Stochastic Approximation for Online Tensorial Independent Component Analysis

Chris Junchi Li JUNCHILI@BERKELEY.EDU
Michael I. Jordan JORDAN@CS.BERKELEY.EDU
Department of EECS & Department of Statistics, University of California, Berkeley

Editors: Mikhail Belkin and Samory Kpotufe

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 provide a dynamics-based analysis to 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 sharp 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 and 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 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; Frieze et al., 1996; 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 as the problem of optimizing an objective function based on the fourth-order cumulant tensor over the unit sphere $\mathcal{D}_1 \equiv \{\mathbf{u} \in \mathbb{R}^d : \|\mathbf{u}\| = 1\}$, where $\|\cdot\|$ denotes the Euclidean norm. The optimization problem is as follows:

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

Such an objective function is referred to as a non-Gaussianity contrast function (Hyvärinen et al., 2001). The formulation (1.1) of ICA via its fourth-order cumulant tensor was initiated by Comon (1994) and Frieze et al. (1996) in the batch case. The landscape of objective (1.1) is featured by its nonconvexity, in the sense that it presents $2d$ local minimizers $\pm a_1, \ldots, \pm a_d$ and exponentially many unstable stationary points in terms of d (Ge et al., 2015; Li et al., 2016; Sun et al., 2015).[1] Here, by unstable stationary points we refer to those points with zero gradient and at least one negative Hessian direction, and hence includes the collection of saddle points and local maximizers. Algorithms that find local minimizers of (1.1) allow us to estimate the columns of the mixing matrix 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 and Lu, 2017). Initialize a unit vector $u^{(0)}$ appropriately, at step $t = 1, 2, \ldots, T$ the algorithm processes an observation $x^{(t)}$ by performing the following update:

$$u^{(t)} = \big(1 - \eta^{(t)}\big)\, u^{(t-1)} + \eta^{(t)} \operatorname{sign}\big(\kappa_4 - 3\big)\, \big(u^{(t-1)}\big)^\top x^{(t)})^3\, x^{(t)}; \tag{1.2}$$

where $\eta^{(t)}$ is a positive stepsize, and the operator $\mathrm{proj}_{\mathbb{S}^{d-1}}[\,\cdot\,] := \cdot / \|\cdot\|$ projects a nonzero vector onto the unit sphere $\mathbb{S}^{d-1}$ 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 including saddle points and local maximizers. 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 and beyond (for a more general class of nonconvex optimization problems). In contrast, we present a dynamics-based analysis that requires no noise injection step. We show that a uniform initialization on the unit sphere $\mathbb{S}^{d-1}$, along with mild scaling conditions, is sufficient to ensure that the algorithm iterate enters a basin of attraction and finds a local (and also global) minimizer with high probability.

Overview of the Main Result and Contributions Our main result states that for estimating a single independent component, the iteration (1.2) with uniform initialization and carefully chosen stepsize $\eta^{(t)}$ enters the basin of attraction of a uniformly drawn independent component and achieves a convergence rate of $O(\sqrt{d/T})$ in terms of angle with respect to an independent component, up to a polylogarithmic factor of d and T. Such a convergence result holds under mild distributional assumptions and scaling conditions that involve the data dimension d and sample size T. Informally, this result is stated as follows.

Theorem 1 (Informal version of Corollary 8 in §3) Let the initial $u^{(0)}$ be sampled uniformly from the unit sphere $\mathbb{S}^{d-1}$, and let appropriate light-tailed distributional assumptions hold. Then, for any fixed positive $\varepsilon \in (0, 1/5]$ satisfying the scaling condition:

$$d \gtrsim 2^{\sqrt{2e}} \log\big(\varepsilon^{-1} + 1\big); \qquad C_{1;T} \log^8(C_{1;T}\, \varepsilon^{-1} d T)\, \frac{B^8}{|\kappa_4 - 3|^2}\, \frac{d^4 \log^2 T}{T} \lesssim \frac{\varepsilon^2}{\log^2 \varepsilon^{-1}};$$

1. In contrast, the rank-one eigenvalue problem presents $2d$ local minimizers and $2d^2 - 2$ unstable stationary points in the typical setting of distinct eigenvalues.

Algorithm 1 Online Tensorial ICA

Initialize $u^{(0)}$ and select stepsize $\eta^{(t)}$ appropriately (to elaborate later)
for $t = 1, 2, \ldots$ do
  Draw one observation $X^{(t)}$ from streaming data, and update iteration $u^{(t)}$ via

$$u^{(t)} = \mathcal{P}_1\left(u^{(t-1)} + \eta^{(t)} \cdot \mathrm{sign}(\kappa_4) \cdot \left((u^{(t-1)})^\top X^{(t)}\right)^3 X^{(t)}\right) \tag{1.3}$$

  where $\mathcal{P}_1(\cdot) = (\cdot)/\|\cdot\|^{-1}$ denotes the projection operator onto the unit sphere centered at the origin $\mathbb{R}^d$
end for

there exists a uniformly distributed random variable $\hat{k} \in [d]$ such that with probability at least $1 - 5\delta$, iteration $u^{(t)}$ of (1.2) with appropriate choice of stepsize $\eta^{(t)}$ satisfies

$$\tan\angle\left(u^{(T)}, a_{\hat{k}}\right) \le C_{1,T} \log^{5/2}(C_{1,T}\, d/\delta) \cdot \frac{B^4}{|\kappa_4|^3} \cdot \sqrt{\frac{d \log^2 T}{T}},$$

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

To the best of our knowledge, this provides the first rigorous analysis of an online tensorial ICA algorithm that achieves a $\mathcal{O}(\sqrt{d/T})$ finite-sample convergence rate, under mild distributional assumptions and scaling conditions. The contributions of this work lie in several aspects:

(i) Partly adapting from the analysis of online principal component estimation in Li et al. (2018), we provide a per-coordinate analysis of the warmly-initialized online tensorial ICA algorithm that achieves a sharp $\mathcal{O}(\sqrt{d/T})$ convergence rate [Theorem 4].

(ii) In contradistinction to existing saddle-point-escaping analysis in Ge et al. (2015); Jin et al. (2017); Li et al. (2018), we developed a novel coordinate-pair analysis of the uniformly-initialized online tensorial ICA algorithm based on our dynamics-based characterization [Theorem 7].

(iii) We combine these two analyses to conclude that a two-stage training procedure provides a finite-sample error bound of $\mathcal{O}(\sqrt{d/T})$ for the uniform initialization case under a mild assumption on the data-generating distribution and also a scaling condition $d = o(T)$ being sufficiently small, up to a polylogarithmic factor of d, T [Theorem 1, presented formally in Corollary 8].

Organization The rest of this paper is organized as follows. §2 presents our main convergence results and finite-sample error bounds for the warm initialization case for estimating one single component. §3 presents the corresponding results for the uniform initialization case. §4 discusses additional related literature. §5 summarizes our results. Limited by space we relegate to the Appendix the proofs of main results, all secondary lemmas and technical results along with preliminary simulation results.

2. Estimating a Single Component: Warm Initialization Case

For the purpose of estimating a single independent component, we introduce our settings and assumptions for tensorial ICA and its stochastic approximation algorithm (1.2), formally stated in

Algorithm 1. Let the dimension $d \geq 2$, let X be the data vector of which $X^{(1)}, X^{(2)}, \ldots \in \mathbb{R}^d$ are independent draws, and assume the following for the distribution of X:

Assumption 2 (Data vector distribution) Let $X = AZ$, where $A \in \mathbb{R}^{d \times d}$ is an orthogonal matrix with $AA^\top = I$ and $Z \in \mathbb{R}^d$ 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 $\gamma_j \coloneqq \mathbb{E}Z_i^j$;

(ii) The $\gamma_1 = \mathbb{E}Z_i = 0$, $\gamma_2 = \mathbb{E}Z_i^2 = 1$, $\gamma_4 = \mathbb{E}Z_i^4 - 3 \neq 0$;

(iii) For all $i \in [d]$, Z_i has an Orlicz-ψ_2 norm bounded by $\sqrt{3/8}B$.[2]

Note that Assumption 2(i) requires the distribution for all independent components to admit identical first, second and fourth moments. As indicated in Assumption 2(ii), the data vectors are assumed to be whitened first in the sense that $\gamma_2 = 1$. The sign of our excess kurtosis $\gamma_4 - 3$ determines the direction of stochastic gradient update, and, as will be seen later, the magnitude of the excess kurtosis $|\gamma_4 - 3|$ plays an important role in our convergence analysis.[3] Assumption 2(iii) generalizes the boundedness assumption $\|Z_i\|_1 \leq O(B)$ made in recent work in order to cover Gaussian mixtures and Bernoulli-Gaussians, which are typical application cases for the tensorial ICA estimation problem. Note that we include a factor of $\sqrt{3/8}$ in the Orlicz-ψ_2 norm (sub-Gaussian parameter) purely 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 $u^{(0)}$ satisfying, for some integer $i \in [d]$,

$$u^{(0)} \in \mathcal{D}_1 \quad \text{and} \quad \tan \angle\left(u^{(0)}, a_i\right) \leq \frac{1}{\sqrt{3}}. \tag{2.1}$$

For any fixed $\eta > 0$, we define a rescaled time variable T_η as

$$T_\eta \coloneqq \left\lceil \frac{\log \frac{|\gamma_4 - 3|}{B^8 \eta}}{\log\left(1 + \frac{1}{3}|\gamma_4 - 3|\right)} \right\rceil. \tag{2.2}$$

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

Lemma 3 (Convergence Result with Warm Initialization) Let the dimension $d \geq 2$, let Assumption 2 hold, and let initialization $u^{(0)}$ satisfy condition (2.1) for some integer $i \in [d]$. Then, for any

2. A random variable Z with mean 0 is light-tailed if there is a positive number σ such that $\mathbb{E}\exp(\lambda Z) \leq \exp(\lambda^2\sigma^2/2)$ for all $\lambda \in \mathbb{R}$. The smallest possible $\sigma > 0$ is referred to as the Orlicz-ψ_2 norm or sub-Gaussian parameter (Wainwright, 2019). Readers shall be warned that the term "sub-Gaussian" here indicates the light-tailed condition and should be distinguished from $\gamma_4 < 3$ (resp. $\gamma_4 > 3$) of sub-Gaussianity (resp. super-Gaussianity) in ICA (Stone, 2004).

3. When the excess kurtosis $\gamma_4 - 3$ is equal to zero, for instance when Z follows an i.i.d. standard normal distribution, the matrix A is non-identifiable in our framework. Non-gaussian independent components with $\gamma_4 = 3$ can be studied via higher-order tensor decomposition with a different contrast function, but is beyond the scope of this paper.

xed positive numbers, and $2 (0; e^{-1}]$ satisfying the scaling condition,

$$C_{3;L} \log^8(T_{;1}^{-1}) \frac{B^8}{j_4 3j} d \log \frac{j_4 3j}{B^8} \quad 1 \quad \frac{1}{+1}; \quad < \min \frac{1}{j_4 3j}; \frac{j_4 3j}{B^8} e^{-1}; \tag{2.3}$$

there exists an event $H_{3;L}$ with $P(H_{3;L}) \quad 1 \quad 6 \quad +12 + \frac{5184}{\log^5 1} d$; such that on $H_{3;L}$, iteration $u^{(t)}$ of Algorithm 1 satis es

$$\tan \setminus u^{(t)}; a_i \quad \tan \setminus u^{(0)}; a_i \quad 1 \quad \frac{}{3} j_4 3j^t + \sqrt{+1} C_{3;L} \log^{5=2} 1 \frac{B^4}{j_4 3j^{1=2}} \sqrt{d \log \frac{j_4 3j}{B^8} 1}; \tag{2.4}$$

for all $t 2 [0; T_{;}]$, where $C_{3;L}$ and $C_{3;L}$ are positive, absolute constants.

Lemma 3 provides, under the warm initialization assumption (2.1) and scaling condition $O(d^{-1})$, an upper bound for $\tan \setminus u^{(t)}; a_i$ which is the sum of two terms: the rst term on the right hand side of (2.4) decays geometrically from $\tan \setminus u^{(0)}; a_i$ at rate $1 j_4 3j = 3$, and the second term $O(\sqrt{d})$ is induced by the noise. To balance these two terms, when we know in advance the sample size T of online data satisfying some scaling condition $T = (d)$, we choose a constant stepsize $= (\log T=T)$ and obtain a nite-sample error bound:

Theorem 4 (Finite-Sample Error Bound with Warm Initialization) Let the dimension $d 2$, let Assumption 2 hold, and let initialization $u^{(0)}$ satisfy condition (2.1) for some integer $i 2 [d]$. For sample size T set the stepsize as

$$(T) = \frac{9 \log \frac{2j_4 3j^2}{9B^8} T}{2j_4 3jT}: \tag{2.5}$$

Then, for any xed positive number T s $100; 2 (0; 1]$ satisfying the scaling condition

$$C_{4;T} \log^8(C^0_{4;T} \quad ^1 dT) \frac{d \log^2 T}{T} \quad \frac{j_4 3j^2}{B^8}; \tag{2.6}$$

there exists an event $H_{4;T}$ with $P(H_{4;T}) \quad 1$ such that on $H_{4;T}$, iteration $u^{(t)}$ of Algorithm 1 satis es

$$\tan \setminus u^{(T)}; a_i \quad C_{4;T} \log^{5=2}(C^0_{4;T} \quad ^1 d) \frac{B^4}{j_4 3j} \sqrt{\frac{d \log^2 T}{T}};$$

where $C_{4;T}; C_{4;T}; C^0_{4;T}$ are positive, absolute constants.

Theorem 4 indicates a $O(\sqrt{d=T})$ nite-sample error bound for online tensorial ICA when it is warmly initialized in the sense that (2.1) holds for $u^{(0)}$ for some integer $i 2 [d]$. In the rest of this section, we prove Lemma 3 and Theorem 4 for the warm initialization case. The section is organized as follows: x2.1 analyzes our algorithm and provide a key lemma [Lemma 5] when it is warmly initialized. Proof of the key Lemma 5 and all proofs of secondary lemmas are deferred to xA.1 and xB 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 $a_i^{(0)}$ satisfies condition (2.1) and define the rotated iteration $\{v^{(t)}\}_{t \geq 0}$ as

$$v^{(t)} \coloneqq PA^\top u^{(t)}; \tag{2.7}$$

where $P \in \mathbb{R}^{d \times d}$ is the permutation matrix corresponding to the cycle $(1\ i)$; i.e., $P(i,1) = P(1,i) = 1$, $P(j,j) = 1$ 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 e_1 is the closest independent components pair at initialization and (with high probability) at convergence. Furthermore, transforming to rotated observations $Y^{(t)} = PA^\top X^{(t)}$ allows us to equivalently translate our online tensorial ICA iteration (1.3) into an analogous form:

$$v^{(t)} = \Big(1 - \frac{3}{4}\eta\Big) v^{(t-1)} + \frac{\eta}{4}\,\mathrm{sign}(\kappa_4)\,\big((v^{(t-1)})^\top Y^{(t)}\big)^3 Y^{(t)}. \tag{2.8}$$

It is easy to verify that the rotated iterations $\{v^{(t)}\}_{t \geq 0}$ and $\{u^{(t)}\}_{t \geq 0}$ satisfy $a_i^\top u^{(t)} = e_1^\top v^{(t)}$, and hence for all $t \geq 0$

$$\tan\angle\big(v^{(t)}, e_1\big) = \frac{\sqrt{1 - (e_1^\top v^{(t)})^2}}{e_1^\top v^{(t)}} = \frac{\sqrt{1 - (a_i^\top u^{(t)})^2}}{a_i^\top u^{(t)}} = \tan\angle\big(u^{(t)}, a_i\big). \tag{2.9}$$

Now, we let the warm initialization region be

$$\mathcal{D}_{\mathrm{warm}} = \Big\{v \in \mathcal{D}_1 : v_1^2 \geq \frac{3}{4}\Big\} = \Big\{v \in \mathcal{D}_1 : \tan\angle(v, e_1) \leq \frac{1}{\sqrt{3}}\Big\}. \tag{2.10}$$

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

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

To prevent iteration $U_k^{(t)}$ from diverging from the warm initialization region, we also define a larger warm-auxiliary region as $\mathcal{D}_{\text{warm-aux}} \coloneqq \big\{v \in \mathcal{D}_1 : v_1^2 \geq \frac{2}{3}\big\} = \big\{v \in \mathcal{D}_1 : \tan\angle(v, e_1) \leq \frac{1}{\sqrt{2}}\big\}$. Suppose the process $\{v^{(t)}\}$ is initialized at $v^{(0)} \in \mathcal{D}_{\mathrm{warm}}$, and we define the first time $v^{(t)}$ exits the warm-auxiliary region as

$$T_{\mathrm{x}} \coloneqq \inf\Big\{t \geq 1 : v^{(t)} \in \mathcal{D}_{\text{warm-aux}}^{\mathrm{c}}\Big\}. \tag{2.12}$$

We state the key lemma as

Lemma 5 Let the settings in Lemma 3 hold, and x the coordinate $k \in [2, d]$ and value $\varepsilon > 0$. Then for any fixed positive η, satisfying the scaling condition (2.3) along with the warm

initialization condition $v^{(0)} \in D_{warm}$, there exists an event $H_{k;5;L}$ satisfying $P(H_{k;5;L}) \geq 1 - 6\exp(-12\log^{5-1} d) - \frac{5184}{\log^{5-1} d}$, such that on event $H_{k;5;L}$ the following holds:

$$\sup_{t \in [T_{;0:5}, T_{;}]} \left| U_k^{(t)} - U_k^{(0)} \prod_{s=0}^{t-1} \left[1 + \eta_s |\lambda_4 - 3| \left((v_1^{(s)})^2 - (v_k^{(s)})^2\right)\right] \right| \leq 2C_{5;L} \log^{5/2-1} d \cdot B^4 \cdot (T_{;})^{1/2}, \tag{2.13}$$

where $C_{5;L}$ is a positive, absolute constant.

This lemma shows that for each coordinate $k \in [2, d]$, with high probability, the dynamics of $U_k^{(t)}$ is tightly controlled within a deterministic vessel whose center converges to zero at least exponentially fast. As we will later see in the proofs of Lemma 3 and Theorem 4, this guarantees that with high probability $v^{(t)}$ will not exit the warm-auxilliary region $D_{warm\text{-}aux}$ and will stay within a small neighborbood of e_1 after $T_{;0:5}$ iterates, where the rescaled time $\tilde{T}_{;}$ was earlier defined in (2.2).

3. Estimating Single Component: Uniform Initialization Case

It is often the case that we are unable to obtain a warm initialization for online tensorial ICA as required in §2, in which case we resort to initializing $u^{(0)}$ uniformly at random from the unit sphere D_1. To analyze such a case, we define a new rescaled time variable, parameterized by γ as follows:

$$T^o_{;\gamma} = \left\lceil \frac{\log \frac{|\lambda_4 - 3|}{B^8 d^{\gamma-1}}}{\log\left(1 - \frac{1}{2d}|\lambda_4 - 3|\right)} \right\rceil. \tag{3.1}$$

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

Lemma 6 (Convergence Result with Uniform Initialization) Let the dimension $d \geq 2$, let Assumption 2 hold, and let $u^{(0)}$ be uniformly sampled from the unit sphere D_1. Then for any fixed positive numbers $d \geq 2$, $\gamma > 0.5$, $\varepsilon > 0$, $\delta \in (0, e^{-1}]$, $\alpha \in (0, 1/3]$ satisfying the scaling condition

$$\begin{aligned} d &\geq 2^{\sqrt{2e \log \frac{1}{\delta}} + 1}, \quad \varepsilon < \min\left\{\frac{1}{|\lambda_4 - 3|}, \frac{|\lambda_4 - 3|}{B^8} e^{-1}\right\}, \text{ and} \\ C_{6;L} \log^8(T^o_{;1} - 1) &\leq \frac{B^8}{|\lambda_4 - 3|} d^2 \log^2 d \cdot \log \frac{|\lambda_4 - 3|}{B^8 \delta^{-1}} \cdot \frac{\varepsilon^2}{(\gamma+1)\log^{2\gamma-1} d}, \end{aligned} \tag{3.2}$$

there exists a uniformly distributed random variable $k \in \{1, \ldots, d\}$ and an event $H_{6;L}$ with

$$P(H_{6;L}) \geq 1 - 6\exp(-27\log^{5-1} d) - \frac{10368}{\log^{5-1} d} - d^{-3},$$

such that on $H_{6;L}$, iteration $u^{(t)}$ of Algorithm 1 satisfies for $t \in [T^o_{;0:5}, T^o_{;\gamma}]$

$$\begin{aligned} \tan\angle\left(u^{(t)}, a_k\right) &\leq \sqrt{d - 1}\left(1 - \frac{1}{2d}|\lambda_4 - 3|\right)^{t - T^o_{;0:5}} \\ &+ \sqrt{\gamma+1}\, C_{6;L} \log^{5/2-1} d \cdot \frac{B^4}{|\lambda_4 - 3|^{1/2}} \sqrt{d^3 \log \frac{|\lambda_4 - 3|}{B^8 \delta^{-1}}}; \end{aligned} \tag{3.3}$$

where $C_{6;L}, C_{6;L}$ are positive, absolute constants.

Under the uniform initialization assumption, Lemma 6 shows that the term $\tan\angle(u^{(t)}, a_l)$ can be upper bounded by the sum of two summands, with the first term geometrically decaying from $\sqrt{d}$ at rate $1 - |\lambda_4| - 3|\lambda|/(2d)$, and the second being the noise-induced error $\widetilde{O}(\sqrt{d^3/T})$. The key idea behind the proof is that the scaling condition (3.2) ensures that the iterate avoids the set where the unstable stationary points lie, and hence efficiently contracts to the basin of attraction of the independent component pairs.

Analogous to Theorem 4, when the sample size satisfies the scaling condition $T = \widetilde{\Omega}(d^3)$, one can carefully choose a stepsize $\eta = \widetilde{\Theta}(d \log T/T)$ and establish a finite-sample bound on $\tan\angle(u^{(t)}, a_l)$ based on Lemma 6. We formulate this fact as our second main theorem.

Theorem 7 (Finite-Sample Error Bound with Uniform Initialization) Let the dimension $d \geq 2$, let Assumption 2 hold, and let initialization $u^{(0)}$ be uniformly sampled from the unit sphere $\mathbb{S}^{d-1}$. For sample size $T \geq 100$, set the stepsize as $\eta(T) = \dfrac{4d \log \frac{|\lambda_4| - 3|\lambda|^2}{4B^8 d} T}{|\lambda_4| - 3|\lambda| T}$. Then, for any fixed positive numbers $d \geq 2, T \geq 100, \delta \in (0, 1/4]$ satisfying the scaling condition

$$d \geq 2\sqrt{2e}\log\frac{1}{\delta} + 1, \qquad C_{7,T}\log^8(C'_{7,T}/\delta d) \leq \frac{|\lambda|^8}{|\lambda_4| - 3|\lambda|^2}\cdot\frac{d^3\log^2 d\log^2 T}{T}\cdot\frac{1}{\log^2\frac{1}{\delta}}, \tag{3.4}$$

there exists a uniformly distributed random variable $l \in \{1, \ldots, d\}$ and an event $\mathcal{H}_{7,T}$ with $\mathbb{P}(\mathcal{H}_{7,T}) \geq 1 - \delta/4$ such that on $\mathcal{H}_{7,T}$, iteration $u^{(t)}$ of Algorithm 1 satisfies

$$\tan\angle(u^{(T)}, a_l) \leq C_{7,T}\log^{5/2}(C'_{7,T}/\delta d)\cdot\frac{|\lambda|^4}{|\lambda_4| - 3|\lambda|}\cdot\sqrt{\frac{d^4\log^2 T}{T}},$$

where $C_{7,T}, C_{7,T}, C'_{7,T}$ are positive, absolute constants.

Theorem 7 achieves, under the scaling condition $T = \widetilde{\Omega}(d^3)$, an $\widetilde{O}(\sqrt{d^4/T})$ finite-sample error bound on $\tan\angle(u^{(T)}, a_l)$ for some l drawn uniformly at random in $[d]$. With Theorems 4 and 7 for warm and uniform initializations respectively, a specific choice of stepsize allows us to have the best of the two worlds. Assuming prior knowledge of the sample size T, we initialize $u^{(0)}$ uniformly at random from the unit sphere $\mathbb{S}^{d-1}$ and run Algorithm 1 in two consecutive phases, each using $T/2$ observations:

- In the first phase, we initialize $u^{(0)}$ uniformly at random on unit sphere $\mathbb{S}^{d-1}$, pick a constant stepsize $\eta_1 = \widetilde{\Theta}(d \log T/T)$ and update iteration $u^{(t)}$ via (1.3) for $T/2$ iterates. Theorem 7 guarantees with high probability that $u^{(T/2)}$ satisfies the warm initialization condition (2.1) under the scaling condition $T = \widetilde{\Omega}(d^4)$;
- In the second phase, we warm-initialize the algorithm using the output of the first phase $u^{(T/2)}$, pick a constant stepsize $\eta_2 = \widetilde{\Theta}(\log T/T)$ and update the iteration $u^{(t)}$ via (1.3) for $T/2$ iterates. The last iterate achieves an error bound $\widetilde{O}(\sqrt{d/T})$ as indicated by Theorem 4.

This two-phase procedure yields an improved finite-sample error bound $\widetilde{O}(\sqrt{d/T})$ under the uniform initialization and scaling condition $T = \widetilde{\Omega}(d^4)$, formally stated in the following corollary.

Corollary 8 (Improved Finite-Sample Error Bound with Uniform Initialization) Let the dimension $d \geq 2$, let Assumption 2 hold, and let initialization $u^{(0)}$ be uniformly sampled from the unit

sphere $\mathbb{D}_1$. Set for sample size $T \geq 200$ the stepsizes as

$$\eta_1(T) = \frac{8d \log \frac{|\lambda_4| - 3|\lambda_2|^2}{8B^8 d} T}{|\lambda_4| - 3|\lambda_2| T}; \qquad \eta_2(T) = \frac{9 \log \frac{|\lambda_4| - 3|\lambda_2|^2}{9B^8} T}{|\lambda_4| - 3|\lambda_2| T}. \tag{3.5}$$

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

$$d \geq 2\sqrt{2e \log \tfrac{1}{\delta}} + 1, \qquad \varepsilon \leq \max\{C_{4,T}, C_{7,T}, C_{7,T}^2\} \log^8(C_{7,T}^0 / \delta) \frac{B^8}{|\lambda_4| - 3|\lambda_2|^2} \frac{d^4 \log^2 T}{T} \leq \frac{1}{\log^2 \frac{1}{\delta}}, \tag{3.6}$$

there exists a uniformly distributed random variable $\xi \in [d]$ and an event $\mathcal{H}_{8,C}$ with $\mathbb{P}(\mathcal{H}_{8,C}) \geq 1 - 5\delta$ such that on the event $\mathcal{H}_{8,C}$, running Algorithm 1 for $T/2$ iterates with stepsize $\eta_1(T)$ followed by $T/2$ iterates with stepsize $\eta_2(T)$ outputs an $u^{(T)}$ satisfying

$$\tan \angle\left(u^{(T)}, a_\xi\right) \leq \sqrt{2} C_{4,T} \log^{5/2}(C_{4,T}^0 / \delta) \frac{B^4}{|\lambda_4| - 3|\lambda_2|} \sqrt{\frac{d \log^2 T}{T}}.$$

where $C_{4,T}, C_{4,T}^0, C_{7,T}, C_{7,T}, C_{7,T}^0$ are positive, absolute constants defined earlier in Theorems 4 and 7.

We make several remarks:

(i) Our dynamics-based analysis of uniformly-initialized online tensorial ICA is different from that of PCA in Li et al. (2018) (see Section 1). In particular, we provide a coordinate-pair analysis of the algorithm to cope with the landscape of ICA. A two-phase procedure is, however, required to achieve a rate that is sharp in dimension d.

(ii) Ge et al. (2015) make use of artificial noise injection in their noisy projected SGD algorithm, which includes online tensorial independent component analysis as an application scenario. The analysis provided in the appendix of Ge et al. (2015), however, is worsened by its unrealistic isotropic covariance assumption for the incurred stochastic noise, which is not satisfied by the tensorial ICA algorithm, noise-injected or not. A straightforward extension to the generic noise satisfied by the tensorial ICA condition was claimed, but the analysis of such a case is not available in Ge et al. (2015).

(iii) By choosing δ in rate analysis above as a small positive constant, say $\delta = 1/12$, we achieve a scaling condition such that d^4/T is sufficiently small up to a polylogarithmic factor, we achieve the finite-sample convergence rate with probability no less than $1 - 12\delta = 0$. In comparison, Ge et al. (2015) achieve an $\varepsilon = T^{-1/4}$ error bound for some $\varepsilon > 2$, under a more stringent assumption that all independent components are almost surely bounded, which indicates a scaling condition of $d^8/T \to 0$ at best.[4] The scaling condition imposed in Theorem 1, however, leads to an error bound that depends polynomially on the inverse unsuccessful probability $1/\delta$.

In the rest of this section we study the uniform initialization case and prove Theorem 7. The key idea behind our analysis is that the uniform initialization is sufficiently far from the set where the

4. Our scaling condition is the best known when ignoring the polylogarithmic factors. Optimistically, if analysis in the ε^{-4} or $\varepsilon^{-3.5}$ state-of-the-art complexity results in Jin et al. (2021); Fang et al. (2019) can be carried over to Riemannian optimization, setting $\varepsilon = 1/d$ implies a complexity of d^9 or d^8 at best (for the ICA objective the smallest Hessian eigenvalue changes from $-1/d$ to (1) so the Hessian-Lipschitz constant is (1) and the stochastic gradient admits a variance of (d)).

unstable stationary points lie (with high probability), and thus a delicate concentration analysis can show that 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 $u^{(t)} = PA^\top u^{(t)}$ and the rotated observations $Y^{(t)} = PA^\top X^{(t)}$, so our online tensorial ICA update rule can still be translated into (2.8). Here the permutation matrix has $P(I, 1) = P(1, I) = 1 = P(j, j)$ for $j \in \{1, I\}^c$ and 0 elsewhere, in which $I = \operatorname{argmin}_{i \in [d]} \tan^2\angle(u^{(0)}, a_i)$. We let the coordinate-wise intermediate initialization region for each $k \in [2, d]$ be

$$\mathcal{D}_{\mathrm{mid};k} := \Big\{ v \in \mathcal{D}_1 : v_1^2 > \max_{i \in [2,d]} v_i^2 \text{ and } v_1^2 > 3v_k^2 \Big\}. \tag{3.7}$$

In addition, we let the cold initialization region be

$$\mathcal{D}_{\mathrm{cold}} = \Big\{ v \in \mathcal{D}^{d-1} : v_1^2 > \max_{i \in [2,d]} v_i^2 \Big\}. \tag{3.8}$$

By definition of the rotated iteration $\{v^{(t)}\}$ and index I, we know that $v^{(0)} \in \mathcal{D}_{\mathrm{cold}}$ always holds. As we will see later in Lemmas 9, 10 and 11 of this section, when initialized uniformly at random on the unit sphere $\mathcal{D}_1$ a gap between $(v_1^{(0)})^2$ and $\max_{2 \le k \le d}(v_k^{(0)})^2$ persists on a high-probability event $(\bigcap_{k \in [2,d]} \mathcal{H}_{k;9;L})$ as the data dimension $d \to \infty$. Moreover on a high-probability event $(\bigcap_{k \in [2,d]} \mathcal{H}_{k;9;L})$, the iterate $v^{(t)}$ enters the intersection of intermediate regions $\bigcap_{k \in [2,d]} \mathcal{D}_{\mathrm{mid};k}$ within $T^{\circ}_{;0.5}$ iterations. After $v^{(t)}$ enters $\bigcap_{k \in [2,d]} \mathcal{D}_{\mathrm{mid};k}$, on a third high-probability event $(\bigcap_{k \in [2,d]} \mathcal{H}_{k;11;L})$ it decays exponentially fast and stays within $\mathcal{O}(\sqrt{d^3})$-neighborhood of the independent component pair $\pm e_1$.

3.1. Initialization in the Intermediate Region

Recall the intermediate initialization region $\mathcal{D}_{\mathrm{mid};k}$ defined in (3.7) for each $k \in [2, d]$. We also define a slightly larger coordinate-wise intermediate auxiliary region for each $k \in [2, d]$ as $\mathcal{D}_{\text{mid-aux};k} = \{ v \in \mathcal{D}_1 : v_1^2 > \max_{i \ge 2} v_i^2 \text{ and } v_1^2 > 2v_k^2 \}$. When $v^{(0)} \in \mathcal{D}_{\mathrm{mid};k}$, we define the first time the iterate exits $\mathcal{D}_{\text{mid-aux};k}$ as

$$T_{w;k} = \inf \Big\{ t \ge 1 : v^{(t)} \in \mathcal{D}^c_{\text{mid-aux};k} \Big\}. \tag{3.9}$$

Thus, for each $k \in [2, d]$, $T_{w;k}$ is a stopping time with respect to the filtration $\mathcal{F}_t = \sigma(v^{(0)}, Y^{(1)}, \ldots, Y^{(t)})$ (we suppose that all k that appear in the rest of our discussion satisfy $k \in [2, d]$, unless stated otherwise).

Our goal is to prove the following high-probability bound for each coordinate. For the analysis of the intermediate iterates, we recall the notation $u_k^{(t)} = v_k^{(t)} / v_1^{(t)}$ previously defined in (2.11).

Lemma 9 Let the settings in Lemma 6 hold, and fix the coordinate $k \in [2, d]$ and value $\varepsilon > 0$. Then for any positive numbers s, L satisfying the scaling condition (3.2), and given the coordinate-wise intermediate initialization condition $v^{(0)} \in \mathcal{D}_{\mathrm{mid};k}$, there exists an event $\mathcal{H}_{k;9;L}$ satisfying $\mathbb{P}(\mathcal{H}_{k;9;L}) \ge 1 - 6\varepsilon + 12\varepsilon + \frac{5184}{\log^5 \frac{1}{\varepsilon}}$; such that on event $\mathcal{H}_{k;9;L}$ the following holds:

$$\sup_{t \le T^{\circ}_{;\varepsilon} \wedge T_{w;k}} \Big| U_k^{(t)} - U_k^{(0)} \prod_{s=0}^{t-1} \Big[1 + \eta^2 \big| (v_1^{(s)})^2 - (v_k^{(s)})^2 \big| \Big] \Big| \le 2C_{9;L} \log^{5/2} \tfrac{1}{\varepsilon} \cdot \eta^4 \cdot d^{1/2} \cdot (T^{\circ}_{;\varepsilon})^{1/2}; \tag{3.10}$$

where $C_{9;L}$ is a positive, absolute constant.

From (3.10) in Lemma 9 we know that, for each coordinate $k \geq 2$ initialized in the intermediate region $D_{mid;k}$, with high probability the iteration $U_k^{(t)}$ fluctuates around a deterministic curve that decays geometrically whenever $(w_1^{(t)})^2 \quad (v_k^{(t)})^2$ is bounded below by a positive number.

3.2. Initialization in the Cold Region

Recall the cold initialization region D_{cold} defined earlier in (3.8). The iteration we study in the old initialization analysis is

$$W_k^{(t)} = \frac{(v_1^{(t)})^2 \quad (v_k^{(t)})^2}{(v_k^{(t)})^2}: \tag{3.11}$$

Under the setting of uniform initialization on the unit sphere in Lemma 6, we have the following lemma. Note that the uniform initialization conditions for $w^{(0)}$ and $v^{(0)}$ are equivalent.

Lemma 10 Let $v^{(0)}$ be uniformly sampled from the unit sphere S^{d-1} and $\quad$ be any fixed positive number, with dimension d and $\quad$ satisfying

$$d \quad 2^{\sqrt{2e}} \log \quad ^{-1} + 1: \tag{3.12}$$

Then there exists an event $H_{10;L}$ with $P(H_{10;L}) \quad 1 \quad 3 \quad$ such that on event $H_{10;L}$ the following holds:

$$\min_{2 \quad k \quad d} W_k^{(0)} \quad \frac{}{8 \log \quad ^{-1} \log d}: \tag{3.13}$$

Our goal is to estimate the time when $v^{(t)}$ enters each coordinate-wise intermediate initialization region starting with the initialization gap given in Lemma 10. For each coordinate $k \geq 2$, we define the first time $v^{(t)}$ enters the coordinate-wise intermediate region $D_{mid;k}$ as

$$T_{c;k} = \inf \left\{ t \quad 1 : v^{(t)} \in D_{mid;k} \right\}; \tag{3.14}$$

and the first time the iterate exits D_{cold} without entering $D_{mid;k}$, earlier defined in (3.7) and (3.8), as

$$T_1 = \inf \left\{ t \quad 1 : v^{(t)} \in D_{cold}^c \right\}: \tag{3.15}$$

In other words, $T_{c;k}$ is the first time t such that $W_k^{(t)} \quad 2$, and T_1 is the first time $\min_{i \in [2;d]} W_i^{(t)} < 0$. By the definition of the rotated iteration $w^{(t)}$ we have $v^{(0)} \in D_{cold}$ always holds. For each coordinate $k \in [2; d]$, if $T_{c;k} = 0$ then $v^{(0)} \in D_{mid;k}$ and the previous analysis directly applies for coordinate k. The following lemma characterizes the opposite case, where the exponential growth of iteration $W_k^{(t)}$ helps us to determine when $v^{(t)}$ enters the intermediate region $D_{mid;k}$.

Lemma 11 Let the settings in Lemma 6 hold, and fix the coordinate $k \in [2; d]$. Then, for any fixed positive numbers; satisfying the scaling condition (3.2) along with the coordinate-wise

cold initialization condition $v^{(0)} \in D_{cold} \setminus D^c_{mid;k}$, there exists an event $H_{k;11;L}$ with $P(H_{k;11;L}) \geq 1 - 15 + \frac{5184}{\log^5 \quad 1}$; such that on event $H_{k;11;L}$ the following holds:

$$\sup_{t \quad T^o_{;\,0:5} \wedge T_{c;k} \wedge T_1} W_k^{(t)} \prod_{s=0}^{t-1} 1 + \quad j \quad 4 \quad 3j(v_1^{(s)})^2 \quad ^1 \quad W_k^{(0)} \quad C_{11;L} \log^{5=2} \quad ^1 \frac{B^4}{j \quad 4 \quad 3j^{1=2}} d \quad ^{1=2}; \tag{3.16}$$

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

From Lemma 11 we know that for each coordinate $k \geq [2; d]$, if the initialization lies outside intermediate region $D_{mid;k}$, then, with high probability, the iterate $W_k^{(t)}$ is controlled within an exponentially growing dynamics for $T^o_{;\,0:5}$ iterates before it either enters the intermediate region $D_{mid;k}$ or exits the cold region D_{cold}. As we will see later in the proof of Lemma 6, by putting all coordinates together we can show that $v^{(t)}$ rarely leaves the cold region D_{cold}, which implies that $v^{(t)}$ will enter the joint intermediate region $\cap_{k \geq [2;d]} D_{mid;k}$ within $T^o_{;\,0:5}$ iterates with high probability.

4. Additional Related Literature

The themes of ICA and tensor decomposition have been studied in numerous statistical and signal-processing literatures (Bach and Jordan, 2002; Chen and Bickel, 2006; Samworth and Yuan, 2012; Bonhomme and Robin, 2009; Eriksson and Koivunen, 2004; Hallin and Mehta, 2015; Hyvärinen et al., 2001; Hyvärinen and Oja, 1997; Hyvarinen, 1999; Hyvärinen and Oja, 2000; Ilmonen and Paindaveine, 2011; Kollo, 2008; Miettinen et al., 2015; Oja et al., 2006; Tichavsky et al., 2006; Wang and Lu, 2017; Ge and Ma, 2017). For a treatment from a spectral learning perspective (mainly for the deterministic scenario), we refer to the recently published monograph of Janzamin et al. (2019) and the bibliography therein. Recent literature studies the ICA setting in the context of speci c parametric families for independent component distributions and shows that parametric (Lee et al., 1999), semi-parametric (Hastie and Tibshirani, 2003; Chen and Bickel, 2006; Ilmonen and Paindaveine, 2011) and nonparametric (Bach and Jordan, 2002; Samarov and Tsybakov, 2004; Samworth and Yuan, 2012) models can be estimated via maximal likelihood estimation or minimization of mutual information between independent components. Our work focuses on a different type of contrast function based on tensor decomposition and kurtosis maximization, and hence our methodology is quite different from this line of work.

Our work is most closely related to Ge et al. (2015), which led to a general line of work on stochastic-gradient-based nonconvex optimization (Dauphin et al., 2014; Ge et al., 2015; Sun et al., 2015, 2017, 2018; Anandkumar and Ge, 2016; Jin et al., 2017, 2021, 2018; Lei et al., 2017; Allen-Zhu, 2018; Daneshmand et al., 2018; Fang et al., 2019; Cutkosky and Orabona, 2019; Cutkosky and Mehta, 2020) as well as the Riemannian manifold (Zhang and Sra, 2016; Zhang et al., 2016; Tripuraneni et al., 2018). A large family of nonconvex landscape has been studied in Mei et al. (2018), where uniform convergence of the empirical loss to the population loss is established. It is also related to work on recursive variance-reduced gradient methods for smooth optimization (Nguyen et al., 2017a,b; Fang et al., 2018; Zhou et al., 2020; Wang et al., 2019; Arjevani et al., 2020). In particular, we note the work of Ge et al. (2015); Jin et al. (2021); Daneshmand et al. (2018); Fang et al. (2019), who study the dynamics of SGD for optimizing generic functions. When applied to speci c statistical estimation problems, however, the results obtained by these methods can be coarse due to their neglect of speci c geometric features of the landscape.

The pioneering work of Ge et al. (2015) studied the convergence rate of SGD for minimizing a large class of nonconvex objectives de ned 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 rst-order gradient descent algorithms without adding noise injection are generally unknown and can be unfavorable (Lee et al., 2019; Du et al., 2017; Pemantle, 1990). Favorable results can be obtained under special spherical distributions for the noise or sophisticated procedures for avoidance of saddle points (Ge et al., 2015; Jin et al., 2017, 2021; Sun et al., 2015; Allen-Zhu, 2018; Fang et al., 2019).

Recent years have witnessed signi cant progress on computational and statistical aspects of low-rank representation methods. A number of recent papers (Carmon and Duchi, 2020; Ge et al., 2017; Li et al., 2018; Bai et al., 2018; Davis et al., 2018; Li and Bresler, 2018; Zhu et al., 2018; Ma et al., 2019; Chen et al., 2019; Gilboa et al., 2019; Kuo et al., 2019; Qu et al., 2019; Tan and Vershynin, 2019; Yang et al., 2019; Na et al., 2019; Chen et al., 2020; Lau et al., 2020; Li et al., 2020; Zhai et al., 2020) study the gradient-descent dynamics of of matrix factorization/completion, principal component pursuit, dictionary learning, phase retrieval, blind deconvolution, and many others, in the setting of batch or online (streaming) data. Notably, Carmon and Duchi (2020); Li et al. (2018) pursue a dynamics-based analysis to study gradient descent and its cubic-regularized variant for eigenvalue problems. Related work on ef cient convergence of Oja's online PCA iteration can be found in Jain et al. (2016); Allen-Zhu and Li (2017).Li et al. (2016); Wang and Wu (2020); Wang and Lu (2017), who study the online tensorial ICA method from the viewpoint of scaling limits and (stochastic) differential equation approximations. While these results provide valuable insights into our problem, straightforward translations to nonasymptotic convergence guarantees are not available, mainly due to the differential equation approximation being a weak convergence formulation instead of a strong one. Our re ned projected stochastic gradient analysis for online tensorial ICA provided in this paper is both dynamics-based and nonasymptotic, and we are able to prove that under some mild scaling conditions, random initialization provides suf cient deviation from the set of unstable stationary points such that a vanilla online tensorial ICA algorithm can be guaranteesdto achieve an sharp convergence rate.

5. Summary

We have studied the dynamics of an algorithm formulated as online stochastic approximation of (orthogonal) tensorial ICA. Our algorithm can be viewed as a method for optimizing a nonconvex objective of excess kurtosis in a given direction. We show that with properly chosen stepsizes and under mild scaling conditions our online tensorial ICA algorithm achieves a $\widetilde{O}(d/T)$-convergence rate, which is superior to the best existing analysis of this problem. Our algorithm requires no noise-injection steps or specially-designed loops for saddle-point avoidance, and our dynamics-based approach enjoys multiple advantages over existing analyses of online stochastic approximation for tensorial ICA estimation.

We believe that our analysis can generalize to a broader class of statistical estimation problems that can be cast as nonconvex stochastic optimization problems. Future directions include further improvements of the convergence rate and scaling conditions or justi cation of the impossibility (or minimax optimality) of such rates, analyzing the mini-batch stochastic approximation algorithm as well as the non-identical kurtosis case for ICA, and nally generalizing our analysis of

the dynamics of stochastic online algorithms to the nonorthogonal tensor decomposition case and over-parameterized cases.

Acknowledgments

We thank Wenlong Mou and Yuren Zhou for valuable discussions. This work was supported in part by the Mathematical Data Science program of the Office of Naval Research under grant number N00014-18-1-2764.

References

Zeyuan Allen-Zhu. Natasha 2: Faster non-convex optimization than SGD. In *Advances in Neural Information Processing Systems*, pages 2676–2687, 2018.

Zeyuan Allen-Zhu and Yuanzhi Li. First efficient convergence for streaming k-PCA: a global, gap-free, and near-optimal rate. In *The 58th Annual Symposium on Foundations of Computer Science*, 2017.

Animashree Anandkumar and Rong Ge. Efficient approaches for escaping higher order saddle points in non-convex optimization. In *Conference on Learning Theory*, pages 81–102, 2016.

Yossi Arjevani, Yair Carmon, John C Duchi, Dylan J Foster, Ayush Sekhari, and Karthik Sridharan. Second-order information in non-convex stochastic optimization: Power and limitations. In *Conference on Learning Theory*, pages 242–299. PMLR, 2020.

Francis R Bach and Michael I Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3:1–48, 2002.

Yu Bai, Qijia Jiang, and Ju Sun. Subgradient descent learns orthogonal dictionaries. In *International Conference on Learning Representations*, 2018.

Stéphane Bonhomme and Jean-Marc Robin. Consistent noisy independent component analysis. *Journal of Econometrics*, 149(1):12–25, 2009.

Yair Carmon and John C Duchi. First-order methods for nonconvex quadratic minimization. *SIAM Review*, 62(2):395–436, 2020.

Aiyou Chen and Peter J Bickel. Efficient independent component analysis. *The Annals of Statistics*, 34(6):2825–2855, 2006.

Yuxin Chen, Yuejie Chi, Jianqing Fan, and Cong Ma. Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval. *Mathematical Programming*, 176(1): 5–37, 2019.

Yuxin Chen, Yuejie Chi, Jianqing Fan, and Cong Ma. Spectral methods for data science: A statistical perspective. *arXiv preprint arXiv:2012.08496*, 2020.

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

Ashok Cutkosky and Harsh Mehta. Momentum improves normalized SGD. In International Conference on Machine Learning, pages 2260–2268. PMLR, 2020.

Ashok Cutkosky and Francesco Orabona. Momentum-based variance reduction in non-convex SGD. In Advances in Neural Information Processing Systems, pages 15236–15245, 2019.

Hadi Daneshmand, Jonas Kohler, Aurelien Lucchi, and Thomas Hofmann. Escaping saddles with stochastic gradients. In International Conference on Machine Learning, pages 1163–1172, 2018.

Yann N Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In Advances in Neural Information Processing Systems, pages 2933–2941, 2014.

Damek Davis, Dmitriy Drusvyatskiy, Kellie J MacPhee, and Courtney Paquette. Subgradient methods for sharp weakly convex functions. Journal of Optimization Theory and Applications, 179 (3):962–982, 2018.

Simon S Du, Chi Jin, Jason D Lee, Michael I Jordan, Barnabás Póczos, and Aarti Singh. Gradient descent can take exponential time to escape saddle points. In Advances in Neural Information Processing Systems, pages 1068–1078, 2017.

Jan Eriksson and Visa Koivunen. Identi ability, separability, and uniqueness of linear ICA models. IEEE Signal Processing Letters, 11(7):601–604, 2004.

Xiequan Fan, Ion Grama, and Quansheng Liu. Large deviation exponential inequalities for supermartingales. Electronic Communications in Probability, 17, 2012.

Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. SPIDER: Near-optimal non-convex optimization via stochastic path integrated differential estimator. In Advances in Neural Information Processing Systems, pages 686–696, 2018.

Cong Fang, Zhouchen Lin, and Tong Zhang. Sharp analysis for nonconvex SGD escaping from saddle points. In Conference on Learning Theory, pages 1192–1234. PMLR, 2019.

Alan Frieze, Mark Jerrum, and Ravi Kannan. Learning linear transformations. In Proceedings of 37th Conference on Foundations of Computer Science, pages 359–368. IEEE, 1996.

Rong Ge and Tengyu Ma. On the optimization landscape of tensor decompositions. In Advances in Neural Information Processing Systems, pages 3653–3663, 2017.

Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points – online stochastic gradient for tensor decomposition. In Conference on Learning Theory, pages 797–842, 2015.

Rong Ge, Chi Jin, and Yi Zheng. No spurious local minima in nonconvex low rank problems: A uni ed geometric analysis. In International Conference on Machine Learning, pages 1233–1242. PMLR, 2017.

Dar Gilboa, Sam Buchanan, and John Wright. Ef cient dictionary learning with gradient descent. In International Conference on Machine Learning, pages 2252–2259. PMLR, 2019.

Marc Hallin and Chintan Mehta. R-estimation for asymmetric independent component analysis. Journal of the American Statistical Association, 110(509):218–232, 2015.

Trevor Hastie and Rob Tibshirani. Independent components analysis through product density estimation. In Advances in Neural Information Processing Systems, pages 665–672, 2003.

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

Aapo Hyvärinen and Erkki Oja. A fast fixed-point algorithm for independent component analysis. Neural Computation, 9(7):1483–1492, 1997.

Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. Neural Networks, 13(4-5):411–430, 2000.

Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. Independent Component Analysis. John Wiley & Sons, 2001.

Pauliina Ilmonen and Davy Paindaveine. Semiparametrically efficient inference based on signed ranks in symmetric independent component models. The Annals of Statistics, 39(5):2448–2476, 2011.

Prateek Jain, Chi Jin, Sham M Kakade, Praneeth Netrapalli, and Aaron Sidford. Streaming PCA: Matching matrix Bernstein and near-optimal finite sample guarantees for Oja's algorithm. In Conference on Learning Theory, pages 1147–1164. PMLR, 2016.

Majid Janzamin, Rong Ge, Jean Kossai , and Anima Anandkumar. Spectral learning on matrices and tensors. Foundations and Trends in Machine Learning, 12(5-6):393–536, 2019.

Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. In International Conference on Machine Learning, pages 1724–1732. PMLR, 2017.

Chi Jin, Praneeth Netrapalli, and Michael I Jordan. Accelerated gradient descent escapes saddle points faster than gradient descent. In Conference on Learning Theory, pages 1042–1085, 2018.

Chi Jin, Praneeth Netrapalli, Rong Ge, Sham M Kakade, and Michael I Jordan. On nonconvex optimization for machine learning: Gradients, stochasticity, and saddle points. Journal of the ACM, 68(2):1–29, 2021.

Tõnu Kollo. Multivariate skewness and kurtosis measures with an application in ICA. Journal of Multivariate Analysis, 99(10):2328–2338, 2008.

Han-Wen Kuo, Yenson Lau, Yuqian Zhang, and John Wright. Geometry and symmetry in short-and-sparse deconvolution. In International Conference on Machine Learning, pages 3570–3580. PMLR, 2019.

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

Yenson Lau, Qing Qu, Han-Wen Kuo, Pengchang Zhou, Yuqian Zhang, and John Wright. Short and sparse deconvolution–a geometric approach. In International Conference on Learning Representations, 2020.

Jason D Lee, Ioannis Panageas, Georgios Piliouras, Max Simchowitz, Michael I Jordan, and Benjamin Recht. First-order methods almost always avoid strict saddle points. Mathematical Programming, 176(1):311–337, 2019.

Te-Won Lee, Mark Girolami, and Terrence J Sejnowski. Independent component analysis using an extended infomax algorithm for mixed subgaussian and supergaussian sources. Neural Computation, 11(2):417–441, 1999.

Lihua Lei, Cheng Ju, Jianbo Chen, and Michael I Jordan. Non-convex finite-sum optimization via SCSG methods. In Advances in Neural Information Processing Systems, pages 2348–2358, 2017.

Emmanuel Lesigne and Dalibor Volný. Large deviations for martingales. Stochastic Processes and Their Applications, 96(1):143–159, 2001.

Chris Junchi Li, Zhaoran Wang, and Han Liu. Online ICA: Understanding global dynamics of nonconvex optimization via diffusion processes. In Advances in Neural Information Processing Systems, pages 4967–4975, 2016.

Chris Junchi Li, Mengdi Wang, Han Liu, and Tong Zhang. Near-optimal stochastic approximation for online principal component estimation. Mathematical Programming, 167(1):75–97, 2018.

Xiao Li, Zhihui Zhu, Anthony Man-Cho So, and Rene Vidal. Nonconvex robust low-rank matrix recovery. SIAM Journal on Optimization, 30(1):660–686, 2020.

Yanjun Li and Yoram Bresler. Global geometry of multichannel sparse blind deconvolution on the sphere. In Advances in Neural Information Processing Systems, pages 1132–1143, 2018.

Cong Ma, Kaizheng Wang, Yuejie Chi, and Yuxin Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion, and blind deconvolution. Foundations of Computational Mathematics, 2019.

Song Mei, Yu Bai, and Andrea Montanari. The landscape of empirical risk for nonconvex losses. The Annals of Statistics, 46(6A):2747–2774, 2018.

Jari Miettinen, Sara Taskinen, Klaus Nordhausen, and Hannu Oja. Fourth moments and independent component analysis. Statistical Science, 30(3):372–390, 2015.

Sen Na, Zhuoran Yang, Zhaoran Wang, and Mladen Kolar. High-dimensional varying index coefficient models via Stein's identity. Journal of Machine Learning Research, 20:1–44, 2019.

Lam M Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. SARAH: A novel method for machine learning problems using stochastic recursive gradient. In International Conference on Machine Learning, pages 2613–2621, 2017a.

Lam M Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. Stochastic recursive gradient algorithm for nonconvex optimization. arXiv preprint arXiv:1705.07261, 2017b.

Hannu Oja, Seija Sirkiä, and Jan Eriksson. Scatter matrices and independent component analysis. Austrian Journal of Statistics, 35(2&3):175–189, 2006.

Robin Pemantle. Nonconvergence to unstable points in urn models and stochastic approximations. The Annals of Probability, pages 698–712, 1990.

Qing Qu, Xiao Li, and Zhihui Zhu. A nonconvex approach for exact and efficient multichannel sparse blind deconvolution. In Advances in Neural Information Processing Systems, pages 4015–4026, 2019.

Alexander Samarov and Alexandre Tsybakov. Nonparametric independent component analysis. Bernoulli, 10(4):565–582, 2004.

Richard J Samworth and Ming Yuan. Independent component analysis via nonparametric maximum likelihood estimation. The Annals of Statistics, 40(6):2973–3002, 2012.

James V Stone. Independent Component Analysis: A Tutorial Introduction. MIT press, 2004.

Ju Sun, Qing Qu, and John Wright. When are nonconvex problems not scary? arXiv preprint arXiv:1510.06096, 2015.

Ju Sun, Qing Qu, and John Wright. Complete dictionary recovery over the sphere I: Overview and the geometric picture. IEEE Transactions on Information Theory, 63(2):853–884, 2017.

Ju Sun, Qing Qu, and John Wright. A geometric analysis of phase retrieval. Foundations of Computational Mathematics, 18(5):1131–1198, 2018.

Yan Shuo Tan and Roman Vershynin. Online stochastic gradient descent with arbitrary initialization solves non-smooth, non-convex phase retrieval. arXiv preprint arXiv:1910.12837, 2019.

Petr Tichavsky, Zbynek Koldovsky, and Erkki Oja. 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, 2006.

Nilesh Tripuraneni, Mitchell Stern, Chi Jin, Jeffrey Regier, and Michael I Jordan. Stochastic cubic regularization for fast nonconvex optimization. In Advances in Neural Information Processing Systems, pages 2899–2908, 2018.

Aad van der Vaart and Jon A Wellner. Weak Convergence and Empirical Processes: With Applications to Statistics. Springer, 1996.

Vincent Q Vu and Jing Lei. Minimax sparse principal subspace estimation in high dimensions. The Annals of Statistics, 41(6):2905–2947, 2013.

Martin J Wainwright. High-Dimensional Statistics: A Non-Asymptotic Viewpoint, volume 48. Cambridge University Press, 2019.

Chuang Wang and Yue Lu. The scaling limit of high-dimensional online independent component analysis. In Advances in Neural Information Processing Systems, pages 6638–6647, 2017.

Yazhen Wang and Shang Wu. Asymptotic analysis via stochastic differential equations of gradient descent algorithms in statistical and computational paradigms. Journal of Machine Learning Research, 21(199):1–103, 2020.

Zhe Wang, Kaiyi Ji, Yi Zhou, Yingbin Liang, and Vahid Tarokh. SpiderBoost and momentum: Faster variance reduction algorithms. Advances in Neural Information Processing Systems, pages 2406–2416, 2019.

Zhuoran Yang, Lin F Yang, Ethan X Fang, Tuo Zhao, Zhaoran Wang, and Matey Neykov. Misspecified nonconvex statistical optimization for sparse phase retrieval. Mathematical Programming, 176(1):545–571, 2019.

Yuexiang Zhai, Zitong Yang, Zhenyu Liao, John Wright, and Yi Ma. Complete dictionary learning via ℓ^4-norm maximization over the orthogonal group. Journal of Machine Learning Research, 21 (165):1–68, 2020.

Hongyi Zhang and Suvrit Sra. First-order methods for geodesically convex optimization. Conference on Learning Theory, pages 1617–1638, 2016.

Hongyi Zhang, Sashank J Reddi, and Suvrit Sra. Riemannian SVRG: Fast stochastic optimization on Riemannian manifolds. In Advances in Neural Information Processing Systems, pages 4592–4600, 2016.

Dongruo Zhou, Pan Xu, and Quanquan Gu. Stochastic nested variance reduction for nonconvex optimization. Journal of Machine Learning Research, 21(103):1–63, 2020.

Zhihui Zhu, Yifan Wang, Daniel Robinson, Daniel Naiman, Rene Vidal, and Manolis C Tsakiris. Dual principal component pursuit: Improved analysis and efficient algorithms. In Advances in Neural Information Processing Systems, pages 2175–2185, 2018.

Appendix

In Appendix, §A proves the main results in the paper, §B and §C provide all secondary lemmas and their proofs for warm and uniform initialization analysis, separately. §D, §E and §F provide necessary tools including a reversed version of Gronwall's inequality, preliminaries and properties of Orlicz ψ_2-norm and a concentration inequality, all of which are theoretical building blocks of this paper. Finally, §G visualizes the tensorial ICA landscape and provides preliminary simulation results that validate our theory.

Notations Throughout this paper, we treat $L, \mu_4, \ldots$ 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 $\boldsymbol{A}$ or vector $\boldsymbol{v}$, $\boldsymbol{A}^\top$ and $\boldsymbol{v}^\top$ denote their transposes. For any vector $\boldsymbol{v}$, v_k denotes its kth coordinate. 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 $M < 1$ 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 $O, \widetilde{\Omega}, \widetilde{\Theta}$ to hide factors that are polylogarithmically dependent on dimension d, stepsize η, sample size T and inverse unsuccessful probability[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 $\boldsymbol{v}$, we use $\|\boldsymbol{v}\|$ to denote its Euclidean norm. For any integer n, we define set $[n] = \{1, \ldots, n\}$. Finally, we use S^c to denote the complement of set (or event) S.

Appendix A. Deferred Proofs of Main Results

This section includes the deferred proofs of the main results §2 and §3. In their order of appearance, §A.1 and §A.2 prove in sequel Lemma 3 and Theorem 4, and §A.3 and §A.4 prove the convergence Lemma 6 and its finite-sample error Theorem 7, respectively. All secondary lemmas are deferred to §B and §C.

A.1. Proof of Lemma 3

We use the key Lemma 5 to tightly estimate the dynamics in all coordinates and thereby obtain a proof of Lemma 3.

Proof [Proof of Lemma 3] We denote $\mathcal{H}_{3,L} = \bigcap_{k=2}^{d} \mathcal{H}_{k,5,L}$ as the intersection of events $\mathcal{H}_{k,5,L}$. Consider now the following $(d-1)$-dimensional vector

$$U_k^{(t)} \Big/ U_k^{(0)} \prod_{s=0}^{t-1} \Big(1 + \eta\big(3 - (v_1^{(s)})^2 - (v_k^{(s)})^2\big)\Big) : k = 2, [2, d] \Big. \tag{A.1}$$

Using Lemma 5, on the event $\mathcal{H}_{3,L} \cap (t \le T_{\star} \wedge T_{\times})$ we bound the Euclidean norm of (A.1) by

$$\sqrt{\sum_{k=2}^{d} \Big(U_k^{(t)} \Big/ U_k^{(0)} \prod_{s=0}^{t-1} \Big(1 + \eta\big(3 - (v_1^{(s)})^2 - (v_k^{(s)})^2\big)\Big)\Big)^2} \le \sqrt{d}\, 2C_{5,L} \log^{5/2} d \cdot \eta^{-1} \mu_4^{-4} (T_{\star})^{1/2}. \tag{A.2}$$

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

$$\sqrt{\sum_{k=2}^{d}\Big(U_k^{(t)}-U_k^{(0)}\prod_{s=0}^{t-1}\Big(1-\tfrac{3}{4}|(v_1^{(s)})^2-(v_k^{(s)})^2|\Big)\Big)^2}-\sqrt{\sum_{k=2}^{d}(U_k^{(t)})^2}\sqrt{\sum_{k=2}^{d}U_k^{(0)}\prod_{s=0}^{t-1}\Big(1-\tfrac{3}{4}|(v_1^{(s)})^2-(v_k^{(s)})^2|\Big)^2}; \tag{A.3}$$

due to triangle inequality of Euclidean norm. The definition of iterates $\{U_k^{(t)}\}$ in (2.11) implies that $\tan\angle(v^{(t)},e_1)=\sqrt{\sum_{k=2}^{d}(U_k^{(t)})^2}$ and the definition of stopping time $T_\star$ in (2.12) 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 $(t<T_\star)$. Combining this with (A.2) and (A.3), we obtain on the event $H_{3,L}\cap(t<T_\varepsilon\wedge T_\star)$

$$\begin{aligned}\tan\angle(v^{(t)},e_1) &= \sqrt{\sum_{k=2}^{d}(U_k^{(t)})^2}\\ &\sqrt{\sum_{k=2}^{d}(U_k^{(0)})^2}\Big(1-\tfrac{1}{3}|\lambda_4-\lambda_3|\Big)^t+\sqrt{d-2}\,C_{5,L}\log^{5/2-1}B\,\lambda_4^{-1}(T_\varepsilon)^{1/2}\\ &=\tan\angle(v^{(0)},e_1)\Big(1-\tfrac{1}{3}|\lambda_4-\lambda_3|\Big)^t+2C_{5,L}\log^{5/2-1}B\,\lambda_4^{-1}(d-2)T_\varepsilon)^{1/2}.\end{aligned} \tag{A.4}$$

The definition of T_ε in (2.2) along with $\log(1-\frac{1}{3}|\lambda_4-\lambda_3|)\le-\frac{1}{3}|\lambda_4-\lambda_3|$ gives

$$T_\varepsilon\le 1+\frac{3}{|\lambda_4-\lambda_3|}\log\frac{|\lambda_4-\lambda_3|}{B^8}\le 1+\frac{3(\lambda_1+1)}{|\lambda_4-\lambda_3|}\log\frac{|\lambda_4-\lambda_3|}{B^8}\le\varepsilon^{-1}; \tag{A.5}$$

where in the second inequality we use $\frac{|\lambda_4-\lambda_3|}{B^8}\le e$ and $\frac{1}{|\lambda_4-\lambda_3|}\ge 1$ implied by scaling condition (2.3). Using relation (A.5), we find that scaling condition (2.3) with constant $713C_{5,L}^2$ indicates

$$2C_{5,L}\log^{5/2-1}B\,\lambda_4^{-1}(d-2)T_\varepsilon)^{1/2}\le\sqrt{4C_{5,L}^2B^8\log^{5-1}\frac{3(\lambda_1+1)}{|\lambda_4-\lambda_3|}\,d\log\frac{|\lambda_4-\lambda_3|}{B^8}}\le\sqrt{\frac{12}{713}(\lambda_1+1)\,C_{3,L}\log^8(T_\varepsilon-1)\,\frac{B^8}{|\lambda_4-\lambda_3|}\,d\log\frac{|\lambda_4-\lambda_3|}{B^8}}\le\sqrt{\frac{12}{713}}; \tag{A.6}$$

where the elementary inequality $\log^{5-1}\le\log^8(T_\varepsilon-1)$ is applied due to $T_\varepsilon-1\ge 1$ and $\varepsilon\le e^{-1}$. Viewing (2.1) (equivalent to $v^{(0)}\in D_{warm}$ in (2.10)) and (A.4), we have for each t on the event $H_{3,L}\cap(T_\star\le T_\varepsilon)\cap(t<T_\star)$ that

$$\tan\angle(v^{(t)},e_1)\le\tan\angle(v^{(0)},e_1)+\sqrt{\frac{12}{713}}<\frac{1}{\sqrt{3}}+\frac{1}{\sqrt{2}}-\frac{1}{\sqrt{3}}=\frac{1}{\sqrt{2}}; \tag{A.7}$$

where in the first inequality we again apply $(y=3)|\gamma - 3| \le 1$ from (2.3). This further indicates that on event $\mathcal{H}_{3;L} \setminus (T_x \le T_{\star;\star})$, inequality $\tan \angle\big(v^{(T_x)}, e_1\big) < 1/\sqrt{2}$ holds, which contradicts with the fact that $v^{(T_x)} \in \mathcal{D}^c_{\text{warm-aux}}$ on the same event. Therefore, we have $\mathcal{H}_{3;L} \setminus (T_x \le T_{\star;\star}) = \emptyset$, and equivalently

$$\mathcal{H}_{3;L} = \mathcal{H}_{3;L} \setminus (T_x > T_{\star;\star}). \tag{A.8}$$

This implies that for all $t \in [0, T_{\star;\star}]$, (A.4) holds on the event $\mathcal{H}_{3;L}$. Plugging in the inequality involving $T_{\star;\star}$ in (A.5), we have on the event $\mathcal{H}_{3;L}$ that for all $t \in [0, T_{\star;\star}]$

$$\begin{aligned}
\tan \angle\big(v^{(t)}, e_1\big) &\le \tan \angle\big(v^{(0)}, e_1\big) + \frac{1}{3}|\gamma - 3|\, t + 2C_{5;L} B^4 \log^{5/2}(1/\delta)\, (d\, T_{\star;\star})^{1/2} \\
&\le \tan \angle\big(v^{(0)}, e_1\big) + \frac{1}{3}|\gamma - 3|\, t \\
&\quad + \sqrt{2}\sqrt{3}C_{5;L} \log^{5/2}(1/\delta)\, \frac{B^4}{|\gamma - 3|^{1/2}} \sqrt{d \log\Big(\frac{|\gamma - 3|}{B^8}\Big)}.
\end{aligned} \tag{A.9}$$

Letting the constant $C_{3;L} = 2\sqrt{3}C_{5;L}$, the scaling relation (2.9) and the above derivation (A.9) prove that (2.4) holds for all $t \in [0, T_{\star;\star}]$ on the event $\mathcal{H}_{3;L}$.

The only left is to estimate the probability of $\mathcal{H}_{3;L}$. Lemma 5 gives for each $k \in [2, d]$ the probability of event $\mathbb{P}(\mathcal{H}_{k;5;L}) \ge 1 - 6\exp(-12 + \frac{5184}{\log^{5/2}(1/\delta)})$, and hence elementary union bound calculation gives

$$\mathbb{P}(\mathcal{H}_{3;L}) = \mathbb{P}\Big(\bigcap_{k \in [2,d]} \mathcal{H}_{k;5;L}\Big) \ge 1 - 6d\exp\Big(-12 + \frac{5184}{\log^{5/2}(1/\delta)}\Big), \tag{A.10}$$

completing the whole proof of Lemma 3. ∎

A.2. Proof of Theorem 4

Now we turn to the proof of the finite-sample error Theorem 4. The idea is to apply Lemma 3 with appropriate stepsize (as in (2.5)) as well as an appropriate ε to obtain the finite-sample error bound.

Proof [Proof of Theorem 4]

(1) We first provide an upper bound on $T_{\varepsilon(T);\star}$. Under scaling condition (2.6) with constant $C_{4;T} = \max\{90C_{3;L}, 10\}$ and some constant $C'_{4;T} > 1$ to be determined later, we have $\frac{|\gamma - 3|^2}{9B^8}T \le \varepsilon(T)$. Plugging in $\varepsilon = \varepsilon(T)$ from (2.5) to relation (A.5), we have

$$\begin{aligned}
T_{\varepsilon(T);\star} &\le \frac{3(\alpha+1)}{|\gamma - 3|}\, \varepsilon(T)^{-1} \log\Big(\frac{|\gamma - 3|}{B^8}\, \varepsilon(T)^{-1}\Big) \\
&= \frac{3(\alpha+1)}{|\gamma - 3|}\, \frac{2|\gamma - 3|T}{9\log\frac{|\gamma - 3|^2}{9B^8}T} \log\Big(\frac{|\gamma - 3|}{B^8}\, \frac{2|\gamma - 3|T}{9\log\frac{|\gamma - 3|^2}{9B^8}T}\Big) \le \frac{2(\alpha+1)}{3}T.
\end{aligned} \tag{A.11}$$

(2) Next we provide a lower bound on $T_{(T);}$. By Taylor expansion, for all $x \in (0; 1=3]$ we know that

$$|\log(1 \quad x) + x| = \left| \sum_{n=2}^{1} \frac{x^n}{n} \right| \quad \frac{x^2}{2} \sum_{n=0}^{1} x^n \quad \frac{x^2}{2} \frac{1}{1 \quad 1=3} = \frac{3x^2}{4};$$

and hence

$$\frac{1}{\log(1 \quad x)} \quad \frac{1}{x + 3x^2=4} \quad \frac{1}{x + x=4} \quad \frac{4}{5x}: \tag{A.12}$$

From the de nition of $T_{;}$ in (2.2), for satisfying $|_4 \quad 3| = 3 \quad 1=3$ we have

$$T_{;} \quad \frac{\log \frac{|_4 \quad 3|}{B^8} \quad 1}{\log 1 \quad \frac{1}{3}|_4 \quad 3|} \quad \frac{12}{5|_4 \quad 3|} \quad ^1 \log \frac{|_4 \quad 3|}{B^8} \quad ^1 : \tag{A.13}$$

Under scaling condition (2.6), along with relation $\frac{B^4}{|_4 \quad 3|} \quad \frac{1}{8}$ given by Lemma 12 in Appendix B and $T \quad 100$, we have

$$\frac{(T)|_4 \quad 3|}{3} = \frac{3 \log \frac{2|_4 \quad 3|^2}{9B^8} T}{2T} < \frac{3 \log T}{T} \quad \frac{1}{3}; \quad \frac{2|_4 \quad 3|^2}{9B^8} T \quad e: \tag{A.14}$$

Plugging in $= (T)$ (as in (2.5)) to (A.13) and we obtain

$$\begin{aligned} T_{(T);} \quad & \frac{12}{5|_4 \quad 3|} \quad (T) \quad ^1 \log \frac{|_4 \quad 3|}{B^8} \quad (T) \quad ^1 \\ & \quad 0 \quad 1 \\ = \frac{12}{5} \quad & \frac{2T}{9 \log \frac{2|_4 \quad 3|^2}{9B^8} T} \log @ \frac{2|_4 \quad 3|^2 T}{9B^8 \log \frac{2|_4 \quad 3|^2}{9B^8} T} A \\ & \frac{8}{15} \quad \frac{T}{\log \frac{2|_4 \quad 3|^2}{9B^8} T} \quad \frac{1}{2} \log \frac{2|_4 \quad 3|^2}{9B^8} T \quad \frac{4}{15} T; \end{aligned} \tag{A.15}$$

where we use the elementary inequality $\log \frac{x}{\log x} \quad \frac{1}{2} \log x$ for all $x \quad e$.

(3) From (A.11) and (A.15) we know that $T \in [T_{(T);0:5}; T_{(T);4}]$. Here we will verify scaling condition (2.3) required in Lemma 3 under our setting. By choosing

$$36 + \frac{5184}{\log^5 \quad 1} \quad d ; \tag{A.16}$$

we have

$$T_{(T);1} \quad ^1 \quad C^0_{4;T} \quad ^1 dT; \tag{A.17}$$

where we de ne constant $C^0_{4;T} \quad (4=3) \ (36 + 5184) = 6960$ and use results $T_{(T);1} \quad 4T=3$, $e \quad ^1$ implied by (A.11), $\quad 1$.

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

$$22:5C_{3;L} \log^8(T_{(T);1} \quad ^1) \quad \frac{d \log \frac{2|_4 \quad 3|^2}{9B^8} T}{T} \log @ \frac{2|_4 \quad 3|^2 T}{9B^8 \log \frac{2|_4 \quad 3|^2}{9B^8} T} A \quad \frac{|_4 \quad 3|^2}{B^8};$$

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

$$C_{4;T}\log^8(C^0_{4;T}\quad{}^1dT)\quad\frac{d\log^2 T}{T}\qquad\frac{j\ {}_4\ \ 3j^2}{B^8};\tag{A.18}$$

which comes from (2.6) and constant de nition $C_{4;T}\quad 22{:}5C_{3;L}\quad 2^2\ =\ 90\,C_{3;L}$, because $T\quad 100$ and the relation (B.1) of $B; j\ {}_4\ \ 3j$ in Appendix Lemma 12 imply $T\quad\frac{2j\ {}_4\ 3j^2}{9B^8}$, and hence

$$1\quad\log\ \frac{2j\ {}_4\ \ 3j^2}{9B^8}T\qquad 2\log T:\tag{A.19}$$

To verify the second condition in (2.3), using scaling condition (2.6) in Theorem 4 and (A.19), we have

$$\frac{B^8}{j\ {}_4\ \ 3j}\ (T)=\frac{9B^8}{2j\ {}_4\ \ 3j^2}\ \frac{\log\ \frac{2j\ {}_4\ 3j^2}{9B^8}T}{T}\qquad\frac{9B^8}{j\ {}_4\ \ 3j^2}\ \frac{\log T}{T}<e\quad{}^1;$$

Note that we have already veri ed $j\ {}_4\ \ 3j\ (T)<1$ in (A.14). So far we have shown that, under scaling condition (2.6) in Theorem 4, for stepsize $=e\ (T)$ given in (2.5) the scaling condition (2.3) in Lemma 3 is guaranteed to hold, which justi es our following act on proving Theorem 4 with Lemma 3.

(4) Using warm initialization condition (2.1) in Lemma 3 and the de nition of $T_{;}$ given in (2.2), for all $t\ 2\ [T_{;\,0:5};T_{;\,4}]$ we have

$$\tan\backslash\ u^{(0)};a_i\quad 1\quad\frac{1}{3}j\ {}_4\ \ 3j^{\ t}\quad p\frac{1}{3}\ \frac{B^4}{j\ {}_4\ \ 3j^{1=2}}\ {}^{1=2}$$
$$p\frac{1}{3}\log^{5=2}\quad{}^1\ \frac{B^4}{j\ {}_4\ \ 3j^{1=2}}\qquad d\ \log\ \frac{j\ {}_4\ \ 3j}{B^8}\quad{}^1\ ;$$

where the first inequality comes from $m\quad T_{;\,0:5}$ and de nition of $T_{;}$ in (2.2), and the second inequality is due to $^1\quad e$ and $\frac{j\ {}_4\ \ 3j}{B^8}\quad{}^1\quad e$ given by scaling condition (2.3). Therefore, on the event $H_{3;L}$ we have for all $t\ 2\ [T_{;\,0:5};T_{;\,4}]$

$$\tan\backslash\ u^{(t)};a_i\qquad 3C_{3;L}\log^{5=2}\quad{}^1\ \frac{B^4}{j\ {}_4\ \ 3j^{1=2}}\qquad d\ \log\ \frac{j\ {}_4\ \ 3j}{B^8}\quad{}^1\ :\tag{A.20}$$

To nalize our proof, we plug in $=\ (T)$ to (A.20) and conclude from $t\ 2\ [T_{\ (T);0:5};T_{\ (T);4}]$ that there exists an event $H_{4;T}$ equivalent to $H_{3;L}$ with, due to (A.16), $P(H_{4;T})\quad 1\quad ;$ such

that on event $\mathcal{H}_{4;T}$ the following holds

$$\begin{aligned}
&\tan\angle\big(u^{(T)}, a_i\big) \\
&\quad \le 3C_{3;L}\log^{5/2}(d)\frac{B^4}{|j-3j'|^{1/2}}\sqrt{\frac{d}{T}(T)\log\frac{|j+3j'|}{B^8}(T)} \\
&\quad = \frac{9\sqrt{2}}{2}C_{3;L}\log^{5/2}(d)\frac{B^4}{|j-3j'|}\sqrt{\frac{d\log\frac{2j+3j^2}{9B^8}T}{T}\log\frac{2j+3j^2T}{9B^8\log\frac{2j+3j^2}{9B^8}T}} \\
&\quad \le C_{4;T}\log^{5/2}(C^0_{4;T}d^{-1})\frac{B^4}{|j-3j'|}\sqrt{\frac{d\log^2 T}{T}};
\end{aligned} \tag{A.21}$$

where in the last step we apply (A.19) $\log^{5/2}(\cdot) \le \log^{5/2}(C^0_{4;T}d^{-1})$ from (A.16), with constants $C_{4;T} = 9\sqrt{2}C_{3;L}$; $C^0_{4;T} = 6960$. This completes the whole proof of the theorem. ■

A.3. Proof of Lemma 6

In uniform initialization analysis, intuitively $v^{(t)}$ needs to enter intermediate region $\mathcal{D}_{\text{mid};k}$ first before we worry about its exit of the intermediate-auxilliary region $\mathcal{D}_{\text{mid-aux};k}$. For each coordinate $k \in [2;d]$, we upper bound the stopping time $T_{e;k}$ earlier defined in (2.12) using cold initialization Lemmas 10 and 11. For each coordinate $k \in [2;d]$, we view the iterative process $\{U^{(t)}_k = v^{(t)}_k/v^{(t)}_1\}$ as a Markov chain with v initialized in region $\mathcal{D}_{\text{mid-aux};k}$ shifted by $T_{c;k}$. We notice that process $\{U^{(t)}_k 1_{(t<T_1)}\}$ is bounded by 1 due to definition of $U^{(t)}_k; T_1$, and we have $U^{(t)}_k 1_{(t<T_1)} = U^{(t)}_k$ for all $(t < T_{w;k})$. Due to strong Markov property the intermediate initialization Lemma 9 applies to the shifted Markov chain.

Proof [Proof of Lemma 6] We let constant $C_{6;L} = \max\{256C^2_{11;L}; 476C^2_{9;L}\}$ in scaling condition (3.2) in Lemma 6.

(1) We start by making coordinate-wise analysis for each $k \in [2;d]$. For all $t \le T^o_{;0.5} \wedge T_{c;k} \wedge T_1$, on the event $\mathcal{H}_{10;L} \cap \mathcal{H}_{k;11;L}$, since $(v^{(t-1)}_1)^2 \ge 1/d$, by applying Lemmas 10 and 11 we have

$$\begin{aligned}
W^{(t)}_k &\le W^{(0)}_k + C_{11;L}\log^{5/2}(d)\frac{B^4}{|j-3j'|^{1/2}}\sqrt{d}\Big(1+\frac{1}{d}|j+3j'|\Big)^t\sqrt{8\log(d^{-1})\log d} \\
&\le C_{11;L}\log^{5/2}(d)\frac{B^4}{|j-3j'|^{1/2}}\sqrt{d}\Big(1+\frac{1}{d}|j+3j'|\Big)^t,
\end{aligned} \tag{A.22}$$

where in the last step we used the following inequality implied by scaling condition (3.2)

$$\sqrt{8\log(d^{-1})\log d} \le 2C_{11;L}\log^{5/2}(d)\frac{B^4}{|j-3j'|^{1/2}}\sqrt{d}. \tag{A.23}$$

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

$$1 + \frac{|\Delta|}{d}\Big|_0^{T^o_{\cdot,0.5}} |\Delta| \exp\Big(\frac{1}{2}\log\Big(\frac{|\Delta|}{B^8}\Big)^{-1} \frac{\log\big(1+\frac{|\Delta|}{d}\big)}{\log\big(1-\frac{|\Delta|}{2d}\big)}\Big)\Big(\frac{|\Delta|}{B^4}\Big)^{1/2}, \tag{A.24}$$

since $\frac{|\Delta|}{2d} \le 1/2$ under scaling condition (3.2). Hence on the event

$$\mathcal{H}_{0,L} \cap \mathcal{H}_{k,1,L} \cap (T_1 > T^o_{\cdot,0.5}) \cap (T_{c,k} > T^o_{\cdot,0.5}),$$

using (A.22), (A.23), (A.24) and $\alpha \in (0, e^{-1}]$, we have

$$W_k^{(T^o_{\cdot,0.5})} \le C_{1,L}\log^{5/2-1}\frac{B^4}{|\Delta|^{1/2}} d^{1/2}\Big(\frac{|\Delta|}{B^4}\Big)^{1/2} = C_{1,L}\log^{5/2-1} d \le 2, \tag{A.25}$$

which indicates $v^{(T^o_{\cdot,0.5})} \in \mathcal{D}_{\text{mid-aux},k}$, i.e. event $\mathcal{H}_{0,L} \cap \mathcal{H}_{k,1,L} \cap (T_1 > T^o_{\cdot,0.5}) \cap (T_{c,k} > T^o_{\cdot,0.5}) \subset (T_{c,k} \le T^o_{\cdot,0.5})$, and hence

$$\mathcal{H}_{0,L} \cap \mathcal{H}_{k,1,L} \cap (T_1 > T^o_{\cdot,0.5}) \subset (T_{c,k} \le T^o_{\cdot,0.5}). \tag{A.26}$$

(2) To study $v^{(t)}$ in the intermediate region, we first assume $T_{c,k} = 0$, that is, the initialization $v^{(0)} \in \mathcal{D}_{\text{mid-aux},k}$. For all $t \le T_{w,k} \wedge T^o_{\cdot}$, on the event $\mathcal{H}_{k,9,L}$, since $(v_1^{(t-1)})^2 \le (v_k^{(t-1)})^2 \le 1/(2d)$ and the following is guaranteed by scaling condition (3.2)

$$\frac{|\Delta|}{2d} \le 1, \qquad 2C_{9,L}B^4\log^{5/2-1} d^{1/2}(T^o_{\cdot})^{1/2} \le \sqrt{\frac{8}{476}}, \tag{A.27}$$

applying Lemma 9 we have

$$\begin{aligned} |U_k^{(t)}| &\le |U_k^{(0)}|\Big(1+\frac{|\Delta|}{2d}\Big)^t + 2C_{9,L}B^4\log^{5/2-1} d^{1/2}(T^o_{\cdot})^{1/2} \\ &< \frac{1}{\sqrt{3}}\Big(1+\sqrt{\frac{8}{476}}\Big) \le \frac{1}{\sqrt{2}}. \end{aligned} \tag{A.28}$$

Now we consider uniform initialization case. Using strong Markov property as discussed earlier, we have for all $t \in [T_{c,k}, T_{w,k} \wedge T^o_{\cdot}]$

$$U_k^{(t)} \le \frac{1}{\sqrt{2}}. \tag{A.29}$$

On the event

$$\mathcal{H}_{k,9,L} \cap \mathcal{H}_{0,L} \cap \mathcal{H}_{k,1,L} \cap (T_1 > T^o_{\cdot,0.5}) \cap (T_{w,k} \le T^o_{\cdot}),$$

we have already proven $T_{c,k} \le T^o_{\cdot}(\alpha > 0.5)$ in (A.26). Applying (A.29) with $t = T_{w,k} = T_{w,k} \wedge T^o_{\cdot}$, we obtain $|U_k^{(T_{w,k})}| \le 1/\sqrt{2}$, which leads to contradiction with the definition of $T_{w,k}$ in (3.9) and indicates $\mathcal{H}_{k,9,L} \cap \mathcal{H}_{0,L} \cap \mathcal{H}_{k,1,L} \cap (T_1 > T^o_{\cdot,0.5}) \cap (T_{w,k} \le T^o_{\cdot}) = \emptyset$, i.e.

$$\mathcal{H}_{k,9,L} \cap \mathcal{H}_{0,L} \cap \mathcal{H}_{k,1,L} \cap (T_1 > T^o_{\cdot,0.5}) \subset (T_{w,k} > T^o_{\cdot}). \tag{A.30}$$

(3) With (A.26) and (A.30) proven, we put all coordinates $k \in [2;d]$ together and define event

$$H_{6;L} \quad \bigcap_{k\in[2;d]} H_{k;9;L} \cap H_{10;L} \cap \bigcap_{k\in[2;d]} H_{k;11;L} :$$

On the event $H_{6;L} \cap (T_1 \quad T^o_{;0:5})$; for each coordinate $k \in [2;d]$ satisfying $T_1 \quad T^o_{;0:5} \wedge T_{c;k}$, we apply (A.22) with $t = T_1 = T^o_{;0:5} \wedge T_{c;k} \wedge T_1$ and obtain $W_k^{(T_1)} \quad 0$. Then due to the definition of T_1 in (3.15), there must exist $k \in [2;d]$ such that $W_k^{(T_1)} < 0$ and $T_{c;k} < T_1 \quad T^o_{;0:5}$. Recall that $W_k^{(t)} < 0$ is equivalent to $|U_k^{(t)}| > 1$. By definitions of stopping times $T_1; T_{w;k}$, we know the existence of $k \in [2;d]$ such that $T_{c;k} < T_{w;k} \quad T_1 \quad T^o_{;0:5}$ on the event $H_{6;L} \cap (T_1 \quad T^o_{;0:5})$. By applying (A.29) with $t = T_{w;k} = T_{w;k} \wedge T^o_{;}$, we find $|U_k^{(T_{w;k})}| \quad 1=\sqrt{2}$, which contradicts with the definition of $T_{w;k}$ and indicates that $H_{6;L} \cap (T_1 \quad T^o_{;0:5}) = ?$, i.e.

$$H_{6;L} \quad (T_1 > T^o_{;0:5}): \tag{A.31}$$

Combining (A.26) and (A.30) for each $k \in [2;d]$ along with (A.31), we have

$$H_{6;L} \quad \sup_{2\ k\ d} T_{c;k} \quad T^o_{;0:5} \quad \cap \quad \inf_{2\ k\ d} T_{w;k} > T^o_{;} \quad \cap (T_1 > T^o_{;0:5}): \tag{A.32}$$

(4) With (A.32) ready at hand, on the event $H_{6;L}$, we apply Lemma 9 for each coordinate $k \in [2;d]$ and all $t \in [T^o_{;0:5}; T^o_{;}]$ to obtain

$$\sqrt{\sum_{k=2}^{d} \Big(U_k^{(t)} \quad U_k^{(T^o_{;0:5})} \Big)^2} \quad \sqrt{\sum_{s=T^o_{;0:5}}^{t} \Big(1 \quad \tfrac{|\gamma|}{4} \tfrac{3|\gamma|}{4} \big((v_1^{(s)})^2 \quad (v_k^{(s)})^2\big)^2 \Big)} \quad 2C_{9;L} \log^{5=2} \quad {}^1 B^4 d \, (T^o_{;})^{1=2}: \tag{A.33}$$

On the event $H_{6;L}$, we have $T_{c;k} \quad T^o_{;0:5}$, $(v_1^{(s)})^2 \quad (v_k^{(s)})^2 \quad 1=(2d)$, $U_k^{(T_{c;k})} \quad 1$ for all $k \in [2;d]$, $s \in [T^o_{;0:5}; T^o_{;})$. Since the left hand of (A.33) is the norm of two vectors subtraction, we use the triangle inequality of Euclidean norms to lower bound it as

$$\sqrt{\sum_{k=2}^{d} \Big(U_k^{(t)} \quad U_k^{(T^o_{;0:5})} \Big)^2} \quad \sqrt{\sum_{s=T^o_{;0:5}}^{t} \Big(1 \quad \tfrac{|\gamma|}{4} \tfrac{3|\gamma|}{4} \big((v_1^{(s)})^2 \quad (v_k^{(s)})^2\big)^2 \Big)}$$
$$\sqrt{\sum_{k=2}^{d} (U_k^{(t)})^2} \quad \sqrt{\sum_{k=2}^{d} \Big(U_k^{(T^o_{;0:5})} \Big)^2} \quad \sqrt{\sum_{s=T^o_{;0:5}}^{t} \Big(1 \quad \tfrac{|\gamma|}{4} \tfrac{3|\gamma|}{4} \big((v_1^{(s)})^2 \quad (v_k^{(s)})^2\big)^2 \Big)}$$
$$\tan \angle \big(v^{(t)}; e_1\big) \quad \sqrt{d} \quad 1 \quad \tfrac{|\gamma|}{\sqrt{2d}} \tfrac{3|\gamma|}{4} \quad t \quad T^o_{;0:5}: \tag{A.34}$$

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

$$\frac{|\gamma| \quad 3|\gamma|}{2d} \quad 1; \qquad \frac{B^8}{|\gamma| \quad 3|\gamma|} \quad e^{\ 1}: \tag{A.35}$$

Since $\log(1-\frac{1}{2d}|\lambda_4-\lambda_3|) \le -\frac{1}{2d}|\lambda_4-\lambda_3|$ and (A.35) holds, for each positive ε we have relation

$$T^{o}_{\varepsilon} \le 1+\frac{2d}{|\lambda_4-\lambda_3|}\log\frac{|\lambda_4-\lambda_3|}{B^8}-1-\frac{2(\varepsilon+1)d}{|\lambda_4-\lambda_3|}\log\frac{|\lambda_4-\lambda_3|}{B^8}-1: \tag{A.36}$$

Combining (A.33), (A.34) together and using relation (A.36), we have

$$\tan\angle\big(v^{(t)},e_1\big) \le \sqrt{d}\,(1-\tfrac{1}{2d}|\lambda_4-\lambda_3|)^{t-T^{o}_{\varepsilon}}\tan\theta_{0.5} + \sqrt{\varepsilon+1}\,C_{6,L}\log^{5/2}\frac{B^4}{|\lambda_4-\lambda_3|^{1/2}}\sqrt{d^3\log\frac{|\lambda_4-\lambda_3|}{B^8}-1}; \tag{A.37}$$

where constant $C_{6,L} \ge 2\sqrt{2}C_{9,L}$. To complete proof of Lemma 6, we provide a lower bound on probability of event $H_{6,L}$ by taking union bound

$$\mathbb{P}(H_{6,L}) \ge 1-\sum_{k=2}^{d}\mathbb{P}(H^{c}_{k,9,L})-\mathbb{P}(H^{c}_{10,L})-\sum_{k=2}^{d}\mathbb{P}(H^{c}_{k,11,L}) \ge 1-6\exp(-27)-\frac{10368}{\log^5 B}d^3.$$

Applying the scaling relation (2.9) of $\{u^{(t)}\}_{t\ge 0}$ and $\{v^{(t)}\}_{t\ge 0}$ to (A.37) on the event H completes the proof of (3.3), and hence Lemma 6.

■

A.4. Proof of Theorem 7

Now we are ready to derive the finite-sample error bound and prove Theorem 7.

Proof [Proof of Theorem 7]

(1) Analogous to the proof of Theorem 4, we sharply upper- and lower-bound the rescaled time $T^{o}_{\varepsilon(T)}$. Using relation $\frac{B^4}{|\lambda_4-\lambda_3|}\ge\frac{1}{8}$ in Lemma 12 and elementary inequality $\frac{\log T}{T}\le\frac{1}{20}$, we have

$$\frac{|\lambda_4-\lambda_3|}{2d}\varepsilon(T) = \frac{2}{T}\log\frac{|\lambda_4-\lambda_3|^2}{4B^8d}T \ge \frac{1}{3}: \tag{A.38}$$

Since $\frac{1}{\log(1-x)}\le\frac{1}{x}$ for $x\in(0,1)$, similar to (A.36) we have

$$\begin{aligned} T^{o}_{\varepsilon(T)} &\le 1+\frac{2d}{|\lambda_4-\lambda_3|}\varepsilon(T)-1\log\frac{|\lambda_4-\lambda_3|}{B^8}\varepsilon(T)-1 \\ &= 1+\frac{1}{2\log\frac{|\lambda_4-\lambda_3|^2}{4B^8d}T}T\log\Big(\frac{|\lambda_4-\lambda_3|^2T}{4B^8d\log\frac{|\lambda_4-\lambda_3|^2}{4B^8d}T}\Big) \\ &= 1+\frac{1}{2}T\,\frac{\log\frac{|\lambda_4-\lambda_3|^2}{4B^8d}T-\log\log\frac{|\lambda_4-\lambda_3|^2}{4B^8d}T}{\log\frac{|\lambda_4-\lambda_3|^2}{4B^8d}T}-\frac{\varepsilon+1}{2}T-1; \end{aligned} \tag{A.39}$$

where the last inequality holds due to $\frac{|\lambda_4-\lambda_3|^2}{4B^8d}T\ge e$ under scaling condition (3.2) with constants $C_{7,T}\ge 96C_{6,L}$; $C'_{7,T}=10425$.

On the lower bound side, previously in (A.12) we show that $\frac{1}{\log(1-x)} \geq \frac{4}{5x}$ for all $x \in (0, 1/3]$, which can be applied to $T^o_{(T),j}$ by replacing x with $\frac{j^4 - 3j}{2d}(T)$ due to (A.38). By noticing that $T \geq 100$ and scaling condition (3.4) guarantee the following inequalities

$$\frac{4\log T}{T} \leq \frac{1}{3}, \qquad \frac{8B^8}{j^4 - 3j^2}\sqrt{\frac{d\log T}{T}} \leq 1, \tag{A.40}$$

we have

$$T^o_{(T),j} \geq \frac{\log\frac{j^4-3j}{B^8}(T)^{-1}}{\log\left(1-\frac{j^4-3j}{2d}(T)\right)} \geq \frac{8\,d\log\frac{j^4-3j}{B^8}(T)^{-1}}{5j^4 - 3j}(T) \geq \frac{2T}{5\log\frac{j^4-3j^2}{4B^8d}T}\log\left(\frac{j^4-3j^2T}{4B^8d\log\frac{j^4-3j^2}{4B^8d}T}\right) \geq \frac{1}{5}T, \tag{A.41}$$

where in the last step we use the elementary inequality $\frac{x}{\log x} \geq \frac{1}{2}\log x$ for all $x \geq e$, since $\frac{j^4-3j^2}{4B^8d}T \geq e$ is satisfied under scaling condition (3.2).

(2) From (A.39) and (A.41) we find that $T \in [T^o_{(T),1}+1, T^o_{(T),5}]$. By letting $\Gamma = 57 + \frac{10368}{\log^5 2}d$ along with (A.39) we have

$$T^o_{(T),1} \geq \frac{1}{C^0_{7,T}}\int_1^T \frac{1}{t}dt,$$

for positive, absolute constant $C^0_{7,T} \leq 10425$ due to $T^o_{(T),1} \geq T$ given by (A.39) and e^{-1} guaranteed by $\delta \leq 1/4$.

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

$$24C_{6,L}\log^8(C^0_{7,T}\int_1^T \frac{1}{t}dt)\frac{B^8}{j^4-3j^2}\frac{d^3\log^2 d\log\frac{j^4-3j^2T}{4B^8d}}{T}\log\left(\frac{j^4-3j^2T}{4B^8d\log\frac{j^4-3j^2}{4B^8d}T}\right)^2 \Big/ \log^2 2 \leq 1. \tag{A.42}$$

From Lemma 12 we have $\frac{B^4}{j^4-3j} \leq \frac{1}{8}$. Along with scaling condition (3.4) and $C_{7,T} \geq 96C_{6,L}$, we have $T/d \geq 16$ and hence

$$1 \geq \log\frac{j^4-3j^2T}{4B^8d} \geq 2\log(T/d), \tag{A.43}$$

implying (A.42) holds under scaling condition (3.4).

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

$$\frac{B^8}{j^4-3j}\Gamma(T) = \frac{4B^8d\log\frac{j^4-3j^2}{4B^8d}T}{j^4-3j^2T}\frac{8B^8d\log T}{j^4-3j^2T} < e^{-1},$$

and

$$\Gamma(T) = \frac{4d\log\frac{j^4-3j^2}{4B^8d}T}{j^4-3jT}\frac{8\log(T/d)}{j^4-3j(T/d)} < \frac{1}{j^4-3j},$$

due to (A.43) and the elementary inequality $\frac{8\log x}{x} < 1$ for all $x \geq 100$.
Therefore, all scaling conditions required in Lemma 6 are satisfied by scaling condition (3.4) in Theorem 7.

(3) Using $B^8 = |\lambda_4 - \lambda_{3j}| \vee e^{-1}$ and $\varepsilon \leq e^{-1}$ given by (3.2) and $T^o_{\cdot,1} + 1 \leq T^o_{\cdot,0.5} \leq T^o_{\cdot,0.5}$ following its definition (3.1), for all $t \in [T^o_{\cdot,1} + 1, T^o_{\cdot,5}]$, on the event $H_{6,L}$ we have

$$\sqrt{d}\Big(1 - \frac{1}{2d}|\lambda_4 - \lambda_{3j}|^{t - T^o_{\cdot,0.5}}\Big) \leq \sqrt{d}\Big(1 - \frac{1}{2d}|\lambda_4 - \lambda_{3j}|^{T^o_{\cdot,0.5}} \frac{B^4}{|\lambda_4 - \lambda_{3j}|^{1/2}}\Big) \leq \sqrt{d} - (3\sqrt{6})C_{6,L}\log^{5/2}\Big(\frac{B^4}{|\lambda_4 - \lambda_{3j}|^{1/2}}\Big)\sqrt{d^3\log\frac{|\lambda_4 - \lambda_{3j}|}{B^8}}.$$

Therefore, from Lemma 6, on the event $H_{6,L}$ we have for all $t \in [T^o_{\cdot,1} + 1, T^o_{\cdot,5}]$ that

$$\tan\angle\big(u^{(t)}, a_j\big) \leq 3C_{6,L}\log^{5/2}\Big(\frac{B^4}{|\lambda_4 - \lambda_{3j}|^{1/2}}\Big)\sqrt{d^3\log\frac{|\lambda_4 - \lambda_{3j}|}{B^8}}. \quad \text{(A.44)}$$

Plugging in our choice of $\varepsilon = 5$ and $\delta(T) = \frac{4d\log\frac{|\lambda_4 - 3j^2|}{4B^8 d}T}{|\lambda_4 - 3j|T}$ to (A.44) and using (A.43), we know that there exists an event $H_{7,T} \subset H_{6,L}$ with $P(H_{7,T}) \geq 1 - 4\varepsilon$ such that on event $H_{7,T}$ we have

$$\tan\angle\big(u^{(T)}, a_j\big) \leq C_{7,T}\log^{5/2}(C^0_{7,T}T^d)\frac{B^4}{|\lambda_4 - \lambda_{3j}|}\sqrt{\frac{d^4\log^2 T}{T}},$$

where constants $C_{7,T} = 12C_{6,L}$; $C^0_{7,T} = 10425$.

■

Appendix B. Secondary Lemmas in Warm Initialization Analysis

For notational simplicity, we denote $v = v^{(t-1)}$ and $Y = Y^{(t)}$. We first provide a lemma on Orlicz ψ_2-norm of $v^\top Y$ and the relation between B and λ_4.

Lemma 12 Let Assumption 2 hold. For each rotated observation Y and any unit vector v, we have Orlicz ψ_2-norm $\|v^\top Y\|_{\psi_2} \leq B$ and the following relation of B and λ_4

$$\frac{B^4}{|\lambda_4 - \lambda_{3j}|} \geq \frac{1}{8}. \quad \text{(B.1)}$$

With the bound on Orlicz ψ_2-norm given in Lemma 12 and $T_{\cdot,1}$ defined in (2.2), we introduce truncation barrier parameter

$$B = B\log^{1/2}(T_{\cdot,1}\varepsilon^{-1}), \quad \text{(B.2)}$$

where $\varepsilon \in (0, e^{-1}]$ 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 as

$$T_{B,k} = \inf\Big\{t \geq 1 : v^{(t-1)\top}Y^{(t)} > B \text{ or } Y_1^{(t)} > B \text{ or } Y_k^{(t)} > B\Big\}; \quad \text{(B.3)}$$

B.1. Proof of Lemma 5

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

$$Q_{U;k}^{(t)} \triangleq U_k^{(t)} - U_k^{(t-1)} \quad \text{sign}(\gamma_4 - 3) \quad (v^\top Y)^3 v_1^2 (v_1 Y_k - v_k Y_1). \tag{B.4}$$

Lemma 13 Let B be any positive value. For each coordinate $k \geq 2$ $[2, d]$ and any $t \geq 1$, on the event

$$H_{k;13;L}^{(t)} \triangleq |v^\top Y| \leq B, |Y_1| \leq B, |Y_k| \leq B, v_1^2 \geq 1/d, v_1^2 \geq v_k^2, \tag{B.5}$$

for stepsize $\eta \leq 1/(2B^4 d^{1/2})$ we have $|Q_{U;k}^{(t)}| \leq 4B^8 \eta^2 v_1^2$.

Lemma 14 Let Assumption 2 hold. For each coordinate $k \geq 2$ $[2, d]$ and any $t \geq 1$, we have

$$\mathbb{E}\left[\text{sign}(\gamma_4 - 3) \quad (v^\top Y)^3 v_1^2 (v_1 Y_k - v_k Y_1) \mid \mathcal{F}_{t-1}\right] = |\gamma_4 - 3| (v_1^2 - v_k^2) U_k^{(t-1)}. \tag{B.6}$$

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

$$e_k^{(t)} \triangleq \text{sign}(\gamma_4 - 3) \quad (v^\top Y)^3 v_1^2 (v_1 Y_k - v_k Y_1) + |\gamma_4 - 3| (v_1^2 - v_k^2) U_k^{(t-1)} \tag{B.7}$$

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

Lemma 15 For each coordinate $k \geq 2$ $[2, d]$ and any $t \geq 1$, $U_k^{(t)}$ has linear representation

$$U_k^{(t)} = U_k^{(0)} \quad |\gamma_4 - 3| \sum_{s=0}^{t-1} (v_1^{(s)})^2 \quad (v_k^{(s)})^2 \; U_k^{(s)} + \sum_{s=1}^{t} Q_{U;k}^{(s)} + \sum_{s=1}^{t} e_k^{(s)}. \tag{B.8}$$

Lemma 16 Let Assumption 2 hold and initialization $v^{(0)} \geq D_{\text{warm}}$. Let $\eta \geq (0, e^{-1}]$ and ζ be any fixed positive. For each coordinate $k \geq 2$ $[2, d]$, there exists an event $H_{k;16;L}$ satisfying

$$\mathbb{P}(H_{k;16;L}) \geq 1 - 6 + \frac{5184}{\log^{5-1}} ;$$

such that on the event $H_{k;16;L}$ the following concentration result holds

$$\max_{1 \leq t \leq T_{\eta,\zeta} \wedge T_\times} \sum_{s=1}^{t} e_k^{(s)} \leq C_{16;L} \log^{5/2-1} B^4 (T_{\eta,\zeta})^{1/2};$$

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

Lemma 17 Let $\eta \geq (0, e^{-1}]$ and ζ be any fixed positive. For each coordinate $k \geq 2$ $[2, d]$, we have

$$\mathbb{P}\left[T_{B,k} \leq T_{\eta,\zeta}\right] \leq 6(\zeta + 1) : \tag{B.9}$$

With the above secondary lemmas at hand, we are now ready to prove Lemma 5.

Proof [Proof of Lemma 5] We recall the de nition of stopping time T_X in (2.12). Because stepsize $1{=}(2B^4 d^{1=2})$ holds under scaling condition (2.3), on the event $(T_{;} \wedge T_X) \setminus (T_{B\,;k} > T_{;})$ $H^{(t)}_{k;13;L}$, we have $v_1^{\,2} \quad \frac{3}{2}$, and applying Lemma 13 gives

$$
|Q^{(t)}_{U;k}| \quad 4B^8\ {}^2 v_1^{\,2} \quad 6B^8\ {}^2 \log^4(T_{;1}\ {}^1):
$$

We de ne event $H_{k;5;L}$ $(T_{B\,;k} > T_{;}) \setminus H_{k;16;L}$. Applying Lemmas 15 and 16, on the event $H_{k;5;L}$ for all t $T_{;} \wedge T_X$ we have

$$
\begin{aligned}
U^{(t)}_k \quad U^{(0)}_k + \quad | \ 4 \quad 3| \sum_{s=0}^{t\ 1} \ (v_1^{(s)})^2 \quad (v_k^{(s)})^2 \ U^{(s)}_k \\
T_{;} \quad 6B^8\ {}^2\log^4(T_{;1}\ {}^1) + C_{16;L} \log^{5=2} \ {}^1\ B^4 \quad (T_{;})^{1=2}:
\end{aligned}
\tag{B.10}
$$

Scaling condition (2.3) and de nition of $T_{;}$ in (2.2) imply that

$$
B^4 \ (T_{;})^{1=2} \log^4(T_{;1}\ {}^1) \quad \log^{5=2}\ {}^1; \tag{B.11}
$$

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

$$
U^{(t)}_k \quad U^{(0)}_k + \quad | \ 4 \quad 3| \sum_{s=0}^{t\ 1} \ (v_1^{(s)})^2 \quad (v_k^{(s)})^2 \ U^{(s)}_k \qquad C_{5;L} \log^{5=2} \ {}^1\ B^4 \quad (T_{;})^{1=2};
$$

where constant $C_{5;L}$ $C_{16;L} + 6$. Scaling condition (2.3) implies $| \ 4 \quad 3|((v_1^{(s)})^2 \quad (v_k^{(s)})^2)$ 2 $[0;1)$ for all $s < T_{;} \wedge T_X$. Using reversed Gronwall Lemma 25, on the event $H_{k;5;L}$ we have the following holds for all t $T_{;} \wedge T_X$

$$
U^{(t)}_k \quad U^{(0)}_k \prod_{s=0}^{t\ 1} \Big[1 \quad | \ 4 \quad 3| \ (v_1^{(s)})^2 \quad (v_k^{(s)})^2 \Big] \qquad 2C_{5;L} \log^{5=2} \ {}^1\ B^4 \quad (T_{;})^{1=2}:
$$

To complete proof of Lemma 5, we apply Lemmas 16 and 17 and take union bound to obtain

$$
P(H_{k;5;L}) \quad 1 \quad P(T_{B\,;k} \quad T_{;}) \quad P(H^c_{k;16;L}) \quad 1 \quad 6 \ + 12 + \frac{5184}{\log^5 \ {}^1} \ :
$$

■

B.2. Proof of Secondary Lemmas

Proof [Proof of Lemma 12]

(1) Because random vector Y is a permutation of Z, from Assumption 2 we know that each Y_i is sub-Gaussian with parameter $\sqrt{3{=}8}B$. By independence of Y_i, for any unit vector v and all 2 R we have

$$
E\exp \quad v^> Y \quad = \prod_{i=1}^{d} E\exp(\ v_i Y_i) \prod_{i=1}^{d} \exp\left(\frac{2 v_i^2}{2} \quad \left(\sqrt{\frac{3}{8}} B \right)^2 \right) = \exp\left(\frac{2}{2} \quad \left(\sqrt{\frac{3}{8}} B \right)^2 \right);
$$

which implies that $v^\top Y$ 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 $\mathbb{E}\exp\left(\frac{X^2}{2\sigma^2}\right) \le \frac{1}{1-\lambda}$ for all $\lambda \in [0,1)$. By choosing $\lambda = 3/4$, $\sigma = \sqrt{3/8}B$ and $X = v^\top Y$, we have

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

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

(2) Applying Markov's inequality to (B.12) with $v = e_i$ gives

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

Hence we have

$$\mu_4 = \mathbb{E}Y_i^4 = \int_0^\infty \mathbb{P}(Y_i^2 \ge t)\, 2t\, dt \le 4\int_0^\infty \exp\left(-\frac{t}{B^2}\right) t\, dt = 4B^4;$$

and hence $\frac{B^4}{|\mu_4 - 3|} \ge \frac{4}{4|\mu_4 - 3|} \ge \frac{1}{8}$ holds due to $\mu_4 = \mathbb{E}Y_i^4 \ge (\mathbb{E}Y_i^2)^2 = 1$. ∎

Proof [Proof of Lemma 13] Let $s = \operatorname{sign}(\mu_4 - 3)$ for notational simplicity. We recall definitions of iteration $U_k^{(t)}$ in (2.11) and random variable $Q_{U,k}^{(t)}$ in (B.4). Using update formula (2.8), we have

$$U_k^{(t)} - U_k^{(t-1)} = \frac{v_k + s(v^\top Y)^3 Y_k}{v_1 + s(v^\top Y)^3 Y_1} - \frac{v_k}{v_1} = \frac{v_k + s(v^\top Y)^3 Y_k - v_1 - v_1 + s(v^\top Y)^3 Y_1 v_k}{(v_1 + s(v^\top Y)^3 Y_1) v_1}$$

$$= s\left(1 + s(v^\top Y)^3 \frac{Y_1}{v_1}\right)^{-1} (v^\top Y)^3 v_1^{-2}(v_1 Y_k - v_k Y_1).$$

Along with (B.4), we obtain

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

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

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

On the event $\mathcal{H}_{k;13;L}^{(t)}$ defined earlier in (B.5), since $s(v^\top Y)^3 Y_1/v_1 \le B^4 d^{1/2} \le 1/2$, we have

$$|Q_{U,k}^{(t)}| \le \left|1 + s(v^\top Y)^3 \frac{Y_1}{v_1}\right|^{-1} \left|1 - \left(1 + s(v^\top Y)^3 \frac{Y_1}{v_1}\right)^{-1}\right| |v^\top Y|^3 v_1^{-2} |v_1 Y_k - v_k Y_1|$$

$$\le 2\left|s(v^\top Y)^3 \frac{Y_1}{v_1}\right| |v^\top Y|^3 v_1^{-2} |v_1 Y_k - v_k Y_1|$$

$$= 2|v_1^{-2}| |Y_1| |v^\top Y|^6 \left|Y_k - \frac{v_k}{v_1} Y_1\right| \le 4B^8 |v_1^{-2}|.$$

■

Proof [Proof of Lemma 14] Under Assumption 2, for all $k \in [d]$ we have

$$\mathbb{E}\left[(v^\top Y)^3 Y_k \mid \mathcal{F}_{t-1}\right] = -4v_k^3 + 3v_k \sum_{i=1, i\neq k}^{d} v_i^2 = (-4 - 3)v_k^3 + 3v_k. \tag{B.13}$$

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

$$\begin{aligned}
&\mathbb{E}\left[\operatorname{sign}(-4 - 3)\, (v^\top Y)^3 v_1^{-2}(v_1 Y_k - v_k Y_1) \mid \mathcal{F}_{t-1}\right] \\
&= \operatorname{sign}(-4 - 3)\, v_1^{-2}\left(-4 - 3)v_1 v_k^3 + 3v_1 v_k - (-4 - 3)v_k v_1^3 - 3v_k v_1\right) \\
&= -|-4 - 3|\,(v_1^2 - v_k^2)U_k^{(t-1)}.
\end{aligned}$$

■

Proof [Proof of Lemma 15] From definitions (B.4) and (B.7), by applying (B.6) in Lemma 14,

$$U_k^{(s)} - U_k^{(s-1)} = -|-4 - 3|\,(v_1^{(s-1)})^2 - (v_k^{(s-1)})^2\, U_k^{(s-1)} + Q_{U,k}^{(s)} + e_k^{(s)}. \tag{B.14}$$

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

Proof [Proof of Lemma 16] Under Assumption 2, we apply (E.1) along with Lemmas 12 and 27 to obtain

$$\begin{aligned}
&\|v_1^{-2}(v^\top Y)^3(v_1 Y_k - v_k Y_1)\|_{\psi_{1/2}} \le \|v_1^{-2}\, k(v^\top Y)^2 k_{\psi_1}\, k(v^\top Y)(v_1 Y_k - v_k Y_1) k_{\psi_1} \\
&\le \|v_1^{-2}\, kv^\top Y k_{\psi_2}^3\, kv_1 Y_k - v_k Y_1 k_{\psi_2} \\
&\le |v_1|^{-1}\, kv^\top Y k_{\psi_2}^3 \left(k Y_k k_{\psi_2} + |v_k / v_1| k Y_1 k_{\psi_2}\right) \le B^4 |v_1|^{-1}(1 + |v_k / v_1|).
\end{aligned}$$

Recall the definition of martingale difference sequence $\{e_k^{(t)}\}_{t\ge 1}$ in (B.7). By applying Lemma 28, we have

$$\begin{aligned}
\|e_k^{(t)}\|_{\psi_{1/2}} &= \|v_1^{-2}(v^\top Y)^3(v_1 Y_k - v_k Y_1) - \mathbb{E}\left[v_1^{-2}(v^\top Y)^3(v_1 Y_k - v_k Y_1)\right]\|_{\psi_{1/2}} \\
&\le C_{28;L}' \|v_1^{-2}(v^\top Y)^3(v_1 Y_k - v_k Y_1)\|_{\psi_{1/2}} \le C_{28;L}' B^4 |v_1|^{-1}(1 + |v_k / v_1|).
\end{aligned} \tag{B.15}$$

Because initialization $v^{(0)} \in \mathcal{D}_{\text{warm}}$, on the event $\mathcal{E}(t, T_x)$ for T_x earlier defined in (2.12), we have $|v_1|^{-1} \le \sqrt{3}/\sqrt{2}$, $|v_k / v_1| \le 1/\sqrt{2}$, and then $\|e_k^{(t)}\|_{\psi_{1/2}} \le 3C_{28;L}' B^4$. Since $\mathbb{1}_{(t \le T_x)} \in \mathcal{F}_{t-1}$, we know that $\{e_k^{(t)} \mathbb{1}_{(t \le T_x)}\}$ forms a martingale difference sequence with respect to $\mathcal{F}_{t-1}$. Additionally, because $\|e_k^{(t)} \mathbb{1}_{(t \le T_x)}\|_{\psi_{1/2}} \le \|e_k^{(t)}\|_{\psi_{1/2}}$, we have

$$\|e_k^{(t)} \mathbb{1}_{(t \le T_x)}\|_{\psi_{1/2}} \le 3C_{28;L}' B^4.$$

With the bound given above, we apply Theorem 29 with $h = 1=2$ and obtain for all $\in (0; e^{-1}]$,

$$P\left(\max_{1 \le t \le T_{;}} \frac{\sum_{s=1}^{t} e_k^{(s)}}{t} \ge C_{16;L} B^4 (T_{;})^{1=2} \log^{5=2}(1=)\right)$$

$$= P\left(\max_{1 \le t \le T_{;}} \frac{\sum_{s=1}^{t} e_k^{(s)} 1_{(s \le T_x)}}{t} \ge C_{16;L} B^4 (T_{;})^{1=2} \log^{5=2}(1=)\right)$$

$$\le 2 \left[3 + 6^4 \frac{64 T_{;} \vee 9C_{28;L}^{02} B^{8 \vee 2}}{C_{16;L}^2 B^{8 \vee 2} T_{;} \log^5(1=)}\right] \exp\left\{-\frac{C_{16;L}^2 B^{8 \vee 2} T_{;} \log^5(1=)}{32 T_{;} \vee 9C_{28;L}^{02} B^{8 \vee 2}}\right\} \le \frac{5184}{\log^5(1=)};$$

where constant $C_{16;L} = 12^{\sqrt{2}} C_{28;L}^{0}$. ∎

Proof [Proof of Lemma 17] Recall de nition of B in (B.2), $T_{B;k}$ in (B.3) and $T_{;}$ in (2.2). Using Markov inequality and Lemma 12, we have

$$P(|v^{>} Y| > B) = P\left(\frac{|v^{>} Y|^2}{B^2} > \frac{B^2}{B^2}\right) \le \exp\left(-\frac{B^2}{B^2}\right) E\exp\left(\frac{|v^{>} Y|^2}{B^2}\right) \le \frac{2}{T_{;1}}:$$

Similarly we also have

$$P(|Y_1| > B) \le \frac{2}{T_{;1}}; \qquad P(|Y_k| > B) \le \frac{2}{T_{;1}}:$$

Taking union bound,

$$P(T_{B;k} \le T_{;}) \le \sum_{t=1}^{T_{;}} P(|v^{(t-1)>} Y^{(t)}| > B) + P(|Y_1^{(t)}| > B) + P(|Y_k^{(t)}| > B)$$

$$\le 3T_{;} \cdot \frac{2}{T_{;1}} \le 6(+1);$$

where we use the elementary inequality $x \le e^{x}$ for all $x \ge 0$ to obtain $T_{;} = T_{;1} + 1$. ∎

Appendix C. Secondary Lemmas in Uniform Initialization Analysis

For notational simplicity, we denote $v \equiv v^{(t-1)}$, $Y \equiv Y^{(t)}$. Recall that we bounded the Orlicz $_2$-norm of each Y_i and $v^{>} Y$ by B using Lemma 12 in Appendix B and introduced truncation barrier B in (B.2) under warm initialization condition, based on rescaled time $T_{;}$. Under uniform initialization condition, we consider a different rescaled time $T_{;}^{0}$ de ned in (3.1). Accordingly, we introduce a slightly larger truncation barrier based on $T_{;}^{0}$

$$B_o \equiv B \log^{1=2}(T_{;1}^{o} \vee 1); \tag{C.1}$$

where $\in (0; e^{-1}]$ is some xed positive. For each coordinate $k \in [2; d]$ de ne the rst time the norm of a data observation exceeds the truncation barrier as

$$T_{B_o;k} = \inf\left\{t \ge 1 : |v^{(t-1)>} Y^{(t)}| > B_o \text{ or } |Y_1^{(t)}| > B_o \text{ or } |Y_k^{(t)}| > B_o\right\}: \tag{C.2}$$

C.1. Proof of Lemma 9

Proof of Lemma 9 shares Lemma 13, 14 and 15 with proof of Lemma 5. With the new truncation barrier B_o given in (C.1) and the new rescaled time $T^o_\cdot$ earlier de ned in (3.1), we plug in $B = B_o$ in Lemma 13 and introduce a new concentration lemma and a new tail probability lemma.

Lemma 18 For any xed coordinate $k \in [2, d]$, let Assumption 2 hold and initialization $v^{(0)} \in D_{mid;k}$. Let $\cdot$, be any xed positives. Then there exists an event $H_{k;18;L}$ satisfying

$$P(H_{k;18;L}) \geq 1 - 6 + \frac{5184}{\log^5 \quad 1} \quad ;$$

such that on event $H_{k;18;L}$ the following concentration result holds

$$\max_{1 \leq t \leq T^o_\cdot \wedge T_{w;k}} \sum_{s=1}^{t} e_k^{(s)} \leq C_{18;L} \log^{5=2} \quad {}^{1} B^4 d^{1=2} (T^o_\cdot)^{1=2};$$

where $C_{18;L}$ is a positive, absolute constant.

Lemma 19 Let be any xed positive. For each coordinate $k \in [2, d]$, we have

$$P\left(T_{B_o;k} \leq T^o_\cdot\right) \leq 6(\ + 1) : \tag{C.3}$$

With the above secondary lemmas at hand, we are now ready for the proof of Lemma 9.

Proof [Proof of Lemma 9] Recall the de nition of stopping time $T_{w;k}$ in (3.9) and $T_{B_o;k}$ in (C.2). Since scaling condition (3.2) implies stepsize $\leq 1=(2B_o^4 d^{1=2})$, we can apply Lemma 13 with $B = B_o$. On the event $(t \leq T^o_\cdot \wedge T_{w;k}) \cap (T_{B_o;k} > T^o_\cdot)$, we have $v_1^{\ 2} \leq d$ and hence

$$Q^{(t)}_{U;k} \leq 4B_o^{\ 8\ 2} v_1^{\ 2} \leq 4B^8 d^{\ 2} \log^4(T^o_{\cdot;1} \quad {}^{1}):$$

We de ne event $H_{k;9;L} := (T_{B_o;k} > T^o_\cdot) \cap H_{k;18;L}$, then on the event $H_{k;9;L}$, by applying Lemma 15 and 18, for all $t \leq T^o_\cdot \wedge T_{w;k}$ we have

$$\begin{aligned} U_k^{(t)} \leq U_k^{(0)} + \left| \ \right| 4 \left| 3 \right| \sum_{s=0}^{t \quad 1} (v_1^{(s)})^2 (v_k^{(s)})^2 U_k^{(s)} \\ \leq T^o_\cdot \cdot 4B^8 d^{\ 2} \log^4(T^o_{\cdot;1} \quad {}^{1}) + C_{16;L} \log^{5=2} \quad {}^{1} B^4 d^{1=2} (T^o_\cdot)^{1=2}: \end{aligned} \tag{C.4}$$

Scaling condition (3.2) in Lemma 6 implies that

$$B^4 d^{1=2} (T^o_\cdot)^{1=2} \log^4(T^o_{\cdot;1} \quad {}^{1}) \leq \log^{5=2} \quad {}^{1}: \tag{C.5}$$

Together with (C.4), on the event $H_{k;9;L}$ we have

$$U_k^{(t)} \leq U_k^{(0)} + \left| \ \right| 4 \left| 3 \right| \sum_{s=0}^{t \quad 1} (v_1^{(s)})^2 (v_k^{(s)})^2 U_k^{(s)} \leq C_{9;L} \log^{5=2} \quad {}^{1} B^4 d^{1=2} (T^o_\cdot)^{1=2};$$

where positive constant $C_{9;L} = C_{16;L} + 4$. Scaling condition (3.2) implies $|4 - 3|((v_1^{(s)})^2 - (v_k^{(s)})^2) \in [0;1)$ for all $s < T_{;}^{o} \wedge T_{w;k}$. From the reversed Gronwall Lemma 25, on the event $H_{k;9;L}$ the following holds for all $t < T_{;}^{o} \wedge T_{w;k}$

$$U_k^{(t)} - U_k^{(0)} \prod_{s=0}^{t-1} \left[1 - |4 - 3|\,(v_1^{(s)})^2 - (v_k^{(s)})^2 \right] \le 2C_{9;L} \log^{5/2}(1/B^4 d)^{1/2} (T_{;}^{o})^{1/2}.$$

To obtain a lower bound on the probability of event $H_{k;9;L}$, we combine Lemmas 18, 19 and take union bound,

$$P(H_{k;9;L}) \ge 1 - 6 - 12 - \frac{5184}{\log^{5}(1/B)}$$

■

C.2. Proof of Lemma 10

Lemma 10 provides a quantitative characterization of the uniform initialization on D_1. As a probabilistic fact, the uniform distribution $v^{(0)}$ in D_1 is equal in distribution to $x/\|x\|^{1/2}$ with $N(0;I)$. In addition, we have

$$\min_{2 \le k \le d} W_k^{(0)} \ge \min_{2 \le k \le d} \log \frac{(v_1^{(0)})^2}{(v_k^{(0)})^2} = \log(v_1^{(0)})^2 - \max_{2 \le k \le d} \log(v_k^{(0)})^2; \tag{C.6}$$

due to definition of $W_k^{(t)}$ in (3.11) and the elementary inequality $x \ge 1 - \log x$ for all $x > 0$. Intuitively, this means that $\min_{2 \le k \le d} W_k^{(0)}$ is lower bounded by the spacing between the largest and second largest order statistics of i.i.d. logarithmic chi-squared distributions.

We provide an elementary probabilistic Lemma 20 to show the spacing on the right hand of (C.6) $W_k^{(0)} = \Omega(\log^{-1} d)$ with high probability.[5]

Lemma 20 Let χ_i^2 $(1 \le i \le n)$ be squares of i.i.d. standard normal variables, and denote their order statistics as $\chi_{(1)}^2 \le \cdots \le \chi_{(n)}^2$. Then for any $\delta \in (0; 1/3)$, when $n \ge 2^{\sqrt{2/e}\log(1/\delta)} + 1$, we have with probability at least $1 - 3\delta$

$$\log \chi_{(n)}^2 - \log \chi_{(n-1)}^2 \ge \frac{1}{8 \log(1/\delta) \log n}. \tag{C.7}$$

We can apply Lemma 20 and prove Lemma 10.

Proof [Proof of Lemma 10] Following assumptions in Theorem 7, we know that $v^{(0)}$ has the same distribution as $x/\|x\|^{1/2}$ where $x \sim N(0;I)$. Under scaling condition (3.2) in Lemma 6, Lemma 10 is straightforward if we apply Lemma 20 with $n = d$ and use (C.6). ■

5. In retrospect, a similar lemma characterizing the lower bound of spacings was achieved using a different method in Bai et al. (2018).

C.3. Proof of Lemma 11

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

$$Q_{W,k}^{(t)} = W_k^{(t)} - W_k^{(t-1)} + \text{sign}(\cdot_4 - 3) \cdot 2(v^\top Y)^3 v_1 v_k^3 (v_1 Y_k - v_k Y_1). \tag{C.8}$$

Lemma 21 For each coordinate $k \in [2, d]$ and any $t \geq 1$, on the event

$$H_{k,21,L}^{(t)} := |v^\top Y| \leq B_o, |Y_1| \leq B_o, |Y_k| \leq B_o, v_1^2 < 3v_k^2, v_1^2 \geq \max_{i \in [2,d]} v_i^2, \tag{C.9}$$

under condition

$$12B^8 d^{-2} \log^4(T_{\cdot,1}^{-1}) \leq 1, \tag{C.10}$$

we have

$$|Q_{W,k}^{(t)}| \leq C_{21,L} B_o^8 d^{-2},$$

where $C_{21,L}$ is a positive, absolute constant.

Lemma 22 Let Assumption 2 hold. For each coordinate $k \in [2, d]$ and any $t \geq 1$, we have

$$\mathbb{E}\left[\text{sign}(\cdot_4 - 3) \cdot 2(v^\top Y)^3 v_1 v_k^3 (v_1 Y_k - v_k Y_1) \,\middle|\, \mathcal{F}_{t-1}\right] = -2|\cdot_4 - 3|v_1^2 W_k. \tag{C.11}$$

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

$$f_k^{(t)} = \text{sign}(\cdot_4 - 3) \cdot 2(v^\top Y)^3 v_1 v_k^3 (v_1 Y_k - v_k Y_1) - 2|\cdot_4 - 3|v_1^2 W_k, \tag{C.12}$$

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

$$W_k^{(t)} = (1 + 2|\cdot_4 - 3|(v_1^{(t-1)})^2) W_k^{(t-1)} + Q_{W,k}^{(t)} + f_k^{(t)}. \tag{C.13}$$

By letting

$$P_t^o = \prod_{s=0}^{t-1} (1 + 2|\cdot_4 - 3|(v_1^{(s)})^2)^{-1} \preceq \mathcal{F}_{t-1}, \tag{C.14}$$

we conclude the following lemma.

Lemma 23 For each coordinate $k \in [2, d]$ and any $t \geq 1$, the iteration $W_k^{(t)}$ can be represented linearly as

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

Lemma 24 Let Assumption 2 hold and initialization $w^{(0)} \in D_{cold} \setminus D_{mid,k}^c$. Let $\cdot$ be a fixed positive. For each coordinate $k \in [2, d]$, there exists an event $H_{k,24,L}$ satisfying

$$\mathbb{P}(H_{k,24,L}) \geq 1 - 6 + \frac{5184}{\log^5 \cdot^{-1}}$$

such that on event $H_{k;24;L}$ the following concentration result holds

$$\max_{1 \le t \le T^o_{;0:5} \wedge T_{c;k} \wedge T_1} \left| \sum_{s=1}^{t} P_s^o f_k^{(s)} \right| \le C_{24;L} \log^{5/2}(1/\delta) \frac{B^4}{|j|^4 \vee 3|j|^{1/2}} d^{1/2}$$

where $C_{24;L}$ is a positive, absolute constant.

Along with the tail probability Lemma 19, we are ready to present the proof of Lemma 11.

Proof [Proof of Lemma 11] Recall the definition of truncation barrier B_0 in (C.1), stopping times $T_{B_o;k}$ in (C.2), $T_{c;k}$ in (3.14) and T_1 in (3.15). We notice that (C.10) holds under scaling condition (3.2), and event $\{t \le T^o_{;0:5} \wedge T_{c;k} \wedge T_1\} \setminus (T_{B_o;k} > T^o_{;0:5}) \cap H^{(t)}_{k;21;L}$. We define event $H_{k;11;L} = (T_{B_o;k} > T^o_{;0:5}) \setminus H_{k;24;L}$, then on event $H_{k;11;L}$, by applying Lemma 21 for all $t \le T^o_{;0:5} \wedge T_{c;k} \wedge T_1$ we obtain

$$v_1^2 \le \frac{1}{d}; \quad |Q^{(t)}_{W;k}| \le C_{21;L} B^8 d^2 \log^4(T^o_{;1}/\delta).$$

Scaling condition (3.2) also guarantees

$$\frac{2|j|^4 \vee 3|j|}{d} < 1; \quad \frac{B^4}{|j|^4 \vee 3|j|^{1/2}} d^{1/2} \log^4(T^o_{;1}/\delta) \le \log^{5/2}(1/\delta); \tag{C.16}$$

and hence by summation of geometric series, on event $H_{k;11;L}$ we have for all $t \le T^o_{;0:5} \wedge T_{c;k} \wedge T_1$

$$\begin{aligned}
&\left| \sum_{s=1}^{t} P_s^o Q^{(s)}_{W;k} \right| \\
&\le \sum_{s=1}^{t} \left(1 + \frac{2|j|^4 \vee 3|j|}{d}\right)^s |Q^{(s)}_{W;k}| \le \frac{1 + \frac{2|j|^4 \vee 3|j|}{d}}{1 - \left(1 + \frac{2|j|^4 \vee 3|j|}{d}\right)^{-1}} C_{21;L} B^8 d^2 \log^4(T^o_{;1}/\delta) \\
&= \frac{C_{21;L}}{2} \frac{B^8}{|j|^4 \vee 3|j|} d^2 \log^4(T^o_{;1}/\delta) \le \frac{C_{21;L}}{2} \log^{5/2}(1/\delta) \frac{B^4}{|j|^4 \vee 3|j|^{1/2}} d^{1/2}.
\end{aligned}$$

Combining with Lemmas 23 and 24, on event $H_{k;11;L}$ we know that for all $t \le T^o_{;0:5} \wedge T_{c;k} \wedge T_1$ the following holds for constant $C_{11;L} = C_{24;L} + C_{21;L}/2$

$$W_k^{(t)} \le \prod_{s=0}^{t-1} \left(1 + |j|^4 \vee 3|j| (v_1^{(s)})^2\right) W_k^{(0)} \le C_{11;L} \log^{5/2}(1/\delta) \frac{B^4}{|j|^4 \vee 3|j|^{1/2}} d^{1/2}.$$

We verify the remaining claims in Lemma 11 by applying Lemma 19 with $\eta = 0.5$, Lemma 24 and taking union bound

$$\mathbb{P}(H_{k;11;L}) \ge 1 - \mathbb{P}(T_{B_o;k} > T^o_{;0:5}) - \mathbb{P}(H^c_{k;24;L}) \ge 1 - 15 \vee \frac{5184}{\log^{-1}(5/\delta)}.$$

. ■

C.4. Proof of Secondary Lemmas

Proof [Proof of Lemma 18] If $\{e_k^{(t)}\}_{t-1}$ earlier defined in (B.7) forms a martingale difference sequence with respect to $\mathcal{F}_{t-1}$. Recall (B.15) in proof of Lemma 16, we have derived the following Orlicz $\psi_{1/2}$-norm bound under Assumption 2

$$\|e_k^{(t)}\|_{\psi_{1/2}} \le C_{28;L}^0 B^4 |v_1|^{-1}(1 + |v_k/v_1|).$$

Because initialization $w^{(0)} \in D_{mid;k}$, on the event $\{t \le T_{w;k}\}$, where $T_{w;k}$ is earlier defined in (3.9), we have $|v_1|^{-1} \le d^{1/2}$, $|v_k/v_1| \le 1/\sqrt{2}$, and then

$$\|e_k^{(t)}\|_{\psi_{1/2}} \le 2C_{28;L}^0 B^4 d^{1/2}.$$

Because $1_{(t \le T_{w;k})} \in \mathcal{F}_{t-1}$, we know that $\{e_k^{(t)} 1_{(t \le T_{w;k})}\}$ forms a martingale difference sequence with respect to $\mathcal{F}_{t-1}$, and $\|e_k^{(t)} 1_{(t \le T_{w;k})}\|_{\psi_{1/2}} \le \|e_k^{(t)}\|_{\psi_{1/2}} \le 2C_{28;L}^0 B^4 d^{1/2}$. Hence we can apply Theorem 29 with $\psi = \psi_{1/2}$ and obtain for all $\delta \in (0, e^{-1}]$,

$$\begin{aligned}
&P\Big(\max_{1 \le t \le T_{\epsilon}^o \wedge T_{w;k}} \sum_{s=1}^t e_k^{(s)} \ge C_{18;L} B^4 d^{1/2} (T_{\epsilon}^o)^{1/2} \log^{5/2}(\tfrac{1}{\delta})\Big) \\
&= P\Big(\max_{1 \le t \le T_{\epsilon}^o} \sum_{s=1}^t e_k^{(s)} 1_{(s \le T_{w;k})} \ge C_{18;L} B^4 d^{1/2} (T_{\epsilon}^o)^{1/2} \log^{5/2}(\tfrac{1}{\delta})\Big) \\
&\le 2\Big[3 + 6^4 \frac{64\, T_{\epsilon}^o\, 4C_{28;L}^{02} B^8 d}{C_{18;L}^2 B^8 d\, T_{\epsilon}^o \log^5(\frac{1}{\delta})}\Big] \exp\Big\{-\frac{C_{18;L}^2 B^8 d\, T_{\epsilon}^o \log^5(\frac{1}{\delta})}{32\, T_{\epsilon}^o\, 4C_{28;L}^{02} B^8 d}\Big\} \\
&= \delta\Big(6 + \frac{5184}{\log^5(\frac{1}{\delta})}\Big),
\end{aligned}$$

where constant $C_{18;L} \ge 8\sqrt{2}C_{28;L}^0$. ∎

Proof [Proof of Lemma 19] Applying Markov inequality and Lemma 12 gives

$$P(|v^\top Y| > B_0) = P\Big(\frac{|v^\top Y|^2}{B^2} > \frac{B_0^2}{B^2}\Big) \le \exp\Big(-\frac{B_0^2}{B^2}\Big) E\exp\Big(\frac{|v^\top Y|^2}{B^2}\Big) \le \frac{2}{T_{\epsilon;1}^o}.$$

With the same procedure we also obtain

$$P(|Y_1^{(t)}| > B_0) \le \frac{2}{T_{\epsilon;1}^o}, \quad P(|Y_k^{(t)}| > B_0) \le \frac{2}{T_{\epsilon;1}^o}.$$

Taking union bound,

$$\begin{aligned}
&P\{T_{B_0;k} \le T_{\epsilon}^o\} \le \sum_{t=1}^{T_{\epsilon}^o} P(|v^{(t-1)\top} Y^{(t)}| > B_0) + P(|Y_1^{(t)}| > B_0) + P(|Y_k^{(t)}| > B_0) \\
&\le 3T_{\epsilon}^o \cdot \frac{2}{T_{\epsilon;1}^o} \le 6(\psi + 1)\delta;
\end{aligned}$$

due to $T^{o}_{;} = T^{o}_{;1} + 1$, which comes from its definition in (3.1) and the elementary inequality $\lceil x \rceil \le (x+1) \le \lceil x \rceil + 1$ for all $x \ge 0$. ∎

Proof [Proof of Lemma 20] Let $F(x)$ be the cumulative distribution function of $\log(\chi^2)$, where χ^2 denote the chi-squared distribution, then

$$F(x) = 2\Phi(e^{x/2}) - 1, \qquad F'(x) = e^{x/2}\phi(e^{x/2}). \tag{C.17}$$

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$(0,1)$, whose largest and second largest order statistics $U_{(n)}, U_{(n-1)}$ are denoted as $U, \bar{U}$ for notational simplicity. We also define $M = F^{-1}(U)$, $\bar{M} = F^{-1}(\bar{U})$. Since $F(x)$ is concave for $x \ge 0$, under condition

$$M \ge 0, \quad \text{i.e.} \quad U \ge 2\Phi(1) - 1, \tag{C.18}$$

by concavity we have

$$U - \bar{U} = F(M) - F(\bar{M}) \le F'(M)(M - \bar{M}). \tag{C.19}$$

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

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

(1) Taking the derivative for an inverse function gives

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

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

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

(2) The following bound on tail probability of standard gaussian random variable is folklore:

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

We define

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

then for all $x \ge 1$, if $z_1 \ge \exp(x^2/2)$, i.e. $x \ge \sqrt{2\log z_1}$, we have

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

which implies that

$$x \quad \sqrt{2\log z_1 \quad \log\log z_1 \quad 3\log 2} \tag{C.23}$$

If $z_2 \quad \exp(x^2=2)$, we have

$$x \quad \sqrt{2\log z_2}: \tag{C.24}$$

For all $y \ 2 \ [1 \quad 1=\sqrt{2e}; \ 1)$, we can find the solution $x \quad 1$ of $\frac{y+1}{2} = 1 \quad \frac{1}{\sqrt{2z_1}}$: By applying (C.22) and (C.23) with this solution x, we obtain

$$1 \quad \frac{y+1}{2} \quad x \quad \sqrt{2\log\frac{\sqrt{2}}{(1 \quad y)} \quad \log\log\frac{\sqrt{2}}{(1 \quad y)} \quad 3\log 2} \tag{C.25}$$

We can find solution $x \quad 1$ to

$$\frac{y+1}{2} = 1 \quad \frac{1}{\sqrt{2z_2}}:$$

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

$$1 \quad \frac{y+1}{2} \quad x \quad \sqrt{2\log\frac{\sqrt{2}}{(1 \quad y)}}: \tag{C.26}$$

Applying the lower bound (C.25) and upper bound (C.26) of $1 \quad \frac{y+1}{2}$ to (C.21), we have for all $y \ 2 \ [1 \quad 1=\sqrt{2e}; \ 1)$

$$[F \quad ^1]^0(y) \quad \sqrt{2} \quad \frac{1}{\sqrt{2\log\frac{\sqrt{2}}{(1 \quad y)}}} \quad \frac{\sqrt{\frac{\sqrt{2}}{(1 \quad y)}}}{\sqrt{8\log\frac{\sqrt{2}}{(1 \quad y)}}} \quad \frac{1}{2(1 \quad y)\log(1 \quad y) \quad ^1}:$$

Replacing y with U and $F \quad ^1(y)$ by M we have under condition

$$U \quad 1 \quad \frac{1}{\sqrt{2e}} \tag{C.27}$$

we have

$$F^0(M \ ) = \frac{1}{[F \quad ^1]^0(U \ )} \quad 2(1 \quad U \ )\log(1 \quad U \ ) \quad ^1 \tag{C.28}$$

(3) It is standard from order statistics that $U \quad \text{Beta}(1; n)$ and $1 \quad U \quad \text{Beta}(2; n \quad 1)$. Therefore, for any $2 \ (0; 1=3]$ and $n \quad 2$,

$$P \ U \quad U \quad \frac{}{n} \quad \int_0^{\frac{}{n}} n(1 \quad x)^{n \quad 1}dx = 1 \quad 1 \quad \frac{}{n}^n \quad 1 \quad \frac{1}{3} \quad (\log 3)$$

where we used elementary inequalities $(1 \quad 1=x)^x \quad 1=3$ when $x \quad 6$, and $1 \quad 3^{\ x} \quad (\log 3)x$ for all $x \ 2 \ (0; 1=3]$. In addition, for any $2 \ (0; 1=3]$ and $n \quad 2 \quad \sqrt{2e}\log \quad ^1 + 1$ we have

$$P \quad 1 \quad U \quad \frac{2\log \quad ^1}{n \quad 1} = \int_{\frac{2\log \quad ^1}{n \quad 1}}^1 n(n \quad 1)x(1 \quad x)^{n \quad 2}dx$$

$$= 2\log \quad ^1\frac{n}{n \quad 1} \quad 1 \quad \frac{2\log \quad ^1}{n \quad 1}^{n \quad 1} + \quad 1 \quad \frac{2\log \quad ^1}{n \quad 1}^n$$

$$2\log \quad ^1 + 1 \quad 1 \quad \frac{2\log \quad ^1}{n \quad 1}^{n \quad 1} \quad (2\log \quad ^1 + 1) \quad ^2 \quad \frac{2\log 3 + 1}{3};$$

where we used elementary inequalities $(1+1/x)^x < e$ for all $x > 0$, and $(2\log\frac{1}{\delta}+1)/(2\log 3+1) \le 1/\delta$ for all $\delta \in (0, 1/3]$. Taking union bound, we have

$$\mathbb{P}\left\{\bigcup_{k=2}^{d}\left\{\frac{|v_1/v_k|}{\sqrt{n}} \ge 1\right\}\right\} \le \frac{2\log\frac{1}{\delta}}{n-1} \le \delta/3. \tag{C.29}$$

(4) Notice that when $\delta \in (0, 1/3]$ and $n \ge 2\sqrt{2e}\log\frac{1}{\delta}+1$, conditions (C.18) and (C.27) hold automatically when (C.29) holds. Combining (C.19), (C.28) and (C.29), for all $\delta \in (0, 1/3]$, $n \ge 2\sqrt{2e}\log\frac{1}{\delta}+1$ we have with probability at least $1-\delta$ that

$$M - \underline{M} \le \frac{U-\underline{U}}{F'(M)} \le \frac{U-\underline{U}}{2(1-\underline{U})\log(1-\underline{U})^{-1}} \le \frac{(n-1)}{4n\log\frac{1}{\delta}\log\frac{n-1}{2\log\frac{1}{\delta}}} \le \frac{1}{8\log\frac{1}{\delta}\log n}.$$

Seeing that $(M, \underline{M})$ is equidistributed as $(\log\chi^2_{(n)}, \log\chi^2_{(n-1)})$, for all $\delta \in (0, 1/3]$ and $n \ge 2\sqrt{2e}\log\frac{1}{\delta}+1$ we have (C.7) holds with probability $1-\delta$.

■

Proof [Proof of Lemma 21] Recall that we define $\varsigma = \mathrm{sign}(\gamma_4 - 3)$. For any fixed coordinate $k \in [2, d]$, we have

$$\begin{aligned} W_k^{(t)} - W_k^{(t-1)} &= \frac{v_1 + \varsigma(v^\top Y)^3 Y_1)^2 - v_k + \varsigma(v^\top Y)^3 Y_k)^2}{v_k^2} - \frac{v_1^2 - v_k^2}{v_k^2} \\ &= \frac{2\varsigma(v^\top Y)^3 v_1 v_k (v_k Y_1 - v_1 Y_k) + \varsigma^2 (v^\top Y)^6 (v_k^2 Y_1^2 - v_1^2 Y_k^2)}{v_k^2 (v_k + \varsigma(v^\top Y)^3 Y_k)^2} \\ &= \varsigma\left(1 + \varsigma(v^\top Y)^3 \frac{Y_k}{v_k}\right)^{-2} v_k^{-4}\left[2(v^\top Y)^3 v_1 v_k (v_k Y_1 - v_1 Y_k) + \varsigma(v^\top Y)^6 (v_k^2 Y_1^2 - v_1^2 Y_k^2)\right]. \end{aligned}$$

Combining with (C.8), this implies

$$\begin{aligned} Q^{(t)}_{W;k} = 2\varsigma &\left[\left(1 + \varsigma(v^\top Y)^3 \frac{Y_k}{v_k}\right)^{-2} - 1 - (v^\top Y)^3 v_1 v_k^{-2} Y_1 - \frac{v_1}{v_k} Y_k\right] \\ &+ \varsigma^2\left[\left(1 + \varsigma(v^\top Y)^3 \frac{Y_k}{v_k}\right)^{-2} - (v^\top Y)^6 v_k^{-2} Y_1^2 - \frac{v_1^2}{v_k^2} Y_k^2\right]. \end{aligned}$$

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

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

On event $\mathcal{H}^{(t)}_{k;21;L}$, since $|v_1/v_k| < \sqrt{3}$ and $|v_k|^2 \ge 1/(3d)$, (C.10) implies $\varsigma(v^\top Y)^3 \frac{Y_k}{v_k} \le \frac{1}{2}$, and hence we have

$$|Q^{(t)}_{W;k}| \le 6\sqrt{3}B_o^4 d^{1/2} - (3+3\sqrt{3})B_o^4 d^{1/2} + 4 \le 12 B_o^8 d \le C_{21;L} B_o^8 \delta^2 d$$

where constant $C_{21;L} = 156 + 36\sqrt{3}$.

■

Proof [Proof of Lemma 22] Recall that $W_k^{(t-1)} = (v_1^2 - v_k^2)/v_k^2$. Under Assumption 2, using (B.13) we have

$$\begin{aligned}
&E\left[\operatorname{sign}(\Delta_4 - \Delta_3) \cdot 2v_1 v_k^3 (v^\top Y)^3 (v_1 Y_k - v_k Y_1) \mid \mathcal{F}_{t-1}\right] \\
&= 2\operatorname{sign}(\Delta_4 - \Delta_3) \cdot v_1 v_k^3 \cdot (\Delta_4 - \Delta_3) v_1 v_k^3 + 3 v_1 v_k \cdot (\Delta_4 - \Delta_3) v_k v_1^3 - 3 v_k v_1 \\
&= -2|\Delta_4 - \Delta_3| v_1^2 W_k
\end{aligned}$$

■

Proof [Proof of Lemma 23] From (C.13) and (C.14), we have

$$P_s^o W_k^{(s)} - P_s^o \Delta_1 W_k^{(s-1)} = P_s^o Q_{W;k}^{(s)} + f_k^{(s)} \tag{C.30}$$

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

Proof [Proof of Lemma 24] If $\{f_k^{(t)}\}_{t \geq 1}$, earlier defined in (C.12), forms a martingale difference sequence with respect to $\mathcal{F}_{t-1}$. Similar to techniques in proof of Lemma 16, we apply Lemma 27 three times, then use (E.1) and (B.12), in order to obtain the following bound on Orlicz-$\psi_{1/2}$ norm

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

where we use the fact that $|v_1/v_k| \leq \sqrt{3}$ and $v_k^2 \geq 1/(3d)$ on the event $(t \geq T_{c;k} \wedge T_1)$. Then by applying Lemma 28 we further derive

$$\begin{aligned}
\|f_k^{(t)}\|_{\psi_{1/2}} &= \left\|2(v^\top Y)^3 v_1 v_k^3 (v_k Y_1 - v_1 Y_k) - E\left[2(v^\top Y)^3 v_1 v_k^3 (v_k Y_1 - v_1 Y_k)\right]\right\|_{\psi_{1/2}} \\
&\leq C_{28;L}' \left\|2(v^\top Y)^3 v_1 v_k^3 (v_k Y_1 - v_1 Y_k)\right\|_{\psi_{1/2}} \leq 6(1+\sqrt{3}) C_{28;L}' B^4 d^{1/2}
\end{aligned}$$

Because $\mathbb{1}_{(t \geq T_{c;k} \wedge T_1)} \in \mathcal{F}_{t-1}$, we know that $\left\{f_k^{(t)} \mathbb{1}_{(t \geq T_{c;k} \wedge T_1)}\right\}$ forms a martingale difference sequence with respect to $\mathcal{F}_{t-1}$, and $\|f_k^{(t)} \mathbb{1}_{(t \geq T_{c;k} \wedge T_1)}\|_{\psi_{1/2}} \leq \|f_k^{(t)}\|_{\psi_{1/2}}$. Since $v_1^2 \geq 1/d$ on the event $(t \geq T_{c;k} \wedge T_1)$ and $2|\Delta_4 - \Delta_3|/d < 1$ under scaling condition (3.2), by summation of geometric series, for all $t \geq 1$ we have

$$\begin{aligned}
&\left\|\sum_{s=1}^t (P_s^o)^2 f_k^{(s)} \mathbb{1}_{(t \geq T_{c;k} \wedge T_1)}\right\|_{\psi_{1/2}}^2 \leq \sum_{s=1}^t \left(1 + \frac{2|\Delta_4 - \Delta_3|}{d}\right)^{2s} 36(1+\sqrt{3})^2 C_{28;L}'^2 B^8 d^2 \\
&\leq \frac{\left(1 + \frac{2|\Delta_4 - \Delta_3|}{d}\right)^2}{1 - \left(1 + \frac{2|\Delta_4 - \Delta_3|}{d}\right)^2} 36(1+\sqrt{3})^2 C_{28;L}'^2 B^8 d^2 = 18(1+\sqrt{3})^2 C_{28;L}'^2 \frac{B^8}{|\Delta_4 - \Delta_3|} d^2
\end{aligned}$$

With the bound given above, we apply Theorem 29 with $h = 1=2$ as

$$P\left(\max_{1 \le t \le T^{o}_{;0:5} \wedge T_{c;k} \wedge T_1} \left| \sum_{s=1}^{t} P_s^{o} f_k^{(s)} \right| \ge C_{24;L} \log^{5=2}(1/\delta) \sqrt{\frac{B^4}{j^4 - 3j^{1=2}}}\, d^{1=2}\right)$$

$$= P\left(\max_{1 \le t \le T^{o}_{;0:5}} \left| \sum_{s=1}^{t} P_s^{o} f_k^{(s)} 1_{(s \le T_{c;k} \wedge T_1)} \right| \ge C_{24;L} \log^{5=2}(1/\delta) \sqrt{\frac{B^4}{j^4 - 3j^{1=2}}}\, d^{1=2}\right)$$

$$\le 2\left[3 + 6^4 \frac{64 \cdot 18(1+\sqrt{3})^2 C^{o2}_{28;L} B^8 j^4 - 3j^1 d^2}{C^2_{24;L} \log^5(1/\delta) B^8 j^4 - 3j^1 d^2}\right] \exp\left\{ - \frac{C^2_{24;L} \log^5(1/\delta) B^8 j^4 - 3j^1 d^2}{32 \cdot 18(1+\sqrt{3})^2 C^{o2}_{28;L} B^8 j^4 - 3j^1 d^2} \right\}^{1=5}$$

$$= 6 + \frac{5184}{\log^5(1/\delta)}$$

where constant $C_{24;L} = 24(1+\sqrt{3})C^{0}_{28;L}$. ■

Appendix D. 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 literature:

Lemma 25 If for all $t = 1; : : : ; T$, $(t) \in [0; 1)$ and if $u(t)$ satisfies that for some positive constant

$$u(t) \le u(0) + \sum_{0 \le s<t} (s)u(s) \tag{D.1}$$

then for all $t = 1; : : : ; T$

$$u(t) \le u(0) \prod_{0 \le s<t} (1 - (s)) \le 2 \prod_{0 \le s<t} (1 - (s)) \le 2: \tag{D.2}$$

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

Proof [Proof of Lemma 25] Define for all $t = 1; : : : ; T$

$$v(t) = \prod_{0 \le s<t} (1 - (s)) \left[u(0) + \sum_{0 \le s<t} (s)u(s) \right]: \tag{D.3}$$

We have for $s = 0; : : : ; t \quad 1$

$$v(s+1) \quad v(s) = \prod_{0 \ r<s+1} (1 \quad (r))^{\ 1} \quad (s)u(s)$$

$$+ \ @u(0) \sum_{0 \ r<s} (r)u(r)A \ @ \prod_{0 \ r<s+1} (1 \quad (r))^{\ 1} \quad \prod_{0 \ r<s} (1 \quad (r))^{\ 1}A$$

$$= \prod_{0 \ r<s+1} (1 \quad (r))^{\ 1} \quad (s)u(s) + @u(0) \sum_{0 \ r<s} (r)u(r)A \ @ \prod_{0 \ r<s+1} (1 \quad (r))^{\ 1} \quad (s)A$$

$$= @u(s) \quad u(0) + \sum_{0 \ r<s} (r)u(r)A \quad (s) \prod_{0 \ r<s+1} (1 \quad (r))^{\ 1}:$$

Since $(s) \ 2 \ [0; 1)$, and the product is nonnegative, the use of the lower side of (D.1) upper-estimates the difference of $v(s)$

$$v(s+1) \quad v(s) \quad (s) \prod_{0 \ r<s+1} (1 \quad (r))^{\ 1} \qquad (D.4)$$

Since $v(0) = u(0)$, telescoping the above inequality for $s = 0; : : : ; t \quad 1$ gives

$$v(t) = v(0) + \sum_{0 \ s<t} v(s+1) \quad v(s) \quad u(0) + \sum_{0 \ s<t} (s) \prod_{0 \ r<s+1} (1 \quad (r))^{\ 1}$$

and hence from the de nition of $f(t)$ in (D.3)

$$u(0) \sum_{0 \ s<t} (s)u(s) = \prod_{0 \ s<t} (1 \quad (s)) \ v(t)$$

$$\prod_{0 \ s<t} (1 \quad (s)) \ @u(0) + \sum_{0 \ s<t} (s) \prod_{0 \ r<s+1} (1 \quad (r))^{\ 1}A$$

$$= u(0) \prod_{0 \ s<t} (1 \quad (s)) + \sum_{0 \ s<t} (s) \prod_{s+1 \ r<t} (1 \quad (r)):$$

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

$$u(t) \quad + u(0) \sum_{0 \ s<t} (s)u(s) \quad + u(0) \prod_{0 \ s<t} (1 \quad (s)) + \sum_{0 \ s<t} (s) \prod_{s+1 \ r<t} (1 \quad (r)) ;$$

which further reduces to

$$u(t) \quad u(0) \prod_{0 \ s<t} (1 \quad (s)) \quad + \sum_{0 \ s<t} (s) \prod_{s+1 \ r<t} (1 \quad (r)):$$

That is, for all $t = 0; 1; : : : ; T$,

$$
\begin{aligned}
u(t) \quad u(0) \quad & \prod_{0 \; s<t} (1 \quad (s)) \quad + \quad \sum_{0 \; s<t} (s) \quad \prod_{s+1 \quad r<t} (1 \quad (r)) \\
= \quad + \quad & \sum_{0 \; s<t} (1 \quad (1 \quad (s))) \quad \prod_{s+1 \quad r<t} (1 \quad (r)) \\
= \quad + \quad & \sum_{0 \; s<t} \Big[\prod_{s+1 \quad r<t} (1 \quad (r)) \quad \prod_{s \quad r<t} (1 \quad (r)) \Big] \\
= 2 & \prod_{0 \; r<t} (1 \quad (r)) \quad 2 ;
\end{aligned}
$$

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

Appendix E. Orlicz -Norm

In this section, we recap Orlicz-norm and relevant properties. Many of the contributions can be found in standard texts Wainwright (2019); van der Vaart and Wellner (1996), and we derive its (old and new) conclusions for our use in online tensorial ICA analysis. We begin with its de nition

De nition 26 (Orlicz -norm) For a continuous, monotonically increasing and convex function (x) de ned for all $x \quad 0$ satisfying $(0) = 0$ and $\lim_{x!1} (x) = 1$, we de ne the Orlicz -norm for a random variable X as

$$kX k \quad \inf f K > 0 : E \ (jX=K j) \quad 1g :$$

When (x) is monotonically increasing and convex for $x \quad 0$, the Orlicz -norm satis es triangle inequality, i.e. for any random variables X and Y we have $kX + Y k \quad k X k + kY k$. The central case of interest in this paper is the Orlicz -norm de ned for a random variable X as

$$kX k \quad \inf \quad K > 0 : E \exp \frac{jX j}{K} \quad 2 :$$

In above we adopt the function $(x) \quad \exp(x) \quad 1$. When 1, (x) is monotonically increasing and convex for $x \quad 0$, and consequently -norm satis es triangle inequality (van der Vaart and Wellner, 1996)

$$kX + Y k \quad k X k + kY k : \tag{E.1}$$

We state a multiplicative property of the Orlicz -norm which extends a standard result (e.g. Proposition D.3 in Vu and Lei (2013)):

Lemma 27 (Multiplicative property) Let X and Y be random variables then for some 1

$$kXY k_{ = 2} \quad k X k \quad kY k :$$

Proof [Proof of Lemma 27] The inequality is trivial when (i) $\|X\|_\Psi = 0$ or $\|Y\|_\Psi = 0$ since one of the variables is 0 a.s., or (ii) either X or Y has an infinite Ψ-norm. Otherwise let $A = X/\|X\|_\Psi$ and $B = Y/\|Y\|_\Psi$ then $\|A\|_\Psi = \|B\|_\Psi = 1$. Using $|AB| \le \frac{1}{4}(|A| + |B|)^2$ and triangle inequality in (E.1) we have

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

Multiplying both sides of the inequality by the positive $\|X\|_\Psi \|Y\|_\Psi$ gives the desired result. ■

In the case of $\alpha \in (0, 1)$, $\psi(x)$ is no longer convex; in fact it is only convex after a modification of the function in a neighborhood of $x = 0$. Strictly speaking, Ψ-"norm" in this case is not a norm (defined in a Banach space) in the sense that the triangle inequality (E.1) does not hold for this set of Ψ's. For the purpose of our applications, a generalized triangle inequality holds for the $\alpha = 1/2$ case in the following Lemma 28:

Lemma 28 For any random variables X, Y we have the following inequalities for Orlicz $\Psi_{1/2}$-norm

$$\|X + Y\|_{1/2} \le C_{28,L}(\|X\|_{1/2} + \|Y\|_{1/2}) \quad \text{and} \quad \|\mathbb{E}X\|_{1/2} \le C_{28,L}\|X\|_{1/2} \tag{E.2}$$

with $C_{28,L} \le 1.3937$. In addition

$$\|X - \mathbb{E}X\|_{1/2} \le C'_{28,L}\|X\|_{1/2}; \tag{E.3}$$

where $C'_{28,L} \le 3.3359$.

Proof [Proof of Lemma 28] We first rule out some simple cases. Analogously, we only need to prove (E.2) and (E.3) under the setting that both $\|X\|_\Psi$ and $\|Y\|_\Psi$ belongs to the two-point set $\{0, +1\}$. Recall that when $\alpha \in (0, 1)$, $\psi(x)$ does not satisfy convexity when x is around 0. Let the modified Orlicz-Ψ function be

$$\tilde{\psi}(x) := \begin{cases} \exp(x^\alpha) - 1 & x \ge x_0 \\ \frac{x}{x_0}(\exp(x_0^\alpha) - 1) & x \in [0, x_0) \end{cases}$$

for some appropriate $x_0 > 0$, so as to make the function convex. Here x_0 is chosen such that the tangent line of function ψ at x_0 passes through origin, i.e.

$$x_0^{1-\alpha}\exp(x_0^\alpha) = \frac{\exp(x_0^\alpha) - 1}{x_0}.$$

Simplifying it gives us a transcendental equation $(1 - \alpha)\ln x_0)\exp(x_0^\alpha) = 1$ which (admits no analytic solution but) can be solved numerically. When $\alpha = 1/2$, we have $x_{0,1/2} = 2.5396$. Some numerical calculation yields

$$0 \le \psi_{1/2}(x) - \tilde{\psi}_{1/2}(x) \le 0.2666. \tag{E.4}$$

From (E.4) we immediately have

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

(1) Let K_1, K_2 denote the $\psi_{1/2}$ norms of X and Y, then $E\,\psi_{1/2}(|X/K_1|) \le 1$ and $E\,\psi_{1/2}(|Y/K_2|) \le 1$. Based on (E.4) we have $E\,\tilde{\psi}_{1/2}(|X/K_1|) \le 1$ and $E\,\tilde{\psi}_{1/2}(|Y/K_2|) \le 1$. Applying triangle inequality (E.1) to Orlicz $\tilde{\psi}_{1/2}$-norm, we have

$$E\,\tilde{\psi}_{1/2}\left(\frac{X+Y}{K_1+K_2}\right) \le 1 \quad \Rightarrow \quad E\,\psi_{1/2}\left(\frac{X+Y}{K_1+K_2}\right) \le 1.2666$$

where we applied (E.4). Now, applying Jensen's inequality to concave function $f(z) = z^{\log_{2.2666} 2}$ gives that, for constant $C_{28;L} \equiv (\log_2 2.2666)^2 = 1.3937$,

$$E\,\psi_{1/2}(|(X+Y)/(C_{28;L}(K_1+K_2))|) = E\exp(|(X+Y)/(K_1+K_2)|^{1/2})^{\log_{2.2666} 2} \le 1$$
$$E\exp(|(X+Y)/(K_1+K_2)|^{1/2})^{\log_{2.2666} 2} \le 1 \le 1,$$

which implies the first conclusion of (E.2):

$$\|X+Y\|_{\psi_{1/2}} \le C_{28;L}(\|X\|_{\psi_{1/2}} + \|Y\|_{\psi_{1/2}}).$$

(2) Let K denote the $\psi_{1/2}$ norm of X, then $E\,\psi_{1/2}(|X/K|) \le 1$. Based on (E.4) we have $E\,\tilde{\psi}_{1/2}(|X/K|) \le 1$. Because $\tilde{\psi}_{1/2}$ is a convex function, we can apply Jensen's inequality as

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

Combining with (E.4), it holds that $\psi_{1/2}(|EX/K|) \le 1.2666$ which is equivalent to $\psi_{1/2}(|EX/(C_{28;L}K)|) \le 1$, where $C_{28;L} = (\log_2 2.2666)^2 = 1.3937$. Noticing that $E\,\psi_{1/2}(|EX/(CK)|) = \psi_{1/2}(|EX/(CK)|)$, it holds that $\|EX\|_{\psi_{1/2}} \le C_{28;L}\|X\|_{\psi_{1/2}}$ which concludes (E.2).

(3) To conclude (E.3), we have

$$\|X - EX\|_{\psi_{1/2}} \le C_{28;L}(\|X\|_{\psi_{1/2}} + \|EX\|_{\psi_{1/2}}) \le C_{28;L}(1 + C_{28;L})\|X\|_{\psi_{1/2}} = C'_{28;L}\|X\|_{\psi_{1/2}}$$

where positive constant $C'_{28;L} \equiv C_{28;L}(1 + C_{28;L}) = 3.3359$.

■

Appendix F. A Concentration Inequality for Martingales with Weak Exponential-type Tails

We prove a novel concentration inequality for 1-dimensional supermartingale difference sequence with finite ψ_α-norms that plays an important role in our analysis:

Theorem 29 Let $\alpha \in (0, 1]$ be given. Assume that $(u_i : i \ge 1)$ is a sequence of supermartingale differences with respect to F_i, i.e. $E[u_i \mid F_{i-1}] \le 0$, and it satisfies $\|u_i\|_{\psi_\alpha} < \infty$ for each $i = 1, \ldots, N$. Then for an arbitrary $N \ge 1$ and $z > 0$,

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

Proof [Proof of Theorem 29] To prove Theorem 29, 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 and Volný (2001) and Fan et al. (2012) to truncate the tail and analyze the bounded and unbounded pieces separately.

(1) First of all, for the sake of simplicity and with no loss of generality, throughout the following proof of Theorem 29 we shall pose the following extra condition

$$\sum_{i=1}^{N} \|u_i\|^2 = 1 : \tag{F.2}$$

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

$$P\left(\max_{1 \le n \le N} \sum_{i=1}^{n} u_i \ge z\right) \le \left(3 + \frac{3}{4}\right)^2 \frac{64}{z^2} \exp\left(-\frac{z^2}{32}\right)^{\frac{1}{2}} : \tag{F.3}$$

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

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

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

$$\left(3 + \frac{3}{4}\right)^2 \frac{64 \sum_{i=1}^{N} \|u_i\|^2}{z^2} \exp\left\{-\frac{z^2}{32 \sum_{i=1}^{N} \|u_i\|^2}\right\}^{\frac{1}{2}} ;$$

proving (F.1).

(2) We apply a truncation argument used in Lesigne and Volný (2001) and later in Fan et al. (2012). Let $M > 0$ be arbitrary, and we define

$$u_i^0 = u_i 1_{\{\|u_i\| \le M\}} - E\left(u_i 1_{\{\|u_i\| \le M\}} - \|u_i\| - g \mid \mathcal{F}_{i-1}\right) ; \tag{F.4}$$

$$u_i^{00} = u_i 1_{\{\|u_i\| > M\}} - E\left(u_i 1_{\{\|u_i\| > M\}} \|u_i\| - g \mid \mathcal{F}_{i-1}\right) ; \tag{F.5}$$

$$T_n^0 = \sum_{i=1}^{n} u_i^0, \qquad T_n^{00} = \sum_{i=1}^{n} u_i^{00}, \qquad T_n^{000} = \sum_{i=1}^{n} E\left(u_i \mid \mathcal{F}_{i-1}\right) :$$

Since u_i is $\mathcal{F}_i$-measurable, u_i^0 and u_i^{00} are two martingale difference sequences with respect to $\mathcal{F}_i$, and let T_n be defined as

$$T_n = \sum_{i=1}^{n} u_i \quad \text{and hence} \quad T_n = T_n^0 + T_n^{00} + T_n^{000}. \tag{F.6}$$

Since u_i are supermartingale differences we have that T_n^{000} is $\mathcal{F}_{n-1}$-measurable with T_n^{000} $T_0^{000} = 0$; a.s., and hence for any $z > 0$,

$$P\left(\max_{1 \le n \le N} T_n \ge 2z\right) \le P\left(\max_{1 \le n \le N} T_n^{0} + T_n^{000} \ge z\right) + P\left(\max_{1 \le n \le N} T_n^{00} \ge z\right) \le P\left(\max_{1 \le n \le N} T_n^{0} \ge z\right) + P\left(\max_{1 \le n \le N} T_n^{00} \ge z\right) \tag{F.7}$$

In the following, we analyze the tail bounds for T_n^0 and T_n^{00} separately (Lesigne and Volný, 2001; Fan et al., 2012).

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

Lemma 30 (Laib, 1999) Let $(w_i : 1 \le i \le N)$ be a real-valued martingale difference sequence with respect to some filtration $\mathcal{F}_i$, i.e. $E[w_i \mid \mathcal{F}_{i-1}] = 0$; a.s., and the essential norm $\|w_i\|_1$ is finite. Then for an arbitrary $N \ge 1$ and $z > 0$,

$$P\left(\max_{1 \le n \le N} \sum_{i=1}^{n} w_i \ge z\right) \le \exp\left(\frac{z^2}{2 \sum_{i=1}^{N} \|w_i\|_1^2}\right). \tag{F.8}$$

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

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

The proof of Lemma 30 is given in Laib (1999).

Recall our extra condition (F.2), then from the definition of u_i^0 in (F.4) that $|u_i^0| \le 2M\|u_i\|$, the desired bound follows immediately from Laib's inequality in Lemma 30 by setting $w_i = u_i^0$:

$$P\left(\max_{1 \le n \le N} T_n^0 \ge z\right) = P\left(\max_{1 \le n \le N} \sum_{i=1}^{n} u_i^0 \ge z\right) \le \exp\left(\frac{z^2}{8M^2}\right) \tag{F.9}$$

To obtain the tail bound of T_n^{00} we only need to show

$$E(u_i^{00})^2 \le (6M^2 + 8B^2)\|u_i\|^2 \exp\{-M\}, \tag{F.10}$$

where

$$B \ge \left(\frac{3}{2}\right)^{\frac{1}{2}}; \tag{F.11}$$

from which, Doob's martingale inequality implies immediately that

$$P\left(\max_{1 \le n \le N} T_n^{00} \ge z\right) \le \frac{1}{z^2} \sum_{i=1}^{N} E(u_i^{00})^2 \le \frac{6M^2 + 8B^2}{z^2} \exp\{-M\}. \tag{F.12}$$

To prove (F.10), first recall from the definition of u_i^{00} in (F.5) that

$$u_i^{00} = u_i 1_{\{|u_i| > M\|u_i\|\}} - E\left[u_i 1_{\{|u_i| > M\|u_i\|\}} \mid \mathcal{F}_{i-1}\right];$$

Recall from the property of conditional expectation that for any random variable W and a σ-algebra $\mathcal{G} \subseteq \mathcal{F}$

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

where the last equality is due to a simple application of Fubini's Theorem for nonnegative random variable $|W|$. Plugging in $W = u_i 1_{\{|u_i| > \mathcal{M}\|u_i\| \quad \}}$ and $\mathcal{G} = \mathcal{F}_{i-1}$ we have

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

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

$$\mathbb{P}(|u_i|/\|u_i\|_\psi \quad \geq z) \leq \exp\{-z^\alpha\}\mathbb{E}\exp\{|u_i|^\alpha/\|u_i\|_\psi^\alpha \ \} \leq 2\exp\{-z^\alpha\}. \qquad \text{(F.14)}$$

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

$$\begin{aligned}
&\int_{\mathcal{M}}^\infty t\exp\left\{-t^\alpha\right\} \ dt = \int_{\mathcal{M}}^\infty t^{-2}t^3\exp\left\{-t^\alpha\right\} \ dt \leq \int_{\mathcal{M}}^\infty t^{-2} \, dt \cdot \mathcal{M}^3\exp\left\{-\mathcal{M}^\alpha\right\} \qquad \text{(F.15)} \\
&= \mathcal{M}^{-1} \cdot \mathcal{M}^3\exp\left\{-\mathcal{M}^\alpha\right\} = \mathcal{M}^2\exp\left\{-\mathcal{M}^\alpha\right\}.
\end{aligned}$$

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

$$\begin{aligned}
&\int_{\mathcal{M}}^\infty t\exp\left\{-t^\alpha\right\} \ dt = \int_{\mathcal{M}}^{\mathcal{B}} t\exp\left\{-t^\alpha\right\} \ dt + \int_{\mathcal{B}}^\infty t\exp\left\{-t^\alpha\right\} \ dt \\
&\leq \int_{\mathcal{M}}^{\mathcal{B}} dt \cdot \mathcal{B}\exp\left\{-\mathcal{M}^\alpha\right\} + \mathcal{B}^2\exp\left\{-\mathcal{B}^\alpha\right\} \qquad \text{(F.16)} \\
&\leq (\mathcal{B} - \mathcal{M})\mathcal{B}\exp\left\{-\mathcal{M}^\alpha\right\} + \mathcal{B}^2\exp\left\{-\mathcal{M}^\alpha\right\} \leq 2\mathcal{B}^2\exp\left\{-\mathcal{M}^\alpha\right\}.
\end{aligned}$$

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

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

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

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

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

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

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

Plugging this $\mathcal{M}$ back into (F.17) we obtain

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

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

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

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, (F.3) is concluded by noticing the relation (F.6) and setting $z/2$ in the place of z, which hence proves Theorem 29 via the argument in ((1)) in our proof.

■

Appendix G. Visualization and Numerical Experiments for Online Tensorial ICA

In this section, we provide visualization and conduct a group of experiments in support of our theoretical results. In §G.1 we illustrate the two-phase convergence behavior of our algorithm on two most commonly seen ICA problem distributions: Mixture Gaussian and Gaussian-Bernoulli. In §G.2 we empirically validate our sharp theoretical convergence rate (mainly in both d and T) in our Corollary 8.

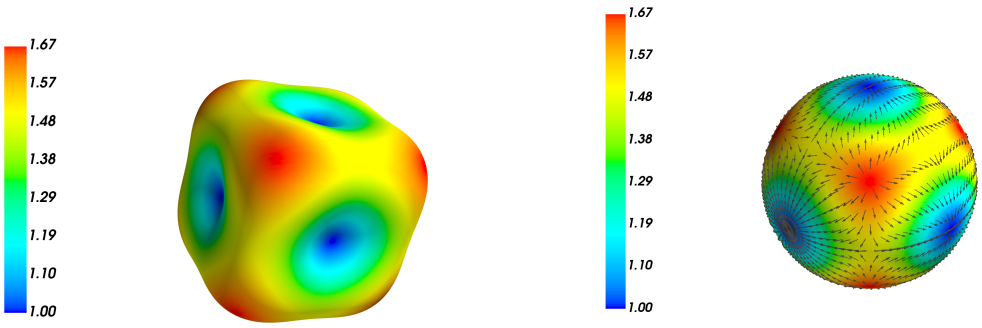

Figure 1: Heatmap visualization of the tensorial ICA objective function. Left: a graph plot where the radius denotes the corresponding function value (a global offset of 2 is added for illustration purposes). Right: a quiver plot with arrows denoting the negative gradient direction.

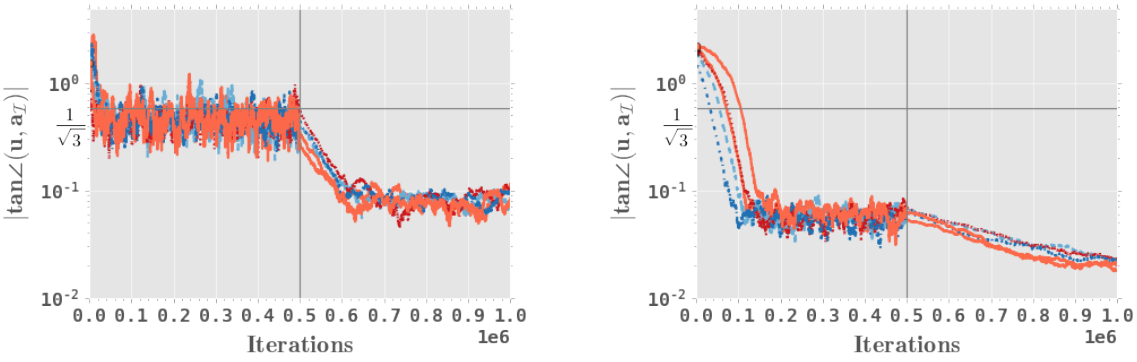

Figure 2: Convergence of our two-phase online tensorial ICA algorithm implemented under two settings. Left: each independent component being a mixture-Gaussian distribution with $\mu_4 = 2.5$. Right: each component being a Gaussian-Bernoulli distribution with $\mu_4 = 6$. Both settings are plotted in the semilog-scale, where the horizontal axis is the iteration number, and the vertical axis is the distance to the closest independent component in metric $\min_{i\in[d]} \left|\tan \angle(\mathbf{u}^{(T)}, \mathbf{a}_i)\right|$.

G.1. Two-Phase Algorithm for Online Tensorial ICA

In this subsection, we present an initial experimental result that demonstrates the two-phase convergence behavior of our Algorithm 1. We arbitrarily set dimension as $d = 20$ and total sample size as $T = 1 \times 10^6$, and the mixing matrix $\mathbf{A}$ randomly drawn from the Haar measure over the d-dimensional orthogonal group. We conduct experiments on two separate instances of the one-dimensional distribution Z_i of independent component in ICA: mixture Gaussian with $\mu_4 < 3$, and Gaussian-Bernoulli with $\mu_4 > 3$, to validate the convergence theory of our two-phase algorithm.

(a) **Mixture Gaussian** We adopt $Z = \delta Y_1 + (1-\delta)Y_2$ as the mixture Gaussian component of Gaussian variables Y_1, Y_2 with

$$Y_1 \sim N\left(-\frac{1}{\sqrt{2}}, \frac{1}{2}\right), \qquad Y_2 \sim N\left(\frac{1}{\sqrt{2}}, \frac{1}{2}\right), \qquad \delta \sim \text{Bernoulli}(1/2) \quad \text{independently.}$$

It is easy to verify that the distribution of Z_i satisfies Assumption 2 with $\mu_4 = 2.5$.

(b) **Gaussian-Bernoulli** where we adopt $Z = \delta Y$ with

$$Y \sim N(0,2), \qquad \delta \sim \text{Bernoulli}(1/2) \quad \text{independently.}$$

It is straightforward to verify that $\mu_4 = 6$ in this case.

The algorithm is initialized at $\mathbf{u}_0$ uniformly drawn from the unit sphere $\mathcal{D}_1$, and we run our two-phase algorithm scheduled as in Corollary 8. That is, in the first half of the training when the number of iterates $\leq T/2$ we choose stepsize as $\frac{8d}{|\mu_4-3|T}$ and in the second half choose stepsize as $\frac{9}{|\mu_4-3|T}$, both omitting logarithmic factors. We measure the convergence by the tangent of the angle between $\mathbf{u}^{(T)}$ and its closest independent component $\mathbf{a}_i$. The resulting 5 independent runs are shown in Figure 2. (The horizontal line represents $y = \frac{1}{\sqrt{3}}$ which is the barrier of the warm region, and the vertical line is $x = T/2$ which is the separation between two phases.) Our result in Figure 2 exemplifies consistency with the main theoretical result of our paper, especially the algorithm's two-phase demonstration. Regardless of the initialization all 10 independent runs share similar trajectories in terms of our measure: in Phase I the algorithm decays fast until oscillating into the warm region (the absolute tangent value below $\frac{1}{\sqrt{3}}$), and in Phase II the algorithm continue to converge linearly until reaching the desired accuracy.

G.2. Validating the Finite-Sample Convergence Rate

In this subsection, we validate our main convergence rate result, Corollary 8, via simulations for a range of values d and T satisfying the scaling condition

$$d \geq 2\sqrt{2\pi e}\log\epsilon^{-1} + 1 \quad \text{and} \quad \frac{d^4}{T} \leq C\epsilon^2, \tag{G.1}$$

so with probability $\geq 1 - O(\epsilon)$ the following $\tilde{\mathcal{O}}(\sqrt{d/T})$-convergence rate result holds for all sufficiently large d and T:

$$|\tan \measuredangle(\mathbf{u}^{(T)}, \mathbf{a}_{\mathcal{I}})| \leq C\sqrt{\frac{d}{T}}, \tag{G.2}$$

where C includes a polylogarithmic factor in d, T, ϵ.

We continue to generate random data $\boldsymbol{Z}$ as we did in the Gaussian-Bernoulli distributions case in §G.1, with parameter value of either d or T allowed to vary in each run while freezing the other. By ergodicity we may empirically estimate $\mathbb{E}|\tan \measuredangle(\mathbf{u}^{(T)}, \mathbf{a}_{\mathcal{I}})|$ by averaging the last few iterates of $|\tan \measuredangle(\mathbf{u}^{(T)}, \mathbf{a}_{\mathcal{I}})|$ — the final $0.6T$ iterates are taken. Then we scatter-plot the empirical $\mathbb{E}|\tan \measuredangle(\mathbf{u}^{(T)}, \mathbf{a}_{\mathcal{I}})|$ under a range of parameters in the figure, gauged by the theoretical rate of (G.2). We first run for $T/2$ epochs with the stepsize propotional to $8d/T$ and run for $T/2$ epochs with the stepsize propotional to $9/T$, as shown in §G.1 and in Corollary 8, forgoing the logarithmic factor. The range of values one takes is $d \in \{7, 12, 20, 33, 54\}$, $T \in \{2\times 10^2, 2\times 10^3, 1\times$