Stochastic Approximation for Online Tensorial Independent Component Analysis

Yuren Zhou [*] Chris Junchi Li [†] Michael I. Jordan [‡]

February 3, 2020

Abstract

Independent component analysis (ICA) has been a popular dimension reduction tool in statistical machine learning and signal processing. In this paper, we present a convergence analysis for an online tensorial ICA algorithm, by viewing the problem as a nonconvex stochastic approximation problem. For estimating one component, we prove that our online tensorial ICA algorithm with a specific choice of stepsize achieves a sharp finite-sample error bound. In particular, under a mild assumption on the data-generating distribution and a scaling condition such that d^4/T is sufficiently small up to a polylogarithmic factor of data dimension d and sample size T, a finite-sample error bound of $\tilde{O}(\sqrt{d/T})$ can be obtained.

Keywords: Independent component analysis, tensor decomposition, non-Gaussianity, finite-sample error bound, online learning

1 Introduction

Independent Component Analysis (ICA) is a widely used dimension reduction method with diverse applications in the fields of statistical machine learning and signal processing (Hyvärinen et al., 2001; Stone, 2004; Samworth & Yuan, 2012). Let the data vector be modeled as $\boldsymbol{X} = \mathbf{A}\boldsymbol{Z}$, where $\mathbf{A} \equiv (\mathbf{a}_1, \ldots, \mathbf{a}_d) \in \mathbb{R}^{d\times d}$ is a full-rank *mixing matrix* whose columns are orthogonal components in $\mathbb{R}^d$, and $\boldsymbol{Z} \in \mathbb{R}^d$ is a *non-Gaussian latent random vector* consisting of independent entries $\boldsymbol{Z} = (Z_1, \ldots, Z_d)^\top$. The goal of ICA is to recover one or multiple columns among $(\mathbf{a}_1, \ldots, \mathbf{a}_d)$ from independent observations of $\boldsymbol{X} = (X_1, \ldots, X_d)^\top$. Following standard practice, we assume that the random vector $\boldsymbol{X}$ has been *whitened* in the sense that it has zero mean and an identity covariance matrix (Hyvärinen et al., 2001), and we focus on the case where the distributions of $Z_1, \ldots, Z_d$ share a fourth moment $\mu_4 \neq 3$ (i.e., they are of *identical kurtosis* $\mu_4 - 3$). These assumptions

[*]Department of Statistical Science, Duke University, Durham, NC 27710; email: yuren.zhou@duke.edu

[†]Department Electrical Engineering and Computer Sciences, UC Berkeley, Berkeley, CA, USA 94720; email: junchili@berkeley.edu

[‡]Department Electrical Engineering and Computer Sciences & Department of Statistics, UC Berkeley, Berkeley, CA, USA 94720; email: jordan@cs.berkeley.edu

restrict the search of mixing matrix $\mathbf{A}$ to the space of orthogonal matrices and guarantee its identifiability up to signed permutations of its columns $(\mathbf{a}_1, \ldots, \mathbf{a}_d)$ (Comon, 1994; Hyvärinen et al., 2001).

In this paper, we study a stochastic algorithm that estimates independent components for streaming data. Such an algorithm processes and discards one or a small batch of data observations at each iterate and enjoys reduced storage complexity. To begin with, we cast the *tensorial ICA* problem as the problem of optimizing a stochastic function based on the fourth-order cumulant tensor over the unit sphere $\mathcal{D}_1 \equiv \{\mathbf{u} \in \mathbb{R}^d : \|\mathbf{u}\| = 1\}$. This function is referred to as a *non-Gaussianity contrast function*, where $\|\cdot\|$ denotes the Euclidean norm) (Hyvärinen et al., 2001). The optimization problem is as follows:

$$\begin{aligned} &\min \ -\operatorname{sign}(\mu_4 - 3)\mathbb{E}(\mathbf{u}^\top \boldsymbol{X})^4 \\ &\text{subject to } \ \mathbf{u} \in \mathcal{D}_1 \end{aligned} \tag{1.1}$$

The landscape of objective (1.1) is highly *nonconvex*, in the sense that it presents $2d$ local minimizers $\pm\mathbf{a}_1, \ldots, \pm\mathbf{a}_d$ and (in terms of d) exponentially many saddle points (Ge et al., 2015; Li et al., 2016; Sun et al., 2015). Here by saddle points, we refer to those points with zero gradient and at least one negative Hessian direction, and hence includes the collection of local maximizers. Algorithms that find local minimizers of (1.1) allow us to estimate the columns of the mixing matrix $\mathbf{A}$. We analyze and discuss the following stochastic approximation method, which we refer to as *online tensorial ICA* (Ge et al., 2015; Li et al., 2016; Wang & Lu, 2017). Initialized at a unit vector $\mathbf{u}^{(0)}$, at step $t = 1, 2, \ldots, T$ the algorithm processes an observation $\boldsymbol{X}^{(t)}$ and performs the following update:

$$\mathbf{u}^{(t)} = \Pi_1 \left\{ \mathbf{u}^{(t-1)} + \eta^{(t)} \cdot \operatorname{sign}(\mu_4 - 3) \left((\mathbf{u}^{(t-1)})^\top \boldsymbol{X}^{(t)} \right)^3 \boldsymbol{X}^{(t)} \right\}. \tag{1.2}$$

Here, $\eta^{(t)}$ is a positive stepsize, and the operator $\Pi_1[\bullet] := \bullet / \|\bullet\|$ projects a nonzero vector onto the unit sphere $\mathcal{D}_1$ centered at the origin.

Due to the nonconvexity of the optimization problem (1.1), a key issue that arises in analyzing the iteration (1.2) is the avoidance of *unstable stationary points, a.k.a. saddle points*. In earlier work, Ge et al. (2015) introduced an additional artificial noise injection step to the algorithm and developed a *hit-and-escape* convergence analysis, obtaining a polynomial convergence rate for the online tensorial ICA problem (in fact, for a more general class of nonconvex optimization problems). In contrast, we present a *dynamics-based* approach that requires no noise addition step. We show that a uniform initialization on the unit sphere $\mathcal{D}_1$ along with mild scaling conditions is enough to ensure that the algorithm enters a basin of attraction and finds a local minimizer with high probability.

Overview of Main Results For estimating a single independent component, our main result states that under mild distributional assumptions, and the scaling condition that d^4/T is sufficiently small up to a polylogarithmic factor of d and T, the iteration (1.2) with uniform initialization and carefully chosen step-sizes $\eta^{(t)}$ enters the basin of attraction of a uniformly drawn independent component $\pm\mathbf{a}_{\mathscr{I}}$ and achieves a

convergence rate of $\tilde{O}(\sqrt{d/T})$ with high probability. Informally, this result is stated as follows.

Theorem 1 (Informal version of Corollary 4 in §3)**.** *Given appropriate initialization and distributional assumptions, and letting the initial* $\mathbf{u}^{(0)}$ *be uniformly sampled from the unit sphere* $\mathcal{D}_1$*, there exist an appropriate choice of stepsizes* $\eta^{(t)}$ *such that for any fixed positive* $\varepsilon \in (0, 1/5]$ *satisfying the scaling condition*

$$d \geq 2\sqrt{2\pi e}\log\varepsilon^{-1} + 1, \quad C_{1,T}\log^8(C_{1,T}\varepsilon^{-1}dT)\cdot\frac{B^8}{|\mu_4-3|^2}\cdot\frac{d^4\log^2 T}{T} \leq \frac{\varepsilon^2}{\log^2\varepsilon^{-1}},$$

there exists a uniformly distributed random variable $\mathscr{I} \in [d]$ *such that with probability at least* $1-5\varepsilon$*, we have:*

$$\left|\tan\angle\left(\mathbf{u}^{(T)}, \mathbf{a}_{\mathscr{I}}\right)\right| \leq C_{1,T}\log^{5/2}(C_{1,T}\varepsilon^{-1}d)\cdot\frac{B^4}{|\mu_4-3|}\cdot\sqrt{\frac{d\log^2 T}{T}},$$

where $C_{1,T}$ *is a positive, absolute constant.*

To the best of our knowledge, Theorem 1 [presented formally in Corollary 4] provides the first rigorous analysis of an online tensorial ICA algorithm that achieves a $\tilde{O}(\sqrt{d/T})$ finite-sample convergence rate. Our online tensorial ICA analysis proceeds in several stages. Partly adapting from the analysis of online principal component estimation in Li et al. (2018), we provide a single analysis for both warm initialization and uniform initialization cases.[1] The analysis carries through to estimate multiple independent components, where we parallelize our tensorial ICA algorithm on N machines with i.i.d. uniform initializations. As a side result in the Appendix [Theorem 5], we use a simple combinatoric argument to show that we can find multiple independent components (or all components when $N = \Theta(d\log d)$) with desirable error bounds with high probability.

This paper makes two contributions. First, we prove for the first time a finite-sample error bound of $\tilde{O}(\sqrt{d/T})$, under a mild assumption on the data-generating distribution and a scaling condition such that d^4/T is sufficiently small up to a polylogarithmic factor of data dimension d and sample size T. Second, we also design online tensorial ICA algorithm that estimates multiple independent components in parallel, achieving desirable finite-sample error bound for each independent component estimator.

The rest of this paper is organized as follows. §2 presents our main convergence results and finite-sample error bounds for warm initialization for estimating one single component. §3 handles the uniform initialization case. §5 summarizes this paper. Proofs of all secondary lemmas and technical results are deferred to the Appendix.

Notations Throughout this paper, we treat B, μ_4, τ as positive constants. We use bold upper case letters to denote matrices, bold lower case letters to denote vectors and italic letters to denote randomness. For any matrix $\mathbf{A}$ or vector $\mathbf{v}$, $\mathbf{A}^\top$ and $\mathbf{v}^\top$ denote their transposes. For any vector $\mathbf{v}$, v_k denotes its kth coordinate.

[1] Related approaches have been suggested in other nonconvex optimization settings (Ge et al., 2015): see also Li et al. (2016); Wang (2017); Wang & Lu (2017) for related methods from the viewpoint of scaling limits and stochastic differential equation approximations.

Algorithm 1 Online Tensorial ICA, Single Component

Initialize $\mathbf{u}^{(0)}$ and select stepsize $\eta^{(t)}$ appropriately (to elaborate later)

for $t = 1, 2, \ldots$ **do**

Draw one observation $\boldsymbol{X}^{(t)}$ from streaming data, and update iteration $\mathbf{u}^{(t)}$ via

$$\mathbf{u}^{(t)} = \Pi_1 \left\{ \mathbf{u}^{(t-1)} + \eta \cdot \mathrm{sign}(\mu_4 - 3) \left((\mathbf{u}^{(t-1)})^\top \boldsymbol{X}^{(t)} \right)^3 \boldsymbol{X}^{(t)} \right\} \tag{1.3}$$

where $\Pi_1\{\bullet\} = \|\bullet\|^{-1}\bullet$ denotes the projection operator onto the unit sphere centered at the origin $\mathcal{D}_1$

end for

For a sequence of $\{x^{(t)}\}$ and positive $\{y^{(t)}\}$, we write $x^{(t)} = O(y^{(t)})$ if there exists a positive constant M such that $|x^{(t)}| \le M y^{(t)}$, write $x^{(t)} = \Omega(y^{(t)})$ if there exists a positive constant $M < \infty$ such that $|x^{(t)}| \ge M y^{(t)}$, and write $x^{(t)} = \Theta(y^{(t)})$ if both $x^{(t)} = O(y^{(t)})$ and $x^{(t)} = \Omega(y^{(t)})$ hold. We use $\tilde{O}, \tilde{\Theta}, \tilde{\Omega}$ to hide factors that are polylogarithmically dependent on dimension d, stepsize η, sample size T and inverse error probability δ^{-1}. We use $\lfloor x \rfloor$ to denote the floor function and $\lceil x \rceil$ to denote the ceiling function. We let $x \wedge y = \min(x, y)$ and $x \vee y = \max(x, y)$. For any vector $\mathbf{v}$, we use $\|\mathbf{v}\|$ to denote its Euclidean norm. For any integer n, we define set $[n] = \{1, \ldots, n\}$. Finally, we use $\mathscr{S}^c$ to denote the complement of set (or event) $\mathscr{S}$.

2 Estimating Single Component: Warm Initialization Case

For the purpose of estimating one single independent component, we introduce our settings and assumptions of tensorial ICA and its stochastic approximation algorithm (1.2), formally stated in Algorithm 1. Let the dimension $d \ge 2$, let $\boldsymbol{X}$ be the data vector of which $\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}, \ldots \in \mathbb{R}^d$ are independent data observations, and assume the following on the distribution of $\boldsymbol{X}$:

Assumption 1 (Data vector distribution)**.** *Let $\boldsymbol{X} = \mathbf{A}\boldsymbol{Z}$, where $\mathbf{A} \in \mathbb{R}^{d \times d}$ is an orthogonal matrix with $\mathbf{A}\mathbf{A}^\top = \mathbf{I}$ and $\boldsymbol{Z} \in \mathbb{R}^d$ is a random vector satisfying*

(i) The $Z_i, i = 1, \ldots, d$ are independent with identical jth-moment for $j = 1, 2, 4$, denoted as $\mu_j \equiv \mathbb{E}Z_i^j$;

(ii) The $\mu_1 = \mathbb{E}Z_i = 0$, $\mu_2 = \mathbb{E}Z_i^2 = 1$, $\mu_4 = \mathbb{E}Z_i^4 \ne 3$;

(iii) For all $i \in [d]$, Z_i is sub-Gaussian with parameter $\sqrt{3/8}B$.[2]

In above, Assumption 1(i) requires the distribution for all independent components to admit *identical first, second and fourth moments*. As indicated in Assumption 1(ii), the data vector are assumed to be *whitened* first in the sense that $\mu_2 = 1$. The sign of our *excess kurtosis* $\mu_4 - 3$ determines the direction of stochastic gradient update, and as the readers will see later, the magnitude of the excess kurtosis $|\mu_4 - 3|$

[2] A random variable Z with mean 0 is sub-Gaussian if there is a positive number σ such that $\mathbb{E}\exp(\lambda Z) \le \exp(\sigma^2\lambda^2/2)$ for all $\lambda \in \mathbb{R}$. The constant σ is referred to as the sub-Gaussian parameter (Wainwright, 2019).

plays an important role in our convergence analysis.[3] Assumption 1(iii) generalizes the common boundedness assumption $\|Z_i\|_\infty \le O(B)$ to include distributions such as mixture Gaussian and Gaussian-Bernoulli distributions, which are typical models for ICA. We multiply by a factor of $\sqrt{3/8}$ in the subgaussian parameter for notational simplicity in our analysis.

We target to study the convergence of tensorial ICA under certain initialization conditions. For each initialization condition, we first analyze the convergence result for any fixed, plausible stepsizes, and then (by choosing the stepsize according to the number of observations) obtain the finite-sample error bound. We focus in this section the warm initialization condition as any $\mathbf{u}^{(0)}$ satisfying, for some integer $i \in [d]$,

$$\mathbf{u}^{(0)} \in \mathcal{D}_1, \qquad \text{and } \left|\tan\angle\left(\mathbf{u}^{(0)}, \mathbf{a}_i\right)\right| \le \frac{1}{\sqrt{3}}. \tag{2.1}$$

For any fixed $\tau > 0$, we define a rescaled time $T^*_{\eta,\tau}$ as

$$T^*_{\eta,\tau} \equiv \left\lceil \frac{\tau \log\left(\frac{|\mu_4-3|}{B^8}\cdot\eta^{-1}\right)}{-\log\left(1-\frac{\eta}{3}|\mu_4-3|\right)} \right\rceil. \tag{2.2}$$

Then under warm initialization condition (2.1), we have the following convergence lemma.

Lemma 1 (Convergence Result with Warm Initialization)**.** *Let the dimension $d \ge 2$, let Assumption 1 hold, and let initialization $\mathbf{u}^{(0)}$ satisfy condition (2.1) for some integer $i \in [d]$. Then for any fixed positives τ, η and $\delta \in (0, e^{-1}]$ satisfying the scaling condition*

$$C^*_{1,L}\log^8(T^*_{\eta,1}\delta^{-1})\cdot\frac{B^8}{|\mu_4-3|}\cdot d\eta\log\left(\frac{|\mu_4-3|}{B^8}\eta^{-1}\right) \le \frac{1}{\tau+1}, \quad \eta < \min\left(\frac{1}{|\mu_4-3|}, \frac{|\mu_4-3|}{B^8}e^{-1}\right), \tag{2.3}$$

there exists an event $\mathscr{H}_{1,L}$ with

$$\mathbb{P}(\mathscr{H}_{1,L}) \ge 1 - \left(6\tau + 12 + \frac{5184}{\log^5\delta^{-1}}\right)d\delta,$$

such that on $\mathscr{H}_{1,L}$, iteration $\mathbf{u}^{(t)}$ of Algorithm 1 satisfies for all $t \in [0, T^*_{\eta,\tau}]$

$$\begin{aligned}\left|\tan\angle\left(\mathbf{u}^{(t)}, \mathbf{a}_i\right)\right| \le{}& \left|\tan\angle\left(\mathbf{u}^{(0)}, \mathbf{a}_i\right)\right|\left(1-\frac{\eta}{3}|\mu_4-3|\right)^t \\ &+ \sqrt{\tau+1}C_{1,L}\log^{5/2}\delta^{-1}\cdot\frac{B^4}{|\mu_4-3|^{1/2}}\cdot\sqrt{d\eta\log\left(\frac{|\mu_4-3|}{B^8}\eta^{-1}\right)},\end{aligned} \tag{2.4}$$

where $C_{1,L}$ and $C^*_{1,L}$ are positive, absolute constants.

Lemma 1 provides, under the warm initialization assumption (2.1) and scaling condition $\eta = \tilde{O}(d^{-1})$,

[3] When the excess kurtosis $\mu_4 - 3 = 0$, for instance when $\boldsymbol{Z}$ follows the i.i.d. standard normal distribution, matrix $\mathbf{A}$ is *non-identifiable* in our tensorial ICA framework. Non-gaussian independent components with $\mu_4 = 3$ can be studied via higher-order tensor decomposition with a different contrast function, but is beyond the scope of this paper.

an upper bound for $\left|\tan\angle\left(\mathbf{u}^{(t)},\mathbf{a}_i\right)\right|$ which is the addition of two terms: the first term on the right hand side of (A.21) decays geometrically from $\sqrt{d}$ at rate $1-|\mu_4-3|\eta/3$, and the second term $\tilde{O}(\sqrt{d\eta})$ is incurred by the stochastic noise. To balance these two terms, when we know in advance the sample size T of online data satisfying some scaling condition $T=\tilde{\Theta}(d)$, we choose a constant stepsize $\eta=\tilde{\Theta}(\log T/T)$ and establish based on Lemma 1 a finite-sample error bound:

Theorem 2 (Finite-Sample Error Bound with Warm Initialization)**.** *Let the dimension $d\geq 2$, let Assumption 1 hold, and let initialization $\mathbf{u}^{(0)}$ satisfy condition (2.1) for some integer $i\in[d]$. Set for sample size $T\geq 2$ the stepsize η as*

$$\eta(T)=\frac{9\log\left(\frac{2|\mu_4-3|^2}{9B^8}T\right)}{2|\mu_4-3|T}. \tag{2.5}$$

Then for any fixed positives $T\geq 2,\varepsilon\in(0,1]$ satisfying the scaling condition

$$C^*_{2,T}\log^8(C'_{2,T}\varepsilon^{-1}dT)\cdot\frac{d\log^2 T}{T}\leq\frac{|\mu_4-3|^2}{B^8}, \tag{2.6}$$

there exists an event $\mathscr{H}_{2,T}$ with $\mathbb{P}(\mathscr{H}_{2,T})\geq 1-\varepsilon$ such that on $\mathscr{H}_{2,T}$, iteration $\mathbf{u}^{(t)}$ of Algorithm 1 satisfies

$$\left|\tan\angle\left(\mathbf{u}^{(T)},\mathbf{a}_i\right)\right|\leq C_{2,T}\log^{5/2}(C'_{2,T}\varepsilon^{-1}d)\cdot\frac{B^4}{|\mu_4-3|}\cdot\sqrt{\frac{d\log^2 T}{T}},$$

where $C_{2,T},C^*_{2,T},C'_{2,T}$ are positive, absolute constants.

Theorem 2 is a new result. It obtains an $\tilde{O}\left(\sqrt{d/T}\right)$ finite-sample error bound for online tensorial ICA when it is warmly initialized in the sense of (2.1).

For the rest of this section, we target to prove Lemma 1 and Theorem 2 for the warm initialization case, organized as follows. §2.1 analyzes our algorithm and provide a key lemma [Lemma 2] on the iteration when it is warmly initialized. §A.1 and §A.2 (in Appendix due to space limit) prove in sequel Lemma 1 and Theorem 2. Proof of the key Lemma 2 and all proofs of secondary lemmas are deferred to §C in Appendix.

2.1 Key Lemma in the Warm Initialization Analysis

To simplify our problem, we set $i\in[d]$ as the integer such that $\mathbf{u}^{(0)}$ satisfies condition (2.1) and define the *rotated iteration* $\{\mathbf{v}^{(t)}\}_{t\geq 0}$ as

$$\mathbf{v}^{(t)}\equiv\mathbf{P}\mathbf{A}^\top\mathbf{u}^{(t)}, \tag{2.7}$$

where $\mathbf{P}\in\mathbb{R}^{d\times d}$ is the permutation matrix corresponding to the cycle $(1i)$, i.e. $\mathbf{P}(i;1)=\mathbf{P}(1;i)=1,\mathbf{P}(j;j)=1$ for $j\neq 1,i$ and all other elements being zero. Such a matrix, as an operator, maps the component vectors to coordinate vectors and ensures that $\pm\mathbf{e}_1$ is the closest independent components pair at initialization and (with high probability) at convergence. Furthermore, letting the *rotated observations* $\boldsymbol{Y}^{(t)}=\mathbf{P}\mathbf{A}^\top\boldsymbol{X}^{(t)}$ allows us

to equivalently translate our online tensorial ICA iteration (1.3) into an analogous form:

$$\mathbf{v}^{(t)} = \Pi_1 \left\{ \mathbf{v}^{(t-1)} + \eta \cdot \mathrm{sign}(\mu_4 - 3) \left((\mathbf{v}^{(t-1)})^\top \boldsymbol{Y}^{(t)} \right)^3 \boldsymbol{Y}^{(t)} \right\}. \tag{2.8}$$

Indeed, left-multiplying both sides of (1.3) by orthogonal matrix $\mathbf{P}\mathbf{A}^\top$ gives

$$\begin{aligned} \mathbf{v}^{(t)} = \mathbf{P}\mathbf{A}^\top \mathbf{u}^{(t)} &= \Pi_1 \left\{ \mathbf{P}\mathbf{A}^\top \mathbf{u}^{(t-1)} + \eta \cdot \mathrm{sign}(\mu_4 - 3) \left((\mathbf{P}\mathbf{A}^\top \mathbf{u}^{(t-1)})^\top \mathbf{P}\mathbf{A}^\top \boldsymbol{X}^{(t)} \right)^3 \mathbf{P}\mathbf{A}^\top \boldsymbol{X}^{(t)} \right\} \\ &= \Pi_1 \left\{ \mathbf{v}^{(t-1)} + \eta \cdot \mathrm{sign}(\mu_4 - 3) \left((\mathbf{v}^{(t-1)})^\top \boldsymbol{Y}^{(t)} \right)^3 \boldsymbol{Y}^{(t)} \right\}, \end{aligned}$$

proving (2.8). It is easy to verify that the rotated iterations $\{\mathbf{v}^{(t)}\}_{t\geq 0}$ and $\{\mathbf{u}^{(t)}\}_{t\geq 0}$ satisfy $\mathbf{a}_i^\top \mathbf{u}^{(t)} = \mathbf{e}_1^\top \mathbf{v}^{(t)}$, and hence for all $t \geq 0$

$$\tan \angle \left(\mathbf{v}^{(t)}, \mathbf{e}_1 \right) = \frac{\sqrt{1 - (\mathbf{e}_1^\top \mathbf{v}^{(t)})^2}}{\mathbf{e}_1^\top \mathbf{v}^{(t)}} = \frac{\sqrt{1 - (\mathbf{a}_i^\top \mathbf{u}^{(t)})^2}}{\mathbf{a}_i^\top \mathbf{u}^{(t)}} = \tan \angle \left(\mathbf{u}^{(t)}, \mathbf{a}_i \right). \tag{2.9}$$

Now, we let the *warm initialization region* be

$$\mathcal{D}_{\text{warm}} = \left\{ \mathbf{v} \in \mathcal{D}_1 : v_1^2 \geq \frac{3}{4} \right\} = \left\{ \mathbf{v} \in \mathcal{D}_1 : \left| \tan \angle (\mathbf{v}, \mathbf{e}_1) \right| \leq \frac{1}{\sqrt{3}} \right\}. \tag{2.10}$$

Note the warm initialization condition in (2.1) is simply equivalent to $\mathbf{v}^{(0)} \in \mathcal{D}_{\text{warm}}$. Analogous to the warm initialization study in the case of principal component estimation (Li et al., 2018), the iteration we study in our warmly initialized online tensorial ICA is

$$U_k^{(t)} \equiv \frac{v_k^{(t)}}{v_1^{(t)}}. \tag{2.11}$$

To bound iteration U_k from getting far away from the warm initialization region, we also define a slightly larger *warm-auxiliary region* as

$$\mathcal{D}_{\text{warm-aux}} \equiv \left\{ \mathbf{v} \in \mathcal{D}_1 : v_1^2 \geq \frac{2}{3} \right\} = \left\{ \mathbf{v} \in \mathcal{D}_1 : \left| \tan \angle (\mathbf{v}, \mathbf{e}_1) \right| \leq \frac{1}{\sqrt{2}} \right\}. \tag{2.12}$$

Suppose the process $\{\mathbf{v}^{(t)}\}$ is initialized at $\mathbf{v}^{(0)} \in \mathcal{D}_{\text{warm}}$, and we define the first time $\mathbf{v}^{(t)}$ exits the warm-auxiliary region as

$$\mathscr{T}_x \equiv \inf \left\{ t \geq 1 : \mathbf{v}^{(t)} \in \mathcal{D}_{\text{warm-aux}}^c \right\}. \tag{2.13}$$

We state the key lemma as

Lemma 2. *Let the settings in Lemma 1 hold, and fix the coordinate $k \in [2, d]$ and value $\tau > 0$. Then for any fixed positives η, δ satisfying the scaling condition (2.3) along with the warm initialization condition*

$\mathbf{v}^{(0)} \in \mathcal{D}_{warm}$, *there exists an event* $\mathscr{H}_{k;2,L}$ *satisfying*

$$\mathbb{P}(\mathscr{H}_{k;2,L}) \geq 1 - \left(6\tau + 12 + \frac{5184}{\log^5 \delta^{-1}}\right)\delta,$$

such that on event $\mathscr{H}_{k;2,L}$ *the following holds*

$$\sup_{t \leq T^*_{\eta,\tau} \wedge \mathscr{T}_x} \left| U_k^{(t)} - U_k^{(0)} \prod_{s=0}^{t-1} \left[1 - \eta|\mu_4 - 3| \left((v_1^{(s)})^2 - (v_k^{(s)})^2\right)\right] \right| \leq 2C_{2,L} \log^{5/2} \delta^{-1} \cdot B^4 \cdot \eta (T^*_{\eta,\tau})^{1/2}, \quad (2.14)$$

where $C_{2,L}$ *is a positive, absolute constant.*

The key lemma 2 for the warm initialization case shows for each coordinate $k \in [2,d]$ that, with high probability, the dynamics of $U_k^{(t)}$ is tightly controlled within a deterministic vessel whose center converges to 0 at least exponentially fast. As we will later see in the proofs of Lemma 1 and Theorem 2, this guarantees with high probability $\mathbf{v}^{(t)}$ not to exit the warm-auxilliary region $\mathcal{D}_{\text{warm-aux}}$ and to stay within a small neighborbood of $\pm\mathbf{e}_1$ after $T^*_{\eta,0.5}$ iterates, where the rescaled time $T^*_{\eta,\tau}$ was earlier defined in (2.2).

Due to the limitation of space, Lemma 2 is proved in §C.1 in Appendix.

3 Estimating Single Component: Uniform Initialization Case

We are often unable to obtain a warm initialization for online tensorial ICA, and the best we can hope for is to initialize $\mathbf{u}^{(0)}$ uniformly at random from the unit sphere $\mathcal{D}_1$. To proceed with such a case, we define a new rescaled time for any fixed $\tau > 0$, as follows:

$$T^o_{\eta,\tau} \equiv \left\lceil \frac{\tau \log\left(\frac{|\mu_4 - 3|}{B^8} \cdot \eta^{-1}\right)}{-\log\left(1 - \frac{\eta}{2d}|\mu_4 - 3|\right)} \right\rceil. \quad (3.1)$$

Then under uniform initialization, we have the following convergence result:

Lemma 3 (Convergence Result with Uniform Initialization)**.** *Let the dimension* $d \geq 2$*, let Assumption 1 hold, and let* $\mathbf{u}^{(0)}$ *be uniformly sampled from the unit sphere* $\mathcal{D}_1$*. Then for any fixed positives* $d \geq 2, \tau > 0.5, \eta > 0, \delta \in (0, e^{-1}], \varepsilon \in (0, 1/3]$ *satisfying the scaling condition*

$$\begin{aligned} d &\geq 2\sqrt{2\pi e} \log \varepsilon^{-1} + 1, \quad \eta < \min\left(\frac{1}{|\mu_4 - 3|}, \frac{|\mu_4 - 3|}{B^8} e^{-1}\right), \quad \text{and} \\ C^*_{3,L} \log^8(T^o_{\eta,1} \delta^{-1}) \cdot \frac{B^8}{|\mu_4 - 3|} \cdot d^2 \log^2 d \cdot \eta \log&\left(\frac{|\mu_4 - 3|}{B^8} \eta^{-1}\right) \leq \frac{\varepsilon^2}{(\tau + 1)\log^2 \varepsilon^{-1}}, \end{aligned} \quad (3.2)$$

there exist a uniformly distributed random variable $\mathscr{I} \in \{1, \ldots, d\}$ *and an event* $\mathscr{H}_{3,L}$ *with*

$$\mathbb{P}(\mathscr{H}_{3,L}) \geq 1 - \left(6\tau + 27 + \frac{10368}{\log^5 \delta^{-1}}\right) d\delta - 3\varepsilon,$$

such that on $\mathscr{H}_{3,L}$, iteration $\mathbf{u}^{(t)}$ of Algorithm 1 satisfies for $t \in [T^o_{\eta,0.5}, T^o_{\eta,\tau}]$

$$\left|\tan\angle\left(\mathbf{u}^{(t)}, \mathbf{a}_{\mathscr{I}}\right)\right| \le \sqrt{d}\cdot\left(1-\frac{\eta}{2d}|\mu_4-3|\right)^{t-T^o_{\eta,0.5}} + \sqrt{\tau+1}C_{3,L}\log^{5/2}\delta^{-1}\cdot\frac{B^4}{|\mu_4-3|^{1/2}}\cdot\sqrt{d^3\eta\log\left(\frac{|\mu_4-3|}{B^8}\eta^{-1}\right)}, \tag{3.3}$$

where $C_{3,L}, C^*_{3,L}$ are positive, absolute constants.

Under uniform initialization assumption, Lemma 3 shows that under uniform initialization, as long as the stepsize $\eta = \tilde{O}(d^{-2})$ the term $\left|\tan\angle\left(\mathbf{u}^{(t)}, \mathbf{a}_{\mathscr{I}}\right)\right|$ can be upper-bounded by the sum of two summands, with the first term geometrically decaying from $\sqrt{d}$ at rate $1-\eta|\mu_4-3|/(2d)$, and the second being the noise-induced error $\tilde{O}(\sqrt{d^3\eta})$. As mentioned in §1, the key idea behind is that the scaling condition (3.2) ensures the iteration to sufficiently deviate from a manifold called *frame* where all unstable stationary points lies on (Li et al., 2016), and hence traverse fast to the basin of attraction of the independent component pairs.

Analogous to Theorem 2, when the sample size T satisfies the scaling condition $T = \tilde{\Omega}(d^3)$, one can carefully choose a stepsize $\eta = \tilde{\Theta}(d\log T/T)$ and establish a finite-sample bound in T based on Lemma 3, formulated as our second main theorem:

Theorem 3 (Finite-Sample Error Bound with Uniform Initialization)**.** *Let the dimension $d \ge 2$, let Assumption 1 hold, and let initialization $\mathbf{u}^{(0)}$ be uniformly sampled from the unit sphere $\mathcal{D}_1$. Set for sample size $T \ge 2$ the stepsize η as*

$$\eta(T) = \frac{4d\log\left(\frac{|\mu_4-3|^2}{4B^8 d}T\right)}{|\mu_4-3|T}.$$

Then for any fixed positives $d \ge 2, T \ge 2, \varepsilon \in (0, 1/4]$ satisfying the scaling condition

$$d \ge 2\sqrt{2\pi e}\log\varepsilon^{-1}+1, \quad C^*_{3,T}\log^8(C'_{3,T}\varepsilon^{-1}dT)\cdot\frac{B^8}{|\mu_4-3|^2}\cdot\frac{d^3\log^2 d\log^2 T}{T} \le \frac{\varepsilon^2}{\log^2\varepsilon^{-1}}, \tag{3.4}$$

there exists a uniformly distributed random variable $\mathscr{I} \in \{1,\dots,d\}$ and an event $\mathscr{H}_{3,T}$ with $\mathbb{P}(\mathscr{H}_{3,T}) \ge 1-4\varepsilon$ such that on $\mathscr{H}_{3,T}$, iteration $\mathbf{u}^{(t)}$ of Algorithm 1 satisfies

$$\left|\tan\angle\left(\mathbf{u}^{(T)}, \mathbf{a}_{\mathscr{I}}\right)\right| \le C_{3,T}\log^{5/2}(C'_{3,T}\varepsilon^{-1}d)\cdot\frac{B^4}{|\mu_4-3|}\cdot\sqrt{\frac{d^4\log^2 T}{T}},$$

where $C_{3,T}, C^*_{3,T}, C'_{3,T}$ are positive, absolute constants.

Theorem 3 achieves, under the scaling condition $T = \tilde{\Omega}(d^3)$, an $\tilde{O}(\sqrt{d^4/T})$ finite-sample error bound on $\left|\tan\angle\left(\mathbf{u}^{(T)}, \mathbf{a}_{\mathscr{I}}\right)\right|$ for some $\mathscr{I}$ drawn uniformly at random in $[d]$. With Theorems 2 and 3 for warm and uniform initializations respectively, a specific choice of stepsizes allows us to have the best of the two worlds. Assuming the prior knowledge of sample size T, we initialize $\mathbf{u}^{(0)}$ uniformly at random from the unit sphere $\mathcal{D}_1$ and run Algorithm 1 in two consecutive phases, each using $T/2$ observations:

- In the first phase, we initialize $\mathbf{u}^{(0)}$ uniformly at random on unit sphere $\mathcal{D}_1$, pick a constant stepsize $\eta_1 = \tilde{\Theta}(d\log T/T)$ and update iteration $\mathbf{u}^{(t)}$ via (1.3) for $T/2$ iterates. Theorem 3 guarantees with high probability that $\mathbf{u}^{(T/2)}$ satisfies the warm initialization condition (2.1) under the scaling condition $T = \tilde{\Omega}(d^4)$;

- In the second phase, we warm-initialize the algorithm using the output of the first phase $\mathbf{u}^{(T/2)}$, pick a constant stepsize $\eta_2 = \tilde{\Theta}(\log T/T)$ and update the iteration $\mathbf{u}^{(t)}$ via (1.3) for $T/2$ iterates. The last iterate it outputs achieves an error bound of $\tilde{O}(\sqrt{d/T})$ as indicated by Theorem 2.

The above two-phase procedure indicates an improved finite-sample error bound $\tilde{O}(\sqrt{d/T})$ under the uniform initialization and scaling condition $T = \tilde{\Omega}(d^4)$, formally stated in the following:

Corollary 4 (Improved Finite-Sample Error Bound with Uniform Initialization)**.** *Let the dimension $d \geq 2$, let Assumption 1 hold, and let initialization $\mathbf{u}^{(0)}$ be uniformly sampled from the unit sphere $\mathcal{D}_1$. Set for sample size $T \geq 2$ the stepsizes as*

$$\eta_1(T) = \frac{8d\log\left(\frac{|\mu_4-3|^2}{8B^8d}T\right)}{|\mu_4-3|T}, \qquad \eta_2(T) = \frac{9\log\left(\frac{|\mu_4-3|^2}{9B^8}T\right)}{|\mu_4-3|T}. \tag{3.5}$$

Then for any fixed positive $\varepsilon \in (0, 1/5]$ satisfying the scaling condition

$$d \geq 2\sqrt{2\pi e}\log\varepsilon^{-1}+1, \quad 2\max\{C^*_{2,T}, C^*_{3,T}, C^2_{3,T}\}\log^8(C'_{3,T}\varepsilon^{-1}dT)\cdot\frac{B^8}{|\mu_4-3|^2}\cdot\frac{d^4\log^2 T}{T} \leq \frac{\varepsilon^2}{\log^2\varepsilon^{-1}}, \tag{3.6}$$

there exists a uniformly distributed random variable $\mathscr{I} \in [d]$ and an event $\mathscr{H}_{4,C}$ with $\mathbb{P}(\mathscr{H}_{4,C}) \geq 1-5\varepsilon$ such that on the event $\mathscr{H}_{4,C}$, running Algorithm 1 for $T/2$ iterates with stepsize $\eta_1(T)$ followed by $T/2$ iterates with stepsize $\eta_2(T)$ outputs an $\mathbf{u}^{(T)}$ satisfying

$$\left|\tan\angle\left(\mathbf{u}^{(T)}, \mathbf{a}_i\right)\right| \leq \sqrt{2}C_{2,T}\log^{5/2}(C'_{2,T}\varepsilon^{-1}d)\cdot\frac{B^4}{|\mu_4-3|}\cdot\sqrt{\frac{d\log^2 T}{T}}.$$

where $C_{2,T}, C'_{2,T}, C_{3,T}, C^*_{3,T}, C'_{3,T}$ are positive, absolute constants defined earlier in Theorems 2 and 3.

We study in this section the uniform initialization case and prove Theorem 3. The key idea behind our analysis is that, the uniform initialization is sufficiently deviated from set where all unstable stationary points lie on with high probability, and with delicate concentration analysis the saddle-point avoidance is guaranteed throughout the entire online tensorial ICA algorithm.

Inherited from the warm initialization analysis in §2, we recall the rotated iteration $\mathbf{v}^{(t)} \equiv \mathbf{P}\mathbf{A}^\top\mathbf{u}^{(t)}$ and the rotated observations $\boldsymbol{Y}^{(t)} = \mathbf{P}\mathbf{A}^\top\boldsymbol{X}^{(t)}$, so our online tensorial ICA update rule can still be translated into (2.8). Here the permutation matrix $\mathbf{P}$ has $\mathbf{P}(\mathscr{I};1) = \mathbf{P}(1;\mathscr{I}) = 1 = \mathbf{P}(j;j)$ for $j \in \{1,\mathscr{I}\}^c$ and 0 elsewhere, in which $\mathscr{I} \equiv \operatorname{argmin}_{i\in[d]}\tan^2\angle\left(\mathbf{u}^{(0)}, \mathbf{a}_i\right)$.We let the *coordinate-wise intermediate initialization region* for

each $k \in [2,d]$ be

$$\mathcal{D}_{\text{mid},k} \equiv \left\{ \mathbf{v} \in \mathcal{D}_1 : v_1^2 \geq \max_{2\leq i\leq d} v_i^2 \text{ and } v_1^2 \geq 3v_k^2 \right\}. \tag{3.7}$$

In addition, we let the *cold initialization region* be

$$\mathcal{D}_{\text{cold}} = \left\{ \mathbf{v} \in \mathcal{D}^{d-1} : v_1^2 \geq \max_{2\leq i\leq d} v_i^2 \right\} \tag{3.8}$$

By definition of the rotated iteration $\{\mathbf{v}^{(t)}\}$ and index $\mathscr{I}$, we know that $\mathbf{v}^{(0)} \in \mathcal{D}_{\text{cold}}$ always holds.

For the rest of this section, §3.1 and §3.2 analyze in two steps our algorithm in the uniform initialization case. §A.3 and §A.4 (in Appendix, limited by space) prove the convergence Lemma 3 and its finite-sample error Theorem 3, respectively. Analogous to §2, all secondary lemmas are deferred to §D in Appendix.

3.1 Initialization in the Intermediate Region

Recall the intermediate initialization region $\mathcal{D}_{\text{mid},k}$ was defined in (3.7) for each $k \in [2,d]$. We also let a slightly larger *coordinate-wise intermediate auxiliary region* for each $k \in [2,d]$ be

$$\mathcal{D}_{\text{mid-aux},k} = \left\{ \mathbf{v} \in \mathcal{D}_1 : v_1^2 \geq \max_{i\geq 2} v_i^2 \text{ and } v_1^2 \geq 2v_k^2 \right\}. \tag{3.9}$$

When initialization $\mathbf{v}^{(0)} \in \mathcal{D}_{\text{mid},k}$, we define the first time iterate exits $\mathcal{D}_{\text{mid-aux},k}$ as

$$\mathscr{T}_{w,k} = \inf\left\{ t \geq 1 : \mathbf{v}^{(t)} \in \mathcal{D}_{\text{mid-aux},k}^c \right\}. \tag{3.10}$$

Therefore for each $k \in [2,d]$, $\mathscr{T}_{w,k}$ is a stopping time with respect to filtration $\mathscr{F}_t = \sigma(\mathbf{v}^{(0)}, \boldsymbol{Y}^{(1)}, \ldots, \boldsymbol{Y}^{(t)})$ (we suppose all k that appears later satisfies $k \in [2,d]$ by default).

Our goal is to prove the following high probability trajectory bound for each coordinate k. For intermediate initialization analysis, we consider iteration $U_k^{(t)} = v_k^{(t)}/v_1^{(t)}$ previously defined in (2.11).

Lemma 4. *Let the settings in Lemma 3 hold, and fix the coordinate $k \in [2,d]$ and value $\tau > 0$. Then for any positives η, δ satisfying the scaling condition (3.2) along with the coordinate-wise intermediate initialization condition $\mathbf{v}^{(0)} \in \mathcal{D}_{mid,k}$, there exists an event $\mathscr{H}_{k;4,L}$ satisfying*

$$\mathbb{P}(\mathscr{H}_{k;4,L}) \geq 1 - \left(6\tau + 12 + \frac{5184}{\log^5 \delta^{-1}}\right)\delta,$$

such that on event $\mathscr{H}_{k;4,L}$ the following holds

$$\sup_{t\leq T^o_{\eta,\tau}\wedge\mathscr{T}_{w,k}} \left| U_k^{(t)} - U_k^{(0)} \prod_{s=0}^{t-1} \left[1 - \eta|\mu_4 - 3|\left((v_1^{(s)})^2 - (v_k^{(s)})^2\right)\right] \right| \leq 2C_{4,L}\log^{5/2}\delta^{-1} \cdot B^4 \cdot d^{1/2}\eta(T^o_{\eta,\tau})^{1/2}, \tag{3.11}$$

where $C_{4,L}$ is a positive, absolute constant.

From (3.11) in Lemma 4 we know that, for each coordinate $k \in [2,d]$ and initialized in the intermediate region $\mathcal{D}_{\text{mid},k}$, with high probability the iteration $U_k^{(t)}$ fluctuates around a deterministic curve that decays at least exponentially fast, as long as $(v_1^{(t)})^2 > (v_k^{(t)})^2$ holds.

3.2 Initialization in the Cold Region

Recall the *cold initialization region* $\mathcal{D}_{\text{cold}}$ was earlier defined in (3.8). The iteration we study in cold initialization analysis is

$$W_k^{(t)} = \frac{(v_1^{(t)})^2 - (v_k^{(t)})^2}{(v_k^{(t)})^2}. \tag{3.12}$$

Under assumption in Lemma 3 of uniform initialization on the unit sphere, we have the following lemma. Note that the uniform initialization conditions for $\mathbf{u}^{(0)}$ and $\mathbf{v}^{(0)}$ are equivalent.

Lemma 5. *Let $\mathbf{v}^{(0)}$ be uniformly sampled from the unit sphere $\mathcal{D}_1$ and ε be any fixed positive, with dimension d and ε satisfying*

$$d \geq 2\sqrt{2\pi e}\log\varepsilon^{-1} + 1. \tag{3.13}$$

Then there exists an event $\mathscr{H}_{5,L}$ with $\mathbb{P}(\mathscr{H}_{5,L}) \geq 1 - 3\varepsilon$ such that on event $\mathscr{H}_{5,L}$ the following holds

$$\min_{2\leq k\leq d} W_k^{(0)} \geq \frac{\varepsilon}{8\log\varepsilon^{-1}\log d}. \tag{3.14}$$

Our goal is to estimate the time t when $\mathbf{v}^{(t)}$ enters each coordinate-wise intermediate initialization region starting with the initialization gap given in Lemma 5. For each coordinate $k \in [2,d]$, we define the first time $\mathbf{v}^{(t)}$ enters the coordinate-wise intermediate region $\mathcal{D}_{\text{mid},k}$ as

$$\mathscr{T}_{c,k} = \inf\left\{t \geq 0 : \mathbf{v}^{(t)} \in \mathcal{D}_{\text{mid},k}\right\} \tag{3.15}$$

and the first time iterates exit $\mathcal{D}_{\text{cold}}$ without entering $\mathcal{D}_{\text{mid},k}$, earlier defined in (3.7) and (3.8), as

$$\mathscr{T}_1 = \inf\left\{t \geq 1 : \mathbf{v}^{(t)} \in \mathcal{D}_{\text{cold}}^c\right\} \tag{3.16}$$

In other words, $\mathscr{T}_{c,k}$ is the first time $W_k^{(t)} \geq 2$, and $\mathscr{T}_1$ is the first time $\inf_{2\leq i\leq d} W_i^{(t)} < 0$. By the definition of the rotated iteration $\mathbf{v}^{(t)}$, initialization $\mathbf{v}^{(0)} \in \mathcal{D}_{\text{cold}}$ always holds. For each coordinate $k \in [2,d]$, if $\mathscr{T}_{c,k} = 0$ then $\mathbf{v}^{(0)} \in \mathcal{D}_{\text{mid},k}$ and the previous intermediate initialization analysis directly applies for coordinate k. Otherwise we apply the following lemma, which characterizes the exponential growth of iteration $W_k^{(t)}$ and helps us determine the time of $\mathbf{v}^{(t)}$ entering intermediate region $\mathcal{D}_{\text{mid},k}$.

Lemma 6. *Let the settings in Lemma 3 hold, and fix the coordinate $k \in [2,d]$. Then for any fixed positives η, δ satisfying the scaling condition (3.2) along with the coordinate-wise cold initialization condition $\mathbf{v}^{(0)} \in \mathcal{D}_{cold} \cap \mathcal{D}_{mid,k}^c$, there exists an event $\mathscr{H}_{k;6,L}$ with $\mathbb{P}(\mathscr{H}_{k;6,L}) \geq 1 - \left(15 + \frac{5184}{\log^5\delta^{-1}}\right)\delta$, such that on event $\mathscr{H}_{k;6,L}$*

the following holds

$$\sup_{t\le T^o_{\eta,0.5}\wedge\mathscr{T}_{c,k}\wedge\mathscr{T}_1}\left|W_k^{(t)}\prod_{s=0}^{t-1}\left(1+\eta|\mu_4-3|(v_1^{(s)})^2\right)^{-1}-W_k^{(0)}\right|\le C_{6,L}\log^{5/2}\delta^{-1}\cdot\frac{B^4}{|\mu_4-3|^{1/2}}\cdot d\eta^{1/2}, \quad (3.17)$$

where $C_{6,L}$ is a positive, absolute constant.

From Lemma 6 we know that for each coordinate $k\in[2,d]$, if initialization sets foot outside intermediate region $\mathcal{D}_{\text{mid},k}$, with high probability iteration $W_k^{(t)}$ is tightly controlled within an exponentially growing trajectory for $T^o_{\eta,0.5}$ iterates before it either enters the intermediate region $\mathcal{D}_{\text{mid},k}$ or exits the cold region $\mathcal{D}_{\text{cold}}$. As we will see later in proof of Lemma 3, by putting all coordinates together we can show that $\mathbf{v}^{(t)}$ seldomly leaves the cold region $\mathcal{D}_{\text{cold}}$, which implies that $\mathbf{v}^{(t)}$ will enter the joint intermediate region $\cap_{2\le k\le d}\mathcal{D}_{\text{mid},k}$ within $T^o_{\eta,0.5}$ iterates with high probability.

Limited by space in the main text, we provide the proof of the convergence Lemma 3 in §A.3 and its finite-sample error Theorem 3 in §A.4, respectively.

4 More Related Literatures

The themes of ICA and tensor decomposition have been studied in numerous statistics and signal processing literature (Bach & Jordan, 2002; Chen & Bickel, 2006; Samworth & Yuan, 2012; Bonhomme & Robin, 2009; Eriksson & Koivunen, 2004; Hallin & Mehta, 2015; Hyvärinen et al., 2001; Hyvärinen & Oja, 1997; Hyvarinen, 1999; Hyvärinen & Oja, 2000; Ilmonen & Paindaveine, 2011; Kollo, 2008; Miettinen et al., 2015; Oja et al., 2006; Tichavsky et al., 2006; Wang & Lu, 2017). Recent literatures study the ICA setting in the context of specific parametric families for independent component distributions and obtain parametric (Lee et al., 1999), semi-parametric (Hastie & Tibshirani, 2003; Chen & Bickel, 2006; Ilmonen & Paindaveine, 2011) or non-parametric (Bach & Jordan, 2002; Samarov & Tsybakov, 2004; Samworth & Yuan, 2012) models that can be estimated via maximal likelihood estimation or minimization of mutual information between independent components. We pursue here a different type of contrast function based on tensor decomposition and kurtosis maximization, and hence our methodology is different from these works on ICA.

Recently a line of stochastic approximation and nonconvex optimization literature study the convergence properties of SGD as well as its many variants (Ge et al., 2015; Sun et al., 2015; Jin et al., 2019, 2017; Lei et al., 2017; Allen-Zhu, 2018; Zhang & Sra, 2016; Zhang et al., 2016, 2018; Tripuraneni et al., 2018; Daneshmand et al., 2018). The precursor Ge et al. (2015) study the convergence rate of SGD for minimizing a large class of nonconvex objectives defined on a generic Riemannian manifold. Under the bounded distributional assumption, Ge et al. (2015) prove that SGD equipped with projection as well as a special *noise injection* step can escape from all saddle points and land at an approximate local minimizer in polynomial time of relevant parameters. Convergence rates for generic first-order gradient descent algorithms without adding noise injection are generally unknown (Lee et al., 2017; Pemantle, 1990). Recent

nonconvex optimization literature often design special distributions of noise or sophisticated saddle-point avoidance iteration (Allen-Zhu, 2018; Jin et al., 2019; Sun et al., 2015). We prove that these special saddle-point escaping treatments are in fact *not* necessary under mild scaling conditions, in which case our vanilla online tensorial ICA algorithm guarantees to achieve the $\tilde{O}(\sqrt{d/T})$-optimal rate.

5 Summary

In this work, we study the dynamics, convergence and finite-sample error bound of the online stochastic approximation of (orthogonal) tensorial independent component analysis algorithm, which can be viewed as a stochastic approximation method for optimizing a nonconvex objective of expected kurtosis. We show that with properly chosen stepsizes and under mild scaling conditions our online tensorial ICA algorithm achieves the $\tilde{O}(\sqrt{d/T})$-convergence rate, which is superior than the best existing analysis of such. Our algorithm requires no noise-injection steps or specially-designed loops for saddle points avoidance. Future directions include further improvements of the convergence rate (both upper and lower bounds) and scaling condition, analyzing the mini-batch case as well as the non-identical kurtoses case, and extending our approach to other statistical models that can be casted as nonconvex objective landscapes.

References

Allen-Zhu, Z. (2018). Natasha 2: Faster non-convex optimization than sgd. In *Advances in Neural Information Processing Systems* (pp. 2676–2687).

Arnold, B. C., Balakrishnan, N., & Nagaraja, H. N. (1992). *A First Course in Order Statistics*. Society for Industrial and Applied Mathematics (SIAM).

Bach, F. R. & Jordan, M. I. (2002). Kernel independent component analysis. *Journal of Machine Learning Research*, 3, 1–48.

Bonhomme, S. & Robin, J.-M. (2009). Consistent noisy independent component analysis. *Journal of Econometrics*, 149(1), 12–25.

Chen, A. & Bickel, P. J. (2006). Efficient independent component analysis. *The Annals of Statistics*, 34(6), 2825–2855.

Comon, P. (1994). Independent component analysis, a new concept? *Signal processing*, 36(3), 287–314.

Daneshmand, H., Kohler, J., Lucchi, A., & Hofmann, T. (2018). Escaping saddles with stochastic gradients. In *International Conference on Machine Learning* (pp. 1163–1172).

Durrett, R. (2010). *Probability: Theory and Examples (4th edition)*. Cambridge University Press.

Eriksson, J. & Koivunen, V. (2004). Identifiability, separability, and uniqueness of linear ica models. *IEEE signal processing letters*, 11(7), 601–604.

Fan, X., Grama, I., & Liu, Q. (2012). Large deviation exponential inequalities for supermartingales. *Electronic Communications in Probability*, 17.

Ge, R., Huang, F., Jin, C., & Yuan, Y. (2015). Escaping from saddle points – online stochastic gradient for tensor decomposition. In *Proceedings of The 28th Conference on Learning Theory* (pp. 797–842).

Hallin, M. & Mehta, C. (2015). R-estimation for asymmetric independent component analysis. *Journal of the American Statistical Association*, 110(509), 218–232.

Hastie, T. & Tibshirani, R. (2003). Independent components analysis through product density estimation. In *Advances in neural information processing systems* (pp. 665–672).

Hyvarinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE transactions on Neural Networks*, 10(3), 626–634.

Hyvärinen, A., Karhunen, J., & Oja, E. (2001). *Independent Component Analysis*. John Wiley & Sons.

Hyvärinen, A. & Oja, E. (1997). A fast fixed-point algorithm for independent component analysis. *Neural computation*, 9(7), 1483–1492.

Hyvärinen, A. & Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5), 411–430.

Ilmonen, P. & Paindaveine, D. (2011). Semiparametrically efficient inference based on signed ranks in symmetric independent component models. *the Annals of Statistics*, 39(5), 2448–2476.

Jin, C., Netrapalli, P., Ge, R., Kakade, S. M., & Jordan, M. I. (2019). Stochastic gradient descent escapes saddle points efficiently. *arXiv preprint arXiv:1902.04811*.

Jin, C., Netrapalli, P., & Jordan, M. I. (2017). Accelerated gradient descent escapes saddle points faster than gradient descent. *arXiv preprint arXiv:1711.10456*.

Kollo, T. (2008). Multivariate skewness and kurtosis measures with an application in ica. *Journal of Multivariate Analysis*, 99(10), 2328–2338.

Laib, N. (1999). Exponential-type inequalities for martingale difference sequences. application to nonparametric regression estimation. *Communications in Statistics-Theory and Methods*, 28(7), 1565–1576.

Lee, J. D., Panageas, I., Piliouras, G., Simchowitz, M., Jordan, M. I., & Recht, B. (2017). First-order methods almost always avoid saddle points. *arXiv preprint arXiv:1710.07406*.

Lee, T.-W., Girolami, M., & Sejnowski, T. J. (1999). Independent component analysis using an extended infomax algorithm for mixed subgaussian and supergaussian sources. *Neural computation*, 11(2), 417–441.

Lei, L., Ju, C., Chen, J., & Jordan, M. I. (2017). Non-convex finite-sum optimization via scsg methods. In *Advances in Neural Information Processing Systems* (pp. 2348–2358).

Lesigne, E. & Volnỳ, D. (2001). Large deviations for martingales. *Stochastic processes and their applications*, 96(1), 143–159.

Li, C. J., Wang, M., Liu, H., & Zhang, T. (2018). Near-optimal stochastic approximation for online principal component estimation. *Mathematical Programming*, 167(1), 75–97.

Li, C. J., Wang, Z., & Liu, H. (2016). Online ICA: Understanding global dynamics of nonconvex optimization via diffusion processes. In *Advances in Neural Information Processing System* (pp. 4967–4975).

Miettinen, J., Taskinen, S., Nordhausen, K., & Oja, H. (2015). Fourth moments and independent component analysis. *Statistical science*, 30(3), 372–390.

Oja, H., Sirkiä, S., & Eriksson, J. (2006). Scatter matrices and independent component analysis. *Austrian Journal of Statistics*, 35(2&3), 175–189.

Pemantle, R. (1990). Nonconvergence to unstable points in urn models and stochastic approximations. *The Annals of Probability*, (pp. 698–712).

Samarov, A. & Tsybakov, A. (2004). Nonparametric independent component analysis. *Bernoulli*, 10(4), 565–582.

Samworth, R. J. & Yuan, M. (2012). Independent component analysis via nonparametric maximum likelihood estimation. *The Annals of Statistics*, 40(6), 2973–3002.

Stone, J. V. (2004). *Independent component analysis: a tutorial introduction*. MIT press.

Sun, J., Qu, Q., & Wright, J. (2015). When are nonconvex problems not scary? *arXiv preprint arXiv:1510.06096*.

Tichavsky, P., Koldovsky, Z., & Oja, E. (2006). Performance analysis of the fastica algorithm and crame/spl acute/r-rao bounds for linear independent component analysis. *IEEE transactions on Signal Processing*, 54(4), 1189–1203.

Tripuraneni, N., Stern, M., Jin, C., Regier, J., & Jordan, M. I. (2018). Stochastic cubic regularization for fast nonconvex optimization. In *Advances in neural information processing systems* (pp. 2899–2908).

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

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press.

Wang, C. & Lu, Y. (2017). The scaling limit of high-dimensional online independent component analysis. In *Advances in Neural Information Processing Systems* (pp. 6638–6647).

Wang, Y. (2017). Asymptotic analysis via stochastic differential equations of gradient descent algorithms in statistical and computational paradigms. *arXiv preprint arXiv:1711.09514*.

Zhang, H., Reddi, S. J., & Sra, S. (2016). Riemannian svrg: Fast stochastic optimization on riemannian manifolds. In *Advances in Neural Information Processing Systems* (pp. 4592–4600).

Zhang, H. & Sra, S. (2016). First-order methods for geodesically convex optimization. In *Conference on Learning Theory* (pp. 1617–1638).

Zhang, J., Zhang, H., & Sra, S. (2018). R-spider: A fast riemannian stochastic optimization algorithm with curvature independent rate. *arXiv preprint arXiv:1811.04194*.

Appendix

In Appendix, §4 discusses (a non-exhaustive list of) related works. §A proves the main results in the paper. §B studies the case of estimating multiple components of online tensorial ICA algorithms and provide a short proof. §C and §D provide all secondary lemmas and their proofs for warm and uniform initialization analysis, separately. §E, §F and §G provide necessary tools including a reversed Gronwall's inequality, preliminaries and properties of Orlicz ψ_α-norm and a concentration inequality, all of which are theoretical building blocks of this paper.

A Deferred Proofs of Main Results

A.1 Proof of Lemma 1

Now we use the key Lemma 2 to tightly estimate the dynamics in all coordinates and conclude Lemma 1.

Proof of Lemma 1. We denote $\mathscr{H}_{1,L} \equiv \bigcap_{k\in[2,d]} \mathscr{H}_{k;2,L}$ as the intersection of events $\mathscr{H}_{k;2,L}$. Consider now the following $(d-1)$-dimensional vector

$$\left(U_k^{(t)} - U_k^{(0)} \prod_{s=0}^{t-1} \left[1 - \eta|\mu_4 - 3| \left((v_1^{(s)})^2 - (v_k^{(s)})^2 \right) \right] : k \in [2,d] \right). \tag{A.1}$$

Using Lemma 2, on the event $\mathscr{H}_{1,L} \cap (t \le T^*_{\eta,\tau} \wedge \mathscr{T}_x)$ we bound the Euclidean norm of (A.1) by

$$\sqrt{\sum_{k=2}^{d} \left(U_k^{(t)} - U_k^{(0)} \prod_{s=0}^{t-1} \left[1 - \eta|\mu_4 - 3| \left((v_1^{(s)})^2 - (v_k^{(s)})^2 \right) \right] \right)^2} \le \sqrt{d} \cdot 2C_{2,L} \log^{5/2} \delta^{-1} \cdot B^4 \cdot \eta (T^*_{\eta,\tau})^{1/2}. \tag{A.2}$$

Additionally, the left hand of (A.2) is the norm of subtraction of two vectors and hence lower bounded by

$$\begin{aligned} &\sqrt{\sum_{k=2}^{d} \left(U_k^{(t)} - U_k^{(0)} \prod_{s=0}^{t-1} \left[1 - \eta|\mu_4 - 3| \left((v_1^{(s)})^2 - (v_k^{(s)})^2 \right) \right] \right)^2} \\ \ge &\sqrt{\sum_{k=2}^{d} \left(U_k^{(t)} \right)^2} - \sqrt{\sum_{k=2}^{d} \left(U_k^{(0)} \prod_{s=0}^{t-1} \left[1 - \eta|\mu_4 - 3| \left((v_1^{(s)})^2 - (v_k^{(s)})^2 \right) \right] \right)^2}, \end{aligned} \tag{A.3}$$

due to triangle inequality of Euclidean norms. The definition of iteration $\{U_k^{(t)}\}$ in (2.11) implies that $\left|\tan\angle\left(\mathbf{v}^{(t)}, \mathbf{e}_1\right)\right| = \sqrt{\sum_{k=2}^d (U_k^{(t)})^2}$ and the definition of stopping time $\mathscr{T}_x$ in (2.13) implies that $(v_1^{(s)})^2 - (v_k^{(s)})^2 \ge 1/3$ holds for all $k \in [2,d]$ and $0 \le s < t$ on the event $(t \le \mathscr{T}_x)$. Combining this with (A.2) and

(A.3), we obtain on the event $\mathscr{H}_{1,L} \cap (t \le T^*_{\eta,\tau} \wedge \mathscr{T}_x)$

$$\begin{aligned}
\left|\tan\angle\left(\mathbf{v}^{(t)}, \mathbf{e}_1\right)\right| &= \sqrt{\sum_{k=2}^{d}\left(U_k^{(t)}\right)^2} \\
&\le \sqrt{\sum_{k=2}^{d}\left(U_k^{(0)}\right)^2}\left(1-\frac{\eta}{3}|\mu_4-3|\right)^t + \sqrt{d}\cdot 2C_{2,L}\log^{5/2}\delta^{-1}\cdot B^4\cdot\eta(T^*_{\eta,\tau})^{1/2} \\
&= \left|\tan\angle\left(\mathbf{v}^{(0)}, \mathbf{e}_1\right)\right|\left(1-\frac{\eta}{3}|\mu_4-3|\right)^t + 2C_{2,L}\log^{5/2}\delta^{-1}\cdot B^4\cdot(d\eta^2 T^*_{\eta,\tau})^{1/2}.
\end{aligned} \tag{A.4}$$

The definition of $T^*_{\eta,\tau}$ in (2.2) along with $-\log\left(1-\frac{\eta}{3}|\mu_4-3|\right) \ge \frac{\eta}{3}|\mu_4-3|$ gives

$$T^*_{\eta,\tau} \le 1 + \frac{3\tau}{|\mu_4-3|}\cdot\eta^{-1}\log\left(\frac{|\mu_4-3|}{B^8}\eta^{-1}\right) \le \frac{3(\tau+1)}{|\mu_4-3|}\cdot\eta^{-1}\log\left(\frac{|\mu_4-3|}{B^8}\eta^{-1}\right), \tag{A.5}$$

where we use $\frac{|\mu_4-3|}{B^8}\eta^{-1} \ge e$ and $\frac{1}{|\mu_4-3|}\eta^{-1} \ge 1$ implied by scaling condition (2.3). Using relation (A.5), we find that scaling condition (2.3) with constant $C^*_{1,L} \equiv 713C^2_{2,L}$ indicates

$$\begin{aligned}
2C_{2,L}\log^{5/2}\delta^{-1}\cdot B^4\cdot(d\eta^2 T^*_{\eta,\tau})^{1/2} &\le \sqrt{4C^2_{2,L}B^8\log^5\delta^{-1}\cdot\frac{3(\tau+1)}{|\mu_4-3|}\cdot d\eta\log\left(\frac{|\mu_4-3|}{B^8}\eta^{-1}\right)} \\
\le \sqrt{\frac{12}{713}(\tau+1)\cdot C^*_{1,L}\log^8(T^*_{\eta,1}\delta^{-1})\cdot\frac{B^8}{|\mu_4-3|}\cdot d\eta\log\left(\frac{|\mu_4-3|}{B^8}\eta^{-1}\right)} &\le \sqrt{\frac{12}{713}},
\end{aligned} \tag{A.6}$$

where the $\log^5\delta^{-1} \le \log^8(T^*_{\eta,1}\delta^{-1})$ was applied due to $T^*_{\eta,1} \ge 1$ and $\delta \le e^{-1}$. Viewing (2.1) (equivalent to $\mathbf{v}^{(0)} \in \mathcal{D}_{\text{warm}}$ in (2.10)) and (A.4), we have for each t on the event $\mathscr{H}_{1,L} \cap (\mathscr{T}_x \le T^*_{\eta,\tau}) \cap (t \le \mathscr{T}_x)$ that

$$\left|\tan\angle\left(\mathbf{v}^{(t)}, \mathbf{e}_1\right)\right| \le \left|\tan\angle\left(\mathbf{v}^{(0)}, \mathbf{e}_1\right)\right| + \sqrt{\frac{12}{713}} < \frac{1}{\sqrt{3}} + \left(\frac{1}{\sqrt{2}} - \frac{1}{\sqrt{3}}\right) = \frac{1}{\sqrt{2}}, \tag{A.7}$$

where we again applied $(\eta/3)|\mu_4-3| \le 1$ from (2.3). This further indicates that on $\mathscr{H}_{1,L} \cap (\mathscr{T}_x \le T^*_{\eta,\tau})$ $\left|\tan\angle\left(\mathbf{v}^{(\mathscr{T}_x)}, \mathbf{e}_1\right)\right| < 1/\sqrt{2}$ holds, contradicting the fact that $\mathbf{v}^{(\mathscr{T}_x)} \in \mathcal{D}^c_{\text{warm-aux}}$ on the same event. Therefore, we have $\mathscr{H}_{1,L} \cap (\mathscr{T}_x \le T^*_{\eta,\tau}) = \varnothing$, and equivalently

$$\mathscr{H}_{1,L} = \mathscr{H}_{1,L} \cap (\mathscr{T}_x > T^*_{\eta,\tau}). \tag{A.8}$$

This implies that for all $t \in [0, T^*_{\eta,\tau}]$, (A.4) holds on the event $\mathscr{H}_{1,L}$. Plugging in the inequality involving $T^*_{\eta,\tau}$

in (A.5), we have on the event $\mathscr{H}_{1,L}$ that for all $t \in [0, T^*_{\eta,\tau}]$

$$\begin{aligned}\left|\tan\angle\left(\mathbf{v}^{(t)}, \mathbf{e}_1\right)\right| &\le \left|\tan\angle\left(\mathbf{v}^{(0)}, \mathbf{e}_1\right)\right| \left(1 - \frac{\eta}{3}|\mu_4 - 3|\right)^t + 2C_{2,L}B^4 \log^{5/2}\delta^{-1} \cdot (d\eta^2 T^*_{\eta,\tau})^{1/2} \\ &\le \left|\tan\angle\left(\mathbf{v}^{(0)}, \mathbf{e}_1\right)\right| \left(1 - \frac{\eta}{3}|\mu_4 - 3|\right)^t \\ &\quad + \sqrt{\tau+1}\left(2\sqrt{3}C_{2,L}\right)\log^{5/2}\delta^{-1} \cdot \frac{B^4}{|\mu_4-3|^{1/2}} \cdot \sqrt{d\eta \log\left(\frac{|\mu_4-3|}{B^8}\eta^{-1}\right)}.\end{aligned} \tag{A.9}$$

Letting the constant $C_{1,L} \equiv 2\sqrt{3}C_{2,L}$, the scaling relation (2.9) and the above derivation (A.9) prove that (2.4) holds for all $t \in [0, T^*_{\eta,\tau}]$ on the event $\mathscr{H}_{1,L}$.

The only left is to estimate the probability of $\mathscr{H}_{1,L}$. Lemma 2 gives for each $k \in [2,d]$ the probability of event $\mathbb{P}(\mathscr{H}_{k;2,L}) \ge 1 - \left(6\tau + 12 + \frac{5184}{\log^5 \delta^{-1}}\right)\delta$, and hence elementary union bound calculation gives

$$\mathbb{P}(\mathscr{H}_{1,L}) = \mathbb{P}\left(\bigcap_{k\in[2,d]} \mathscr{H}_{k;2,L}\right) \ge 1 - \left(6\tau + 12 + \frac{5184}{\log^5 \delta^{-1}}\right) d\delta, \tag{A.10}$$

completing the whole proof of Lemma 1. □

A.2 Proof of Theorem 2

Now we turn to the proof of the finite-sample error Theorem 2. The idea is to apply Lemma 1 with appropriate stepsize η (as in (2.5)) as well as an appropriate τ to obtain the finite-sample error bound.

Proof of Theorem 2. (i) We first provide an upper bound on $T^*_{\eta(T),\tau}$. Under scaling condition (2.6) with constants $C^*_{2,T} \equiv 90C^*_{1,L} > 10$ and $C'_{2,T} > 1$ to be determined later, we have $\frac{2|\mu_4-3|^2}{9B^8}T \ge e$. Plugging in $\eta = \eta(T)$ from (2.5) to relation (A.5), we have

$$\begin{aligned}T^*_{\eta(T),\tau} &\le \frac{3(\tau+1)}{|\mu_4-3|} \cdot \eta(T)^{-1} \log\left(\frac{|\mu_4-3|}{B^8}\eta(T)^{-1}\right) \\ &= \frac{3(\tau+1)}{|\mu_4-3|} \cdot \frac{2|\mu_4-3|T}{9\log\left(\frac{2|\mu_4-3|^2}{9B^8}T\right)} \log\left(\frac{|\mu_4-3|}{B^8} \cdot \frac{2|\mu_4-3|T}{9\log\left(\frac{2|\mu_4-3|^2}{9B^8}T\right)}\right) \le \frac{2(\tau+1)}{3}T.\end{aligned} \tag{A.11}$$

(ii) Next we provide a lower bound on $T^*_{\eta(T),\tau}$. By Taylor expansion, for all $x \in (0, 1/3]$ we know that

$$|\log(1-x) + x| = \left|\sum_{n=2}^{\infty} \frac{x^n}{n}\right| \le \frac{x^2}{2}\sum_{n=0}^{\infty} x^n \le \frac{x^2}{2}\frac{1}{1-1/3} = \frac{3x^2}{4},$$

and hence

$$\frac{1}{-\log(1-x)} \ge \frac{1}{x + 3x^2/4} \ge \frac{1}{x + x/4} \ge \frac{4}{5x}. \tag{A.12}$$

From the definition of $T^*_{\eta,\tau}$ in (2.2), for η satisfying $\eta|\mu_4-3|/3 \le 1/3$ we have

$$T^*_{\eta,\tau} \ge \frac{\tau \log\left(\frac{|\mu_4-3|}{B^8}\cdot\eta^{-1}\right)}{-\log\left(1-\frac{\eta}{3}|\mu_4-3|\right)} \ge \frac{12\tau}{5|\mu_4-3|}\cdot\eta^{-1}\log\left(\frac{|\mu_4-3|}{B^8}\cdot\eta^{-1}\right). \tag{A.13}$$

Under scaling condition (2.6), along with relation $\frac{B^4}{|\mu_4-3|} \ge \frac{1}{8}$ given by Lemma 7 in Appendix C and $T \ge 2$, we have

$$\frac{\eta(T)|\mu_4-3|}{3} = \frac{3\log\left(\frac{2|\mu_4-3|^2}{9B^8}T\right)}{2T} \le \frac{3\left(\log T+\log(128/9)\right)}{2T} < \frac{3\log T}{T} \le \frac{1}{3}, \quad \frac{2|\mu_4-3|^2}{9B^8}T \ge e, \tag{A.14}$$

where we apply the elementary inequality $\frac{9\log x}{x} < 1$ for all $x \ge 9$, since $T \ge 9$ is an intrinsic result of (2.6) and (C.1). Plugging in $\eta = \eta(T)$ (as in (2.5)) to (A.13) and we obtain

$$\begin{aligned} T^*_{\eta(T),\tau} &\ge \frac{12\tau}{5|\mu_4-3|}\cdot\eta(T)^{-1}\log\left(\frac{|\mu_4-3|}{B^8}\cdot\eta(T)^{-1}\right) \\ &= \frac{12\tau}{5}\cdot\frac{2T}{9\log\left(\frac{2|\mu_4-3|^2}{9B^8}T\right)}\log\left(\frac{2|\mu_4-3|^2T}{9B^8\log\left(\frac{2|\mu_4-3|^2}{9B^8}T\right)}\right) \\ &\ge \frac{8\tau}{15}\cdot\frac{T}{\log\left(\frac{2|\mu_4-3|^2}{9B^8}T\right)}\cdot\frac{1}{2}\log\left(\frac{2|\mu_4-3|^2}{9B^8}T\right) \ge \frac{4\tau}{15}T, \end{aligned} \tag{A.15}$$

where we use the elementary inequality $\log\left(\frac{x}{\log x}\right) \ge \frac{1}{2}\log x$ for all $x > 1$.

(iii) From (A.11) and (A.15) we know that $T \in [T^*_{\eta(T),0.5}, T^*_{\eta(T),4}]$. Here we will verify scaling condition (2.3) required in Lemma 1 under our setting. By choosing

$$\varepsilon \equiv \left(36+\frac{5184}{\log^5\delta^{-1}}\right)d\delta, \tag{A.16}$$

we have

$$T^*_{\eta(T),1}\delta^{-1} \le C'_{2,T}\varepsilon^{-1}dT, \tag{A.17}$$

where constant $C'_{2,T} \equiv (4/3)\cdot(36+5184) = 6960$, since (A.11) gives $T^*_{\eta(T),1} \le 4T/3$ and $\delta \le e^{-1}$ obviously holds as long as $\varepsilon \le 1$.

Therefore for the first scaling condition in (2.3), our pick of $\tau = 4$ requires

$$22.5C^*_{1,L}\log^8(T^*_{\eta(T),1}\delta^{-1})\cdot\frac{d\log\left(\frac{2|\mu_4-3|^2}{9B^8}T\right)}{T}\log\left(\frac{2|\mu_4-3|^2T}{9B^8\log\left(\frac{2|\mu_4-3|^2}{9B^8}T\right)}\right) \le \frac{|\mu_4-3|^2}{B^8},$$

while a sufficient condition for the above to hold is, due to (A.17),

$$C^*_{2,T}\log^8(C'_{2,T}\varepsilon^{-1}dT)\cdot\frac{d\log^2 T}{T}\le\frac{|\mu_4-3|^2}{B^8}, \tag{A.18}$$

which comes from (2.6) and constant $C^*_{2,T}\equiv 22.5C^*_{1,L}\cdot 2^2 = 90C^*_{1,L}$, because (2.6) and (C.1) imply $T\ge 128/9\ge\frac{2|\mu_4-3|^2}{9B^8}$, and

$$1\le\log\left(\frac{2|\mu_4-3|^2}{9B^8}T\right)\le 2\log T. \tag{A.19}$$

To verify the second condition in (2.3), using (2.6), (A.19) and (C.1), we have

$$\frac{B^8}{|\mu_4-3|}\eta(T)=\frac{9B^8}{2|\mu_4-3|^2}\cdot\frac{\log\left(\frac{2|\mu_4-3|^2}{9B^8}T\right)}{T}\le\frac{9B^8}{|\mu_4-3|^2}\cdot\frac{\log T}{T}<e^{-1},$$

Note that we have already verified $|\mu_4-3|\eta(T)<1$ in (A.14).

(iv) Using warm initialization condition (2.1) in Lemma 1 and the definition of $T^*_{\eta,\tau}$ given in (2.2), for all $t\in[T^*_{\eta,0.5},T^*_{\eta,4}]$ we have

$$\begin{aligned}&\left|\tan\angle\left(\mathbf{u}^{(0)},\mathbf{a}_i\right)\right|\left(1-\frac{\eta}{3}|\mu_4-3|\right)^t\le\frac{1}{\sqrt{3}}\cdot\frac{B^4}{|\mu_4-3|^{1/2}}\eta^{1/2}\\&\le(3-\sqrt{5})C_{1,L}\log^{5/2}\delta^{-1}\cdot\frac{B^4}{|\mu_4-3|^{1/2}}\cdot\sqrt{d\eta\log\left(\frac{|\mu_4-3|}{B^8}\eta^{-1}\right)},\end{aligned}$$

where the first inequality comes from $t\ge T^*_{\eta,0.5}$ and definition of $T^*_{\eta,\tau}$ in (2.2), and the second inequality is due to $\log\delta^{-1}\ge 1$ and $\frac{|\mu_4-3|}{B^8}\eta^{-1}\ge e$ given by scaling condition (2.3). Therefore, on the event $\mathscr{H}_{1,L}$ we have for all $t\in[T^*_{\eta,0.5},T^*_{\eta,4}]$

$$\left|\tan\angle\left(\mathbf{u}^{(t)},\mathbf{a}_i\right)\right|\le 3C_{1,L}\log^{5/2}\delta^{-1}\cdot\frac{B^4}{|\mu_4-3|^{1/2}}\cdot\sqrt{d\eta\log\left(\frac{|\mu_4-3|}{B^8}\eta^{-1}\right)}. \tag{A.20}$$

To finalize our proof, we plug in $\eta=\eta(T)$ to (A.20) and conclude from $T\in[T^*_{\eta(T),0.5},T^*_{\eta(T),4}]$ that there exists an event $\mathscr{H}_{2,T}\equiv\mathscr{H}_{1,L}$ with, due to (A.16), $\mathbb{P}(\mathscr{H}_{2,T})\ge 1-\varepsilon$, such that on $\mathscr{H}_{2,T}$ the follow-

ing holds

$$\begin{aligned}
&\left|\tan\angle\left(\mathbf{u}^{(T)},\mathbf{a}_i\right)\right| \\
&\le 3C_{1,L}\log^{5/2}\delta^{-1}\cdot\frac{B^4}{|\mu_4-3|^{1/2}}\cdot\sqrt{d\eta(T)\log\left(\frac{|\mu_4-3|}{B^8}\eta(T)^{-1}\right)} \\
&= \frac{9\sqrt{2}}{2}C_{1,L}\log^{5/2}\delta^{-1}\cdot\frac{B^4}{|\mu_4-3|}\cdot\sqrt{\frac{d\log\left(\frac{2|\mu_4-3|^2}{9B^8}T\right)}{T}\log\left(\frac{2|\mu_4-3|^2T}{9B^8\log\left(\frac{2|\mu_4-3|^2}{9B^8}T\right)}\right)} \\
&\le C_{2,T}\log^{5/2}(C'_{2,T}d\varepsilon^{-1})\cdot\frac{B^4}{|\mu_4-3|}\cdot\sqrt{\frac{d\log^2 T}{T}},
\end{aligned} \tag{A.21}$$

where in the last step we apply (A.19), $\log^{5/2}\delta^{-1}\le\log^{5/2}(C'_{2,T}d\varepsilon^{-1})$ from (A.16) and $C'_{2,T}=6960$, with constant $C_{2,T}\equiv 9\sqrt{2}C_{1,L}$. This completes the whole proof of the theorem.

□

A.3 Proof of Lemma 3

In uniform initialization analysis, intuitively $\mathbf{v}^{(t)}$ needs to enter intermediate region $\mathcal{D}_{\text{mid},k}$ first before we worry about its exit of the intermediate-auxilliary region $\mathcal{D}_{\text{mid-aux},k}$, or in other words we need $\mathscr{T}_{w,k}>\mathscr{T}_{c,k}$. Hence we slightly modify the definition of $\mathscr{T}_{w,k}$ as

$$\mathscr{T}_{w,k}=\inf\left\{t>\mathscr{T}_{c,k}:\mathbf{v}^{(t)}\in\mathcal{D}^c_{\text{mid-aux},k}\right\}. \tag{A.22}$$

We consider $\mathscr{T}_{c,k}$ as the starting time when applying Lemma 4 for coordinate k. Due to strong Markov property of process $\{\mathbf{v}^{(t)}\}$ (Durrett, 2010), such modification does not affect the application of Lemma 4.

Proof of Lemma 3. We let constant $C^*_{3,L}\equiv\max\{256C^2_{6,L},476C^2_{4,L}\}$ in scaling condition (3.2) in Lemma 3.

(i) We start by making coordinate-wise analysis for each $k\in[2,d]$. On the event

$$\mathscr{H}_{5,L}\cap\mathscr{H}_{k;6,L}\cap(\mathscr{T}_1>T^o_{\eta,\tau})\cap(\mathscr{T}_{c,k}>T^o_{\eta,0.5}),$$

since $(v_1^{(t)})^2\ge 1/d$ for all $t<T^o_{\eta,0.5}$, by applying Lemmas 5 and 6 we have

$$\begin{aligned}
W_k^{(T^o_{\eta,0.5})} &\ge \left(W_k^{(0)}-C_{6,L}\log^{5/2}\delta^{-1}\cdot\frac{B^4}{|\mu_4-3|^{1/2}}\cdot d\eta^{1/2}\right)\left(1+\frac{\eta}{d}|\mu_4-3|\right)^{T^o_{\eta,0.5}} \\
&\ge \left(\frac{\varepsilon}{8\log\varepsilon^{-1}\log d}-C_{6,L}\log^{5/2}\delta^{-1}\cdot\frac{B^4}{|\mu_4-3|^{1/2}}\cdot d\eta^{1/2}\right)\cdot\left(1+\frac{\eta}{d}|\mu_4-3|\right)^{T^o_{\eta,0.5}}.
\end{aligned} \tag{A.23}$$

Using the elementary inequality $-\log(1-x) \le \log(1+2x)$ for all $0 \le x \le 1/2$, we have

$$\left(1+\frac{\eta}{d}|\mu_4-3|\right)^{T^o_{\eta,0.5}} \ge \exp\left(\frac{1}{2}\log\left(\frac{|\mu_4-3|}{B^8}\eta^{-1}\right)\cdot\frac{\log\left(1+\frac{|\mu_4-3|}{d}\eta\right)}{-\log\left(1-\frac{|\mu_4-3|}{2d}\eta\right)}\right) \ge \frac{|\mu_4-3|^{1/2}}{B^4}\eta^{-1/2}, \tag{A.24}$$

since $\frac{|\mu_4-3|}{2d}\eta \le 1/2$ under scaling condition (3.2). Because the following inequality hold,

$$\frac{\varepsilon}{8\log\varepsilon^{-1}\log d} \ge 2C_{6,L}\log^{5/2}\delta^{-1}\cdot\frac{B^4}{|\mu_4-3|^{1/2}}\cdot d\eta^{1/2}, \tag{A.25}$$

along with (A.23) we have

$$W_k^{(T^o_{\eta,0.5})} \ge C_{6,L}\log^{5/2}\delta^{-1}\cdot\frac{B^4}{|\mu_4-3|^{1/2}}\cdot d\eta^{1/2}\cdot\frac{|\mu_4-3|^{1/2}}{B^4}\eta^{-1/2} = C_{6,L}\log^{5/2}\delta^{-1}\cdot d \ge 2, \tag{A.26}$$

due to $\delta \in (0, e^{-1}]$, which indicates $\mathbf{v}^{(T^o_{\eta,0.5})} \in \mathcal{D}_{\text{mid-aux},k}$, i.e. event $\mathscr{H}_{5,L}\cap\mathscr{H}_{k;6,L}\cap(\mathscr{T}_1 > T^o_{\eta,\tau})\cap(\mathscr{T}_{c,k} > T^o_{\eta,0.5}) \subseteq (\mathscr{T}_{c,k} \le T^o_{\eta,0.5})$, and hence

$$\mathscr{H}_{5,L}\cap\mathscr{H}_{k;6,L}\cap(\mathscr{T}_1 > T^o_{\eta,\tau}) \subseteq (\mathscr{T}_{c,k} \le T^o_{\eta,0.5}). \tag{A.27}$$

(ii) On the event

$$\mathscr{H}_{k;4,L}\cap\mathscr{H}_{5,L}\cap\mathscr{H}_{k;6,L}\cap(\mathscr{T}_1 > T^o_{\eta,\tau})\cap(\mathscr{T}_{w,k} \le T^o_{\eta,\tau}),$$

since $(v_1^{(t)})^2 - (v_k^{(t)})^2 \ge 1/(2d)$ for all $\mathscr{T}_{c,k} \le t < \mathscr{T}_{w,k}$ and the following is guaranteed by scaling condition (3.2)

$$\frac{|\mu_4-3|}{2d}\eta \le 1, \quad 2C_{4,L}B^4\log^{5/2}\delta^{-1}\cdot d^{1/2}\eta(T^o_{\eta,\tau})^{1/2} \le \sqrt{\frac{8}{476}}, \tag{A.28}$$

using Lemma 4 we have

$$\begin{aligned}\left|U_k^{(\mathscr{T}_{w,k})}\right| &\le \left|U_k^{(\mathscr{T}_{c,k})}\right|\left(1-\frac{|\mu_4-3|\eta}{2d}\right)^{(\mathscr{T}_{w,k}-\mathscr{T}_{c,k})} + 2C_{4,L}B^4\log^{5/2}\delta^{-1}\cdot d^{1/2}\eta(T^o_{\eta,\tau})^{1/2} \\ &< \frac{1}{\sqrt{3}}\cdot 1 + \sqrt{\frac{8}{476}} \le \frac{1}{\sqrt{2}},\end{aligned} \tag{A.29}$$

which implies $\mathscr{H}_{k;4,L}\cap\mathscr{H}_{5,L}\cap\mathscr{H}_{k;6,L}\cap(\mathscr{T}_1 > T^o_{\eta,\tau})\cap(\mathscr{T}_{w,k} \le T^o_{\eta,\tau}) \subseteq (\mathscr{T}_{w,k} > T^o_{\eta,\tau})$, and hence

$$\mathscr{H}_{k;4,L}\cap\mathscr{H}_{5,L}\cap\mathscr{H}_{k;6,L}\cap(\mathscr{T}_1 > T^o_{\eta,\tau}) \subseteq (\mathscr{T}_{w,k} > T^o_{\eta,\tau}). \tag{A.30}$$

(iii) With (A.27) and (A.30) proven, we put all coordinates $k \in [2,d]$ together and define event

$$\mathscr{H}_{3,L} \equiv (\cap_{2\le k\le d}\mathscr{H}_{k;4,L}) \cap \mathscr{H}_{5,L} \cap (\cap_{2\le k\le d}\mathscr{H}_{k;6,L}).$$

On the event $\mathscr{H}_{3,L}\cap(\mathscr{T}_1 \le T^o_{\eta,\tau})\cap(\mathscr{T}_1 \le \mathscr{T}_{c,k})$, by applying Lemmas 5 and 6 with (A.25) we have

$$W_k^{(\mathscr{T}_1)} \ge \left(W_k^{(0)} - C_{6,L}\log^{5/2}\delta^{-1}\cdot\frac{B^4}{|\mu_4-3|^{1/2}}\cdot d\eta^{1/2}\right)\left(1+\frac{\eta}{d}|\mu_4-3|\right)^{\mathscr{T}_1} \ge 0. \tag{A.31}$$

On the event $\mathscr{H}_{3,L}\cap(\mathscr{T}_1 \le T^o_{\eta,\tau})\cap(\mathscr{T}_1 > \mathscr{T}_{c,k})$, by applying Lemma 4 with (A.28) we have

$$\begin{aligned}\left|U_k^{(\mathscr{T}_1)}\right| &\le \left|U_k^{(\mathscr{T}_{c,k})}\right|\left(1-\frac{|\mu_4-3|\eta}{2d}\right)^{(\mathscr{T}_1-\mathscr{T}_{c,k})} + 2C_{4,L}B^4\log^{5/2}\delta^{-1}\cdot d^{1/2}\eta(T^o_{\eta,\tau})^{1/2}\\ &\le \frac{1}{\sqrt{3}}\cdot 1+\sqrt{\frac{8}{476}} \le 1.\end{aligned} \tag{A.32}$$

Recall that $W_k^{(t)} \ge 0$, $|U_k^{(t)}| \le 1$ and $|v_1^{(t)}| \ge |v_k^{(t)}|$ are equivalent. By combining (A.31) and (A.32) for all $k \in [2,d]$, we obtain

$$\begin{aligned}\mathscr{H}_{3,L}\cap(\mathscr{T}_1 \le T^o_{\eta,\tau}) &= \cap_{2\le k\le d}\left[\left(\mathscr{H}_{3,L}\cap(\mathscr{T}_1 \le T^o_{\eta,\tau})\cap(\mathscr{T}_1 \le \mathscr{T}_{c,k})\right)\cup\left(\mathscr{H}_{3,L}\cap(\mathscr{T}_1 \le T^o_{\eta,\tau})\cap(\mathscr{T}_1 > \mathscr{T}_{c,k})\right)\right]\\ &\subseteq \cap_{2\le k\le d}\left[(W_k^{(\mathscr{T}_1)}\ge 0)\cup(|U_k^{(\mathscr{T}_1)}|\le 1)\right] = \cap_{2\le k\le d}(|v_1^{(\mathscr{T}_1)}| \ge |v_k^{(\mathscr{T}_1)}|) = \varnothing,\end{aligned}$$

and hence

$$\mathscr{H}_{3,L} \subseteq (\mathscr{T}_1 > T^o_{\eta,\tau}) \tag{A.33}$$

By combining (A.27) and (A.30) for each $k \in [2,d]$ along with (A.33), we have

$$\mathscr{H}_{3,L} \subseteq \left(\sup_{2\le k\le d}\mathscr{T}_{c,k} \le T^o_{\eta,0.5}\right)\cap\left(\inf_{2\le k\le d}\mathscr{T}_{w,k} > T^o_{\eta,\tau}\right)\cap(\mathscr{T}_1 > T^o_{\eta,\tau}). \tag{A.34}$$

(iv) Remark 1 interprets (A.34) and depicts the story of convergence with uniform initialization. With (A.34) ready at hand, on the event $\mathscr{H}_{3,L}$, we apply Lemma 4 for each coordinate $k \in [2,d]$ and all $t \in [T^o_{\eta,0.5}, T^o_{\eta,\tau}]$ to obtain

$$\sqrt{\sum_{k=2}^{d}\left(U_k^{(t)} - U_k^{(T^o_{\eta,0.5})}\prod_{s=T^o_{\eta,0.5}}^{t-1}\left[1-\eta|\mu_4-3|\left((v_1^{(s)})^2-(v_k^{(s)})^2\right)\right]\right)^2} \le 2C_{4,L}\log^{5/2}\delta^{-1}\cdot B^4\cdot d\eta(T^o_{\eta,\tau})^{1/2}, \tag{A.35}$$

On the event $\mathscr{H}_{3,L}$, we have $\mathscr{T}_{c,k} \le T^o_{\eta,0.5}$, $(v_1^{(s)})^2-(v_k^{(s)})^2 \ge 1/(2d)$, $\left|U_k^{(\mathscr{T}_{c,k})}\right| \le 1$ for all $k \in [2,d]$, $s \in [T^o_{\eta,0.5}, T^o_{\eta,\tau})$. Since the left hand of (A.35) is the norm of two vectors subtraction, we use the

triangle inequality of Euclidean norms to lower bound it as

$$\begin{aligned}
&\sqrt{\sum_{k=2}^{d}\left(U_k^{(t)}-U_k^{(T^o_{\eta,0.5})}\prod_{s=T^o_{\eta,0.5}}^{t-1}\left[1-\eta|\mu_4-3|\left((v_1^{(s)})^2-(v_k^{(s)})^2\right)\right]\right)^2}\\
&\geq\sqrt{\sum_{k=2}^{d}\left(U_k^{(t)}\right)^2}-\sqrt{\sum_{k=2}^{d}\left(U_k^{(T^o_{\eta,0.5})}\prod_{s=T^o_{\eta,0.5}}^{t-1}\left[1-\eta|\mu_4-3|\left((v_1^{(s)})^2-(v_k^{(s)})^2\right)\right]\right)^2} \qquad \text{(A.36)}\\
&\geq\left|\tan\angle\left(\mathbf{v}^{(t)},\mathbf{e}_1\right)\right|-\sqrt{d}\cdot\left(1-\frac{\eta}{2d}|\mu_4-3|\right)^{t-T^o_{\eta,0.5}}.
\end{aligned}$$

Recall the definition of $T^o_{\eta,\tau}$ in (3.1). Scaling condition (3.2) guarantees the following

$$\frac{|\mu_4-3|}{2d}\eta\leq 1,\quad \frac{B^8}{|\mu_4-3|}\eta\leq e^{-1}. \qquad \text{(A.37)}$$

Since $-\log\left(1-\frac{\eta}{2d}|\mu_4-3|\right)\geq\frac{\eta}{2d}|\mu_4-3|$ and (A.37) holds, for each positive τ we have relation

$$T^o_{\eta,\tau}\leq 1+\frac{2\tau d}{|\mu_4-3|}\cdot\eta^{-1}\log\left(\frac{|\mu_4-3|}{B^8}\eta^{-1}\right)\leq\frac{2(\tau+1)d}{|\mu_4-3|}\cdot\eta^{-1}\log\left(\frac{|\mu_4-3|}{B^8}\eta^{-1}\right). \qquad \text{(A.38)}$$

Combining (A.35), (A.36) together and using relation (A.38), we have

$$\begin{aligned}
\left|\tan\angle\left(\mathbf{v}^{(t)},\mathbf{e}_1\right)\right|&\leq\sqrt{d}\cdot\left(1-\frac{\eta}{2d}|\mu_4-3|\right)^{t-T^o_{\eta,0.5}}\\
&\quad+\sqrt{\tau+1}C_{3,L}\log^{5/2}\delta^{-1}\cdot\frac{B^4}{|\mu_4-3|^{1/2}}\cdot\sqrt{d^3\eta\log\left(\frac{|\mu_4-3|}{B^8}\eta^{-1}\right)} \qquad \text{(A.39)}
\end{aligned}$$

where constant $C_{3,L}\equiv 2\sqrt{2}C_{4,L}$. To complete proof of Lemma 3, we provide a lower bound on probability of event $\mathscr{H}_{3,L}$ by taking union bound

$$\mathbb{P}(\mathscr{H}_{3,L})\geq 1-\sum_{k=2}^{d}\mathbb{P}(\mathscr{H}^c_{k;4,L})-\mathbb{P}(\mathscr{H}^c_{5,L})-\sum_{k=2}^{d}\mathbb{P}(\mathscr{H}^c_{k;6,L})\geq 1-\left(6\tau+27+\frac{10368}{\log^5\delta^{-1}}\right)d\delta-3\varepsilon$$

Applying the scaling relation (2.9) of $\{\mathbf{u}^{(t)}\}_{t\geq 0}$ and $\{\mathbf{v}^{(t)}\}_{t\geq 0}$ to (A.39) on the event $\mathscr{H}$ completes the proof of (3.3), and hence Lemma 3.

□

Remark 1. (A.34) *combines Lemmas 4, 5 and 6 to depict a complete story of uniform initialization in terms of the iterative process* $\{\boldsymbol{v}^{(t)}\}$*:*

(i) When initialized uniformly at random on the unit sphere, on a high probability event $\mathscr{H}_{5,L}$ *initialization satisfies* $\min_{2\leq k\leq d}W_k^{(0)}=\Omega(\log^{-1}d)$.

(ii) *If* $\min_{2\le k\le d} W_k^{(0)} = \Omega(\log^{-1} d)$, *on a high probability event* $\cap_{2\le k\le d}\mathscr{H}_{k;4,L}$ $\mathbf{v}^{(t)}$ *enters joint intermediate region* $\cap_{2\le k\le d}\mathcal{D}_{mid,k}$ *within* $T^o_{\eta,0.5}$ *iterations.*

(iii) *After* $\mathbf{v}^{(t)}$ *enters joint intermediate region* $\cap_{2\le k\le d}\mathcal{D}_{mid,k}$, *on a high probability event* $\cap_{2\le k\le d}\mathscr{H}_{k;6,L}$ *it will move exponentially fast towards the local minimizer (pair)* $\pm\mathbf{e}_1$ *and then stays in a* $\tilde{O}(\sqrt{d^3\eta})$*-neighborhood of* $\mathbf{e}_1$ *or* $-\mathbf{e}_1$.

A.4 Proof of Theorem 3

Now we are ready to derive the finite-sample error bound and prove Theorem 3.

Proof of Theorem 3. (i) We first provide an upper bound on $T^o_{\eta(T),\tau}$. Plugging in $\eta(T)$ to relation (A.38), we have

$$\begin{aligned} T^o_{\eta(T),\tau} &\le \frac{2(\tau+1)d}{|\mu_4-3|}\cdot\eta(T)^{-1}\log\left(\frac{|\mu_4-3|}{B^8}\eta(T)^{-1}\right) \\ &\le \frac{(\tau+1)}{2\log\left(\frac{|\mu_4-3|^2}{4B^8d}T\right)}T\log\left(\frac{|\mu_4-3|^2T}{4B^8d\log\left(\frac{|\mu_4-3|^2}{4B^8d}T\right)}\right) < \frac{\tau+1}{2}T, \end{aligned} \tag{A.40}$$

since scaling condition (3.2) with constants $C^*_{3,T}\equiv 96C^*_{3,L}$ and $C'_{3,T}>1$ to be determined later guarantees $\log\left(\frac{|\mu_4-3|^2}{4B^8d}T\right)>1$.

(ii) Next we provide a lower bound on $T^o_{\eta(T),\tau}$. Recall in proof of Theorem 3, by Taylor expansion we showed in (A.12) for all $x\in(0,1/3]$ that $\frac{1}{-\log(1-x)}\ge\frac{4}{5x}$. Applying to $T^o_{\eta(T),\tau}$, under scaling condition

$$\frac{4\log T}{T}\le\frac{1}{3},\quad \frac{8B^8}{|\mu_4-3|^2}\cdot\frac{d\log T}{\sqrt{T}}\le 1 \tag{A.41}$$

given by (3.4), we have

$$\begin{aligned} T^o_{\eta(T),\tau} &\ge \frac{\tau\log\left(\frac{|\mu_4-3|}{B^8}\eta(T)^{-1}\right)}{-\log\left(1-\frac{|\mu_4-3|}{2d}\eta(T)\right)} \ge \frac{8\tau d\log\left(\frac{|\mu_4-3|}{B^8}\eta(T)^{-1}\right)}{5|\mu_4-3|\eta(T)} \\ &\ge \frac{2\tau T}{5\log\left(\frac{|\mu_4-3|^2}{4B^8d}T\right)}\cdot\log\left(\frac{|\mu_4-3|^2T}{4B^8d\log\left(\frac{|\mu_4-3|^2}{4B^8d}T\right)}\right)\ge\frac{\tau}{5}T, \end{aligned} \tag{A.42}$$

where we use the elementary inequality $\log\left(\frac{x}{\log x}\right)\ge\frac{1}{2}\log x$ for all $x>1$ since $\frac{|\mu_4-3|^2}{4B^8d}T>1$ is satisfied under scaling condition (3.2).

(iii) From (A.40) and (A.42) we find that $T \in [T^o_{\eta(T),1}+1, T^o_{\eta(T),5}]$. By letting

$$\varepsilon = \left(57 + \frac{10368}{\log^5 \delta^{-1}}\right) d\delta,$$

along with (A.40) we have

$$T^o_{\eta(T),1}\delta^{-1} \le C'_{3,T}\varepsilon^{-1} dT,$$

for some positive, absolute constant $1 < C'_{3,T} \le 10425$, due to $T^o_{\eta(T),1} \le T$ given by (A.40) and $\delta \le e^{-1}$ guaranteed by $\varepsilon \le 1/4$.

The third scaling condition in (3.2) with our pick $\tau = 5$ and $\eta = \eta(T)$ is satisfied by

$$24C^*_{3,L}\log^8(C'_{3,T}\varepsilon^{-1}dT)\cdot\frac{B^8}{|\mu_4-3|^2}\cdot\frac{d^3\log^2 d\log\left(\frac{|\mu_4-3|^2T}{4B^8d}\right)}{T}\log\left(\frac{|\mu_4-3|^2T}{4B^8d\log\left(\frac{|\mu_4-3|^2T}{4B^8d}\right)}\right) \le \frac{\varepsilon^2}{\log^2\varepsilon^{-1}}, \tag{A.43}$$

From Lemma 7 we have $\frac{B^4}{|\mu_4-3|} \ge \frac{1}{8}$. Along with scaling condition (3.4) and $C^*_{3,T} \equiv 96C^*_{3,L}$, we have $T/d \ge 16$ and hence

$$1 \le \log\left(\frac{|\mu_4-3|^2T}{4B^8d}\right) \le 2\log(T/d), \tag{A.44}$$

implying (A.43) holds under scaling condition (3.4).

To verify the second scaling condition in (3.2), we notice that the following holds under (3.4)

$$\frac{B^8}{|\mu_4-3|}\eta(T) = \frac{4B^8d\log\left(\frac{|\mu_4-3|^2}{4B^8d}T\right)}{|\mu_4-3|^2T} \le \frac{8B^8d\log T}{|\mu_4-3|^2T} < e^{-1},$$

and

$$\eta(T) = \frac{4d\log\left(\frac{|\mu_4-3|^2}{4B^8d}T\right)}{|\mu_4-3|T} \le \frac{8\log(T/d)}{|\mu_4-3|(T/d)} < \frac{1}{|\mu_4-3|},$$

due to (A.44) and the elementary inequality $\frac{8\log x}{x} < 1$ for all $x \ge 16$.

Therefore, all scaling conditions required in Lemma 3 are satisfied by scaling condition (3.4) in Theorem 3.

(iv) Using $B^8\eta/|\mu_4-3| \le e^{-1}$ and $\delta \le e^{-1}$ given by (3.2) and $T^o_{\eta,1}+1-T^o_{\eta,0.5} \ge T^o_{\eta,0.5}$ following its definition (3.1), for all $t \in [T^o_{\eta,1}+1, T^o_{\eta,5}]$, on the event $\mathscr{H}_{3,L}$ we have

$$\sqrt{d}\cdot\left(1-\frac{\eta}{2d}|\mu_4-3|\right)^{t-T^o_{\eta,0.5}} \le \sqrt{d}\cdot\left(1-\frac{\eta}{2d}|\mu_4-3|\right)^{T^o_{\eta,0.5}} \le \frac{B^4}{|\mu_4-3|^{1/2}}\cdot\sqrt{d\eta}$$
$$\le (3-\sqrt{6})C_{3,L}\log^{5/2}\delta^{-1}\cdot\frac{B^4}{|\mu_4-3|^{1/2}}\cdot\sqrt{d^3\eta\log\left(\frac{|\mu_4-3|}{B^8}\eta^{-1}\right)}.$$

Algorithm 2 Online Tensorial ICA, Estimating Multiple Components

Initialize independent $\mathbf{u}^{(0;1)},\dots,\mathbf{u}^{(0;N)}$ uniformly at random from the unit sphere and select stepsize η appropriately

for $t=1,2,\dots$ **do**

Draw one observation $\boldsymbol{X}^{(t)}$ from streaming data

Update iteration $\mathbf{u}^{(t;j)}$ $(j=1,\dots,N)$ in parallel by

$$\mathbf{u}^{(t;j)}=\Pi_1\left\{\mathbf{u}^{(t-1;j)}+\eta\cdot\operatorname{sign}(\mu_4-3)\left((\mathbf{u}^{(t-1;j)})^\top\boldsymbol{X}^{(t)}\right)^3\boldsymbol{X}^{(t)}\right\}$$

where $\Pi_1\{\bullet\}=\|\bullet\|^{-1}\bullet$ denotes the projection operator onto the unit sphere

end for

Return $\left\{\mathbf{u}^{(T;1)},\dots,\mathbf{u}^{(T;N)}\right\}$

Therefore, from Lemma 3, on the event $\mathscr{H}_{3,L}$ we have for all $t\in[T^o_{\eta,1}+1,T^o_{\eta,5}]$ that

$$\left|\tan\angle\left(\mathbf{u}^{(t)},\mathbf{a}_{\mathscr{I}}\right)\right|\le 3C_{3,L}\log^{5/2}\delta^{-1}\cdot\frac{B^4}{|\mu_4-3|^{1/2}}\cdot\sqrt{d^3\eta\log\left(\frac{|\mu_4-3|}{B^8}\eta^{-1}\right)}. \tag{A.45}$$

Plugging in our choice of $\tau=5$ and $\eta(T)=\frac{4d\log\left(\frac{|\mu_4-3|^2}{4B^8d}T\right)}{|\mu_4-3|T}$ to (A.45) and using (A.44), we know that there exists an event $\mathscr{H}_{3,T}\equiv\mathscr{H}_{3,L}$ with $\mathbb{P}(\mathscr{H}_{3,T})\ge 1-4\varepsilon$ such that on event $\mathscr{H}_{3,T}$ we have

$$\left|\tan\angle\left(\mathbf{u}^{(T)},\mathbf{a}_{\mathscr{I}}\right)\right|\le C_{3,T}\log^{5/2}(C'_{3,T}\varepsilon^{-1}d)\cdot\frac{B^4}{|\mu_4-3|}\cdot\sqrt{\frac{d^4\log^2 T}{T}},$$

where constant $C_{3,T}\equiv 12C_{3,L}$.

□

B Estimating Multiple Components

Let us turn to the case of estimating multiple components. Instead of using a classical deflationary orthogonalization device (Hyvarinen, 1999), here since the (standardized) kurtoses μ_4-3 are identical for all Z_i's we take a considerably simpler approach: we initialize $\{\mathbf{u}^{(0;1)},\dots,\mathbf{u}^{(0;N)}\}$ independently and uniformly from $\mathcal{D}_1$, and then run our online tensorial ICA iteration (1.3) in parallel on N machines using the same stream of data observations, obtaining N estimators $\{\mathbf{u}^{(T;1)},\dots,\mathbf{u}^{(T;N)}\}$. We formally write the above procedures as Algorithm 2.

As readers may recall in the proofs of Lemma 3 and Theorem 3, the random variable $\mathscr{I}$ is determined purely at initialization of Algorithm 1. Mathematically, it is the index of independent component pairs $\pm\mathbf{a}_i$

which it has minimal spherical distance to at initialization, and is uniformly distributed in $[d]$, i.e.

$$\mathscr{I} \equiv \operatorname*{argmin}_{i\in[d]} \tan^2 \angle\left(\mathbf{u}^{(0)}, \mathbf{a}_i\right). \tag{B.1}$$

Here the argmin operator produces a unique solution almost surely due to continuous distribution. In this subsection, we sample multiple independent initializations to parallelize Algorithm 1 and obtain sharp estimates for multiple (or all d) independent component pairs $\pm\mathbf{a}_1, \ldots, \pm\mathbf{a}_d$. Under proper scaling conditions with high probability each $\mathbf{u}^{(T;j)}$ is a sharp estimate of an independent component pair equiprobabilistically drawn from $\{\pm\mathbf{a}_1, \ldots, \pm\mathbf{a}_d\}$, and hence a simple combinatoric argument allows us to obtain multiple (or all) components simultaneously:

Theorem 5 (Finite-Sample Error Bound for Multiple Components Estimation)**.** *Let the dimension $d \geq 2$, let the number of machines $N \geq 1$, let Assumption 1 hold, and let initializations $\mathbf{u}^{(0;1)}, \ldots, \mathbf{u}^{(0;N)}$ be i.i.d. uniformly sampled from the unit sphere $\mathcal{D}_1$. For any sample size $T \geq 2$ we pick the stepsizes as in (3.5). Then for fixed positive $\varepsilon \in (0, 1/(5N)]$ satisfying the scaling condition (3.6), there exist N i.i.d. uniformly distributed random variable $\mathscr{I}_1, \ldots, \mathscr{I}_N \in [d]$ and an event $\mathscr{H}_{5,T}$ with $\mathbb{P}(\mathscr{H}_{5,T}) \geq 1 - 5N\varepsilon$ such that on $\mathscr{H}_{5,T}$, running Algorithm 2 for $T/2$ iterates in parallel with stepsize $\eta_1(T)$ followed by $T/2$ iterates in parallel with stepsize $\eta_2(T)$ outputs $\{\mathbf{u}^{(T;1)}, \ldots, \mathbf{u}^{(T;N)}\}$ satisfying, for all $j = 1, \ldots, N$,*

$$\left|\tan\angle\left(\mathbf{u}^{(T;j)}, \mathbf{a}_{\mathscr{I}_j}\right)\right| \leq C_{5,T} \log^{5/2}(C'_{5,T} d\varepsilon^{-1}) \cdot \frac{B^4}{|\mu_4 - 3|} \cdot \sqrt{\frac{d\log^2 T}{T}}, \tag{B.2}$$

where $C_{5,T}, C'_{5,T}$ are positive, absolute constants. Additionally, when $N \geq \lceil d\log\varepsilon^{-1}\rceil$, with probability at least $1 - 6N\varepsilon$, (B.2) is satisfied with the set of random variables $\{\mathscr{I}_j : j \in [N]\}$ containing $[d]$.

Theorem 5 states that under the scaling condition $T = \tilde{\Omega}(N^2 d^4)$, with high probability $\mathbf{u}^{(T;j)}$ provides an $\tilde{O}(\sqrt{d/T})$ finite-sample error estimate for each independent component pair $\pm\mathbf{a}_{\mathscr{I}_j}$ where $\mathscr{I}_j$ are i.i.d. uniforms in $[d]$. When $N = \tilde{\Theta}(d\log d)$ the set of N estimators $\{\mathbf{u}^{(T;1)}, \ldots, \mathbf{u}^{(T;N)}\}$ contains at least one sharp estimate of each independent component $\mathbf{a}_i, i \in [d]$ with high probability.

The analysis of multiple component estimation is relatively simple with the results of single component estimation at hand. We proceed with the proof:

Proof of Theorem 5. Choosing constants $C_{5,T} \equiv \sqrt{2}C_{2,T}, C'_{5,T} \equiv 2\max\{C'_{2,T}, C'_{3,T}\}$, applying Corollary 4 and taking union bound, we know that there exists an event $\mathscr{H}_{5,T}$ with $\mathbb{P}(\mathscr{H}_{5,T}) \geq 1 - 5N\varepsilon$, such that on event $\mathscr{H}_{5,T}$, (B.2) holds for each $j \in [N]$, and $\mathscr{I}_1, \ldots, \mathscr{I}_N$ are i.i.d. random variables uniformly distributed on $[d]$.

In addition, if the number of machines $N \geq \lceil d\log\varepsilon^{-1}\rceil$, then the probability that event $\mathscr{H}_{5,T}$ occurs but component pair $\pm\mathbf{a}_i$ ($i \in [d]$) has *not* been estimated by all parallel iterations is

$$\mathbb{P}\left(\mathscr{H}_{5,T} \cap \left(\cap_{1\leq j\leq N}(\mathscr{I}_j \neq i)\right)\right) \leq \mathbb{P}\left(\cap_{1\leq j\leq N}(\mathscr{I}_j \neq i)\right) \leq \left(1 - \frac{1}{d}\right)^N \leq \exp\left(-\frac{N}{d}\right) \leq \varepsilon.$$

By taking union bound, we know that the probability that event $\mathscr{H}_{5,T}$ occurs and $\{\mathscr{I}_j : j \in [N]\}$ contains all component indices $[d]$ satisfies

$$\mathbb{P}\bigg(\mathscr{H}_{5,T} \cap (\{\mathscr{I}_j : j \in [N]\} \supseteq [d])\bigg) + \sum_{i=1}^{d} \mathbb{P}\left(\mathscr{H}_{5,T} \cap (\cap_{1\le j\le N}(\mathscr{I}_j \neq i))\right) \ge \mathbb{P}(\mathscr{H}_{5,T}),$$

and hence

$$\mathbb{P}\bigg(\mathscr{H}_{5,T} \cap (\{\mathscr{I}_j : j \in [N]\} \supseteq [d])\bigg) \ge (1-5N-d)\varepsilon \ge 1-6N\varepsilon,$$

concluding this proof. □

C Secondary Lemmas in Warm Initialization Analysis

For notational simplicity, we denote $\mathbf{v} \equiv \mathbf{v}^{(t-1)}$ and $\boldsymbol{Y} \equiv \boldsymbol{Y}^{(t)}$. We first provide a lemma on Orlicz ψ_2-norm of $\mathbf{v}^\top \boldsymbol{Y}$ and the relation between B and μ_4.

Lemma 7. *Let Assumption 1 hold. For each rotated observation $\boldsymbol{Y}$ and any unit vector $\mathbf{v}$, we have Orlicz ψ_2-norm $\|\mathbf{v}^\top \boldsymbol{Y}\|_{\psi_2} \le B$ and the following relation of B and μ_4*

$$\frac{B^4}{|\mu_4 - 3|} \ge \frac{1}{8}. \tag{C.1}$$

With the bound on Orlicz ψ_2-norm given in Lemma 7 and $T^*_{\eta,1}$ defined in (2.2), we introduce truncation barrier parameter

$$\mathcal{B}_* \equiv B\log^{1/2}(T^*_{\eta,1}\delta^{-1}), \tag{C.2}$$

where $\delta \in (0, e^{-1}]$ is some fixed positive. For each coordinate $k \in [2,d]$ define the first time the norm of a data observation exceeds the truncation barrier $\mathcal{B}_*$ as

$$\mathscr{T}_{\mathcal{B}_*,k} = \inf\left\{t \ge 1 : \left|\mathbf{v}^{(t-1)\top}\boldsymbol{Y}^{(t)}\right| > \mathcal{B}_* \text{ or } \left|Y_1^{(t)}\right| > \mathcal{B}_* \text{ or } \left|Y_k^{(t)}\right| > \mathcal{B}_*\right\}, \tag{C.3}$$

C.1 Proof of Lemma 2

For each $t \ge 1$, we define random variable

$$Q^{(t)}_{U,k} \equiv U_k^{(t)} - U_k^{(t-1)} - \operatorname{sign}(\mu_4 - 3)\eta \cdot (\mathbf{v}^\top \boldsymbol{Y})^3 v_1^{-2}(v_1 Y_k - v_k Y_1). \tag{C.4}$$

Lemma 8. *Let $\mathcal{B}$ be any positive value. For each coordinate $k \in [2,d]$ and any $t \ge 1$, on the event*

$$\mathscr{H}^{(t)}_{k;8,L} \equiv \left(|\mathbf{v}^\top \boldsymbol{Y}| \le \mathcal{B},\ |Y_1| \le \mathcal{B},\ |Y_k| \le \mathcal{B},\ v_1^2 \ge 1/d,\ v_1^2 \ge v_k^2\right), \tag{C.5}$$

for stepsize $\eta \le 1/(2\mathcal{B}^4 d^{1/2})$ we have $|Q^{(t)}_{U,k}| \le 4\mathcal{B}^8\eta^2 v_1^{-2}$.

Lemma 9. *Let Assumption 1 hold. For each coordinate $k \in [2,d]$ and any $t \geq 1$, we have*

$$\mathbb{E}\left[\operatorname{sign}(\mu_4-3)\eta\cdot(\mathbf{v}^\top \boldsymbol{Y})^3 v_1^{-2}(v_1 Y_k - v_k Y_1)\,\middle|\,\mathscr{F}_{t-1}\right] = -\eta|\mu_4-3|\cdot(v_1^2-v_k^2)U_k^{(t-1)}. \tag{C.6}$$

For each $k \in [2,d]$ and $t \geq 1$, at the t-th iteration we define

$$e_k^{(t)} \equiv \operatorname{sign}(\mu_4-3)\eta\cdot(\mathbf{v}^\top \boldsymbol{Y})^3 v_1^{-2}(v_1 Y_k - v_k Y_1) + \eta|\mu_4-3|\cdot(v_1^2-v_k^2)U_k^{(t-1)} \tag{C.7}$$

which, indexed by t, forms a sequence of martingale differences with respect to $\mathscr{F}_{t-1}$.

Lemma 10. *For each coordinate $k \in [2,d]$ and any $t \geq 1$, $U_k^{(t)}$ has linear representation*

$$U_k^{(t)} = U_k^{(0)} - \eta|\mu_4-3|\sum_{s=0}^{t-1}\left((v_1^{(s)})^2-(v_k^{(s)})^2\right)U_k^{(s)} + \sum_{s=1}^{t} Q_{U,k}^{(s)} + \sum_{s=1}^{t} e_k^{(s)}. \tag{C.8}$$

Lemma 11. *Let Assumption 1 hold and initialization $\mathbf{v}^{(0)} \in \mathcal{D}_{warm}$. Let $\delta \in (0, e^{-1}]$ and τ be any fixed positive. For each coordinate $k \in [2,d]$, there exists an event $\mathscr{H}_{k;11,L}$ satisfying*

$$\mathbb{P}(\mathscr{H}_{k;11,L}) \geq 1 - \left(6 + \frac{5184}{\log^5 \delta^{-1}}\right)\delta,$$

such that on the event $\mathscr{H}_{k;11,L}$ the following concentration result holds

$$\max_{1\leq t\leq T^*_{\eta,\tau}\wedge\mathscr{T}_x}\left|\sum_{s=1}^{t} e_k^{(s)}\right| \leq C_{11,L}\log^{5/2}\delta^{-1}\cdot B^4\cdot\eta(T^*_{\eta,\tau})^{1/2},$$

where $C_{11,L}$ is a positive, absolute constant.

Lemma 12. *Let $\delta \in (0, e^{-1}]$ and τ be any fixed positive. For each coordinate $k \in [2,d]$, we have*

$$\mathbb{P}\left(\mathscr{T}_{\mathcal{B}_*,k} \leq T^*_{\eta,\tau}\right) \leq 6(\tau+1)\delta. \tag{C.9}$$

With the above secondary lemmas at hand, we are now ready to prove Lemma 2.

Proof of Lemma 2. We recall the definition of stopping time $\mathscr{T}_x$ in (2.13). Because stepsize $\eta \leq 1/(2\mathcal{B}_*^4 d^{1/2})$ holds under scaling condition (2.3), on the event $(t \leq T^*_{\eta,\tau}\wedge\mathscr{T}_x)\cap(\mathscr{T}_{\mathcal{B}_*,k} > T^*_{\eta,\tau}) \subseteq \mathscr{H}^{(t)}_{k;8,L}$, we have $v_1^{-2} \leq \frac{3}{2}$, and applying Lemma 8 gives

$$|Q_{U,k}^{(t)}| \leq 4\mathcal{B}_*^8\eta^2 v_1^{-2} \leq 6B^8\eta^2\log^4(T^*_{\eta,1}\delta^{-1}).$$

We define event $\mathscr{H}_{k;2,L} \equiv (\mathscr{T}_{\mathcal{B}_*,k} > T^*_{\eta,\tau})\cap\mathscr{H}_{k;11,L}$. Applying Lemmas 10 and 11, on the event $\mathscr{H}_{k;2,L}$ for all

$t \le T^*_{\eta,\tau} \wedge \mathscr{T}_x$ we have

$$\begin{aligned}&\left|U_k^{(t)} - U_k^{(0)} + \eta|\mu_4 - 3| \sum_{s=0}^{t-1} \left((v_1^{(s)})^2 - (v_k^{(s)})^2\right) U_k^{(s)}\right| \\ \le & T^*_{\eta,\tau} \cdot 6B^8\eta^2 \log^4(T^*_{\eta,1}\delta^{-1}) + C_{11,L}\log^{5/2}\delta^{-1} \cdot B^4 \cdot \eta (T^*_{\eta,\tau})^{1/2}.\end{aligned} \tag{C.10}$$

Scaling condition (2.3) and definition of $T^*_{\eta,\tau}$ in (2.2) imply that

$$B^4\eta(T^*_{\eta,\tau})^{1/2}\log^4(T^*_{\eta,1}\delta^{-1}) \le \log^{5/2}\delta^{-1}, \tag{C.11}$$

Combining (C.10) and (C.11), we have

$$\left|U_k^{(t)} - U_k^{(0)} + \eta|\mu_4 - 3| \sum_{s=0}^{t-1} \left((v_1^{(s)})^2 - (v_k^{(s)})^2\right) U_k^{(s)}\right| \le C_{2,L}\log^{5/2}\delta^{-1} \cdot B^4 \cdot \eta (T^*_{\eta,\tau})^{1/2},$$

where constant $C_{2,L} \equiv C_{11,L} + 6$. Scaling condition (2.3) implies $\eta|\mu_4 - 3|((v_1^{(s)})^2 - (v_k^{(s)})^2) \in [0,1)$ for all $s < T^*_{\eta,\tau} \wedge \mathscr{T}_x$. Using reversed Gronwall Lemma 20, on the event $\mathscr{H}_{k;2,L}$ we have the following holds for all $t \le T^*_{\eta,\tau} \wedge \mathscr{T}_x$

$$\left|U_k^{(t)} - U_k^{(0)} \prod_{s=0}^{t-1} \left[1 - \eta|\mu_4 - 3|\left((v_1^{(s)})^2 - (v_k^{(s)})^2\right)\right]\right| \le 2C_{2,L}\log^{5/2}\delta^{-1} \cdot B^4 \cdot \eta (T^*_{\eta,\tau})^{1/2}.$$

To complete proof of Lemma 2, we apply Lemmas 11 and 12 and take union bound to obtain

$$\mathbb{P}(\mathscr{H}_{k;2,L}) \ge 1 - \mathbb{P}(\mathscr{T}_{\mathcal{B}_*,k} \le T^*_{\eta,\tau}) - \mathbb{P}(\mathscr{H}^c_{k;11,L}) \ge 1 - \left(6\tau + 12 + \frac{5184}{\log^5\delta^{-1}}\right)\delta.$$

□

C.2 Proof of Secondary Lemmas

Proof of Lemma 7. (i) Because random vector $\boldsymbol{Y}$ is a permutation of $\boldsymbol{Z}$, from Assumption 1 we know that each Y_i is sub-Gaussian with parameter $\sqrt{3/8}B$. By independence of Y_i, for any unit vector $\mathbf{v}$ and all $\lambda \in \mathbb{R}$ we have

$$\mathbb{E}\exp\left(\lambda \mathbf{v}^\top \boldsymbol{Y}\right) = \prod_{i=1}^d \mathbb{E}\exp(\lambda v_i Y_i) \le \prod_{i=1}^d \exp\left(\frac{\lambda^2 v_i^2}{2} \cdot \left(\sqrt{\frac{3}{8}}B\right)^2\right) = \exp\left(\frac{\lambda^2}{2} \cdot \left(\sqrt{\frac{3}{8}}B\right)^2\right),$$

which implies that $\mathbf{v}^\top \boldsymbol{Y}$ is also sub-Gaussian with parameter $\sqrt{3/8}B$. Theorem 2.6 in Wainwright (2019) shows that any sub-Gaussian random variable X with parameter σ satisfies

$$\mathbb{E}\exp\left(\frac{\lambda X^2}{2\sigma^2}\right) \le \frac{1}{\sqrt{1-\lambda}}$$

for all $\lambda \in [0,1)$. By choosing $\lambda = 3/4$, $\sigma = \sqrt{3/8}B$ and $X = \mathbf{v}^\top \boldsymbol{Y}$, we have

$$\mathbb{E}\exp\left(\frac{(\mathbf{v}^\top \boldsymbol{Y})^2}{B^2}\right) \le 2 \quad \text{i.e.} \quad \|\mathbf{v}^\top \boldsymbol{Y}\|_{\psi_2} \le B. \tag{C.12}$$

We refer the readers to §F in Appendix for more details on Orlicz ψ_2-norm.

(ii) Applying Markov's inequality to (C.12) with $\mathbf{v} = \mathbf{e}_i$ gives

$$\mathbb{P}(Y_i^2 \ge \lambda) = \mathbb{P}\left(\exp\left(\frac{Y_i^2}{B^2}\right) \ge \exp\left(\frac{\lambda}{B^2}\right)\right) \le \exp\left(-\frac{\lambda}{B^2}\right)\mathbb{E}\exp\left(\frac{Y_i^2}{B^2}\right) \le 2\exp\left(-\frac{\lambda}{B^2}\right).$$

Hence we have

$$\mu_4 = \mathbb{E}Y_i^4 = \int_0^\infty \mathbb{P}(Y_i^2 \ge \lambda)\cdot 2\lambda \mathrm{d}\lambda \le 4\int_0^\infty \lambda \exp\left(-\frac{\lambda}{B^2}\right)\mathrm{d}\lambda = 4B^4,$$

and hence unconditionally

$$\frac{B^4}{|\mu_4 - 3|} \ge \frac{\mu_4}{4|\mu_4 - 3|} \ge \frac{1}{8},$$

due to $\mu_4 = \mathbb{E}Y_i^4 \ge (\mathbb{E}Y_i^2)^2 = 1$.

□

Proof of Lemma 8. Let $\eta_S \equiv \eta \cdot \mathrm{sign}(\mu_4 - 3)$ for notational simplicity. We recall definitions of iteration $U_k^{(t)}$ in (2.11) and random variable $Q_{U,k}^{(t)}$ in (C.4). Using update formula (2.8), we have

$$\begin{aligned} U_k^{(t)} - U_k^{(t-1)} &= \frac{v_k + \eta_S \cdot (\mathbf{v}^\top \boldsymbol{Y})^3 Y_k}{v_1 + \eta_S \cdot (\mathbf{v}^\top \boldsymbol{Y})^3 Y_1} - \frac{v_k}{v_1} = \frac{\left(v_k + \eta_S \cdot (\mathbf{v}^\top \boldsymbol{Y})^3 Y_k\right) v_1 - \left(v_1 + \eta_S \cdot (\mathbf{v}^\top \boldsymbol{Y})^3 Y_1\right) v_k}{\left(v_1 + \eta_S \cdot (\mathbf{v}^\top \boldsymbol{Y})^3 Y_1\right) v_1} \\ &= \eta_S \cdot \left(1 + \eta_S \cdot (\mathbf{v}^\top \boldsymbol{Y})^3 \frac{Y_1}{v_1}\right)^{-1} \cdot (\mathbf{v}^\top \boldsymbol{Y})^3 v_1^{-2} \left(v_1 Y_k - v_k Y_1\right). \end{aligned}$$

Along with (C.4), we obtain

$$Q_{U,k}^{(t)} = \eta_S \cdot \left[\left(1 + \eta_S \cdot (\mathbf{v}^\top \boldsymbol{Y})^3 \frac{Y_1}{v_1}\right)^{-1} - 1\right] \cdot (\mathbf{v}^\top \boldsymbol{Y})^3 v_1^{-2} \left(v_1 Y_k - v_k Y_1\right).$$

For any $|x| \le \frac{1}{2}$, summation of geometric series gives

$$\left|(1+x)^{-1}-1\right| = |x|\left|\sum_{k=0}^{\infty}(-x)^k\right| \le 2|x|.$$

On the event $\mathscr{H}^{(t)}_{k;8,L}$ defined earlier in (C.5), since $\left|\eta_S\cdot(\mathbf{v}^\top \boldsymbol{Y})^3 Y_1/v_1\right| \le B^4 d^{1/2}\eta \le 1/2$, we have

$$\begin{aligned}
|Q^{(t)}_{U,k}| &\le \eta\cdot\left|\left(1+\eta_S\cdot(\mathbf{v}^\top\boldsymbol{Y})^3\frac{Y_1}{v_1}\right)^{-1}-1\right|\cdot|\mathbf{v}^\top\boldsymbol{Y}|^3 v_1^{-2}|v_1Y_k - v_kY_1| \\
&\le \eta\cdot 2\left|\eta_S\cdot(\mathbf{v}^\top\boldsymbol{Y})^3\frac{Y_1}{v_1}\right|\cdot|\mathbf{v}^\top\boldsymbol{Y}|^3 v_1^{-2}|v_1Y_k - v_kY_1| \\
&= 2\eta^2 v_1^{-2}|Y_1||\mathbf{v}^\top\boldsymbol{Y}|^6\left|Y_k - \frac{v_k}{v_1}Y_1\right| \\
&\le 4\mathcal{B}^8\eta^2 v_1^{-2}.
\end{aligned}$$

□

Proof of Lemma 9. Under Assumption 1, for all $k\in[d]$ we have

$$\mathbb{E}\left[(\mathbf{v}^\top\boldsymbol{Y})^3 Y_k\middle|\mathscr{F}_{t-1}\right] = \mu_4 v_k^3 + 3v_k\sum_{i=1,i\neq k}^{d} v_i^2 = (\mu_4-3)v_k^3 + 3v_k. \tag{C.13}$$

Recall that $U_k^{(t-1)} = v_k/v_1$, then

$$\begin{aligned}
&\mathbb{E}\left[\operatorname{sign}(\mu_4-3)\eta\cdot(\mathbf{v}^\top\boldsymbol{Y})^3 v_1^{-2}(v_1Y_k - v_kY_1)\middle|\mathscr{F}_{t-1}\right] \\
&= \operatorname{sign}(\mu_4-3)\eta\cdot v_1^{-2}\left((\mu_4-3)v_1v_k^3 + 3v_1v_k - (\mu_4-3)v_kv_1^3 - 3v_kv_1\right) \\
&= -\eta|\mu_4-3|\cdot(v_1^2-v_k^2)U_k^{(t-1)}.
\end{aligned}$$

□

Proof of Lemma 10. From definitions (C.4) and (C.7), by applying (C.6) in Lemma 9,

$$U_k^{(s)} - U_k^{(s-1)} = -\eta|\mu_4-3|\cdot\left((v_1^{(s-1)})^2 - (v_k^{(s-1)})^2\right)U_k^{(s-1)} + Q^{(s)}_{U,k} + e_k^{(s)}. \tag{C.14}$$

Iteratively applying (C.14) for $s=1,\ldots,t$ gives (C.8). □

Proof of Lemma 11. Under Assumption 1, we apply Lemmas 7, 21 and 22 to obtain

$$\begin{aligned}\left\|\eta\cdot v_1^{-2}(\mathbf{v}^\top \boldsymbol{Y})^3(v_1Y_k-v_kY_1)\right\|_{\psi_{1/2}} &\le \eta v_1^{-2}\cdot\|(\mathbf{v}^\top\boldsymbol{Y})^2\|_{\psi_1}\cdot\|(\mathbf{v}^\top\boldsymbol{Y})(v_1Y_k-v_kY_1)\|_{\psi_1}\\ &\le \eta v_1^{-2}\cdot\|\mathbf{v}^\top\boldsymbol{Y}\|_{\psi_2}^3\cdot\|v_1Y_k-v_kY_1\|_{\psi_2}\\ &\le \eta|v_1|^{-1}\cdot\|\mathbf{v}^\top\boldsymbol{Y}\|_{\psi_2}^3\cdot(\|Y_k\|_{\psi_2}+|v_k/v_1|\|Y_1\|_{\psi_2})\\ &\le B^4\eta\cdot|v_1|^{-1}(1+|v_k/v_1|).\end{aligned}$$

Recall the definition of martingale difference sequence $\{e_k^{(t)}\}_{t\ge1}$ in (C.7). By applying Lemma 23, we have

$$\begin{aligned}\left\|e_k^{(t)}\right\|_{\psi_{1/2}} &= \left\|\eta\cdot v_1^{-2}(\mathbf{v}^\top\boldsymbol{Y})^3(v_1Y_k-v_kY_1)-\mathbb{E}\left[\eta\cdot v_1^{-2}(\mathbf{v}^\top\boldsymbol{Y})^3(v_1Y_k-v_kY_1)\right]\right\|_{\psi_{1/2}}\\ &\le C'_{23,L}\left\|\eta\cdot v_1^{-2}(\mathbf{v}^\top\boldsymbol{Y})^3(v_1Y_k-v_kY_1)\right\|_{\psi_{1/2}}\\ &\le C'_{23,L}B^4\eta\cdot|v_1|^{-1}(1+|v_k/v_1|).\end{aligned}\tag{C.15}$$

Because initialization $\mathbf{v}^{(0)}\in\mathcal{D}_{\text{warm}}$, on the event $(t\le\mathscr{T}_x)$ for $\mathscr{T}_x$ earlier defined in (2.13), we have $|v_1|^{-1}\le\sqrt{3}/\sqrt{2}$, $|v_k/v_1|\le1/\sqrt{2}$, and then

$$\left\|e_k^{(t)}\right\|_{\psi_{1/2}}\le 3C'_{23,L}B^4\eta.$$

Because $1_{(t\le\mathscr{T}_x)}\in\mathscr{F}_{t-1}$, we know that $\{e_k^{(t)}1_{(t\le\mathscr{T}_x)}\}$ forms a martingale difference sequence with respect to $\mathscr{F}_{t-1}$. Additionally, because $\|e_k^{(t)}1_{(t\le\mathscr{T}_x)}\|_{\psi_{1/2}}\le\|e_k^{(t)}\|_{\psi_{1/2}}$, we have

$$\left\|e_k^{(t)}1_{(t\le\mathscr{T}_x)}\right\|_{\psi_{1/2}}\le 3C'_{23,L}B^4\eta.$$

With the bound given above, we apply Theorem 6 with $\alpha=1/2$ and obtain for all $\delta\in(0,e^{-1}]$,

$$\begin{aligned}&\mathbb{P}\left(\max_{1\le t\le T^*_{\eta,\tau}\wedge\mathscr{T}_x}\left|\sum_{s=1}^t e_k^{(s)}\right|\ge C_{11,L}B^4\eta(T^*_{\eta,\tau})^{1/2}\log^{5/2}\delta^{-1}\right)\\ =&\,\mathbb{P}\left(\max_{1\le t\le T^*_{\eta,\tau}}\left|\sum_{s=1}^t e_k^{(s)}1_{(s\le\mathscr{T}_x)}\right|\ge C_{11,L}B^4\eta(T^*_{\eta,\tau})^{1/2}\log^{5/2}\delta^{-1}\right)\\ \le&\,2\left[3+6^4\frac{64\cdot T^*_{\eta,\tau}\cdot 9C'^2_{23,L}B^8\eta^2}{C^2_{11,L}B^8\eta^2T^*_{\eta,\tau}\log^5\delta^{-1}}\right]\exp\left\{-\left(\frac{C^2_{11,L}B^8\eta^2T^*_{\eta,\tau}\log^5\delta^{-1}}{32\cdot T^*_{\eta,\tau}\cdot 9C'^2_{23,L}B^8\eta^2}\right)^{\frac15}\right\}\\ =&\left(6+\frac{5184}{\log^5\delta^{-1}}\right)\delta,\end{aligned}$$

where constant $C_{11,L}=12\sqrt{2}C'_{23,L}$. □

Proof of Lemma 12. Recall definition of $\mathcal{B}_*$ in (C.2), $\mathscr{T}_{\mathcal{B}_*,k}$ in (C.3) and $T^*_{\eta,\tau}$ in (2.2). Using Markov in-

equality and Lemma 7, we have

$$\mathbb{P}(|\mathbf{v}^\top \boldsymbol{Y}| > \mathcal{B}_*) = \mathbb{P}\left(\frac{|\mathbf{v}^\top \boldsymbol{Y}|^2}{B^2} > \frac{\mathcal{B}_*^2}{B^2}\right) \le \exp\left(-\frac{\mathcal{B}_*^2}{B^2}\right) \cdot \mathbb{E}\exp\left(\frac{|\mathbf{v}^\top \boldsymbol{Y}|^2}{B^2}\right) \le \frac{2\delta}{T^*_{\eta,1}}.$$

Similarly we also have

$$\mathbb{P}(|Y_1| > \mathcal{B}_*) \le \frac{2\delta}{T^*_{\eta,1}}, \quad \mathbb{P}(|Y_k| > \mathcal{B}_*) \le \frac{2\delta}{T^*_{\eta,1}}.$$

Taking union bound,

$$\begin{aligned}\mathbb{P}\left(\mathscr{T}_{\mathcal{B}_*,k} \le T^*_{\eta,\tau}\right) &\le \sum_{t=1}^{T^*_{\eta,\tau}} \left(\mathbb{P}(|\mathbf{v}^{(t-1)\top} \boldsymbol{Y}^{(t)}| > \mathcal{B}_*) + \mathbb{P}(|Y_1^{(t)}| > \mathcal{B}_*) + \mathbb{P}(|Y_k^{(t)}| > \mathcal{B}_*)\right) \\ &\le 3T^*_{\eta,\tau} \cdot \frac{2\delta}{T^*_{\eta,1}} \le 6(\tau+1)\delta,\end{aligned}$$

where we use the elementary inequality $\lceil \tau x \rceil \le (\tau+1)\lceil x \rceil$ for all $\tau, x \ge 0$ to obtain $T^*_{\eta,\tau}/T^*_{\eta,1} \le \tau+1$. □

D Secondary Lemmas in Uniform Initialization Analysis

For notational simplicity, we denote $\mathbf{v} \equiv \mathbf{v}^{(t-1)}$, $\boldsymbol{Y} \equiv \boldsymbol{Y}^{(t)}$. Recall that we bounded the Orlicz ψ_2-norm of each Y_i and $\mathbf{v}^\top \boldsymbol{Y}$ by B using Lemma 7 in Appendix C and introduced truncation barrier $\mathcal{B}_*$ in (C.2) under warm initialization condition, based on rescaled time $T^*_{\eta,\tau}$. Under uniform initialization condition, we consider a different rescaled time $T^o_{\eta,\tau}$ defined in (3.1). Accordingly, we introduce a slightly larger truncation barrier based on $T^o_{\eta,\tau}$

$$\mathcal{B}_o \equiv B\log^{1/2}(T^o_{\eta,1}\delta^{-1}), \tag{D.1}$$

where $\delta \in (0, e^{-1}]$ is some fixed positive. For each coordinate $k \in [2,d]$ define the first time the norm of a data observation exceeds the truncation barrier $\mathcal{B}_o$ as

$$\mathscr{T}_{\mathcal{B}_o,k} = \inf\left\{t \ge 1 : |\mathbf{v}^{(t-1)\top} \boldsymbol{Y}^{(t)}| > \mathcal{B}_o \text{ or } |Y_1^{(t)}| > \mathcal{B}_o \text{ or } |Y_k^{(t)}| > \mathcal{B}_o\right\}. \tag{D.2}$$

D.1 Proof of Lemma 4

Proof of Lemma 4 shares Lemma 8, 9 and 10 with proof of Lemma 2. With the new truncation barrier $\mathcal{B}_o$ given in (D.1) and the new rescaled time $T^o_{\eta,\tau}$ earlier defined in (3.1), we plug in $\mathcal{B} = \mathcal{B}_o$ in Lemma 8 and introduce a new concentration lemma and a new tail probability lemma.

Lemma 13. *For any fixed coordinate $k \in [2,d]$, let Assumption 1 hold and initialization $\mathbf{v}^{(0)} \in \mathcal{D}_{mid,k}$. Let δ, τ be any fixed positives. Then there exists an event $\mathscr{H}_{k;13,L}$ satisfying*

$$\mathbb{P}(\mathscr{H}_{k;13,L}) \ge 1 - \left(6 + \frac{5184}{\log^5 \delta^{-1}}\right)\delta,$$

such that on event $\mathscr{H}_{k;13,L}$ the following concentration result holds

$$\max_{1\le t\le T^o_{\eta,\tau}\wedge\mathscr{T}_{w,k}}\left|\sum_{s=1}^{t} e_k^{(s)}\right| \le C_{13,L}\log^{5/2}\delta^{-1}\cdot B^4\cdot d^{1/2}\eta(T^o_{\eta,\tau})^{1/2},$$

where $C_{13,L}$ is a positive, absolute constant.

Lemma 14. *Let τ be any fixed positive. For each coordinate $k\in[2,d]$, we have*

$$\mathbb{P}\left(\mathscr{T}_{\mathcal{B}_o,k}\le T^o_{\eta,\tau}\right)\le 6(\tau+1)\delta. \tag{D.3}$$

With the above secondary lemmas at hand, we are now ready for the proof of Lemma 4.

Proof of Lemma 4. Recall the definition of stopping time $\mathscr{T}_{w,k}$ in (3.10) and $\mathscr{T}_{\mathcal{B}_o,k}$ in (D.2). Since scaling condition (3.2) implies stepsize $\eta\le 1/(2\mathcal{B}_o^4 d^{1/2})$, we could apply Lemma 8 with $\mathcal{B}=\mathcal{B}_o$. On the event $(t\le T^o_{\eta,\tau}\wedge\mathscr{T}_{w,k})\cap(\mathscr{T}_{\mathcal{B}_o,k}>T^o_{\eta,\tau})$, we have $v_1^{-2}\le d$ and hence

$$\left|Q^{(t)}_{U,k}\right| = 4\mathcal{B}_o^8\eta^2 v_1^{-2}\le 4B^8 d\eta^2\log^4(T^o_{\eta,1}\delta^{-1}).$$

We define event $\mathscr{H}_{k;4,L}:=(\mathscr{T}_{\mathcal{B}_o,k}>T^o_{\eta,\tau})\cap\mathscr{H}_{k;13,L}$, then on the event $\mathscr{H}_{k;4,L}$, by applying Lemma 10 and 13, for all $t\le T^o_{\eta,\tau}\wedge\mathscr{T}_{w,k}$ we have

$$\begin{aligned}&\left|U_k^{(t)}-U_k^{(0)}+\eta|\mu_4-3|\sum_{s=0}^{t-1}\left((v_1^{(s)})^2-(v_k^{(s)})^2\right)U_k^{(s)}\right|\\ \le& T^o_{\eta,\tau}\cdot 4B^8 d\eta^2\log^4(T^o_{\eta,1}\delta^{-1})+C_{11,L}\log^{5/2}\delta^{-1}\cdot B^4\cdot d^{1/2}\eta(T^o_{\eta,\tau})^{1/2}.\end{aligned} \tag{D.4}$$

Scaling condition (3.2) in Lemma 3 implies that

$$B^4 d^{1/2}\eta(T^o_{\eta,\tau})^{1/2}\log^4(T^o_{\eta,1}\delta^{-1})\le\log^{5/2}\delta^{-1}. \tag{D.5}$$

Together with (D.4), on the event $\mathscr{H}_{k;4,L}$ we have

$$\left|U_k^{(t)}-U_k^{(0)}+\eta|\mu_4-3|\sum_{s=0}^{t-1}\left((v_1^{(s)})^2-(v_k^{(s)})^2\right)U_k^{(s)}\right|\le C_{4,L}\log^{5/2}\delta^{-1}\cdot B^4\cdot d^{1/2}\eta(T^o_{\eta,\tau})^{1/2},$$

where positive constant $C_{4,L}=C_{11,L}+4$. Scaling condition (3.2) implies $\eta|\mu_4-3|((v_1^{(s)})^2-(v_k^{(s)})^2)\in[0,1)$ for all $s<T^o_{\eta,\tau}\wedge\mathscr{T}_{w,k}$. From the reversed Gronwall Lemma 20, on the event $\mathscr{H}_{k;4,L}$ the following holds for all $t\le T^o_{\eta,\tau}\wedge\mathscr{T}_{w,k}$

$$\left|U_k^{(t)}-U_k^{(0)}\prod_{s=0}^{t-1}\left[1-\eta|\mu_4-3|\left((v_1^{(s)})^2-(v_k^{(s)})^2\right)\right]\right|\le 2C_{4,L}\log^{5/2}\delta^{-1}\cdot B^4\cdot d^{1/2}\eta(T^o_{\eta,\tau})^{1/2}.$$

To obtain a lower bound on the probability of event $\mathscr{H}_{k;4,L}$, we combine Lemmas 13, 14 and take union bound,

$$\mathbb{P}(\mathscr{H}_{k;4,L}) \geq 1 - \left(6\tau + 12 + \frac{5184}{\log^5 \delta^{-1}}\right)\delta$$

□

D.2 Proof of Lemma 5

Lemma 5 provides a quantitative characterization of the uniform initialization on $\mathcal{D}_1$. As a probabilistic fact, the uniform distribution $\mathbf{v}^{(0)}$ in $\mathcal{D}_1$ is equal in distribution to $\|\chi\|^{-1}\chi$ with $\chi \sim N(0,\mathbf{I})$. In addition, we have

$$\min_{2\leq k\leq d} W_k^{(0)} \geq \min_{2\leq k\leq d} \log\left(\frac{(v_1^{(0)})^2}{(v_k^{(0)})^2}\right) = \log(v_1^{(0)})^2 - \max_{2\leq k\leq d} \log(v_k^{(0)})^2, \tag{D.6}$$

due to definition of $W_k^{(t)}$ in (3.12) and the elementary inequality $x-1 \geq \log x$ for all $x > 0$. Intuitively, this means that $\min_{2\leq k\leq d} W_k^{(0)}$ is lower bounded by the spacing between the largest and second largest order statistics of d i.i.d. logarithmic chi-squared distributions.

We provide an elementary probabilistic Lemma 15 to show the spacing on the right hand of (D.6) $W_k^{(0)} = \Omega(\log^{-1} d)$ with high probability. To our best knowledge, there is no existing literature characterizing such bound.

Lemma 15. *Let χ_i^2 $(1 \leq i \leq n)$ be squares of i.i.d. standard normal variables, and denote their order statistics as $\chi_{(1)}^2 \leq \cdots \leq \chi_{(n)}^2$. Then for any $\varepsilon \in (0,1/3)$, when $n \geq 2\sqrt{2\pi e}\log\varepsilon^{-1} + 1$, we have with probability at least $1-3\varepsilon$*

$$\log\chi_{(n)}^2 - \log\chi_{(n-1)}^2 \geq \frac{\varepsilon}{8\log\varepsilon^{-1}\log n}. \tag{D.7}$$

We can apply Lemma 15 and prove Lemma 5.

Proof of Lemma 5. Following assumptions in Theorem 3, we know that $\mathbf{v}^{(0)}$ has the same distribution as $\|\chi\|^{-1}\chi$ where $\chi \sim N(0,\mathbf{I})$. Under scaling condition (3.2) in Lemma 3, Lemma 5 is straightforward if we apply Lemma 15 with $n = d$ and use (D.6). □

D.3 Proof of Lemma 6

Let $k \in [2,d]$ be any fixed coordinate. For each $t \geq 1$, we define random variable

$$Q_{W,k}^{(t)} = W_k^{(t)} - W_k^{(t-1)} + \mathrm{sign}(\mu_4 - 3)\eta \cdot 2(\mathbf{v}^\top \boldsymbol{Y})^3 v_1 v_k^{-3}(v_1 Y_k - v_k Y_1). \tag{D.8}$$

Lemma 16. *For each coordinate $k \in [2,d]$ and any $t \geq 1$, on the event*

$$\mathscr{H}_{k;16,L}^{(t)} \equiv \left(|\mathbf{v}^\top \boldsymbol{Y}| \leq \mathcal{B}_o,\ |Y_1| \leq \mathcal{B}_o,\ |Y_k| \leq \mathcal{B}_o,\ v_1^2 < 3v_k^2,\ v_1^2 \geq \max_{2\leq i\leq d} v_i^2\right), \tag{D.9}$$

under condition

$$12\mathcal{B}^8 d\eta^2 \log^4(T^o_{\eta,1}\delta^{-1}) \le 1, \tag{D.10}$$

we have

$$|Q^{(t)}_{W,k}| \le C_{16,L}\mathcal{B}^8_o d\eta^2,$$

where $C_{16,L}$ is a positive, absolute constant.

Lemma 17. *Let Assumption 1 hold. For each coordinate $k \in [2,d]$ and any $t \ge 1$, we have*

$$\mathbb{E}\left[\operatorname{sign}(\mu_4-3)\eta \cdot 2(\mathbf{v}^\top \boldsymbol{Y})^3 v_1 v_k^{-3}(v_1 Y_k - v_k Y_1)\middle| \mathscr{F}_{t-1}\right] = -2\eta|\mu_4-3|v_1^2 W_k. \tag{D.11}$$

For each $k \in [2,d]$ and $t \ge 1$, at the t-th iterate we let

$$f_k^{(t)} = -\operatorname{sign}(\mu_4-3)\eta \cdot 2(\mathbf{v}^\top \boldsymbol{Y})^3 v_1 v_k^{-3}(v_1 Y_k - v_k Y_1) - 2\eta|\mu_4-3|v_1^2 W_k. \tag{D.12}$$

which, indexed by t, forms a sequence of martingale differences with respect to $\mathscr{F}_{t-1}$. By combining (D.8) and (D.11) together, we have

$$W_k^{(t)} = \left(1 + 2\eta|\mu_4-3|(v_1^{(t-1)})^2\right) W_k^{(t-1)} + Q^{(t)}_{W,k} + f_k^{(t)}. \tag{D.13}$$

By letting

$$P_t^o = \prod_{s=0}^{t-1}\left(1 + 2\eta|\mu_4-3|(v_1^{(s)})^2\right)^{-1} \in \mathscr{F}_{t-1}, \tag{D.14}$$

we conclude the following lemma.

Lemma 18. *For each coordinate $k \in [2,d]$ and any $t \ge 1$, the iteration $W_k^{(t)}$ can be represented linearly as*

$$P_t^o W_k^{(t)} = W_k^{(0)} + \sum_{s=1}^{t} P_s^o Q^{(s)}_{W,k} + \sum_{s=1}^{t} P_s^o f_k^{(s)}. \tag{D.15}$$

Lemma 19. *Let Assumption 1 hold and initialization $\mathbf{v}^{(0)} \in \mathcal{D}_{cold} \cap \mathcal{D}^c_{mid,k}$. Let δ be a fixed positive. For each coordinate $k \in [2,d]$, there exists an event $\mathscr{H}_{k;19,L}$ satisfying*

$$\mathbb{P}(\mathscr{H}_{k;19,L}) \ge 1 - \left(6 + \frac{5184}{\log^5 \delta^{-1}}\right)\delta$$

such that on event $\mathscr{H}_{k;19,L}$ the following concentration result holds

$$\max_{1\le t\le T^o_{\eta,0.5}\wedge\mathscr{T}_{c,k}\wedge\mathscr{T}_1}\left|\sum_{s=1}^{t} P_s^o f_k^{(s)}\right| \le C_{19,L}\log^{5/2}\delta^{-1}\cdot\frac{B^4}{|\mu_4-3|^{1/2}}\cdot d\eta^{1/2}$$

where $C_{19,L}$ is a positive, absolute constant.

Along with the tail probability Lemma 14, we are ready to present the proof of Lemma 6.

Proof of Lemma 6. Recall the definition of truncation barrier $\mathcal{B}_o$ in (D.1), stopping times $\mathscr{T}_{\mathcal{B}_o,k}$ in (D.2), $\mathscr{T}_{c,k}$ in (3.15) and $\mathscr{T}_1$ in (3.16). We notice that (D.10) holds under scaling condition (3.2), and event $(t \le T^o_{\eta,0.5} \wedge \mathscr{T}_{c,k} \wedge \mathscr{T}_1) \cap (\mathscr{T}_{\mathcal{B}_o,k} > T^o_{\eta,0.5}) \subseteq \mathscr{H}^{(t)}_{k;16,L}$. We define event $\mathscr{H}_{k;6,L} \equiv (\mathscr{T}_{\mathcal{B}_o,k} > T^o_{\eta,0.5}) \cap \mathscr{H}_{k;19,L}$, then on event $\mathscr{H}_{k;6,L}$, by applying Lemma 16 for all $t \le T^o_{\eta,0.5} \wedge \mathscr{T}_{c,k} \wedge \mathscr{T}_1$ we obtain

$$v_1^2 \ge \frac{1}{d}, \quad |Q^{(t)}_{W,k}| \le C_{16,L} B^8 d\eta^2 \log^4(T^o_{\eta,1}\delta^{-1}).$$

Scaling condition (3.2) also guarantees

$$\frac{2|\mu_4-3|}{d}\eta < 1, \quad \frac{B^4}{|\mu_4-3|^{1/2}} \cdot d\eta^{1/2} \log^4(T^o_{\eta,1}\delta^{-1}) \le \log^{5/2}\delta^{-1}, \tag{D.16}$$

and hence by summation of geometric series, on event $\mathscr{H}_{k;6,L}$ we have for all $t \le T^o_{\eta,0.5} \wedge \mathscr{T}_{c,k} \wedge \mathscr{T}_1$

$$\begin{aligned}\left|\sum_{s=1}^t P^o_s Q^{(s)}_{W,k}\right| &\le \sum_{s=1}^t \left(1+\frac{2|\mu_4-3|}{d}\eta\right)^{-s} \cdot |Q^{(s)}_{W,k}| \le \frac{\left(1+\frac{2|\mu_4-3|}{d}\eta\right)^{-1}}{1-\left(1+\frac{2|\mu_4-3|}{d}\eta\right)^{-1}} \cdot C_{16,L} B^8 d\eta^2 \log^4(T^o_{\eta,1}\delta^{-1}) \\ &= \frac{C_{16,L}}{2} \cdot \frac{B^8}{|\mu_4-3|} \cdot d^2\eta \log^4(T^o_{\eta,1}\delta^{-1}) \le \frac{C_{16,L}}{2} \log^{5/2}\delta^{-1} \cdot \frac{B^4}{|\mu_4-3|^{1/2}} \cdot d\eta^{1/2}.\end{aligned}$$

Combining with Lemmas 18 and 19, on event $\mathscr{H}_{k;6,L}$ we know that for all $t \le T^o_{\eta,0.5} \wedge \mathscr{T}_{c,k} \wedge \mathscr{T}_1$ the following holds for constant $C_{6,L} \equiv C_{19,L} + C_{16,L}/2$

$$\left|W^{(t)}_k \prod_{s=0}^{t-1}\left(1+\eta|\mu_4-3|(v^{(s)}_1)^2\right)^{-1} - W^{(0)}_k\right| \le C_{6,L}\log^{5/2}\delta^{-1} \cdot \frac{B^4}{|\mu_4-3|^{1/2}} \cdot d\eta^{1/2}.$$

We verify the remaining claims in Lemma 6 by applying Lemma 14 with $\tau = 0.5$, Lemma 19 and taking union bound

$$\mathbb{P}(\mathscr{H}_{k;6,L}) \ge 1 - \mathbb{P}(\mathscr{T}_{\mathcal{B}_o,k} > T^o_{\eta,0.5}) - \mathbb{P}(\mathscr{H}^c_{k;19,L}) \ge 1 - \left(15 + \frac{5184}{\log^5\delta^{-1}}\right)\delta$$

. □

D.4 Proof of Secondary Lemmas

Proof of Lemma 13. $\{e^{(t)}_k\}_{t\ge 1}$ earlier defined in (C.7) forms a martingale difference sequence with respect to $\mathscr{F}_{t-1}$. Recall (C.15) in proof of Lemma 11, we have derived the following Orlicz $\psi_{1/2}$-norm bound under Assumption 1

$$\left\|e^{(t)}_k\right\|_{\psi_{1/2}} \le C'_{23,L}\eta B^4 \cdot |v_1|^{-1}(1+|v_k/v_1|).$$

Because initialization $\mathbf{v}^{(0)} \in \mathcal{D}_{\text{mid},k}$, on the event $(t \le \mathscr{T}_{w,k})$, where $\mathscr{T}_{w,k}$ is earlier defined in (3.10), we have $|v_1|^{-1} \le d^{1/2}$, $|v_k/v_1| \le 1/\sqrt{2}$, and then

$$\left\| e_k^{(t)} \right\|_{\psi_{1/2}} \le 2C'_{23,L} B^4 d^{1/2} \eta.$$

Because $1_{(t\le\mathscr{T}_{w,k})} \in \mathscr{F}_{t-1}$, we know that $\{e_k^{(t)} 1_{(t\le\mathscr{T}_{w,k})}\}$ forms a martingale difference sequence with respect to $\mathscr{F}_{t-1}$, and $\|e_k^{(t)} 1_{(t\le\mathscr{T}_{w,k})}\|_{\psi_{1/2}} \le \|e_k^{(t)}\|_{\psi_{1/2}} \le 2C'_{23,L} B^4 d^{1/2} \eta$. Hence we could apply Theorem 6 with $\alpha = 1/2$ and obtain for all $\delta \in (0, e^{-1}]$,

$$\begin{aligned}
&\mathbb{P}\left(\max_{1\le t\le T^o_{\eta,\tau}\wedge\mathscr{T}_{w,k}} \left| \sum_{s=1}^{t} e_k^{(s)} \right| \ge C_{13,L} B^4 d^{1/2} \eta (T^o_{\eta,\tau})^{1/2} \log^{5/2} \delta^{-1} \right) \\
&= \mathbb{P}\left(\max_{1\le t\le T^o_{\eta,\tau}} \left| \sum_{s=1}^{t} e_k^{(s)} 1_{(s\le\mathscr{T}_{w,k})} \right| \ge C_{13,L} B^4 d^{1/2} \eta (T^o_{\eta,\tau})^{1/2} \log^{5/2} \delta^{-1} \right) \\
&\le 2\left[3 + 6^4 \frac{64 \cdot T^o_{\eta,\tau} \cdot 4C'^2_{23,L} B^8 d\eta^2}{C^2_{13,L} B^8 d\eta^2 T^o_{\eta,\tau} \log^5 \delta^{-1}} \right] \exp\left\{ -\left(\frac{C^2_{13,L} B^8 d\eta^2 T^o_{\eta,\tau} \log^5 \delta^{-1}}{32 \cdot T^o_{\eta,\tau} \cdot 4C'^2_{23,L} B^8 d\eta^2} \right)^{\frac{1}{5}} \right\} \\
&= \left(6 + \frac{5184}{\log^5 \delta^{-1}} \right) \delta,
\end{aligned}$$

where constant $C_{13,L} \equiv 8\sqrt{2} C'_{23,L}$. □

Proof of Lemma 14. Applying Markov inequality and Lemma 7 gives

$$\mathbb{P}(|\mathbf{v}^\top \boldsymbol{Y}| > \mathcal{B}_o) = \mathbb{P}\left(\frac{|\mathbf{v}^\top \boldsymbol{Y}|^2}{B^2} > \frac{\mathcal{B}_o^2}{B^2} \right) \le \exp\left(-\frac{\mathcal{B}_o^2}{B^2} \right) \cdot \mathbb{E} \exp\left(\frac{|\mathbf{v}^\top \boldsymbol{Y}|^2}{B^2} \right) \le \frac{2\delta}{T^o_{\eta,1}}.$$

With the same procedure we also obtain

$$\mathbb{P}(|Y_1^{(t)}| > \mathcal{B}_o) \le \frac{2\delta}{T^o_{\eta,1}}, \quad \mathbb{P}(|Y_k^{(t)}| > \mathcal{B}_o) \le \frac{2\delta}{T^o_{\eta,1}}.$$

Taking union bound,

$$\begin{aligned}
\mathbb{P}\left(\mathscr{T}_{\mathcal{B}_o,k} \le T^o_{\eta,\tau} \right) &\le \sum_{t=1}^{T^o_{\eta,\tau}} \left(\mathbb{P}(|\mathbf{v}^{(t-1)\top} \boldsymbol{Y}^{(t)}| > \mathcal{B}_o) + \mathbb{P}(|Y_1^{(t)}| > \mathcal{B}_o) + \mathbb{P}(|Y_k^{(t)}| > \mathcal{B}_o) \right) \\
&\le 3T^o_{\eta,\tau} \cdot \frac{2\delta}{T^o_{\eta,1}} \le 6(\tau+1)\delta,
\end{aligned}$$

due to $T^o_{\eta,\tau}/T^o_{\eta,1} \le \tau + 1$, which comes from its definition in (3.1) and the elementary inequality $\lceil \tau x \rceil \le (\tau+1)\lceil x \rceil$ for all $\tau, x \ge 0$. □

Proof of Lemma 15. Let $F(x)$ be the cumulative distribution function of distribution $x \sim \log(\chi^2)$, where χ^2

denote the chi-squared distribution, then

$$F(x) = 2\Phi\left(e^{x/2}\right) - 1, \quad F'(x) = \phi\left(e^{x/2}\right)e^{x/2}. \tag{D.17}$$

where $\Phi(x)$ and $\phi(x)$ are cumulative distribution function and probability density function of standard Gaussian random variables. Let U_i $(1 \le i \le n)$ be i.i.d. samples from Uniform$(0,1)$. We denote order statistics $U_{(n)}, U_{(n-1)}$ as U, U_- for notational simplicity, and we define $M \equiv F^{-1}(U)$, $M_- \equiv F^{-1}(U_-)$. When $x \ge 0$, $F(x)$ is concave. Therefore, under condition

$$M_- \ge 0, \quad \text{i.e.} \quad U_- \ge 2\Phi(1) - 1, \tag{D.18}$$

by concavity we have

$$U - U_- = F(M) - F(M_-) \le F'(M_-)(M - M_-). \tag{D.19}$$

We denote the inverse functions of F, Φ by F^{-1}, Φ^{-1}, then from (D.17) we have

$$F^{-1}(y) = 2\log\Phi^{-1}\left(\frac{y+1}{2}\right). \tag{D.20}$$

(i) By inverse function theorem,

$$[\Phi^{-1}]'(y) = \frac{1}{\phi(\Phi^{-1}(y))} = \sqrt{2\pi}\exp\left(\frac{\Phi^{-1}(y)^2}{2}\right).$$

Applying chain rule of derivative to (D.20), we have

$$[F^{-1}]'(y) = \frac{[\Phi^{-1}]'\left(\frac{y+1}{2}\right)}{\Phi^{-1}\left(\frac{y+1}{2}\right)} = \frac{\sqrt{2\pi}\exp\left(\Phi^{-1}\left(\frac{y+1}{2}\right)^2/2\right)}{\Phi^{-1}\left(\frac{y+1}{2}\right)}. \tag{D.21}$$

(ii) The following bound on tail probability of standard gaussian random variable is folklore (see e.g. Durrett (2010))

$$\frac{x}{x^2+1}\cdot\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right) \le 1 - \Phi(x) \le \frac{1}{x}\cdot\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right). \tag{D.22}$$

We define

$$z_1 \equiv \frac{x^2+1}{x}\exp\left(\frac{x^2}{2}\right), \quad z_2 \equiv x\exp\left(\frac{x^2}{2}\right),$$

then for all $x \ge 1$,

- Because $z_1 \ge \exp(x^2/2)$, i.e. $x \le \sqrt{2\log z_1}$, we have

$$\exp\left(\frac{x^2}{2}\right) \ge \frac{z_1}{2x} \ge \frac{z_1}{2\sqrt{2\log z_1}},$$

which implies that

$$x \geq \sqrt{2\log z_1 - \log\log z_1 - 3\log 2}. \tag{D.23}$$

- Because $z_2 \geq \exp(x^2/2)$, we have

$$x \leq \sqrt{2\log z_2}. \tag{D.24}$$

For all $y \in [1-1/\sqrt{2\pi e}, 1)$,

- We can find solution $x \geq 1$ to

$$\frac{y+1}{2} = 1 - \frac{1}{\sqrt{2\pi} z_1}.$$

By applying (D.22) and (D.23) with this solution x, we obtain

$$\Phi^{-1}\left(\frac{y+1}{2}\right) \geq x \geq \sqrt{2\log\frac{\sqrt{2}}{\sqrt{\pi}(1-y)} - \log\log\frac{\sqrt{2}}{\sqrt{\pi}(1-y)} - 3\log 2}. \tag{D.25}$$

- We can find solution $x \geq 1$ to

$$\frac{y+1}{2} = 1 - \frac{1}{\sqrt{2\pi} z_2}.$$

By applying (D.22) and (D.24) with this solution x, we obtain

$$\Phi^{-1}\left(\frac{y+1}{2}\right) \leq x \leq \sqrt{2\log\frac{\sqrt{2}}{\sqrt{\pi}(1-y)}}. \tag{D.26}$$

By applying the lower bound (D.25) and upper bound (D.26) of $\Phi^{-1}\left(\frac{y+1}{2}\right)$ to (D.21), we have for all $y \in [1-1/\sqrt{2\pi e}, 1)$

$$[F^{-1}]'(y) \geq \sqrt{2\pi} \cdot \frac{1}{\sqrt{2\log\frac{\sqrt{2}}{\sqrt{\pi}(1-y)}}} \cdot \frac{\frac{\sqrt{2}}{\sqrt{\pi}(1-y)}}{\sqrt{8\log\frac{\sqrt{2}}{\sqrt{\pi}(1-y)}}} \geq \frac{1}{2(1-y)\log(1-y)^{-1}}.$$

Replace y with U_- and $F^{-1}(y)$ with M_-, then under condition

$$U_- \geq 1 - \frac{1}{\sqrt{2\pi e}} \tag{D.27}$$

we have

$$F'(M_-) = \frac{1}{[F^{-1}]'(U_-)} \leq 2(1-U_-)\log(1-U_-)^{-1} \tag{D.28}$$

(iii) From Arnold et al. (1992), we know that the spacing of order statistics $U - U_- \sim \text{Beta}(1,n)$ and $1 - U_- \sim \text{Beta}(2, n-1)$. Therefore, for any $\varepsilon \in (0, 1/3]$ and $n \geq 2$,

$$\mathbb{P}\left(U - U_- \leq \frac{\varepsilon}{n}\right) \leq \int_0^{\frac{\varepsilon}{n}} n(1-x)^{n-1}\mathrm{d}x = 1 - \left(1 - \frac{\varepsilon}{n}\right)^n \leq 1 - \left(\frac{1}{3}\right)^{\varepsilon} \leq (\log 3)\varepsilon$$

where we used elementary inequalities $(1-1/x)^x \geq 1/3$ when $x \geq 6$, and $1-3^{-x} \leq (\log 3)x$ for all $x \in (0,1/3]$.

In addition, for any $\varepsilon \in (0,1/3]$ and $n \geq 2\sqrt{2\pi e}\log\varepsilon^{-1}+1$ we have

$$\begin{aligned}
\mathbb{P}\left(1-U_- \geq \frac{2\log\varepsilon^{-1}}{n-1}\right) &= \int_{\frac{2\log\varepsilon^{-1}}{n-1}}^1 n(n-1)x(1-x)^{n-2}\mathrm{d}x \\
&= 2\log\varepsilon^{-1}\frac{n}{n-1}\left(1-\frac{2\log\varepsilon^{-1}}{n-1}\right)^{n-1} + \left(1-\frac{2\log\varepsilon^{-1}}{n-1}\right)^n \\
&\leq \left(2\log\varepsilon^{-1}+1\right)\cdot\left(1-\frac{2\log\varepsilon^{-1}}{n-1}\right)^{n-1} \\
&\leq (2\log\varepsilon^{-1}+1)\cdot\varepsilon^2 \leq \frac{2\log 3+1}{3}\varepsilon,
\end{aligned}$$

where we used elementary inequalities $(1-1/x)^x \leq 1/e$ for all $x \geq \sqrt{2\pi e}$, and $(2\log\varepsilon^{-1}+1)\varepsilon \leq (2\log 3+1)/3$ for all $\varepsilon \in (0,1/3]$. Taking union bound, we have

$$\mathbb{P}\left(U-U_- \geq \frac{\varepsilon}{n},\ 1-U_- \leq \frac{2\log\varepsilon^{-1}}{n-1}\right) \geq 1-3\varepsilon. \tag{D.29}$$

(iv) Notice that when $\varepsilon \in (0,1/3]$ and $n \geq 2\sqrt{2\pi e}\log\varepsilon^{-1}+1$, conditions (D.18) and (D.27) hold automatically when (D.29) holds. By combining (D.19), (D.28) and (D.29), for all $\varepsilon \in (0,1/3]$, $n \geq 2\sqrt{2\pi e}\log\varepsilon^{-1}+1$ we have with probability at least $1-3\varepsilon$ that

$$M-M_- \geq \frac{U-U_-}{F'(M_-)} \geq \frac{U-U_-}{2(1-U_-)\log(1-U_-)^{-1}} \geq \frac{(n-1)\varepsilon}{4n\log\varepsilon^{-1}\log\frac{n-1}{2\log\varepsilon^{-1}}} \geq \frac{\varepsilon}{8\log\varepsilon^{-1}\log n}$$

By noticing that the distributions of M and M_- are respectively equivalent to the distributions of $\log(\chi^2_{(n)})$ and $\log(\chi^2_{(n-1)})$, for all $\varepsilon \in (0,1/3]$ and $n \geq 2\sqrt{2\pi e}\log\varepsilon^{-1}+1$ we have with probability at least $1-3\varepsilon$ that (D.7) holds.

□

Proof of Lemma 16. Recall that we define $\eta_S = \eta\cdot\mathrm{sign}(\mu_4-3)$. For any fixed coordinate $k \in [2,d]$, we have

$$\begin{aligned}
W_k^{(t)} - W_k^{(t-1)} &= \frac{\left(v_1+\eta_S(\mathbf{v}^\top\boldsymbol{Y})^3Y_1\right)^2 - \left(v_k+\eta_S(\mathbf{v}^\top\boldsymbol{Y})^3Y_k\right)^2}{\left(v_k+\eta_S(\mathbf{v}^\top\boldsymbol{Y})^3Y_k\right)^2} - \frac{v_1^2-v_k^2}{v_k^2} \\
&= \frac{2\eta_S(\mathbf{v}^\top\boldsymbol{Y})^3v_1v_k(v_kY_1-v_1Y_k)+\eta_S^2(\mathbf{v}^\top\boldsymbol{Y})^6(v_k^2Y_1^2-v_1^2Y_k^2)}{v_k^2\left(v_k+\eta_S(\mathbf{v}^\top\boldsymbol{Y})^3Y_k\right)^2} \\
&= \eta_S\cdot\left(1+\eta_S(\mathbf{v}^\top\boldsymbol{Y})^3\frac{Y_k}{v_k}\right)^{-2}\cdot v_k^{-4}\left(2(\mathbf{v}^\top\boldsymbol{Y})^3v_1v_k(v_kY_1-v_1Y_k)+\eta_S(\mathbf{v}^\top\boldsymbol{Y})^6(v_k^2Y_1^2-v_1^2Y_k^2)\right).
\end{aligned}$$

Combining with (D.8), this implies

$$Q_{W,k}^{(t)} = 2\eta_S \cdot \left[\left(1+\eta_S(\mathbf{v}^\top \boldsymbol{Y})^3 \frac{Y_k}{v_k}\right)^{-2} - 1\right] \cdot (\mathbf{v}^\top \boldsymbol{Y})^3 v_1 v_k^{-2} \left(Y_1 - \frac{v_1}{v_k} Y_k\right)$$
$$+ \eta_S^2 \cdot \left(1+\eta_S(\mathbf{v}^\top \boldsymbol{Y})^3 \frac{Y_k}{v_k}\right)^{-2} \cdot (\mathbf{v}^\top \boldsymbol{Y})^6 v_k^{-2} \left(Y_1^2 - \frac{v_1^2}{v_k^2} Y_k^2\right).$$

Taylor series give for all $|x| \le \frac{1}{2}$ that

$$\left|(1+x)^{-2} - 1\right| = |x| \left|\sum_{i=0}^{\infty} (-1)^i (i+2) x^i\right| \le 6|x|, \quad (1+x)^{-2} \le 4.$$

On event $\mathscr{H}_{k;16,L}^{(t)}$, since $|v_1/v_k| < \sqrt{3}$ and $|v_k|^2 \ge 1/(3d)$, (D.10) implies $\left|\eta_S (\mathbf{v}^\top \boldsymbol{Y})^3 \frac{Y_k}{v_k}\right| \le \frac{1}{2}$, and hence we have

$$|Q_{W,k}^{(t)}| \le 2\eta \cdot 6\sqrt{3}\mathcal{B}_o^4 d^{1/2} \eta \cdot (3+3\sqrt{3})\mathcal{B}_o^4 d^{1/2} + \eta^2 \cdot 4 \cdot 12\mathcal{B}_o^8 d \le C_{16,L} \mathcal{B}_o^8 \eta^2 d$$

where constant $C_{16,L} = 156 + 36\sqrt{3}$. □

Proof of Lemma 17. Recall that $W_k^{(t-1)} = (v_1^2 - v_k^2)/v_k^2$. Under Assumption 1, using (C.13) we have

$$\begin{aligned}
&\mathbb{E}\left[\operatorname{sign}(\mu_4 - 3)\eta \cdot 2 v_1 v_k^{-3} (\mathbf{v}^\top \boldsymbol{Y})^3 (v_1 Y_k - v_k Y_1) \middle| \mathscr{F}_{t-1}\right] \\
&= 2\operatorname{sign}(\mu_4 - 3)\eta \cdot v_1 v_k^{-3} \left((\mu_4 - 3) v_1 v_k^3 + 3 v_1 v_k - (\mu_4 - 3) v_k v_1^3 - 3 v_k v_1\right) \\
&= -2\eta |\mu_4 - 3| \cdot v_1^2 W_k
\end{aligned}$$

□

Proof of Lemma 18. From (D.13) and (D.14), we have

$$P_s^o W_k^{(s)} - P_{s-1}^o W_k^{(s-1)} = P_s^o \left(Q_{W,k}^{(s)} + f_k^{(s)}\right) \tag{D.30}$$

We iteratively apply (D.30) for $s = 1, \ldots, t$ and obtain (D.15). □

Proof of Lemma 19. $\{f_k^{(t)}\}_{t \ge 1}$, earlier defined in (D.12), forms a martingale difference sequence with respect to $\mathscr{F}_{t-1}$. Similar to techniques in proof of Lemma 11, we apply Lemma 22 three times, then use

Lemma 21 and (C.12), in order to obtain the following bound on Orlicz $\psi_{1/2}$-norm

$$\begin{aligned}\left\|\eta\cdot 2(\mathbf{v}^\top \boldsymbol{Y})^3 v_1 v_k^{-3}(v_k Y_1 - v_1 Y_k)\right\|_{\psi_{1/2}} &\le 2\eta|v_1||v_k|^{-2}\cdot\left\|(\mathbf{v}^\top\boldsymbol{Y})^2\right\|_{\psi_1}\cdot\left\|(\mathbf{v}^\top\boldsymbol{Y})\left(Y_1-\frac{v_1}{v_k}Y_k\right)\right\|_{\psi_1}\\ &\le 2\eta|v_1||v_k|^{-2}\cdot\left\|\mathbf{v}^\top\boldsymbol{Y}\right\|_{\psi_2}^3\cdot\left\|Y_1-\frac{v_1}{v_k}Y_k\right\|_{\psi_2}\\ &\le 6\eta d^{1/2}\cdot\left\|\mathbf{v}^\top\boldsymbol{Y}\right\|_{\psi_2}^3\cdot(\|Y_1\|_{\psi_2}+\sqrt{3}\|Y_k\|)\\ &\le 6(1+\sqrt{3})B^4 d^{1/2}\eta\end{aligned}$$

where we use the fact that $|v_1/v_k|\le\sqrt{3}$ and $v_k^2\ge 1/(3d)$ on the event $(t\le\mathscr{T}_{c,k}\wedge\mathscr{T}_1)$. Then by applying Lemma 23 we further derive

$$\begin{aligned}\|f_k^{(t)}\|_{\psi_{1/2}} &= \left\|\eta\cdot 2(\mathbf{v}^\top\boldsymbol{Y})^3 v_1 v_k^{-3}(v_kY_1-v_1Y_k)-\mathbb{E}\left[\eta\cdot 2(\mathbf{v}^\top\boldsymbol{Y})^3 v_1 v_k^{-3}(v_kY_1-v_1Y_k)\right]\right\|_{\psi_{1/2}}\\ &\le C'_{23,L}\left\|\eta\cdot 2(\mathbf{v}^\top\boldsymbol{Y})^3 v_1 v_k^{-3}(v_kY_1-v_1Y_k)\right\|_{\psi_{1/2}}\le 6(1+\sqrt{3})C'_{23,L}B^4 d^{1/2}\eta\end{aligned}$$

Because $1_{(t\le\mathscr{T}_{c,k}\wedge\mathscr{T}_1)}\in\mathscr{F}_{t-1}$, we know that $\left\{f_k^{(t)}1_{(t\le\mathscr{T}_{c,k}\wedge\mathscr{T}_1)}\right\}$ forms a martingale difference sequence with respect to $\mathscr{F}_{t-1}$, and $\|f_k^{(t)}1_{(t\le\mathscr{T}_{c,k}\wedge\mathscr{T}_1)}\|_{\psi_{1/2}}\le\|f_k^{(t)}\|_{\psi_{1/2}}$. Since $v_1^2\ge 1/d$ on the event $(t\le\mathscr{T}_{c,k}\wedge\mathscr{T}_1)$ and $2\eta|\mu_4-3|/d<1$ under scaling condition (3.2), by summation of geometric series, for all $t\ge 1$ we have

$$\begin{aligned}\sum_{s=1}^t(P_s^o)^2\left\|f_k^{(s)}1_{(t\le\mathscr{T}_{c,k}\wedge\mathscr{T}_1)}\right\|_{\psi_{1/2}}^2 &\le \sum_{s=1}^t\left(1+\frac{2\eta|\mu_4-3|}{d}\right)^{-2s}\cdot 36(1+\sqrt{3})^2C'^2_{23,L}B^8 d\eta^2\\ &\le \frac{\left(1+\frac{2\eta|\mu_4-3|}{d}\right)^{-2}}{1-\left(1+\frac{2\eta|\mu_4-3|}{d}\right)^{-2}}\cdot 36(1+\sqrt{3})^2C'^2_{23,L}B^8 d\eta^2\\ &= 18(1+\sqrt{3})^2C'^2_{23,L}\cdot\frac{B^8}{|\mu_4-3|}\cdot d^2\eta\end{aligned}$$

With the bound given above, we apply Theorem 6 with $\alpha=1/2$ as

$$\begin{aligned}&\mathbb{P}\left(\max_{1\le t\le T^o_{\eta,0.5}\wedge\mathscr{T}_{c,k}\wedge\mathscr{T}_1}\left|\sum_{s=1}^t P_s^o f_k^{(s)}\right|\ge C_{19,L}\log^{5/2}\delta^{-1}\cdot\frac{B^4}{|\mu_4-3|^{1/2}}\cdot d\eta^{1/2}\right)\\ &=\mathbb{P}\left(\max_{1\le t\le T^o_{\eta,0.5}}\left|\sum_{s=1}^t P_s^o f_k^{(s)}1_{(s\le\mathscr{T}_{c,k}\wedge\mathscr{T}_1)}\right|\ge C_{19,L}\log^{5/2}\delta^{-1}\cdot\frac{B^4}{|\mu_4-3|^{1/2}}\cdot d\eta^{1/2}\right)\\ &\le 2\left[3+6^4\frac{64\cdot 18(1+\sqrt{3})^2C'^2_{23,L}B^8|\mu_4-3|^{-1}d^2\eta}{C^2_{19,L}\log^5\delta^{-1}B^8|\mu_4-3|^{-1}d^2\eta}\right]\exp\left\{-\left(\frac{C^2_{19,L}\log^5\delta^{-1}B^8|\mu_4-3|^{-1}d^2\eta}{32\cdot 18(1+\sqrt{3})^2C'^2_{23,L}B^8|\mu_4-3|^{-1}d^2\eta}\right)^{\frac{1}{5}}\right\}\\ &=\left(6+\frac{5184}{\log^5\delta^{-1}}\right)\delta\end{aligned}$$

where constant $C_{19,L} = 24(1+\sqrt{3})C'_{23,L}$. □

E A Reversed Gronwall's Inequality

We present a discrete generalization of Gronwall's inequality, which is sharper than a straightforward application of Gronwall's inequality. Such sharp estimation plays a key role in our analysis. Although elementary, the lemma seems not recorded in relevant literatures:

Lemma 20. *If for all $t = 1, \dots, T$, $\beta(t) \in [0,1)$ and if $u(t)$ satisfies that for some positive constant α*

$$\left| u(t) - u(0) + \sum_{0 \le s < t} \beta(s) u(s) \right| \le \alpha \tag{E.1}$$

then for all $t = 1, \dots, T$

$$\left| u(t) - u(0) \prod_{0 \le s < t} (1 - \beta(s)) \right| \le 2\alpha - \alpha \prod_{0 \le s < t} (1 - \beta(s)) \le 2\alpha. \tag{E.2}$$

Note unlike analogous Gronwall-type results, here we pose *no* assumption on the sign of $u(t)$.

Proof of Lemma 20. Define for all $t = 1, \dots, T$

$$v(t) = \prod_{0 \le s < t} (1 - \beta(s))^{-1} \left(u(0) - \sum_{0 \le s < t} \beta(s) u(s) \right). \tag{E.3}$$

We have for $s = 0, \dots, t-1$

$$\begin{aligned}
v(s+1) - v(s) &=: A_{s+1}(B_{s+1} - B_s) + B_s(A_{s+1} - A_s) \\
&= - \prod_{0 \le r < s+1} (1 - \beta(r))^{-1} \cdot \beta(s) u(s) \\
&\quad + \left(u(0) - \sum_{0 \le r < s} \beta(r) u(r) \right) \left(\prod_{0 \le r < s+1} (1 - \beta(r))^{-1} - \prod_{0 \le r < s} (1 - \beta(r))^{-1} \right) \\
&= - \prod_{0 \le r < s+1} (1 - \beta(r))^{-1} \cdot \beta(s) u(s) + \left(u(0) - \sum_{0 \le r < s} \beta(r) u(r) \right) \cdot \left(\prod_{0 \le r < s+1} (1 - \beta(r))^{-1} \cdot \beta(s) \right) \\
&= - \left(u(s) - u(0) + \sum_{0 \le r < s} \beta(r) u(r) \right) \cdot \beta(s) \prod_{0 \le r < s+1} (1 - \beta(r))^{-1}.
\end{aligned}$$

Since $\beta(s) \in [0,1)$, and the product is nonnegative, the use of the *lower side* of (E.1) upper-estimates the difference of $v(s)$

$$v(s+1) - v(s) \le \alpha \beta(s) \prod_{0 \le r < s+1} (1 - \beta(r))^{-1} \tag{E.4}$$

Since $v(0) = u(0)$, telescoping the above inequality for $s = 0, \ldots, t-1$ gives

$$v(t) = v(0) + \sum_{0 \le s < t} v(s+1) - v(s) \le u(0) + \alpha \sum_{0 \le s < t} \beta(s) \prod_{0 \le r < s+1} (1 - \beta(r))^{-1}$$

and hence from the definition of $v(t)$ in (E.3)

$$\begin{aligned} u(0) - \sum_{0 \le s < t} \beta(s) u(s) &= \prod_{0 \le s < t} (1 - \beta(s))\, v(t) \\ &\le \prod_{0 \le s < t} (1 - \beta(s)) \left(u(0) + \alpha \sum_{0 \le s < t} \beta(s) \prod_{0 \le r < s+1} (1 - \beta(r))^{-1} \right) \\ &= u(0) \prod_{0 \le s < t} (1 - \beta(s)) + \alpha \sum_{0 \le s < t} \beta(s) \prod_{s+1 \le r < t} (1 - \beta(r)). \end{aligned}$$

Taking the above result into the *upper side* of (E.1) gives

$$u(t) \le \alpha + u(0) - \sum_{0 \le s < t} \beta(s) u(s) \le \alpha + u(0) \prod_{0 \le s < t} (1 - \beta(s)) + \alpha \sum_{0 \le s < t} \beta(s) \prod_{s+1 \le r < t} (1 - \beta(r)),$$

which further reduces to

$$u(t) - u(0) \prod_{0 \le s < t} (1 - \beta(s)) \le \alpha + \alpha \sum_{0 \le s < t} \beta(s) \prod_{s+1 \le r < t} (1 - \beta(r)).$$

That is, for all $t = 0, 1, \ldots, T$,

$$\begin{aligned} u(t) - u(0) \prod_{0 \le s < t} (1 - \beta(s)) &\le \alpha + \alpha \sum_{0 \le s < t} \beta(s) \prod_{s+1 \le r < t} (1 - \beta(r)) \\ &= \alpha + \alpha \sum_{0 \le s < t} (1 - (1 - \beta(s))) \prod_{s+1 \le r < t} (1 - \beta(r)) \\ &= \alpha + \alpha \sum_{0 \le s < t} \left[\prod_{s+1 \le r < t} (1 - \beta(r)) - \prod_{s \le r < t} (1 - \beta(r)) \right] \\ &= 2\alpha - \alpha \prod_{0 \le r < t} (1 - \beta(r)) \\ &\le 2\alpha, \end{aligned}$$

which proves the upper side of (E.2). For the lower side, applying the same inequality to $-u$ in the place of u gives the desired result. □

F Orlicz norm and Properties

Definition 1 (Orlicz ψ_α-norm)**.** *For a continuous, monotonically increasing and convex function $\psi(x)$ defined for all $x > 0$ satisfying $\psi(0) = 0$ and $\lim_{x \to \infty} \psi(x) = \infty$, we define the Orlicz ψ-norm for a random*

variable X as

$$\|X\|_\psi \equiv \inf\{K > 0 : \mathbb{E}\psi(|X/K|) \leq 1\}$$

As a commonly used special case, we consider function $\psi_\alpha(x) \equiv \exp(x^\alpha) - 1$ and define the Orlicz ψ_α-norm for a random variable X as

$$\|X\|_{\psi_\alpha} \equiv \inf\left\{K > 0 : \mathbb{E}\exp\left(\frac{|X|^\alpha}{K^\alpha}\right) \leq 2\right\}$$

In the case of $0 < \alpha < 1$, $\psi_\alpha(x)$ is not convex in a neignborhood of 0. We choose to keep this definition and call it ψ_α-norm, but it is actually not a norm and does not satisfy the triangle inequality given in Lemma 21. In Lemma 23, we will find that, when $\alpha = 1/2$, it follows a generalized triangle inequality.

Lemma 21 (Triangle inequality)**.** *When $\psi(x)$ is monotonically increasing and convex for $x > 0$, the Orlicz ψ-norm satisfies triangle inequality, i.e. for any random variables with $\|X\|_\psi < \infty, \|Y\|_\psi < \infty$, we have*

$$\|X + Y\|_\psi \leq \|X\|_\psi + \|Y\|_\psi$$

Consequently, when $\alpha \geq 1$, $\psi_\alpha(x)$ is monotonically increasing and convex for $x > 0$, and therefore ψ_α-norm satisifies triangle inequality, i.e.

$$\|X + Y\|_{\psi_\alpha} \leq \|X\|_{\psi_\alpha} + \|Y\|_{\psi_\alpha}$$

We state the multiplicative property of the Orlicz ψ_α-norm, which serves as an extension of Proposition D.3 in Vu & Lei (2013):

Lemma 22 (Multiplicative property)**.** *Let X and Y be random variables with finite ψ_α-norm for some $\alpha \geq 1$, then*

$$\|XY\|_{\psi_{\alpha/2}} \leq \|X\|_{\psi_\alpha}\|Y\|_{\psi_\alpha}$$

Lemma 23. *For any random variables X, Y with $\|X\|_{\psi_{1/2}} < \infty$ and $\|Y\|_{\psi_{1/2}} < \infty$, we have the following inequalities for Orlicz $\psi_{1/2}$-norm*

$$\|X + Y\|_{\psi_{1/2}} \leq C_{23,L}(\|X\|_{\psi_{1/2}} + \|Y\|_{\psi_{1/2}}) \quad \text{and} \quad \|\mathbb{E}X\|_{\psi_{1/2}} \leq C_{23,L}\|X\|_{\psi_{1/2}}$$

where positive constant $C_{23,L} \equiv 1.3937$. In addition, we have

$$\|X - \mathbb{E}X\|_{\psi_{1/2}} \leq C'_{23,L}\|X\|_{\psi_{1/2}}$$

where positive constant $C'_{23,L} \equiv 3.3359$.

F.1 Proof of Orlicz norm Properties

Proof of Lemma 21. We denote $K_1 \equiv \|X\|_\psi$ and $K_2 \equiv \|Y\|_\psi$. Because $\psi(x)$ is monotonically increasing and convex, we have

$$\psi\left(\left|\frac{X+Y}{K_1+K_2}\right|\right) \le \psi\left(\frac{K_1}{K_1+K_2}\cdot\left|\frac{X}{K_1}\right| + \frac{K_2}{K_1+K_2}\cdot\left|\frac{Y}{K_2}\right|\right) \le \frac{K_1}{K_1+K_2}\cdot\psi\left(\left|\frac{X}{K_1}\right|\right) + \frac{K_2}{K_1+K_2}\cdot\psi\left(\left|\frac{Y}{K_2}\right|\right),$$

which implies that

$$\mathbb{E}\psi\left(\left|\frac{X+Y}{K_1+K_2}\right|\right) \le 1, \quad \text{i.e.} \quad \|X+Y\|_\psi \le K_1+K_2 = \|X\|_\psi + \|Y\|_\psi.$$

□

Proof of Lemma 22. Denote $A \equiv X/\|X\|_{\psi_\alpha}$, $B \equiv Y/\|Y\|_{\psi_\alpha}$, then $\|A\|_{\psi_\alpha} = \|B\|_{\psi_\alpha} = 1$. Using the elementary inequality

$$|AB| \le \frac{1}{4}(|A|+|B|)^2$$

and triangle inequality in Lemma 21 we have that

$$\|AB\|_{\psi_{\alpha/2}} \le \frac{1}{4}\|(|A|+|B|)^2\|_{\psi_{\alpha/2}} = \frac{1}{4}\||A|+|B|\|^2_{\psi_\alpha} \le \frac{1}{4}(\|A\|_{\psi_\alpha}+\|B\|_{\psi_\alpha})^2 = 1$$

Multiplying both sides of the inequality by $\|X\|_{\psi_\alpha}\|Y\|_{\psi_\alpha}$ gives the desired result. □

Proof of Lemma 23. Recall that when $\alpha \in (0,1)$, $\psi_\alpha(x)$ does *not* satisfy convexity when x is around 0. Let $\tilde{\psi}_\alpha(x)$ be

$$\tilde{\psi}_\alpha(x) = \begin{cases} \exp(x^\alpha) - 1 & x \ge x_\alpha \\ \frac{x}{x_\alpha}(\exp(x_\alpha^\alpha) - 1) & x \in [0, x_\alpha) \end{cases}$$

for some appropriate $x_\alpha > 0$, so as to make the function convex. Here x_α is chosen such that the tangent line of function ψ_α at x_α passes through origin, i.e.

$$\alpha x_\alpha^{\alpha-1}\exp(x_\alpha^\alpha) = \frac{\exp(x_\alpha^\alpha)-1}{x_\alpha}$$

Simplifying it gives us a transcendental equation

$$(1-\alpha x_\alpha^\alpha)\exp(x_\alpha^\alpha) = 1$$

which does not possess an analytic solution, but can be solved numerically. When $\alpha = 1/2$, we have $x_{1/2} = 2.5396$. With numerical calculation, we also find that

$$0 \le \psi_{1/2}(x) - \tilde{\psi}_{1/2}(x) \le 0.2666 \tag{F.1}$$

As an application of (F.1), we have

$$\mathbb{E}\tilde{\psi}_{1/2}(|X|) \le 1 \quad \Longrightarrow \quad \mathbb{E}\psi_{1/2}(|X|) \le 1.2666 \quad \text{i.e. } \mathbb{E}\exp(|X|^{1/2}) \le 2.2666$$

(i) Let K_1, K_2 denote the $\psi_{1/2}$ norms of X and Y, then

$$\mathbb{E}\psi_{1/2}(|X/K_1|) \le 1 \quad \text{and} \quad \mathbb{E}\psi_{1/2}(|Y/K_2|) \le 1$$

Based on (F.1) we have

$$\mathbb{E}\tilde{\psi}_{1/2}(|X/K_1|) \le 1 \quad \text{and} \quad \mathbb{E}\tilde{\psi}_{1/2}(|Y/K_2|) \le 1$$

Then by applying triangle inequality from Lemma 21 to Orlicz $\tilde{\psi}_{1/2}$-norm,

$$\mathbb{E}\tilde{\psi}_{1/2}(|(X+Y)/(K_1+K_2)|) \le 1$$

Along with (F.1) we find

$$\mathbb{E}\psi_{1/2}(|(X+Y)/(K_1+K_2)|) \le 1.2666$$

By applying Jensen's inequality to concave function $f(z) = z^{\log_{2.2666} 2}$, for constant $C_{23,L} \equiv (\log_2 2.2666)^2 = 1.3937$, we have

$$\begin{aligned}
\mathbb{E}\psi_{1/2}(|(X+Y)/(C_{23,L}(K_1+K_2))|) &= \mathbb{E}\exp(|(X+Y)/(K_1+K_2)|^{1/2})^{\log_{2.2666} 2} - 1 \\
&\le \left(\mathbb{E}\exp(|(X+Y)/(K_1+K_2)|^{1/2})\right)^{\log_{2.2666} 2} - 1 \\
&\le 1
\end{aligned}$$

which implies that

$$\|X+Y\|_{\psi_{1/2}} \le C_{23,L}(\|X\|_{\psi_{1/2}} + \|Y\|_{\psi_{1/2}})$$

(ii) Let K denote the $\psi_{1/2}$ norm of X, then

$$\mathbb{E}\psi_{1/2}(|X/K|) \le 1$$

Based on (F.1) we have

$$\mathbb{E}\tilde{\psi}_{1/2}(|X/K|) \le 1$$

Because $\tilde{\psi}_{1/2}$ is a convex function, we can apply Jensen's inequality as

$$\tilde{\psi}_{1/2}(|\mathbb{E}X/K|) \le \tilde{\psi}_{1/2}(\mathbb{E}|X/K|) \le \mathbb{E}\tilde{\psi}_{1/2}(|X/K|) \le 1$$

Combining with (F.1),

$$\psi_{1/2}(|\mathbb{E}X/K|) \le 1.2666$$

which is equivalent to

$$\psi_{1/2}(|\mathbb{E}X/(C_{23,L}K)|) \le 1$$

where $C_{23,L} = (\log_2 2.2666)^2 = 1.3937$. By noticing that $\mathbb{E}\psi_{1/2}(\mathbb{E}X/(CK)) = \psi_{1/2}(\mathbb{E}X/(CK))$, we find that

$$\|\mathbb{E}X\|_{\psi_{1/2}} \le C_{23,L}\|X\|_{\psi_{1/2}}$$

(iii) As an application, for any random variable X with $\|X\|_{\psi_{1/2}} < \infty$, we have

$$\|X - \mathbb{E}X\|_{\psi_{1/2}} \le C_{23,L}(\|X\|_{\psi_{1/2}} + \|\mathbb{E}X\|_{\psi_{1/2}}) \le C_{23,L}(1 + C_{23,L})\|X\|_{\psi_{1/2}} = C'_{23,L}\|X\|_{\psi_{1/2}}$$

where positive constant $C'_{23,L} \equiv C_{23,L}(1 + C_{23,L}) = 3.3359$.

□

G A Useful Concentration Inequality

We prove the following concentration inequality for 1-dimensional supermartingale difference sequence with finite ψ_α-norms:

Theorem 6. *Let $\alpha \in (0,\infty)$ be given. Assume that $(u_i : i \ge 1)$ is a sequence of supermartingale differences with respect to $\mathscr{F}_i$, i.e. $\mathbb{E}[u_i \mid \mathscr{F}_{i-1}] \le 0$, and it satisfies $\|u_i\|_{\psi_\alpha} < \infty$ for each $i = 1,\ldots,N$. Then for an arbitrary $N \ge 1$ and $z > 0$,*

$$\mathbb{P}\left(\max_{1\le n\le N}\sum_{i=1}^{n} u_i \ge z\right) \le \left[3 + \left(\frac{3}{\alpha}\right)^{\frac{2}{\alpha}} \frac{64\sum_{i=1}^{N}\|u_i\|_{\psi_\alpha}^2}{z^2}\right] \exp\left\{-\left(\frac{z^2}{32\sum_{i=1}^{N}\|u_i\|_{\psi_\alpha}^2}\right)^{\frac{\alpha}{\alpha+2}}\right\} \tag{G.1}$$

Proof of Theorem 6. To prove Theorem 6, we will use a maxima version of the classical Azuma-Hoeffding's inequality proposed by Laib (1999) for bounded martingale differences, and then apply an argument of Lesigne & Volnỳ (2001) and Fan et al. (2012) to truncate the tail and analyze the bounded and unbounded pieces separately.

(i) First of all, for the sake of simplicity and with no loss of gen erality, throughout the following proof of Theorem 6 we shall pose the following extra condition

$$\sum_{i=1}^{N}\|u_i\|_{\psi_\alpha}^2 = 1. \tag{G.2}$$

In other words, under the additional (G.2) condition proving (G.1) reduces to showing

$$\mathbb{P}\left(\max_{1\le n\le N}\sum_{i=1}^{n} u_i \ge z\right) \le \left[3+\left(\frac{3}{\alpha}\right)^{\frac{2}{\alpha}}\frac{64}{z^2}\right]\exp\left\{-\left(\frac{z^2}{32}\right)^{\frac{\alpha}{\alpha+2}}\right\}. \tag{G.3}$$

This can be made more clear from the following rescaling argument: one can put in the left of (G.3) $u_i/\left(\sum_{i=1}^N \|u_i\|_{\psi_\alpha}^2\right)^{1/2}$ in the place of u_i, and $z/\left(\sum_{i=1}^N \|u_i\|_{\psi_\alpha}^2\right)^{1/2}$ in the place of z, the left hand of (G.1) is just

$$\mathbb{P}\left(\max_{1\le n\le N}\sum_{i=1}^{n}\frac{u_i}{\left(\sum_{i=1}^N \|u_i\|_{\psi_\alpha}^2\right)^{1/2}} \ge \frac{z}{\left(\sum_{i=1}^N \|u_i\|_{\psi_\alpha}^2\right)^{1/2}}\right)$$

which, by (G.3), is upper-bounded by

$$\le \left[3+\left(\frac{3}{\alpha}\right)^{\frac{2}{\alpha}}\frac{64\sum_{i=1}^N \|u_i\|_{\psi_\alpha}^2}{z^2}\right]\exp\left\{-\left(\frac{z^2}{32\sum_{i=1}^N \|u_i\|_{\psi_\alpha}^2}\right)^{\frac{\alpha}{\alpha+2}}\right\},$$

proving (G.1).

(ii) We apply a truncating argument used in Lesigne & Volnỳ (2001) and later in Fan et al. (2012). Let $\mathscr{M} > 0$ be arbitrary, and we define

$$u_i' = u_i 1_{\{|u_i|\le \mathscr{M}\|u_i\|_{\psi_\alpha}\}} - \mathbb{E}\left(u_i 1_{\{|u_i|\le \mathscr{M}\|u_i\|_{\psi_\alpha}\}} \mid \mathscr{F}_{i-1}\right), \tag{G.4}$$

$$u_i'' = u_i 1_{\{|u_i|> \mathscr{M}\|u_i\|_{\psi_\alpha}\}} - \mathbb{E}\left(u_i 1_{\{|u_i|> \mathscr{M}\|u_i\|_{\psi_\alpha}\}} \mid \mathscr{F}_{i-1}\right), \tag{G.5}$$

$$T_n' = \sum_{i=1}^n u_i', \quad T_n'' = \sum_{i=1}^n u_i'', \quad T_n''' = \sum_{i=1}^n \mathbb{E}\left(u_i \mid \mathscr{F}_{i-1}\right).$$

Since u_i is $\mathscr{F}_i$-measurable, u_i' and u_i'' are two martingale difference sequences with respect to $\mathscr{F}_i$, and let T_n be defined as

$$T_n = \sum_{i=1}^n u_i \quad \text{and hence} \quad T_n = T_n' + T_n'' + T_n'''. \tag{G.6}$$

Since u_i are supermartingale differences we have that T_n''' is $\mathscr{F}_{n-1}$-measurable with $T_n''' \le T_0''' = 0$, *a.s.*, and hence for any $z > 0$,

$$\begin{aligned}\mathbb{P}\left(\max_{1\le n\le N} T_n \ge 2z\right) &\le \mathbb{P}\left(\max_{1\le n\le N} T_n' + T_n''' \ge z\right) + \mathbb{P}\left(\max_{1\le n\le N} T_n'' \ge z\right)\\ &\le \mathbb{P}\left(\max_{1\le n\le N} T_n' \ge z\right) + \mathbb{P}\left(\max_{1\le n\le N} T_n'' \ge z\right)\end{aligned} \tag{G.7}$$

In the following, we analyze the tail bounds for T_n' and T_n'' separately (Lesigne & Volnỳ, 2001; Fan et al., 2012).

(iii) To obtain the first bound, we recap Laib's inequality as follows:

Lemma 24. *(Laib, 1999) Let $(w_i : 1 \le i \le N)$ be a real-valued martingale difference sequence with respect to some filtration $\mathscr{F}_i$, i.e. $\mathbb{E}[w_i \mid \mathscr{F}_{i-1}] = 0, a.s.$, and the essential norm $\|w_i\|_\infty$ is finite. Then for an arbitrary $N \ge 1$ and $z > 0$,*

$$\mathbb{P}\left(\max_{n\le N}\sum_{i=1}^{n} w_i \ge z\right) \le \exp\left\{-\frac{z^2}{2\sum_{i=1}^{N}\|w_i\|_\infty^2}\right\}. \tag{G.8}$$

(G.8) generalizes the folklore Azuma-Hoeffding's inequality, where the latter can be concluded from

$$\max_{n\le N}\sum_{i=1}^{n} w_i \ge \sum_{i=1}^{N} w_i.$$

The proof of Lemma 24 is given in Laib (1999).

Recall our extra condition (G.2), then from the definition of u_i' in (G.4) that $|u_i'| \le 2\mathscr{M}\|u_i\|_{\psi_\alpha}$, the desired bound follows immediately from Laib's inequality in Lemma 24 by setting $w_i = u_i'$:

$$\mathbb{P}\left(\max_{1\le n\le N} T_n' \ge z\right) = \mathbb{P}\left(\max_{1\le n\le N}\sum_{i=1}^{n} u_i' \ge z\right) \le \exp\left\{-\frac{z^2}{8\mathscr{M}^2}\right\} \tag{G.9}$$

To obtain the tail bound of T_n'' we only need to show

$$\mathbb{E}(u_i'')^2 \le (6\mathscr{M}^2 + 8\mathscr{B}^2)\|u_i\|_{\psi_\alpha}^2 \exp\{-\mathscr{M}^\alpha\}, \tag{G.10}$$

where

$$\mathscr{B} \equiv \left(\frac{3}{\alpha}\right)^{\frac{1}{\alpha}}, \tag{G.11}$$

from which, Doob's martingale inequality (Durrett, 2010, §5) implies immediately that

$$\mathbb{P}\left(\max_{1\le n\le N} T_n'' \ge z\right) \le \frac{1}{z^2}\sum_{i=1}^{N}\mathbb{E}(u_i'')^2 \le \frac{6\mathscr{M}^2 + 8\mathscr{B}^2}{z^2}\exp\{-\mathscr{M}^\alpha\}. \tag{G.12}$$

To prove (G.10), first recall from the definition of u_i'' in (G.5) that

$$u_i'' = u_i 1_{\{|u_i| > \mathscr{M}\|u_i\|_{\psi\alpha}\}} - \mathbb{E}\left(u_i 1_{\{|u_i| > \mathscr{M}\|u_i\|_{\psi\alpha}\}} \mid \mathscr{F}_{i-1}\right).$$

Recall from the property of conditional expectation (Durrett, 2010) that for any random variable W and a σ-algebra $\mathscr{G} \subseteq \mathscr{F}$

$$\mathbb{E}\left[W - \mathbb{E}(W \mid \mathscr{G})\right]^2 = \mathbb{E}W^2 - \mathbb{E}\left[\mathbb{E}(W \mid \mathscr{G})\right]^2 \le \mathbb{E}W^2 = \int_0^\infty 2y\mathbb{P}(|W| > y)dy$$

where the last equality is due to the second-moment formula for nonnegative random variable $|W|$ (Durrett, 2010). Plugging in $W = u_i 1_{\{|u_i|>\mathscr{M}\|u_i\|_{\psi_\alpha}\}}$ and $\mathscr{G} = \mathscr{F}_{i-1}$ we have

$$\begin{aligned}
\mathbb{E}(u_i'')^2 &= \mathbb{E}\left[u_i 1_{\{|u_i|>\mathscr{M}\|u_i\|_{\psi_\alpha}\}} - \mathbb{E}\left(u_i 1_{\{|u_i|>\mathscr{M}\|u_i\|_{\psi_\alpha}\}} \mid \mathscr{F}_{i-1}\right)\right]^2 \\
&\le \int_0^\infty 2y\mathbb{P}(|u_i| 1_{|u_i|>\mathscr{M}\|u_i\|_{\psi_\alpha}} > y)dy \\
&= \int_0^{\mathscr{M}\|u_i\|_{\psi_\alpha}} 2ydy \cdot \mathbb{P}(|u_i| > \mathscr{M}\|u_i\|_{\psi_\alpha}) + \int_{\mathscr{M}\|u_i\|_{\psi_\alpha}}^\infty 2y\mathbb{P}(|u_i| > y)dy \\
&= \mathscr{M}^2\|u_i\|_{\psi_\alpha}^2 \mathbb{P}(|u_i| > \mathscr{M}\|u_i\|_{\psi_\alpha}) + \int_{\mathscr{M}}^\infty 2t\|u_i\|_{\psi_\alpha}\mathbb{P}(|u_i| > t\|u_i\|_{\psi_\alpha})\ \|u_i\|_{\psi_\alpha} dt \\
&\le 2\mathscr{M}^2\|u_i\|_{\psi_\alpha}^2 \exp\{-\mathscr{M}^\alpha\} + 4\|u_i\|_{\psi_\alpha}^2 \int_{\mathscr{M}}^\infty t\exp\{-t^\alpha\}\ dt,
\end{aligned} \tag{G.13}$$

where the last inequality is due to Markov's inequality that for all $z > 0$

$$\mathbb{P}(|u_i|/\|u_i\|_{\psi_\alpha} \ge z) \le \exp\{-z^\alpha\}\mathbb{E}\exp\{|u_i|^\alpha/\|u_i\|_{\psi_\alpha}^\alpha\} \le 2\exp\{-z^\alpha\}. \tag{G.14}$$

It can be shown from Calculus I that the function $g(t) = t^3\exp\{-t^\alpha\}$ is decreasing in $[\mathscr{B}, +\infty)$ and is increasing in $[0, \mathscr{B}]$, where $\mathscr{B}$ was earlier defined in (G.11) (Fan et al., 2012). If $\mathscr{M} \in [\mathscr{B}, \infty)$ we have

$$\begin{aligned}
\int_{\mathscr{M}}^\infty t\exp\{-t^\alpha\}\ dt &= \int_{\mathscr{M}}^\infty t^{-2}t^3\exp\{-t^\alpha\}\ dt \\
&\le \int_{\mathscr{M}}^\infty t^{-2}\ dt \cdot \mathscr{M}^3\exp\{-\mathscr{M}^\alpha\} \\
&= \mathscr{M}^{-1}\cdot\mathscr{M}^3\exp\{-\mathscr{M}^\alpha\} = \mathscr{M}^2\exp\{-\mathscr{M}^\alpha\}.
\end{aligned} \tag{G.15}$$

If $\mathscr{M} \in (0, \mathscr{B})$, we have by setting $\mathscr{M}$ as $\mathscr{B}$ in above

$$\begin{aligned}
\int_{\mathscr{M}}^\infty t\exp\{-t^\alpha\}\ dt &= \int_{\mathscr{M}}^{\mathscr{B}} t\exp\{-t^\alpha\}\ dt + \int_{\mathscr{B}}^\infty t\exp\{-t^\alpha\}\ dt \\
&\le \int_{\mathscr{M}}^{\mathscr{B}} dt \cdot \mathscr{B}\exp\{-\mathscr{M}^\alpha\} + \mathscr{B}^2\exp\{-\mathscr{B}^\alpha\} \\
&\le (\mathscr{B}-\mathscr{M})\mathscr{B}\exp\{-\mathscr{M}^\alpha\} + \mathscr{B}^2\exp\{-\mathscr{M}^\alpha\} \\
&\le 2\mathscr{B}^2\exp\{-\mathscr{M}^\alpha\}.
\end{aligned} \tag{G.16}$$

Combining (G.13) with the two above displays (G.15) and (G.16) we obtain

$$\begin{aligned}
\mathbb{E}(u_i'')^2 &\le 2\mathscr{M}^2\|u_i\|_{\psi_\alpha}^2\exp\{-\mathscr{M}^\alpha\} + 4\|u_i\|_{\psi_\alpha}^2\int_{\mathscr{M}}^\infty t\exp\{-t^\alpha\}\ dt \\
&\le (6\mathscr{M}^2 + 8\mathscr{B}^2)\|u_i\|_{\psi_\alpha}^2\exp\{-\mathscr{M}^\alpha\},
\end{aligned}$$

completing the proof of (G.10) and hence (G.12).

(iv) Putting the pieces together: combining (G.7), (G.9) and (G.12) we obtain for an arbitrary $u \in (0,\infty)$ that

$$\begin{aligned}\mathbb{P}\left(\max_{1\le n\le N} T_n \ge 2z\right) &\le \mathbb{P}\left(\max_{1\le n\le N} T_n' \ge z\right) + \mathbb{P}\left(\max_{1\le n\le N} T_n'' \ge z\right) \\ &\le \exp\left\{-\frac{z^2}{8\mathscr{M}^2}\right\} + \frac{6\mathscr{M}^2+8\mathscr{B}^2}{z^2}\exp\{-\mathscr{M}^\alpha\}\end{aligned} \tag{G.17}$$

We choose $\mathscr{M}$ as, by making the exponents equal in above,

$$\mathscr{M} = \left(\frac{z^2}{8}\right)^{\frac{1}{\alpha+2}} \quad \text{such that} \quad \frac{z^2}{8\mathscr{M}^2} = \mathscr{M}^\alpha = \left(\frac{z^2}{8}\right)^{\frac{\alpha}{\alpha+2}}.$$

Plugging this $\mathscr{M}$ back into (G.17) we obtain

$$\begin{aligned}\mathbb{P}\left(\max_{1\le n\le N} T_n \ge 2z\right) &\le \exp\left\{-\left(\frac{z^2}{8}\right)^{\frac{\alpha}{\alpha+2}}\right\} + \frac{6\mathscr{M}^2+8\mathscr{B}^2}{z^2}\exp\left\{-\left(\frac{z^2}{8}\right)^{\frac{\alpha}{\alpha+2}}\right\} \\ &\le \left[1+\left(\frac{1}{8}\right)^{\frac{2}{\alpha+2}}\frac{6}{z^{\frac{2\alpha}{\alpha+2}}} + \left(\frac{3}{\alpha}\right)^{\frac{2}{\alpha}}\frac{8}{z^2}\right]\exp\left\{-\left(\frac{z^2}{8}\right)^{\frac{\alpha}{\alpha+2}}\right\}\end{aligned} \tag{G.18}$$

where we plugged in the expression of $\mathscr{B}$ in (G.11). We can further simplify the square-bracket prefactor in the last line of (G.18) which can be tightly bounded by

$$\begin{aligned}1+\left(\frac{1}{8}\right)^{\frac{2}{\alpha+2}}\frac{6}{z^{\frac{2\alpha}{\alpha+2}}} + \left(\frac{3}{\alpha}\right)^{\frac{2}{\alpha}}\frac{8}{z^2} &\le 1 + \frac{6\cdot\frac{2}{\alpha+2}}{(8)^{\frac{2}{\alpha+2}}} + \frac{6\cdot\frac{\alpha}{\alpha+2}}{(8)^{\frac{2}{\alpha+2}}z^2} + \left(\frac{3}{\alpha}\right)^{\frac{2}{\alpha}}\frac{8}{z^2} \\ &\le 3 + \left(\frac{0.75\cdot\frac{\alpha}{\alpha+2}}{(8)^{\frac{2}{\alpha+2}}} + \left(\frac{3}{\alpha}\right)^{\frac{2}{\alpha}}\right)\frac{8}{z^2} \\ &\le 3 + \left(0.75 + \left(\frac{3}{\alpha}\right)^{\frac{2}{\alpha}}\right)\frac{8}{z^2} \\ &\le 3 + \left(\frac{3}{\alpha}\right)^{\frac{2}{\alpha}}\frac{16}{z^2}.\end{aligned}$$

where we used an implication of Jensen's inequality: for $\gamma = \alpha/(\alpha+2) \in (0,1)$ one has $x^\gamma \le 1-\gamma+\gamma x$ for all $x \ge 0$ (where the equality holds for $x=1$), as well as a few elementary algebraic inequalities, including $\gamma 8^{-\gamma} < 0.177$, $(1-\gamma)8^{-\gamma} < 1$, $(3/\alpha)^{2/\alpha} > 0.78$ for all $\alpha > 0$ and $0 < \gamma = 2/(\alpha+2) < 1$. Thus, (G.3) is concluded by noticing the relation (G.6) and setting $z/2$ in the place of z, which hence proves Theorem 6 via the argument in (i) in our proof.

□