Lower Bounds for the Convergence of Tensor Power Iteration on Random Overcomplete Models

Yuchen Wu WUYC14@STANFORD.EDU and **Kangjie Zhou** KANGJIE@STANFORD.EDU
Stanford University

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

Tensor decomposition serves as a powerful primitive in statistics and machine learning, and has numerous applications in problems such as learning latent variable models or mixture of Gaussians. In this paper, we focus on using power iteration to decompose an overcomplete random tensor. Past work studying the properties of tensor power iteration either requires a non-trivial data-independent initialization, or is restricted to the undercomplete regime. Moreover, several papers implicitly suggest that logarithmically many iterations (in terms of the input dimension) are sufficient for the power method to recover one of the tensor components.

Here we present a novel analysis of the dynamics of tensor power iteration from random initialization in the overcomplete regime, where the tensor rank is much greater than its dimension. Surprisingly, we show that polynomially many steps are necessary for convergence of tensor power iteration to any of the true component, which refutes the previous conjecture. On the other hand, our numerical experiments suggest that tensor power iteration successfully recovers tensor components for a broad range of parameters in polynomial time. To further complement our empirical evidence, we prove that a popular objective function for tensor decomposition is strictly increasing along the power iteration path.

Our proof is based on the Gaussian conditioning technique, which has been applied to analyze the approximate message passing (AMP) algorithm. The major ingredient of our argument is a conditioning lemma that allows us to generalize AMP-type analysis to non-proportional limit and polynomially many iterations of the power method.

Keywords: Tensor decomposition, power method, approximate message passing

1. Introduction

Tensors of order m are multidimensional arrays with m indices, with $m = 1$ corresponding to vectors and $m = 2$ corresponding to matrices. The notion of *rank* naturally generalizes from matrices to tensors: An m-th order tensor $\mathbf{T} \in (\mathbb{R}^d)^{\otimes m}$ is said to be rank-1 if it can be written as

$$\mathbf{T} = v_1 \otimes \cdots \otimes v_m \iff \mathbf{T}(i_1, \cdots, i_m) = v_1(i_1) \cdots v_m(i_m),$$

where $v_1, \cdots, v_m \in \mathbb{R}^d$. Past results imply that any tensor can be expressed as the sum of rank-1 tensors (Kiers, 2000; Carroll and Chang, 1970). Namely, given $\mathbf{T} \in (\mathbb{R}^d)^{\otimes m}$ we can find vectors $\{v_i^{(j)}\}_{i\in[m],\ j\in[k]}$ such that

$$\mathbf{T} = \sum_{j=1}^{k} v_1^{(j)} \otimes \cdots \otimes v_m^{(j)}.$$

The above decomposition is referred to as *tensor decomposition*, and the (CP) rank of a tensor is defined as the minimum number of rank-1 tensors required in such decomposition. Unlike matrix decomposition, tensor decomposition with $m \geq 3$ is in many cases unique (Kruskal, 1977). This is often true even in the overcomplete case, where the rank of the tensor is much larger than its ambient dimension. The uniqueness of tensor decomposition makes its application suitable in many practical settings, which we discuss below.

Tensor decomposition serves as a powerful primitive in statistics and machine learning, especially for algorithms that leverage *the method of moments* (Pearson, 1894) to learn model parameters. Applications of tensor decomposition include dictionary learning (Barak et al., 2015; Ma et al., 2016; Schramm and Steurer, 2017), Gaussian mixture models (Anandkumar et al., 2014a; Ge et al., 2015; Hsu and Kakade, 2013), independent component analysis (De Lathauwer et al., 2007; Comon and Jutten, 2010), and learning two-layer neural networks (Novikov et al., 2015; Mondelli and Montanari, 2019). Despite the fact that tensor decomposition is NP-hard in the worst case (Hillar and Lim, 2013), researchers have designed polynomial-time algorithms that successfully approximate the tensor components under natural distributional assumptions. Exemplary algorithms of this kind include the classical Jennrich's algorithm (Harshman et al., 1970; De Lathauwer et al., 1996), algebraic methods (De Lathauwer, 2006; De Lathauwer et al., 2007), iterative methods (Zhang and Golub, 2001; Anandkumar et al., 2014a,b, 2015, 2017; Kileel and Pereira, 2019; Kileel et al., 2021), sum-of-squares (SOS) algorithms (Hopkins et al., 2015; Barak et al., 2015; Ge and Ma, 2015; Ma et al., 2016) and their spectral analogues (Hopkins et al., 2016; Schramm and Steurer, 2017; Hopkins et al., 2019; Ding et al., 2022).

SOS algorithms and their spectral counterparts provably achieve strong guarantees of recovering tensor components, and can be implemented in polynomial time. However, they are often computationally prohibitive on large-scale problems due to the high-degree polynomial running time. Therefore, in practice it is often more preferable to resort to simple iterative algorithms (Celentano et al., 2020; Montanari and Wu, 2022b), such as gradient descent and its variants. These algorithms are computationally efficient in terms of both runtime and memory, and are typically easy to implement. In the case of tensors, popular iterative algorithms include tensor power iteration (Anandkumar et al., 2017), gradient descent on non-convex losses (Ge and Ma, 2017), and alternating minimization (Anandkumar et al., 2014b).

We focus in this paper on the tensor power iteration method, which can be regarded as a generalization of matrix power iteration. This method can also be viewed as gradient ascent on a polynomial objective function with infinite step size (see Eq. (2) for a formal definition). However, unlike the matrix case where the convergence properties are well understood theoretically, in the tensor case the dynamics of power iteration still remains mysterious due to non-convexity of the corresponding optimization problem. To unveil the mystery behind tensor power iteration, the present work proposes to study its asymptotic behavior on decomposing a random fourth order symmetric tensor

$$\mathbf{T} = \sum_{i=1}^{k} a_i \otimes a_i \otimes a_i \otimes a_i,\ a_i \in \mathbb{R}^d,\ \forall i \in [k]$$

in the overcomplete regime $k \gg d$, where we assume $a_i \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0, I_d/d)$. We note that this is a well-studied model in the literature, while its properties are not yet fully understood. Given the entries of $\mathbf{T}$, our goal is to recover one or all of the tensor components $\{a_i\}_{i\le k}$, up to potential sign flips.

We denote by $A \in \mathbb{R}^{k\times d}$ the matrix whose i-th row is a_i^{T}. Initialized at $x_0 \in \mathbb{S}^{d-1}(\sqrt{d})$ that is independent of the tensor components $\{a_i\}_{i\le k}$, tensor power iteration is defined recursively as follows:

$$x_t = \frac{\sqrt{d}\,\mathbf{T}(I, x_{t-1}, x_{t-1}, x_{t-1})}{\|\mathbf{T}(I, x_{t-1}, x_{t-1}, x_{t-1})\|_2},\ t \ge 1, \tag{1}$$

where

$$\mathbf{T}(I, x, x, x) := \sum_{i,j,l\in[d]} x(i)x(j)x(l)\mathbf{T}(:, i, j, l).$$

For $x \in \mathbb{S}^{d-1}(\sqrt{d})$, we introduce the following polynomial objective function:

$$\mathcal{S}(x) = \sum_{i,j,k,l\in[d]} \mathbf{T}(i,j,k,l)x(i)x(j)x(k)x(l) = \sum_{i=1}^{k}\langle a_i, x\rangle^4 = \|Ax\|_4^4. \tag{2}$$

Notice that Eq. (1) can be reformulated as $x_t = \sqrt{d} \cdot \nabla \mathcal{S}(x_{t-1}) / \|\nabla \mathcal{S}(x_{t-1})\|_2$, i.e., tensor power iteration can be regarded as gradient ascent on $\mathcal{S}$ with infinite learning rate.

In the undercomplete regime where $k \leq d$, if the components $\{a_i\}_{i \leq k}$ are orthogonal to each other, then $\mathcal{S}$ has only $2k$ local maximizers that are close to $\{\pm\sqrt{d}a_i\}_{i \leq k}$. In this case, tensor power iteration provably converges to one of the $\pm\sqrt{d}a_i$'s (Anandkumar et al., 2014a). Indeed, tensors with linearly independent components can be orthogonalized, thus suggesting the existence of efficient algorithms in the undercomplete regime.

Things become far more challenging in the overcomplete regime where k is much greater than d. When $k \ll d^2$, it is known that any global maximizer of $\mathcal{S}$ must be close to one of the $\pm\sqrt{d}a_i$'s. However, algebraic geometry techniques show that $\mathcal{S}$ has exponentially many other critical points (Cartwright and Sturmfels, 2013). Further, if $k \gg d^2$, then $\mathcal{S}(x)$ concentrates tightly around its expectation, uniformly for all $x \in \mathbb{S}^{d-1}(\sqrt{d})$, thus making it hard to identify the tensor components from the information encoded in $\mathcal{S}$. As far as we know, there is no polynomial-time algorithm known for tensor decomposition within this regime.

We thus focus on the regime $d \ll k \ll d^2$, where algebraic methods and SOS-based algorithms are proven to succeed (Ge and Ma, 2015; Ma et al., 2016; Hopkins et al., 2016; Bhaskara et al., 2019; Ding et al., 2022), and numerical experiments indicate that the performance of randomly initialized power iteration matches that of these methods (see Fig. 1 for details). However, from a theoretical standpoint, the dynamics of tensor power iteration in the overcomplete regime still remain elusive. A reasonable first step towards solving this puzzle would be to understand how many iterations are necessary for power method to find one of the true components.

1.1. Main results

We hereby give a partial answer to the above question. In particular, we establish several new results on the behavior of tensor power iteration in the overcomplete regime. Our first theorem states that randomly initialized tensor power iteration requires at least polynomially many steps to converge to a true component:

Theorem 1.1 (Slow convergence from random start, informal, see Theorem 3.1) *Assume that k, d are large enough, and that $k \asymp d^c$ for some $c \in (3/2, 2)$. Then, there exists some $\eta > 0$ that only depends on c, such that with high probability the following happens: Tensor power iteration from random initialization fails to identify any true component of $\mathbf{T}$ within d^η steps.*

Related work. Let us pause here to make some comparisons between our result and prior work on the same model: The seminal paper Anandkumar et al. (2015) shows that tensor power iteration with an SVD-based initialization converges in $O(\log d)$ steps for $k = O(d)$. In Anandkumar et al. (2017), the authors prove that tensor power method successfully recovers one of the a_i's in $O(\log \log d)$ iterations, given that its initialization has non-trivial correlations with the true components. As a comparison, our Theorem 1.1 shows that the $O(\log d)$ bound on the number of iterations does not hold for randomly-initialized power iteration, and establishes that polynomially many steps are necessary for convergence in the overcomplete regime. To the best of our knowledge, this is the first result that provides a lower bound on the computational complexity of tensor power iteration. From a more fundamental point of view, we also show that tensor power iterates are "trapped" in a small neighborhood around its initialization for polynomially many steps.

Although the power method fails to converge in logarithmic many steps as conjectured, we still believe that it will correctly learn one of the tensor components when $k \ll d^2$ within polynomial time. We present numerical evidence as Fig. 1 in Section 4 to support our claim. Establishing a rigorous positive result for tensor power iteration in this regime is challenging, and we leave it as an interesting open question for future work. As an alternative, we present here a weaker result suggesting the correctness of tensor power iteration.

Theorem 1.2 (Increasing objective function, informal, see Theorem 3.2) *For $d^{3/2} \ll k \ll d^2$, starting from random initialization, with high probability the objective function $\mathcal{S}$ is strictly increasing along the power iteration path up to finitely many steps.*

According to Ge and Ma (2017), the tensor components are the only local maximizers of $\mathcal{S}$ on a superlevel set that is slightly better than random initialization. Their result, together with Theorem 1.2 suggest that, in order to prove convergence of tensor power iteration, it suffices to show that the objective function $\mathcal{S}$ eventually surpasses a small threshold determined in Ge and Ma (2017).

Proof technique. Our proof is based on the Gaussian conditioning technique, and is similar to the analysis of the Approximate Message Passing (AMP) algorithm (Bayati and Montanari, 2011). The majority of prior AMP theory can only accommodate a constant number of iterations (i.e., the number of iterations does not grow with the input size) and proportional asymptotics (in our case, this corresponds to assuming $k/d \to \delta \in (0, \infty)$). Rush and Venkataramanan (2018) moves beyond the constant regime and extends Gaussian conditioning analysis to $O(\log d/\log\log d)$ many steps. However, their results fall short of validity when targeting for polynomially many iterations, which is essential in our context. More recently, Li and Wei (2022) develops a non-asymptotic framework that enables the analysis of AMP up to $O(n/\mathrm{polylog}(n))$ many iterations, while they focus exclusively on symmetric spiked model and require nontrivial initialization.

The technical innovation in this paper is that we successfully apply the Gaussian conditioning scheme to analyze the tensor power iteration up to polynomially many steps. Indeed, our conditioning lemma (Lemma 2.2) gives an exact non-asymptotic characterization of the power iterates, thus allowing for precisely tracking the values of the objective function along the iteration path. It is noteworthy that the same argument can be used to prove that polynomially many power iterates are necessary for general even-order tensors (see Remark 3.1). To the best of our knowledge, this is the first result that generalizes AMP-type analysis to non-proportional asymptotics and polynomially many iterations. From a technical perspective, we believe our results will help to push forward the development of AMP theory and enrich the toolbox for theoretical analysis of general iterative algorithms.

Organization. The rest of this paper is organized as follows. Section 2 introduces some preliminaries regarding power iteration and Gaussian conditioning technique. In Section 3 we state formally our main theorems and sketch their proofs, with all technical details deferred to the appendices. Section 4 provides some useful numerical experiments that support our theoretical results. We provide in Section 5 several concluding remarks and discuss possible future directions.

2. Preliminaries

2.1. Notation

For $x \in \mathbb{R}^d$ and $S \in \mathbb{R}^{d\times k}$, we denote by $\Pi_S(x)$ the projection of x onto the column space of S, and let $\Pi_S^{\perp}(x) = x - \Pi_S(x)$. For two vectors u and v of the same dimension, we use $\langle u, v\rangle$ to represent their inner product, and denote by $\|u\|_p$ the ℓ^p-norm of u for $p \geq 1$. We use the notation $\odot$ to represent the Hadamard product of vectors and matrices. Furthermore, for $x \in \mathbb{R}^d$ we define $x^k = x \odot x \cdots \odot x$, the Hadamard product of k copies of x. For a matrix X, we denote by $X^\dagger$ the pseudoinverse of X. We denote by e_k the k-th standard basis in any Euclidean space.

We denote by $\mathbb{S}^{d-1}(r)$ the sphere in $\mathbb{R}^d$ centered at the origin with radius r. For random variables X, Y, we write $X \perp Y$ if X is independent of Y. For a collection of random variables $\{X_i\}_{i\in I}$, we use $\sigma(\{X_i\}_{i\in I})$ to represent the σ-algebra generated by these random variables.

For $n \in \mathbb{N}_+$, we define the set $[n] := \{1, 2, \cdots, n\}$. For two sequences of positive numbers $\{a_n\}_{n\in\mathbb{N}_+}$, $\{b_n\}_{n\in\mathbb{N}_+}$, we say $a_n \ll b_n$ if $a_n/b_n \to 0$ as $n \to \infty$, and $a_n \asymp b_n$ if $a_n = O(b_n)$ and $b_n = O(a_n)$. For

$\{c_n\}_{n\in\mathbb{N}_+} \subseteq \mathbb{R}$, we say $c_n = o_n(1)$ if $c_n \to 0$ as $n \to \infty$. For a sequence of events $\{E_n\}_{n\in\mathbb{N}_+}$, we say that E_n happens with high probability if $\mathbb{P}(E_n) = 1 - o_n(1)$.

We reserve O_P and o_P as the standard big-O/small-o in probability notations: For a set of random variables $\{X_n\}_{n\geq 1}$ and a sequence of positive numbers $\{a_n\}_{n\geq 1}$, we say $X_n = o_P(a_n)$ if and only if for all $\delta > 0$, $\lim_{n\to\infty} \mathbb{P}(|X_n/a_n| > \delta) \to 0$. Similarly, we say $X_n = O_P(a_n)$ if and only if for all $\delta > 0$, there exists $M, N \in \mathbb{R}_+$, such that $\mathbb{P}(|X_n/a_n| > M) < \delta$ for all $n > N$.

Throughout the proof, with a slight abuse of notation, we use capital letter C to represent various numerical constants, the values of which might not necessarily be the same in each occurrence.

2.2. Tensor power iteration

Using the rank-one decomposition of $\mathbf{T}$, we can reformulate the tensor power iteration stated in Eq. (1) as follows:

$$x_t = \sqrt{d} \cdot \frac{\sum_{i=1}^k \langle a_i, x_{t-1}\rangle^3 a_i}{\left\|\sum_{i=1}^k \langle a_i, x_{t-1}\rangle^3 a_i\right\|_2}, \qquad t \geq 1.$$

For notational convenience, we recast x_t as $\widetilde{x}_t$, and redefine

$$x_t := \sum_{i=1}^k \langle a_i, \widetilde{x}_{t-1}\rangle^3 a_i = A^\mathsf{T}(A\widetilde{x}_{t-1})^3,$$

where we recall that $A \in \mathbb{R}^{k\times d}$ is the matrix whose i-th row is a_i^T. In other words, x_t is the gradient of $\mathcal{S}(x) = \|Ax\|_4^4$ at $\widetilde{x}_{t-1}$, and $\widetilde{x}_t$ is the projection of x_t onto $\mathbb{S}^{d-1}(\sqrt{d})$:

$$x_t = A^\mathsf{T}(A\widetilde{x}_{t-1})^3, \ \widetilde{x}_t = \sqrt{d} \cdot \frac{x_t}{\|x_t\|_2}. \tag{3}$$

We introduce some useful intermediate variables: $y_t = A\widetilde{x}_{t-1}$, $f_t = y_t^3$. Using these intermediate variables, tensor power iteration can be decomposed into the following steps:

$$\begin{aligned} y_t &= A\widetilde{x}_{t-1}, \qquad & f_t &= y_t^3, \\ x_t &= A^\mathsf{T} f_t, \qquad & \widetilde{x}_t &= \sqrt{d} \cdot \frac{x_t}{\|x_t\|_2}. \end{aligned} \tag{4}$$

2.3. Intuition behind slow convergence

We now provide a heuristic justification for Theorem 1.1 through analyzing the first step of power iteration. Indeed, we show that for any initialization $x_0 \in \mathbb{S}^{d-1}(\sqrt{d})$ that is independent of $\mathbf{T}$, the normalized first iterate $\widetilde{x}_1$ (defined in Eq. (3)) must lie in a small neighborhood of x_0 with high probability. This claim is made precise by the following proposition:

Proposition 2.1 *Let $x_0 \in \mathbb{S}^{d-1}(\sqrt{d})$ be independent of A, and $\widetilde{x}_1$ be defined as per Eq. (3). Furthermore, we assume $k \asymp d^c$ for some $c \in (3/2, 2)$, and let η be a small positive constant satisfying $\eta < (c-1)/2$. Then, with probability at least $1 - C_0\exp(-C_1 d^{\min\{2\eta, c/4+\eta/2\}})$, it holds that $\|\widetilde{x}_1 - x_0\|_2 \leq C_2 \cdot d^{(1-c)/2+\eta}$. $\|x_0\|_2$, where $C_0, C_1, C_2 > 0$ are absolute constants.*

Proof By rotational invariance, we may assume without loss of generality that $x_0 = \sqrt{d}\cdot e_1 = (\sqrt{d}, 0, \cdots, 0)^\mathsf{T}$, which implies that

$$\sum_{i=1}^k \langle a_i, x_0\rangle^3 a_i = \sum_{i=1}^k d^{3/2} a_{i1}^3 a_i = \left(d^{3/2}\sum_{i=1}^k a_{i1}^4, d^{3/2}\sum_{i=1}^k a_{i1}^3 a_{i2}, \cdots, d^{3/2}\sum_{i=1}^k a_{i1}^3 a_{id}\right). \tag{5}$$

Since $a_{ij} \sim_{\text{i.i.d.}} \mathsf{N}(0, 1/d)$, applying Theorem D.3 gives the following concentration bounds:

$$\mathbb{P}\left(\left|\sum_{i=1}^{k} a_{i1}^4 - \frac{3k}{d^2}\right| \geq t\right) \leq C_0 \exp\left(-C_1 \cdot \min\left\{\frac{t^2 d^4}{k},\ d\sqrt{t}\right\}\right), \tag{6}$$

$$\mathbb{P}\left(\left|\sum_{i=1}^{k} a_{i1}^3 a_{il}\right| \geq t\right) \leq C_0 \exp\left(-C_1 \cdot \min\left\{\frac{t^2 d^4}{k},\ d\sqrt{t}\right\}\right), \text{ for } l = 2, \cdots, d. \tag{7}$$

As a consequence, we know that

$$\mathbb{P}\left(\left\|\sum_{i=1}^{k} \frac{1}{d^{3/2}} \langle a_i, x_0\rangle^3 a_i - \frac{3k}{d^2} e_1\right\|_2 \geq t\sqrt{d}\right) \leq C_0 d \exp\left(-C_1 \cdot \min\left\{\frac{t^2 d^4}{k},\ d\sqrt{t}\right\}\right) \tag{8}$$

$$\Longrightarrow \mathbb{P}\left(\left\|\frac{d}{3k} \cdot \sum_{i=1}^{k} \langle a_i, x_0\rangle^3 a_i - x_0\right\|_2 \geq \frac{d^3 t}{3k}\right) \leq C_0 d \exp\left(-C_1 \cdot \min\left\{\frac{t^2 d^4}{k},\ d\sqrt{t}\right\}\right). \tag{9}$$

Therefore, setting $t = 3ks/d^{2.5}$, we obtain that

$$\mathbb{P}\left(\left|\|x_1\|_2 - \frac{3k}{\sqrt{d}}\right| \geq \frac{3ks}{\sqrt{d}}\right) \leq C_0 d \exp\left(-C_1 \cdot \min\left\{\frac{ks^2}{d},\ \frac{\sqrt{ks}}{d^{1/4}}\right\}\right),$$

$$\mathbb{P}\left(\left\|x_1 - \frac{3k}{\sqrt{d}} e_1\right| \geq \frac{3ks}{\sqrt{d}}\right) \leq C_0 d \exp\left(-C_1 \cdot \min\left\{\frac{ks^2}{d},\ \frac{\sqrt{ks}}{d^{1/4}}\right\}\right).$$

Hence, for any $s > 0$,

$$\begin{aligned}
\mathbb{P}\left(\|\widetilde{x}_1 - x_0\|_2 \geq s\sqrt{d}\right) \leq& \mathbb{P}\left(\left\|\widetilde{x}_1 - \frac{d}{3k} \cdot x_1\right\|_2 \geq \frac{s\sqrt{d}}{2}\right) + \mathbb{P}\left(\left\|\frac{d}{3k} \cdot x_1 - x_0\right\|_2 \geq \frac{s\sqrt{d}}{2}\right) \\
=& \mathbb{P}\left(\left|\|x_1\|_2 - \frac{3k}{\sqrt{d}}\right| \geq \frac{3ks}{2\sqrt{d}}\right) + \mathbb{P}\left(\left\|\frac{d}{3k} \cdot x_1 - x_0\right\|_2 \geq \frac{s\sqrt{d}}{2}\right) \\
\leq& C_0 d \exp\left(-C_1 \cdot \min\left\{\frac{ks^2}{d},\ \frac{\sqrt{ks}}{d^{1/4}}\right\}\right),
\end{aligned}$$

which further implies that for any $s > 0$,

$$\mathbb{P}\left(\|\widetilde{x}_1 - x_0\|_2 \geq s\sqrt{d}\right) \leq C_0 d \exp\left(-C_1 \cdot \min\left\{\frac{ks^2}{d}, \frac{\sqrt{ks}}{d^{1/4}}\right\}\right). \tag{10}$$

Recall that $k \asymp d^c$, then choosing $s = d^{(1-c)/2+\eta}$ yields the desired result. ■

The above proposition implies that $\|\widetilde{x}_1 - x_0\|_2 / \|x_0\|_2$ is polynomially small in d with high probability $(1 - \exp(-\Omega(d^{\varepsilon}))$. If we can establish similar upper bounds for power iterations up to $t = \text{poly}(d)$, then we are in good shape as it requires at least polynomially many iterates for the power method to escape the neighborhood of x_0 of (an arbitrarily small) constant radius and in turn converges to any of the tensor components. In the following sections, we show that this is indeed the case by leveraging the Gaussian conditioning argument and extend its analysis to polynomially many steps.

2.4. Gaussian conditioning

In this section, we present a Gaussian conditioning lemma, which enables us to decompose each step of the power iteration as the sum of projections onto its previous iterates and an independent component. This lemma can be viewed as a multi-step generalization of Lemma 3.1 in Montanari and Wu (2022a), and is proved using the properties of Gaussian conditional distribution.

Recalling the variables defined in Eq. (4), we further denote by $F_{1:t} \in \mathbb{R}^{k\times t}$ the matrix whose i-th column is f_i, and $X_{0:t} \in \mathbb{R}^{d\times(t+1)}$ the matrix whose j-th column is x_{j-1}. We define $f_t^\perp = \Pi^\perp_{F_{1:t-1}}(f_t)$, $x_t^\perp = \Pi^\perp_{X_{0:t-1}}(x_t)$, and $\widetilde{x}_t^\perp = \Pi^\perp_{X_{0:t-1}}(\widetilde{x}_t)$. Note that $f_t^\perp \in \mathbb{R}^k$ and $x_t^\perp, \widetilde{x}_t^\perp \in \mathbb{R}^d$. The following lemma will be used several times throughout our proof:

Lemma 2.2 (Gaussian conditioning) *For $s, t \in \mathbb{N}$, we define the sigma-algebra:*

$$\mathcal{F}_{s,t} := \sigma(x_0, \cdots, x_s, y_1, \cdots, y_t).$$

Then, we have the following decompositions:

$$x_t = \sum_{i=0}^{t-1} \widetilde{x}_i^\perp \cdot \frac{\langle h_{i+1}, f_t\rangle}{\|\widetilde{x}_i^\perp\|_2^2} + \Pi^\perp_{X_{0:t-1}} \overline{A}_t^\mathsf{T} f_t^\perp, \tag{11}$$

$$\begin{aligned} y_{t+1} &= \sum_{i=1}^{t} h_i \cdot \frac{\langle \widetilde{x}_{i-1}^\perp, \widetilde{x}_t\rangle}{\|\widetilde{x}_{i-1}^\perp\|_2^2} + h_{t+1} \\ &= \sum_{i=1}^{t+1} \Pi^\perp_{F_{1:i-1}} \widetilde{A}_i \widetilde{x}_{i-1}^\perp \frac{\langle \widetilde{x}_{i-1}^\perp, \widetilde{x}_t\rangle}{\|\widetilde{x}_{i-1}^\perp\|_2^2} + \sum_{i=2}^{t+1} f_{i-1}^\perp \cdot \frac{\langle x_{i-1}, \widetilde{x}_{i-1}^\perp\rangle}{\|f_{i-1}^\perp\|_2^2} \cdot \frac{\langle \widetilde{x}_{i-1}^\perp, \widetilde{x}_t\rangle}{\|\widetilde{x}_{i-1}^\perp\|_2^2}, \end{aligned} \tag{12}$$

where $\overline{A}_t, \widetilde{A}_{t+1} \in \mathbb{R}^{k\times d}$ satisfy $\overline{A}_t \overset{d}{=} \widetilde{A}_{t+1} \overset{d}{=} A$, and

$$\overline{A}_t \perp \sigma(\mathcal{F}_{t-1,t} \cup \sigma(\overline{A}_1, \cdots, \overline{A}_{t-1}, \widetilde{A}_1, \ldots, \widetilde{A}_t)),\ \widetilde{A}_{t+1} \perp \sigma(\mathcal{F}_{t,t} \cup \sigma(\overline{A}_1, \cdots, \overline{A}_t, \widetilde{A}_1, \ldots, \widetilde{A}_t)).$$

Further, the sequence $\{h_t\}$ is defined via

$$h_{t+1} = f_t^\perp \cdot \frac{\langle x_t, \widetilde{x}_t^\perp\rangle}{\|f_t^\perp\|_2^2} + \Pi^\perp_{F_{1:t}} \widetilde{A}_{t+1} \widetilde{x}_t^\perp. \tag{13}$$

The proof of Lemma 2.2 is provided in Appendix C.

3. Proof overview

We give in this section a formal statement of our main theorem, and provide an overview of its proof. We postpone the full version of the proof to Appendix A.

Theorem 3.1 *Assume $d, k, T \to \infty$ simultaneously satisfying $d^{3/2} \ll k \ll d^2$, and that*

$$T = T(k,d) \ll (\log k)^{-1/3} \cdot \frac{k^{2/3}}{d}.$$

Then for any $\varepsilon > 0$, with probability $1 - o_d(1)$, the following result holds for all $0 \le t \le T$:

$$\frac{1}{\sqrt{d}} \max_{i\in[k]} |\langle \widetilde{x}_t, a_i\rangle| \le \varepsilon. \tag{14}$$

That is to say, tensor power iteration fails to identify any true component in $T(k,d)$ steps.

Remark 3.1 *The above conclusion extends to the case of arbitrary even-order tensors: For general $m \geq 2$ and $a_i \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0, I_d/d)$, let $\mathbf{T} = \sum_{i=1}^k a_i^{\otimes 2m}$. Tensor power iteration for decomposing $\mathbf{T}$ is defined as the following iteration:*

$$x_t = A^\mathsf{T}(A\widetilde{x}_{t-1})^{2m-1},\ \widetilde{x}_t = \sqrt{d} \cdot \frac{x_t}{\|x_t\|_2}.$$

Similar to the proof of Theorem 3.1, we obtain that under the condition $d^{(2m-1)/m} \ll k \ll d^m$, as long as tensor power iteration starts from a random initialization and

$$T \ll (\log k)^{-1/3} \cdot \min\left\{k^{1/2m},\ \frac{k^{m/(2m-1)}}{d}\right\},$$

then for all fixed $\varepsilon > 0$, with probability $1 - o_d(1)$, Eq. (14) holds for all $0 \leq t \leq T$. The proof of this claim is similar to the proof of Theorem 3.1, and we skip it here for the sake of simplicity.

Remark 3.2 *Using a similar argument, we can show that projected gradient descent requires at least polynomially many steps to converge to any tensor component as well.*

We also state here a formal version of Theorem 1.2, whose proof is provided in Appendix B:

Theorem 3.2 *Recall that $\mathcal{S}(x) = \sum_{i=1}^k \langle a_i, x\rangle^4$ for $x \in \mathbb{S}^{d-1}(\sqrt{d})$. Assume $d, k \to \infty$ simultaneously satisfying $d^{3/2} \ll k \ll d^2$. Then for any $T_c \in \mathbb{N}_+$ that does not grow with k and d, the following holds with high probability: For all $t = 0, 1, \cdots, T_c$, we have*

$$S(\widetilde{x}_t) = 3k + 20td + o_P(d).$$

As a consequence, with probability $1 - o_d(1)$ we have $\mathcal{S}(\widetilde{x}_{t+1}) > \mathcal{S}(\widetilde{x}_t)$ for all $0 \leq t \leq T_c - 1$.

3.1. Proof sketch of Theorem 3.1

By definition, we have $y_t = A\widetilde{x}_{t-1}$. Therefore, in order to prove Theorem 3.1, it suffices to show with probability $1 - o_d(1)$, $\|y_t\|_\infty \leq \varepsilon\sqrt{d}$ for all $0 \leq t \leq T$. We will prove a stronger version of this result, namely $\|y_t\|_4 \leq \varepsilon\sqrt{d}$. To this end, we use Eq. (12) in Lemma 2.2 and decompose y_t as $y_t = w_t - \eta_t + v_t$, where

$$\begin{aligned}
w_t &= \sum_{i=1}^t \widetilde{A}_i\widetilde{x}_{i-1}^\perp \cdot \frac{\langle \widetilde{x}_{i-1}^\perp, \widetilde{x}_{t-1}\rangle}{\|\widetilde{x}_{i-1}^\perp\|_2^2},\\
\eta_t &= \sum_{i=1}^t \sum_{j=1}^{i-1} f_j^\perp \cdot \frac{\langle f_j^\perp, \widetilde{A}_i\widetilde{x}_{i-1}^\perp\rangle}{\|f_j^\perp\|_2^2} \cdot \frac{\langle \widetilde{x}_{i-1}^\perp, \widetilde{x}_{t-1}\rangle}{\|\widetilde{x}_{i-1}^\perp\|_2^2},\\
v_t &= \sum_{i=2}^t f_{i-1}^\perp \cdot \frac{\langle x_{i-1}, \widetilde{x}_{i-1}^\perp\rangle}{\|f_{i-1}^\perp\|_2^2} \cdot \frac{\langle \widetilde{x}_{i-1}^\perp, \widetilde{x}_{t-1}\rangle}{\|\widetilde{x}_{i-1}^\perp\|_2^2}.
\end{aligned}$$

By triangle inequality, we have

$$\|y_t\|_4 \leq \|w_t\|_4 + \|\eta_t\|_4 + \|v_t\|_4 \leq \|w_t\|_4 + \|\eta_t\|_2 + \|v_t\|_2.$$

Then, it suffices to upper bound $\|w_t\|_4$, $\|\eta_t\|_2$, and $\|v_t\|_2$ respectively.

Upper bounding $\|w_t\|_4$ and $\|\eta_t\|_2$. For future convenience, define

$$z_i = \widetilde{A}_i\widetilde{x}_{i-1}^\perp \cdot \frac{\sqrt{d}}{\|\widetilde{x}_{i-1}^\perp\|_2}, \qquad \alpha_{i,t} = \frac{\langle \widetilde{x}_{i-1}^\perp, \widetilde{x}_{t-1}\rangle}{\sqrt{d}\|\widetilde{x}_{i-1}^\perp\|_2}.$$

Then, we can write $w_t = \sum_{i=1}^{t} \alpha_{i,t} z_i$ and $\eta_t = \sum_{i=2}^{t} \alpha_{i,t} \Pi_{F_{1:i-1}} z_i$. Moreover, one can show that $z_i \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0, I_k)$ and $\sum_{i=1}^{t} \alpha_{i,t}^2 = 1$. Using standard probability tools, we obtain that there exists a numerical constant $C > 0$, such that the following result holds with high probability (see Lemma A.1 and A.2 for more details):

$$\|w_t\|_4 \le Ck^{1/4} \ll \sqrt{d}, \qquad \|\eta_t\|_2 \le CT^{3/4}(\log k)^{1/4} \ll \sqrt{d} \qquad \forall t \in [T].$$

Upper bounding $\|v_t\|_2$. First, using Lemma A.3 in the appendix, we can show that $\|v_{t+1}\|_2^2$ is very close to $\left\|\Pi_{x_0}^{\perp}(\widetilde{x}_t)\right\|_2^2$. Therefore, it suffices to prove that

$$\left\|\Pi_{x_0}^{\perp}(\widetilde{x}_t)\right\|_2^2 \ll d = \|\widetilde{x}_t\|_2^2, \qquad \forall t \in [T],$$

i.e., tensor power iteration stays close to its initialization in the first T steps. To this end, we express $\left\|\Pi_{x_0}^{\perp}(\widetilde{x}_t)\right\|_2^2$ as $d \cdot \left\|\Pi_{x_0}^{\perp}(x_t)\right\|_2^2 / \|x_t\|_2^2$, and estimate $\left\|\Pi_{x_0}^{\perp}(x_t)\right\|_2^2$ and $\|x_t\|_2^2$, respectively. Leveraging Lemma 2.2, we can express x_t in terms of the previous iterations, and finally establish the following recurrence relation (see Lemma A.4-A.6 for details):

$$\frac{\left\|\Pi_{x_0}^{\perp}(x_t)\right\|_2^2}{\|x_t\|_2^2} \le \frac{\left\|\Pi_{x_0}^{\perp}(x_t)\right\|_2^2}{\|\Pi_{x_0}(x_t)\|_2^2} \le U_{k,d,T}\left(\frac{\left\|\Pi_{x_0}^{\perp}(x_{t-1})\right\|_2^2}{\|\Pi_{x_0}(x_{t-1})\|_2^2}\right),$$

where $U_{k,d,T}$ is an increasing function. See the proof of Theorem 3.1 for its explicit definition.

Analysis of $U_{k,d,T}$. Note that the above inequality and the properties of $U_{k,d,T}$ imply:

$$\frac{\left\|\Pi_{x_0}^{\perp}(x_t)\right\|_2^2}{\|\Pi_{x_0}(x_t)\|_2^2} \le U_{k,d,T}^t\left(\frac{\left\|\Pi_{x_0}^{\perp}(x_0)\right\|_2^2}{\|\Pi_{x_0}(x_0)\|_2^2}\right) = U_{k,d,T}^t(0), \quad \text{for all } t \in [T].$$

Hence, it suffices to show that $U_{k,d,T}^t(0) \ll 1$ for all $t \in [T]$. This is achieved by establishing that

$$U_{k,d,T}(x) \le \left(1 + \frac{1}{2T}\right) x, \qquad \forall x \in [\varepsilon_{k,d}, 2\varepsilon_{k,d}], \tag{15}$$

where $\varepsilon_{k,d} \ll 1$ explicitly depends on k and d. It finally follows that

$$U_{k,d,T}^t(0) \le U_{k,d,T}^t(\varepsilon_{k,d}) \le \left(1 + \frac{1}{2T}\right)^t \cdot \varepsilon_{k,d} \le \sqrt{e}\varepsilon_{k,d} \le 2\varepsilon_{k,d},$$

where the last inequality guarantees that Eq. (15) can be repeatedly applied to control $U_{k,d,T}^t(\varepsilon_{k,d})$. This completes the proof sketch of Theorem 3.1.

4. Numerical experiments

We present in this section several numerical experiments that support our theory. In the following experiments, we generate random tensors, and run tensor power method (1) starting from random initialization over the unit sphere. If the iterates converge to one of the components (up to signs), then we call it a success.

We investigate the success region for tensor power method in our first experiment. In this experiment, we run power method for 1000 iterations, and repeat this procedure for 1000 times independently for each pair of (k, d). We record the corresponding empirical success rates, and plot these success rates for different

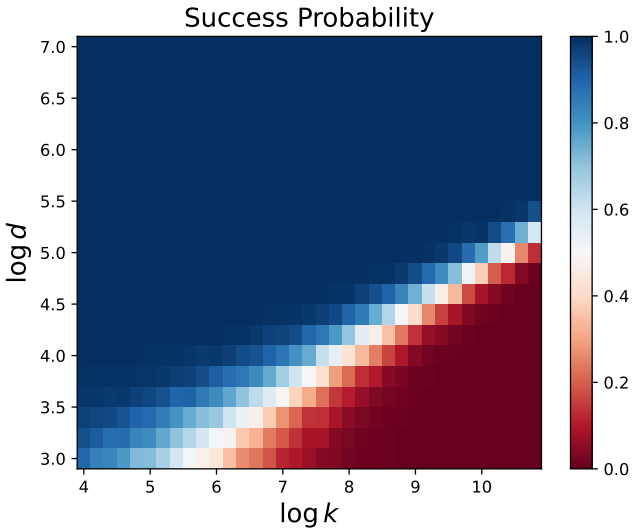

Figure 1: Success probability of tensor power iteration for varying k and d.

(k, d) as a heat map shown by Fig. 1. From the plot, we see that the success rate undergoes a sharp phase transition around the boundary $\log k = 2 \log d$. Our experiment suggests that for $k \ll d^2$, tensor power iteration succeeds with high probability, and when $k \gg d^2$ it fails with high probability. This matches the success region of SOS-based methods, which is also the conjectured region for possible recovery with polynomial-time algorithms.

In our second experiment, we fix $\log k / \log d = 1.8$, and varies d and T. Again we repeat the experiment 1000 times for each pair of (d, T), and record the estimated success probability after T steps of power iteration. We present the outcomes in Fig. 2. The heatmap shows that polynomially many steps are necessary for tensor power iteration to converge, since $\log T$ scales linearly with $\log d$ around the success/failure boundary. This observation validates Theorem 3.1.

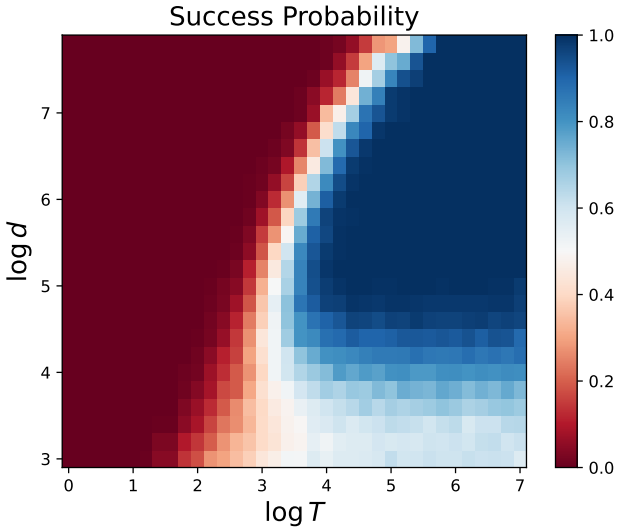

Figure 2: Success probability of tensor power iteration for varying d and T.

We illustrate Theorem 3.2 in our third experiment. In this experiment, we record the values of the score function $\mathcal{S}$ for the first few iterations along the power iteration path, and compare them with the corresponding theoretical predictions given by Theorem 3.2. We repeat such procedure for 1000 times independently, and present the outcomes in Fig. 3, which justifies the correctness of the theorem.

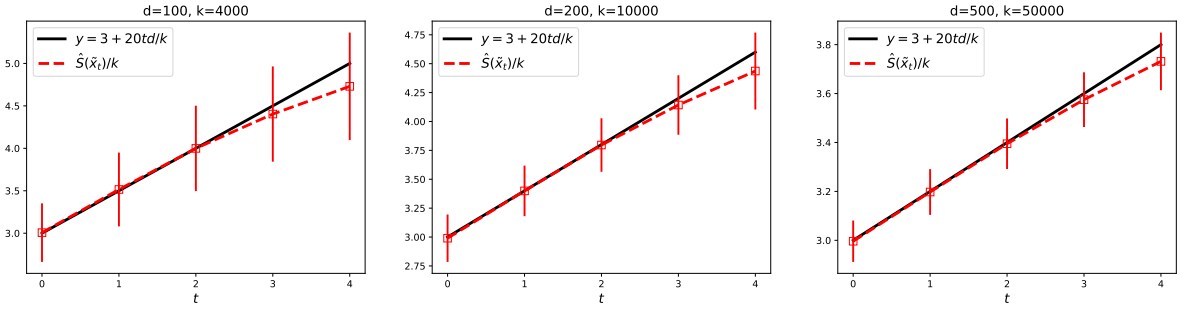

Figure 3: Theoretical predictions of $\mathcal{S}$ along the power iteration paths versus the corresponding empirical values. Outcomes are averaged over 1000 independent experiments. The error bars reflect the intervals determined by two times the empirical standard deviation.

5. Conclusion

We analyze the dynamics of randomly initialized tensor power iteration in the overcomplete regime, using the Gaussian conditioning technique. We prove that it takes polynomially many iterates for the power method to find a true component of a random even-order tensor, thus refuting the previous conjecture that tensor power iteration converges in logarithmically many steps. We also establish that along the power iteration path, a popular polynomial objective function for tensor decomposition is strictly increasing for finitely many steps. Extensive numerical studies verify our theoretical results.

Our work leads to a number of fascinating open problems. First, it would be interesting to understand whether the power method indeed converges to a tensor component in polynomially many steps, within the regime where SOS-based methods succeed. One possible direction is to extend our analysis in Theorem 3.2 to polynomially many iterations. Another appealing direction is to generalize our results to the case of odd-order tensors, for which the power iterates will no longer stay in a small neighborhood of the initialization.

Acknowledgments

The authors would like to thank Tselil Schramm for suggesting this topic, as well as for many helpful conversations regarding the technical contents and the presentation of this work. Y.W. and K.Z. were 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, Rong Ge, Daniel Hsu, Sham M Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *Journal of machine learning research*, 15:2773–2832, 2014a.

Animashree Anandkumar, Rong Ge, and Majid Janzamin. Guaranteed non-orthogonal tensor decomposition via alternating rank-1 updates. *arXiv preprint arXiv:1402.5180*, 2014b.

Animashree Anandkumar, Rong Ge, and Majid Janzamin. Learning overcomplete latent variable models through tensor methods. In *Conference on Learning Theory*, pages 36–112. PMLR, 2015.

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

Boaz Barak, Jonathan A Kelner, and David Steurer. Dictionary learning and tensor decomposition via the sum-of-squares method. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 143–151, 2015.

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.

Aditya Bhaskara, Aidao Chen, Aidan Perreault, and Aravindan Vijayaraghavan. Smoothed analysis in unsupervised learning via decoupling. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 582–610. IEEE, 2019.

J Douglas Carroll and Jih-Jie Chang. Analysis of individual differences in multidimensional scaling via an n-way generalization of "eckart-young" decomposition. *Psychometrika*, 35(3):283–319, 1970.

Dustin Cartwright and Bernd Sturmfels. The number of eigenvalues of a tensor. *Linear algebra and its applications*, 438(2):942–952, 2013.

Michael Celentano, Andrea Montanari, and Yuchen Wu. The estimation error of general first order methods. In *Conference on Learning Theory*, pages 1078–1141. PMLR, 2020.

Pierre Comon and Christian Jutten. *Handbook of Blind Source Separation: Independent component analysis and applications*. Academic press, 2010.

Lieven De Lathauwer. A link between the canonical decomposition in multilinear algebra and simultaneous matrix diagonalization. *SIAM journal on Matrix Analysis and Applications*, 28(3):642–666, 2006.

Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle. Blind source separation by simultaneous third-order tensor diagonalization. In *1996 8th European Signal Processing Conference (EUSIPCO 1996)*, pages 1–4. IEEE, 1996.

Lieven De Lathauwer, Josphine Castaing, and Jean-Franois Cardoso. Fourth-order cumulant-based blind identification of underdetermined mixtures. *IEEE Transactions on Signal Processing*, 55(6):2965–2973, 2007.

Jingqiu Ding, Tommaso d'Orsi, Chih-Hung Liu, David Steurer, and Stefan Tiegel. Fast algorithm for overcomplete order-3 tensor decomposition. In *Conference on Learning Theory*, pages 3741–3799. PMLR, 2022.

Rong Ge and Tengyu Ma. Decomposing overcomplete 3rd order tensors using sum-of-squares algorithms. *arXiv preprint arXiv:1504.05287*, 2015.

Rong Ge and Tengyu Ma. On the optimization landscape of tensor decompositions. *Advances in Neural Information Processing Systems*, 30, 2017.

Rong Ge, Qingqing Huang, and Sham M Kakade. Learning mixtures of gaussians in high dimensions. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 761–770, 2015.

Botao Hao, Yasin Abbasi Yadkori, Zheng Wen, and Guang Cheng. Bootstrapping upper confidence bound. *Advances in neural information processing systems*, 32, 2019.

Richard A Harshman et al. Foundations of the parafac procedure: Models and conditions for an" explanatory" multimodal factor analysis. 1970.

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.

Samuel B Hopkins, Tselil Schramm, and Jonathan Shi. A robust spectral algorithm for overcomplete tensor decomposition. In *Conference on Learning Theory*, pages 1683–1722. PMLR, 2019.

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

Henk AL Kiers. Towards a standardized notation and terminology in multiway analysis. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 14(3):105–122, 2000.

Joe Kileel and Joao M Pereira. Subspace power method for symmetric tensor decomposition and generalized pca. *arXiv preprint arXiv:1912.04007*, 2019.

Joe Kileel, Timo Klock, and João M Pereira. Landscape analysis of an improved power method for tensor decomposition. *Advances in Neural Information Processing Systems*, 34:6253–6265, 2021.

Joseph B Kruskal. Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear algebra and its applications*, 18(2):95–138, 1977.

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

Tengyu Ma, Jonathan Shi, and David Steurer. Polynomial-time tensor decompositions with sum-of-squares. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 438–446. IEEE, 2016.

Marco Mondelli and Andrea Montanari. On the connection between learning two-layer neural networks and tensor decomposition. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1051–1060. PMLR, 2019.

Andrea Montanari and Yuchen Wu. Adversarial examples in random neural networks with general activations. *arXiv preprint arXiv:2203.17209*, 2022a.

Andrea Montanari and Yuchen Wu. Statistically optimal first order algorithms: A proof via orthogonalization. *arXiv preprint arXiv:2201.05101*, 2022b.

Alexander Novikov, Dmitrii Podoprikhin, Anton Osokin, and Dmitry P Vetrov. Tensorizing neural networks. *Advances in neural information processing systems*, 28, 2015.

Karl Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 185:71–110, 1894.

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

Tselil Schramm and David Steurer. Fast and robust tensor decomposition with applications to dictionary learning. In *Conference on Learning Theory*, pages 1760–1793. PMLR, 2017.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Tong Zhang and Gene H Golub. Rank-one approximation to high order tensors. *SIAM Journal on Matrix Analysis and Applications*, 23(2):534–550, 2001.

Appendix A. Analysis of tensor power iteration: Proof of Theorem 3.1

This section will de devoted to proving Theorem 3.1. For $t \in \mathbb{N}_+$, we define

$$g_t = \sum_{i=1}^{t} \Pi^{\perp}_{F_{1:i-1}} \widetilde{A}_i \widetilde{x}^{\perp}_{i-1} \cdot \frac{\langle \widetilde{x}^{\perp}_{i-1}, \widetilde{x}_{t-1} \rangle}{\|\widetilde{x}^{\perp}_{i-1}\|_2^2},$$
$$v_t = \sum_{i=2}^{t} f^{\perp}_{i-1} \cdot \frac{\langle x_{i-1}, \widetilde{x}^{\perp}_{i-1} \rangle}{\|f^{\perp}_{i-1}\|_2^2} \cdot \frac{\langle \widetilde{x}^{\perp}_{i-1}, \widetilde{x}_{t-1} \rangle}{\|\widetilde{x}^{\perp}_{i-1}\|_2^2}.$$

By Theorem 2.2 we have $y_t = g_t + v_t$. Applying Minkowski's inequality and power mean inequality, we deduce that $\|y_t\|_4^4 \le 8(\|g_t\|_4^4 + \|v_t\|_4^4)$. Notice that if Eq. (14) does not hold, since $y_t = A\widetilde{x}_{t-1}$, then there exists $1 \le t \le T$ such that $\|y_t\|_4^4 \ge \varepsilon^4 d^2$. As a result, in order to prove Theorem 3.1, it suffices to show with high probability, $\|y_t\|_4^4 < \varepsilon^4 d^2$ for all $t \in [T]$. This further reduces to upper bounding $\|g_t\|_4^4$ and $\|v_t\|_4^4$.

We first provide an upper bound for $\|g_t\|_4^4$. To this end, we define

$$w_t = \sum_{i=1}^{t} \widetilde{A}_i \widetilde{x}^{\perp}_{i-1} \cdot \frac{\langle \widetilde{x}^{\perp}_{i-1}, \widetilde{x}_{t-1} \rangle}{\|\widetilde{x}^{\perp}_{i-1}\|_2^2}, \tag{16}$$
$$\eta_t = \sum_{i=1}^{t} \sum_{j=1}^{i-1} f^{\perp}_j \cdot \frac{\langle f^{\perp}_j, \widetilde{A}_i \widetilde{x}^{\perp}_{i-1} \rangle}{\|f^{\perp}_j\|_2^2} \cdot \frac{\langle \widetilde{x}^{\perp}_{i-1}, \widetilde{x}_{t-1} \rangle}{\|\widetilde{x}^{\perp}_{i-1}\|_2^2}. \tag{17}$$

We immediately see that $g_t = w_t - \eta_t$, thus upper bounding $\|g_t\|_4^4$ can be achieved via upper bounding $\|w_t\|_4^4$ and $\|\eta_t\|_4^4$, respectively. As $\widetilde{A}_i \perp \mathcal{F}_{i-1,i-1}$, intuitively speaking, this suggests that w_t behaves like a k-dimensional random vector with i.i.d. standard Gaussian entries. Therefore, with high probability w_t has ℓ^4-norm of order k. On the other hand, observe that η_t is the sum of projections of random vectors onto low-dimensional subspaces, which only accounts for a small proportion of the total variation. As a result, we expect $\|\eta_t\|_4^4$ to be small.

To make these heuristic arguments concrete, we establish the following two lemmas:

Lemma A.1 *Assume $T \ll k^{1/2}$. Then there exists a numerical constant $C > 0$, such that with probability $1 - o_d(1)$, for all $1 \le t \le T$ we have*

$$\|w_t\|_4^4 \le Ck.$$

Lemma A.2 *Assume $T \ll k^{1/2}$. Then with probability $1 - o_d(1)$, for all $1 \le t \le T$ we have*

$$\|\eta_t\|_2 \le CT^{3/4} (\log k)^{1/4} \frac{\|\Pi^{\perp}_{x_0} \widetilde{x}_{t-1}\|_2}{\sqrt{d}}.$$

Note that $\|\eta_t\|_4 \le \|\eta_t\|_2$. Invoking Theorem A.1, Theorem A.2, power mean inequality and Minkowski's inequality, we obtain that with high probability, for all $1 \le t \le T$,

$$\|g_t\|_4^4 \le 8(\|w_t\|_4^4 + \|\eta_t\|_2^4) \le C\left(k + \frac{\|\Pi^{\perp}_{x_0} \widetilde{x}_{t-1}\|_2^2}{d^2} \cdot T^3 \log k\right) \le C(k + T^3 \log k) \ll d^2 \tag{18}$$

for some numerical constant $C > 0$, since under the conditions of Theorem 3.1 we have $T \ll d^{1/2}$. Therefore, in order to prove Theorem 3.1, it remains to upper bound $\|v_{t+1}\|_4^4$. In what follows, we perform a crude analysis which uses the ℓ^2-norm to control the ℓ^4-norm. We comment that a more careful analysis might lead to an improved estimate.

The next lemma establishes that the ℓ^2-norm of v_{t+1} is close to the ℓ^2-norm of $\Pi^{\perp}_{x_0}(\widetilde{x}_t)$.

Lemma A.3 *Under the condition of Theorem 3.1, with probability* $1 - o_d(1)$*, the following result holds for all* $0 \le t \le T-1$*:*

$$\left(1 - \frac{C\log k}{\sqrt{d}}\right)\|\Pi_{x_0}^{\perp}(\widetilde{x}_t)\|_2^2 \le \|v_{t+1}\|_2^2 \le \left(1 + \frac{C\log k}{\sqrt{d}}\right)\|\Pi_{x_0}^{\perp}(\widetilde{x}_t)\|_2^2.$$

The rest of the analysis is devoted to upper bounding $\|\Pi_{x_0}^{\perp}(\widetilde{x}_t)\|_2^2$. By definition, we have

$$\|\Pi_{x_0}^{\perp}(\widetilde{x}_t)\|_2^2 = \frac{d}{\|x_t\|_2^2}\cdot\|\Pi_{x_0}^{\perp}(x_t)\|_2^2. \tag{19}$$

Using Eq. (11) we see that

$$x_t = \sum_{i=0}^{t-1}\widetilde{x}_i^{\perp}\cdot\frac{\langle h_{i+1}, f_t\rangle}{\|\widetilde{x}_i^{\perp}\|_2^2} + \Pi_{X_{0:t-1}}^{\perp}\overline{A}_t^{\mathsf{T}} f_t^{\perp},$$
$$\Pi_{x_0}^{\perp}(x_t) = \sum_{i=1}^{t-1}\widetilde{x}_i^{\perp}\cdot\frac{\langle h_{i+1}, f_t\rangle}{\|\widetilde{x}_i^{\perp}\|_2^2} + \Pi_{X_{0:t-1}}^{\perp}\overline{A}_t^{\mathsf{T}} f_t^{\perp}.$$

According to Pythagorean theorem,

$$\|x_t\|_2^2 = \sum_{i=0}^{t-1}\frac{\langle h_{i+1}, f_t\rangle^2}{\|\widetilde{x}_i^{\perp}\|_2^2} + \|\Pi_{X_{0:t-1}}^{\perp}\overline{A}_t^{\mathsf{T}} f_t^{\perp}\|_2^2, \tag{20}$$

$$\|\Pi_{x_0}^{\perp}(x_t)\|_2^2 = \sum_{i=1}^{t-1}\frac{\langle h_{i+1}, f_t\rangle^2}{\|\widetilde{x}_i^{\perp}\|_2^2} + \|\Pi_{X_{0:t-1}}^{\perp}\overline{A}_t^{\mathsf{T}} f_t^{\perp}\|_2^2. \tag{21}$$

Using the definition of h_{i+1} given in Eq. (13), we deduce that

$$\frac{\langle h_{i+1}, f_t\rangle}{\|\widetilde{x}_i^{\perp}\|_2} = \frac{\langle f_i^{\perp}, f_t\rangle}{\|\widetilde{x}_i^{\perp}\|_2}\cdot\frac{\langle x_i, \widetilde{x}_i^{\perp}\rangle}{\|f_i^{\perp}\|_2^2} + \frac{\langle \Pi_{F_{1:i}}^{\perp}\widetilde{A}_{i+1}\widetilde{x}_i^{\perp}, f_t\rangle}{\|\widetilde{x}_i^{\perp}\|_2}. \tag{22}$$

Combining Eqs. (20) to (22), we obtain the following decomposition:

$$\begin{aligned}\|x_t\|_2^2 &= \mathsf{I} + \mathsf{II} + \mathsf{III},\\ \|\Pi_{x_0}^{\perp}(x_t)\|_2^2 &= \mathsf{I}' + \mathsf{II} + \mathsf{III},\end{aligned} \tag{23}$$

where

$$\mathsf{I} = \sum_{i=0}^{t-1}\frac{\langle \Pi_{F_{1:i}}^{\perp}\widetilde{A}_{i+1}\widetilde{x}_i^{\perp}, f_t\rangle^2}{\|\widetilde{x}_i^{\perp}\|_2^2}, \qquad \mathsf{I}' = \sum_{i=1}^{t-1}\frac{\langle \Pi_{F_{1:i}}^{\perp}\widetilde{A}_{i+1}\widetilde{x}_i^{\perp}, f_t\rangle^2}{\|\widetilde{x}_i^{\perp}\|_2^2}, \tag{24}$$

$$\mathsf{II} = \sum_{i=1}^{t-1}\frac{\langle f_i^{\perp}, f_t\rangle^2}{\|f_i^{\perp}\|_2^2}\cdot\frac{\langle x_i, x_i^{\perp}\rangle}{\|f_i^{\perp}\|_2^2} + \|\Pi_{X_{0:t-1}}^{\perp}\overline{A}_t^{\mathsf{T}} f_t^{\perp}\|_2^2, \tag{25}$$

$$\mathsf{III} = \sum_{i=1}^{t-1}\frac{2\langle x_i, \widetilde{x}_i^{\perp}\rangle\langle f_i^{\perp}, f_t\rangle\langle \Pi_{F_{1:i}}^{\perp}\widetilde{A}_{i+1}\widetilde{x}_i^{\perp}, f_t\rangle}{\|\widetilde{x}_i^{\perp}\|_2^2\|f_i^{\perp}\|_2^2}. \tag{26}$$

Next, we will show that terms II and III are negligible comparing to terms I and I'. Note that by Cauchy-Schwarz inequality, $|\mathsf{III}| \le 2\sqrt{\mathsf{I}'\times\mathsf{II}}$, namely the term III is controlled by II. This motivates us to first provide an upper bound for term II.

Lemma A.4 *Under the condition of Theorem 3.1, with probability $1-o_d(1)$, for all $1\le t\le T$ we have (note that II depends on t as well)*

$$\left(1-\frac{C\log k}{\sqrt{d}}\right)\|f_t\|_2^2\le \mathsf{II}\le\left(1+\frac{C\log k}{\sqrt{d}}\right)\|f_t\|_2^2.$$

The following lemma establishes an upper bound on $\|f_t\|_2^2$:

Lemma A.5 *Under the condition of Theorem 3.1, there exists a numerical constant $C>0$, such that with probability $1-o_d(1)$, for all $1\le t\le T$ we have*

$$\|f_t\|_2^2\le C\left(k+\|\Pi_{x_0}^{\perp}(\widetilde{x}_{t-1})\|_2^6\right).$$

With the aid of Theorem A.4 and Theorem A.5, we obtain that with high probability

$$\mathsf{II}\le C\left(k+\|\Pi_{x_0}^{\perp}(\widetilde{x}_{t-1})\|_2^6\right)$$

for all $1\le t\le T$ and some positive numerical constant C. Next, we analyze terms I and I'. To achieve this goal, we find the following lemma useful:

Lemma A.6 *Under the condition of Theorem 3.1, there exists a numerical constant $C>0$, such that with probability $1-o_d(1)$, for all $0\le i<t\le T$, we have*

$$\begin{aligned}&\left|\frac{\langle\Pi_{F_{1:i}}^{\perp}\widetilde{A}_{i+1}\widetilde{x}_i^{\perp},f_t\rangle}{\|\widetilde{x}_i^{\perp}\|_2}-\frac{3k}{d}\cdot\frac{\langle\widetilde{x}_i^{\perp},\widetilde{x}_{t-1}\rangle}{\|\widetilde{x}_i^{\perp}\|_2}\right|\\ \le& C\left(\sqrt{\frac{kT\log k}{d}}+\sqrt{\frac{T}{d}}\,\|\Pi_{x_0}^{\perp}(\widetilde{x}_{t-1})\|_2^3+\sqrt{\frac{k\log k}{d}}\,\|\Pi_{x_0}^{\perp}\widetilde{x}_{t-1}\|_2\right).\end{aligned}$$

Now we are in position to finish the proof of Theorem 3.1. For future convenience, we hereby establish a general framework for the analysis of tensor power iteration dynamics, based on Theorem A.1 to Theorem A.6. To begin with, let us denote

$$P_t=\left\|\Pi_{x_0}^{\perp}(\widetilde{x}_t)\right\|_2^2,\quad Q_t=\|\Pi_{x_0}(\widetilde{x}_t)\|_2^2.$$

Then, we know that $P_t+Q_t=\|\widetilde{x}_t\|_2^2=d$, and that

$$\frac{P_t}{Q_t}=\frac{\left\|\Pi_{x_0}^{\perp}(x_t)\right\|_2^2}{\|\Pi_{x_0}(x_t)\|_2^2}=\frac{\mathsf{I}'+\mathsf{II}+\mathsf{III}}{\mathsf{I}-\mathsf{I}'}.$$

Recall that our aim is to show that $\|v_t\|_2^4\ll d^2$ for all $1\le t\le T$. According to Theorem A.3, this amounts to proving that $P_t\ll Q_t$ for all $t\in[T]$. Define the stopping time

$$T_k=\inf\{t\in\mathbb{N}_+:P_t\ge k^{1/3}\}.$$

Since $k^{1/3}\ll d$, it then suffices to show that $T_k\ge T(k,d)$ with high probability, where $T(k,d)$ is defined in the statement of Theorem 3.1.

Step 1. A lower bound for $\mathsf{I}-\mathsf{I}'$. By definition, we know that

$$\mathsf{I}-\mathsf{I}'=\frac{\langle\widetilde{A}_1\widetilde{x}_0,f_t\rangle^2}{\|\widetilde{x}_0\|_2^2}.$$

Note that the proof of Theorem A.6 also implies that

$$\begin{aligned}\left|\frac{\langle \widetilde{A}_1\widetilde{x}_0, f_t\rangle}{\|\widetilde{x}_0\|_2} - \frac{3k}{d}\sqrt{Q_{t-1}}\right| &\le C\left(\sqrt{\frac{kT\log k}{d}} + \sqrt{\frac{\log k}{d}}\,\|\Pi_{x_0}^{\perp}(\widetilde{x}_{t-1})\|_2^3 + \sqrt{\frac{k\log k}{d}}\,\|\Pi_{x_0}^{\perp}\widetilde{x}_{t-1}\|_2\right)\\ &= C\left(\sqrt{\frac{kT\log k}{d}} + \sqrt{\frac{\log k}{d}}\,P_{t-1}^{3/2} + \sqrt{\frac{k\log k}{d}}\,P_{t-1}^{1/2}\right)\\ &= C(\log k)^{1/2}\left(\sqrt{\frac{kT}{d}} + \frac{1}{\sqrt{d}}\,P_{t-1}^{3/2} + \sqrt{\frac{k}{d}}\,P_{t-1}^{1/2}\right).\end{aligned}$$

For $t \le T_k$, we have $Q_{t-1} \asymp d$, thus leading to the estimate:

$$\begin{aligned}\mathsf{I} - \mathsf{I}' = \frac{\langle \widetilde{A}_1\widetilde{x}_0, f_t\rangle^2}{\|\widetilde{x}_0\|_2^2} &\ge \frac{9k^2}{d^2}Q_{t-1}\cdot\left(1 - \frac{C(\log k)^{1/2}\left(\sqrt{dkT} + \sqrt{d}\,P_{t-1}^{3/2} + \sqrt{dk}\,P_{t-1}^{1/2}\right)}{3k\sqrt{Q_{t-1}}}\right)^2\\ &\ge \frac{9k^2}{d^2}Q_{t-1}\cdot\left(1 - C(\log k)^{1/2}\left(k^{-1/2}\,T^{1/2} + k^{-1}P_{t-1}^{3/2} + k^{-1/2}\,P_{t-1}^{1/2}\right)\right)^2.\end{aligned}$$

Hence, it follows that

$$\frac{1}{\mathsf{I} - \mathsf{I}'} \le \frac{d^2}{9k^2Q_{t-1}}\cdot\left(1 + C(\log k)^{1/2}\left(k^{-1/2}\,T^{1/2} + k^{-1}P_{t-1}^{3/2} + k^{-1/2}\,P_{t-1}^{1/2}\right)\right),$$

where the last line is due to the fact that

$$(\log k)^{1/2}\left(k^{-1/2}\,T^{1/2} + k^{-1}P_{t-1}^{3/2} + k^{-1/2}\,P_{t-1}^{1/2}\right) = o(1).$$

Since by our assumption, $T \ll k^{1/3}$, and $P_{t-1} \le k^{1/3}$.

Step 2. An upper bound for $\mathsf{I}' + \mathsf{II} + \mathsf{III}$. Using Cauchy-Schwarz inequality, we get

$$\mathsf{I}' + \mathsf{II} + \mathsf{III} \le \mathsf{I}' + \mathsf{II} + 2\sqrt{\mathsf{I}'\times\mathsf{II}} = \left(\sqrt{\mathsf{I}'} + \sqrt{\mathsf{II}}\right)^2.$$

To upper bound $\sqrt{\mathsf{I}'}$, we note that

$$\begin{aligned}
&\left|\sqrt{\mathsf{I}'} - \frac{3k}{d}\sqrt{P_{t-1}}\right| = \frac{|\mathsf{I}' - 9k^2 P_{t-1}/d^2|}{\sqrt{\mathsf{I}'} + 3k\sqrt{P_{t-1}}/d} \\
&\le \frac{1}{\sqrt{\mathsf{I}'} + 3k\sqrt{P_{t-1}}/d} \cdot \left|\sum_{i=1}^{t-1} \frac{\langle \Pi^{\perp}_{F_{1:i}} \widetilde{A}_{i+1}\widetilde{x}_i^{\perp}, f_t\rangle^2}{\|\widetilde{x}_i^{\perp}\|_2^2} - \frac{9k^2}{d^2}\sum_{i=1}^{t-1}\frac{\langle \widetilde{x}_i^{\perp}, \widetilde{x}_{t-1}\rangle^2}{\|\widetilde{x}_i^{\perp}\|_2^2}\right| \\
&= \frac{1}{\sqrt{\mathsf{I}'} + 3k\sqrt{P_{t-1}}/d} \cdot \left|\sum_{i=1}^{t-1} \left(\frac{\langle \Pi^{\perp}_{F_{1:i}} \widetilde{A}_{i+1}\widetilde{x}_i^{\perp}, f_t\rangle}{\|\widetilde{x}_i^{\perp}\|_2} - \frac{3k}{d}\frac{\langle \widetilde{x}_i^{\perp}, \widetilde{x}_{t-1}\rangle}{\|\widetilde{x}_i^{\perp}\|_2}\right)\left(\frac{\langle \Pi^{\perp}_{F_{1:i}} \widetilde{A}_{i+1}\widetilde{x}_i^{\perp}, f_t\rangle}{\|\widetilde{x}_i^{\perp}\|_2} + \frac{3k}{d}\frac{\langle \widetilde{x}_i^{\perp}, \widetilde{x}_{t-1}\rangle}{\|\widetilde{x}_i^{\perp}\|_2}\right)\right| \\
&\le \frac{1}{\sqrt{\mathsf{I}'} + 3k\sqrt{P_{t-1}}/d} \cdot \sqrt{\sum_{i=1}^{t-1}\left(\frac{\langle \Pi^{\perp}_{F_{1:i}} \widetilde{A}_{i+1}\widetilde{x}_i^{\perp}, f_t\rangle}{\|\widetilde{x}_i^{\perp}\|_2} - \frac{3k}{d}\frac{\langle \widetilde{x}_i^{\perp}, \widetilde{x}_{t-1}\rangle}{\|\widetilde{x}_i^{\perp}\|_2}\right)^2} \\
&\quad \times \sqrt{\sum_{i=1}^{t-1}\left(\frac{\langle \Pi^{\perp}_{F_{1:i}} \widetilde{A}_{i+1}\widetilde{x}_i^{\perp}, f_t\rangle}{\|\widetilde{x}_i^{\perp}\|_2} + \frac{3k}{d}\frac{\langle \widetilde{x}_i^{\perp}, \widetilde{x}_{t-1}\rangle}{\|\widetilde{x}_i^{\perp}\|_2}\right)^2} \\
&\overset{(i)}{\le} C\sqrt{\sum_{i=1}^{t-1}\left(\frac{\langle \Pi^{\perp}_{F_{1:i}} \widetilde{A}_{i+1}\widetilde{x}_i^{\perp}, f_t\rangle}{\|\widetilde{x}_i^{\perp}\|_2} - \frac{3k}{d}\frac{\langle \widetilde{x}_i^{\perp}, \widetilde{x}_{t-1}\rangle}{\|\widetilde{x}_i^{\perp}\|_2}\right)^2} \\
&\overset{(ii)}{\le} C\left(\sqrt{\frac{k\log k}{d}}\,T + \sqrt{\frac{T^2}{d}}\,\|\Pi^{\perp}_{x_0}(\widetilde{x}_{t-1})\|_2^3 + \sqrt{\frac{Tk\log k}{d}}\,\|\Pi^{\perp}_{x_0}\widetilde{x}_{t-1}\|_2\right) \\
&\le C(\log k)^{1/2}\left(\sqrt{\frac{k}{d}}\,T + \frac{T}{\sqrt{d}}\,P_{t-1}^{3/2} + \sqrt{\frac{Tk}{d}}\,P_{t-1}^{1/2}\right),
\end{aligned}$$

where (i) follows from Minkowski's inequality, and (ii) follows from Theorem A.6. Using Theorem A.4 and Theorem A.5, we obtain that with high probability,

$$\sqrt{\mathsf{II}} \le C\,\|f_t\|_2 \le C\left(\sqrt{k} + \|\Pi^{\perp}_{x_0}(\widetilde{x}_{t-1})\|_2^3\right) = C\left(\sqrt{k} + P_{t-1}^{3/2}\right),$$

thus leading to the estimate:

$$\sqrt{\mathsf{I}'} + \sqrt{\mathsf{II}} \le \frac{3k}{d}\sqrt{P_{t-1}} + C(\log k)^{1/2}\left(\sqrt{\frac{k}{d}}\,T + \sqrt{\frac{Tk}{d}}\,P_{t-1}^{1/2}\right) + C\left(\sqrt{k} + P_{t-1}^{3/2}\right),$$

since $T \ll d^{1/3}$ by our assumption. We finally obtain that

$$\begin{aligned}
\mathsf{I}' + \mathsf{II} + \mathsf{III} &\le \left(\frac{3k}{d}\sqrt{P_{t-1}} + C(\log k)^{1/2}\left(\sqrt{\frac{k}{d}}\,T + \sqrt{\frac{Tk}{d}}\,P_{t-1}^{1/2}\right) + C\left(\sqrt{k} + P_{t-1}^{3/2}\right)\right)^2 \\
&\le \left(\frac{3k}{d}\sqrt{P_{t-1}} + C(\log k)^{1/2}\sqrt{\frac{Tk}{d}}\,P_{t-1}^{1/2} + C\left(\sqrt{k} + P_{t-1}^{3/2}\right)\right)^2.
\end{aligned}$$

Step 3. Write a recurrence inequality for P_t/Q_t. Combining our results from the previous steps gives the following recurrence relationship:

$$\begin{aligned}
\frac{P_t}{Q_t} = \frac{\mathsf{I}' + \mathsf{II} + \mathsf{III}}{\mathsf{I} - \mathsf{I}'} &\le \frac{d^2}{9k^2 Q_{t-1}} \cdot \left(\frac{3k}{d}\sqrt{P_{t-1}} + C(\log k)^{1/2}\sqrt{\frac{Tk}{d}}\,P_{t-1}^{1/2} + C\left(\sqrt{k} + P_{t-1}^{3/2}\right)\right)^2 \\
&\quad \times \left(1 + C(\log k)^{1/2}\left(k^{-1/2}\,T^{1/2} + k^{-1}P_{t-1}^{3/2} + k^{-1/2}\,P_{t-1}^{1/2}\right)\right).
\end{aligned}$$

Denote the right hand side of the above inequality as $U_{k,d,T}(P_{t-1}/Q_{t-1})$, then we know that $U_{k,d,T}$ is an increasing function: As P_{t-1}/Q_{t-1} increases, P_{t-1} increases and Q_{t-1} decreases. Hence, the right hand side of the above inequality will also increase. As a consequence, we deduce that

$$\frac{P_T}{Q_T} \leq U_{k,d,T}^T\left(\frac{P_0}{Q_0}\right) = U_{k,d,T}^T(0) \leq U_{k,d,T}^T\left(\frac{k^{1/3}/2}{d-k^{1/3}/2}\right).$$

By definition of $U_{k,d,T}$, whenever

$$\frac{k^{1/3}/2}{d-k^{1/3}/2} \leq x \leq \frac{k^{1/3}}{d-k^{1/3}},$$

we have the following estimate:

$$\begin{aligned}
U_{k,d,T}(x) \leq & x \times \left(1 + 2C \cdot \frac{d}{k^{2/3}} + C(\log k)^{1/2} \cdot \frac{d^{1/2}T^{1/2}}{k^{1/2}}\right)^2 \\
& \times \left(1 + C(\log k)^{1/2}\left(\frac{T^{1/2}}{k^{1/2}} + 2k^{-1/3}\right)\right) \\
\overset{(i)}{\leq} & \left(1 + \frac{Cd}{k^{2/3}} + C(\log k)^{1/2}\left(\frac{d^{1/2}T^{1/2}}{k^{1/2}} + k^{-1/3}\right)\right) \cdot x \\
\overset{(ii)}{\leq} & \left(1 + \frac{Cd}{k^{2/3}} + C(\log k)^{1/2} \cdot \frac{d^{1/2}T^{1/2}}{k^{1/2}}\right) \cdot x,
\end{aligned}$$

where (i) and (ii) both follow from our assumption: $d^{3/2} \ll k \ll d^2$. Since $k^{1/3} \ll d$, it suffices to show that for $T = T(k,d) \ll (\log k)^{-1/3} \cdot (k^{2/3}/d)$,

$$U_{k,d,T}^T\left(\frac{k^{1/3}/2}{d-k^{1/3}/2}\right) \leq \frac{k^{1/3}}{d-k^{1/3}}.$$

Let T' be the maximum integer such that

$$U_{k,d,T}^{T'}\left(\frac{k^{1/3}/2}{d-k^{1/3}/2}\right) \leq \frac{k^{1/3}}{d-k^{1/3}},$$

then by monotonicity of $U_{k,d,T}$ and maximality of T' we know that

$$\begin{aligned}
\frac{k^{1/3}}{d-k^{1/3}} < U_{k,d,T}^{T'+1}\left(\frac{k^{1/3}/2}{d-k^{1/3}/2}\right) \leq & \left(1 + \frac{Cd}{k^{2/3}} + C(\log k)^{1/2} \cdot \frac{d^{1/2}T^{1/2}}{k^{1/2}}\right)^{T'+1} \times \frac{k^{1/3}/2}{d-k^{1/3}/2} \\
\leq & \exp\left(C(T'+1)\left(\frac{d}{k^{2/3}} + (\log k)^{1/2} \cdot \frac{d^{1/2}T^{1/2}}{k^{1/2}}\right)\right) \times \frac{k^{1/3}/2}{d-k^{1/3}/2},
\end{aligned}$$

which further implies that

$$\begin{aligned}
T'+1 \geq & (\log 2/C) \cdot \left(\frac{d}{k^{2/3}} + (\log k)^{1/2} \cdot \frac{d^{1/2}T^{1/2}}{k^{1/2}}\right)^{-1} \geq \frac{\log 2}{2C} \cdot \min\left\{\frac{k^{2/3}}{d},\ (\log k)^{-1/3}k^{1/6}\right\} \\
\geq & \frac{\log 2}{2C} \cdot (\log k)^{-1/3} \cdot \frac{k^{2/3}}{d} \implies T' \gtrsim (\log k)^{-1/3} \cdot \frac{k^{2/3}}{d}.
\end{aligned}$$

This completes the proof of Theorem 3.1.

A.1. Proof of Lemma A.1

For $i, t \in [T]$, we define

$$z_i = \widetilde{A}_i \widetilde{x}_{i-1}^{\perp} \cdot \frac{\sqrt{d}}{\|\widetilde{x}_{i-1}^{\perp}\|_2}, \qquad \alpha_{i,t} = \frac{\langle \widetilde{x}_{i-1}^{\perp}, \widetilde{x}_{t-1} \rangle}{\sqrt{d}\|\widetilde{x}_{i-1}^{\perp}\|_2}. \tag{27}$$

We immediately see that for all $t \in [T]$, we have $\sum_{i=1}^{t} \alpha_{i,t}^2 = 1$, and that $w_t = \sum_{i=1}^{t} \alpha_{i,t} z_i$. According to Theorem 2.2, for all $i \in [T]$ we have $\widetilde{A}_i \perp \sigma(\mathcal{F}_{i-1,i-1} \cup \sigma(\overline{A}_1, \cdots, \overline{A}_{i-1}, \widetilde{A}_1, \cdots, \widetilde{A}_{i-1}))$, we then obtain that for any $i \in [T]$ and $\{z_1, \cdots, z_{i-1}\}$, $z_i \mid z_1, \cdots, z_{i-1} \overset{d}{=} \mathsf{N}(0, I_k)$. As a result, we see that $z_1, \cdots, z_T$ are independent and identically distributed random vectors with marginal distribution $\mathsf{N}(0, I_k)$. Hence, it suffices to prove that

$$\sup_{\alpha \in \mathbb{S}^{T-1}} \left\| \sum_{i=1}^{T} \alpha_i z_i \right\|_4^4 \le Ck$$

with high probability. To this end, we use a covering argument. Fix $\varepsilon \in (0, 1)$ (to be determined later), let $N_\varepsilon(\mathbb{S}^{T-1})$ be an ε-covering of $\mathbb{S}^{T-1}$. Then for any $\alpha \in \mathbb{S}^{T-1}$, there exists $\alpha' \in N_\varepsilon(\mathbb{S}^{T-1})$ such that $\|\alpha - \alpha'\|_2 \le \varepsilon$, thus leading to

$$\left\| \sum_{i=1}^{T} \alpha_i z_i \right\|_4 \le \left\| \sum_{i=1}^{T} \alpha_i' z_i \right\|_4 + \left\| \sum_{i=1}^{T} (\alpha_i - \alpha_i') z_i \right\|_4 \le \sup_{\alpha \in N_\varepsilon(\mathbb{S}^{T-1})} \left\| \sum_{i=1}^{T} \alpha_i z_i \right\|_4 + \varepsilon \cdot \sup_{\alpha \in \mathbb{S}^{T-1}} \left\| \sum_{i=1}^{T} \alpha_i z_i \right\|_4,$$

which further implies that

$$\begin{aligned} &\sup_{\alpha \in \mathbb{S}^{T-1}} \left\| \sum_{i=1}^{T} \alpha_i z_i \right\|_4 \le \sup_{\alpha \in N_\varepsilon(\mathbb{S}^{T-1})} \left\| \sum_{i=1}^{T} \alpha_i z_i \right\|_4 + \varepsilon \cdot \sup_{\alpha \in \mathbb{S}^{T-1}} \left\| \sum_{i=1}^{T} \alpha_i z_i \right\|_4 \\ \Longrightarrow \quad &\sup_{\alpha \in \mathbb{S}^{T-1}} \left\| \sum_{i=1}^{T} \alpha_i z_i \right\|_4 \le \frac{1}{1-\varepsilon} \cdot \sup_{\alpha \in N_\varepsilon(\mathbb{S}^{T-1})} \left\| \sum_{i=1}^{T} \alpha_i z_i \right\|_4. \end{aligned}$$

Now, for any fixed $\alpha \in N_\varepsilon(\mathbb{S}^{T-1})$, we know that $\sum_{i=1}^{T} \alpha_i z_i \sim \mathsf{N}(0, I_k)$. According to Lemma D.3, we know that there exists a constant $C > 0$ such that

$$\mathbb{P}\left(\left\| \sum_{i=1}^{T} \alpha_i z_i \right\|_4^4 \ge Ck \right) \le C \exp(-Ck^{1/2}).$$

Applying a union bound then gives

$$\mathbb{P}\left(\sup_{\alpha \in N_\varepsilon(\mathbb{S}^{T-1})} \left\| \sum_{i=1}^{T} \alpha_i z_i \right\|_4^4 \ge Ck \right) \le C \left(\frac{C}{\varepsilon} \right)^T \exp(-Ck^{1/2}).$$

By our assumption, $T \ll k^{1/2}$. Now we choose $\varepsilon = 1/2$, it follows that

$$\sup_{\alpha \in \mathbb{S}^{T-1}} \left\| \sum_{i=1}^{T} \alpha_i z_i \right\|_4^4 \le 2^4 \cdot \sup_{\alpha \in N_\varepsilon(\mathbb{S}^{T-1})} \left\| \sum_{i=1}^{T} \alpha_i z_i \right\|_4^4 \le Ck$$

with high probability. This completes the proof of Lemma A.1.

A.2. Proof of Lemma A.2

Recall the definition of η_t from Eq. (17):

$$\eta_t = \sum_{i=1}^{t}\sum_{j=1}^{i-1} f_j^{\perp}\cdot\frac{\langle f_j^{\perp}, \widetilde{A}_i\widetilde{x}_{i-1}^{\perp}\rangle}{\|f_j^{\perp}\|_2^2}\cdot\frac{\langle \widetilde{x}_{i-1}^{\perp}, \widetilde{x}_{t-1}\rangle}{\|\widetilde{x}_{i-1}^{\perp}\|_2^2} = \sum_{i=1}^{t}\alpha_{i,t}\Pi_{F_{1:i-1}}z_i = \sum_{i=2}^{t}\alpha_{i,t}\Pi_{F_{1:i-1}}z_i,$$

where the last equality follows from the fact that $\Pi_{F_{1:0}} = 0$, and the $\alpha_{i,t}$'s and z_i's are defined in the proof of Theorem A.1. We thus obtain that

$$\begin{aligned}\|\eta_t\|_2^2 &= \sum_{i=2}^{t}\sum_{j=2}^{t}\alpha_{i,t}\alpha_{j,t}\left\langle\Pi_{F_{1:i-1}}z_i, \Pi_{F_{1:j-1}}z_j\right\rangle\\ &\le \sum_{i=2}^{t}\alpha_{i,t}^2\left\|\Pi_{F_{1:i-1}}z_i\right\|_2^2 + 2\sum_{2\le i<j\le t}\left|\alpha_{i,t}\alpha_{j,t}\left\langle\Pi_{F_{1:i-1}}z_i, \Pi_{F_{1:j-1}}z_j\right\rangle\right|,\end{aligned}$$

where for $i<j$, we know that

$$\left\langle\Pi_{F_{1:i-1}}z_i, \Pi_{F_{1:j-1}}z_j\right\rangle = z_j^{\mathsf{T}}\Pi_{F_{1:j-1}}\Pi_{F_{1:i-1}}z_i = z_j^{\mathsf{T}}\Pi_{F_{1:i-1}}z_i = \mathrm{Tr}\left(\Pi_{F_{1:i-1}}z_iz_j^{\mathsf{T}}\right).$$

Using the same argument as in the proof of Theorem A.1, we see that for $i<j$, $\Pi_{F_{1:i-1}}$, $z_i\sim\mathsf{N}(0,I_k)$, and $z_j\sim\mathsf{N}(0,I_k)$ are mutually independent. In fact, given $(\Pi_{F_{1:i-1}}, z_i)$, the conditional distribution of z_j is always $\mathsf{N}(0,I_k)$, and given $\Pi_{F_{1:i-1}}$, the conditional distribution of z_i is always $\mathsf{N}(0,I_k)$. This further implies that

$$\begin{aligned}&\mathbb{P}\left(\left|z_j^{\mathsf{T}}\Pi_{F_{1:i-1}}z_i\right| \ge \sqrt{C\log k}\left\|\Pi_{F_{1:i-1}}z_i\right\|_2\left|(\Pi_{F_{1:i-1}}, z_i)\right.\right) \le k^{-C},\\ &\mathbb{P}\left(\left\|\Pi_{F_{1:i-1}}z_i\right\|_2 \ge \sqrt{CT}\right) \le C\exp(-CT),\end{aligned}$$

where $C>0$ is an absolute constant. Therefore, we conclude that with probability $1-o_d(1)$, for all $2\le i<j\le T$, one has

$$\left|\left\langle\Pi_{F_{1:i-1}}z_i, \Pi_{F_{1:j-1}}z_j\right\rangle\right| \le \sqrt{CT\log k},\ \left\|\Pi_{F_{1:i-1}}z_i\right\|_2^2 \le CT,$$

thus leading to the following estimate:

$$\begin{aligned}\|\eta_t\|_2^2 &\le CT\sum_{i=2}^{t}\alpha_{i,t}^2 + 2\sqrt{CT\log k}\sum_{2\le i<j\le t}|\alpha_{i,t}\alpha_{j,t}| \le CT\left(1-\alpha_{1,t}^2\right) + \sqrt{CT\log k}\left(\sum_{i=2}^{t}|\alpha_{i,t}|\right)^2\\ &\le CT\left(1-\alpha_{1,t}^2\right) + t\sqrt{CT\log k}\cdot\sum_{i=2}^{t}\alpha_{i,t}^2 \le CT^{3/2}(\log k)^{1/2}\left(1-\alpha_{1,t}^2\right).\end{aligned}$$

Note that by definition, we have

$$1-\alpha_{1,t}^2 = 1-\frac{\langle\widetilde{x}_0^{\perp}, \widetilde{x}_{t-1}\rangle^2}{d\left\|\widetilde{x}_0^{\perp}\right\|_2^2} = 1-\frac{1}{d}\left\|\Pi_{x_0}\widetilde{x}_{t-1}\right\|_2^2 = \frac{1}{d}\left\|\Pi_{x_0}^{\perp}\widetilde{x}_{t-1}\right\|_2^2.$$

Hence, we finally deduce that with high probability,

$$\|\eta_t\|_2 \le CT^{3/4}(\log k)^{1/4}\frac{\left\|\Pi_{x_0}^{\perp}\widetilde{x}_{t-1}\right\|_2}{\sqrt{d}}$$

for all $t\in[T]$, as desired. This concludes the proof.

A.3. Proof of Lemma A.3

Recall the definition of v_{t+1}:

$$v_{t+1} = \sum_{i=1}^{t} f_i^{\perp} \cdot \frac{\langle x_i, \widetilde{x}_i^{\perp} \rangle}{\|f_i^{\perp}\|_2^2} \cdot \frac{\langle \widetilde{x}_i^{\perp}, \widetilde{x}_t \rangle}{\|\widetilde{x}_i^{\perp}\|_2^2}.$$

Since $\{f_i^{\perp}\}_{1 \le i \le t}$ is an orthogonal set, we readily see that

$$\begin{aligned}
\|v_{t+1}\|_2^2 &= \sum_{i=1}^{t} \frac{\langle x_i, \widetilde{x}_i^{\perp} \rangle^2}{\|f_i^{\perp}\|_2^2} \cdot \frac{\langle \widetilde{x}_i^{\perp}, \widetilde{x}_t \rangle^2}{\|\widetilde{x}_i^{\perp}\|_2^4} = \sum_{i=1}^{t} \frac{\|x_i\|_2^2}{d} \cdot \frac{\langle \widetilde{x}_i, \widetilde{x}_i^{\perp} \rangle^2}{\|f_i^{\perp}\|_2^2} \cdot \frac{\langle \widetilde{x}_i^{\perp}, \widetilde{x}_t \rangle^2}{\|\widetilde{x}_i^{\perp}\|_2^4} \\
&= \sum_{i=1}^{t} \frac{\|x_i\|_2^2}{d} \cdot \frac{\langle \widetilde{x}_i^{\perp}, \widetilde{x}_i^{\perp} \rangle^2}{\|f_i^{\perp}\|_2^2} \cdot \frac{\langle \widetilde{x}_i^{\perp}, \widetilde{x}_t \rangle^2}{\|\widetilde{x}_i^{\perp}\|_2^4} = \sum_{i=1}^{t} \frac{\|x_i\|_2^2}{d} \cdot \frac{\langle \widetilde{x}_i^{\perp}, \widetilde{x}_t \rangle^2}{\|f_i^{\perp}\|_2^2} = \sum_{i=1}^{t} \frac{\|x_i^{\perp}\|_2^2}{\|f_i^{\perp}\|_2^2} \cdot \frac{\langle \widetilde{x}_i^{\perp}, \widetilde{x}_t \rangle^2}{\|\widetilde{x}_i^{\perp}\|_2^2}.
\end{aligned}$$

According to Theorem 2.2, we have $x_i^{\perp} = \Pi_{X_{0:i-1}}^{\perp} \overline{A}_i^{\mathsf{T}} f_i^{\perp}$ where $\overline{A}_i \perp (\Pi_{X_{0:i-1}}^{\perp}, f_i^{\perp})$. Therefore, given $(\Pi_{X_{0:i-1}}^{\perp}, f_i^{\perp})$, the conditional distribution of $x_i^{\perp}$ is specified as

$$x_i^{\perp} \overset{d}{=} \frac{\|f_i^{\perp}\|_2}{\sqrt{d}} \cdot \mathsf{N}\left(0, \Pi_{X_{0:i-1}}^{\perp}\right) \Big| \left(\Pi_{X_{0:i-1}}^{\perp}, f_i^{\perp}\right),$$

which further implies that $\|x_i^{\perp}\|_2^2 / \|f_i^{\perp}\|_2^2 \sim \chi^2(d-i)/d$. Using standard concentration arguments, we know that with high probability for all $i \in [T]$:

$$\left| \frac{\|x_i^{\perp}\|_2^2}{\|f_i^{\perp}\|_2^2} - 1 \right| \le C\left(\frac{T}{d} + \frac{\log k}{\sqrt{d}}\right) \le \frac{C \log k}{\sqrt{d}},$$

where the last inequality follows from the condition of Theorem 3.1: $T \ll \sqrt{d}$. With the aid of the above estimation, we deduce that

$$\begin{aligned}
\left| \|v_{t+1}\|_2^2 - \|\Pi_{x_0}^{\perp}(\widetilde{x}_t)\|_2^2 \right| &= \left| \sum_{i=1}^{t} \left(\frac{\|x_i^{\perp}\|_2^2}{\|f_i^{\perp}\|_2^2} - 1 \right) \frac{\langle \widetilde{x}_i^{\perp}, \widetilde{x}_t \rangle^2}{\|\widetilde{x}_i^{\perp}\|_2^2} \right| \le \sum_{i=1}^{t} \left| \frac{\|x_i^{\perp}\|_2^2}{\|f_i^{\perp}\|_2^2} - 1 \right| \cdot \frac{\langle \widetilde{x}_i^{\perp}, \widetilde{x}_t \rangle^2}{\|\widetilde{x}_i^{\perp}\|_2^2} \\
&\le \frac{C \log k}{\sqrt{d}} \sum_{i=1}^{t} \frac{\langle \widetilde{x}_i^{\perp}, \widetilde{x}_t \rangle^2}{\|\widetilde{x}_i^{\perp}\|_2^2} = \frac{C \log k}{\sqrt{d}} \|\Pi_{x_0}^{\perp}(\widetilde{x}_t)\|_2^2,
\end{aligned}$$

which completes the proof of this lemma.

A.4. Proof of Lemma A.4

The proof is similar to that of Theorem A.3. By definition, $x_t^{\perp} = \Pi_{X_{0:t-1}}^{\perp} \overline{A}_t^{\mathsf{T}} f_t^{\perp}$, we get that

$$\mathsf{II} = \sum_{i=1}^{t} \frac{\langle f_i^{\perp}, f_t \rangle^2}{\|f_i^{\perp}\|_2^2} \cdot \frac{\langle x_i, x_i^{\perp} \rangle}{\|f_i^{\perp}\|_2^2} = \sum_{i=1}^{t} \frac{\langle f_i^{\perp}, f_t \rangle^2}{\|f_i^{\perp}\|_2^2} \cdot \frac{\|x_i^{\perp}\|_2^2}{\|f_i^{\perp}\|_2^2}.$$

From the proof of Theorem A.3 we know that, with probability $1 - o_d(1)$, for all $i \in [T]$ one has

$$\frac{\|x_i^{\perp}\|_2^2}{\|f_i^{\perp}\|_2^2} \in \left[1 - \frac{C \log k}{\sqrt{d}}, 1 + \frac{C \log k}{\sqrt{d}}\right],$$

which immediately implies that

$$\mathsf{II} \leq \left(1 + \frac{C \log k}{\sqrt{d}}\right) \sum_{i=1}^{t} \frac{\langle f_i^{\perp}, f_t \rangle^2}{\|f_i^{\perp}\|_2^2} = \left(1 + \frac{C \log k}{\sqrt{d}}\right) \|f_t\|_2^2,$$
$$\mathsf{II} \geq \left(1 - \frac{C \log k}{\sqrt{d}}\right) \sum_{i=1}^{t} \frac{\langle f_i^{\perp}, f_t \rangle^2}{\|f_i^{\perp}\|_2^2} = \left(1 - \frac{C \log k}{\sqrt{d}}\right) \|f_t\|_2^2.$$

This completes the proof.

A.5. Proof of Lemma A.5

Note that $y_t = w_t - \eta_t + v_t$, then by power mean inequality and Minkowski's inequality,

$$\|f_t\|_2^2 = \|y_t\|_6^6 = \|w_t - \eta_t + v_t\|_6^6 \leq 243 \cdot (\|w_t\|_6^6 + \|\eta_t\|_6^6 + \|v_t\|_6^6).$$

It follows from Theorem A.2 that with high probability,

$$\|\eta_t\|_6^6 \leq \|\eta_t\|_2^6 \leq C T^{9/2} (\log k)^{3/2} \frac{\left\|\Pi_{x_0}^{\perp} \widetilde{x}_{t-1}\right\|_2^6}{d^3}$$

for all $1 \leq t \leq T$. Leveraging Theorem A.3, we know that with high probability, for all $1 \leq t \leq T$ we have $\|v_t\|_6^6 \leq \|v_t\|_2^6 \leq 2\|\Pi_{x_0}^{\perp}(\widetilde{x}_{t-1})\|_2^6$. Finally, we upper bound $\|w_t\|_6^6$. Recall that z_i and $\alpha_{i,t}$ are defined in Eq. (27) in the proof of Theorem A.1. Furthermore, the following properties are satisfied: (i) $z_1, z_2, \cdots, z_T \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0, I_k)$; (ii) For all $1 \leq t \leq T$ we have $\sum_{i=1}^{t} \alpha_{i,t}^2 = 1$; (iii) $w_t = \sum_{i=1}^{t} \alpha_{i,t} z_i$. Then, we can use the same covering argument as in the proof of Theorem A.1 to show that

$$\mathbb{P}\left(\sup_{\alpha \in \mathbb{S}^{T-1}} \left\|\sum_{i=1}^{T} \alpha_i z_i\right\|_6^6 \geq Ck\right) \leq C^T \exp(-Ck^{1/3}).$$

By our assumption, $T \ll k^{1/3}$. Therefore, $\|w_t\|_6^6 \leq Ck$ with high probability. As a consequence, we deduce that

$$\|f_t\|_2^2 \leq C\left(k + T^{9/2}(\log k)^{3/2} \frac{\left\|\Pi_{x_0}^{\perp} \widetilde{x}_{t-1}\right\|_2^6}{d^3} + \|\Pi_{x_0}^{\perp}(\widetilde{x}_{t-1})\|_2^6\right) \leq C\left(k + \|\Pi_{x_0}^{\perp}(\widetilde{x}_{t-1})\|_2^6\right),$$

since $T \ll d^{1/2}$. This completes the proof.

A.6. Proof of Lemma A.6

Notice that

$$\frac{\langle \Pi_{F_{1:i}}^{\perp} \widetilde{A}_{i+1} \widetilde{x}_i^{\perp}, f_t \rangle}{\|\widetilde{x}_i^{\perp}\|_2} = \frac{\langle \widetilde{A}_{i+1} \widetilde{x}_i^{\perp}, f_t \rangle}{\|\widetilde{x}_i^{\perp}\|_2} - \frac{\langle \Pi_{F_{1:i}} \widetilde{A}_{i+1} \widetilde{x}_i^{\perp}, f_t \rangle}{\|\widetilde{x}_i^{\perp}\|_2}. \tag{28}$$

We first upper bound the second term on the right hand side of Eq. (28). Applying Cauchy-Schwarz inequality implies that

$$\frac{|\langle \Pi_{F_{1:i}} \widetilde{A}_{i+1} \widetilde{x}_i^{\perp}, f_t \rangle|}{\|\widetilde{x}_i^{\perp}\|_2} \leq \frac{\|\Pi_{F_{1:i}} \widetilde{A}_{i+1} \widetilde{x}_i^{\perp}\|_2}{\|\widetilde{x}_i^{\perp}\|_2} \cdot \|f_t\|_2. \tag{29}$$

Since $\widetilde{A}_{i+1} \perp \mathcal{F}_{i,i}$, we can deduce that $\sqrt{d}\|\Pi_{F_{1:i}}\widetilde{A}_{i+1}\widetilde{x}_i^{\perp}\|_2/\|\widetilde{x}_i^{\perp}\|_2 \overset{d}{=} \sqrt{X_D}$, where $1 \le D \le i$ is the rank of $F_{1:i}$ and X_D is a chi-squared random variable with D degrees of freedom. By Bernstein's inequality (Theorem D.2), we obtain that with high probability for all $0 \le i \le T$, $\|\Pi_{F_{1:i}}\widetilde{A}_{i+1}\widetilde{x}_i^{\perp}\|_2/\|\widetilde{x}_i^{\perp}\|_2 \le C\sqrt{T/d}$ for some absolute constant $C > 0$. Applying this result and Theorem A.5, we conclude from Eq. (29) that, there exists a positive absolute constant C, such that with high probability for all $0 \le i < t \le T$:

$$\frac{|\langle \Pi_{F_{1:i}}\widetilde{A}_{i+1}\widetilde{x}_i^{\perp}, f_t\rangle|}{\|\widetilde{x}_i^{\perp}\|_2} \le C\sqrt{\frac{T}{d}} \cdot \left(\sqrt{k} + \|\Pi_{x_0}^{\perp}(\widetilde{x}_{t-1})\|_2^3\right).$$

Next, we consider $\langle \widetilde{A}_{i+1}\widetilde{x}_i^{\perp}, f_t\rangle/\|\widetilde{x}_i^{\perp}\|_2$. Direct computation implies that

$$\begin{aligned}\frac{\langle \widetilde{A}_{i+1}\widetilde{x}_i^{\perp}, f_t\rangle}{\|\widetilde{x}_i^{\perp}\|_2} =& \frac{\langle \widetilde{A}_{i+1}\widetilde{x}_i^{\perp}, (w_t - \eta_t + v_t)^3\rangle}{\|\widetilde{x}_i^{\perp}\|_2} \\ =& \frac{\langle \widetilde{A}_{i+1}\widetilde{x}_i^{\perp}, w_t^3\rangle}{\|\widetilde{x}_i^{\perp}\|_2} + \frac{3\langle \widetilde{A}_{i+1}\widetilde{x}_i^{\perp}, w_t^2(v_t - \eta_t)\rangle}{\|\widetilde{x}_i^{\perp}\|_2} + \frac{3\langle \widetilde{A}_{i+1}\widetilde{x}_i^{\perp}, w_t(v_t - \eta_t)^2\rangle}{\|\widetilde{x}_i^{\perp}\|_2} + \frac{\langle \widetilde{A}_{i+1}\widetilde{x}_i^{\perp}, (v_t - \eta_t)^3\rangle}{\|\widetilde{x}_i^{\perp}\|_2}.\end{aligned}$$

In what follows, we analyze each of the terms above, separately. Recall that we have defined z_i and $\alpha_{i,t}$ in Eq. (27). Using the representation $w_t = \sum_{i=1}^t \alpha_{i,t} z_i$, we can then reformulate the first summand above as follow:

$$\frac{\langle \widetilde{A}_{i+1}\widetilde{x}_i^{\perp}, w_t^3\rangle}{\|\widetilde{x}_i^{\perp}\|_2} = \frac{1}{\sqrt{d}}\langle z_{i+1}, (\textstyle\sum_{j=1}^t \alpha_{j,t} z_j)^3\rangle.$$

We then show that the above quantity concentrates around its expectation uniformly for $\alpha_t \in \mathbb{S}^{t-1}$ and $t \in [T]$, via a covering argument similar to that in the proof of Theorem A.1. First, note that for any fixed $\alpha_t \in \mathbb{S}^{t-1}$, one has

$$\left(z_{i+1}, \sum_{j=1}^t \alpha_{j,t} z_j\right) \overset{d}{=} \left(\alpha_{i+1,t} z + \sqrt{1 - \alpha_{i+1,t}^2}\, g, z\right),$$

where $z, g \sim \mathsf{N}(0, I_k)$ are mutually independent. This further implies that

$$\mathbb{E}\left[\frac{\langle \widetilde{A}_{i+1}\widetilde{x}_i^{\perp}, w_t^3\rangle}{\|\widetilde{x}_i^{\perp}\|_2}\right] = \frac{1}{\sqrt{d}}\mathbb{E}\left[\langle \alpha_{i+1,t} z, z^3\rangle\right] = \frac{3k}{\sqrt{d}}\alpha_{i+1,t} = \frac{3k}{d} \cdot \frac{\langle \widetilde{x}_i^{\perp}, \widetilde{x}_{t-1}\rangle}{\|\widetilde{x}_i^{\perp}\|_2}.$$

Moreover, using Theorem D.3, we deduce that there exist constants $C_0, C_1, C_2 > 0$, such that

$$\mathbb{P}\left(\left|\frac{1}{\sqrt{d}}\langle z_{i+1}, (\textstyle\sum_{j=1}^t \alpha_{j,t} z_j)^3\rangle - \frac{3k}{\sqrt{d}}\alpha_{i+1,t}\right| \ge C_0\sqrt{\frac{k}{d}} \cdot \sqrt{T \log k}\right) \le C_1 \exp(-C_2 T \log k), \tag{30}$$

where C_1 and C_2 depend on C_0, and $C_2 \to \infty$ as $C_0 \to \infty$. Let $\varepsilon > 0$ be a small constant (to be determined later), for $\alpha_t, \alpha_t' \in \mathbb{S}^{t-1}$ satisfying $\|\alpha_t - \alpha_t'\|_2 \le \varepsilon$, we have

$$\begin{aligned}&\left|\left(\frac{1}{\sqrt{d}}\langle z_{i+1}, (\textstyle\sum_{j=1}^t \alpha_{j,t} z_j)^3\rangle - \frac{3k}{\sqrt{d}}\alpha_{i+1,t}\right) - \left(\frac{1}{\sqrt{d}}\langle z_{i+1}, (\textstyle\sum_{j=1}^t \alpha_{j,t}' z_j)^3\rangle - \frac{3k}{\sqrt{d}}\alpha_{i+1,t}'\right)\right| \\ \le& \frac{3k\varepsilon}{\sqrt{d}} + \frac{1}{\sqrt{d}}\left|\left\langle z_{i+1}, (\textstyle\sum_{j=1}^t \alpha_{j,t} z_j)^3 - (\sum_{j=1}^t \alpha_{j,t}' z_j)^3\right\rangle\right| \\ \le& \frac{3k\varepsilon}{\sqrt{d}} + \frac{1}{\sqrt{d}}\|z_{i+1}\|_\infty \cdot \left\|(\textstyle\sum_{j=1}^t \alpha_{j,t} z_j)^3 - (\sum_{j=1}^t \alpha_{j,t}' z_j)^3\right\|_1 \\ \le& \frac{3k\varepsilon}{\sqrt{d}} + \frac{1}{\sqrt{d}}\|z_{i+1}\|_\infty \cdot \frac{3}{2}\left(\left\|\textstyle\sum_{j=1}^t \alpha_{j,t} z_j\right\|_2^2 + \left\|\sum_{j=1}^t \alpha_{j,t}' z_j\right\|_2^2\right) \cdot \left\|\textstyle\sum_{j=1}^t \alpha_{j,t} z_j - \sum_{j=1}^t \alpha_{j,t}' z_j\right\|_\infty \\ \le& \frac{3k\varepsilon}{\sqrt{d}} + \frac{3\sqrt{t}\varepsilon}{\sqrt{d}} \sup_{1 \le j \le t}\|z_j\|_\infty^2 \cdot \sup_{\alpha_t \in \mathbb{S}^{t-1}}\left\|\textstyle\sum_{j=1}^t \alpha_{j,t} z_j\right\|_2^2,\end{aligned}$$

where the last line follows from Cauchy-Schwarz inequality. According to Theorem D.1, we know that there exists a numerical constant $C > 0$, such that with probability $1 - o_d(1)$, we have $\|z_i\|_\infty \le C\sqrt{\log k}$ for all $i \in [T]$. Moreover, using a covering argument similar to the proof of Lemma A.1, we deduce that with high probability,

$$\sup_{\alpha_t \in \mathbb{S}^{t-1}} \left\| \sum_{j=1}^t \alpha_{j,t} z_j \right\|_2^2 \le Ck \text{ for all } t \in [T],$$

thus leading to the following estimate:

$$\left| \left(\frac{1}{\sqrt{d}} \langle z_{i+1}, (\sum_{j=1}^t \alpha_{j,t} z_j)^3 \rangle - \frac{3k}{\sqrt{d}} \alpha_{i+1,t} \right) - \left(\frac{1}{\sqrt{d}} \langle z_{i+1}, (\sum_{j=1}^t \alpha'_{j,t} z_j)^3 \rangle - \frac{3k}{\sqrt{d}} \alpha'_{i+1,t} \right) \right| \le Ck\varepsilon \sqrt{\frac{T}{d}} \log k.$$

Therefore, we can apply an ε-net covering argument with $\varepsilon = 1/k$ on $\mathbb{S}^{t-1}$, and choose C_0 to be large enough so that $C_2 > 1$. This finally implies that with high probability, for all $t \in [T]$ we have

$$\sup_{\alpha_t \in \mathbb{S}^{t-1}} \left| \frac{1}{\sqrt{d}} \langle z_{i+1}, (\sum_{j=1}^t \alpha_{j,t} z_j)^3 \rangle - \frac{3k}{\sqrt{d}} \alpha_{i+1,t} \right| \le C \sqrt{\frac{k}{d}} \cdot \sqrt{T \log k},$$

which further implies that

$$\left| \frac{\langle \widetilde{A}_{i+1} \widetilde{x}_i^\perp, w_t^3 \rangle}{\|\widetilde{x}_i^\perp\|_2} - \frac{3k}{d} \cdot \frac{\langle \widetilde{x}_i^\perp, \widetilde{x}_{t-1} \rangle}{\|\widetilde{x}_i^\perp\|_2} \right| \le C \sqrt{\frac{k}{d}} \cdot \sqrt{T \log k}.$$

Now we try to upper bound the remainders. We already know that there exists a numerical constant $C > 0$, such that with probability $1 - o_d(1)$, we have $\|z_i\|_\infty \le C\sqrt{\log k}$ for all $i \in [T]$. Therefore, with probability $1 - o_d(1)$, the following holds for all $t \in [T]$:

$$\begin{aligned}
\|w_t\|_\infty &\le \sum_{i=1}^t |\alpha_{i,t}| \cdot \|z_i\|_\infty \le C\sqrt{\log k} \left(|\alpha_{1,t}| + \sum_{i=2}^t |\alpha_{i,t}| \right) \le C\sqrt{\log k} \left(|\alpha_{1,t}| + \sqrt{t \sum_{i=2}^t \alpha_{i,t}^2} \right) \\
&= C\sqrt{\log k} \left(\frac{1}{\sqrt{d}} \|\Pi_{x_0}(\widetilde{x}_{t-1})\|_2 + \sqrt{\frac{t}{d}} \|\Pi_{x_0}^\perp(\widetilde{x}_{t-1})\|_2 \right) \le C\sqrt{\log k} \left(1 + \sqrt{\frac{T}{d}} \|\Pi_{x_0}^\perp(\widetilde{x}_{t-1})\|_2 \right).
\end{aligned}$$

According to power mean inequality, there exists a numerical constant $C > 0$, such that with probability $1 - o_d(1)$, for all $1 \le i + 1 \le t \le T$ we have

$$\begin{aligned}
\frac{3|\langle \widetilde{A}_{i+1} \widetilde{x}_i^\perp, w_t (v_t - \eta_t)^2 \rangle|}{\|\widetilde{x}_i^\perp\|_2} &\le \frac{3}{\sqrt{d}} \|z_{i+1}\|_\infty \cdot \|w_t\|_\infty \cdot \|v_t - \eta_t\|_2^2 \\
&\le \frac{C \log k}{\sqrt{d}} \cdot \left(1 + \sqrt{\frac{T}{d}} \|\Pi_{x_0}^\perp(\widetilde{x}_{t-1})\|_2 \right) \left(\|\eta_t\|_2^2 + \|v_t\|_2^2 \right) \qquad (31) \\
&\overset{(i)}{\le} \frac{C \log k}{\sqrt{d}} \cdot \left(1 + \sqrt{\frac{T}{d}} \|\Pi_{x_0}^\perp(\widetilde{x}_{t-1})\|_2 \right) \|\Pi_{x_0}^\perp(\widetilde{x}_{t-1})\|_2^2.
\end{aligned}$$

In the above equation, *(i)* follows from Theorem A.2 and Theorem A.3. Similarly, we can conclude that there exists a numerical constant $C > 0$, such that with probability $1 - o_d(1)$, for all $1 \le i + 1 \le t \le T$ the

following holds:

$$\begin{aligned}\frac{3|\langle \widetilde{A}_{i+1}\widetilde{x}_i^{\perp}, w_t^2(v_t-\eta_t)\rangle|}{\|\widetilde{x}_i^{\perp}\|_2} \leq & \frac{3}{\sqrt{d}}\|z_{i+1}\|_\infty \cdot \|w_t^2(v_t-\eta_t)\|_1 \\ \overset{(ii)}{\leq} & \frac{C\sqrt{\log k}}{\sqrt{d}}\left\|w_t^2\right\|_2 \cdot \|v_t-\eta_t\|_2 \\ \leq & \frac{C\sqrt{\log k}}{\sqrt{d}}\|w_t\|_4^2 \cdot \left(\|\eta_t\|_2+\|v_t\|_2\right) \\ \overset{(iii)}{\leq} & \frac{C\sqrt{k\log k}}{\sqrt{d}} \cdot \left\|\Pi_{x_0}^{\perp}(\widetilde{x}_{t-1})\right\|_2 .\end{aligned} \tag{32}$$

In the above inequalities, *(ii)* is due to Hölder's inequality and $\|z_{i+1}\|_\infty \leq C\sqrt{\log k}$, and *(iii)* is due to Theorem A.1, Theorem A.2, and Theorem A.3. Finally, according to the power mean inequality, we obtain that with probability $1-o_d(1)$,

$$\begin{aligned}\frac{|\langle \widetilde{A}_{i+1}\widetilde{x}_i^{\perp}, (v_t-\eta_t)^3\rangle|}{\|\widetilde{x}_i^{\perp}\|_2} \leq & \frac{4}{\sqrt{d}}\|z_{i+1}\|_\infty \cdot \left(\|v_t\|_3^3+\|\eta_t\|_3^3\right) \\ \leq & C\sqrt{\frac{\log k}{d}} \cdot \left(\|v_t\|_2^3+\|\eta_t\|_2^3\right) \\ \leq & 2C\sqrt{\frac{\log k}{d}} \cdot \left\|\Pi_{x_0}^{\perp}(\widetilde{x}_{t-1})\right\|_2^3 .\end{aligned} \tag{33}$$

Theorem A.6 then follows from Eqs. (29) to (33) and our assumptions.

Appendix B. Increasing objective function: Proof of Theorem 3.2

This section will be devoted to proving Theorem 3.2. The main technical lemma employed to prove the theorem can be stated as follows:

Lemma B.1 *Under the condition of Theorem 3.2, for any $t \in [T_c]$ we have*

$$y_t = \sum_{i=1}^{t} \alpha_{i,t} z_i + \sum_{j=1}^{t-1} \zeta_{j,t} f_j + \varepsilon_t, \tag{34}$$

$$\frac{d}{9k^2}\|x_t\|_2^2 = \frac{\langle x_0, x_t\rangle^2}{9k^2} + o_P(1) = 1 + o_P(1), \tag{35}$$

where as in the proof of Theorem A.1 we have

$$z_i = \widetilde{A}_i \widetilde{x}_{i-1}^{\perp} \cdot \frac{\sqrt{d}}{\|\widetilde{x}_{i-1}^{\perp}\|_2}, \qquad \alpha_{i,t} = \frac{\langle \widetilde{x}_{i-1}^{\perp}, \widetilde{x}_{t-1}\rangle}{\sqrt{d}\|\widetilde{x}_{i-1}^{\perp}\|_2}.$$

Furthermore, the above quantities satisfy:

$$\alpha_{1,t} = 1 + o_P(1), \tag{36}$$

$$\frac{3k}{d}\zeta_{j,t} = 1 + o_P(1), \tag{37}$$

$$\|\varepsilon_t\|_2 = o_P(d/\sqrt{k}). \tag{38}$$

In addition, with probability $1-o_d(1)$, for all $t \in [T_c]$ we have

$$\frac{1}{k}\|f_t\|_2^2 \leq 20. \tag{39}$$

We then proceed to prove Theorem 3.2. For the sake of simplicity, define

$$w_t = \sum_{i=1}^{t} \alpha_{i,t} z_i, \quad \nu_t = \sum_{j=1}^{t-1} \zeta_{j,t} f_j + \varepsilon_t.$$

Then, Theorem B.1 implies that

$$\|\nu_t\|_2 \le \sum_{j=1}^{t-1} |\zeta_{j,t}| \cdot \|f_j\|_2 + \|\varepsilon_t\|_2 \le C \cdot \sum_{j=1}^{t-1} O_P\left(\frac{d}{\sqrt{k}}\right) + o_P\left(\frac{d}{\sqrt{k}}\right) = O_P\left(\frac{d}{\sqrt{k}}\right),$$

i.e., $\|\nu_t\|_2^2 = O_P(d^2/k)$. Recall that in the proof of Lemma A.1, we have shown that $\sum_{i=1}^{t} \alpha_{i,t}^2 = 1$ and $z_i \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0, I_k)$. Since $\alpha_{1,t} = 1 + o_P(1)$ and T_c is a constant, we can use standard concentration arguments for Gaussian random variables to deduce that

$$\|w_t^3\|_2^2 = \|w_t\|_6^6 = 15k + O_P(\sqrt{k}), \quad \|w_t\|_\infty = O_P(\log k).$$

Using Eq. (34) from Theorem B.1, we see that

$$\begin{aligned}
\mathcal{S}(\widetilde{x}_{t-1}) &= \|y_t\|_4^4 = \langle w_t^4, 1\rangle + 4\langle w_t^3, \nu_t\rangle + 6\langle w_t^2, \nu_t^2\rangle + 4\langle w_t, \nu_t^3\rangle + \langle \nu_t^4, 1\rangle \\
&= \langle w_t^4, 1\rangle + 4\sum_{j=1}^{t-1} \zeta_{j,t}\langle w_t^3, f_j\rangle + 4\langle w_t^3, \varepsilon_t\rangle + 6\langle w_t^2, \nu_t^2\rangle + 4\langle w_t, \nu_t^3\rangle + \langle \nu_t^4, 1\rangle.
\end{aligned}$$

Next, we analyze the terms above, respectively. Similar to the previous argument, we can show that $\langle w_t^4, 1\rangle = \|w_t\|_4^4 = 3k + O_P(\sqrt{k})$. Leveraging Eqs. (36) and (37), we obtain that

$$4\sum_{j=1}^{t-1} \zeta_{j,t}\langle w_t^3, f_j\rangle = 4\sum_{j=1}^{t-1} \zeta_{j,t}\langle w_t^3, (y_j)^3\rangle = 4\sum_{j=1}^{t-1} \zeta_{j,t}\langle w_t^3, (w_j + \nu_j)^3\rangle.$$

For any $1 \le j \le t-1$, using again standard concentration arguments, we obtain that $\left\langle w_t^3, (w_j + \nu_j)^3\right\rangle = 15k + o_P(k)$. Note that $\zeta_{j,t} = d/((3 + o_P(1))k)$, it follows that

$$4\sum_{j=1}^{t-1} \zeta_{j,t}\langle w_t^3, f_j\rangle = 4\sum_{j=1}^{t-1}(5 + o_P(1))d = 20(t-1)d + o_P(d).$$

Applying Cauchy–Schwarz inequality we see that

$$4|\langle w_t^3, \varepsilon_t\rangle| \le 4\|w_t^3\|_2 \cdot \|\varepsilon_t\|_2 = O_P(\sqrt{k}) \cdot o_P\left(\frac{d}{\sqrt{k}}\right) = o_P(d).$$

Furthermore, the following results hold:

$$\begin{aligned}
6|\langle w_t^2, \nu_t^2\rangle| &\le 6\|w_t\|_\infty^2 \cdot \|\nu_t\|_2^2 = O_P(d^2(\log k)^2/k), \\
4|\langle w_t, \nu_t^3\rangle| &\le 4\|w_t\|_\infty \cdot \|\nu_t\|_3^3 \le 4\|w_t\|_\infty \cdot \|\nu_t\|_2^3 = O_P(d^3 \log k/k^{3/2}), \\
|\langle \nu_t^4, 1\rangle| &= \|\nu_t\|_4^4 \le \|\nu_t\|_2^4 = O_P(d^4/k^2).
\end{aligned}$$

Combining these estimates and using the assumption that $d^{3/2} \ll k \ll d^2$, we conclude that $S(\widetilde{x}_{t-1}) = 3k + 20(t-1)d + o_P(d)$, thus completing the proof of the theorem.

B.1. Proof of Theorem B.1

We prove the lemma via induction over t.

BASE CASE

For the base case $t = 1$, from Theorem 2.2 we immediately see that $y_1 = \alpha_{1,1} z_1$, thus proves Eq. (34). Furthermore, by definition $\alpha_{1,1} = 1$, which justifies Eq. (36) for the base case. Eq. (39) follows immediately from the law of large numbers. Again by Theorem 2.2, we have

$$x_1 = \widetilde{x}_0 \cdot \frac{\langle y_1, f_1 \rangle}{d} + \Pi_{x_0}^{\perp} \overline{A}_1^{\mathsf{T}} f_1. \tag{40}$$

Using the law of large numbers, we obtain that $\langle y_1, f_1 \rangle / d = 3k/d + o_P(k/d)$, $\|\Pi_{x_0}^{\perp} \overline{A}_1^{\mathsf{T}} f_1\|_2^2 = \|f_1\|_2^2 + o_P(k) = 15k + o_P(k)$. We then discover that Eq. (35) for the base case follows, since by Eq. (40) we have $\|x_1\|_2^2 = \langle y_1, f_1 \rangle^2 / d + \|\Pi_{x_0}^{\perp} \overline{A}_1^{\mathsf{T}} f_1\|_2^2$ and $\langle x_0, x_1 \rangle^2 / 9k^2 = \langle y_1, f_1 \rangle^2 / 9k^2$. This completes the proof of Eq. (35) for the base case. We note that Eq. (37) does not apply for the base case.

PROOF OF EQ. (34), (36), (37), (38) FOR $t = s + 1$

Suppose the claims hold for $t = s$. Next, we prove that they also hold for $t = s+1$ via induction. Leveraging Eq. (12) in Theorem 2.2, we obtain that

$$y_{s+1} = \sum_{i=1}^{s+1} \alpha_{i,s+1} z_i + \sum_{i=2}^{s+1} \eta_{i,s+1} f_{i-1}^{\perp}, \tag{41}$$

where for $2 \le i \le s+1$

$$\eta_{i,s+1} = \underbrace{\frac{\langle x_{i-1}, \widetilde{x}_{i-1}^{\perp} \rangle}{\|f_{i-1}^{\perp}\|_2^2} \cdot \frac{\langle \widetilde{x}_{i-1}^{\perp}, \widetilde{x}_s \rangle}{\|\widetilde{x}_{i-1}^{\perp}\|_2^2}}_{a_{i,s+1}} - \underbrace{\sum_{j=i}^{s+1} \frac{\langle f_{i-1}^{\perp}, \widetilde{A}_j \widetilde{x}_{j-1}^{\perp} \rangle}{\|f_{i-1}^{\perp}\|_2^2} \cdot \frac{\langle \widetilde{x}_{j-1}^{\perp}, \widetilde{x}_s \rangle}{\|\widetilde{x}_{j-1}^{\perp}\|_2^2}}_{b_{i,s+1}}.$$

We then proceed to prove that $\|b_{i,s+1} f_{i-1}^{\perp}\|_2 = O_P(1)$. To this end, it suffices to show for every $j \in \{i, i+1, \cdots, s+1\}$,

$$\frac{|\langle f_{i-1}^{\perp}, \widetilde{A}_j \widetilde{x}_{j-1}^{\perp} \rangle|}{\|f_{i-1}^{\perp}\|_2^2} \cdot \frac{|\langle \widetilde{x}_{j-1}^{\perp}, \widetilde{x}_s \rangle|}{\|\widetilde{x}_{j-1}^{\perp}\|_2^2} \cdot \|f_{i-1}^{\perp}\|_2 = O_P(1). \tag{42}$$

Note that for $j \ge i$ we have $\widetilde{A}_j \perp \mathcal{F}_{j-1,j-1}$, thus

$$\frac{\langle f_{i-1}^{\perp}, \widetilde{A}_j \widetilde{x}_{j-1}^{\perp} \rangle}{\|f_{i-1}^{\perp}\|_2 \|\widetilde{x}_{j-1}^{\perp}\|_2} \overset{d}{=} \mathsf{N}(0, d^{-1}). \tag{43}$$

Applying Cauchy-Schwarz inequality, we see that

$$\frac{|\langle \widetilde{x}_{j-1}^{\perp}, \widetilde{x}_s \rangle|}{\|\widetilde{x}_{j-1}^{\perp}\|_2} \le \|\widetilde{x}_s\|_2 = \sqrt{d}. \tag{44}$$

Combining Eqs. (43) and (44), we deduce that Eq. (42) holds for all $j \in \{i, i+1, \cdots, s+1\}$. Then we switch to consider $a_{i,s+1}$. Using Eq. (11) and Eq. (13), we have the following decomposition:

$$\begin{aligned} a_{i,s+1} =& \frac{\langle x_{i-1}, \widetilde{x}_{i-1}^{\perp} \rangle}{\|f_{i-1}^{\perp}\|_2^2} \cdot \frac{\langle h_i, f_s \rangle}{\|\widetilde{x}_{i-1}^{\perp}\|_2^2} \cdot \frac{\sqrt{d}}{\|x_s\|_2} \\ =& \underbrace{\frac{\langle x_{i-1}, \widetilde{x}_{i-1}^{\perp} \rangle}{\|f_{i-1}^{\perp}\|_2^2} \cdot \frac{\langle f_{i-1}^{\perp}, f_s \rangle}{\|\widetilde{x}_{i-1}^{\perp}\|_2^2} \cdot \frac{\sqrt{d}}{\|x_s\|_2} \cdot \frac{\langle x_{i-1}, \widetilde{x}_{i-1}^{\perp} \rangle}{\|f_{i-1}^{\perp}\|_2^2}}_{p_{i,s+1}} + \underbrace{\frac{\langle x_{i-1}, \widetilde{x}_{i-1}^{\perp} \rangle}{\|f_{i-1}^{\perp}\|_2^2} \cdot \frac{\langle \Pi_{F_{1:i-1}}^{\perp} \widetilde{A}_i \widetilde{x}_{i-1}^{\perp}, f_s \rangle}{\|\widetilde{x}_{i-1}^{\perp}\|_2^2} \cdot \frac{\sqrt{d}}{\|x_s\|_2}}_{q_{i,s+1}}. \end{aligned}$$

We then analyze $p_{i,s+1}$ and $q_{i,s+1}$, respectively. Combining Eq. (11), the law of large numbers and the fact that $\overline{A}_t \perp \mathcal{F}_{t-1,t}$, we see that for all $t \in \mathbb{N}_+$

$$\frac{\langle x_t, x_t^{\perp} \rangle}{\|f_t^{\perp}\|_2^2} = 1 + O_P(d^{-1/2}). \tag{45}$$

Combining the above analysis with Eq. (35) from previous induction steps, we conclude that

$$p_{i,s+1} = (1 + o_P(1))) \times \frac{\langle f_{i-1}^{\perp}, f_s \rangle}{\|f_{i-1}^{\perp}\|_2^2} \times \frac{d}{3k}. \tag{46}$$

Next, we consider $q_{i,s+1}$. We further decompose $q_{i,s+1}$ as the difference of the following two terms:

$$q_{i,s+1} = \underbrace{\frac{\langle x_{i-1}, \widetilde{x}_{i-1}^{\perp} \rangle}{\|f_{i-1}^{\perp}\|_2^2} \cdot \frac{\langle \widetilde{A}_i \widetilde{x}_{i-1}^{\perp}, f_s \rangle}{\|\widetilde{x}_{i-1}^{\perp}\|_2^2} \cdot \frac{\sqrt{d}}{\|x_s\|_2}}_{c_{i,s+1}} - \underbrace{\sum_{j=1}^{i-1} \frac{\langle x_{i-1}, \widetilde{x}_{i-1}^{\perp} \rangle}{\|f_{i-1}^{\perp}\|_2^2} \cdot \frac{\langle f_j^{\perp}, \widetilde{A}_i \widetilde{x}_{i-1}^{\perp} \rangle}{\|f_j^{\perp}\|_2^2} \cdot \frac{\langle f_j^{\perp}, f_s \rangle}{\|\widetilde{x}_{i-1}^{\perp}\|_2^2} \cdot \frac{\sqrt{d}}{\|x_s\|_2}}_{e_{i,s+1}}.$$

Note that for all $1 \le j \le i-1 \le s$, leveraging Eq. (43), (45) and Eq. (35) from previous induction steps, we obtain that

$$\begin{aligned} & \|f_{i-1}^{\perp}\|_2 \cdot \frac{|\langle x_{i-1}, \widetilde{x}_{i-1}^{\perp} \rangle|}{\|f_{i-1}^{\perp}\|_2^2} \cdot \frac{|\langle f_j^{\perp}, \widetilde{A}_i \widetilde{x}_{i-1}^{\perp} \rangle|}{\|f_j^{\perp}\|_2^2} \cdot \frac{|\langle f_j^{\perp}, f_s \rangle|}{\|\widetilde{x}_{i-1}^{\perp}\|_2^2} \cdot \frac{\sqrt{d}}{\|x_s\|_2} \\ =& \frac{|\langle f_j^{\perp}, f_s \rangle|}{\|f_j^{\perp}\|_2} \cdot \frac{\sqrt{d}}{\|x_s\|_2} \cdot \frac{|\langle f_j^{\perp}, \widetilde{A}_i \widetilde{x}_{i-1}^{\perp} \rangle|}{\|f_j^{\perp}\|_2 \|\widetilde{x}_{i-1}^{\perp}\|_2} \cdot (1 + o_P(1)) \\ \le& \frac{\sqrt{d}}{k} \|f_s\|_2 \cdot O_P(1). \end{aligned}$$

Leveraging Eq. (39) from previous induction steps, we obtain that with probability $1 - o_d(1)$, the above quantity is no larger than $O_P(\sqrt{d/k})$. This further implies that $\|e_{i,s+1} f_{i-1}^{\perp}\|_2 = O_P(\sqrt{d/k})$.

Finally, we analyze $c_{i,s+1}$. By Eq. (41) from previous induction steps, we see that

$$\begin{aligned} \frac{\langle \widetilde{A}_i \widetilde{x}_{i-1}^{\perp}, f_s \rangle}{\|\widetilde{x}_{i-1}^{\perp}\|_2} =& \frac{1}{\sqrt{d}} \Big\langle z_i, \big(\sum_{j=1}^{s} \alpha_{j,s} z_j + \sum_{j=1}^{s-1} \zeta_{j,s} f_j + \varepsilon_s \big)^3 \Big\rangle \\ =& \frac{1}{\sqrt{d}} \Big\langle z_i, \big(\sum_{j=1}^{s} \alpha_{j,s} z_j \big)^3 \Big\rangle + \frac{3}{\sqrt{d}} \Big\langle z_i (\sum_{j=1}^{s} \alpha_{j,s} z_j)^2, \sum_{j=1}^{s-1} \zeta_{j,s} f_j + \varepsilon_s \Big\rangle + \\ & \frac{3}{\sqrt{d}} \Big\langle z_i (\sum_{j=1}^{s} \alpha_{j,s} z_j), \big(\sum_{j=1}^{s-1} \zeta_{j,s} f_j + \varepsilon_s \big)^2 \Big\rangle + \frac{1}{\sqrt{d}} \Big\langle z_i, \big(\sum_{j=1}^{s-1} \zeta_{j,s} f_j + \varepsilon_s \big)^3 \Big\rangle. \end{aligned} \tag{47}$$

Below we analyze terms in Eq. (47), separately.

$$\frac{1}{\sqrt{d}}\Big\langle z_i, \big(\sum_{j=1}^{s}\alpha_{j,s}z_j\big)^3\Big\rangle = \frac{3k\langle \widetilde{x}_{i-1}^{\perp}, \widetilde{x}_{s-1}\rangle}{d\|\widetilde{x}_{i-1}^{\perp}\|_2} + O_P\Big(\sqrt{\frac{k}{d}}\Big), \tag{48}$$

$$\begin{aligned}\frac{3}{\sqrt{d}}\Big|\Big\langle z_i(\sum_{j=1}^{s}\alpha_{j,s}z_j)^2, \sum_{j=1}^{s-1}\zeta_{j,s}f_j+\varepsilon_s\Big\rangle\Big| \le& \frac{3}{\sqrt{d}}\big\|z_i^2(\sum_{j=1}^{s}\alpha_{j,s}z_j)^2\big\|_2 \times \left(\sum_{j=1}^{s-1}|\zeta_{j,s}|\cdot\|f_j\|_2+\|\varepsilon_s\|_2\right)\\ =& O_P(\sqrt{d}),\end{aligned} \tag{49}$$

$$\begin{aligned}\frac{3}{\sqrt{d}}\Big|\Big\langle z_i(\sum_{j=1}^{s}\alpha_{j,s}z_j), \big(\sum_{j=1}^{s-1}\zeta_{j,s}f_j+\varepsilon_s\big)^2\Big\rangle\Big| \le& \frac{3T_c\|z_i\|_\infty \max_{1\le j\le T_c}\|z_j\|_\infty}{\sqrt{d}}\cdot\Big\|\big(\sum_{j=1}^{s-1}\zeta_{j,s}f_j+\varepsilon_s\big)^2\Big\|_1\\ =& \frac{3T_c\|z_i\|_\infty \max_{1\le j\le T_c}\|z_j\|_\infty}{\sqrt{d}}\cdot\Big\|\sum_{j=1}^{s-1}\zeta_{j,s}f_j+\varepsilon_s\Big\|_2^2\\ \le& \frac{3T_c^2\|z_i\|_\infty \max_{1\le j\le T_c}\|z_j\|_\infty}{\sqrt{d}}\cdot\Big(\sum_{j=1}^{s-1}\zeta_{j,s}^2\|f_j\|_2^2+\|\varepsilon_s\|_2^2\Big)\\ \overset{w.h.p.}{\le}& \frac{100T_c^3(\log k)^2}{\sqrt{d}}\times\frac{d^2}{k} = O_P(d^{1.5001}/k).\end{aligned} \tag{50}$$

Eq. (48) is by the law of large numbers. In Eq. (49), we employ Eqs. (37) to (39) from previous induction steps. Eq. (50) is by Eqs. (37) to (39) from induction and the fact that with high probability, $\|z_i\|_\infty \le \log k$ for all $i \in [T_c]$.

Combining Eqs. (47) to (50) and Eq. (35) from induction, we obtain that

$$\Big\|c_{i,s+1}f_{i-1}^{\perp} - \frac{\langle \widetilde{x}_{i-1}^{\perp}, \widetilde{x}_{s-1}\rangle}{\|\widetilde{x}_{i-1}^{\perp}\|_2\|f_{i-1}^{\perp}\|_2}f_{i-1}^{\perp}\Big\|_2 \le O_P(d^{3/2}/k). \tag{51}$$

Plugging the definitions of $\{a_{i,s+1}, b_{i,s+1}, p_{i,s+1}, q_{i,s+1}, c_{i,s+1}, e_{i,s+1}\}$ into Eq. (41), we have

$$\begin{aligned}y_{s+1} - \sum_{i=1}^{s+1}\alpha_{i,s+1}z_i =& \sum_{i=2}^{s+1}(a_{i,s+1}-b_{i,s+1})f_{i-1}^{\perp}\\ =& \sum_{i=2}^{s+1}(p_{i,s+1}+q_{i,s+1}-b_{i,s+1})f_{i-1}^{\perp}\\ =& \sum_{i=2}^{s+1}(p_{i,s+1}+c_{i,s+1}-e_{i,s+1}-b_{i,s+1})f_{i-1}^{\perp}.\end{aligned}$$

Recall that we have proved $\|b_{i,s+1}f_{i-1}^{\perp}\|_2 = O_P(1)$ and $\|e_{i,s+1}f_{i-1}^{\perp}\|_2 \le O_P(\sqrt{d/k})$. Using these results, together with Eq. (46), (51) and Eq. (39) with $t = s$, we have

$$\Big\|y_{s+1} - \sum_{i=1}^{s+1}\alpha_{i,s+1}z_i - \sum_{i=2}^{s+1}\frac{\langle \widetilde{x}_{i-1}^{\perp}, \widetilde{x}_{s-1}\rangle}{\|\widetilde{x}_{i-1}^{\perp}\|_2\|f_{i-1}^{\perp}\|_2}f_{i-1}^{\perp} - \sum_{i=2}^{s+1}\frac{d\langle f_{i-1}^{\perp}, f_s\rangle}{3k\|f_{i-1}^{\perp}\|_2^2}f_{i-1}^{\perp}\Big\|_2 = o_P(d/\sqrt{k}). \tag{52}$$

Notice that

$$\sum_{i=2}^{s+1}\frac{d\langle f_{i-1}^{\perp}, f_s\rangle}{3k\|f_{i-1}^{\perp}\|_2^2}f_{i-1}^{\perp} = \frac{d}{3k}f_s.$$

Using triangle inequality, we see that

$$\begin{aligned}&\Big\|\sum_{i=2}^{s+1}\left(\frac{\langle\widetilde{x}_{i-1}^{\perp},\widetilde{x}_{s-1}\rangle}{\|\widetilde{x}_{i-1}^{\perp}\|_2\|f_{i-1}^{\perp}\|_2}-\frac{\langle x_{i-1},\widetilde{x}_{i-1}^{\perp}\rangle}{\|f_{i-1}^{\perp}\|_2^2}\cdot\frac{\langle\widetilde{x}_{i-1}^{\perp},\widetilde{x}_{s-1}\rangle}{\|\widetilde{x}_{i-1}^{\perp}\|_2^2}\right)f_{i-1}^{\perp}\Big\|_2\\ \le&\sum_{i=2}^{s+1}\Big|\frac{\langle\widetilde{x}_{i-1}^{\perp},\widetilde{x}_{s-1}\rangle}{\|\widetilde{x}_{i-1}^{\perp}\|_2}\cdot\left(1-\frac{\|x_{i-1}^{\perp}\|_2}{\|f_{i-1}^{\perp}\|_2}\right)\Big|,\end{aligned}\tag{53}$$

which by Eq. (45) is $O_P(1)$. Eqs. (52) and (53) and the condition $d \ll k \ll d^2$ together imply that

$$\Big\|y_{s+1}-\sum_{i=1}^{s+1}\alpha_{i,s+1}z_i-\sum_{i=2}^{s+1}\frac{\langle x_{i-1},\widetilde{x}_{i-1}^{\perp}\rangle}{\|f_{i-1}^{\perp}\|_2^2}\cdot\frac{\langle\widetilde{x}_{i-1}^{\perp},\widetilde{x}_{s-1}\rangle}{\|\widetilde{x}_{i-1}^{\perp}\|_2^2}f_{i-1}^{\perp}-\frac{d}{3k}f_s\Big\|_2=o_P(d/\sqrt{k}).\tag{54}$$

Using the definitions of $a_{i,s}, b_{i,s}$, we obtain that

$$\begin{aligned}\sum_{i=2}^{s+1}\frac{\langle x_{i-1},\widetilde{x}_{i-1}^{\perp}\rangle}{\|f_{i-1}^{\perp}\|_2^2}\cdot\frac{\langle\widetilde{x}_{i-1}^{\perp},\widetilde{x}_{s-1}\rangle}{\|\widetilde{x}_{i-1}^{\perp}\|_2^2}f_{i-1}^{\perp}=&\sum_{i=2}^{s}a_{i,s}f_{i-1}^{\perp}\\ =&y_s-\sum_{i=1}^{s}\alpha_{i,s}z_i+\sum_{i=2}^{s}b_{i,s}f_{i-1}^{\perp}\\ =&\sum_{j=1}^{s-1}\zeta_{j,s}f_j+\varepsilon_s+\sum_{i=2}^{s}b_{i,s}f_{i-1}^{\perp}.\end{aligned}$$

Since we have proved $\|\sum_{i=2}^{s}b_{i,s}f_{i-1}^{\perp}\|_2=O_P(1)$ and by induction $\|\varepsilon_s\|_2=o_P(d/\sqrt{k})$, we can set

$$\begin{aligned}&\varepsilon_{s+1}=y_{s+1}-\sum_{i=1}^{s+1}\alpha_{i,s+1}z_i-\sum_{j=1}^{s-1}\zeta_{j,s}f_j-\frac{d}{3k}f_s,\\ &\zeta_{j,s+1}=\zeta_{j,s}\quad\text{for } j=1,2,\cdots,s-1,\\ &\zeta_{s,s+1}=\frac{d}{3k}.\end{aligned}$$

Combining the above analysis, we find that $\|\varepsilon_{s+1}\|_2=o_P(d/\sqrt{k})$, thus proves Eq. (38). Eq. (37) for $t=s+1$ is a direct consequence of our induction hypothesis. Thus, we have completed the proof of Eq. (34) for $t=s+1$. Furthermore, using Eq. (35) for $t=s$, which holds by induction, we see that

$$\alpha_{1,s+1}=\frac{\langle\widetilde{x}_0,\widetilde{x}_s\rangle}{d}=1+o_P(1).\tag{55}$$

PROOF OF EQ. (39) FOR $t=s+1$

Next, we prove Eq. (39) for $t=s+1$. We have showed that $y_{s+1}=\sum_{i=1}^{s+1}\alpha_{i,s+1}z_i+v_{s+1}$ with $\|v_{s+1}\|_2=O_P(d/\sqrt{k})$. For the sake of simplicity, we let $g_{s+1}=\sum_{i=1}^{s+1}\alpha_{i,s+1}z_i$. We claim without proof that

$$\begin{aligned}\|f_{s+1}\|_2^2\le&\|g_{s+1}\|_6^6+6\|g_{s+1}\|_\infty^5\|v_{s+1}\|_1+15\|g_{s+1}\|_\infty^4\|v_{s+1}\|_2^2+20\|g_{s+1}\|_\infty^3\|v_{s+1}\|_3^3\\ &+15\|g_{s+1}\|_\infty^2\|v_{s+1}\|_4^4+6\|g_{s+1}\|_\infty\|v_{s+1}\|_5^5+\|v_{s+1}\|_6^6.\end{aligned}\tag{56}$$

Standard application of Gaussian concentration reveals that with high probability, $\|g_{s+1}\|_\infty\le\log k$. Furthermore, for all $j\in\{2,3,4,5,6\}$ we have $\|v_{s+1}\|_j\le\|v_{s+1}\|_2$ and $\|v_{s+1}\|_1\le\sqrt{k}\|v_{s+1}\|_2$. Using the

law of large numbers, we see that $\|g_{s+1}\|_6^6 = 15k + o_P(k)$ and $\|g_{s+1}\|_2^2 = k + o_P(k)$. Plugging the above analysis into Eq. (56), we see that

$$\|f_{s+1}\|_2^2 \le 15k + o_P(k) + O_P(d^6/k^3) \overset{(i)}{=} 15k + o_P(k), \tag{57}$$

where in *(i)* we use the assumption that $d^{3/2} \ll k$. This concludes the proof of Eq. (39) for $t = s + 1$.

PROOF OF EQ. (35) FOR $t = s + 1$

Finally, we prove Eq. (35) for $t = s + 1$. Leveraging Eq. (11) in Theorem 2.2, we have

$$\begin{aligned}
&\frac{\langle x_0, x_{s+1}\rangle}{3k} = \frac{\langle h_1, f_{s+1}\rangle}{3k} = \frac{\langle \widetilde{A}_1 \widetilde{x}_0, f_{s+1}\rangle}{3k} = \frac{1}{3k}\langle z_1, (\sum_{i=1}^{s+1} \alpha_{i,s+1} z_i + v_{s+1})^3\rangle \\
=&\frac{\alpha_{1,s+1}^3}{3k}\langle z_1^4, 1\rangle + \frac{1}{3k}\Big(\langle z_1, (\sum_{i=1}^{s+1} \alpha_{i,s+1} z_i)^3\rangle - \alpha_{1,s+1}^3 \langle z_1^4, 1\rangle\Big) + \frac{1}{k}\langle z_1 g_{s+1}^2 v_{s+1}, 1\rangle \\
&+ \frac{1}{k}\langle z_1 g_{s+1} v_{s+1}^2, 1\rangle + \frac{1}{3k}\langle z_1 v_{s+1}^3, 1\rangle.
\end{aligned}$$

Using Eq. (55) and the fact that $\sum_{i=1}^{s+1} \alpha_{i,s+1}^2 = 1$, we obtain that $\alpha_{i,s+1} = o_P(1)$ for all $2 \le i \le s+1$. Therefore, straightforward computation reveals that

$$\frac{1}{3k}\Big(\langle z_1, (\sum_{i=1}^{s+1} \alpha_{i,s+1} z_i)^3\rangle - \langle z_1^4, 1\rangle\Big) = o_P(1).$$

Furthermore,

$$\begin{aligned}
&\frac{1}{k}|\langle z_1 g_{s+1}^2 v_{s+1}, 1\rangle| + |\frac{1}{k}\langle z_1 g_{s+1} v_{s+1}^2, 1\rangle| + |\frac{1}{3k}\langle z_1 v_{s+1}^3, 1\rangle| \\
\le&\frac{1}{\sqrt{k}}\|z_1\|_\infty \|g_{s+1}\|_\infty^2 \|v_{s+1}\|_2 + \frac{1}{k}\|z_1\|_\infty \|g_{s+1}\|_\infty \|v_{s+1}\|_2^2 + \frac{1}{3k}\|z_1\|_\infty \|v_{s+1}\|_2^3 = o_P(1).
\end{aligned} \tag{58}$$

As a result, we conclude that

$$\frac{\langle x_0, x_{s+1}\rangle}{3k} = \frac{\alpha_{1,s+1}^3}{3k}\langle z_1^4, 1\rangle + o_P(1) = 1 + o_P(1).$$

This proves the second part of Eq. (35) for $s = t + 1$.

Again by Eq. (11) in Theorem 2.2, we see that

$$\frac{d}{9k^2}\|x_{s+1}\|_2^2 = \frac{d}{9k^2}\sum_{i=0}^{s} \frac{\langle h_{i+1}, f_{s+1}\rangle^2}{\|\widetilde{x}_i^\perp\|_2^2} + \frac{d}{9k^2}\|\Pi_{X_{0:s}}^\perp \overline{A}_{s+1}^\mathsf{T} f_{s+1}^\perp\|_2^2.$$

Recall that we just proved $\langle h_1, f_{s+1}\rangle^2/9k^2 = 1 + o_P(1)$. Furthermore, by the law of large numbers and Eq. (39) for $t = s + 1$, we have $d\|\Pi_{X_{0:s}}^\perp \overline{A}_{s+1}^\mathsf{T} f_{s+1}^\perp\|_2^2/9k^2 = O_P(d/k) = o_P(1)$. Therefore, in order to prove the first part of Eq. (35) for $s = t + 1$, it suffices to show

$$\frac{d\langle h_{i+1}, f_{s+1}\rangle^2}{9k^2\|\widetilde{x}_i^\perp\|_2^2} = o_P(1) \tag{59}$$

for all $1 \le i \le s$. By Eq. (13)

$$\begin{aligned}\frac{\sqrt{d}\langle h_{i+1}, f_{s+1}\rangle}{3k\|\widetilde{x}_i^\perp\|_2} =& \frac{\sqrt{d}\langle f_i^\perp, f_{s+1}\rangle}{3k\|\widetilde{x}_i^\perp\|_2}\frac{\langle x_i, \widetilde{x}_i^\perp\rangle}{\|f_i^\perp\|_2^2} + \frac{\sqrt{d}\langle \Pi_{F_{1:i}}^\perp \widetilde{A}_{i+1}\widetilde{x}_i^\perp, f_{s+1}\rangle}{3k\|\widetilde{x}_i^\perp\|_2} \\ =& \frac{\sqrt{d}\langle f_i^\perp, f_{s+1}\rangle}{3k\|f_i^\perp\|_2}\frac{\|x_i^\perp\|_2}{\|f_i^\perp\|_2} + \frac{\sqrt{d}\langle \widetilde{A}_{i+1}\widetilde{x}_i^\perp, f_{s+1}\rangle}{3k\|\widetilde{x}_i^\perp\|_2} - \sum_{j=1}^{i}\frac{\sqrt{d}\langle f_j^\perp, \widetilde{A}_{i+1}\widetilde{x}_i^\perp\rangle\langle f_j^\perp, f_{s+1}\rangle}{3k\|\widetilde{x}_i^\perp\|_2\|f_j^\perp\|_2^2}.\end{aligned}$$

Using Eq. (45), (57) and Cauchy-Schwarz inequality, we see that

$$\frac{\sqrt{d}\langle f_i^\perp, f_{s+1}\rangle}{3k\|f_i^\perp\|_2}\frac{\|x_i^\perp\|_2}{\|f_i^\perp\|_2} \le \frac{\sqrt{d}\|f_{s+1}\|_2}{3k}\frac{\|x_i^\perp\|_2}{\|f_i^\perp\|_2} = O_P(\sqrt{d/k}) = o_P(1).$$

For all $1 \le j \le i$, we have

$$\frac{\sqrt{d}\langle f_j^\perp, \widetilde{A}_{i+1}\widetilde{x}_i^\perp\rangle}{\|\widetilde{x}_i^\perp\|_2\|f_j^\perp\|_2} \overset{d}{=} \mathsf{N}(0,1).$$

Therefore,

$$\frac{\sqrt{d}\langle f_j^\perp, \widetilde{A}_{i+1}\widetilde{x}_i^\perp\rangle\langle f_j^\perp, f_{s+1}\rangle}{3k\|\widetilde{x}_i^\perp\|_2\|f_j^\perp\|_2^2} = O_P(1)\cdot\frac{\langle f_j^\perp, f_{s+1}\rangle}{k\|f_j^\perp\|_2} = o_P(1).$$

Note that

$$\begin{aligned}&\frac{\sqrt{d}\langle \widetilde{A}_{i+1}\widetilde{x}_i^\perp, f_{s+1}\rangle}{3k\|\widetilde{x}_i^\perp\|_2} \\ =& \frac{1}{3k}\langle z_{i+1}, (\sum_{i=1}^{s+1}\alpha_{i,s+1}z_i + v_{s+1})^3\rangle \\ =& \frac{\alpha_{i+1,s+1}^3}{3k}\langle z_{i+1}^4, 1\rangle + \frac{1}{3k}\Big(\langle z_{i+1}, (\sum_{i=1}^{s+1}\alpha_{i,s+1}z_i)^3\rangle - \alpha_{i+1,s+1}^3\langle z_{i+1}^4, 1\rangle\Big) + \frac{1}{k}\langle z_{i+1}g_{s+1}^2 v_{s+1}, 1\rangle \\ &+ \frac{1}{k}\langle z_{i+1}g_{s+1}v_{s+1}^2, 1\rangle + \frac{1}{3k}\langle z_{i+1}v_{s+1}^3, 1\rangle,\end{aligned}$$

which is $o_P(1)$ via a similar argument that is similar to the derivation of Eq. (58) and $\alpha_{i+1,s+1} = o_P(1)$, which we have already proved. Thus, we have completed the proof of Eq. (35) for $t = s+1$.

Appendix C. Gaussian conditioning: Proof of Lemma 2.2

The proof idea comes from (Montanari and Wu, 2022a, Lemma 3.1). We first show that for all $t \in \mathbb{N}$,

$$A = \Pi_{F_{1:t}}^\perp \widetilde{W}_{t+1}\Pi_{X_{0:t-1}}^\perp + \Pi_{F_{1:t}}A\Pi_{X_{0:t-1}}^\perp + \Pi_{F_{1:t}}^\perp A\Pi_{X_{0:t-1}} + \Pi_{F_{1:t}}A\Pi_{X_{0:t-1}}, \tag{60}$$

$$A = \Pi_{F_{1:t}}^\perp \overline{W}_{t+1}\Pi_{X_{0:t}}^\perp + \Pi_{F_{1:t}}A\Pi_{X_{0:t}}^\perp + \Pi_{F_{1:t}}^\perp A\Pi_{X_{0:t}} + \Pi_{F_{1:t}}A\Pi_{X_{0:t}}, \tag{61}$$

where $\widetilde{W}_{t+1} \overset{d}{=} \overline{W}_{t+1} \overset{d}{=} A$, and satisfy $\widetilde{W}_{t+1} \perp \mathcal{F}_{t,t}$, $\overline{W}_{t+1} \perp \mathcal{F}_{t,t+1}$. Next, we prove Eq. (60) and Eq. (61) via induction. For the base case $t = 0$, Eq. (60) holds as we can take $\widetilde{W}_1 = A$, which is independent of $\mathcal{F}_{0,0}$ by dedinition. Eq. (61) with $t = 0$ is a direct consequence of Lemma 3.1 in Montanari and Wu (2022a).

Suppose decompositions (60) and (61) hold for $t = s-1$, we then prove it also holds for $t = s$ based on induction hypothesis. We first decompose A as the sum of the following four terms:

$$A = \Pi^{\perp}_{F_{1:s}} A \Pi^{\perp}_{X_{0:s-1}} + \Pi_{F_{1:s}} A \Pi^{\perp}_{X_{0:s-1}} + \Pi^{\perp}_{F_{1:s}} A \Pi_{X_{0:s-1}} + \Pi_{F_{1:s}} A \Pi_{X_{0:s-1}}.$$

Using induction hypothesis, we see that

$$\begin{aligned}
&A = \Pi^{\perp}_{F_{1:s-1}} \overline{W}_s \Pi^{\perp}_{X_{0:s-1}} + \Pi_{F_{1:s-1}} A \Pi^{\perp}_{X_{0:s-1}} + \Pi^{\perp}_{F_{1:s-1}} A \Pi_{X_{0:s-1}} + \Pi_{F_{1:s-1}} A \Pi_{X_{0:s-1}} \\
\Longrightarrow\ &\Pi^{\perp}_{F_{1:s-1}} A \Pi^{\perp}_{X_{0:s-1}} = \Pi^{\perp}_{F_{1:s-1}} \overline{W}_s \Pi^{\perp}_{X_{0:s-1}} \Longrightarrow \Pi^{\perp}_{F_{1:s}} A \Pi^{\perp}_{X_{0:s-1}} = \Pi^{\perp}_{F_{1:s}} \overline{W}_s \Pi^{\perp}_{X_{0:s-1}} \\
\Longrightarrow\ &A = \Pi^{\perp}_{F_{1:s}} \overline{W}_s \Pi^{\perp}_{X_{0:s-1}} + \Pi_{F_{1:s}} A \Pi^{\perp}_{X_{0:s-1}} + \Pi^{\perp}_{F_{1:s}} A \Pi_{X_{0:s-1}} + \Pi_{F_{1:s}} A \Pi_{X_{0:s-1}}.
\end{aligned}$$

Since by induction we have $\overline{W}_s \perp \mathcal{F}_{s-1,s}$, we can then conclude that $\overline{W}_s \perp \{F_{1:s}, Y_{1:s}, X_{0:s-1}\}$. We take

$$\widetilde{W}_{s+1} = \Pi^{\perp}_{F_{1:s}} \overline{W}_s \Pi^{\perp}_{X_{0:s-1}} + \Pi_{F_{1:s}} W' \Pi^{\perp}_{X_{0:s-1}} + \Pi^{\perp}_{F_{1:s}} W' \Pi_{X_{0:s-1}} + \Pi_{F_{1:s}} W' \Pi_{X_{0:s-1}},$$

where W' is an independent copy of A that is independent of $\mathcal{F}_{s,s}$. We immediately see that given any specific value of $\{F_{1:s}, X_{0:s-1}, Y_{1:s}, \Pi_{F_{1:s}} \overline{W}_s, \overline{W}_s \Pi_{X_{0:s-1}}\}$, the conditional distribution of $\widetilde{W}_{s+1}$ is equal to the law of A. As a result, we deduce that $\widetilde{W}_{s+1} \perp \{F_{1:s}, X_{0:s-1}, Y_{1:s}, \Pi_{F_{1:s}} \overline{W}_s, \overline{W}_s \Pi_{X_{0:s-1}}\}$. Again by induction, we know that

$$A = \Pi^{\perp}_{F_{1:s-1}} \overline{W}_s \Pi^{\perp}_{X_{0:s-1}} + \Pi_{F_{1:s-1}} A \Pi^{\perp}_{X_{0:s-1}} + \Pi^{\perp}_{F_{1:s-1}} A \Pi_{X_{0:s-1}} + \Pi_{F_{1:s-1}} A \Pi_{X_{0:s-1}}.$$

It then follows that

$$\begin{aligned}
A^{\mathsf{T}} f_s =& \Pi^{\perp}_{X_{0:s-1}} \overline{W}_s^{\mathsf{T}} \Pi^{\perp}_{F_{1:s-1}} f_s + \Pi^{\perp}_{X_{0:s-1}} A^{\mathsf{T}} \Pi_{F_{1:s-1}} f_s + \Pi_{X_{0:s-1}} A^{\mathsf{T}} \Pi^{\perp}_{F_{1:s-1}} f_s + \Pi_{X_{0:s-1}} A^{\mathsf{T}} \Pi_{F_{1:s-1}} f_s \\
=& \overline{W}_s^{\mathsf{T}} f_s - \overline{W}_s^{\mathsf{T}} F_{1:s-1} (F_{1:s-1}^{\mathsf{T}} F_{1:s-1})^{\dagger} F_{1:s-1}^{\mathsf{T}} f_s - \Pi_{X_{0:s-1}} \overline{W}_s^{\mathsf{T}} (I - F_{1:s-1} (F_{1:s-1}^{\mathsf{T}} F_{1:s-1})^{\dagger} F_{1:s-1}^{\mathsf{T}}) f_s + \\
& X_{0:s-1} (X_{0:s-1}^{\mathsf{T}} X_{0:s-1})^{\dagger} D Y_{1:s}^{\mathsf{T}},
\end{aligned}$$

where $D = \mathrm{diag}(\{\|x_i\|_2^2/d\}_{0 \le i \le s-1})$. Therefore, we see that

$$x_s = A^{\mathsf{T}} f_s \in \sigma(\{F_{1:s}, X_{0:s-1}, Y_{1:s}, F_{1:s} \overline{W}_s, \overline{W}_s X_{0:s-1}\}).$$

This further implies that $\widetilde{W}_{s+1} \perp \sigma(\{A X_{0:s-1}, A^{\mathsf{T}} F_{1:s}, X_{0:s-1}, Y_{1:s}, F_{1:s}\}) = \sigma(\{X_{0:s}, Y_{1:s}\}) = \mathcal{F}_{s,s}$, which concludes the proof of Eq. (60) for $t = s$. The proof for Eq. (61) for $t = s$ can be shown similarly.

Next, we prove Theorem 2.2 using Eq. (60) and Eq. (61). For $t \in \mathbb{N}$, we define

$$\widetilde{A}_{t+1} = \Pi^{\perp}_{F_{1:t}} \widetilde{W}_{t+1} \Pi_{\widetilde{x}_t^{\perp}} + \Pi_{F_{1:t}} \widetilde{M}_{t+1} \Pi_{\widetilde{x}_t^{\perp}} + \Pi_{F_{1:t}} \widetilde{M}_{t+1} \Pi^{\perp}_{\widetilde{x}_t^{\perp}} + \Pi^{\perp}_{F_{1:t}} \widetilde{M}_{t+1} \Pi^{\perp}_{\widetilde{x}_t^{\perp}}, \tag{62}$$

$$\overline{A}_{t+1} = \Pi^{\perp}_{X_{0:t}} \overline{W}^{\mathsf{T}}_{t+1} \Pi_{f_{t+1}^{\perp}} + \Pi^{\perp}_{X_{0:t}} \overline{M}^{\mathsf{T}}_{t+1} \Pi^{\perp}_{f_{t+1}^{\perp}} + \Pi_{X_{0:t}} \overline{M}^{\mathsf{T}}_{t+1} \Pi^{\perp}_{f_{t+1}^{\perp}} + \Pi_{X_{0:t}} \overline{M}^{\mathsf{T}}_{t+1} \Pi_{f_{t+1}^{\perp}}. \tag{63}$$

where $\widetilde{M}_{t+1} \overset{d}{=} \overline{M}_{t+1} \overset{d}{=} A$ are i.i.d., and independent of $\sigma(\mathcal{F}_{T,T} \cup \sigma((\overline{W}_t, \widetilde{W}_t)_{0 \le t \le T+1}))$. From Eqs. (60) and (61) we see that

$$\Pi^{\perp}_{F_{1:t}} A \Pi_{\widetilde{x}_t^{\perp}} = \Pi^{\perp}_{F_{1:t}} \widetilde{W}_{t+1} \Pi_{\widetilde{x}_t^{\perp}} = \Pi^{\perp}_{F_{1:t}} \widetilde{A}_{t+1} \Pi_{\widetilde{x}_t^{\perp}} \in \mathcal{F}_{t,t+1}, \tag{64}$$

$$\Pi^{\perp}_{X_{0:t}} A^{\mathsf{T}} \Pi_{f_{t+1}^{\perp}} = \Pi^{\perp}_{X_{0:t}} \overline{W}^{\mathsf{T}}_{t+1} \Pi_{f_{t+1}^{\perp}} = \Pi^{\perp}_{X_{0:t}} \overline{A}^{\mathsf{T}}_{t+1} \Pi_{f_{t+1}^{\perp}} \in \mathcal{F}_{t+1,t+1}. \tag{65}$$

We then show that $\{\widetilde{A}_t, \overline{A}_t : t \in [T]\}$ are i.i.d. with marginal distribution A. To this end, we only need to show: (1) For all $t \in [T]$, $\widetilde{A}_t$ is independent of $\widetilde{A}_1, \cdots, \widetilde{A}_{t-1}, \overline{A}_1, \cdots, \overline{A}_{t-1}$ and has marginal distribution A; (2) For all $t \in [T]$, $\overline{A}_t$ is independent of $\widetilde{A}_1, \cdots, \widetilde{A}_{t-1}, \overline{A}_1, \cdots, \overline{A}_{t-1}$ and has marginal distribution A.

We prove this result via induction. For the base case $t = 1$, $\widetilde{A}_1 = \widetilde{W}_1\Pi_{x_0} + \widetilde{M}_1\Pi_{x_0}^{\perp}$ which obviously has marginal distribution equal to A as $\widetilde{W}_1 \perp \mathcal{F}_{0,0}$. On the other hand, $\overline{A}_1 = \Pi_{x_0}^{\perp}\overline{W}_1^{\mathsf{T}}\Pi_{f_1} + \Pi_{x_0}^{\perp}\overline{M}_1^{\mathsf{T}}\Pi_{f_1}^{\perp} + \Pi_{x_0}\overline{M}_1^{\mathsf{T}}\Pi_{f_1} + \Pi_{x_0}\overline{M}_1^{\mathsf{T}}\Pi_{f_1}^{\perp}$ and $\overline{W}_1$ is independent of $\mathcal{F}_{0,1}$, thus the conditional distribution of $\overline{A}_1$ conditioning on $\mathcal{F}_{0,1}$ is always equal to A. The marginal distribution of $\overline{A}_1$ being A follows as a simple corollary. Notice that $\widetilde{A}_1 \in \sigma(\mathcal{F}_{0,1} \cup \sigma(\widetilde{M}_1))$ and $\mathcal{F}_{0,1} \perp \widetilde{M}_1$. Therefore, in order to show $\overline{A}_1 \perp \widetilde{A}_1$, it suffices to prove $\overline{A}_1 \perp \widetilde{M}_1$ and $\overline{A}_1 \perp \mathcal{F}_{0,1}$. The first independence follows by definition. As for the second independence, since $x_0, f_1 \in \mathcal{F}_{0,1}$ and $\overline{W}_1, \overline{M}_1 \perp \mathcal{F}_{0,1}$, we can conclude that conditioning on $\mathcal{F}_{0,1}$, the conditional distribution of $\overline{A}_1$ is always equal to its marginal distribution, thus concludes the proof of this step.

Suppose the result holds for the first t steps, we then prove it also holds for $t+1$ via induction. Conditioning on $\mathcal{F}_{t,t}$, we see that the conditional distribution of $\widetilde{A}_{t+1}$ is always equal to the marginal distribution of A, thus $\widetilde{A}_{t+1}$ has marginal distribution equal to A and $\widetilde{A}_{t+1} \perp \mathcal{F}_{t,t}$. Notice that $\widetilde{A}_i \in \sigma(\mathcal{F}_{i-1,i} \cup \sigma(\widetilde{M}_i))$ and $\overline{A}_i \in \sigma(\mathcal{F}_{i,i} \cup \sigma(\overline{M}_i))$. Therefore, in order to prove $\widetilde{A}_{t+1} \perp \{\widetilde{A}_1, \cdots, \widetilde{A}_t, \overline{A}_1, \cdots, \overline{A}_t\}$, it suffices to prove $\widetilde{A}_{t+1} \perp \mathcal{F}_{t,t}$ and $\widetilde{A}_{t+1} \perp \sigma(\{\widetilde{M}_i, \overline{M}_i : i \in [t]\})$. We have already proved the first independence. The second independence follows by definition.

As for $\overline{A}_{t+1}$, first notice that the conditional distribution of $\overline{A}_{t+1}$ conditioning on $\mathcal{F}_{t,t+1}$ is always equal to the marginal distribution of A, thus $\overline{A}_{t+1}$ has marginal distribution equal to A and $\overline{A}_{t+1} \perp \mathcal{F}_{t,t+1}$. Similarly, notice that $\widetilde{A}_i \in \sigma(\mathcal{F}_{i-1,i} \cup \sigma(\widetilde{M}_i))$ and $\overline{A}_i \in \sigma(\mathcal{F}_{i,i} \cup \sigma(\overline{M}_i))$. As a result, in order to prove $\overline{A}_{t+1} \perp \{\widetilde{A}_1, \cdots, \widetilde{A}_{t+1}, \overline{A}_1, \cdots, \overline{A}_t\}$, we only need to show $\overline{A}_{t+1} \perp \mathcal{F}_{t,t+1}$ and $\overline{A}_{t+1} \perp \sigma(\{\widetilde{M}_1, \cdots, \widetilde{M}_{t+1}, \overline{M}_1, \cdots, \overline{M}_t\})$. We have already proved the first independence. The second independence follows by definition. Thus, we have concluded the proof of arguments (1) and (2) via induction.

Finally, we are ready to prove the lemma. We first show that

$$h_{t+1} = A\widetilde{x}_t^{\perp} \tag{66}$$

for all $0 \le t \le T-1$. By Eq. (60) and Eq. (64), we have

$$\begin{aligned}
A\widetilde{x}_t^{\perp} =& \Pi_{F_{1:t}}^{\perp}\widetilde{W}_{t+1}\Pi_{X_{0:t-1}}^{\perp}\widetilde{x}_t^{\perp} + \Pi_{F_{1:t}}A\widetilde{x}_t^{\perp} \\
=& \Pi_{F_{1:t}}^{\perp}\widetilde{A}_{t+1}\widetilde{x}_t^{\perp} + \sum_{i=1}^{t} f_i^{\perp} \cdot \frac{\langle A^{\mathsf{T}}f_i^{\perp}, \widetilde{x}_t^{\perp}\rangle}{\|f_i^{\perp}\|_2^2} \\
=& \Pi_{F_{1:t}}^{\perp}\widetilde{A}_{t+1}\widetilde{x}_t^{\perp} + f_t^{\perp} \cdot \frac{\langle x_t, \widetilde{x}_t^{\perp}\rangle}{\|f_t^{\perp}\|_2^2} = h_{t+1},
\end{aligned}$$

which concludes the proof of Eq. (66). Now, by definition, we deduce that

$$\begin{aligned}
x_t = A^{\mathsf{T}}f_t =& \Pi_{X_{0:t-1}}^{\perp}A^{\mathsf{T}}f_t + \sum_{i=0}^{t-1} \frac{\langle \widetilde{x}_i^{\perp}, A^{\mathsf{T}}f_t\rangle}{\|\widetilde{x}_i^{\perp}\|_2^2} \cdot \widetilde{x}_i^{\perp} \\
\overset{(i)}{=}& \Pi_{X_{0:t-1}}^{\perp}A^{\mathsf{T}}f_t^{\perp} + \sum_{i=0}^{t-1} \frac{\langle A\widetilde{x}_i^{\perp}, f_t\rangle}{\|\widetilde{x}_i^{\perp}\|_2^2} \cdot \widetilde{x}_i^{\perp} \\
=& \Pi_{X_{0:t-1}}^{\perp}A^{\mathsf{T}}f_t^{\perp} + \sum_{i=0}^{t-1} \frac{\langle h_{i+1}, f_t\rangle}{\|\widetilde{x}_i^{\perp}\|_2^2} \cdot \widetilde{x}_i^{\perp} \\
\overset{(ii)}{=}& \Pi_{X_{0:t-1}}^{\perp}\overline{A}_t^{\mathsf{T}}f_t^{\perp} + \sum_{i=0}^{t-1} \frac{\langle h_{i+1}, f_t\rangle}{\|\widetilde{x}_i^{\perp}\|_2^2} \cdot \widetilde{x}_i^{\perp},
\end{aligned}$$

where (i) follows from the fact that $\Pi^{\perp}_{X_{0:t-1}} A^{\mathsf{T}} f_i = \Pi^{\perp}_{X_{0:t-1}} x_i = 0$ for $0 \le i \le t-1$, and (ii) follows from Eq. (65). Similarly, we obtain that

$$\begin{aligned} y_{t+1} = A\widetilde{x}_t &= \sum_{i=1}^{t+1} \frac{\langle \widetilde{x}^{\perp}_{i-1}, \widetilde{x}_t \rangle}{\|\widetilde{x}^{\perp}_{i-1}\|_2^2} \cdot A\widetilde{x}^{\perp}_{i-1} = \sum_{i=1}^{t+1} \frac{\langle \widetilde{x}^{\perp}_{i-1}, \widetilde{x}_t \rangle}{\|\widetilde{x}^{\perp}_{i-1}\|_2^2} \cdot h_i \\ &= \sum_{i=1}^{t+1} \Pi^{\perp}_{F_{1:i-1}} \widetilde{A}_i \widetilde{x}^{\perp}_{i-1} \frac{\langle \widetilde{x}^{\perp}_{i-1}, \widetilde{x}_t \rangle}{\|\widetilde{x}^{\perp}_{i-1}\|_2^2} + \sum_{i=2}^{t+1} f^{\perp}_{i-1} \cdot \frac{\langle x_{i-1}, \widetilde{x}^{\perp}_{i-1} \rangle}{\|f^{\perp}_{i-1}\|_2^2} \cdot \frac{\langle \widetilde{x}^{\perp}_{i-1}, \widetilde{x}_t \rangle}{\|\widetilde{x}^{\perp}_{i-1}\|_2^2}. \end{aligned}$$

This completes the proof of Theorem 2.2.

Appendix D. Supporting lemmas

This section contains supporting lemmas required by our proof.

Lemma D.1 (Tails of the normal distribution, Proposition 2.1.2 of Vershynin (2018)) *Let $g \sim \mathsf{N}(0,1)$. Then for all $t > 0$ we have*

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

Lemma D.2 (Bernstein's inequality, Theorem 2.8.1 of Vershynin (2018)) *Let $X_1, \cdots, X_N$ be independent, mean zero, sub-exponential random variables. Then, for every $t \ge 0$, we have*

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

where $c > 0$ is an absolute constant, and $\|\cdot\|_{\Psi_1}$ is the Orlicz norm.

Lemma D.3 (Concentration inequality for sub-Weibull distribution, adapted from Theorem 3.1 of Hao et al. (2019)) *Let $\{X_i\}_{1\le i\le n}$ be a sequence of i.i.d. random variables such that for some constants $C_1, C_2 > 0$ and $q \in (0,1)$, $\mathbb{P}(|X_1| \ge t) \le C_1 \exp(-C_2 t^q)$ for all sufficiently large $t > 0$. Then, there exists a constant $C > 0$, such that the following holds for all $t > 0$:*

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} (X_i - \mathbb{E}[X_i])\right| \ge t\right) \le C\exp\left(-\min\left\{\frac{C^{-2}t^2}{n},\ C^{-q}t^q\right\}\right).$$

Proof Choosing a to be the vector of all ones in (Hao et al., 2019, Theorem 3.1), we know that there exists a constant $C > 0$, such that

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} (X_i - \mathbb{E}[X_i])\right| \ge C\left(\sqrt{n\log(1/\alpha)} + (\log(1/\alpha))^{1/q}\right)\right) \le \alpha.$$

Set $t = C\left(\sqrt{n\log(1/\alpha)} + (\log(1/\alpha))^{1/q}\right)$, then one of the followings must happen:

(a) $C\sqrt{n\log(1/\alpha)} \ge t/2$. This implies that

$$\alpha \le \exp\left(-\frac{t^2}{4C^2 n}\right).$$

(b) $C(\log(1/\alpha))^{1/q} \geq t/2$, which is equivalent to

$$\alpha \leq \exp\left(-\frac{t^q}{(2C)^q}\right).$$

Hence, we finally obtain that

$$\begin{aligned}\mathbb{P}\left(\left|\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i])\right| \geq t\right) &\leq \alpha \leq \max\left\{\exp\left(-\frac{t^2}{4C^2n}\right),\ \exp\left(-\frac{t^q}{(2C)^q}\right)\right\} \\ &= \exp\left(-\min\left\{\frac{t^2}{4C^2n},\ \frac{t^q}{(2C)^q}\right\}\right).\end{aligned}$$

This completes the proof. ■