathsf{stop}} \leq T_\varepsilon + 4$.

We first prove with probability $1 - o_n(1)$ we have $|\langle \boldsymbol{v}, \tilde{\boldsymbol{v}}^{T_\varepsilon+i}\rangle| \geq 1 - \delta$ for all $i \in \{1, 2, 3, 4\}$. According to Eq. (26), we see that with probability $1 - o_n(1)$ we have $|\langle \tilde{\boldsymbol{v}}^{T_\varepsilon+1}, \boldsymbol{v}\rangle| \geq 1 - Cn^{-1/24}$. This already proves the desired result for $i = 1$.

Once again we apply standard Gaussian concentration inequalities, and obtain that with probability $1 - o_n(1)$

$$\left|\langle \boldsymbol{w}_{i_1, i_2, \cdots, i_{k-1}}, \boldsymbol{v}\rangle\right| \leq \log n, \qquad |\langle \mathbf{g}_t, \boldsymbol{v}\rangle| \leq \log n \tag{80}$$

for all $t \in \{0, 1, \cdots, \lceil n^{1/10(k-1)}\rceil\}$ and $(i_1, \cdots, i_{k-1}) \in \boldsymbol{H}_t$.

Applying Eq. (78) and the triangle inequality to Eq. (8), we get

$$\begin{aligned}
\|\boldsymbol{v}^{T_\varepsilon+2}\|_2 \geq& |\alpha_{T_\varepsilon+2}| - \sum_{(i_1, \cdots, i_{k-1}) \in \boldsymbol{H}_{T_\varepsilon}} |\beta^{(T_\varepsilon+1)}_{i_1, i_2, \cdots, i_{k-1}}| \cdot \|\boldsymbol{w}_{i_1, i_2, \cdots, i_{k-1}}\|_2 - c_{T_\varepsilon+2}\|\mathbf{g}_{T_\varepsilon+2}\|_2 \\
\geq& \lambda_n (1 - Cn^{-1/24})^{k-1} - n^{1/10} \cdot \sqrt{n \log n},
\end{aligned}$$

where to get the second lower bound we use the equality $|\alpha_{T_\varepsilon+2}| = \lambda_n \cdot |\langle \boldsymbol{v}, \tilde{\boldsymbol{v}}^{T_\varepsilon+1}\rangle^{k-1}|$. Similarly, we obtain $\|\boldsymbol{v}^{T_\varepsilon+2}\|_2 \leq \lambda_n + n^{1/10} \cdot \sqrt{n \log n}$. Therefore,

$$\begin{aligned}
&|\langle \tilde{\boldsymbol{v}}^{T_\varepsilon+2}, \boldsymbol{v}\rangle| = \frac{|\langle \boldsymbol{v}^{T_\varepsilon+2}, \boldsymbol{v}\rangle|}{\|\boldsymbol{v}^{T_\varepsilon+2}\|_2} \\
=& \frac{1}{\|\boldsymbol{v}^{T_\varepsilon+2}\|_2} \cdot \left|\alpha_{T_\varepsilon+2} + \sum_{(i_1, \cdots, i_{k-1}) \in \boldsymbol{H}_{T_\varepsilon}} \beta^{(T_\varepsilon+1)}_{i_1, i_2, \cdots, i_{k-1}} \langle \boldsymbol{w}_{i_1, i_2, \cdots, i_{k-1}}, \boldsymbol{v}\rangle + c_{T_\varepsilon+2}\langle \mathbf{g}_{T_\varepsilon+2}, \boldsymbol{v}\rangle\right| \\
\geq& \frac{1}{\lambda_n + n^{1/10} \cdot \sqrt{n \log n}} \cdot \left(\lambda_n (1 - Cn^{-1/24})^{k-1} - n^{1/10} \log n\right) \geq 1 - o_n(1)
\end{aligned}$$

with probability $1 - o_n(1)$. Following the same route, we are able to conclude that with high probability $|\langle \tilde{\boldsymbol{v}}^{T_\varepsilon+3}, \boldsymbol{v}\rangle| = 1 - o_n(1)$ and $|\langle \tilde{\boldsymbol{v}}^{T_\varepsilon+3}, \boldsymbol{v}\rangle| = 1 - o_n(1)$. As a direct consequence, we see that with high probability $\min_{i \in \{1,2,3,4\}} |\langle \tilde{\boldsymbol{v}}^{T_\varepsilon+i}, \boldsymbol{v}\rangle| \geq 1 - \delta$.

Finally, we show $|\langle \tilde{\boldsymbol{v}}^{T_\varepsilon+1}, \tilde{\boldsymbol{v}}^{T_\varepsilon+2}\rangle| > 1/2$. Let $s = \operatorname{sign}(\langle \boldsymbol{v}, \tilde{\boldsymbol{v}}^{T_\varepsilon+1}\rangle)$, then

$$|\langle \tilde{\boldsymbol{v}}^{T_\varepsilon+1}, \tilde{\boldsymbol{v}}^{T_\varepsilon+2}\rangle| = |\langle \tilde{\boldsymbol{v}}^{T_\varepsilon+1}, \tilde{\boldsymbol{v}}^{T_\varepsilon+2} - s\boldsymbol{v}\rangle + s\langle \boldsymbol{v}, \tilde{\boldsymbol{v}}^{T_\varepsilon+1}\rangle| \geq |\langle \boldsymbol{v}, \tilde{\boldsymbol{v}}^{T_\varepsilon+1}\rangle| - \|\tilde{\boldsymbol{v}}^{T_\varepsilon+2} - s\boldsymbol{v}\|_2. \tag{81}$$

With high probability

$$\|\tilde{\boldsymbol{v}}^{T_\varepsilon+2} - s\boldsymbol{v}\|_2^2 = 2 - 2s\langle \boldsymbol{v}, \tilde{\boldsymbol{v}}^{T_\varepsilon+2}\rangle = 2 - 2|\langle \boldsymbol{v}, \tilde{\boldsymbol{v}}^{T_\varepsilon+2}\rangle| = o_n(1).$$

Plugging the above upper bound into Eq. (81), we get

$$|\langle \tilde{\boldsymbol{v}}^{T_\varepsilon+1}, \tilde{\boldsymbol{v}}^{T_\varepsilon+2}\rangle| \geq 1 - o_n(1),$$

which further implies that with probability $1 - o_n(1)$ the inner product $|\langle \tilde{\boldsymbol{v}}^{T_\varepsilon+1}, \tilde{\boldsymbol{v}}^{T_\varepsilon+2} \rangle|$ is larger than $1/2$. Hence, $T_{\mathsf{stop}} \le T_\varepsilon + 4$. Recall that we have proved $\min_{i\in\{1,2,3,4\}} |\langle \tilde{\boldsymbol{v}}^{T_\varepsilon+i}, \boldsymbol{v} \rangle| \ge 1 - \delta$ with high probability, which further implies that with high probability $|\langle \boldsymbol{v}, \tilde{\boldsymbol{v}}^{T_{\mathsf{stop}}} \rangle| \ge 1 - \delta$. The proof is complete.

Appendix G. Additional experiments

G.1 Comparing alignment and the polynomial recurrence process

We collect in this section additional experiments for Section 4.1. The basic setup is the same as that presented in the main text. Throughout the experiment, we fix $\lambda_n = n^{(k-1)/2}$, and use different combinations of (n, k). As before, we compare the marginal distributions of $\{\alpha_t\}_{t\ge 0}$ and $\{X_t\}_{t\ge 0}$ by comparing the histograms of their empirical marginal distributions, generated from 1000 independent experiments. Observing the simulation outcomes, we see that they all match well.

Setting I: $\mathbf{n = 100, k = 3}$

Simulation results are plotted as Figure 5.

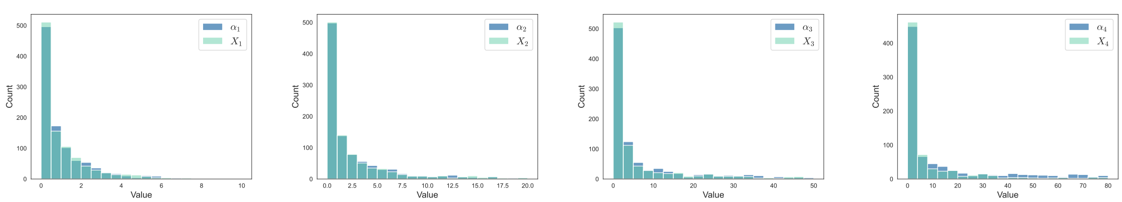

Figure 5: Comparison of the marginal distributions between α_t and X_t, for $t \in \{1, 2, 3, 4\}$ from left to right. Here, we set $n = 100$, $k = 3$, $\lambda_n = n^{(k-1)/2}$, and run tensor power iteration from random initialization on independent datasets for 1000 times.

Setting II: $\mathbf{n = 500, k = 3}$

Simulation results are plotted as Figure 6.

Setting III: $\mathbf{n = 1000, k = 3}$

Simulation results are plotted as Figure 7.

Setting IV: $\mathbf{n = 100, k = 4}$

Simulation results are plotted as Figure 8.

Setting IV: $\mathbf{n = 200, k = 4}$

Simulation results are plotted as Figure 9.

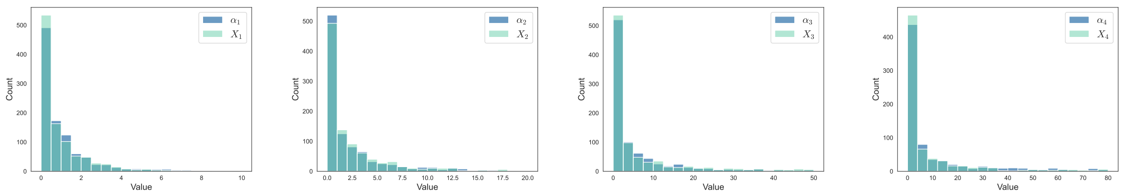

Figure 6: Comparison of the marginal distributions between α_t and X_t, for $t \in \{1, 2, 3, 4\}$ from left to right. Here, we set $n = 500$, $k = 3$, $\lambda_n = n^{(k-1)/2}$, and run tensor power iteration from random initialization on independent datasets for 1000 times.

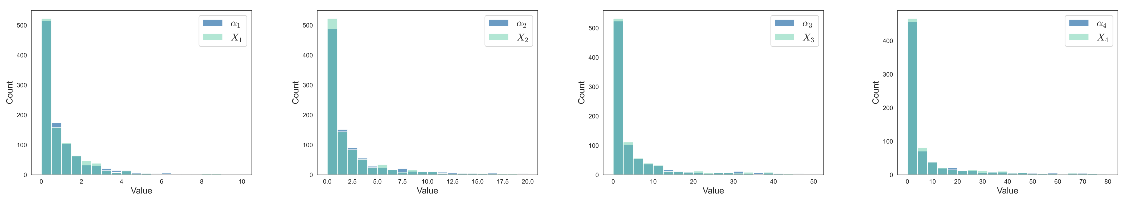

Figure 7: Comparison of the marginal distributions between α_t and X_t, for $t \in \{1, 2, 3, 4\}$ from left to right. Here, we set $n = 1000$, $k = 3$, $\lambda_n = n^{(k-1)/2}$, and run tensor power iteration from random initialization on independent datasets for 1000 times.

G.2 Experiments with different stopping thresholds

In this section, we adjust the stopping threshold value, and test the proposed stopping criterion under these values. To be specific, we consider here thresholds 0.3 and 0.7, and conduct the corresponding experiments. Counterparts of Figure 4 are presented as Figures 10 and 11. These figures suggest that the effectiveness of the stopping rule is not sensitive to the choice of the stopping threshold.

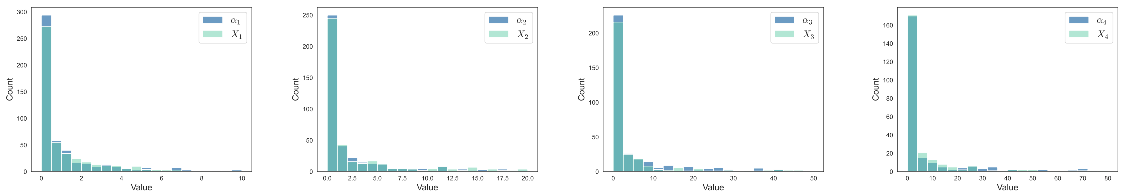

Figure 8: Comparison of the marginal distributions between α_t and X_t, for $t \in \{1, 2, 3, 4\}$ from left to right. Here, we set $n = 100$, $k = 4$, $\lambda_n = n^{(k-1)/2}$, and run tensor power iteration from random initialization on independent datasets for 1000 times.

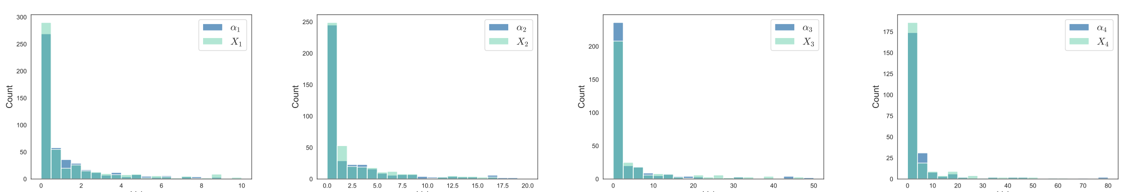

Figure 9: Comparison of the marginal distributions between α_t and X_t, for $t \in \{1, 2, 3, 4\}$ from left to right. Here, we set $n = 200$, $k = 4$, $\lambda_n = n^{(k-1)/2}$, and run tensor power iteration from random initialization on independent datasets for 1000 times.

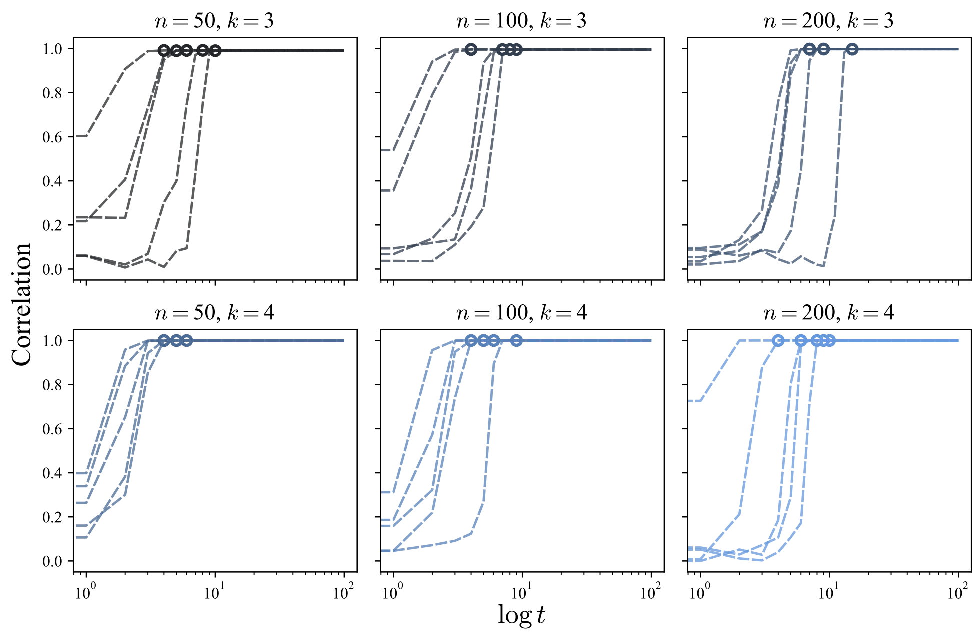

Figure 10: Illustration of the effectiveness of the stopping rule with stopping threshold 0.3.

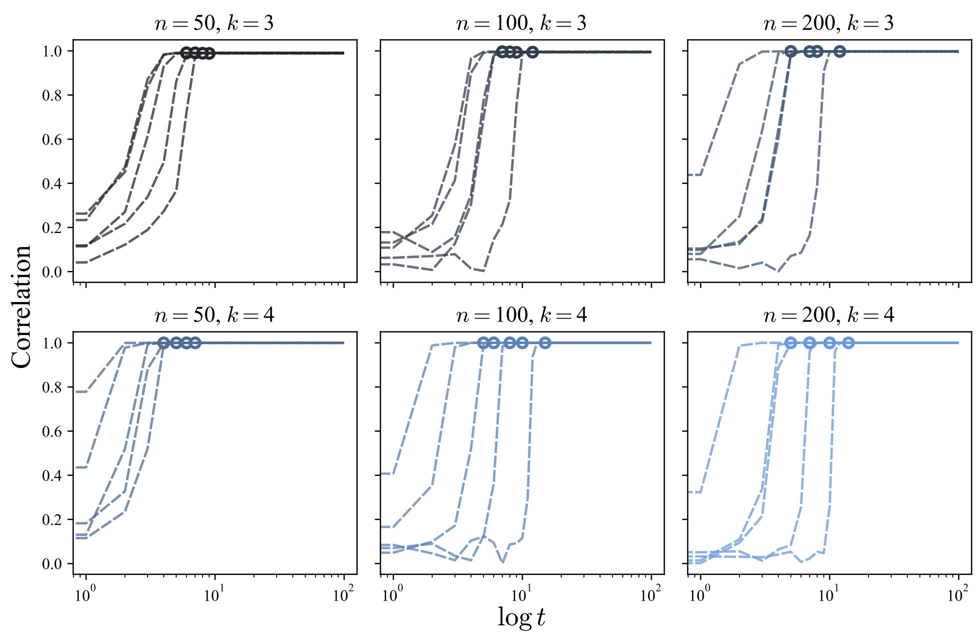

Figure 11: Illustration of the effectiveness of the stopping rule with stopping threshold 0.7.