Proceedings of Machine Learning Research vol TBD:1–28, 2020 33rd Annual Conference on Learning Theory

Stochastic Scaled-Gradient Descent with Applications to Generalized Eigenvector Problems

author names withheld

Editor: Under Review for COLT 2020

Abstract

Motivated by the problem of stochastic generalized eigenvalue problem and online canonical correlation analysis, we propose in this paper the Stochastic Scaled-Gradient Descent (SSGD) algorithm for minimizing a nonconvex stochastic function on a generic Riemannian manifold, which generalizes the idea of projected stochastic gradient descent and allows the access of unbiased stochastic scaled-gradients instead of stochastic gradients. By exploiting the local strong convexity we establish a nonasymptotic finite-sample bound of $\sqrt{1/T}$ for the spherical constraint case (applicable to GEV/CCA) which is optimal up to a polylogarithmic factor. On the asymptotic side, a special trajectory averaging argument allows us to achieve the local asymptotic normality whose rate matches the original Ruppert-Polyak-Juditsky asymptotic covariance. This is the first time that both a provably optimal one-loop one-time-scale online CCA algorithm is proposed, and a local asymptotic normality rate is achieved.

Keywords: Local convergence, Polyak-Juditsky trajectory averaging, stochastic gradient descent, double stochastic sampling

1. Introduction

Nonconvex optimization has become the algorithmic engine powering many recent developments in statistics and machine learning. Advances in both theoretical understanding and algorithmic implementation have motivated the use of nonconvex optimization formulations with very large datasets, and the striking empirical discovery is that nonconvex models can be successful in this setting, despite the pessimism of classical worst-case analysis(Jin et al., 2019). In this paper, we consider the following general constrained nonconvex optimization problem:

$$\min_{\boldsymbol{v}} F(\boldsymbol{v}), \qquad \text{subject to } \boldsymbol{v} \in \mathcal{C}, \tag{1}$$

where $F(\boldsymbol{v})$ is a smooth and possibly non-convex objective function and $\mathcal{C}$ is a feasible set. The workhorse algorithm in this setting is stochastic gradient descent (SGD) and its variants (Robbins and Monro, 1951; Qian, 1999; Duchi et al., 2011; Kingma and Ba, 2015; Zhang and Sra, 2016). Given an unbiased estimate $\widetilde{\nabla} F(\boldsymbol{v};\boldsymbol{\zeta})$ of the gradient $\nabla F(\boldsymbol{v})$, SGD performs the following update at the t-th step ($t \geq 1$):

$$\boldsymbol{v}_t = \Pi_{\mathcal{C}}\left[\boldsymbol{v}_{t-1} - \eta \widetilde{\nabla} F(\boldsymbol{v}_{t-1};\boldsymbol{\zeta}_t)\right], \tag{2}$$

where $\eta > 0$ is a step size and $\Pi_{\mathcal{C}}$ is a projection operator onto the feasible set $\mathcal{C}$. SGD updates use only a single data point, or a small number of data points, and thus significantly reduce computational and storage complexities compared with offline algorithms, which require storing the full data set and evaluating the full gradient at each iteration.

In many applications, however, we do not have access to an unbiased estimate of $\nabla F(x)$ when we restrict access to a small number of data points. Instead, for each $\boldsymbol{v} \in \mathcal{C}$ we have access only to a stochastic vector $\Gamma(\boldsymbol{v};\boldsymbol{\zeta})$ which is an unbiased estimate of some *scaled* gradient:

$$\mathbb{E}_{\boldsymbol{\zeta}}\left[\Gamma(\boldsymbol{v};\boldsymbol{\zeta})\right] = D(\boldsymbol{v})\nabla F(\boldsymbol{v}), \tag{3}$$

where $D(\boldsymbol{v})$ is a deterministic positive scalar that depends on the current state $\boldsymbol{v}$. Examples of this setup arise most notably in generalized eigenvector (GEV) computation, which finds its applications in principal component analysis (Hotelling, 1933; Fan et al., 2018), partial least squares regression (Stone and Brooks, 1990), Fisher's linear discriminant analysis (Fisher, 1936; Welling, 2005), canonical correlation analysis (CCA) (Hotelling, 1936; Witten et al., 2009), sufficient dimension reduction (Li, 1991), and mixture models (Fan et al., 2018). Despite this wide range of applications, and their particular relevance to large-scale machine learning problems, there exist few rigorous general frameworks for SGD-based *online* learning using such models.

Our approach is a conceptually straightforward extension of SGD. We propose to continue to use (2) but with $\widetilde{\nabla} F(\boldsymbol{v}_{t-1};\boldsymbol{\zeta}_t)$ there replaced by $\Gamma(\boldsymbol{v}_{t-1};\boldsymbol{\zeta}_t)$. We refer this algorithm as the *Stochastic Scaled-Gradient Descent* (SSGD) algorithm. Specifically, at each step, SSGD performs the update:

$$\boldsymbol{v}_t = \Pi_{\mathcal{C}}\left[\boldsymbol{v}_{t-1} - \eta\Gamma(\boldsymbol{v}_{t-1};\boldsymbol{\zeta}_t)\right]. \tag{4}$$

We provide a theoretical analysis of this algorithm. While some of our analysis applies to the algorithm in full generality, our most useful results arise when we specialize to the online GEV problem. In this case we aim to minimize the generalized Rayleigh quotient given a unit spherical constraint:

$$\min_{\boldsymbol{v}} -\frac{\boldsymbol{v}^\top\mathbf{A}\boldsymbol{v}}{\boldsymbol{v}^\top\mathbf{B}\boldsymbol{v}}, \qquad \text{subject to } \boldsymbol{v}\in\mathbb{R}^d,\ \|\boldsymbol{v}\|=1. \tag{5}$$

The first-order derivative of the generalized Rayleigh quotient with respect to $\boldsymbol{v}$ is

$$\frac{\partial}{\partial\boldsymbol{v}}\left[-\frac{\boldsymbol{v}^\top\mathbf{A}\boldsymbol{v}}{\boldsymbol{v}^\top\mathbf{B}\boldsymbol{v}}\right] = -\frac{(\boldsymbol{v}^\top\mathbf{B}\boldsymbol{v})\mathbf{A}\boldsymbol{v} - (\boldsymbol{v}^\top\mathbf{A}\boldsymbol{v})\mathbf{B}\boldsymbol{v}}{(1/2)(\boldsymbol{v}^\top\mathbf{B}\boldsymbol{v})^2}. \tag{6}$$

As pointed out by Arora et al. (2012), the major stumbling block in applying SGD to this problem lies in obtaining an unbiased stochastic sample of the gradient (6), due to the fact that the objective function takes a fractional form of two expectations. In our approach we circumvent this issue by simply replacing the denominator on the right hand side of (6) by the constant 1 and use the following update:

$$\boldsymbol{v}_t = \Pi_{\mathcal{S}^{d-1}}\left[\boldsymbol{v}_{t-1} + \eta\left((\boldsymbol{v}_{t-1}^\top\widetilde{\mathbf{B}}'\boldsymbol{v}_{t-1})\widetilde{\mathbf{A}}\boldsymbol{v}_{t-1} - (\boldsymbol{v}_{t-1}^\top\widetilde{\mathbf{A}}\boldsymbol{v}_{t-1})\widetilde{\mathbf{B}}'\boldsymbol{v}_{t-1}\right)\right]. \tag{7}$$

We refer to the rule (7) as an *online GEV iteration*. In the special case where $\widetilde{\mathbf{B}} = \mathbf{I}$ a.s., (7) is identical (up to $O(\eta^2)$) to the Oja's algorithm (Oja, 1982; Jain et al., 2016; Li et al., 2018; Allen-Zhu and Li, 2017b).

To identify the iterative algorithm in (7) as a manifestation of SSGD, we rewrite the term in parentheses in the algorithm as follows:

$$\frac{(\boldsymbol{v}^\top\mathbf{B}\boldsymbol{v})^2}{2}\cdot\frac{(\boldsymbol{v}^\top\widetilde{\mathbf{B}}'\boldsymbol{v})\widetilde{\mathbf{A}}\boldsymbol{v} - (\boldsymbol{v}^\top\widetilde{\mathbf{A}}\boldsymbol{v})\widetilde{\mathbf{B}}'\boldsymbol{v}}{(1/2)(\boldsymbol{v}^\top\mathbf{B}\boldsymbol{v})^2}. \tag{8}$$

To proceed, we take $\widetilde{\mathbf{A}}$ and $\widetilde{\mathbf{B}}'$ as mutually independent and unbiased stochastic samples of $\mathbf{A}$ and $\mathbf{B}$ respectively. It can be easily seen that the expectation of (8) is a scaled gradient of the generalized Rayleigh quotient, where the scaling is the factor $(\boldsymbol{v}^\top\mathbf{B}\boldsymbol{v})^2/2$. This approach, which has been referred to as *double stochastic sampling* in the setting of kernel methods (Dai et al., 2014, 2016), makes it possible to develop an efficient stochastic approximation algorithm. Indeed, often $\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}}$ are of rank-1, so the computation of matrix-vector products $\tilde{\mathbf{A}}\boldsymbol{v}, \tilde{\mathbf{B}}'\boldsymbol{v}$ only invokes vector-vector inner products and is hence efficient.

Our contributions relative to previous work on nonconvex stochastic optimization as are follows. First, we propose a novel algorithm—the stochastic scaled-gradient descent algorithm (SSGD)—which generalizes the classical SGD algorithm and has a wider range of applications. Second, we provide a local convergence analysis for spherical-constraint objective functions that are locally convex. Starting with a warm initialization, our local convergence rate matches a known information-theoretic lower bound (Mei et al., 2018). Third, by applying SSGD to the GEV problem, we give a positive answer to the question raised by Arora et al. (2012) regarding to the existence of an efficient online GEV algorithm. Specifically, in the case of CCA, our algorithm (SSGD-CCA) uses as few as two samples at each update, does not incur intermediate and expensive computational cost while achieving a polynomial convergence rate guarantee (cf. Gao et al., 2019).

1.1. Related Literature

Several recent papers have focused on developing efficient algorithms for particular instances of generalized eigenvalue problems (Ge et al., 2016; Allen-Zhu and Li, 2017a; Yuan et al., 2018; Ma et al., 2015; Chaudhuri et al., 2009). For example, Yuan and Zhang (2013) and Ma (2013) studied iterative algorithms for sparse principal component analysis based on a power method with soft-thresholding steps. Tan et al. (2018) proposed a truncated Rayleigh flow algorithm to estimate the leading sparse generalized eigenvector which achieves a linear convergence rate.

Online learning of canonical eigenvectors has also been of interest. Recently, Gao et al. (2019) developed a streaming canonical correlation analysis (CCA) algorithm which involves solving a large linear system at each iteration. Independently, Arora et al. (2017) proposed a different online CCA algorithm which has temporal and spatial complexities that are quadratic in d.

In a recent concurrent work, Bhatia et al. (2018) studied the CCA problem and proposed a two-time-scale online iteration that they refer to as "Gen-Oja." The notion of two-time-scale analysis has been used widely in stochastic control and reinforcement learning (Borkar, 2008; Kushner and Yin, 2003), and the slow process in Gen-Oja is essentially Oja's iteration for online principal component estimation with Markovian noise (Oja, 1982; Jain et al., 2016; Li et al., 2018). Bhatia et al. (2018) obtained a convergence rate under the bounded sample assumptions that achieves the minimax rate $1/\sqrt{N}$ in terms of the sample size N. In comparison, our proposed SSGD algorithm is a one-time-scale algorithm with a single step size and requiring only two (independent) samples per iterate. Nonetheless it is minimax optimal with respect to local convergence and hence theoretically comparable with Gen-Oja. We will provide additional comparative remarks in Section 3.

1.2. Notation

For two sequences $\{a_n\}$ and $\{b_n\}$ of positive scalars, we denote $a_n \gtrsim b_n$ (resp. $a_n \lesssim b_n$) if $a_n \geq Cb_n$ (resp. $a_n \leq Cb_n$) for all n, and $a_n \asymp b_n$ if $a_n \gtrsim b_n$ and $a_n \lesssim b_n$ hold simultaneously. We also write $a_n = O(b_n), a_n = \Theta(b_n), a_n = \Omega(b_n)$ as $a_n \lesssim b_n, a_n \asymp b_n, a_n \gtrsim b_n$, respectively. We use $\|\boldsymbol{v}\|$ to denote the ℓ_2-norm of $\boldsymbol{v}$. Let $\lambda_{\max}(\mathbf{A}) = \|\mathbf{A}\|$ and $\lambda_{\min}(\mathbf{A})$ denote the maximal and minimal eigenvalues of a real symmetric matrix $\mathbf{A}$.

2. Main Result

In this section, we present a theoretical analysis of the SSGD algorithm for nonconvex optimization. To illustrate the core idea we focus on the case of a spherical constraint, $\boldsymbol{v} \in \mathcal{S}^{d-1}$, in which case SSGD reduces to the following update:

$$\boldsymbol{v}_t = \Pi_{\mathcal{S}^{d-1}} \left[\boldsymbol{v}_{t-1} - \eta \Gamma(\boldsymbol{v}_{t-1}; \boldsymbol{\zeta}_t) \right]. \tag{9}$$

Let $\mathcal{F}_t = \sigma\big(\boldsymbol{\zeta}_s : s \le t\big)$ be the filtration generated by the stochastic process $\boldsymbol{\zeta}_t$. Then, from (3), we ahve $\mathbb{E}[\Gamma(\boldsymbol{v}_{t-1}; \boldsymbol{\zeta}_t) \mid \mathcal{F}_{t-1}] = D(\boldsymbol{v}_{t-1}) \nabla F(\boldsymbol{v}_{t-1})$. That is, the conditional expectation is a scaled gradient. The ensuing analysis is analogous to that of locally convex SGD (Ge et al., 2015; Li et al., 2018) given we have appropriate Lipschitz-smoothness of the scalar function $D(\boldsymbol{v})$, but it requires delicate treatment given that SSGD effectively has a varying step size embodied in the scaling factor.

We begin by stating our assumptions in Section 2.1 and then state our main convergence results in Sections 2.2 and 2.3.

2.1. Settings and Assumptions

There is a rich literature on stochastic gradient methods on Riemannian manifolds (Ge et al., 2015; Zhang and Sra, 2016). Since we are working on unit sphere, we follow this literature and introduce a definition of manifold gradient and manifold Hessian in the presence of the unit spherical constraint $\mathcal{C} : c(\boldsymbol{v}) = (1/2)(\boldsymbol{v}^\top \boldsymbol{v} - 1) = 0$ (to simplify the derivation we incorporate a factor of $1/2$). For this equality-constrained optimization problem, we utilize the method of Lagrange multipliers and introduce the following Lagrangian function:

$$L(\boldsymbol{v}; \mu) = F(\boldsymbol{v}) - \frac{\mu}{2} \left(\boldsymbol{v}^\top \boldsymbol{v} - 1 \right).$$

We define the manifold gradient:

$$g(\boldsymbol{v}) = \nabla L(\boldsymbol{v}; \mu)\big|_{\mu = \mu^*(\boldsymbol{v})} = \nabla F(\boldsymbol{v}) - \frac{\boldsymbol{v}^\top \nabla F(\boldsymbol{v})}{\|\boldsymbol{v}\|^2} \boldsymbol{v}, \tag{10}$$

and the manifold Hessian:

$$\mathcal{H}(\boldsymbol{v}) = \nabla^2 L(\boldsymbol{v}; \mu)\big|_{\mu = \mu^*(\boldsymbol{v})} = \nabla^2 F(\boldsymbol{v}) - \frac{\boldsymbol{v}^\top \nabla F(\boldsymbol{v})}{\|\boldsymbol{v}\|^2} \mathbf{I}, \tag{11}$$

where $\mu^*(\boldsymbol{v}) = \|\boldsymbol{v}\|^{-2} \boldsymbol{v}^\top \nabla F(\boldsymbol{v})$ is the *optimal Lagrangian multiplier* defined by

$$\frac{\boldsymbol{v}^\top \nabla F(\boldsymbol{v})}{\|\boldsymbol{v}\|^2} = \operatorname*{argmin}_{\mu} \|\nabla L(\boldsymbol{v}; \mu)\| = \operatorname*{argmin}_{\mu} \|\nabla F(\boldsymbol{v}) - \mu \boldsymbol{v}\|.$$

For $\boldsymbol{v} \in \mathcal{S}^{d-1}$, we let $\mathcal{T}(\boldsymbol{v}) = \{\boldsymbol{u} : \boldsymbol{u}^\top \boldsymbol{v} = 0\}$ denote the tangent space of $\mathcal{S}^{d-1}$ at $\boldsymbol{v}$. To prove our main theoretical result, we need the following definitions and assumptions.

Definition 1 (Lipschitz Continuity) *Let $\mathcal{M}$ be a Banach space. The map $M : \mathbb{R}^d \mapsto \mathcal{M}$ is called L_M-Lipschitz, if for any two points $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{R}^d$*

$$\|M(\boldsymbol{v}) - M(\boldsymbol{v}')\|_{\mathcal{M}} \le L_M \|\boldsymbol{v} - \boldsymbol{v}'\|,$$

where $\|\cdot\|_{\mathcal{M}}$ is any norm properly defined in the Banach space $\mathcal{M}$.

For a fixed $\boldsymbol{v}$, define the state-dependent covariance $\boldsymbol{\Sigma}(\boldsymbol{v})$ to be

$$\boldsymbol{\Sigma}(\boldsymbol{v}) = \operatorname{var}\left(\Gamma(\boldsymbol{v};\boldsymbol{\zeta})\right) = \mathbb{E}\left[\left(\Gamma(\boldsymbol{v};\boldsymbol{\zeta}) - D(\boldsymbol{v})\nabla F(\boldsymbol{v})\right)\left(\Gamma(\boldsymbol{v};\boldsymbol{\zeta}) - D(\boldsymbol{v})\nabla F(\boldsymbol{v})\right)^{\top}\right]. \tag{12}$$

For the purposes of our analysis, we assume that the state-dependent parameter $D(\boldsymbol{v})$ and the Hessian $\nabla^2 F(\boldsymbol{v})$ are Lipschitz continuous within $\{\boldsymbol{v} : \|\boldsymbol{v}\| \le 1, \|\boldsymbol{v} - \boldsymbol{v}^*\| \le \delta\}$, where $\boldsymbol{v}^*$ is a local minimizer of the constrained optimization problem (5) and where $\delta \in (0, 1]$ is a fixed constant. Within this convex bounded compact space, we can also show $F(\boldsymbol{v}), \nabla F(\boldsymbol{v})$ to be Lipschitz continuous. We explicitly specify these constants in the following assumption.

Assumption 1 (Smoothness Assumption) *For any $\boldsymbol{v} \in \{\boldsymbol{v} : \|\boldsymbol{v}\| \le 1, \|\boldsymbol{v} - \boldsymbol{v}^*\| \le \delta\}$, we assume that $D(\boldsymbol{v})$ is L_D-Lipschitz, $F(\boldsymbol{v})$ is L_F-Lipschitz, $\nabla F(\boldsymbol{v})$ is L_K-Lipschitz and $\nabla^2 F(\boldsymbol{v})$ is L_Q-Lipschitz, where L_D, L_F, L_K, L_Q are fixed positive constants.*

For some fixed $\alpha > 0$, we assume that the stochastic vectors $\Gamma(\boldsymbol{v}_{t-1}; \boldsymbol{\zeta}_t), t \ge 1$ are *vector α-sub-Weibull* according to the following definition.

Assumption 2 (Sub-Weibull Tail) *For some fixed $\mathcal{V} \in (0, \infty)$ and for all $\boldsymbol{v} \in \mathcal{C}$, we assume that the stochastic vectors $\Gamma(\boldsymbol{v}; \boldsymbol{\zeta})$ satisfy*

$$\mathbb{E}\exp\left(\left\|\frac{\Gamma(\boldsymbol{v};\boldsymbol{\zeta})}{\mathcal{V}}\right\|^{\alpha}\right) \le 2.$$

The class of sub-Weibull distributions contains the common sub-Gaussian ($\alpha = 2$) and sub-Exponential ($\alpha = 1$) distribution classes as special cases (Kuchibhotla and Chakrabortty, 2018). Background on vector α-sub-Weibull distributions (and the associated notion of Orlicz ψ_α-norm) are provided in Appendix D.

2.2. Main Results

For notational simplicity, we denote

$$D = D(\boldsymbol{v}^*), \qquad \rho = D\left(2L_Q + \frac{5}{2}L_F + \frac{9}{2}L_K\right) + L_D(L_K + 2L_F). \tag{13}$$

For our local convergence analysis, we assume that the initialization $\boldsymbol{v}_0$ falls into the neighborhood of a local minimizer $\boldsymbol{v}^*$ of the constrained optimization problem; that is,

$$\|\boldsymbol{v}_0 - \boldsymbol{v}^*\| \le \min\left\{\frac{D\mu}{2^5\rho}, \delta\right\}, \tag{14}$$

where μ denotes the minimum positive eigenvalue of the manifold hessian $\mathcal{H}(\boldsymbol{v}^*)$:

$$\boldsymbol{v}_1^{\top}\mathcal{H}(\boldsymbol{v}^*)\boldsymbol{v}_1 \ge \mu, \quad \forall \boldsymbol{v}_1 \in \mathcal{T}(\boldsymbol{v}^*) \text{ and } \|\boldsymbol{v}_1\| = 1.$$

We note that the initialization condition (14) has a constant neighborhood radius that does not depend on dimension d.

To state our first main theorem on local convergence, we take $\epsilon \in (0, 1)$ and define the following quantities:

$$K_{\eta,\epsilon} = \left\lceil \log_2\left\{\frac{\sqrt{D^3\mu^3}}{2^5\rho\mathcal{V}\log^{\frac{\alpha+2}{2\alpha}}\epsilon^{-1}\cdot\eta^{1/2}}\right\}\right\rceil + 1, \tag{15}$$

and

$$T_\eta^* = \left\lceil \frac{2\log 2}{-\log(1 - D\mu\eta)} \right\rceil. \tag{16}$$

In the ensuing Theorem 2, $K_{\eta,\epsilon}T_\eta^*$ can be interpreted as the burn-in time for $\boldsymbol{v}_t$ to arrive in a $O(\eta^{1/2})$ neighborhood of local minimizer $\boldsymbol{v}^*$. We view every $T_\eta^* = \Theta\left((D\mu)^{-1}\eta^{-1}\right)$ iterations as one round and interpret $K_{\eta,\epsilon} = \Theta\left(\log \eta^{-1}\right)$ as the number of rounds.

Theorem 2 *Given Assumptions 1 and 2, assuming that anthed initialization condition (14) holds, for any positive constants η, ϵ that satisfy the scaling condition*

$$\eta \le \min\left\{ \frac{D^3\mu^3}{2^{24}G_\alpha^2\mathcal{V}^2\rho^2} \log^{-\frac{\alpha+2}{\alpha}} \epsilon^{-1},\ \frac{1}{D\mu} \right\}, \tag{17}$$

for all $T \ge K_{\eta,\epsilon}T_\eta^*$, there exists an event $\mathcal{H}_2$ with

$$\mathbb{P}(\mathcal{H}_2) \ge 1 - \left(14 + 8\left(\frac{3}{\alpha}\right)^{\frac{2}{\alpha}} \log^{-\frac{\alpha+2}{\alpha}} \epsilon^{-1} \right) T\epsilon, \tag{18}$$

such that on event $\mathcal{H}_2$ the iterates generated by the SSGD algorithm satisfy:

$$\|\boldsymbol{v}_t - \boldsymbol{v}^*\| \le \frac{2^{\frac{17}{2}}G_\alpha\mathcal{V}}{\sqrt{D\mu}} \log^{\frac{\alpha+2}{2\alpha}} \epsilon^{-1} \cdot \eta^{1/2},$$

for all $t \in [K_{\eta,\epsilon}T_\eta^*, T]$, where $G_\alpha \equiv \log_2^{1/\alpha}(1+e^{1/\alpha})\left(1 + \log_2^{1/\alpha}(1+e^{1/\alpha})\right)$ is a positive factor depending on α.

To prove Theorem 2, we define Δ_t as the projection of $\boldsymbol{v}_t - \boldsymbol{v}^*$ onto the tangent space $\mathcal{T}(\boldsymbol{v}^*)$, namely

$$\Delta_t = (\mathbf{I} - \boldsymbol{v}^*\boldsymbol{v}^{*\top})(\boldsymbol{v}_t - \boldsymbol{v}^*).$$

We first present a proposition that provides an upper bound on $\|\Delta_t\|$ over T iterations and characterizes the descent in $\|\Delta_t\|$ at the end of each round.

Proposition 3 *Assume Assumptions 1, 2 and initialization condition (14) hold. For any positive constants η, ϵ satisfying the scaling condition (17) and $T \ge 1$, with probability at least*

$$1 - \left(14 + 8\left(\frac{3}{\alpha}\right)^{\frac{2}{\alpha}} \log^{-\frac{\alpha+2}{\alpha}} \epsilon^{-1} \right) T\epsilon,$$

the algorithm iterates satisfy, for all $t \in [0, T]$,

$$\|\Delta_t\| \le \|\boldsymbol{v}_t - \boldsymbol{v}^*\| \le \sqrt{2}\|\Delta_t\|, \tag{19}$$

$$\|\Delta_t\| \le 4\max\left\{ \frac{\|\Delta_0\|}{2},\ \frac{2^6G_\alpha\mathcal{V}}{\sqrt{D\mu}} \log^{\frac{\alpha+2}{2\alpha}} \epsilon^{-1} \cdot \eta^{1/2} \right\}. \tag{20}$$

Moreover, if $T_\eta^* \in [0, T]$, we have:

$$\|\Delta_{T_\eta^*}\| \le \max\left\{ \frac{\|\Delta_0\|}{2},\ \frac{2^6G_\alpha\mathcal{V}}{\sqrt{D\mu}} \log^{\frac{\alpha+2}{2\alpha}} \epsilon^{-1} \cdot \eta^{1/2} \right\}. \tag{21}$$

The proof of Proposition 3 is provided in Appendix A.

By choosing an asymptotic regime such that $T\epsilon\log(1/\varepsilon)\to 0$, Proposition 3 states that (19), (20) and (21) hold with probability tending to one. On that high-probability event, (19) indicates that $\|\boldsymbol{v}_t-\boldsymbol{v}^*\|$ and its projection in the tangent space $\|\Delta_t\|$ are bounded by each other up to constant factors, (20) guarantees that $\|\Delta_t\|$ does not exceed $\max\left\{2\|\Delta_0\|,\Theta(\eta^{1/2})\right\}$—that is, $\boldsymbol{v}_t$ stays in a neighborhood of local minimizer $\boldsymbol{v}^*$—and (21) states that, for $\|\Delta_0\|=\Omega(\eta^{1/2})$, $\|\Delta_t\|$ decreases by half after T_η^* iterations: $\|\Delta_{T_\eta^*}\|\le\max\left\{\|\Delta_0\|/2,\Theta(\eta^{1/2})\right\}$.

Proposition 3 studies Δ_t in a single round, i.e., for T_η^* iterations. Theorem 2 is proved by applying Proposition 3 repeatedly for $K_{\eta,\epsilon}$ rounds.

Proof [Proof of Theorem 2] We recall the definition of $K_{\eta,\epsilon}$ in (15). Because $\{\Delta_t\}$ is a Markov process, we can apply Proposition 3 repeatedly for $K_{\eta,\epsilon}$ rounds, initializing each round with the output $\Delta_{T_\eta^*}$ from the previous round. For any $t\in[K_{\eta,\epsilon}T_\eta^*,T]$, we first apply (21) in Proposition 3 for $K_{\eta,\epsilon}$ rounds, then apply (20) for $t-K_{\eta,\epsilon}T_\eta^*$ iterations, and use (19) to conclude that

$$\begin{aligned}\|\boldsymbol{v}_t-\boldsymbol{v}^*\|&\le\sqrt{2}\|\Delta_t\|\le\sqrt{2}\cdot 4\max\left\{\frac{\|\Delta_{K_{\eta,\epsilon}T_\eta^*}\|}{2},\ \frac{2^6G_\alpha\mathcal{V}}{\sqrt{D\mu}}\log^{\frac{\alpha+2}{2\alpha}}\epsilon^{-1}\cdot\eta^{1/2}\right\}\\&\le 4\sqrt{2}\cdot\max\left\{\frac{\|\Delta_0\|}{2^{K_{\eta,\epsilon}}},\frac{2^6G_\alpha\mathcal{V}}{\sqrt{D\mu}}\log^{\frac{\alpha+2}{2\alpha}}\epsilon^{-1}\cdot\eta^{1/2}\right\}\le\frac{2^{\frac{17}{2}}G_\alpha\mathcal{V}}{\sqrt{D\mu}}\log^{\frac{\alpha+2}{2\alpha}}\epsilon^{-1}\cdot\eta^{1/2},\end{aligned}$$

where the last inequality is due to initialization condition (14). By taking a union bound over $K_{\eta,\epsilon}$ rounds and $T-K_{\eta,\epsilon}T_\eta^*$ iterations, we obtain

$$\mathbb{P}(\mathcal{H}_2)\ge 1-\left(14+8\left(\frac{3}{\alpha}\right)^{\frac{2}{\alpha}}\log^{-\frac{\alpha+2}{\alpha}}\epsilon^{-1}\right)T\epsilon.$$

■

Theorem 2 establishes the local convergence of $\boldsymbol{v}_t$ in a neighborhood of $\boldsymbol{v}^*$ for a fixed stepsize η and a number of iterations $T\ge K_{\eta,\epsilon}T_\eta^*$. The following corollary provides a finite-sample version of Theorem 2.

Corollary 4 (Finite Sample) *Assume Assumptions 1 and 2 and the initialization condition* (14)*. For sample size T set the step size as follows:*

$$\eta(T)=\Theta\left(\frac{\log T}{D\mu T}\right).$$

For fixed positive constants ϵ, T satisfying the scaling condition

$$\eta(T)\le\min\left\{\frac{D^3\mu^3}{2^{24}G_\alpha\mathcal{V}^2\rho^2}\log^{-\frac{\alpha+2}{\alpha}}\epsilon^{-1},\ \frac{1}{D\mu}\right\},$$

there exists an event $\mathcal{H}_4$ with

$$\mathbb{P}(\mathcal{H}_4)\ge 1-\left(14+8\left(\frac{3}{\alpha}\right)^{\frac{2}{\alpha}}\log^{-\frac{\alpha+2}{\alpha}}\epsilon^{-1}\right)T\epsilon,$$

such that on the event $\mathcal{H}_4$ the iterates generated by the SSGD algorithm satisfy

$$\|\boldsymbol{v}_T-\boldsymbol{v}^*\|\lesssim\frac{G_\alpha\mathcal{V}}{D\mu}\log^{\frac{\alpha+2}{2\alpha}}\epsilon^{-1}\sqrt{\frac{\log T}{T}}.$$

We notice that our Theorem 2 and Corollary 4 provide a *dimension-free* local convergence rate when $\mathcal{V}$ is $O(1)$. As we will see later in the example of CCA, the ($\alpha = 1/2$) sub-Weibull parameter $\mathcal{V}$ in that case scales with $\sqrt{d}$ and thus the local rate is the minimax-optimal rate $O(\sqrt{d/T})$ up to a polylogarithmic factor.

2.3. Asymptotic Normality via Trajectory Averaging

This subsection is devoted to studying trajectory averaging for SSGD. We generalize Polyak-Juditsky's classical asymptotic normality result for a locally convex objective (Polyak and Juditsky, 1992).

We denote $\mathcal{H}_* \equiv \mathcal{H}(\boldsymbol{v}^*), \boldsymbol{\Sigma}_* \equiv \boldsymbol{\Sigma}(\boldsymbol{v}^*)$ and $D \equiv D(\boldsymbol{v}^*)$. We define

$$\mathcal{M}_* = (\mathbf{I} - \boldsymbol{v}^* \boldsymbol{v}^{*\top}) \mathcal{H}_* (\mathbf{I} - \boldsymbol{v}^* \boldsymbol{v}^{*\top}).$$

From the initialization condition (14), we have $\boldsymbol{u}^\top \mathcal{M}_* \boldsymbol{u} \ge \mu \|\boldsymbol{u}\|^2$ for all $\boldsymbol{u} \in \mathbb{R}^d$. We consider the eigen-decomposition $\mathcal{M}_* = \boldsymbol{P} \mathrm{diag}(\lambda_1, \dots, \lambda_{d-1}, 0) \boldsymbol{P}^\top$ for an orthogonal matrix $\boldsymbol{P} \in \mathbb{R}^{d \times d}$ and eigenvalues $\lambda_1 \ge \dots \ge \lambda_{d-1} > 0$ with minimum positive eigenvalue $\lambda_{d-1} = \mu$. We take the inverse of all positive eigenvalues and define the following matrix

$$\mathcal{M}_*^- \equiv \boldsymbol{P} \mathrm{diag}(\lambda_1^{-1}, \dots, \lambda_{d-1}^{-1}, 0) \boldsymbol{P}^\top. \tag{22}$$

$\mathcal{M}_*^-$ can be interpreted as the inverse of $\mathcal{M}_*$ in the $(d-1)$-dimensional tangent space $\mathcal{T}(\boldsymbol{v}^*)$, and we can easily find $\mathcal{M}_*^- \boldsymbol{v}^* = \mathbf{0}$.

As shown in Theorem 2, we need $K_{\eta,\epsilon} T_\eta^*$ iterations for $\boldsymbol{v}_t$ to fall in a $\Theta(\eta^{1/2})$ neighborhood of the local minimizer $\boldsymbol{v}^*$. For $T \ge K_{\eta,\epsilon} T_\eta^*$, we define the trajectory average over time $K_{\eta,\epsilon} T_\eta^* + 1, \dots, T$ as follows:

$$\overline{\boldsymbol{v}}_T^{(\eta)} \equiv \frac{1}{T - K_{\eta,\epsilon} T_\eta^*} \sum_{t = K_{\eta,\epsilon} T_\eta + 1}^{T} \boldsymbol{v}_t, \tag{23}$$

where we add the superscript (η) to emphasize the choice of η. We notice that $\{\overline{\boldsymbol{v}}_T^{(\eta)}\}_{T,\eta}$ is a triangular array over a continuum η. To obtain asymptotic normality of trajectory average $\overline{\boldsymbol{v}}_T^{(\eta)}$, we additionally make the following local Lipschitz continuous assumption on stochastic scaled-gradient $\Gamma(\boldsymbol{v}; \zeta)$ in the neighborhood of $\boldsymbol{v}^*$.

Assumption 3 *There exists a positive constant L_S such that for all $\boldsymbol{v}, \boldsymbol{v}' \in \{\boldsymbol{v} : \|\boldsymbol{v}\| \le 1, \|\boldsymbol{v} - \boldsymbol{v}^*\| \le \delta\}$ and $t \ge 1$, we have*

$$\mathbb{E}\left[\left\| \Gamma(\boldsymbol{v}; \boldsymbol{\zeta}_t) - \Gamma(\boldsymbol{v}'; \boldsymbol{\zeta}_t) \right\|^2 \middle| \mathcal{F}_{t-1} \right] \le L_S^2 \|\boldsymbol{v} - \boldsymbol{v}'\|^2. \tag{24}$$

The following theorem states that the trajectory average $\overline{\boldsymbol{v}}_T^{(\eta)}$ converges in distribution to a $(d-1)$-dimensional normal distribution in the tangent space $\mathcal{T}(\boldsymbol{v}^*)$.

Theorem 5 (Asymptotic Normality) *Assume Assumptions 1, 2, 3 and initialization condition* (14). *If we choose the step size η such that $\eta \to 0$ as the total sample size $T \to \infty$, where*

$$T\eta^2 \log^{\frac{2\alpha+4}{\alpha}} T \to 0, \quad T\eta \log^{-\frac{\alpha+2}{\alpha}} T \to \infty \quad \textit{a.s.} \tag{25}$$

we obtain convergence in distribution:

$$\sqrt{T} \left(\overline{\boldsymbol{v}}_T^{(\eta)} - \boldsymbol{v}^* \right) \xrightarrow{d} N\left(\mathbf{0}, D^{-2} \cdot \mathcal{M}_*^- \boldsymbol{\Sigma}_* \mathcal{M}_*^- \right). \tag{26}$$

Theorem 5 is broadly consistent with the classical asymptotic normality result that is obtained when minimizing a strongly convex objective function in an Euclidean space using stochastic gradient descent Polyak and Juditsky (1992). Indeed, in the case of a diminishing step size, $\eta(t) \propto t^{-\alpha}$, $\alpha \in (1/2, 1)$, SGD with trajectory averaging converges in distribution to an analogous normal distribution. In contrast, due to our choice of a constant step size that is asymptotically small with $\eta \propto T^{-\alpha}$ up to a polylogarithmic factor, our trajectory averaging begins only after "the burn-in phase"; that is, after $K_{\eta,\varepsilon} T_\eta^*$ iterates.

To prove Theorem 5, we first present the following lemma on a linear representation of $\mathcal{M}_*(\overline{\boldsymbol{v}}_T^{(\eta)} - \boldsymbol{v}^*)$. The proof is deferred to Appendix B.9.

Lemma 6 (Representation Lemma) *Under Assumptions 1, 2 and given initialization condition (14), for any $T \geq K_{\eta,\epsilon} T_\eta^*$ and positive constants η, ϵ satisfying the scaling condition*

$$5\mathcal{V} \log^{1/\alpha} \epsilon^{-1} \cdot \eta \leq 1,$$

we have

$$\begin{aligned}\mathcal{M}_*\left(\overline{\boldsymbol{v}}_T^{(\eta)} - \boldsymbol{v}^*\right) = \frac{1}{D(T - K_{\eta,\epsilon}T_\eta^*)} \sum_{t=K_{\eta,\epsilon}T_\eta^*+1}^{T} \boldsymbol{\chi}_{t+1} + \frac{1}{D(T - K_{\eta,\epsilon}T_\eta^*)} \sum_{t=K_{\eta,\epsilon}T_\eta^*+1}^{T} \boldsymbol{S}_{t+1} \\ + \frac{\eta}{D(T - K_{\eta,\epsilon}T_\eta^*)} \sum_{t=K_{\eta,\epsilon}T_\eta^*+1}^{T} \boldsymbol{P}_{t+1} + \frac{1}{D(T - K_{\eta,\epsilon}T_\eta^*)\eta}(\Delta_{K_{\eta,\epsilon}T_\eta^*+1} - \Delta_{T+1}),\end{aligned} \tag{27}$$

where $\boldsymbol{\chi}_t, \boldsymbol{S}_t, \boldsymbol{P}_t$ are vectors in the tangent space $\mathcal{T}(\boldsymbol{v}^*)$. Here $\boldsymbol{\chi}_t$ is defined as

$$\boldsymbol{\chi}_t \equiv (\mathbf{I} - \boldsymbol{v}^* \boldsymbol{v}^{*\top})(\Gamma(\boldsymbol{v}_{t-1}; \boldsymbol{\zeta}_t) - D(\boldsymbol{v}_{t-1})\nabla F(\boldsymbol{v}_{t-1})), \tag{28}$$

which is α-sub-Weibull with parameter $G_\alpha \mathcal{V}$. The sequence $\{\boldsymbol{\chi}_t\}$ forms a vector-valued martingale difference sequence with respect to $\mathcal{F}_t$. $\boldsymbol{S}_t$ satisfies $\|\boldsymbol{S}_t\| \leq \rho \|\boldsymbol{v}_{t-1} - \boldsymbol{v}^*\|^2$. On the event $\mathcal{H}_2$ defined in Theorem 2, using a total sample size $T+1$, each $\boldsymbol{P}_t$ satisfies $\|\boldsymbol{P}_t\| \leq 7\mathcal{V}^2 \log^{2/\alpha} \epsilon^{-1}$.

With Lemma 6 in hand, we are ready to prove Theorem 5.

Proof [Proof of Theorem 5] For a given T, we apply Theorem 2 and Lemma 6 with $\epsilon = 1/T^2$, such that $\mathbb{P}(\mathcal{H}_2) \to 1$ and the scaling condition (17) is satisfied under condition (25). Using a coupling approach we can safely ignore the small probability event and concentrate on the event $\mathcal{H}_2$, where we have

$$\left\| \frac{1}{D(T - K_{\eta,\epsilon}T_\eta^*)} \sum_{t=K_{\eta,\epsilon}T_\eta^*+1}^{T} \boldsymbol{S}_{t+1} \right\| \leq \frac{2^{\frac{\alpha+2}{\alpha}+17} \rho G_\alpha^2 \mathcal{V}^2}{D^2 \mu} \eta \log^{\frac{\alpha+2}{\alpha}} T,$$

$$\left\| \frac{\eta}{D(T - K_{\eta,\epsilon}T_\eta^*)} \sum_{t=K_{\eta,\epsilon}T_\eta^*+1}^{T} \boldsymbol{P}_{t+1} \right\| \leq \frac{7 \cdot 2^{\frac{2}{\alpha}} \mathcal{V}^2}{D} \eta \log^{\frac{2}{\alpha}} T.$$

Using the relation $\|\Delta_t\| \leq \|\boldsymbol{v}_t - \boldsymbol{v}^*\| \leq \sqrt{2}\|\Delta_t\|$, given in Proposition 3, and applying Theorem 2, on event $\mathcal{H}_2$ we also have

$$\left\| \frac{1}{D(T - K_{\eta,\epsilon}T_\eta^*)\eta}(\Delta_{K_{\eta,\epsilon}T_\eta^*+1} - \Delta_{T+1}) \right\| \leq \frac{2^{\frac{\alpha+2}{2\alpha}+\frac{17}{2}+1} G_\alpha \mathcal{V}}{\sqrt{D^3 \mu}} \frac{\log^{\frac{\alpha+2}{2\alpha}} T}{(T - K_{\eta,\epsilon}T_\eta^*)\eta^{1/2}}.$$

Under condition (25), as $T \to \infty, \eta \to 0$, we have the following almost sure convergences

$$\frac{\sqrt{T}}{D(T - K_{\eta,\epsilon}T_\eta^*)} \sum_{t=K_{\eta,\epsilon}T_\eta^*+1}^{T} \boldsymbol{S}_{t+1} \to \mathbf{0} \quad \text{a.s.}$$

$$\frac{\eta\sqrt{T}}{D(T - K_{\eta,\epsilon}T_\eta^*)} \sum_{t=K_{\eta,\epsilon}T_\eta^*+1}^{T} \boldsymbol{P}_{t+1} \to \mathbf{0} \quad \text{a.s.}$$

$$\frac{\sqrt{T}}{D(T - K_{\eta,\epsilon}T_\eta^*)\eta} (\Delta_{K_{\eta,\epsilon}T_\eta^*+1} - \Delta_{T+1}) \to \mathbf{0} \quad \text{a.s.}$$

From (12) and (28), the covariance matrix of $\boldsymbol{\xi}_t$—i.e., the projection of scaled-gradient noise onto the tangent space $\mathcal{T}(\boldsymbol{v}^*)$—can be denoted by

$$\boldsymbol{\Phi}(\boldsymbol{v}_{t-1}) \equiv (\mathbf{I} - \boldsymbol{v}^*\boldsymbol{v}^{*\top})\boldsymbol{\Sigma}(\boldsymbol{v}_{t-1})(\mathbf{I} - \boldsymbol{v}^*\boldsymbol{v}^{*\top}).$$

We denote the covariance matrix at local minimizer $\boldsymbol{v}^*$ as $\boldsymbol{\Phi}_* \equiv \boldsymbol{\Phi}(\boldsymbol{v}^*) = (\mathbf{I} - \boldsymbol{v}^*\boldsymbol{v}^{*\top})\boldsymbol{\Sigma}_*(\mathbf{I} - \boldsymbol{v}^*\boldsymbol{v}^{*\top})$. Using the central limit theorem and the Slutsky theorem, we have the following convergence-in-distribution result under the condition (25) as $T \to \infty, \eta \to 0$:

$$\frac{1}{\sqrt{T}} \sum_{t=K_{\eta,\epsilon}T_\eta^*+1}^{T} \boldsymbol{\chi}_{t+1} \xrightarrow{d} N\left(\mathbf{0}, \boldsymbol{\Phi}_*\right).$$

Combining these results with (27) in Lemma 6, under condition (25), as $T \to \infty, \eta \to 0$ we have convergence in distribution:

$$\sqrt{T}\mathcal{M}_* \left(\overline{\boldsymbol{v}}_T^{(\eta)} - \boldsymbol{v}^*\right) \xrightarrow{d} N(\mathbf{0}, D^{-2} \cdot \boldsymbol{\Phi}_*). \tag{29}$$

Since $\mathcal{M}_*^-\mathcal{M}_* = \mathbf{I} - \boldsymbol{v}^*\boldsymbol{v}^{*\top}$ and $\mathcal{M}_*^-\boldsymbol{\Phi}_*\mathcal{M}_*^- = \mathcal{M}_*^-\boldsymbol{\Sigma}_*\mathcal{M}_*^-$, (29) is equivalent to

$$\sqrt{T}(\mathbf{I} - \boldsymbol{v}^*\boldsymbol{v}^{*\top})\left(\overline{\boldsymbol{v}}_T^{(\eta)} - \boldsymbol{v}^*\right) \xrightarrow{d} N\left(\mathbf{0}, D^{-2} \cdot \mathcal{M}_*^-\boldsymbol{\Sigma}_*\mathcal{M}_*^-\right), \tag{30}$$

which omits the asymptotic analysis in the direction parallel to $\boldsymbol{v}^*$. To study the asymptotic property of $\boldsymbol{v}^*\boldsymbol{v}^{*\top}(\overline{\boldsymbol{v}}_T^{(\eta)} - \boldsymbol{v}^*)$, we first notice that for all $\boldsymbol{v} \in \mathbb{R}^d$ with $\|\boldsymbol{v}\| = 1$, $\|\boldsymbol{v}^*\boldsymbol{v}^{*\top}(\boldsymbol{v} - \boldsymbol{v}^*)\| = 1 - \boldsymbol{v}^{*\top}\boldsymbol{v} = \frac{1}{2}\|\boldsymbol{v} - \boldsymbol{v}^*\|^2$. Applying Theorem 2, on event $\mathcal{H}_2$ we have:

$$\begin{aligned} \left\|\sqrt{T} \cdot \boldsymbol{v}^*\boldsymbol{v}^{*\top}(\overline{\boldsymbol{v}}_T^{(\eta)} - \boldsymbol{v}^*)\right\| &= \frac{1}{2\sqrt{T}} \sum_{t=K_{\eta,\epsilon}T_\eta^*+1}^{T} \|\boldsymbol{v}_t - \boldsymbol{v}^*\|^2 \\ &\leq \frac{2^{\frac{\alpha+2}{\alpha}+17}G_\alpha^2\mathcal{V}^2}{D\mu} \cdot \frac{\eta(T - K_{\eta,\epsilon}T_\eta^*)\log^{\frac{\alpha+2}{\alpha}} T}{\sqrt{T}} \lesssim \sqrt{\eta^2 T \log^{\frac{2\alpha+4}{\alpha}} T} \to 0, \end{aligned}$$

where in the second line we used the first condition in (25). Under condition (25), as $T \to \infty, \eta \to 0$, we have almost sure convergence

$$\sqrt{T} \cdot \boldsymbol{v}^*\boldsymbol{v}^{*\top}\left(\overline{\boldsymbol{v}}_T^{(\eta)} - \boldsymbol{v}^*\right) \to \mathbf{0} \quad \text{a.s.} \tag{31}$$

Summing (30) and (31) and applying the Slutsky theorem, we conclude (26) and Theorem 5. ■

3. An Application to the GEV Problem

For the generalized eigenvector problem, the objective function of interest is

$$F(\boldsymbol{v}) = -\frac{\boldsymbol{v}^\top \mathbf{A}\boldsymbol{v}}{\boldsymbol{v}^\top \mathbf{B}\boldsymbol{v}}, \qquad \text{such that } \ c(\boldsymbol{v}) = \|\boldsymbol{v}\|^2 - 1, \tag{32}$$

where $\mathbf{A}$ and $\mathbf{B}$ are two real symmetric positive-definite matrices. In the following lemma, we verify that the objective function $F(\boldsymbol{v})$ in (32) satisfies Assumption 1; that is, $D(\boldsymbol{v}), F(\boldsymbol{v}), \nabla F(\boldsymbol{v}), \nabla^2 F(\boldsymbol{v})$ are Lipschitz continuous within $\{\boldsymbol{v} : \|\boldsymbol{v}\| \le 1, \|\boldsymbol{v} - \boldsymbol{v}^*\| \le \delta\}$.

Lemma 7 *Assumption 1 holds for $F(\boldsymbol{v})$ in GEV problem (32) with constants*

$$L_D = 2\|\mathbf{B}\|^2, L_F = \frac{4\|\mathbf{A}\|\|\mathbf{B}\|}{(1-\delta)^2\lambda_{\min}^2(\mathbf{B})}, L_K = \frac{28\|\mathbf{A}\|\|\mathbf{B}\|^2}{(1-\delta)^3\lambda_{\min}^3(\mathbf{B})}, L_Q = \frac{232\|\mathbf{A}\|\|\mathbf{B}\|^3}{(1-\delta)^4\lambda_{\min}^4(\mathbf{B})}.$$

The proof of Lemma 7 is deferred to Appendix C.1. With the Lipschitz parameters given above, we consider the initialization condition (14). The neighborhood radius on the right-hand side of (14) can be viewed as a function of δ that is maximized at some $\delta^* \in (0, 1)$, when all other constants are fixed. The region covered in the local convergence analysis is maximized with such a choice of δ^*.

3.1. CCA Example

The GEV problem is widely applicable to many statistical machine learning tasks. We focus on the example of CCA as a core application; we refer to Tan et al. (2018) for other (sparse, high-dimension) applications including linear discriminant analysis, sliced inverse regression, etc.

Canonical Correlation Analysis (CCA) aims at maximizing the correlation between two transformed vectors. Given $\mathbf{X}$ and $\mathbf{Y}$ as two column vectors, let $\mathbf{\Sigma_{XY}}$ be the cross-covariance matrix between $\mathbf{X}$ and $\mathbf{Y}$. $\mathbf{\Sigma_{XX}}$ and $\mathbf{\Sigma_{YY}}$ are the covariance matrices of $\mathbf{X}$ and $\mathbf{Y}$, respectively. The CCA problem is a special case of the GEV problem (5) with

$$\mathbf{A} = \begin{pmatrix} \mathbf{0} & \mathbf{\Sigma_{XY}} \\ \mathbf{\Sigma_{YX}} & \mathbf{0} \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} \mathbf{\Sigma_{XX}} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma_{YY}} \end{pmatrix}.$$

To obtain $\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}}'$ as mutually independent and unbiased stochastic samples of $\mathbf{A}$ and $\mathbf{B}$, we draw two independent pairs of samples $(\mathbf{X}, \mathbf{Y}), (\mathbf{X}', \mathbf{Y}')$ at each iteration and compute

$$\widetilde{\mathbf{A}} = \begin{pmatrix} \mathbf{0} & \mathbf{X}\mathbf{Y}^\top \\ \mathbf{Y}\mathbf{X}^\top & \mathbf{0} \end{pmatrix}, \qquad \widetilde{\mathbf{B}}' = \begin{pmatrix} \mathbf{X}'\mathbf{X}'^\top & \mathbf{0} \\ \mathbf{0} & \mathbf{Y}'\mathbf{Y}'^\top \end{pmatrix},$$

where all samples of $\mathbf{X}, \mathbf{Y}$ are centralized such that they have expectation zero. As a special case of (7), the SSGD update of CCA is formalized in Algorithm 3.1.

In order to apply the convergence results of SSGD algorithm to the CCA problem, it remains to verify Assumption 2. It is standard to assume that the samples $\mathbf{X} \in \mathbb{R}^{d_x}, \mathbf{Y} \in \mathbb{R}^{d_y}$ follow sub-Gaussian distributions (Gao et al., 2019; Li et al., 2018); that is, there exist fixed positive constants $\mathcal{V}_x, \mathcal{V}_y$ such that

$$\mathbb{E}\exp\left(\left\|\frac{\mathbf{X}}{\mathcal{V}_x}\right\|^2\right) \le 2, \quad \mathbb{E}\exp\left(\left\|\frac{\mathbf{Y}}{\mathcal{V}_y}\right\|^2\right) \le 2.$$

With these standard assumptions for the samples $\mathbf{X}, \mathbf{Y}$, the following lemma shows that the scaled-gradient noise in the CCA problem satisfies Assumption 2 with parameter $\alpha = 1/2$. The proof is provided in Appendix C.2.

Algorithm 1 Online Canonical Correlation Analysis by Stochastic Scaled-Gradient Descent (CCA-SSGD)

```
Pick the total sample size $T$ and step size $\eta$ appropriately
Obtain warm initialization $\boldsymbol{v}_0$ using prior information
for $t = 1, \dots, T/2$ do
    Draw mutually independent sample pairs $(\mathbf{X}, \mathbf{Y})$ and $(\mathbf{X}', \mathbf{Y}')$ from the stochastic oracle
    Compute unbiased estimates
        $\widetilde{\mathbf{A}} = \begin{pmatrix} \mathbf{0} & \mathbf{X}\mathbf{Y}^\top \\ \mathbf{Y}\mathbf{X}^\top & \mathbf{0} \end{pmatrix}, \qquad \widetilde{\mathbf{B}}' = \begin{pmatrix} \mathbf{X}'\mathbf{X}'^\top & \mathbf{0} \\ \mathbf{0} & \mathbf{Y}'\mathbf{Y}'^\top \end{pmatrix}$
    Update $\boldsymbol{v}$ using the following rule
        $\boldsymbol{v}_t \leftarrow \Pi_{\mathcal{S}^{d-1}} \left[ \boldsymbol{v}_{t-1} + \eta \left( (\boldsymbol{v}_{t-1}^\top \widetilde{\mathbf{B}}' \boldsymbol{v}_{t-1}) \widetilde{\mathbf{A}} \boldsymbol{v}_{t-1} - (\boldsymbol{v}_{t-1}^\top \widetilde{\mathbf{A}} \boldsymbol{v}_{t-1}) \widetilde{\mathbf{B}}' \boldsymbol{v}_{t-1} \right) \right]$
end for
Return $\boldsymbol{v}_T$
```

Lemma 8 *Assumption 2 holds in the CCA problem with constants* $\alpha = \frac{1}{2}$, $\mathcal{V} = 400(\mathcal{V}_x^2 + \mathcal{V}_y^2)\mathcal{V}_x\mathcal{V}_y$.

Lemmas 7 and 8 certify that Assumptions 1 and 2 hold in CCA settings and hence local convergence Corollary 4 applies. As a comparison, Gao et al. (2019) considers a measure of error

$$\operatorname{align}(\boldsymbol{v}, \boldsymbol{v}^*) \equiv \frac{1}{2} \left(\frac{\boldsymbol{v}_x^\top \boldsymbol{\Sigma}_{\mathbf{XX}} \boldsymbol{v}_x^*}{\sqrt{\boldsymbol{v}_x{}^\top \boldsymbol{\Sigma}_{\mathbf{XX}} \boldsymbol{v}_x} \sqrt{\boldsymbol{v}_x^*{}^\top \boldsymbol{\Sigma}_{\mathbf{XX}} \boldsymbol{v}_x^*}} + \frac{\boldsymbol{v}_y^\top \boldsymbol{\Sigma}_{\mathbf{YY}} \boldsymbol{v}_y^*}{\sqrt{\boldsymbol{v}_y{}^\top \boldsymbol{\Sigma}_{\mathbf{YY}} \boldsymbol{v}_y} \sqrt{\boldsymbol{v}_y^*{}^\top \boldsymbol{\Sigma}_{\mathbf{YY}} \boldsymbol{v}_y^*}} \right) \asymp \boldsymbol{v}^\top \boldsymbol{v}^*,$$

where $\boldsymbol{v}^\top = (\boldsymbol{v}_x^\top, \boldsymbol{v}_y^\top)$ and $\boldsymbol{v}^{*\top} = (\boldsymbol{v}_x^{*\top}, \boldsymbol{v}_y^{*\top})$ are partitioned by dimensions d_x, d_y. A lower bound for Gaussian inputs is proved $1 - \operatorname{align}(\boldsymbol{v}, \boldsymbol{v}^*) \gtrsim d/T$. By noticing $1 - \boldsymbol{v}^\top \boldsymbol{v}^* = \|\boldsymbol{v} - \boldsymbol{v}^*\|/2$ under our spherical constraint, the lower bound can be translated into our setting as $\|\boldsymbol{v}_T - \boldsymbol{v}^*\| \gtrsim \sqrt{d/T}$, which matches Corollary 4 in terms of sample size T. We notice that our Corollary 4 and Gao et al. (2019) appear with different dimensional dependency, which is due to assumption differences. The parameter $\mathcal{V}$ of vector sub-Weibull distribution in our Assumption 2 implicitly contains a factor $\sqrt{d}$, indicating that Corollary 4 matches the lower bound in Gao et al. (2019) in terms of dimension d as well.

4. Discussion

We have presented Stochastic Scaled-Gradient Descent algorithm for minimizing a constrained nonconvex objective function. Comparing with classical stochastic gradient descent, our method only requires access to an unbiased estimate of a scaled gradient, providing access to a broader range of applications. The proposed algorithm requires only a single pass through the data and is memory-efficient, with storage complexity linearly dependent on the ambient dimensionality of the problem. For a class of nonconvex stochastic optimization problems, we establish local convergence rates of the proposed algorithm to local minimizers and we prove asymptotic normality of trajectory average. An application to the generalized eigenvalue problem is investigated. In the near future we will investigate the rate of escape of saddle points for SSGD, and study global convergence for generic Riemannian manifolds.

References

Zeyuan Allen-Zhu and Yuanzhi Li. Doubly accelerated methods for faster cca and generalized eigendecomposition. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 98–106. JMLR. org, 2017a.

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

Raman Arora, Andrew Cotter, Karen Livescu, and Nathan Srebro. Stochastic optimization for PCA and PLS. In *Communication, Control, and Computing (Allerton), 2012 50th Annual Allerton Conference on*, pages 861–868, 2012.

Raman Arora, Teodor Vanislavov Marinov, Poorya Mianjy, and Nati Srebro. Stochastic approximation for canonical correlation analysis. In *Advances in Neural Information Processing Systems*, pages 4775–4784, 2017.

Kush Bhatia, Aldo Pacchiano, Nicolas Flammarion, Peter L Bartlett, and Michael I Jordan. Gen-oja: Simple & efficient algorithm for streaming generalized eigenvector computation. In *Advances in Neural Information Processing Systems*, pages 7016–7025, 2018.

Vivek S Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge University Press, 2008.

Kamalika Chaudhuri, Sham M Kakade, Karen Livescu, and Karthik Sridharan. Multi-view clustering via canonical correlation analysis. In *Proceedings of the 26th annual international conference on machine learning*, pages 129–136. ACM, 2009.

Bo Dai, Bo Xie, Niao He, Yingyu Liang, Anant Raj, Maria-Florina F Balcan, and Le Song. Scalable kernel methods via doubly stochastic gradients. In *Advances in Neural Information Processing Systems*, pages 3041–3049, 2014.

Bo Dai, Niao He, Yunpeng Pan, Byron Boots, and Le Song. Learning from conditional distributions via dual embeddings. *arXiv preprint arXiv:1607.04579*, 2016.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011.

Jianqing Fan, Qiang Sun, Wen-Xin Zhou, and Ziwei Zhu. Principal component analysis for big data. *arXiv preprint arXiv:1801.01602*, 2018.

Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of human genetics*, 7 (2):179–188, 1936.

Chao Gao, Dan Garber, Nathan Srebro, Jialei Wang, and Weiran Wang. Stochastic canonical correlation analysis. *Journal of Machine Learning Research*, 20(167):1–46, 2019.

Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points – online stochastic gradient for tensor decomposition. In *Proceedings of The 28th Conference on Learning Theory*, pages 797–842, 2015.

Rong Ge, Chi Jin, Sham M Kakade, Praneeth Netrapalli, and Aaron Sidford. Efficient algorithms for large-scale generalized eigenvector computation and canonical correlation analysis. In *Proceedings of the 33th International Conference on Machine Learning*, pages 2741–2750, 2016.

Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

Prateek Jain, Chi Jin, Sham M Kakade, Praneeth Netrapalli, and Aaron Sidford. Matching matrix Bernstein and near-optimal finite sample guarantees for Oja's algorithm. In *Proceedings of The 29th Conference on Learning Theory*, pages 1147–1164, 2016.

Chi Jin, Praneeth Netrapalli, Rong Ge, Sham M Kakade, and Michael I Jordan. On nonconvex optimization for machine learning: Gradients, stochasticity, and saddle points. *arXiv preprint arXiv:1902.04811*, 2019.

D. Kingma and J. Ba. Adam: A method for stochastic optimization. *Proceedings of the 3rd International Conference on Learning Representations*, 2015.

Arun Kumar Kuchibhotla and Abhishek Chakrabortty. Moving beyond sub-gaussianity in high-dimensional statistics: Applications in covariance estimation and linear regression. *arXiv preprint arXiv:1804.02605*, 2018.

Harold Kushner and G George Yin. *Stochastic approximation and recursive algorithms and applications*, volume 35. Springer, 2003.

Chris Junchi Li. A note on concentration inequality for vector-valued martingales with weak exponential-type tails. *arXiv preprint arXiv:1809.02495*, 2018.

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.

Ker-Chau Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991.

Zhuang Ma, Yichao Lu, and Dean Foster. Finding linear structure in large datasets with scalable canonical correlation analysis. In *International Conference on Machine Learning*, pages 169–178, 2015.

Zongming Ma. Sparse principal component analysis and iterative thresholding. *The Annals of Statistics*, 41 (2):772–801, 2013.

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

Erkki Oja. Simplified neuron model as a principal component analyzer. *Journal of mathematical biology*, 15(3):267–273, 1982.

Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

Ning Qian. On the momentum term in gradient descent learning algorithms. *Neural Netw*, 12(1):145–151, 1999.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

Mervyn Stone and Rodney J Brooks. Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 237–269, 1990.

Kean Ming Tan, Zhaoran Wang, Han Liu, Tong Zhang, et al. Sparse generalized eigenvalue problem: Optimal statistical rates via truncated rayleigh flow. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(5):1057–1086, 2018.

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

Max Welling. Fisher linear discriminant analysis. *Department of Computer Science, University of Toronto*, 3:1–4, 2005.

Daniela M Witten, Robert Tibshirani, and Trevor Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–534, 2009.

Ganzhao Yuan, Li Shen, and Wei-Shi Zheng. A decomposition algorithm for sparse generalized eigenvalue problem. *arXiv preprint arXiv:1802.09303*, 2018.

Xiao-Tong Yuan and Tong Zhang. Truncated power method for sparse eigenvalue problems. *Journal of Machine Learning Research*, 14:899–925, 2013.

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

Appendix A. Proof of Proposition 3

This subsection provides a proof for Proposition 3 on the convergence to a local minimizer. Under the initialization condition (14), there exists a local minimizer $\boldsymbol{v}^* \in \mathbf{B}_{\underline{\delta}}(\boldsymbol{v}_0)$ of $F(\boldsymbol{v})$ such that $\boldsymbol{u}^\top \mathcal{H}(\boldsymbol{v}^*)\boldsymbol{u} \geq \mu\|\boldsymbol{u}\|^2$ for all $\boldsymbol{u} \in \mathcal{T}(\boldsymbol{v}^*)$.

For a positive quantity M to be determined later, let

$$\mathcal{T}_M = \inf\left\{t \geq 1 : \|\Gamma(\boldsymbol{v}_{t-1}; \boldsymbol{\zeta}_t)\| > M\right\}. \tag{33}$$

In words, $\mathcal{T}_M$ is the first t such that the norm of the stochastic scaled-gradient $\Gamma(\boldsymbol{v}_{t-1}; \boldsymbol{\zeta}_t)$ exceeds M. We first provide the following lemma.

Lemma 9 *Assume all conditions in Theorem 2. For any positive ϵ, let*

$$M = \mathcal{V}\log^{1/\alpha}\epsilon^{-1}, \tag{34}$$

Then, we have

$$\mathbb{P}(\mathcal{T}_M \leq T^*_\eta) \leq 2T^*_\eta\epsilon.$$

Proof of Lemma 9 is a straightforward corollary of a union bound and Assumption 2, and is provided in §B.1.

Recall the definitions of the manifold gradient $g(\boldsymbol{v})$ and the Hessian $\mathcal{H}(\boldsymbol{v})$ in (10) and (11). Under a unit spherical constraint $c(\boldsymbol{v}) = \|\boldsymbol{v}\|^2 - 1 = 0$, their definitions simplify to

$$g(\boldsymbol{v}) = (\mathbf{I} - \boldsymbol{v}\boldsymbol{v}^\top)\nabla F(\boldsymbol{v}) \quad \text{and} \quad \mathcal{H}(\boldsymbol{v}) = \nabla^2 F(\boldsymbol{v}) - (\boldsymbol{v}^\top \nabla F(\boldsymbol{v}))\mathbf{I}. \tag{35}$$

Taking derivatives, we decompose

$$\nabla g(\boldsymbol{v}) = \mathcal{H}(\boldsymbol{v}) + \mathcal{N}(\boldsymbol{v}), \tag{36}$$

where the additional term $\mathcal{N}(\boldsymbol{v})$ is defined as

$$\mathcal{N}(\boldsymbol{v}) = -\boldsymbol{v}(\nabla F(\boldsymbol{v}) + \nabla^2 F(\boldsymbol{v})\boldsymbol{v})^\top. \tag{37}$$

The following lemma shows that $g(\boldsymbol{v}), \mathcal{H}(\boldsymbol{v}), \mathcal{N}(\boldsymbol{v})$ are Lipschitz continuous.

Lemma 10 *Given Assumption 1, we have that $g(\boldsymbol{v}), \mathcal{H}(\boldsymbol{v}), \mathcal{N}(\boldsymbol{v})$ are L_G, L_H, L_N-Lipschitz and $\|\mathcal{H}(\boldsymbol{v})\| \leq B_H$ within $\{\boldsymbol{v} : \|\boldsymbol{v}\| \leq 1, \|\boldsymbol{v} - \boldsymbol{v}^*\| \leq \delta\}$, where the constants are defined as $L_G \equiv L_K + 2L_F$, $L_H \equiv L_Q + L_F + L_K$, $L_N \equiv L_F + 3L_K + L_Q$, $B_H \equiv L_F + L_K$.*

A proof of Lemma 10 is deferred to Appendix B.2.

For notational simplicity, we denote $\mathcal{H}_* = \mathcal{H}(\boldsymbol{v}^*)$ and $\mathcal{N}_* = \mathcal{N}(\boldsymbol{v}^*)$, and recall that $\mathcal{F}_t$ is the filtration generated by $\boldsymbol{\zeta}_t$. Then we have the following lemma.

Lemma 11 *Under Assumptions 1 and 2, when $\eta \leq 1/(5M)$, on the event $(\|\Gamma(\boldsymbol{v}_{t-1}; \boldsymbol{\zeta}_t)\| \leq M)$, the update rule (9) of $\boldsymbol{v}_t$ can be written as*

$$\boldsymbol{v}_t - \boldsymbol{v}^* = \left(\mathbf{I} - \eta D\mathcal{H}_* - \eta D\mathcal{N}_*\right)(\boldsymbol{v}_{t-1} - \boldsymbol{v}^*) + \eta\boldsymbol{\xi}_t + \eta\boldsymbol{R}_t + \eta^2\boldsymbol{Q}_t; \tag{38}$$

where $\{\boldsymbol{\xi}_t\}$ forms a vector-valued martingale difference sequence with respect to $\mathcal{F}_t$, $\boldsymbol{\xi}_t$ is α-sub-Weibull with parameter $G_\alpha\mathcal{V}$, $\boldsymbol{R}_t$ satisfies $\|\boldsymbol{R}_t\| \leq (DL_H + DL_N + L_D L_G)\|\boldsymbol{v}_{t-1} - \boldsymbol{v}^*\|^2$ and $\boldsymbol{Q}_t$ sastisfies $\|\boldsymbol{Q}_t\| \leq 7M^2$.

The proof of Lemma 11 is deferred to Appendix B.3.

We define the projection of $\boldsymbol{v}_t - \boldsymbol{v}^*$ on $\mathcal{T}(\boldsymbol{v}^*)$ as

$$\Delta_t = (\mathbf{I} - \boldsymbol{v}^* \boldsymbol{v}^{*\top})(\boldsymbol{v}_t - \boldsymbol{v}^*), \tag{39}$$

and the projection of $\mathcal{H}_*$ on $\mathcal{T}(\boldsymbol{v}^*)$ as

$$\mathcal{M}_* = (\mathbf{I} - \boldsymbol{v}^* \boldsymbol{v}^{*\top})\mathcal{H}_*(\mathbf{I} - \boldsymbol{v}^* \boldsymbol{v}^{*\top}). \tag{40}$$

Lemma 12 *Under initialization condition (14), the following properties hold:*

(i) For all $t \geq 0$,

$$\|(\boldsymbol{v}^* \boldsymbol{v}^{*\top})(\boldsymbol{v}_t - \boldsymbol{v}^*)\| = \frac{1}{2}\|\boldsymbol{v}_t - \boldsymbol{v}^*\|^2, \quad \|\Delta_t\|^2 = \|\boldsymbol{v}_t - \boldsymbol{v}^*\|^2 - \frac{1}{4}\|\boldsymbol{v}_t - \boldsymbol{v}^*\|^4.$$

If $\boldsymbol{v}_t^\top \boldsymbol{v}^* \geq 0$,

$$\|\Delta_t\|^2 \leq \|\boldsymbol{v}_t - \boldsymbol{v}^*\|^2 \leq 2\|\Delta_t\|^2. \tag{41}$$

(ii) When $\eta \leq 1/(DB_H)$, for all $\boldsymbol{u} \in \mathcal{T}(\boldsymbol{v}^*)$,

$$\|(\mathbf{I} - \eta D \mathcal{M}_*)^t \Delta_0\| \leq (1 - \eta D\mu)^t \|\Delta_0\|. \tag{42}$$

Remark 13 *To interpret Lemma 12(i), we denote $\theta \equiv \angle(\boldsymbol{v}_t, \boldsymbol{v}^*) \in [0, \pi/2]$, then $\|\boldsymbol{v}_t - \boldsymbol{v}^*\| = 2\sin(\theta/2)$, $\Delta_t = (\mathbf{I} - \boldsymbol{v}^*\boldsymbol{v}^*)^\top(\boldsymbol{v}_t - \boldsymbol{v}^*) = \sin\theta$, and (41) is equivalent to the trigonometric inequality*

$$\sin^2\theta = 4\sin^2(\theta/2)\cos^2(\theta/2) \leq 4\sin^2(\theta/2) = 2(1-\cos\theta) \leq 2(1-\cos\theta)(1+\cos\theta) = 2\sin^2\theta.$$

Proof of Lemma 12 is deferred to §B.4. By combining Lemmas 11 and 12, we have the following Lemma on update rule in terms of Δ_t.

Lemma 14 *Under Assumptions 1, 2 and initialization condition (14), when $\eta \leq 1/(5M)$, on the event $(\|\Gamma(\boldsymbol{v}_{t-1}; \boldsymbol{\zeta}_t)\| \leq M)$, the update (9) can be written in terms of Δ_t as*

$$\Delta_t = (\mathbf{I} - \eta D \mathcal{M}_*)\,\Delta_{t-1} + \eta \boldsymbol{\chi}_t + \eta \boldsymbol{S}_t + \eta^2 \boldsymbol{P}_t; \tag{43}$$

where $\boldsymbol{\chi}_t, \boldsymbol{S}_t, \boldsymbol{P}_t \in \mathcal{T}(\boldsymbol{v}^*)$, $\{\boldsymbol{\chi}_t\}$ forms a vector-valued martingale difference sequence with respect to $\mathcal{F}_t$, $\boldsymbol{\chi}_t$ is α-sub-Weibull with parameter $G_\alpha \mathcal{V}$, $\boldsymbol{S}_t$ satisfies $\|\boldsymbol{S}_t\| \leq \rho\|\boldsymbol{v}_{t-1} - \boldsymbol{v}^*\|^2$ and $\boldsymbol{P}_t$ sastisfies $\|\boldsymbol{P}_t\| \leq 7M^2$.

Here we have $\rho = D(L_H + L_N + B_H/2) + L_D L_G$, which agrees with its definition in (13). Proof of Lemma 14 is deferred to §B.5.

If we analyze the iteration Δ_t itself, its tail behavior cannot be analyzed properly. We let in parallel

$$\widetilde{\boldsymbol{S}}_t = \boldsymbol{S}_t 1_{(\mathcal{T}_M > t)}, \quad \widetilde{\boldsymbol{P}}_t = \boldsymbol{P}_t 1_{(\mathcal{T}_M > t)}, \tag{44}$$

let $\overline{\Delta}_0 = \Delta_0$, and define the coupled process iteratively

$$\overline{\Delta}_t = (\mathbf{I} - \eta D \mathcal{M}_*)\,\overline{\Delta}_{t-1} + \eta \boldsymbol{\chi}_t + \eta \widetilde{\boldsymbol{S}}_t + \eta^2 \widetilde{\boldsymbol{P}}_t. \tag{45}$$

The $\overline{\Delta}_t$ iteration avoids the potential issues of summation over $\boldsymbol{P}_t$. We conclude the following lemma that characterizes the coupling relation $\overline{\Delta}_t = \Delta_t$, which allows us to analyze the coupled iteration $\overline{\Delta}_t$.

Lemma 15 *For each $t \geq 0$ we have $\overline{\Delta}_t = \Delta_t$ on the event $(\mathcal{T}_M > t)$. Furthermore, we have for all $t \geq 1$*

$$\overline{\Delta}_t = (\mathbf{I} - \eta D\mathcal{M}_*)^t \Delta_0 + \eta \sum_{s=1}^{t} (\mathbf{I} - \eta D\mathcal{M}_*)^{t-s} \boldsymbol{\chi}_s + \eta \sum_{s=1}^{t} (\mathbf{I} - \eta D\mathcal{M}_*)^{t-s} \widetilde{\boldsymbol{S}}_s + \eta^2 \sum_{s=1}^{t} (\mathbf{I} - \eta D\mathcal{M}_*)^{t-s} \widetilde{\boldsymbol{P}}_s. \tag{46}$$

We defer the proof of Lemma 15 in Appendix B.6.

We provide a lemma that tightly characterize the approximations in (46) that $\overline{\Delta}_t \approx (\mathbf{I} - \eta D\mathcal{M}_*)^t \Delta_0$.

Lemma 16 *Let $\eta \leq \min\{1/(DB_H), 1/(5M)\}$ and $T \geq 1$. Then with probability at least*

$$1 - \left(12 + 8\left(\frac{3}{\alpha}\right)^{\frac{2}{\alpha}} \log^{-\frac{\alpha+2}{\alpha}} \epsilon^{-1}\right) T\epsilon,$$

the algorithm satisfies for each $t \in [0, T]$, conditioning on $\|\boldsymbol{v}_s - \boldsymbol{v}^*\| \leq r$ for all $s = 0, \ldots, t-1$ for some $r > 0$

$$\left\|\overline{\Delta}_t - (\mathbf{I} - \eta D\mathcal{M}_*)^t \Delta_0\right\| \leq \frac{8G_\alpha \mathcal{V}}{\sqrt{D\mu}} \log^{\frac{\alpha+2}{2\alpha}} \epsilon^{-1} \cdot \eta^{1/2} + \frac{\rho r^2}{D\mu} + \frac{7\mathcal{V}^2}{D\mu} \log^{\frac{2}{\alpha}} \epsilon^{-1} \cdot \eta. \tag{47}$$

Proof of Lemma 16 is provided in Appendix B.7.

In the following lemma we prove that when the initial iterate $\boldsymbol{v}_0$ is sufficiently close to the minimizer $\boldsymbol{v}^*$ and r is appropriately chosen to be dependent on Δ_0 and $\widetilde{\Theta}(\eta^{1/2})$, the conditioning event occurs almost surely on a high-probability event.

Lemma 17 *When initialization*

$$\|\Delta_0\| \leq \left\{\frac{D\mu}{2^5 G_\alpha \rho}, \delta\right\},$$

for any positives η, ϵ satisfying scaling condition (17), with probability at least

$$1 - \left(14 + 8\left(\frac{3}{\alpha}\right)^{\frac{2}{\alpha}} \log^{-\frac{\alpha+2}{\alpha}} \epsilon^{-1}\right) T\epsilon,$$

for all $t \in [0, T]$ we have

$$\|\Delta_t\| \leq 2\max\left\{\|\Delta_0\|, \frac{2^7 G_\alpha \mathcal{V}}{\sqrt{D\mu}} \log^{\frac{\alpha+2}{2\alpha}} \epsilon^{-1} \cdot \eta^{1/2}\right\},$$

and if $T_\eta^* \in [0, T]$, at time T_η^* we have

$$\|\Delta_{T_\eta^*}\| \leq \frac{1}{2}\max\left\{\|\Delta_0\|, \frac{2^7 G_\alpha \mathcal{V}}{\sqrt{D\mu}} \log^{\frac{\alpha+2}{2\alpha}} \epsilon^{-1} \cdot \eta^{1/2}\right\}.$$

Recall definition of Δ_t in (39). Lemma 17, whose proof is given in Appendix B.8, implies that the iteration keeps $\|\Delta_t\| \leq 2\|\Delta_0\|$ unless $\boldsymbol{v}$ is within a noisy neighborhood of the local minimizer $\boldsymbol{v}^*$.

Finally, we are ready to give the proof of Proposition 3.

Proof [Proof of Proposition 3] Proposition 3 is proved by Lemmas 12 and 17 directly. ■

Appendix B. Proofs of Technical Lemmas

B.1. Proof of Lemma 9

Proof [Proof of Lemma 9] Since $M = \mathcal{V} \log^{\frac{1}{\alpha}} \epsilon^{-1}$, we have from Assumption 2 that for each $t \geq 1$,

$$\begin{aligned}\mathbb{P}\left(\|\Gamma(\boldsymbol{v}_{t-1}; \boldsymbol{\zeta}_t)\| > M\right) &= \mathbb{P}\left(\exp\left(\frac{\|\Gamma(\boldsymbol{v}_{t-1}; \boldsymbol{\zeta}_t)\|^\alpha}{\mathcal{V}^\alpha}\right) > \exp\left(\frac{M^\alpha}{\mathcal{V}^\alpha}\right)\right)\\ &\leq \exp\left(-\frac{M^\alpha}{\mathcal{V}^\alpha}\right) \mathbb{E}\exp\left(\frac{\|\Gamma(\boldsymbol{v}_{t-1}; \boldsymbol{\zeta}_t)\|^\alpha}{\mathcal{V}^\alpha}\right) \leq 2\epsilon.\end{aligned}$$

where we apply Markov inequality and Assumption 2 (with law of total expectation applied). Taking union bound,

$$\mathbb{P}(\mathcal{T}_M \leq T_\eta^*) \leq \sum_{t=1}^{T_\eta^*} \mathbb{P}\left(\|\Gamma(\boldsymbol{v}_{t-1}; \boldsymbol{\zeta}_t)\| > M\right) \leq 2T_\eta^* \epsilon.$$

$\blacksquare$

B.2. Proof of Lemma 10

Proof [Proof of Lemma 10] For all $\boldsymbol{u}, \boldsymbol{v} \in \mathcal{S}^{d-1}$, we have

$$\begin{aligned}\|g(\boldsymbol{u}) - g(\boldsymbol{v})\| &\leq \|\mathbf{I} - \boldsymbol{u}\boldsymbol{u}^\top\| \|\nabla F(\boldsymbol{u}) - \nabla F(\boldsymbol{v})\| + \|\boldsymbol{v}\boldsymbol{v}^\top - \boldsymbol{u}\boldsymbol{u}^\top\| \|\nabla F(\boldsymbol{v})\|\\ &\leq 1 \cdot L_K \|\boldsymbol{u} - \boldsymbol{v}\| + 2\|\boldsymbol{u} - \boldsymbol{v}\| \cdot L_F\\ &= (L_K + 2L_F)\|\boldsymbol{u} - \boldsymbol{v}\|,\\ \|\mathcal{H}(\boldsymbol{u}) - \mathcal{H}(\boldsymbol{v})\| &\leq \|\nabla^2 F(\boldsymbol{u}) - \nabla^2 F(\boldsymbol{v})\| + (\|\boldsymbol{u} - \boldsymbol{v}\| \|\nabla F(\boldsymbol{u})\| + \|\boldsymbol{v}\| \|\nabla F(\boldsymbol{u}) - \nabla F(\boldsymbol{v})\|)\|\mathbf{I}\|\\ &\leq L_Q\|\boldsymbol{u} - \boldsymbol{v}\| + (\|\boldsymbol{u} - \boldsymbol{v}\| \cdot L_F + 1 \cdot L_K\|\boldsymbol{u} - \boldsymbol{v}\|) \cdot 1\\ &= (L_Q + L_F + L_K)\|\boldsymbol{u} - \boldsymbol{v}\|,\\ \|\mathcal{N}(\boldsymbol{u}) - \mathcal{N}(\boldsymbol{v})\| &\leq \|\boldsymbol{u} - \boldsymbol{v}\|(\|\nabla F(\boldsymbol{u})\| + \|\nabla^2 F(\boldsymbol{u})\| \|\boldsymbol{u}\|)\\ &\qquad + \|\boldsymbol{v}\|(\|\nabla F(\boldsymbol{u}) - \nabla F(\boldsymbol{v})\| + \|\nabla^2 F(\boldsymbol{u}) - \nabla^2 F(\boldsymbol{v})\| \|\boldsymbol{u}\| + \|\nabla^2 F(\boldsymbol{v})\| \|\boldsymbol{u} - \boldsymbol{v}\|)\\ &\leq \|\boldsymbol{u} - \boldsymbol{v}\|(L_F + L_K \cdot 1) + 1 \cdot (L_K\|\boldsymbol{u} - \boldsymbol{v}\| + L_Q\|\boldsymbol{u} - \boldsymbol{v}\| \cdot 1 + L_K \cdot \|\boldsymbol{u} - \boldsymbol{v}\|)\\ &= (L_F + 3L_K + L_Q)\|\boldsymbol{u} - \boldsymbol{v}\|,\\ \|\mathcal{H}(\boldsymbol{v})\| &\leq \|\nabla^2 F(\boldsymbol{v})\| + \|\boldsymbol{v}\| \|\nabla F(\boldsymbol{v})\| \|\mathbf{I}\| \leq L_K + 1 \cdot L_F \cdot 1 = L_K + L_F.\end{aligned}$$

which implies that $g(\boldsymbol{v})$ is $(L_G \equiv L_K + 2L_F)$-Lipschitz, $\mathcal{H}(\boldsymbol{v})$ is $(L_H \equiv L_Q + L_F + L_K)$-Lipschitz, $\mathcal{N}(\boldsymbol{v})$ is $(L_N \equiv L_F + 3L_K + L_Q)$-Lipschitz and $\|\mathcal{H}(\boldsymbol{v})\| \leq B_H \equiv L_F + L_K$ within $\{\boldsymbol{v} : \|\boldsymbol{v}\| \leq 1, \|\boldsymbol{v} - \boldsymbol{v}^*\| \leq \delta\}$. $\blacksquare$

B.3. Proof of Lemma 11

Proof [Proof of Lemma 11] We have by Taylor series that for any $y \in \mathbb{R}$ satisfying $|y| \leq 1/2$

$$\left|(1-y)^{-1/2} - 1 - \frac{y}{2}\right| \leq \frac{3y^2}{8} \sum_{k=0}^{\infty} |y|^k \leq \frac{3y^2}{4}.$$

When $\eta \le 1/(5M)$, on the event $(\|\Gamma(\boldsymbol{v}_{t-1};\boldsymbol{\zeta}_t)\| \le M)$, by letting $y = 2\eta \boldsymbol{v}_{t-1}^\top \Gamma(\boldsymbol{v}_{t-1};\boldsymbol{\zeta}_t) - \eta^2\|\Gamma(\boldsymbol{v}_{t-1};\boldsymbol{\zeta}_t)\|^2$ we have

$$|y| \le 2\eta \left|\boldsymbol{v}_{t-1}^\top \Gamma(\boldsymbol{v}_{t-1};\boldsymbol{\zeta}_t)\right| + \eta^2\|\Gamma(\boldsymbol{v}_{t-1};\boldsymbol{\zeta}_t)\|^2 \le 2\eta M + \eta^2 M^2 \le (11/5)\eta M < 1/2,$$

and hence combining the above two displays gives

$$\begin{aligned}
&\left|\|\boldsymbol{v}_{t-1} - \eta\Gamma(\boldsymbol{v}_{t-1};\boldsymbol{\zeta}_t)\|^{-1} - 1 - \eta\boldsymbol{v}_{t-1}^\top \Gamma(\boldsymbol{v}_{t-1};\boldsymbol{\zeta}_t)\right| \\
&\le \left|\left(1 - 2\eta \boldsymbol{v}_{t-1}^\top \Gamma(\boldsymbol{v}_{t-1};\boldsymbol{\zeta}_t) + \eta^2\|\Gamma(\boldsymbol{v}_{t-1};\boldsymbol{\zeta}_t)\|^2\right)^{-1/2} - 1 - \eta\boldsymbol{v}_{t-1}^\top \Gamma(\boldsymbol{v}_{t-1};\boldsymbol{\zeta}_t)\right| \\
&\le \left|(1-y)^{-1/2} - 1 - \frac{y}{2}\right| + \frac{\eta^2\|\Gamma(\boldsymbol{v}_{t-1};\boldsymbol{\zeta}_t)\|^2}{2} \\
&\le \frac{3y^2}{4} + \frac{1}{2}\eta^2M^2 \le \frac{3}{4}\cdot\frac{121}{25}\eta^2M^2 + \frac{1}{2}\eta^2M^2 \le 5\eta^2M^2.
\end{aligned} \tag{48}$$

By defining

$$\boldsymbol{\xi}_t = (\mathbf{I} - \boldsymbol{v}_{t-1}\boldsymbol{v}_{t-1}^\top)(\Gamma(\boldsymbol{v}_{t-1};\boldsymbol{\zeta}_t) - D(\boldsymbol{v}_{t-1})\nabla F(\boldsymbol{v}_{t-1})), \tag{49}$$

and

$$\begin{aligned}
\boldsymbol{Q}_t = \eta^{-2}\cdot &\left(\|\boldsymbol{v}_{t-1} - \eta\Gamma(\boldsymbol{v}_{t-1};\boldsymbol{\zeta}_t)\|^{-1} - 1 - \eta\boldsymbol{v}_{t-1}^\top\Gamma(\boldsymbol{v}_{t-1};\boldsymbol{\zeta}_t)\right)(\boldsymbol{v}_{t-1} - \eta\Gamma(\boldsymbol{v}_{t-1};\boldsymbol{\zeta}_t)) \\
&- (\boldsymbol{v}_{t-1}^\top\Gamma(\boldsymbol{v}_{t-1};\boldsymbol{\zeta}_t))\Gamma(\boldsymbol{v}_{t-1};\boldsymbol{\zeta}_t),
\end{aligned} \tag{50}$$

the update formula (9) is equivalent to

$$\boldsymbol{v}_t = \boldsymbol{v}_{t-1} - \eta D(\boldsymbol{v}_{t-1})g(\boldsymbol{v}_{t-1}) + \eta\boldsymbol{\xi}_t + \eta^2\boldsymbol{Q}_t. \tag{51}$$

Using (48), we have

$$\|\boldsymbol{Q}_t\| \le \eta^{-2}\cdot 5\eta^2M^2\cdot(1+\eta M) + M^2 \le 7M^2.$$

Recall that we denote $D = D(\boldsymbol{v}^*), \mathcal{H}_* = \mathcal{H}(\boldsymbol{v}^*), \mathcal{N}_* = \mathcal{N}(\boldsymbol{v}^*)$. By defining

$$\boldsymbol{R}_t = D(\mathcal{H}_* + \mathcal{N}_*)(\boldsymbol{v}_{t-1} - \boldsymbol{v}^*) - D(\boldsymbol{v}_{t-1})g(\boldsymbol{v}_{t-1}), \tag{52}$$

we have

$$\boldsymbol{v}_t = \boldsymbol{v}_{t-1} - \eta D(\mathcal{H}_* + \mathcal{N}_*)(\boldsymbol{v}_{t-1} - \boldsymbol{v}^*) + \eta\boldsymbol{\xi}_t + \eta\boldsymbol{R}_t + \eta^2\boldsymbol{Q}_t.$$

Since $(\mathbf{I} - \boldsymbol{v}_{t-1}\boldsymbol{v}_{t-1}^\top)$ is $\mathcal{F}_{t-1}$-measurable, we know that $\mathbb{E}[\boldsymbol{\xi}_t \mid \mathcal{F}_{t-1}] = 0$ and hence $\{\boldsymbol{\xi}_t\}$ is a vector-valued martingale difference sequence. Additionally, we have $\|\mathbf{I} - \boldsymbol{v}_{t-1}\boldsymbol{v}_{t-1}^\top\| \le 1$, and hence from Assumption 2 and Lemma 22 we know

$$\mathbb{E}\exp\left(\frac{\|\boldsymbol{\xi}_t\|^\alpha}{(G_\alpha\mathcal{V})^\alpha}\right) \le \mathbb{E}\exp\left(\frac{\|\Gamma(\boldsymbol{v}_{t-1};\boldsymbol{\zeta}_t) - D(\boldsymbol{v}_{t-1})\nabla F(\boldsymbol{v}_{t-1})\|^\alpha}{(G_\alpha\mathcal{V})^\alpha}\right) \le 2$$

which implies that $\boldsymbol{\xi}$ is α-sub-Weibull with parameter $G_\alpha\mathcal{V}$.

Finally, we apply mean value theorem using (36) and $g(\boldsymbol{v}^*) = 0$ to obtain

$$\begin{aligned}
\|\boldsymbol{R}_t\| &= \|D(\mathcal{H}_* + \mathcal{N}_*)(\boldsymbol{v}_{t-1} - \boldsymbol{v}^*) - D(\boldsymbol{v}_{t-1})g(\boldsymbol{v}_{t-1})\| \\
&\le D\left\|(\mathcal{H}_* + \mathcal{N}_*)(\boldsymbol{v}_{t-1} - \boldsymbol{v}^*) - \int_0^1 \mathcal{H}(\boldsymbol{v}^* + \theta(\boldsymbol{v}_{t-1} - \boldsymbol{v}^*)) + \mathcal{N}(\boldsymbol{v}^* + \theta(\boldsymbol{v}_{t-1} - \boldsymbol{v}^*))d\theta\,(\boldsymbol{v}_{t-1} - \boldsymbol{v}^*)\right\| \\
&\quad + \|D - D(\boldsymbol{v}_{t-1})\|\|g(\boldsymbol{v}_{t-1})\| \\
&\le D(L_H + L_N)\|\boldsymbol{v}_{t-1} - \boldsymbol{v}^*\|^2 + L_DL_G\|\boldsymbol{v}_{t-1} - \boldsymbol{v}^*\|^2
\end{aligned}$$

where we use the Lipschitz continuity of $D(\boldsymbol{v}), g(\boldsymbol{v}), \mathcal{H}(\boldsymbol{v}), \mathcal{N}(\boldsymbol{v})$. This completes the proof of Lemma 11. ■

B.4. Proof of Lemma 12

Proof [Proof of Lemma 12] Under initialization condition (14), we have the following:

(i) For all unit vector $\boldsymbol{v}$, since $\|\boldsymbol{v}\| = \|\boldsymbol{v}^*\| = 1$ we have

$$\|(\boldsymbol{v}^*\boldsymbol{v}^{*\top})(\boldsymbol{v}-\boldsymbol{v}^*)\| = -\boldsymbol{v}^{*\top}(\boldsymbol{v}-\boldsymbol{v}^*) = \frac{1}{2}\|\boldsymbol{v}\|^2 - \boldsymbol{v}^{*\top}\boldsymbol{v} + \frac{1}{2}\|\boldsymbol{v}^*\|^2 = \frac{1}{2}\|\boldsymbol{v}-\boldsymbol{v}^*\|^2.$$

Because

$$\left((\boldsymbol{v}^*\boldsymbol{v}^{*\top})(\boldsymbol{v}-\boldsymbol{v}^*)\right)^\top \left((\mathbf{I}-\boldsymbol{v}^*\boldsymbol{v}^{*\top})(\boldsymbol{v}-\boldsymbol{v}^*)\right) = 0,$$

by Pythagorean theorem we have

$$\|(\boldsymbol{v}^*\boldsymbol{v}^{*\top})(\boldsymbol{v}-\boldsymbol{v}^*)\|^2 + \|(\mathbf{I}-\boldsymbol{v}^*\boldsymbol{v}^{*\top})(\boldsymbol{v}-\boldsymbol{v}^*)\|^2 = \|\boldsymbol{v}-\boldsymbol{v}^*\|^2$$

Combining the above equalities and plugging in $\boldsymbol{v} = \boldsymbol{v}_t$ give

$$\|\Delta_t\|^2 = \|\boldsymbol{v}_t - \boldsymbol{v}^*\|^2 - \frac{1}{4}\|\boldsymbol{v}_t - \boldsymbol{v}^*\|^4,$$

which admits the following solution given $\boldsymbol{v}_t^\top \boldsymbol{v}^* \geq 0$

$$\|\boldsymbol{v}_t - \boldsymbol{v}^*\|^2 = 2 - \sqrt{4 - 4\|\Delta_t\|^2},$$

and hence

$$\|\Delta_t\|^2 \leq \|\boldsymbol{v}_t - \boldsymbol{v}^*\|^2 = \frac{4\|\Delta_t\|^2}{2 + \sqrt{4 - 4\|\Delta_t\|^2}} \leq 2\|\Delta_t\|^2.$$

(ii) Under initialization condition (14), for all $\boldsymbol{u} \in \mathcal{T}(\boldsymbol{v}^*)$, we have $\boldsymbol{u}^\top \mathcal{H}_* \boldsymbol{u} \geq \mu\|\boldsymbol{u}\|^2$. Hence for $\eta \leq 1/(DB_H)$, we have

$$\|(\mathbf{I} - \eta D\mathcal{M}_*)^{1/2}\boldsymbol{u}\| \leq (1 - \eta D\mu)^{1/2}\|\boldsymbol{u}\|. \tag{53}$$

By noticing that $(\mathbf{I} - \eta D\mathcal{M}_*)^{(t-1)/2}\boldsymbol{u} \in \mathcal{T}(\boldsymbol{v}^*)$, for all $t \geq 1$, we could inductively plug in $(\mathbf{I} - \eta D\mathcal{M}_*)^{(t-1)/2}\boldsymbol{u}$ to $\boldsymbol{u}$ in (53) and obtain for each $t \geq 0$

$$\|(\mathbf{I} - \eta D\mathcal{M}_*)^t\boldsymbol{u}\| \leq (1 - \eta D\mu)^t\|\boldsymbol{u}\|.$$

■

B.5. Proof of Lemma 14

Proof [Proof of Lemma 14] By left multiplying (38) in Lemma 11 by $(\mathbf{I} - \boldsymbol{v}^*\boldsymbol{v}^{*\top})$ and noticing $(\mathbf{I} - \boldsymbol{v}^*\boldsymbol{v}^{*\top})\mathcal{N}_* = 0$, we obtain

$$\begin{aligned}\Delta_t = \Delta_{t-1} &- \eta D(\mathbf{I} - \boldsymbol{v}^*\boldsymbol{v}^{*\top})\mathcal{H}_*(\boldsymbol{v}_{t-1} - \boldsymbol{v}^*) + \eta(\mathbf{I} - \boldsymbol{v}^*\boldsymbol{v}^{*\top})\boldsymbol{\xi}_t \\ &+ \eta(\mathbf{I} - \boldsymbol{v}^*\boldsymbol{v}^{*\top})\boldsymbol{R}_t + \eta^2(\mathbf{I} - \boldsymbol{v}^*\boldsymbol{v}^{*\top})\boldsymbol{Q}_t.\end{aligned}$$

We have decomposition

$$(\mathbf{I} - \boldsymbol{v}^*\boldsymbol{v}^{*\top})\mathcal{H}_*(\boldsymbol{v}_{t-1} - \boldsymbol{v}^*) = (\mathbf{I} - \boldsymbol{v}^*\boldsymbol{v}^{*\top})\mathcal{H}_*\Delta_t + (\mathbf{I} - \boldsymbol{v}^*\boldsymbol{v}^{*\top})\mathcal{H}_* \cdot (\boldsymbol{v}^*\boldsymbol{v}^{*\top})(\boldsymbol{v}_{t-1} - \boldsymbol{v}^*),$$

where $(\mathbf{I} - \boldsymbol{v}^*\boldsymbol{v}^{*\top})\mathcal{H}_*\Delta_t = \mathcal{M}_*\Delta_t$, and based on Lemma 12 and $\|\mathcal{H}_*\| \le B_H$,

$$\|(\mathbf{I} - \boldsymbol{v}^*\boldsymbol{v}^{*\top})\mathcal{H}_* \cdot (\boldsymbol{v}^*\boldsymbol{v}^{*\top})(\boldsymbol{v}_{t-1} - \boldsymbol{v}^*)\| \le \frac{B_H}{2}\|\boldsymbol{v}_{t-1} - \boldsymbol{v}^*\|^2.$$

We set

$$\begin{aligned}\boldsymbol{\chi}_t &= (\mathbf{I} - \boldsymbol{v}^*\boldsymbol{v}^{*\top})\boldsymbol{\xi}_t, \\ \boldsymbol{S}_t &= (\mathbf{I} - \boldsymbol{v}^*\boldsymbol{v}^{*\top})\boldsymbol{R}_t - D \cdot (\mathbf{I} - \boldsymbol{v}^*\boldsymbol{v}^{*\top})\mathcal{H}_* \cdot (\boldsymbol{v}^*\boldsymbol{v}^{*\top})(\boldsymbol{v}_{t-1} - \boldsymbol{v}^*), \\ \boldsymbol{P}_t &= (\mathbf{I} - \boldsymbol{v}^*\boldsymbol{v}^{*\top})\boldsymbol{Q}_t.\end{aligned}$$

Then by combining all of the results above, we have

$$\Delta_t = (\mathbf{I} - \eta D\mathcal{M}_*)\,\Delta_{t-1} + \eta\boldsymbol{\chi}_t + \eta\boldsymbol{S}_t + \eta^2\boldsymbol{P}_t,$$

which proves (43). The rest of Lemma 14 could be easily verified in steps similar to proof of Lemma 11. ■

B.6. Proof of Lemma 15

Proof [Proof of Lemma 15] For $t = 0$ the lemma holds from definition. In general if it holds for $t - 1$ then from the definitions in (44) we have on $(t < \mathcal{T}_M)$ that $\widetilde{\boldsymbol{S}}_s = \boldsymbol{S}_s, \widetilde{\boldsymbol{P}}_s = \boldsymbol{P}_s$ for all $s \le t$, so the conclusion holds for t. Iteratively applying (45) we obtain (46), which concludes our lemma. ■

B.7. Proof of Lemma 16

Proof [Proof of Lemma 16] For any fixed $t \ge 0$, we have the following:

(i) For the first term on the right hand of (46), since $\boldsymbol{\chi}_s \in \mathcal{T}(\boldsymbol{v}^*)$, (42) in Lemma 12 implies $\|(\mathbf{I} - \eta D\mathcal{M}_*)^{t-s}\boldsymbol{\chi}_s\| \le (1-\eta D\mu)^{t-s}\|\boldsymbol{\chi}_s\|$. Hence we have $\|(\mathbf{I}-\eta D\mathcal{M}_*)^{t-s}\boldsymbol{\chi}_s\|_{\psi_\alpha} \le (1-\eta D\mu)^{t-s}\|\boldsymbol{\chi}_s\|_{\psi_\alpha} \le (1 - \eta D\mu)^{t-s}G_\alpha\mathcal{V}$ and

$$\sum_{s=1}^{t}\left\|\eta(\mathbf{I} - \eta D\mathcal{M}_*)^{t-s}\boldsymbol{\chi}_s\right\|_{\psi_\alpha}^2 \le \eta^2\sum_{s=1}^{t}(1 - \eta D\mu)^{2(t-s)}G_\alpha^2\mathcal{V}^2 \le \frac{G_\alpha^2\mathcal{V}^2}{D\mu} \cdot \eta$$

Theorem 2 in Li (2018) provides a concentration inequality for α-sub-Weibull random vectors, which gives

$$\mathbb{P}\left(\left\|\eta\sum_{s=1}^{t}(\mathbf{I} - \eta D\mathcal{M}_*)^{t-s}\boldsymbol{\chi}_s\right\| \ge \frac{8G_\alpha\mathcal{V}}{\sqrt{D\mu}}\log^{\frac{\alpha+2}{2\alpha}}\epsilon^{-1} \cdot \eta^{1/2}\right) \le \left(12 + 8\left(\frac{3}{\alpha}\right)^{\frac{2}{\alpha}}\log^{-\frac{\alpha+2}{\alpha}}\epsilon^{-1}\right)\epsilon.$$

(ii) For the second term on the right hand side of (46), by applying (42) in Lemma 12 and using Lemma 14, given $\|\boldsymbol{v}_{s-1} - \boldsymbol{v}^*\| \le r$ for all $s = 1, \dots, t$ we have,

$$\left\| \eta \sum_{s=1}^{t} \left(\mathbf{I} - \eta D \mathcal{M}_*\right)^{t-s} \widetilde{\boldsymbol{S}}_s \right\| \le \eta \sum_{s=1}^{t} (1 - \eta D\mu)^{t-s} \cdot \rho r^2 \le \frac{\rho r^2}{D\mu}. \tag{54}$$

(iii) For the third term on the right hand side of (46), from Lemma 14 we know $\|\widetilde{\boldsymbol{P}}_t\| \le 7M^2$ and

$$\left\| \eta^2 \sum_{s=1}^{t} \left(\mathbf{I} - \eta D \mathcal{M}_*\right)^{t-s} \widetilde{\boldsymbol{P}}_s \right\| \le \eta^2 \sum_{s=1}^{t} (1 - \eta D\mu)^{t-s} \cdot 7M^2 = \frac{7\mathcal{V}^2}{D\mu} \log^{\frac{2}{\alpha}} \epsilon^{-1} \cdot \eta,$$

where we use the definition of M in (34).

The lemma is concluded by combining the above three items and taking union bound on probability. ■

B.8. Proof of Lemma 17

Proof [Proof of Lemma 17] From the given assumptions, under scaling condition (17), we have

$$r = 2 \max \left\{ \|\Delta_0\|, \ \frac{2^7 G_\alpha \mathcal{V}}{\sqrt{D\mu}} \log^{\frac{\alpha+2}{2\alpha}} \epsilon^{-1} \cdot \eta^{1/2} \right\} \le \frac{D\mu}{16\rho}$$

We let event $\mathfrak{J}$ be (47) holding for each $t \in [0, T]$, i.e.

$$\left\| \overline{\Delta}_t - (\mathbf{I} - \eta D \mathcal{M}_*)^t \Delta_0 \right\| \le \frac{8 G_\alpha \mathcal{V}}{\sqrt{D\mu}} \log^{\frac{\alpha+2}{2\alpha}} \epsilon^{-1} \cdot \eta^{1/2} + \frac{\rho r^2}{D\mu} + \frac{7\mathcal{V}^2}{D\mu} \log^{\frac{2}{\alpha}} \epsilon^{-1} \cdot \eta.$$

Then on event $\mathfrak{J}$, under scaling condition (17), because $\|\Delta_0\| \le \frac{r}{2}$, for each $t \in [0, T]$ we have

$$\|\overline{\Delta}_t\| \le \|\Delta_0\| + \frac{16 G_\alpha \mathcal{V}}{\sqrt{D\mu}} \log^{\frac{\alpha+2}{2\alpha}} \epsilon^{-1} \cdot \eta^{1/2} + \frac{\rho r^2}{D\mu} \le \frac{r}{2} + \frac{r}{16} + \frac{r}{16} \le r.$$

Applying Lemma 16 and taking union bound gives

$$\mathbb{P}(\mathfrak{J}) \ge 1 - \left(12 + 8 \left(\frac{3}{\alpha}\right)^{\frac{2}{\alpha}} \log^{-\frac{\alpha+2}{\alpha}} \epsilon^{-1} \right) T\epsilon$$

Furthermore, using (42) in Lemma 12 and definition of T_η^* in (16), if $T_\eta^* \in [0, T]$, on event $\mathfrak{J}$ we have at time T_η^*

$$\|\overline{\Delta}_{T_\eta^*}\| \le \|(\mathbf{I} - \eta D \mathcal{M}_*)^{T_\eta^*} \Delta_0\| + \frac{16 G_\alpha \mathcal{V}}{\sqrt{D\mu}} \log^{\frac{\alpha+2}{2\alpha}} \epsilon^{-1} \cdot \eta^{1/2} + \frac{\rho r^2}{D\mu} \le \frac{r}{8} + \frac{r}{16} + \frac{r}{16} \le \frac{r}{4}$$

In Lemma 15 we have shown that, on the event $(T < \mathcal{T}_M)$, we have $\overline{\Delta}_t = \Delta_t$. In Lemma 9, we have proved $\mathbb{P}(T < \mathcal{T}_M) \ge 1 - 2T\epsilon$. Along with Lemma 16, we take intersection and obtain

$$\mathbb{P}(\mathfrak{J} \cap (T < \mathcal{T}_M)) \ge 1 - \mathbb{P}(\mathfrak{J}^c) - \mathbb{P}(T \ge \mathcal{T}_M) \ge 1 - \left(14 + 8 \left(\frac{3}{\alpha}\right)^{\frac{2}{\alpha}} \log^{-\frac{\alpha+2}{\alpha}} \epsilon^{-1} \right) T\epsilon$$

At this point we have proved every argument in Lemma 17. ■

B.9. Proof of Lemma 6

Proof [Proof of Lemma 6] Telescoping (43) in Lemma 14 for $t = K_{\eta,\epsilon}T_\eta^* + 2, \ldots, T+1$ gives

$$\eta D\mathcal{M}_* \sum_{t=K_{\eta,\epsilon}T_\eta^*+1}^{T} \Delta_t = (\Delta_{K_{\eta,\epsilon}T_\eta^*+1} - \Delta_{T+1}) + \eta \sum_{t=K_{\eta,\epsilon}T_\eta^*+1}^{T} \chi_{t+1} + \eta \sum_{t=K_{\eta,\epsilon}T_\eta^*+1}^{T} \boldsymbol{S}_{t+1} + \eta^2 \sum_{t=K_{\eta,\epsilon}T_\eta^*+1}^{T} \boldsymbol{P}_{t+1}.$$

Plugging in definitions of $\Delta_t, \overline{\boldsymbol{v}}_T^{(\eta)}$ in (39), (23) gives (27). For event $\mathcal{H}_2$ defined in Theorem 2 using total sample size $T+1$, the proof of Lemma 17 in §B.8 shows that $\mathcal{H}_2 \subseteq \big\{\|\Gamma(\boldsymbol{v}_{t-1}; \boldsymbol{\zeta}_t)\| \le M : 1 \le t \le T+2\big\}$. The rest of Lemma 6 directly follows Lemma 14. ■

Appendix C. Proofs in Application to GEV Problem

C.1. Proof of Lemma 7

Proof [Proof of Lemma 7] In GEV problem setting, the gradient and the Hessian of the objective function $F(\boldsymbol{v})$ are

$$\nabla F(\boldsymbol{v}) = -2\frac{(\boldsymbol{v}^\top \mathbf{B}\boldsymbol{v})\mathbf{A}\boldsymbol{v} - (\boldsymbol{v}^\top \mathbf{A}\boldsymbol{v})\mathbf{B}\boldsymbol{v}}{(\boldsymbol{v}^\top \mathbf{B}\boldsymbol{v})^2},$$

$$\nabla^2 F(\boldsymbol{v}) = -2\frac{(\boldsymbol{v}^\top \mathbf{B}\boldsymbol{v})\mathbf{A} - (\boldsymbol{v}^\top \mathbf{A}\boldsymbol{v})\mathbf{B} + 2(\mathbf{A}\boldsymbol{v}\boldsymbol{v}^\top\mathbf{B} - \mathbf{B}\boldsymbol{v}\boldsymbol{v}^\top\mathbf{A})}{(\boldsymbol{v}^\top \mathbf{B}\boldsymbol{v})^2} + 8\frac{\left[(\boldsymbol{v}^\top \mathbf{B}\boldsymbol{v})\mathbf{A} - (\boldsymbol{v}^\top \mathbf{A}\boldsymbol{v})\mathbf{B}\right]\boldsymbol{v}\boldsymbol{v}^\top\mathbf{B}}{(\boldsymbol{v}^\top \mathbf{B}\boldsymbol{v})^3}.$$

We first notice that, for $\boldsymbol{v} \in \{\boldsymbol{v} : \|\boldsymbol{v}\| \le 1, \|\boldsymbol{v} - \boldsymbol{v}^*\| \le \delta\}$,

$$\|\nabla D(\boldsymbol{v})\| = \left\|2(\boldsymbol{v}^\top \mathbf{B}\boldsymbol{v})\mathbf{B}\boldsymbol{v}\right\| \le 2\|\mathbf{B}\|^2,$$

which indicates that $D(\boldsymbol{v})$ has Lipschitz constant $L_D \equiv 2\|\mathbf{B}\|^2$.

Secondly, we introduce an arbitrary unit vector $\boldsymbol{w}$ and take derivative of vector $\nabla^2 F(\boldsymbol{v})\boldsymbol{w}$ w.r.t. $\boldsymbol{v}$ as

$$\begin{aligned}\nabla_{\boldsymbol{v}}\left[\nabla^2 F(\boldsymbol{v})\boldsymbol{w}\right] = -2&\frac{2\mathbf{A}\boldsymbol{w}\boldsymbol{v}^\top\mathbf{B} - 2\mathbf{B}\boldsymbol{v}\boldsymbol{v}^\top\mathbf{A} + 2(\boldsymbol{v}^\top\mathbf{B}\boldsymbol{w})\mathbf{A} + 2\mathbf{A}\boldsymbol{v}\boldsymbol{w}^\top\mathbf{B} - 2(\boldsymbol{v}^\top\mathbf{A}\boldsymbol{w})\mathbf{B} - 2\mathbf{B}\boldsymbol{v}\boldsymbol{w}^\top\mathbf{A}}{(\boldsymbol{v}^\top \mathbf{B}\boldsymbol{v})^2} \\ &+ 8\frac{\left[(\boldsymbol{v}^\top \mathbf{B}\boldsymbol{v})\mathbf{A} - (\boldsymbol{v}^\top \mathbf{A}\boldsymbol{v})\mathbf{B} + 2(\mathbf{A}\boldsymbol{v}\boldsymbol{v}^\top\mathbf{B} - \mathbf{B}\boldsymbol{v}\boldsymbol{v}^\top\mathbf{A})\right]\boldsymbol{w}\boldsymbol{v}^\top\mathbf{B}}{(\boldsymbol{v}^\top \mathbf{B}\boldsymbol{v})^3} \\ &+ 8\frac{\left[(\boldsymbol{v}^\top \mathbf{B}\boldsymbol{v})\mathbf{A} - (\boldsymbol{v}^\top \mathbf{A}\boldsymbol{v})\mathbf{B}\right]\boldsymbol{v}\boldsymbol{w}^\top\mathbf{B}}{(\boldsymbol{v}^\top \mathbf{B}\boldsymbol{v})^3} \\ &+ 8\frac{(\boldsymbol{v}^\top \mathbf{B}\boldsymbol{w})\left[(\boldsymbol{v}^\top \mathbf{B}\boldsymbol{v})\mathbf{A} - (\boldsymbol{v}^\top \mathbf{A}\boldsymbol{v})\mathbf{B} + 2(\mathbf{A}\boldsymbol{v}\boldsymbol{v}^\top\mathbf{B} - \mathbf{B}\boldsymbol{v}\boldsymbol{v}^\top\mathbf{A})\right]}{(\boldsymbol{v}^\top \mathbf{B}\boldsymbol{v})^3} \\ &- 48\left[\frac{(\boldsymbol{v}^\top \mathbf{B}\boldsymbol{w})\left[(\boldsymbol{v}^\top \mathbf{B}\boldsymbol{v})\mathbf{A} - (\boldsymbol{v}^\top \mathbf{A}\boldsymbol{v})\mathbf{B}\right]\boldsymbol{v}\boldsymbol{v}^\top\mathbf{B}}{(\boldsymbol{v}^\top \mathbf{B}\boldsymbol{v})^4}\right].\end{aligned}$$

The five terms on the right hand side have norm bounded by $\frac{24\|\mathbf{A}\|\|\mathbf{B}\|}{(1-\delta)^2\lambda_{\min}^2(\mathbf{B})}$, $\frac{48\|\mathbf{A}\|\|\mathbf{B}\|^2}{(1-\delta)^3\lambda_{\min}^3(\mathbf{B})}$, $\frac{16\|\mathbf{A}\|\|\mathbf{B}\|^2}{(1-\delta)^3\lambda_{\min}^3(\mathbf{B})}$, $\frac{48\|\mathbf{A}\|\|\mathbf{B}\|^2}{(1-\delta)^3\lambda_{\min}^3(\mathbf{B})}$, $\frac{96\|\mathbf{A}\|\|\mathbf{B}\|^3}{(1-\delta)^4\lambda_{\min}^4(\mathbf{B})}$ respectively, which implies that

$$\left\|\nabla_{\boldsymbol{v}}\left[\nabla^2 F(\boldsymbol{v})\boldsymbol{w}\right]\right\| \le \frac{232\|\mathbf{A}\|\|\mathbf{B}\|^3}{\lambda_{\min}^4(\mathbf{B})}.$$

Therefore, for all $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \{\boldsymbol{v} : \|\boldsymbol{v}\| \leq 1, \|\boldsymbol{v} - \boldsymbol{v}^*\| \leq \delta\}$, we have

$$\left\|\nabla^2 F(\boldsymbol{v}_1) - \nabla^2 F(\boldsymbol{v}_2)\right\| = \max_{\|\boldsymbol{w}\|=1} \left\|\nabla^2 F(\boldsymbol{v}_1)\boldsymbol{w} - \nabla^2 F(\boldsymbol{v}_2)\boldsymbol{w}\right\| \leq \frac{232\|\mathbf{A}\|\|\mathbf{B}\|^3}{(1-\delta)^4 \lambda_{\min}^4(\mathbf{B})} \|\boldsymbol{v}_1 - \boldsymbol{v}_2\|,$$

indicating $\nabla^2 F(\boldsymbol{v})$ has Lipschitz constant $L_Q \equiv \frac{232\|\mathbf{A}\|\|\mathbf{B}\|^3}{(1-\delta)^4 \lambda_{\min}^4(\mathbf{B})}$.

Similarly, we also notice for all $\boldsymbol{v} \in \{\boldsymbol{v} : \|\boldsymbol{v}\| \leq 1, \|\boldsymbol{v} - \boldsymbol{v}^*\| \leq \delta\}$,

$$\|\nabla F(\boldsymbol{v})\| \leq \frac{4\|\mathbf{A}\|\|\mathbf{B}\|}{(1-\delta)^2 \lambda_{\min}^2(\mathbf{B})}, \quad \left\|\nabla^2 F(\boldsymbol{v})\right\| \leq \frac{28\|\mathbf{A}\|\|\mathbf{B}\|^2}{(1-\delta)^3 \lambda_{\min}^3(\mathbf{B})},$$

which indicates that $F(\boldsymbol{v})$ has Lipschitz constant $L_F \equiv \frac{4\|\mathbf{A}\|\|\mathbf{B}\|}{(1-\delta)^2 \lambda_{\min}^2(\mathbf{B})}$ and $\nabla F(\boldsymbol{v})$ has Lipschitz constant $L_K \equiv \frac{28\|\mathbf{A}\|\|\mathbf{B}\|^2}{(1-\delta)^3 \lambda_{\min}^3(\mathbf{B})}$. ■

C.2. Proof of Lemma 8

Proof [Proof of Lemma 8] For notational simplicity, we denote vector $\boldsymbol{v} \in \mathbb{R}^{d_x + d_y}$ as $\boldsymbol{v}^\top = (\boldsymbol{v}_x^\top, \boldsymbol{v}_y^\top)$ for $\boldsymbol{v}_x \in \mathbb{R}^{d_x}, \boldsymbol{v}_y \in \mathbb{R}^{d_y}$. For any vectors $\boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathbb{R}^{d_x}$ with $\|\boldsymbol{w}_1\| \leq 1, \|\boldsymbol{w}_2\| \leq 1$, using Lemma 21 we have

$$\left\|\boldsymbol{w}_1^\top \mathbf{X}\mathbf{X}^\top \boldsymbol{w}_2\right\|_{\psi_1} \leq \left\|\boldsymbol{w}_1^\top \mathbf{X}\right\|_{\psi_2} \left\|\boldsymbol{w}_2^\top \mathbf{X}\right\|_{\psi_2} \leq \mathcal{V}_x^2,$$

which indicates that

$$\left\|\boldsymbol{v}_x^\top \mathbf{X}\mathbf{X}^\top \boldsymbol{v}_x\right\|_{\psi_1} \leq \mathcal{V}_x^2, \quad \left\|\mathbf{X}\mathbf{X}^\top \boldsymbol{v}_x\right\|_{\psi_1} = \max_{\boldsymbol{w} \in \mathbb{R}^{d_x}, \|\boldsymbol{w}\|=1} \left\|\boldsymbol{w}^\top \mathbf{X}\mathbf{X}^\top \boldsymbol{v}_x\right\|_{\psi_1} \leq \mathcal{V}_x^2.$$

Similarly, we can show

$$\left\|\boldsymbol{v}_y^\top \mathbf{Y}\mathbf{Y}^\top \boldsymbol{v}_y\right\|_{\psi_1} \leq \mathcal{V}_y^2, \quad \left\|\mathbf{Y}\mathbf{Y}^\top \boldsymbol{v}_y\right\|_{\psi_1} \leq \mathcal{V}_y^2,$$

and

$$\left\|\boldsymbol{v}_x^\top \mathbf{X}\mathbf{Y}^\top \boldsymbol{v}_y\right\|_{\psi_1} \leq \mathcal{V}_x \mathcal{V}_y, \quad \left\|\mathbf{X}\mathbf{Y}^\top \boldsymbol{v}_y\right\|_{\psi_1} \leq \mathcal{V}_x \mathcal{V}_y, \quad \left\|\mathbf{Y}\mathbf{X}^\top \boldsymbol{v}_x\right\|_{\psi_1} \leq \mathcal{V}_x \mathcal{V}_y.$$

Combining all above inequalities and using Corollary 20, we find

$$\left\|\boldsymbol{v}^\top \widetilde{\mathbf{A}} \boldsymbol{v}\right\|_{\psi_1} \leq 2\mathcal{V}_x \mathcal{V}_y, \quad \left\|\widetilde{\mathbf{A}} \boldsymbol{v}\right\|_{\psi_1} \leq 2\mathcal{V}_x \mathcal{V}_y, \quad \left\|\boldsymbol{v}^\top \widetilde{\mathbf{B}}' \boldsymbol{v}\right\|_{\psi_1} \leq \mathcal{V}_x^2 + \mathcal{V}_y^2, \quad \left\|\widetilde{\mathbf{B}}' \boldsymbol{v}\right\|_{\psi_1} \leq \mathcal{V}_x^2 + \mathcal{V}_y^2.$$

By applying Lemmas 21 and 22, in CCA problem we have stochastic scaled-gradient satisfying

$$\begin{aligned}\left\|(\boldsymbol{v}^\top \widetilde{\mathbf{B}}' \boldsymbol{v})\widetilde{\mathbf{A}} \boldsymbol{v} - (\boldsymbol{v}^\top \widetilde{\mathbf{A}} \boldsymbol{v})\widetilde{\mathbf{B}}' \boldsymbol{v}\right\|_{\psi_{1/2}} &\leq G_{1/2} \left(\left\|\boldsymbol{v}^\top \widetilde{\mathbf{B}}' \boldsymbol{v}\right\|_{\psi_1} \left\|\widetilde{\mathbf{A}} \boldsymbol{v}\right\|_{\psi_1} + \left\|\boldsymbol{v}^\top \widetilde{\mathbf{A}} \boldsymbol{v}\right\|_{\psi_1} \left\|\widetilde{\mathbf{B}}' \boldsymbol{v}\right\|_{\psi_1}\right) \\ &\leq 400(\mathcal{V}_x^2 + \mathcal{V}_y^2)\mathcal{V}_x \mathcal{V}_y.\end{aligned}$$

Hence Assumption 2 holds for $\mathcal{V} = 400(\mathcal{V}_x^2 + \mathcal{V}_y^2)\mathcal{V}_x \mathcal{V}_y$ and $\alpha = 1/2$. ■

Appendix D. Preliminary Lemmas

Definition 18 (Orlicz ψ_α-norm) *For a continuous, monotonically increasing and convex function $\psi(x)$ defined for all $x > 0$ satisfying $\psi(0) = 0$ and $\lim_{x\to\infty} \psi(x) = \infty$, we define the Orlicz ψ-norm for a random variable X as*

$$\|X\|_\psi \equiv \inf\left\{K > 0 : \mathbb{E}\psi\left(\left|\frac{X}{K}\right|\right) \le 1\right\}.$$

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

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

Lemma 19 *When $\psi(x)$ is monotonically increasing and convex for $x > 0$, for any random variables X, Y with finite Orlicz ψ-norm, the triangle inequality holds*

$$\|X + Y\|_\psi \le \|X\|_\psi + \|Y\|_\psi.$$

Corollary 20 *For all $\alpha \ge 1$, Orlicz ψ_α-norm satisfies triangle inequality.*

Lemma 21 *Let X and Y be random variables with finite ψ_α-norm for some $\alpha \ge 1$, then*

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

Lemma 22 *For any random variables X, Y with finite Orlicz ψ_α-norm, the following inequalities hold*

$$\|X + Y\|_{\psi_\alpha} \le \log_2^{1/\alpha}(1 + e^{1/\alpha})(\|X\|_{\psi_\alpha} + \|Y\|_{\psi_\alpha}), \quad \|\mathbb{E}X\|_{\psi_\alpha} \le \log_2^{1/\alpha}(1 + e^{1/\alpha})\|X\|_{\psi_\alpha},$$

and

$$\|X - \mathbb{E}X\|_{\psi_\alpha} \le \log_2^{1/\alpha}(1 + e^{1/\alpha})\left(1 + \log_2^{1/\alpha}(1 + e^{1/\alpha})\right)\|X\|_{\psi_\alpha}.$$

Definition 23 *For a random vector $\mathbf{X} \in \mathbb{R}^d$, its Orlicz ψ_α-norm is defined as*

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

A random variable $\mathbf{X}$ is sub-Gaussian if $\|\mathbf{X}\|_{\psi_2} < \infty$, and is sub-Exponential if $\|\mathbf{X}\|_{\psi_1} < \infty$ (Wainwright, 2019).

Remark 24 *We notice that $\|\mathbf{X}\|_{\psi_\alpha}$ equals to the Orlicz ψ_α-norm of $\|\mathbf{X}\|$. Using this relation, we can easily extend all above results of random variables to random vectors with the same positive factors and dependency on α.*

Proof [Proof of Lemma 19] Let K_1, K_2 denote the Orlicz ψ-norms of X and Y. Because $\psi(x)$ is monotonically increasing and convex, we have

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

which implies

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

■

Proof [Proof of Lemma 21] Denote $A \equiv X/\|X\|_{\psi_\alpha}$, $B \equiv Y/\|Y\|_{\psi_\alpha}$, then $\|A\|_{\psi_\alpha} = \|B\|_{\psi_\alpha} = 1$. Using the elementary inequality

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

and triangle inequality in Lemma 19 we have that

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

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

Proof [Proof of Lemma 22] Recall that when $\alpha \in (0,1)$, $\psi_\alpha(x)$ does *not* satisfy convexity when x is around 0. Let $\widetilde{\psi}_\alpha(x)$ be

$$\widetilde{\psi}_\alpha(x) = \begin{cases} \exp(x^\alpha)-1 & x \geq x_* \\ \frac{x}{x_*}\left(\exp(x_*^\alpha)-1\right) & x \in [0, x_*) \end{cases}.$$

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

$$\psi'_\alpha(x_*) = \alpha x_*^{\alpha-1}\exp(x_*^\alpha) = \frac{\exp(x_*^\alpha)-1}{x_*} = \widetilde{\psi}'_\alpha(x_*).$$

Simplifying it gives us a transcendental equation

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

We easily find that $x_*^\alpha \leq 1/\alpha$. Because $\psi_\alpha(x)$ is concave on $\left(0, (\frac{1}{\alpha}-1)^{1/\alpha}\right)$ and convex on $((\frac{1}{\alpha}-1)^{1/\alpha}, \infty)$, we have $\psi_\alpha(x) \geq \widetilde{\psi}_\alpha(x) \geq 0$ for all $x \geq 0$, and hence

$$0 \leq \psi_\alpha(x) - \widetilde{\psi}_\alpha(x) \leq \psi_\alpha(x_*) \leq e^{1/\alpha}-1. \tag{55}$$

Let K_1, K_2 denote the Orlicz ψ_α-norms of X and Y, then

$$\mathbb{E}\widetilde{\psi}_\alpha\left(\frac{|X|}{K_1}\right) \leq \mathbb{E}\psi_\alpha\left(\frac{|X|}{K_1}\right) \leq 1, \quad \mathbb{E}\widetilde{\psi}_\alpha\left(\frac{|Y|}{K_2}\right) \leq \mathbb{E}\psi_\alpha\left(\frac{|Y|}{K_2}\right) \leq 1.$$

By applying triangle inequality in Corollary 20 and using (55), we have

$$\mathbb{E}\psi_\alpha\left(\frac{|X+Y|}{K_1+K_2}\right) \leq \mathbb{E}\widetilde{\psi}_\alpha\left(\frac{|X+Y|}{K_1+K_2}\right) + e^{1/\alpha} - 1 \leq e^{1/\alpha},$$
$$\mathbb{E}\psi_\alpha\left(\frac{|\mathbb{E}X|}{K_1}\right) \leq \mathbb{E}\widetilde{\psi}_\alpha\left(\frac{|\mathbb{E}X|}{K_1}\right) + e^{1/\alpha} - 1 \leq e^{1/\alpha}.$$

By applying Jensen's inequality to concave function $J_\alpha(z) = z^{\log_{1+e^{1/\alpha}} 2}$, we have

$$\begin{aligned}\mathbb{E}\psi_\alpha\left(\frac{|X+Y|}{\log_2^{1/\alpha}(1+e^{1/\alpha})(K_1+K_2)}\right) &= \mathbb{E}J_\alpha\left(\exp\left(\frac{|X+Y|^\alpha}{(K_1+K_2)^\alpha}\right)\right) - 1 \\ &\le J_\alpha\left(\mathbb{E}\exp\left(\frac{|X+Y|^\alpha}{(K_1+K_2)^\alpha}\right)\right) - 1 \le 1,\end{aligned}$$

and

$$\mathbb{E}\psi_\alpha\left(\frac{|\mathbb{E}X|}{\log_2^{1/\alpha}(1+e^{1/\alpha})K_1}\right) = \mathbb{E}J_\alpha\left(\exp\left(\frac{|\mathbb{E}X|^\alpha}{K_1^\alpha}\right)\right) - 1 \le J_\alpha\left(\mathbb{E}\exp\left(\frac{|\mathbb{E}X|^\alpha}{K_1^\alpha}\right)\right) - 1 \le 1,$$

which implies

$$\|X+Y\|_{\psi_\alpha} \le \log_2^{1/\alpha}(1+e^{1/\alpha})(\|X\|_{\psi_\alpha} + \|Y\|_{\psi_\alpha}), \quad \|\mathbb{E}X\|_{\psi_\alpha} \le \log_2^{1/\alpha}(1+e^{1/\alpha})\|X\|_{\psi_\alpha},$$

and

$$\|X-\mathbb{E}X\|_{\psi_\alpha} \le \log_2^{1/\alpha}(1+e^{1/\alpha})(\|X\|_{\psi_\alpha}+\|\mathbb{E}X\|_{\psi_\alpha}) \le \log_2^{1/\alpha}(1+e^{1/\alpha})\left(1+\log_2^{1/\alpha}(1+e^{1/\alpha})\right)\|X\|_{\psi_\alpha}.$$

■