Nonconvex Stochastic Scaled Gradient Descent and Generalized Eigenvector Problems

Chris Junchi Li[1] **Michael I. Jordan**[1,2]

[1]Department of Electrical Engineering and Computer Sciences, UC Berkeley, Berkeley, California, USA
[2]Department of Statistics, UC Berkeley, Berkeley, California, USA

Abstract

Motivated by the problem of online canonical correlation analysis, we propose the *Stochastic Scaled-Gradient Descent* (SSGD) algorithm for minimizing the expectation of a stochastic function over a generic Riemannian manifold. SSGD generalizes the idea of projected stochastic gradient descent and allows the use of scaled stochastic gradients instead of stochastic gradients. In the special case of a spherical constraint, which arises in generalized eigenvector problems, we establish a nonasymptotic finite-sample bound of $\sqrt{1/T}$, and show that this rate is minimax optimal, up to a polylogarithmic factor of relevant parameters. On the asymptotic side, a novel trajectory-averaging argument allows us to achieve local asymptotic normality with a rate that matches that of Ruppert-Polyak-Juditsky averaging. We bring these ideas together in an application to online canonical correlation analysis, deriving, for the first time in the literature, an optimal one-time-scale algorithm with an explicit rate of local asymptotic convergence to normality. Numerical studies of canonical correlation analysis are also provided for synthetic data.

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. 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 nonconvex 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 $\nabla F(\boldsymbol{v})$ 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 $\boldsymbol{v}$, dubbed as *scaled factor*. Examples of this setup arise most notably in generalized eigenvector (GEV) computation, which finds its applications in principal component analysis, partial least squares regression, Fisher's linear discriminant analysis, canonical correlation analysis (CCA), etc. 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

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

$\widetilde{\nabla} F(\boldsymbol{v}_{t-1};\boldsymbol{\zeta}_t)$ 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

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

As pointed out by recent works e.g. 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. At each step we take $\widetilde{\mathbf{A}}$ and $\widetilde{\mathbf{B}}'$ as mutually independent and unbiased stochastic samples of $\mathbf{A}$ and $\mathbf{B}$ respectively and proceed with the following update:

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

We refer to the rule (7) as an *online GEV iteration*. In the special case where the stochastic sample $\widetilde{\mathbf{B}}'$ is taken as $\mathbf{I}$, (7) essentially reproduces Oja's online PCA algorithm [Oja, 1982] with an incurred $O(\eta^2)$ higher-order error term.

To identify the iterative algorithm in (7) as a manifestation of SSGD, we rewrite the term in parentheses in the algorithm as follows (we set $\boldsymbol{v} = \boldsymbol{v}_{t-1}$ for brevity):

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

It can be easily seen that the expectation of (8) is a scaled gradient of the generalized Rayleigh quotient, where the scaled factor $D(\boldsymbol{v}) \equiv (\boldsymbol{v}^\top\mathbf{B}\boldsymbol{v})^2/2$. This approach, which has been referred to as *double stochastic sampling* in the setting of kernel methods [Dai et al., 2014, 2017], makes it possible to develop an efficient stochastic approximation algorithm. Indeed, often $\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}}'$ are of rank one, so the computation of matrix-vector products $\tilde{\mathbf{A}}\boldsymbol{v}, \tilde{\mathbf{B}}'\boldsymbol{v}$ only invokes inner products of vectors and is hence computationally efficient in the face of high dimensionality (i.e. when d is high).

Our contributions relative to previous work on nonconvex stochastic optimization as are follows. First, we propose a novel algorithm—the stochastic scaled-gradient descent (SSGD) algorithm—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 SSGD algorithm uses as few as two samples at each update, does not incur intermediate and expensive computational cost while achieving a polynomial convergence rate guarantee.

Related Literature The generalized eigenvector problem is at the core of many statistical problems such as principal component analysis [Pearson, 1901, Hotelling, 1933], canonical correlation analysis [Hotelling, 1936], Fisher's linear discriminant analysis [Fisher, 1936, Welling, 2005], partial least squares regression [Stone and Brooks, 1990], sufficient dimension reduction [Li, 1991], mixture models [Balakrishnan et al., 2017], along with their sparse counterparts. Iterative algorithms for sparse principal component analysis has been proposed by Ma [2013] and Yuan and Zhang [2013] as a special case of the eigenvalue problem: by adding a soft-thresholding step to each power method step their algorithms achieve linear convergence. In follow-up work, Tan et al. [2018] proposed a truncated Rayleigh flow algorithm to estimate the leading sparse generalized eigenvector that also achieves a linear convergence rate. Additional work on generalized eigenvector computation includes Ge et al. [2016], Allen-Zhu and Li [2017a], Yuan et al. [2019], Ma et al. [2015], Chaudhuri et al. [2009].

Some recent work has focused on developing efficient online procedures for particular instances of generalized eigenvector problems, among which online principal and canonical eigenvectors estimation has been of particular interest. Oja's online PCA iteration [Oja, 1982], which can be reproduced from (7) when $\widetilde{\mathbf{B}}$ is taken as $\mathbf{I}$ as a special case, up to an incurred $O(\eta^2)$ error term, has been shown to provably match the minimax information lower bound [Jain et al., 2016, Li et al., 2018, Allen-Zhu and Li, 2017b]. There is also a rich literature on stochastic gradient methods for convex and nonconvex minimization that takes place on Riemannian manifolds [Ge et al., 2015, Zhang and Sra, 2016]; we refer the readers to Hosseini and Sra [2020] for a recent survey study. More related to our work, procedures for efficient online canonical eigenvectors estimation have been

explored [Arora et al., 2017, Gao et al., 2019, Chen et al., 2019]. Among these works, Gao et al. [2019] developed a streaming canonical correlation analysis (CCA) algorithm which involves solving a large linear system at each iteration, and independently Arora et al. [2017] proposed a different stochastic CCA algorithm which has temporal and spatial complexities that are quadratic in d. Chen et al. [2019] present a landscape analysis of GEV/CCA and provide a continuous-time insight for a class of primal-dual algorithms when the two matrices in GEV commute; the convergence analysis of Chen et al. [2019], however, does *not* directly translate to discrete-time convergence rate bounds and no explicit analysis has been provided when two matrices do *not* commute.

In a recent paper, 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 [Oja, 1982] for online principal component estimation with Markovian noise [Shamir, 2016, Jain et al., 2016, Li et al., 2018, Allen-Zhu and Li, 2017b]. Bhatia et al. [2018] obtained a convergence rate under a bounded sample assumption that achieves the minimax rate $1/\sqrt{N}$ in terms of the sample size N. In comparison, our proposed SSGD algorithm is a single time-scale algorithm with a single step-size and an extra requirement of two (independent) samples per iterate. The algorithm is minimax optimal with respect to local convergence and hence theoretically comparable with Gen-Oja.

Organization The rest of this paper is organized as follows. §2 states our settings and assumptions throughout the theoretical analysis of our paper. §3 presents our local convergence results under the warm initialization condition. §4 presents our two-phase convergence results for arbitrary initialization. §5 investigates the asymptotic property of our algorithm. §6 uses the example of Canonical Correlation Analysis to demonstrate the practical computation and experimental performance of our algorithm. §7 summarizes the entire paper. Limited by space we relegate to Appendix all our theoretical analysis and secondary lemmas.

Notation Unless indicated otherwise, C denotes some positive, absolute constant which may change from line to line. 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 $\lambda_{\max}(\mathbf{A})$, $\lambda_{\min}(\mathbf{A})$ and $\|\mathbf{A}\|$ denote the maximal, minimal eigenvalues and the operator norm of a real symmetric matrix $\mathbf{A}$. We will explain other notation at its first appearance.

2 SETTINGS AND ASSUMPTIONS

In this section, we present the settings and assumptions required by our 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 our proposed SSGD iteration (4) 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 $\mathcal{F}_t = \sigma\big(\boldsymbol{\zeta}_s : s \leq t\big)$ be the filtration generated by the stochastic process $\boldsymbol{\zeta}_t$. Then, from (3), we have $\mathbb{E}[\Gamma(\boldsymbol{v}_{t-1}; \boldsymbol{\zeta}_t) \mid \mathcal{F}_{t-1}] = D(\boldsymbol{v}_{t-1})\nabla F(\boldsymbol{v}_{t-1})$. That is, the conditional expectation is a scaled gradient. The ensuing analysis is analogous to that of locally convex SGD 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.

Following the classical theory of constrained optimization [Nocedal and Wright, 2006] we introduce a definition of *manifold gradient* and *manifold Hessian* in the presence of a unit spherical constraint, $\mathcal{C} : c(\boldsymbol{v}) = (1/2)(\boldsymbol{v}^\top \boldsymbol{v} - 1) = 0$.[1] For this equality-constrained optimization problem, we utilize the method of Lagrange multipliers and introduce the following Lagrangian function: $L(\boldsymbol{v}; \mu) = F(\boldsymbol{v}) - \frac{\mu}{2}\left(\|\boldsymbol{v}\|^2 - 1\right)$. 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. We first define the Lipschitz continuity for a generic mapping:

Definition 1 (Lipschitz Continuity) *Let $\mathcal{M}$ be a finite-dimensional normed vector space. The map $M : \mathbb{R}^d \mapsto \mathcal{M}$ is called L_M-Lipschitz, if for any two points $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{R}^d$ $\|M(\boldsymbol{v}) - M(\boldsymbol{v}')\|_{\mathcal{M}} \leq L_M\|\boldsymbol{v} - \boldsymbol{v}'\|$, where $\|\cdot\|_{\mathcal{M}}$ is any norm properly defined in space $\mathcal{M}$.*

[1] Here for notational simplicity we incorporate a factor of $1/2$.

In addition, we need the following assumption on the state-dependent scalar $D(\boldsymbol{v})$ and covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{v})$. For a fixed $\boldsymbol{v}$, define the state-dependent covariance $\boldsymbol{\Sigma}(\boldsymbol{v})$ to be

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

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 that $F(\boldsymbol{v})$ and $\nabla F(\boldsymbol{v})$ are 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.*

Now we pose some tail behavior of the stochastic vectors $\Gamma(\boldsymbol{v}_{t-1};\boldsymbol{\zeta}_t), t \ge 1$ as *vector α-sub-Weibull*, as in the following definition:

Assumption 2 (Sub-Weibull Tail) *For some fixed $\alpha \in (0, 2]$ and for all $\boldsymbol{v} \in \mathcal{C}$, we assume that the stochastic vectors $\Gamma(\boldsymbol{v};\boldsymbol{\zeta})$ satisfy $\mathbb{E}\exp\left(\|\Gamma(\boldsymbol{v};\boldsymbol{\zeta})\|^\alpha/\mathcal{V}^\alpha\right) \le 2$, where $\mathcal{V}$ is called the* sub-Weibull parameter *of stochastic vector $\Gamma(\boldsymbol{v};\boldsymbol{\zeta})$.*

Note here the sub-Weibull parameter is in the vector-norm sense instead of the maximal projected scalar sense. The class of sub-Weibull distributions contains the sub-Gaussian ($\alpha = 2$) and sub-Exponential ($\alpha = 1$) distribution classes as special cases [Wainwright, 2019, Kuchibhotla and Chakrabortty, 2018]. Background on vector α-sub-Weibull distributions (and the associated notion of Orlicz ψ_α-norm) are provided in Appendix §C.

3 LOCAL CONVERGENCE ANALYSIS

In this section we provide the main local convergence result for our SSGD algorithm. Our local analysis is inspired from both generic [Ge et al., 2015] and dynamics-based [Li et al., 2018, Li and Jordan, 2021] analyses for nonconvex stochastic gradient descent, which we further adapt to our scaled-gradient setup.

For notational simplicity, we denote

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

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. In the ensuing Theorem 2 on local convergence, we take $\epsilon \in (0, 1)$ and define the following quantities:

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

and for $\eta < 1/(D\mu)$, define

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

We state our local convergence theorem.

Theorem 2 (Local Convergence) *Given Assumptions 1 and 2 as well as the initialization condition* (14)*, for any positive constants η, ϵ that satisfy the scaling condition*

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

and for any $T \ge K_{\eta,\epsilon}T_\eta^*$, 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 for all $t \in [K_{\eta,\epsilon}T_\eta^*, T]$:

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

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 $\boldsymbol{\Delta}_t$ as the projection of $\boldsymbol{v}_t - \boldsymbol{v}^*$ onto the tangent space $\mathcal{T}(\boldsymbol{v}^*)$, namely $\boldsymbol{\Delta}_t = (\mathbf{I} - \boldsymbol{v}^*\boldsymbol{v}^{*\top})(\boldsymbol{v}_t - \boldsymbol{v}^*)$. 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. Note that $K_{\eta,\epsilon}T_\eta^*$ 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 present a proposition that provides an upper bound on $\|\boldsymbol{\Delta}_t\|$ over T iterations and characterizes the descent in $\|\boldsymbol{\Delta}_t\|$ 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 $T \geq 1$, 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 $t \in [0, T]$,*

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

and

$$\|\boldsymbol{\Delta}_t\| \leq 4 \max\left\{\frac{\|\boldsymbol{\Delta}_0\|}{2},\ \frac{2^6 G_\alpha \mathcal{V}}{\sqrt{D\mu}} \log^{\frac{\alpha+2}{2\alpha}} \epsilon^{-1} \cdot \eta^{1/2}\right\}. \tag{20}$$

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

$$\|\boldsymbol{\Delta}_{T_\eta^*}\| \leq \max\left\{\frac{\|\boldsymbol{\Delta}_0\|}{2},\ \frac{2^6 G_\alpha \mathcal{V}}{\sqrt{D\mu}} \log^{\frac{\alpha+2}{2\alpha}} \epsilon^{-1} \cdot \eta^{1/2}\right\}. \tag{21}$$

The proof of Proposition 3 is provided in §B.1.

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 $\|\boldsymbol{\Delta}_t\|$ are bounded by each other up to constant factors, (20) guarantees that $\|\boldsymbol{\Delta}_t\|$ does not exceed $\max\left\{2\|\boldsymbol{\Delta}_0\|, \Theta(\eta^{1/2})\right\}$—that is, $\boldsymbol{v}_t$ stays in a neighborhood of local minimizer $\boldsymbol{v}^*$—and (21) states that, for $\|\boldsymbol{\Delta}_0\| = \Omega(\eta^{1/2})$, $\|\boldsymbol{\Delta}_t\|$ decreases by half after T_η^* iterations: $\|\boldsymbol{\Delta}_{T_\eta^*}\| \leq \max\left\{\|\boldsymbol{\Delta}_0\|/2, \Theta(\eta^{1/2})\right\}$.

Proposition 3 studies $\boldsymbol{\Delta}_t$ in a single round, i.e., for T_η^* iterations. We are ready to provide the proof of Theorem 2 by applying Proposition 3 repeatedly for $K_{\eta,\epsilon}$ rounds, detailed as follows:

Proof of Theorem 2 Since the algorithm iteration (4) can be viewed as a (strong) discrete-time Markov process, We recall the definition of $K_{\eta,\epsilon}$ in (15) and repeatedly apply Proposition 3 to the sequence of $\{\boldsymbol{\Delta}_t\}$ for $K_{\eta,\epsilon}$ rounds, initializing each round with the output $\boldsymbol{\Delta}_{T_\eta^*}$ from the previous round. We adopt an adaptive argument of shrinkage in multiple rounds.

More specifically, 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

$$\begin{aligned}
&\|\boldsymbol{v}_t - \boldsymbol{v}^*\| \leq \sqrt{2}\|\boldsymbol{\Delta}_t\| \\
&\leq \sqrt{2} \cdot 4 \max\left\{\frac{\|\boldsymbol{\Delta}_{K_{\eta,\epsilon}T_\eta^*}\|}{2}, \frac{2^6 G_\alpha \mathcal{V}}{\sqrt{D\mu}} \log^{\frac{\alpha+2}{2\alpha}} \epsilon^{-1} \cdot \eta^{1/2}\right\} \\
&\leq 4\sqrt{2} \cdot \max\left\{\frac{\|\boldsymbol{\Delta}_0\|}{2^{K_{\eta,\epsilon}}}, \frac{2^6 G_\alpha \mathcal{V}}{\sqrt{D\mu}} \log^{\frac{\alpha+2}{2\alpha}} \epsilon^{-1} \cdot \eta^{1/2}\right\} \\
&\leq \frac{2^{\frac{17}{2}} G_\alpha \mathcal{V}}{\sqrt{D\mu}} \log^{\frac{\alpha+2}{2\alpha}} \epsilon^{-1} \cdot \eta^{1/2},
\end{aligned}$$

where the last inequality is due to initialization condition (14). Here G_α is a fixed positive factor depending on α, as defined in Theorem 2. By taking a union bound over $K_{\eta,\epsilon}$ rounds and $T - K_{\eta,\epsilon}T_\eta^*$ iterations, we obtain

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

completing the proof of Theorem 2. □

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

Corollary 4 (Finite-Sample) *Assume Assumptions 1 and 2 and the initialization condition* (14). *For fixed positive constants ϵ and sample size T, set the step-size as $\eta(T) = \Theta\left(\frac{\log T}{D\mu T}\right)$ satisfying scaling condition*

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

there exists an event $\mathcal{H}_4$ with $\mathbb{P}(\mathcal{H}_4) \geq 1 - \left(14 + 8\left(\frac{3}{\alpha}\right)^{\frac{2}{\alpha}} \log^{-\frac{\alpha+2}{\alpha}} \epsilon^{-1}\right) T\epsilon$, 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.

4 GLOBAL CONVERGENCE ANALYSIS

In many situations, solving the warm initialization problem itself can be a difficult problem. We borrow the techniques from Ge et al. [2015] and establish a global convergence result for *escaping saddle points* via SSGD. In this section we consider a variant of SSGD with a unit spherical constraint and equipped with an artificial noise injection step: let $\boldsymbol{n}_t$ be an independent spherical noise at each step that is independent of $\mathcal{F}_{t-1}$ and $\boldsymbol{\zeta}_t$, and let

$$\boldsymbol{v}_t = \Pi_{\mathcal{S}^{d-1}}\left[\boldsymbol{v}_{t-1} - \eta\widetilde{\nabla}F(\boldsymbol{v}_{t-1}; \boldsymbol{\zeta}_t) + \eta\boldsymbol{n}_t\right]. \tag{22}$$

Motivated by recent work on escaping saddle points [Ge et al., 2015, Lee et al., 2016, Jin et al., 2019], one can show that SSGD algorithm equipped with the aforementioned artificial noise injection escapes from all saddle points, and hence the initialization condition (14) can be dropped.

First, we generalize Assumption 1 for local convergence to the following for global convergence:

Assumption 3 (Global Smoothness and Boundedness) *For any $\boldsymbol{v} \in \{\boldsymbol{v} : \|\boldsymbol{v}\| \le 1\}$, 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. Also, assume there exists $D_-, D_+ > 0$ such that $D_- \le D(\boldsymbol{v}) \le D_+$ for all $\boldsymbol{v}$.*

Definition 5 (Strict-Saddle Function) *A twice differentiable function $F(\boldsymbol{v})$ with constraint $c(\boldsymbol{v}) = 0$ is called an $(\mu, \beta, \gamma, \delta)$-strict-saddle function, if an arbitrary point $\boldsymbol{v}$ with $c(\boldsymbol{v}) = 0$ satisfies at least one of the following:*

(i) $\|g(\boldsymbol{v})\| \ge \beta$;

(ii) There is a local minimizer $\boldsymbol{v}^*$ such that $\|\boldsymbol{v} - \boldsymbol{v}^*\| \le \delta$. Additionally, for all $\boldsymbol{v}' \in B_{2\delta}(\boldsymbol{v}^*)$, we have

$$\boldsymbol{v}_1^\top \mathcal{H}(\boldsymbol{v}')\boldsymbol{v}_1 \ge \mu, \quad \forall \boldsymbol{v}_1 \in \mathcal{T}(\boldsymbol{v}') \text{ and } \|\boldsymbol{v}_1\| = 1.$$

(iii) There exists a unit vector $\boldsymbol{v}_0 \in \mathcal{T}(\boldsymbol{v})$ such that $\boldsymbol{v}_0^\top \mathcal{H}(\boldsymbol{v})\boldsymbol{v}_0 \le -\gamma$.

In what follows, we show that our algorithms can escape from all saddle points and thus the local initialization is no longer required. We are ready to present the saddle-point escaping result:

Theorem 6 (Escaping from Saddle Points) *Let Assumptions 2 and 3 hold. Let $F(\boldsymbol{v})$ be a $(\mu, \beta, \gamma, \delta)$-strict-saddle function with finite sup-norm $\|F\|_\infty$. Let*

$$T_1 = 4\|F\|_\infty \cdot \left[\min\left(0.5dL_G, \gamma \log^{-1}\left(\frac{6d\mathcal{V}}{\sigma}\right)\right) \cdot \sigma^2 D_-^2 \eta^2\right]^{-1} \tag{23}$$

Then for any $\kappa > 0$ and any step-size $\eta > 0$ satisfying

$$\sqrt{2d\mathcal{V}^2 L_G D_+ \eta} \le \beta, \tag{24}$$

within $T_1 \cdot \lceil \log_2(\kappa^{-1}) \rceil$ iterates, (22) *outputs $\boldsymbol{v}_t$ that satisfies (ii) in Definition 5 with probability no less than $1 - \kappa$.*

The proof of Theorem 6 is collected in §B.2. Motivated by this saddle-point escaping result, one can run SSGD first with a *burn-in* phase and once it enters the warm initialization region, one can re-run SSGD with step-sizes chosen so that the local convergence theorem applies immediately. Using the strong Markov property and combining Theorems 2 and 6 we immediately obtain the following main theorem. Recall that T_1 is defined as in (23).

Theorem 7 (Two-Phase Global Convergence) *Let Assumptions 2 and 3 hold. Let η satisfy*

$$\eta \le \min\left\{\frac{D^3\mu^3}{2^{24}G_\alpha^2 \mathcal{V}^2 \rho^2} \log^{-\frac{\alpha+2}{\alpha}} \epsilon^{-1}, \frac{1}{D\mu}, \frac{\beta^2}{2d\mathcal{V}^2 L_G D_+}\right\}, \tag{25}$$

and for any $T \ge K_{\eta,\epsilon} T_\eta^* + T_1 \cdot \lceil \log_2(\kappa^{-1}) \rceil$, there exists an event $\mathcal{A}_T$ with

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

such that on event $\mathcal{A}_T$ the iterates generated by the SSGD algorithm satisfy for all $t \in \left[K_{\eta,\epsilon} T_\eta^* + T_1 \cdot \lceil \log_2(\kappa^{-1}) \rceil, T\right]$

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

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.

Note the function class of strict-saddle functions is strictly more general than the local convergence Theorem 2. We find the final complexity by interpreting Theorem 7. In the asymptotic relations below we write out the dependency on d, η, and let $\mathcal{L}$ be a generic quantity that only involves a polylogarithmic factor of d, η and T, which is allowed to vary at each appearance. From (15), (16) and (23) we have

$$K_{\eta,\epsilon} T_\eta^* \asymp \mathcal{L} \cdot \eta^{-1}, \quad T_1 \cdot \lceil \log_2(\kappa^{-1}) \rceil \asymp \mathcal{L} \cdot d^{-1}\eta^{-2},$$

and if $\mathcal{V}$ is set as the model scaling $\sqrt{d}$, the iteration achieves a high-probability bound of $\mathcal{L} \cdot \sqrt{d\eta}$ after $K_{\eta,\epsilon} T_\eta^* + T_1 \cdot \lceil \log_2(\kappa^{-1}) \rceil$ steps. We conclude that under the scaling condition $\mathcal{L} \cdot d/T \to 0$, if the total number of samples T is given, we can optimize the choice of step-size $\eta = \eta(d, T)$ to conclude the following convergence rate results:

(i) **Local Convergence:** Given a *warm* initialization, and choosing $\eta(T) \asymp \mathcal{L} \cdot (1/T)$, SSGD (4) has the following *local* convergence rate

$$\|\boldsymbol{v}_t - \boldsymbol{v}^*\| \lesssim \mathcal{L} \cdot \sqrt{\frac{d}{T}}.$$

(ii) **Global Convergence:** Given *any* initialization, and choosing $\eta(T) \asymp \mathcal{L} \cdot (1/\sqrt{dT})$, SSGD with noise injection (22) has the following *global* convergence rate

$$\|\boldsymbol{v}_t - \boldsymbol{v}^*\| \lesssim \mathcal{L} \cdot \sqrt[4]{\frac{d}{T}}.$$

We defer the arguments for the proof to §B.2, and turn to the application to GEV problem.

5 ASYMPTOTIC NORMALITY VIA TRAJECTORY AVERAGING

In this section, we return to the warm initialization as in §3. Ruppert [1988] and Polyak and Juditsky [1992] introduced the idea of trajectory averaging for stochastic

gradient descent in order to provide fine-grained convergence rates along with an asymptotic normality result. Our goal is to generalize the Polyak-Juditsky analysis of SGD with trajectory averaging to SSGD for nonconvex objective that is initialized in a local convex region. We denote $\mathcal{H}_* \equiv \mathcal{H}(\boldsymbol{v}^*), \boldsymbol{\Sigma}_* \equiv \boldsymbol{\Sigma}(\boldsymbol{v}^*)$ and $D \equiv D(\boldsymbol{v}^*)$. 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 $\boldsymbol{u}^\top \mathcal{M}_* \boldsymbol{u} \geq \mu\|\boldsymbol{u}\|^2$ for all $\boldsymbol{u} \in \mathcal{T}(\boldsymbol{v}^*)$. We consider the eigendecomposition $\mathcal{M}_* = \boldsymbol{P}\,\mathrm{diag}(\lambda_1, \ldots, \lambda_{d-1}, 0)\boldsymbol{P}^\top$ for an orthogonal matrix $\boldsymbol{P} \in \mathbb{R}^{d\times d}$ and eigenvalues $\lambda_1 \geq \ldots \geq \lambda_{d-1} > 0$ with minimum positive eigenvalue $\lambda_{d-1} \geq \mu$. We take the inverse of all positive eigenvalues and define the following matrix

$$\mathcal{M}_*^- \equiv \boldsymbol{P}\,\mathrm{diag}(\lambda_1^{-1}, \ldots, \lambda_{d-1}^{-1}, 0)\boldsymbol{P}^\top. \tag{26}$$

Here, $\mathcal{M}_*^-$ 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 $T \geq K_{\eta,\epsilon}T_\eta^*$, we define the trajectory average over time $K_{\eta,\epsilon}T_\eta^* + 1, \ldots, T$ as follows:

$$\overline{\boldsymbol{v}}_T^{(\eta)} \equiv \frac{1}{T - K_{\eta,\epsilon}T_\eta^*} \sum_{t=K_{\eta,\epsilon}T_\eta+1}^{T} \boldsymbol{v}_t, \tag{27}$$

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

Assumption 4 (Mean-Squared Smoothness) *There exists a positive constant L_S such that for all $\boldsymbol{v}, \boldsymbol{v}' \in \{\boldsymbol{v} : \|\boldsymbol{v}\| \leq 1, \|\boldsymbol{v} - \boldsymbol{v}^*\| \leq \delta\}$ and $t \geq 1$, we have for $\boldsymbol{\zeta}$*

$$\mathbb{E}\,\|\Gamma(\boldsymbol{v};\boldsymbol{\zeta}) - \Gamma(\boldsymbol{v}';\boldsymbol{\zeta})\|^2 \leq L_S^2\|\boldsymbol{v} - \boldsymbol{v}'\|^2. \tag{28}$$

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 8 (Asymptotic Normality) *Given Assumptions 1, 2, 4 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*

$$T\eta^2 \log^{\frac{2\alpha+4}{\alpha}} T \to 0, \quad T\eta \log^{-\frac{\alpha+2}{\alpha}} T \to \infty \quad a.s., \tag{29}$$

we obtain Gaussian convergence in distribution:

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

We relegate the proof details of Theorem 8 to §B.4.[2] The analysis has the same rationale as the classical asymptotic normality result that is obtained when minimizing a strongly convex objective function in an Euclidean space using stochastic gradient descent [Ruppert, 1988, 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 a 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, we base our analysis on the idea that trajectory averaging begins only after "the burn-in phase"; that is, after $K_{\eta,\varepsilon}T_\eta^*$ iterates.

6 CASE STUDIES OF CANONICAL CORRELATION ANALYSIS

The GEV problem arises in many statistical machine learning tasks. We focus on the example of (rank-one) Canonical Correlation Analysis (CCA) as a core application; we refer to Tan et al. [2018] for other (sparse, high-dimensional) applications including linear discriminant analysis and sliced inverse regression. Recall that CCA aims at maximizing the correlation between two transformed vectors. Given $\boldsymbol{X}$ and $\boldsymbol{Y}$ as two column vectors, let $\boldsymbol{\Sigma}_{\boldsymbol{XY}}$ be the cross-covariance matrix between $\boldsymbol{X}$ and $\boldsymbol{Y}$, and let $\boldsymbol{\Sigma}_{\boldsymbol{XX}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{YY}}$ be the covariance matrices of $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively. CCA is a special case of the GEV problem (5) with

$$\mathbf{A} = \begin{pmatrix} \mathbf{0} & \boldsymbol{\Sigma}_{\boldsymbol{XY}} \\ \boldsymbol{\Sigma}_{\boldsymbol{YX}} & \mathbf{0} \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} \boldsymbol{\Sigma}_{\boldsymbol{XX}} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{\boldsymbol{YY}} \end{pmatrix}.$$

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 $(\boldsymbol{X}, \boldsymbol{Y}), (\boldsymbol{X}', \boldsymbol{Y}')$ at each iteration and compute

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

where all samples of $\boldsymbol{X}, \boldsymbol{Y}$ are centered such that they have expectation zero.

In order to apply the convergence results for the SSGD algorithm to the CCA problem, it remains to verify Assumption 2. We assume that the samples $\boldsymbol{X} \in \mathbb{R}^{d_x}, \boldsymbol{Y} \in \mathbb{R}^{d_y}$ follow sub-Gaussian distributions [Gao et al., 2019, Li et al., 2018] with parameters $\mathcal{V}_x, \mathcal{V}_y$; that is, $\mathbb{E}\exp\left(\|\boldsymbol{X}\|^2/\mathcal{V}_x^2\right) \leq 2$ and $\mathbb{E}\exp\left(\|\boldsymbol{Y}\|^2/\mathcal{V}_y^2\right) \leq 2$. With these standard assumptions for the samples $\boldsymbol{X}, \boldsymbol{Y}$, the following lemma shows that the scaled-gradient noise in the CCA problem satisfies Assumption 2 with appropriate $\mathcal{V}$ and α. The proof is provided in §B.5.

[2] The limiting distribution is supported on a submanifold of the Euclidean space $\mathbb{R}^d$. The convergence in distribution is hence rigorously characterized by the pointwise convergence of the characteristic functions.

Algorithm 1 Online Canonical Correlation Analysis via Noise-Injected Stochastic Scaled-Gradient Descent

input total sample size T, proper stepsize η, initialize $\boldsymbol{v}_0$
for $t = 1, \ldots, T/2$ **do**
Draw mutually independent sample pairs $(\boldsymbol{X}, \boldsymbol{Y})$ and $(\boldsymbol{X}', \boldsymbol{Y}')$ from the stochastic oracle
Compute unbiased estimates

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

Sample a uniformly spherical noise $\boldsymbol{n}_t$ of covariance $\sigma^2 \mathbf{I}_d$ and update $\boldsymbol{g}_t, \boldsymbol{v}_t$ using the following rule

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

end for
return $\boldsymbol{v}_T$

Proposition 9 *Assumption 2 holds for CCA with parameters* $\mathcal{V} = 400(\mathcal{V}_x^2 + \mathcal{V}_y^2)\mathcal{V}_x \mathcal{V}_y$ *and* $\alpha = 1/2$.

Lemmas 10 and 9 certify that Assumptions 1 and 2 hold in CCA settings and hence local convergence Corollary 4 applies, which establishes a $\sqrt{d/T}$-rate up to a polylogarithmic since the vector sub-Weibull parameter $\mathcal{V}$ in our Assumption 2 implicitly contains a factor $\sqrt{d}$.

Now we demonstrate that our bounds in Corollary 4 match the lower bound. Gao et al. [2019] derived a lower bound for Gaussian variables, $1 - \mathrm{align}(\boldsymbol{v}, \boldsymbol{v}^*) \gtrsim d/T$, in terms of a new measure of error:

$$\mathrm{align}(\boldsymbol{v}, \boldsymbol{v}^*) \equiv \frac{1}{2} \left(\frac{\boldsymbol{v}_x^\top \boldsymbol{\Sigma}_{\boldsymbol{X}\boldsymbol{X}} \boldsymbol{v}_x^*}{\sqrt{\boldsymbol{v}_x{}^\top \boldsymbol{\Sigma}_{\boldsymbol{X}\boldsymbol{X}} \boldsymbol{v}_x} \sqrt{\boldsymbol{v}_x^*{}^\top \boldsymbol{\Sigma}_{\boldsymbol{X}\boldsymbol{X}} \boldsymbol{v}_x^*}} + \frac{\boldsymbol{v}_y^\top \boldsymbol{\Sigma}_{\boldsymbol{Y}\boldsymbol{Y}} \boldsymbol{v}_y^*}{\sqrt{\boldsymbol{v}_y{}^\top \boldsymbol{\Sigma}_{\boldsymbol{Y}\boldsymbol{Y}} \boldsymbol{v}_y} \sqrt{\boldsymbol{v}_y^*{}^\top \boldsymbol{\Sigma}_{\boldsymbol{Y}\boldsymbol{Y}} \boldsymbol{v}_y^*}} \right),$$

where $\boldsymbol{v} = (\boldsymbol{v}_x^\top, \boldsymbol{v}_y^\top)^\top$ and $\boldsymbol{v}^* = (\boldsymbol{v}_x^{*\top}, \boldsymbol{v}_y^{*\top})^\top$ are partitioned in dimensions d_x, d_y. It is easy to verify that $1 - \mathrm{align}(\boldsymbol{v}, \boldsymbol{v}^*) \asymp 1 - \boldsymbol{v}^\top \boldsymbol{v}^2 = \|\boldsymbol{v} - \boldsymbol{v}^*\|^2/2$ when both $\boldsymbol{v}, \boldsymbol{v}^*$ lie on the unit sphere, in which case our lower bound translates into $\|\boldsymbol{v}_T - \boldsymbol{v}^*\| \gtrsim \sqrt{d/T}$ for any estimator $\boldsymbol{v}_T$ that consumes T samples, which matches the upper bound of Corollary 4 in terms of both d and T.

We note that our Corollary 4 and the results of Gao et al. [2019] have different dimension dependency, which is due to a distinct but connected set of assumptions. We have assumed that each sample $\boldsymbol{X}, \boldsymbol{Y}$ follows a vector sub-Gaussian distribution and verifies Assumption 2 required by Proposition 9, whereas Gao et al. [2019] assume that each coordinate of $\boldsymbol{X}, \boldsymbol{Y}$ is sub-Gaussian with a constant parameter. Hence, the vector sub-Gaussian parameter $\mathcal{V}$ in our case suffers a dimension-dependent prefactor.

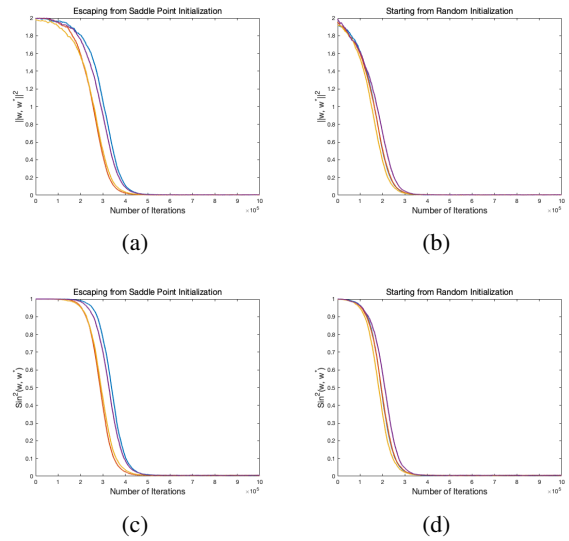


Figure 1: Comparison between saddle point initialization and random initialization

6.1 NUMERICAL STUDIES USING SYNTHETIC DATA

In this subsection, we present simulation results for SSGD for the case of rank-one CCA [Algorithm 1]. The dimensions of the synthetic data samples are picked as $d_1 = 65$ of $\boldsymbol{X}$ and $d_2 = 70$ of $\boldsymbol{Y}$. We generate the covariance matrix for $\boldsymbol{X}, \boldsymbol{Y}$ as

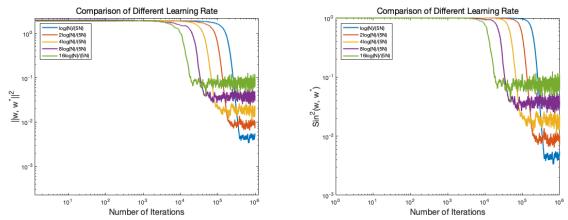
$$\boldsymbol{\Sigma}_{\boldsymbol{X}\boldsymbol{X}} = 3\mathbf{I}_{d_1} + \mathbf{A}_1, \quad \boldsymbol{\Sigma}_{\boldsymbol{Y}\boldsymbol{Y}} = 3\mathbf{I}_{d_2} + \mathbf{A}_2, \tag{31}$$

where $\mathbf{A}_1, \mathbf{A}_2$ are diagonal matrices with each entry along the diagonal obtained as an independent uniform draw from $[0, 1]$. To ensure the eigengap of $\boldsymbol{\Sigma}_{\boldsymbol{X}\boldsymbol{X}}^{-\frac{1}{2}} \boldsymbol{\Sigma}_{\boldsymbol{X}\boldsymbol{Y}} \boldsymbol{\Sigma}_{\boldsymbol{Y}\boldsymbol{Y}}^{-\frac{1}{2}}$ is significantly large, in particular, no less than 0.5, we set

$$\boldsymbol{\Sigma}_{\boldsymbol{X}\boldsymbol{Y}} = \mathbf{A}_3 + \boldsymbol{\Sigma}_{\boldsymbol{X}\boldsymbol{X}}^{1/2} \mathbf{U} \, \mathrm{diag}(0.5, \mathbf{O}) \mathbf{V}^\top \boldsymbol{\Sigma}_{\boldsymbol{Y}\boldsymbol{Y}}^{1/2}. \tag{32}$$

Here $\mathbf{A}_3$ is a $d_1 \times d_2$ matrix where each entry is generated from an independent $N(0, 1/(d_1 + d_2))$ variable with SVD decomposition $\boldsymbol{\Sigma}_{\boldsymbol{X}\boldsymbol{X}}^{1/2} \mathbf{A}_3 \boldsymbol{\Sigma}_{\boldsymbol{Y}\boldsymbol{Y}}^{1/2} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$, and $\mathbf{O}$ is a $(d_1 - 1) \times (d_2 - 1)$ zero matrix. Note that each step of Algorithm 1 can be computed in time $\mathcal{O}(d_1 + d_2)$. Given this setup, we report our numerical findings of Algorithm 1 as follows:

Saddle-point escaping We first discuss the behavior of our algorithm in the presence of saddle points. When $\boldsymbol{v}_0$ is exactly chosen as a saddle point, we show that SSGD escapes from a plateau of saddle points in the landscape and converges to the local (and global) minimizer. For illustrative purposes, the initialization $\boldsymbol{v}_0$ is chosen from four



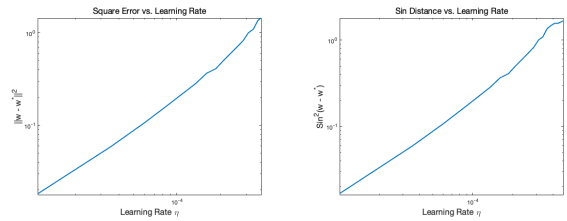
(a) (b)

Figure 2: Log-log plot regarding the convergence with respect to a range of step-sizes η. Figure 2(a) illustrates the squared errors in terms of squared distance to optimality $\|\boldsymbol{v} - \boldsymbol{v}^*\|^2$, and Figure 2(b) does so in terms of $\sin^2(\boldsymbol{v}, \boldsymbol{v}^*)$

saddle points, each of which corresponds to a component of CCA. We choose the total sample size $T =$ 1e6 and set the (constant) step-size $\eta = \log(T)/(5T)$. In Figure 1 we plot the error of the current solution to the optimal solution, where the error is measured both in squared Euclidean distance and in sine-squared. The first two plots shows the behavior initialized from four different saddle points, and the last two plots shows the behavior initialized from four uniform seeds. The horizontal axis is the number of iterates and the vertical axis is error $\|\boldsymbol{v}_t - \boldsymbol{v}^*\|^2$.

Relationship between the step-size and squared error We study the role of step-size η in our SSGD algorithm. Set sample size $T =$ 1e6 and choose 20 η's from 1e–5 to 5e–4 from $\{\log(T)/(5T), 2\log(T)/(5T), 4\log(T)/(5T), 8\log(T)/(5T), 16\log(T)/(5T)\}$ and plot the squared error $\|\boldsymbol{v} - \boldsymbol{v}^*\|^2$ on a log-log scale. It is clearly observed from Figure 2 that smaller step-sizes lead to slower convergence to a stationary point of smaller variance.

We now numerically demonstrate that at stationarity SSGD presents a squared error $\|\boldsymbol{v} - \boldsymbol{v}^*\|^2$ or $\sin^2(\boldsymbol{v}, \boldsymbol{v}^*)$ that has a linear relationship with η. We compute the averaged squared error of the last 10% iterates for each run and plot the result in Figure 3 in a log-log scale. The horizontal axes of both Figures 3(a) and 3(b) represent the step-size η, and the vertical axes of both figures are the squared error $\|\boldsymbol{v} - \boldsymbol{v}^*\|^2$ and $\sin^2(\boldsymbol{v}, \boldsymbol{v}^*)$, respectively. We compute an averaged squared error of the last 10% iterates for each η. Due to ergodicity in the algorithmic final phase, this provides a feasible estimate of its variance around the local (and global) minimizer. Also, the fitting slope of Figure 3 provided by the least-square method is 0.9921 (fairly close to 1), which corroborates our theoretical convergence results in Theorems 2 and 7. These numerical findings are consistent with our theory that the squared error $\|\boldsymbol{v} - \boldsymbol{v}^*\|^2$ at stationarity has a linear relationship with η.



(a) (b)

Figure 3: Relationship between step-size η and the squared error of our algorithmic estimator to the optimal solution

7 SUMMARY

We have presented the Stochastic Scaled-Gradient Descent (SSGD) 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, allowing 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 the trajectory average. An application to the generalized eigenvector 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.

Acknowledgements

We thank the Department of Electrical Engineering and Computer Sciences at UC Berkeley for COVID-19 accommodations during which time this work is completed. We thank Tong Zhang, Huizhuo Yuan, Yuren Zhou for inspiring discussions at various stages of this project. This work was supported in part by the Mathematical Data Science program of the Office of Naval Research under grant number N00014-18-1-2764.

References

Zeyuan Allen-Zhu 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, 2017a.

Zeyuan Allen-Zhu and Yuanzhi Li. First 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.

Sivaraman Balakrishnan, Martin J Wainwright, and Bin Yu. Statistical guarantees for the EM algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45(1):77–120, 2017.

Kush Bhatia, Aldo Pacchiano, Nicolas Flammarion, Peter L Bartlett, and Michael I Jordan. Gen-Oja: A two-time-scale approach for streaming CCA. *arXiv preprint arXiv:1811.08393*, 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.

Zhehui Chen, Xingguo Li, Lin Yang, Jarvis Haupt, and Tuo Zhao. On constrained nonconvex stochastic optimization: A case study for generalized eigenvalue decomposition. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 916–925, 2019.

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. In *Artificial Intelligence and Statistics*, pages 1458–1467, 2017.

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.

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

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.

Reshad Hosseini and Suvrit Sra. Recent advances in stochastic Riemannian optimization. In *Handbook of Variational Methods for Nonlinear Geometric Data*, pages 527–554. Springer, 2020.

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. Streaming PCA: Matching matrix bernstein and near-optimal finite sample guarantees for Oja's algorithm. In *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.

Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent only converges to minimizers. In *Conference on Learning Theory*, pages 1246–1257, 2016.

Chris Junchi Li and Michael I Jordan. Stochastic approximation for online tensorial independent component analysis. In *Conference on Learning Theory*, pages 3051–3106. PMLR, 2021.

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

Jorge Nocedal and Stephen J Wright. *Numerical Optimization*. Springer, 2006.

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

Karl Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2:559–572, 1901.

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 Networks*, 12(1):145–151, 1999.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407, 1951.

David Ruppert. Efficient estimations from a slowly convergent Robbins-Monro process. *Technical Report, Cornell University Operations Research and Industrial Engineering*, 1988.

Ohad Shamir. Convergence of stochastic gradient descent for PCA. In *International Conference on Machine Learning*, pages 257–265, 2016.

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, and Tong Zhang. 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.

Ganzhao Yuan, Li Shen, and Wei-Shi Zheng. A decomposition algorithm for the sparse generalized eigenvalue problem. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6113–6122, 2019.

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.

A PROBLEM-DEPENDENT PARAMETERS FOR GEV

We need to verify that the objective function for the GEV problems is indeed in the class of strict-saddle functions. For the GEV problem, the objective function of interest is

$$F(\boldsymbol{v}) = -\frac{\boldsymbol{v}^\top \mathbf{A}\boldsymbol{v}}{\boldsymbol{v}^\top \mathbf{B}\boldsymbol{v}}, \qquad \text{such that } \; c(\boldsymbol{v}) = \|\boldsymbol{v}\|^2 - 1, \tag{33}$$

where $\mathbf{A}$ and $\mathbf{B}$ are two symmetric matrices. We make one additional mild assumption on the eigenstructure of matrices $\mathbf{A}$ and $\mathbf{B}$.

Assumption 5 *The matrix $\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2}$ is diagonalizable with eigenvalues $\lambda_1 > \lambda_2 > \cdots > \lambda_d$. Moreover, $\lambda_{\min}(\mathbf{B}) > 0$.*

As our argument proceeds, one can safely assume $\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2}$ being diagonal without loss of generality, so we will proceed with such. Under Assumption 5 we denote the minimal gap of λ_i's as

$$\lambda_{\text{gap}} = \min_{1\le i\le d-1} (\lambda_i - \lambda_{i+1}) > 0. \tag{34}$$

We prove that under the mild Assumption 5, the objective function for generalized eigenvector problem is strict-saddle as in Definition 5 if the parameters are chosen properly:

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

$$F(\boldsymbol{v}) = -\frac{\boldsymbol{v}^\top \mathbf{A}\boldsymbol{v}}{\boldsymbol{v}^\top \mathbf{B}\boldsymbol{v}}, \qquad \text{such that } \; c(\boldsymbol{v}) = \|\boldsymbol{v}\|^2 - 1 = 0, \tag{35}$$

where $\mathbf{A}$ and $\mathbf{B}$ are two real symmetric matrices with $\mathbf{B}$ being strictly positive-definite. In the following lemma, we verify that the objective function $F(\boldsymbol{v})$ in (35) satisfies Assumption 1; that is, $D(\boldsymbol{v}), F(\boldsymbol{v}), \nabla F(\boldsymbol{v}), \nabla^2 F(\boldsymbol{v})$ are Lipschitz continuous within $\{\boldsymbol{v} : \|\boldsymbol{v}\| \le 1, \|\boldsymbol{v} - \boldsymbol{v}^*\| \le \delta\}$.

Proposition 10 *Assumption 1 holds for $F(\boldsymbol{v})$ in GEV problem* (35) *with constants*

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

The proof of Proposition 10 is deferred to §B.3. 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^*.

Proposition 11 *Under Assumption 5, the only local minimizers of* (35) *are $\pm\mathbf{e}_1$, and the function satisfies the $(\mu, \beta, \gamma, \delta)$-strict saddle condition for*

$$\begin{aligned} \mu &= (\lambda_1 - \lambda_2)\frac{\lambda_{\min}(\mathbf{B})}{\|\mathbf{B}\|}, \qquad \beta = (\lambda_1 - \lambda_2)\frac{\lambda_{\min}(\mathbf{B})}{\|\mathbf{B}\|}, \\ \gamma &= \lambda_{gap}^3 \frac{\lambda_{\min}^8(\mathbf{B})}{(8)84^2\|\mathbf{A}\|^2\|\mathbf{B}\|^6}, \qquad \delta = (\lambda_1 - \lambda_2)\frac{\lambda_{\min}^4(\mathbf{B})}{168\|\mathbf{A}\|\|\mathbf{B}\|^3}. \end{aligned} \tag{36}$$

To verify the strict-saddle parameters and conclude Proposition 11, we first conclude the parameters for the objective function of the eigenvector problem:

Lemma 12 *Under Assumption 5, and with the choices of parameters as in* (36)*, we have the following:*

(i) Suppose $\|g(\mathbf{x})\| \le \gamma$ *and* $|\mathbf{e}_1^\top \mathbf{B}^{1/2}\mathbf{x}| \le (1/2)\|\mathbf{B}^{1/2}\mathbf{x}\|$. *Let the vector*

$$\boldsymbol{v} \equiv \frac{P_{\mathcal{T}(\mathbf{x})}\mathbf{B}^{-1/2}\mathbf{e}_1}{\|P_{\mathcal{T}(\mathbf{x})}\mathbf{B}^{-1/2}\mathbf{e}_1\|},$$

then $\boldsymbol{v} \in \mathcal{T}(\mathbf{x})$, $\|\boldsymbol{v}\| = 1$, *and we have*

$$\boldsymbol{v}^\top \mathcal{H}(\mathbf{x})\boldsymbol{v} \le -\beta. \tag{37}$$

(ii) Suppose $\|g(\mathbf{x})\| \le \gamma$ *and* $|\mathbf{e}_1^\top \mathbf{B}^{1/2}\mathbf{x}| > (1/2)\|\mathbf{B}^{1/2}\mathbf{x}\|$. *Then there is a local minimizer* $\mathbf{x}^*$ *such that* $\|\mathbf{x} - \mathbf{x}^*\| \le \delta$, *and for all* $\mathbf{x}' \in \mathbf{B}_{2\delta}(\mathbf{x}^*)$ *we have for all* $\hat{\boldsymbol{v}} \in \mathcal{T}(\mathbf{x}')$ *and* $\|\hat{\boldsymbol{v}}\| = 1$

$$\hat{\boldsymbol{v}}^\top \mathcal{H}(\mathbf{x}')\hat{\boldsymbol{v}} \ge \mu. \tag{38}$$

It is straightforward from Definition 5 of strict-saddle property that Lemma 12 leads to Proposition 11 immediately. We postpone the details to §D. Intuitively, the parameters are only dependent on the differences of the consecutive eigenvalues $\lambda_1 - \lambda_2, \ldots, \lambda_{d-1} - \lambda_d$, since we can always add each eigenvalue λ_i by an arbitrary constant and keep the constrained optimization problem (35) unchanged. We also remark that restricted to our analysis, the parameters in (36) might not be the sharpest possible choices. However, we do provide, to the best of our knowledge, a first identification of strict-saddle parameters for the GEV problem, and hence Theorems 6 and 7 apply.

B PROOFS

In this section, we provide detailed proofs of our main results.

B.1 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 $\boldsymbol{v}^* \in \mathbf{B}_\delta(\boldsymbol{v}_0)$ of $F(\boldsymbol{v})$ such that $\boldsymbol{u}^\top \mathcal{H}(\boldsymbol{v}^*)\boldsymbol{u} \ge \mu\|\boldsymbol{u}\|^2$ for all $\boldsymbol{u} \in \mathcal{T}(\boldsymbol{v}^*)$.

For a positive quantity M to be determined later, let

$$\mathcal{T}_M = \inf\{t \ge 1 : \|\Gamma(\boldsymbol{v}_{t-1}; \boldsymbol{\zeta}_t)\| > M\}. \tag{39}$$

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

$$M = \mathcal{V}\log^{1/\alpha}\epsilon^{-1}. \tag{40}$$

Then, we have

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

The proof of Lemma 13 is a straightforward corollary of a union bound and Assumption 2, and is provided in §E.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

$$g(\boldsymbol{v}) = (\mathbf{I} - \boldsymbol{v}\boldsymbol{v}^\top)\nabla F(\boldsymbol{v}) \quad \text{and} \quad \mathcal{H}(\boldsymbol{v}) = \nabla^2 F(\boldsymbol{v}) - (\boldsymbol{v}^\top \nabla F(\boldsymbol{v}))\mathbf{I}. \tag{41}$$

Taking derivatives, we decompose

$$\nabla g(\boldsymbol{v}) = \mathcal{H}(\boldsymbol{v}) + \mathcal{N}(\boldsymbol{v}), \tag{42}$$

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

$$\mathcal{N}(\boldsymbol{v}) = -\boldsymbol{v}(\nabla F(\boldsymbol{v}) + \nabla^2 F(\boldsymbol{v})\boldsymbol{v})^\top. \tag{43}$$

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

Lemma 14 *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 14 is deferred to §E.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 15 *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*

$$\boldsymbol{v}_t - \boldsymbol{v}^* = (\mathbf{I} - \eta D\mathcal{H}_* - \eta D\mathcal{N}_*)(\boldsymbol{v}_{t-1} - \boldsymbol{v}^*) + \eta\boldsymbol{\xi}_t + \eta\boldsymbol{R}_t + \eta^2\boldsymbol{Q}_t, \tag{44}$$

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 $\|\boldsymbol{R}_t\| \leq (DL_H + DL_N + L_DL_G)\|\boldsymbol{v}_{t-1} - \boldsymbol{v}^*\|^2$ and $\boldsymbol{Q}_t$ satisfies $\|\boldsymbol{Q}_t\| \leq 7M^2$.

The proof of Lemma 15 is deferred to §E.3. We define the projection of $\boldsymbol{v}_t - \boldsymbol{v}^*$ on $\mathcal{T}(\boldsymbol{v}^*)$ as

$$\boldsymbol{\Delta}_t = (\mathbf{I} - \boldsymbol{v}^*\boldsymbol{v}^{*\top})(\boldsymbol{v}_t - \boldsymbol{v}^*), \tag{45}$$

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

$$\mathcal{M}_* = (\mathbf{I} - \boldsymbol{v}^*\boldsymbol{v}^{*\top})\mathcal{H}_*(\mathbf{I} - \boldsymbol{v}^*\boldsymbol{v}^{*\top}). \tag{46}$$

Lemma 16 *Under initialization condition (14), the following properties hold:*

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

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

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

$$\|\boldsymbol{\Delta}_t\|^2 \leq \|\boldsymbol{v}_t - \boldsymbol{v}^*\|^2 \leq 2\|\boldsymbol{\Delta}_t\|^2. \tag{47}$$

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

$$\|(\mathbf{I} - \eta D\mathcal{M}_*)^t\boldsymbol{\Delta}_0\| \leq (1 - \eta D\mu)^t\|\boldsymbol{\Delta}_0\|. \tag{48}$$

The proof of Lemma 16 is deferred to §E.4. To interpret Lemma 16(i), we denote $\theta \equiv \angle(\boldsymbol{v}_t, \boldsymbol{v}^*) \in [0, \pi/2]$, such that $\|\boldsymbol{v}_t - \boldsymbol{v}^*\| = 2\sin(\theta/2)$, $\boldsymbol{\Delta}_t = (\mathbf{I} - \boldsymbol{v}^*\boldsymbol{v}^*)^\top(\boldsymbol{v}_t - \boldsymbol{v}^*) = \sin\theta$, and (47) is equivalent to the trigonometric inequality

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

By combining Lemmas 15 and 16, we have the following lemma for the update rule in terms of $\boldsymbol{\Delta}_t$:

Lemma 17 *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 $\boldsymbol{\Delta}_t$ as*

$$\boldsymbol{\Delta}_t = (\mathbf{I} - \eta D\mathcal{M}_*)\boldsymbol{\Delta}_{t-1} + \eta\boldsymbol{\chi}_t + \eta\boldsymbol{S}_t + \eta^2\boldsymbol{P}_t; \tag{49}$$

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 $\boldsymbol{P}_t$ satisfies $\|\boldsymbol{P}_t\| \leq 7M^2$.

Proof of Lemma 17 is deferred to §E.5. Here we have $\rho = D(L_H + L_N + B_H/2) + L_DL_G$, which is consistent with its definition in (13).

Now, to analyze the iteration $\mathbf{\Delta}_t$ we need to control its tail behavior. We define the truncated version

$$\widetilde{\boldsymbol{S}}_t = \boldsymbol{S}_t 1_{(\mathcal{T}_M > t)}, \qquad \widetilde{\boldsymbol{P}}_t = \boldsymbol{P}_t 1_{(\mathcal{T}_M > t)}, \tag{50}$$

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

$$\overline{\mathbf{\Delta}}_t = (\mathbf{I} - \eta D \mathcal{M}_*)\, \overline{\mathbf{\Delta}}_{t-1} + \eta \boldsymbol{\chi}_t + \eta \widetilde{\boldsymbol{S}}_t + \eta^2 \widetilde{\boldsymbol{P}}_t. \tag{51}$$

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

Lemma 18 *For each $t \geq 0$ we have $\overline{\mathbf{\Delta}}_t = \mathbf{\Delta}_t$ on the event $(\mathcal{T}_M > t)$. Furthermore, we have for all $t \geq 1$*

$$\overline{\mathbf{\Delta}}_t = (\mathbf{I} - \eta D \mathcal{M}_*)^t\, \mathbf{\Delta}_0 + \eta \sum_{s=1}^{t} (\mathbf{I} - \eta D \mathcal{M}_*)^{t-s}\, \boldsymbol{\chi}_s + \eta \sum_{s=1}^{t} (\mathbf{I} - \eta D \mathcal{M}_*)^{t-s}\, \widetilde{\boldsymbol{S}}_s + \eta^2 \sum_{s=1}^{t} (\mathbf{I} - \eta D \mathcal{M}_*)^{t-s}\, \widetilde{\boldsymbol{P}}_s. \tag{52}$$

We defer the proof of Lemma 18 in §E.6.

Next we provide a lemma that tightly characterizes the approximations in (52) that $\overline{\mathbf{\Delta}}_t \approx (\mathbf{I} - \eta D \mathcal{M}_*)^t \mathbf{\Delta}_0$.

Lemma 19 *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 $r > 0$

$$\left\|\overline{\mathbf{\Delta}}_t - (\mathbf{I} - \eta D \mathcal{M}_*)^t \mathbf{\Delta}_0\right\| \leq \frac{8 G_\alpha \mathcal{V}}{\sqrt{D\mu}} \log^{\frac{\alpha+2}{2\alpha}} \epsilon^{-1} \cdot \eta^{1/2} + \frac{\rho r^2}{D\mu} + \frac{7\mathcal{V}^2}{D\mu} \log^{\frac{2}{\alpha}} \epsilon^{-1} \cdot \eta. \tag{53}$$

The proof of Lemma 19 is provided in §E.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 $\mathbf{\Delta}_0$ and $\widetilde{\Theta}(\eta^{1/2})$, the conditioning event occurs almost surely on a high-probability event.

Lemma 20 *When initialization*

$$\|\mathbf{\Delta}_0\| \leq \left\{\frac{D\mu}{2^5 G_\alpha \rho}, \delta\right\},$$

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

$$\|\mathbf{\Delta}_t\| \leq 2 \max\left\{\|\mathbf{\Delta}_0\|,\ \frac{2^7 G_\alpha \mathcal{V}}{\sqrt{D\mu}} \log^{\frac{\alpha+2}{2\alpha}} \epsilon^{-1} \cdot \eta^{1/2}\right\},$$

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

$$\|\mathbf{\Delta}_{T_\eta^*}\| \leq \frac{1}{2} \max\left\{\|\mathbf{\Delta}_0\|,\ \frac{2^7 G_\alpha \mathcal{V}}{\sqrt{D\mu}} \log^{\frac{\alpha+2}{2\alpha}} \epsilon^{-1} \cdot \eta^{1/2}\right\}.$$

Lemma 20, whose proof is given in §E.8, implies that the iteration keeps $\|\mathbf{\Delta}_t\| \leq 2\|\mathbf{\Delta}_0\|$ unless $\boldsymbol{v}$ is within a noisy neighborhood of the local minimizer $\boldsymbol{v}^*$, where we recall the definition of $\mathbf{\Delta}_t$ in (45).

Finally, Proposition 3 is proved by combining Lemmas 16 and 20.

B.2 PROOF OF THEOREMS 6 AND 7

In this subsection, we aim to prove Theorem 6. To deal with points with strong gradient corresponding to (i) in Definition 5, we use the following lemma that is adapted from Ge et al. [2015, Lemma 38].

Proposition 21 *Assume all conditions in Theorem 6 as well as* $\sqrt{2d\mathcal{V}^2 L_G D_+ \eta} < \beta$*, we have on the event* $\left(\|\nabla F(\boldsymbol{v}_t)\| \geq \sqrt{2d\mathcal{V}^2 L_G D_+ \eta}\right)$ *that*

$$\mathbb{E}\left[F(\boldsymbol{v}_{t+1}) - F(\boldsymbol{v}_t) \mid \mathcal{F}_t\right] \leq -0.5 d\sigma^2 L_G D_-^2 \eta^2. \tag{54}$$

A core problem involves escaping from saddle points that corresponds to (iii) in Definition 5, we conclude the following modification from Ge et al. [2015, Lemma 40].

Proposition 22 *Assume all conditions in Theorem 6 as well as* $\sqrt{2\eta\sigma^2 L_G d D_+} < \beta$*. Then on the event*

$$\left\{\|\nabla F(\boldsymbol{v}_0)\| < \sqrt{2d\mathcal{V}^2 L_G D_+ \eta},\ \lambda_{\min}(\mathcal{H}(\boldsymbol{v}_0)) \leq -\gamma\right\},$$

there is a stopping time $\mathcal{T}(\boldsymbol{v}_0) \leq T_{\max}$ *almost surely such that*

$$\mathbb{E}F(\boldsymbol{v}_{\mathcal{T}(\boldsymbol{v}_0)}) - F(\boldsymbol{v}_0) \leq -0.5\sigma^2 D_- \eta, \tag{55}$$

where $T_{\max}$ *is fixed and independent of* $\boldsymbol{v}_0$ *defined as*

$$T_{\max} = 0.5\gamma^{-1} D_-^{-1} \eta^{-1} \log\left(\frac{6d\mathcal{V}}{\sigma}\right).$$

Proofs of Propositions 21 and 22 are straightforward generalization of relevant proofs of [Ge et al., 2015], and hence we omit the details.

Proof [Proof of Theorem 6] While this proof can be done in a similar fashion as Theorem 36 in Ge et al. [2015], here we provide a different proof using stopping-time techniques.

(i) Given (24), we split the state space $\mathcal{S}^{d-1}$ into three distinct regions: let

$$\mathcal{Q}_1 = \left\{\boldsymbol{v} \in \mathcal{S}^{d-1} : \|\nabla F(\boldsymbol{v})\| \geq \sqrt{2d\mathcal{V}^2 L_G D_+ \eta}\right\},$$

and let

$$\mathcal{Q}_2 = \left\{\boldsymbol{v} \in \mathcal{S}^{d-1} : \|\nabla F(\boldsymbol{v})\| < \sqrt{2d\mathcal{V}^2 L_G D_+ \eta},\ \lambda_{\min}(\mathcal{H}(\boldsymbol{v})) \leq -\gamma\right\}.$$

Define a stochastic process $\{\mathcal{T}_i\}$ s.t. $\mathcal{T}_0 = 0$, and

$$\mathcal{T}_{i+1} = \mathcal{T}_i + 1_{\mathcal{Q}_1}(\boldsymbol{v}_{\mathcal{T}_i}) + \mathcal{T}(\boldsymbol{v}_{\mathcal{T}_i}) 1_{\mathcal{Q}_2}(\boldsymbol{v}_{\mathcal{T}_i}), \tag{56}$$

where $\mathcal{T}(\boldsymbol{v}_{\mathcal{T}_i}) \leq T_{\max}$ is defined in Proposition 22. By (54) in Lemma 21 and (55) in Proposition 22, we know that on $(\boldsymbol{v}_{\mathcal{T}_i} \in \mathcal{Q}_1)$

$$\mathbb{E}[F(\boldsymbol{v}_{\mathcal{T}_{i+1}}) - F(\boldsymbol{v}_{\mathcal{T}_i}) \mid \mathcal{F}_{\mathcal{T}_i}] \leq -0.5 d\sigma^2 L_G D_-^2 \eta^2,$$

and on $(\boldsymbol{v}_{\mathcal{T}_i} \in \mathcal{Q}_2)$

$$\mathbb{E}[F(\boldsymbol{v}_{\mathcal{T}_{i+1}}) - F(\boldsymbol{v}_{\mathcal{T}_i}) \mid \mathcal{F}_{\mathcal{T}_i}] \leq -0.5\sigma^2 D_- \eta.$$

Combining the above two displays and (56), we have

$$\begin{aligned}
&\mathbb{E}[F(\boldsymbol{v}_{\mathcal{T}_{i+1}}) - F(\boldsymbol{v}_{\mathcal{T}_i}) \mid \mathcal{F}_{\mathcal{T}_i}] \\
&\leq -\min\left(0.5 d\sigma^2 L_G D_-^2 \eta^2, \frac{0.5\sigma^2 D_- \eta}{0.5\gamma^{-1} D_-^{-1} \eta^{-1} \log\left(\frac{6d\mathcal{V}}{\sigma}\right)}\right) \cdot \mathbb{E}\left[\mathcal{T}_{i+1} - \mathcal{T}_i \mid \mathcal{F}_{\mathcal{T}_i}\right] \\
&\leq -\min\left(0.5 d L_G, \gamma \log^{-1}\left(\frac{6d\mathcal{V}}{\sigma}\right)\right) \sigma^2 D_-^2 \eta^2 \cdot \mathbb{E}\left[\mathcal{T}_{i+1} - \mathcal{T}_i \mid \mathcal{F}_{\mathcal{T}_i}\right],
\end{aligned} \tag{57}$$

on $\{\boldsymbol{v}_{\mathcal{T}_i} \in \mathcal{Q}_1 \cup \mathcal{Q}_2\}$.

(ii) Let $\mathcal{I} \in [0, \infty]$ be the (random) first index i such that $\boldsymbol{v}_{\mathcal{T}_i} \in (\mathcal{Q}_1 \cup \mathcal{Q}_2)^c$. We conclude immediately that $(\mathcal{I} > i) \in \mathcal{F}_{\mathcal{T}_i}$, and $(\mathcal{I} > i) \subseteq (\boldsymbol{v}_{\mathcal{T}_i} \in \mathcal{Q}_1 \cup \mathcal{Q}_2)$. Applying (57) gives

$$\begin{aligned}
\mathbb{E}\left[F(\boldsymbol{v}_{\mathcal{T}_\mathcal{I}}) - F(\boldsymbol{v}_0)\right] &= \mathbb{E}\left[\sum_{i=0}^{\infty}\left(F(\boldsymbol{v}_{\mathcal{T}_{i+1}}) - F(\boldsymbol{v}_{\mathcal{T}_i})\right) 1_{\mathcal{I}>i}\right] \\
&\leq -\min\left(0.5dL_G, \gamma\log^{-1}\left(\frac{6d\mathcal{V}}{\sigma}\right)\right)\sigma^2 D_-^2 \eta^2 \cdot \mathbb{E}\mathcal{T}_\mathcal{I} \\
&\leq -\min\left(0.5dL_G, \gamma\log^{-1}\left(\frac{6d\mathcal{V}}{\sigma}\right)\right)\sigma^2 D_-^2 \eta^2 \cdot T \cdot \mathbb{P}\left(\mathcal{T}_\mathcal{I} \geq T\right),
\end{aligned}$$

where $T \geq 0$ is any constant. Plugging in $T = T_1$ as in (23) gives

$$\mathbb{P}\left(\mathcal{T}_\mathcal{I} \geq T_1\right) \leq \frac{\mathbb{E}\left[F(\boldsymbol{v}_0) - F(\boldsymbol{v}_{\mathcal{T}_\mathcal{I}})\right]}{\min\left(0.5dL_G, \gamma\log^{-1}\left(\frac{6d\mathcal{V}}{\sigma}\right)\right)\sigma^2 D_-^2 \eta^2 \cdot T_1} \leq \frac{2\|F\|_\infty}{4\|F\|_\infty} = \frac{1}{2}.$$

In words, event $(\mathcal{T}_\mathcal{I} < T_1)$ has at least $1/2$ probability, on which the iteration $\boldsymbol{v}_t$ must enter $(\mathcal{Q}_1 \cup \mathcal{Q}_2)^c$ by time T_1 at least once.

(iii) Noting that the argument above holds for all initial points $\boldsymbol{v}_0 \in \mathcal{Q}_1 \cup \mathcal{Q}_2$, so one can use Markov property and conclude that within $T_1 \cdot \lceil \log_2(\kappa^{-1}) \rceil$ steps where T_1 was defined in (23), iteration $\{\boldsymbol{v}_t\}$ must enter $(\mathcal{Q}_1 \cup \mathcal{Q}_2)^c$ at least once with probability at least $1 - \kappa$. The rest of our proof follows from the definition of strict-saddle function.

■

Proof [Proof of Theorem 7] The conclusion is reached by directly combining Theorems 2 and 6, setting $\mathcal{A}_T = \mathcal{H}_2$, along with an application of strong Markov property. ■

B.3 PROOF OF PROPOSITION 10

Proof [Proof of Proposition 10] For the 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 $\boldsymbol{v} \in \{\boldsymbol{v} : \|\boldsymbol{v}\| \leq 1, \|\boldsymbol{v} - \boldsymbol{v}^*\| \leq \delta\}$,

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

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

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

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 $\boldsymbol{v} \in \{\boldsymbol{v} : \|\boldsymbol{v}\| \leq 1, \|\boldsymbol{v}-\boldsymbol{v}^*\| \leq \delta\}$,

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

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 $L_K \equiv \frac{28\|\mathbf{A}\|\|\mathbf{B}\|^2}{(1-\delta)^3\lambda_{\min}^3(\mathbf{B})}$. ■

B.4 PROOF OF THEOREM 8

To prove Theorem 8, we first present the following Lemma 23 on a linear representation of $\mathcal{M}_*(\overline{\boldsymbol{v}}_T^{(\eta)} - \boldsymbol{v}^*)$.

Lemma 23 (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

$$\begin{aligned}\mathcal{M}_*\left(\overline{\boldsymbol{v}}_T^{(\eta)} - \boldsymbol{v}^*\right) = &\frac{1}{D(T-K_{\eta,\epsilon}T_\eta^*)}\sum_{t=K_{\eta,\epsilon}T_\eta^*+1}^{T}\boldsymbol{\chi}_{t+1} + \frac{1}{D(T-K_{\eta,\epsilon}T_\eta^*)}\sum_{t=K_{\eta,\epsilon}T_\eta^*+1}^{T}\boldsymbol{S}_{t+1} \\ &+ \frac{\eta}{D(T-K_{\eta,\epsilon}T_\eta^*)}\sum_{t=K_{\eta,\epsilon}T_\eta^*+1}^{T}\boldsymbol{P}_{t+1} + \frac{1}{D(T-K_{\eta,\epsilon}T_\eta^*)\eta}(\boldsymbol{\Delta}_{K_{\eta,\epsilon}T_\eta^*+1} - \boldsymbol{\Delta}_{T+1}),\end{aligned} \tag{58}$$

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

$$\boldsymbol{\chi}_t \equiv (\mathbf{I} - \boldsymbol{v}^*\boldsymbol{v}^{*\top})(\Gamma(\boldsymbol{v}_{t-1};\boldsymbol{\zeta}_t) - D(\boldsymbol{v}_{t-1})\nabla F(\boldsymbol{v}_{t-1})), \tag{59}$$

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 $\|\boldsymbol{P}_t\| \leq 7\mathcal{V}^2\log^{2/\alpha}\epsilon^{-1}$.

Proof [Proof of Lemma 23] Telescoping (49) in Lemma 17 for $t = K_{\eta,\epsilon}T_\eta^* + 2, \ldots, T+1$ gives

$$\begin{aligned}\eta D\mathcal{M}_*\sum_{t=K_{\eta,\epsilon}T_\eta^*+1}^{T}\boldsymbol{\Delta}_t = &(\boldsymbol{\Delta}_{K_{\eta,\epsilon}T_\eta^*+1} - \boldsymbol{\Delta}_{T+1}) + \eta\sum_{t=K_{\eta,\epsilon}T_\eta^*+1}^{T}\boldsymbol{\chi}_{t+1} \\ &+ \eta\sum_{t=K_{\eta,\epsilon}T_\eta^*+1}^{T}\boldsymbol{S}_{t+1} + \eta^2\sum_{t=K_{\eta,\epsilon}T_\eta^*+1}^{T}\boldsymbol{P}_{t+1}.\end{aligned}$$

Plugging in the definitions of $\boldsymbol{\Delta}_t, \overline{\boldsymbol{v}}_T^{(\eta)}$ in (45), (27) gives (58). For event $\mathcal{H}_2$ defined in Theorem 2 using total sample size $T+1$, the proof of Lemma 20 in §E.8 shows that $\mathcal{H}_2 \subseteq \left\{\|\Gamma(\boldsymbol{v}_{t-1};\boldsymbol{\zeta}_t)\| \le M : 1 \le t \le T+2\right\}$. The rest of Lemma 23 directly follows Lemma 17. ■

With Lemma 23 in hand, we are ready to prove Theorem 8.

Proof [Proof of Theorem 8] For a given T, we apply Theorem 2 and Lemma 23 with $\epsilon = 1/T^2$, such that $\mathbb{P}(\mathcal{H}_2) \to 1$ and the scaling condition (17) is satisfied under condition (29). Using a coupling approach we can safely ignore the small probability event and concentrate on the event $\mathcal{H}_2$, where we have

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

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

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

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

Under condition (29), as $T \to \infty, \eta \to 0$, we have the following almost-sure convergences

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

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

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

From (12) and (59), 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 (29) as $T \to \infty, \eta \to 0$:

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

Combining these results with (58) in Lemma 23, under condition (29), as $T \to \infty, \eta \to 0$ we have convergence in distribution:

$$\sqrt{T}\mathcal{M}_*\left(\overline{\boldsymbol{v}}_T^{(\eta)} - \boldsymbol{v}^*\right) \xrightarrow{d} \mathcal{N}(\mathbf{0}, D^{-2}\cdot\boldsymbol{\Phi}_*). \tag{60}$$

Since $\mathcal{M}_*^-\mathcal{M}_* = \mathbf{I} - \boldsymbol{v}^*\boldsymbol{v}^{*\top}$ and $\mathcal{M}_*^-\boldsymbol{\Phi}_*\mathcal{M}_*^- = \mathcal{M}_*^-\boldsymbol{\Sigma}_*\mathcal{M}_*^-$, (60) is equivalent to

$$\sqrt{T}(\mathbf{I} - \boldsymbol{v}^*\boldsymbol{v}^{*\top})\left(\overline{\boldsymbol{v}}_T^{(\eta)} - \boldsymbol{v}^*\right) \xrightarrow{d} N\left(\mathbf{0}, D^{-2}\cdot\mathcal{M}_*^-\boldsymbol{\Sigma}_*\mathcal{M}_*^-\right), \tag{61}$$

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 in Lemma 16 in §B.1 we know 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:

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

where in the second line we used the first condition in (29). Under condition (29), as $T\to\infty, \eta\to 0$, we have almost-sure convergence

$$\sqrt{T}\cdot\boldsymbol{v}^*\boldsymbol{v}^{*\top}\left(\overline{\boldsymbol{v}}_T^{(\eta)}-\boldsymbol{v}^*\right)\to\mathbf{0}\quad\text{a.s.} \tag{62}$$

Adding up (61) and (62) and applying the Slutsky theorem, we conclude (30) and Theorem 8. ■

B.5 PROOF OF PROPOSITION 9

Proof [Proof of Proposition 9] 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 26 we have

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

which indicates that

$$\left\|\boldsymbol{v}_x^\top\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{v}_x\right\|_{\psi_1}\leq\mathcal{V}_x^2,\qquad\left\|\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{v}_x\right\|_{\psi_1}\leq\mathcal{V}_x^2.$$

Similarly, we can show

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

and

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

Combining all above inequalities and using Lemma 25 yields

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

By applying Lemmas 26 and 27, 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 $\alpha=1/2$. ■

C PRELIMINARIES FOR ORLICZ-ψ_α NORM

Of similar style as [Li and Jordan, 2021, §E] we collect in this section some facts for Orlicz-ψ_α norm for our usage. We start with its definition:

Definition 24 (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*

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

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

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

Lemma 25 *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*

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

For all $\alpha \ge 1$, the above inequality holds when $\|\cdot\|_\psi$ is taken as the Orlicz ψ_α-norm.

Proof [Proof of Lemma 25] 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

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

yielding the lemma. ■

Lemma 26 *Let X and Y be random variables with finite ψ_α-norm for some $\alpha \ge 1$, then*

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

Proof [Proof of Lemma 26] Denote $A \equiv X/\|X\|_{\psi_\alpha}$, $B \equiv Y/\|Y\|_{\psi_\alpha}$, then $\|A\|_{\psi_\alpha} = \|B\|_{\psi_\alpha} = 1$. Using the elementary inequality

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

and the triangle inequality in Lemma 25 we have that

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

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

Lemma 27 *For any random variables X, Y with finite Orlicz ψ_α-norm, the following inequalities hold*

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

and

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

Proof [Proof of Lemma 27] 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

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

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 $x_*^\alpha \le 1/\alpha$. 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 $\psi_\alpha(x) \ge \widetilde{\psi}_\alpha(x) \ge 0$ for all $x \ge 0$, and hence

$$0 \le \psi_\alpha(x) - \widetilde{\psi}_\alpha(x) \le \psi_\alpha(x_*) \le e^{1/\alpha} - 1. \tag{63}$$

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

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

By applying the triangle inequality in Lemma 25 and using (63), we have

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

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

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

and

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

■

Now we proceed with the definition of Orlicz ψ_α-norm for random vectors.

Definition 28 *For a random vector $\boldsymbol{X} \in \mathbb{R}^d$, its Orlicz ψ_α-norm is defined as*

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

Seeing the above definition, a random vector $\boldsymbol{X}$ is called *sub-Gaussian* if $\|\boldsymbol{X}\|_{\psi_2} < \infty$, and is called *sub-Exponential* if $\|\boldsymbol{X}\|_{\psi_1} < \infty$.

Remark 29 *We notice that $\|\boldsymbol{X}\|_{\psi_\alpha}$ equals to the Orlicz ψ_α-norm of random variable (scalar) $\|\boldsymbol{X}\|$. 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.*

D ESTIMATION OF THE STRICT-SADDLE PARAMETERS

The goal of this section is to detail the proof of Lemma 12 that estimates the strict-saddle parameters. We first compute the manifold gradient and Hessian in the following Lemma 30:

Lemma 30 *The manifold gradient and Hessian can be computed as*

$$g(\mathbf{x}) = -2\frac{(\mathbf{x}^\top\mathbf{B}\mathbf{x})\mathbf{A} - (\mathbf{x}^\top\mathbf{A}\mathbf{x})\mathbf{B}}{(\mathbf{x}^\top\mathbf{B}\mathbf{x})^2}\mathbf{x}, \tag{64}$$

$$\mathcal{H}(\mathbf{x}) = -2\frac{(\mathbf{x}^\top\mathbf{B}\mathbf{x})\mathbf{A} - (\mathbf{x}^\top\mathbf{A}\mathbf{x})\mathbf{B} + 2(\mathbf{A}\mathbf{x}\mathbf{x}^\top\mathbf{B} - \mathbf{B}\mathbf{x}\mathbf{x}^\top\mathbf{A})}{(\mathbf{x}^\top\mathbf{B}\mathbf{x})^2} + 8\frac{\left[(\mathbf{x}^\top\mathbf{B}\mathbf{x})\mathbf{A} - (\mathbf{x}^\top\mathbf{A}\mathbf{x})\mathbf{B}\right]\mathbf{x}\mathbf{x}^\top\mathbf{B}}{(\mathbf{x}^\top\mathbf{B}\mathbf{x})^3}. \tag{65}$$

Proof The constrained optimization problem has $c(\mathbf{x}) = \|\mathbf{x}\|^2 - 1$ so the Lagrangian is

$$\mathcal{L}(\mathbf{x};\mu) = -\frac{\mathbf{x}^\top\mathbf{A}\mathbf{x}}{\mathbf{x}^\top\mathbf{B}\mathbf{x}} - \mu(\mathbf{x}^\top\mathbf{x} - 1).$$

According to the constrained optimization theory in Nocedal and Wright [2006], since (i) there is one constraint (ii) the gradient $g(\mathbf{x}) = 2\mathbf{x}$ on constraint has constant norm 2, it satisfies some 2-RLICQ condition. The feasible value of Lagrangian multiplier $\mu^*(\mathbf{x})$ has

$$\mu^*(\mathbf{x}) = \arg\min_{\mu} \|\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x},\mu)\|^2.$$

Let

$$\Lambda(\mathbf{x}) = \frac{(\mathbf{x}^\top\mathbf{B}\mathbf{x})\mathbf{A} - (\mathbf{x}^\top\mathbf{A}\mathbf{x})\mathbf{B}}{(\mathbf{x}^\top\mathbf{B}\mathbf{x})^2}.$$

Then we have

$$\nabla\mathcal{L}(\mathbf{x};\mu) = -2\Lambda(\mathbf{x})\mathbf{x} - 2\mu\mathbf{x},$$

and hence

$$\begin{aligned}\|\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x};\mu)\|^2 = 4\,\|\Lambda(\mathbf{x})\mathbf{x} + \mu\mathbf{x}\|^2 &= 4\|\Lambda(\mathbf{x})\mathbf{x} + \mu\mathbf{x}\|^2 \\ &= 4\left(\mathbf{x}^\top\Lambda(\mathbf{x})\Lambda(\mathbf{x})\mathbf{x} + 2(\mathbf{x}^\top\Lambda(\mathbf{x})\mathbf{x})\mu + (\mathbf{x}^\top\mathbf{x})\mu^2\right).\end{aligned}$$

Solving this problem gives $\mu^*(\mathbf{x})$ for $\mathbf{x} \in \mathcal{S}^{d-1}$:

$$\mu^*(\mathbf{x}) = -\mathbf{x}^\top\Lambda(\mathbf{x})\mathbf{x} = -\frac{(\mathbf{x}^\top\mathbf{B}\mathbf{x})\mathbf{x}^\top\mathbf{A}\mathbf{x} - (\mathbf{x}^\top\mathbf{A}\mathbf{x})\mathbf{x}^\top\mathbf{B}\mathbf{x}}{(\mathbf{x}^\top\mathbf{B}\mathbf{x})^2} = 0.$$

The manifold gradient can hence be computed as

$$g(\mathbf{x}) = \nabla L(\mathbf{x};\mu)\big|_{\mu=\mu^*(\mathbf{x})} = -2\Lambda(\mathbf{x})\mathbf{x} - 2\mu^*(\mathbf{x})\mathbf{x} = -2\frac{(\mathbf{x}^\top\mathbf{B}\mathbf{x})\mathbf{A} - (\mathbf{x}^\top\mathbf{A}\mathbf{x})\mathbf{B}}{(\mathbf{x}^\top\mathbf{B}\mathbf{x})^2}\mathbf{x},$$

concluding (64). For manifold Hessian, we can compute it as

$$\begin{aligned}\mathcal{H}(\mathbf{x}) = \nabla^2 L(\mathbf{x};\mu)\big|_{\mu=\mu^*(\mathbf{x})} &= -2\nabla\left[\frac{(\mathbf{x}^\top\mathbf{B}\mathbf{x})\mathbf{A} - (\mathbf{x}^\top\mathbf{A}\mathbf{x})\mathbf{B}}{(\mathbf{x}^\top\mathbf{B}\mathbf{x})^2}\mathbf{x}\right] \\ &= -2\frac{(\mathbf{x}^\top\mathbf{B}\mathbf{x})\mathbf{A} - (\mathbf{x}^\top\mathbf{A}\mathbf{x})\mathbf{B} + 2(\mathbf{A}\mathbf{x}\mathbf{x}^\top\mathbf{B} - \mathbf{B}\mathbf{x}\mathbf{x}^\top\mathbf{A})}{(\mathbf{x}^\top\mathbf{B}\mathbf{x})^2} + 4\frac{\left[(\mathbf{x}^\top\mathbf{B}\mathbf{x})\mathbf{A} - (\mathbf{x}^\top\mathbf{A}\mathbf{x})\mathbf{B}\right]\mathbf{x}\mathbf{x}^\top(2\mathbf{B})}{(\mathbf{x}^\top\mathbf{B}\mathbf{x})^3}.\end{aligned}$$

This proves (65) and concludes the lemma.

■

We prove the Hessian smoothness and give the Lipschitz constant for both manifold gradient and Hessian, as in the following lemmas.

Lemma 31 *There are Lipschitz constants*

$$L_G \equiv \frac{28\|\mathbf{A}\|\|\mathbf{B}\|^2}{\lambda_{\min}^3(\mathbf{B})}, \qquad L_H \equiv \frac{56\|\mathbf{A}\|\|\mathbf{B}\|^3}{\lambda_{\min}^4(\mathbf{B})},$$

such that for all $\mathbf{z}, \mathbf{z}_1, \mathbf{z}_2 \in \mathcal{S}^{d-1}$ *we have*

$$\|\mathcal{H}(\mathbf{z})\| \le L_G, \tag{66}$$

and

$$\|\mathcal{H}(\mathbf{z}_1) - \mathcal{H}(\mathbf{z}_2)\| \le L_H \|\mathbf{z}_1 - \mathbf{z}_2\|. \tag{67}$$

In addition, we have from above two

$$\left\| P_{\mathcal{T}(\mathbf{z})}^\top \mathcal{H}(\mathbf{z}) P_{\mathcal{T}(\mathbf{z})} - P_{\mathcal{T}(\mathbf{z}')}^\top \mathcal{H}(\mathbf{z}') P_{\mathcal{T}(\mathbf{z}')} \right\| \le (2L_G + L_H)\|\mathbf{z} - \mathbf{z}'\|. \tag{68}$$

In fact, in this lemma one can replace $\|\mathbf{A}\|$ *by the norm* $\|\mathbf{A} - c\mathbf{B}\|$ *for any constant scalar* c*.*

Proof [Proof of Lemma 31] Note

$$\|g(\mathbf{x})\| \le \frac{\|\mathbf{B}\|}{\lambda_{\min}^2(\mathbf{B})}\|\mathbf{A}\|,$$

and

$$\|\mathcal{H}(\mathbf{x})\| \le 2\frac{2\|\mathbf{B}\|\|\mathbf{A}\| + 4\|\mathbf{A}\|\|\mathbf{B}\|}{\lambda_{\min}^2(\mathbf{B})} + 8\frac{2\|\mathbf{B}\|\|\mathbf{A}\|\|\mathbf{B}\|}{\lambda_{\min}^3(\mathbf{B})} \le \frac{28\|\mathbf{B}\|^2}{\lambda_{\min}^3(\mathbf{B})}\|\mathbf{A}\|,$$

so we conclude (66) from mean-value theorem.

Moreover, for an arbitrary unit vector $\boldsymbol{v}$,

$$\begin{aligned}
&\|\mathcal{H}(\mathbf{x})\boldsymbol{v} - \mathcal{H}(y)\boldsymbol{v}\| \\
&\le 2\left\| \frac{(\mathbf{x}^\top\mathbf{B}\mathbf{x})\mathbf{A} - (\mathbf{x}^\top\mathbf{A}\mathbf{x})\mathbf{B} + 2(\mathbf{A}\mathbf{x}\mathbf{x}^\top\mathbf{B} - \mathbf{B}\mathbf{x}\mathbf{x}^\top\mathbf{A})}{(\mathbf{x}^\top\mathbf{B}\mathbf{x})^2}\boldsymbol{v} - \frac{(y^\top\mathbf{B}y)\mathbf{A} - (y^\top\mathbf{A}y)\mathbf{B} + 2(\mathbf{A}yy^\top\mathbf{B} - \mathbf{B}yy^\top\mathbf{A})}{(y^\top\mathbf{B}y)^2}\boldsymbol{v} \right\| \\
&\quad + 8\left\| \frac{\left[(\mathbf{x}^\top\mathbf{B}\mathbf{x})\mathbf{A} - (\mathbf{x}^\top\mathbf{A}\mathbf{x})\mathbf{B}\right]\mathbf{x}\mathbf{x}^\top\mathbf{B}}{(\mathbf{x}^\top\mathbf{B}\mathbf{x})^3}\boldsymbol{v} - \frac{\left[(y^\top\mathbf{B}y)\mathbf{A} - (y^\top\mathbf{A}y)\mathbf{B}\right]yy^\top\mathbf{B}}{(y^\top\mathbf{B}y)^3}\boldsymbol{v} \right\| \equiv \mathrm{I} + \mathrm{II}.
\end{aligned}$$

Note

$$\begin{aligned}
&\nabla\left[\frac{(\mathbf{x}^\top\mathbf{B}\mathbf{x})\mathbf{A} - (\mathbf{x}^\top\mathbf{A}\mathbf{x})\mathbf{B} + 2(\mathbf{A}\mathbf{x}\mathbf{x}^\top\mathbf{B} - \mathbf{B}\mathbf{x}\mathbf{x}^\top\mathbf{A})}{(\mathbf{x}^\top\mathbf{B}\mathbf{x})^2}\boldsymbol{v}\right] \\
&= \frac{2(\mathbf{x}^\top\mathbf{B})\mathbf{A}\boldsymbol{v} - 2(\mathbf{x}^\top\mathbf{A})\mathbf{B}\boldsymbol{v} + 2((\mathbf{x}^\top\mathbf{B}\mathbf{x}\boldsymbol{v})\mathbf{A} - (\mathbf{x}^\top\mathbf{A}\mathbf{v})\mathbf{B}) + 2(\mathbf{A}\mathbf{x}\boldsymbol{v}^\top\mathbf{B} - \mathbf{B}\mathbf{x}\boldsymbol{v}^\top\mathbf{A})}{(\mathbf{x}^\top\mathbf{B}\mathbf{x})^2} \\
&\quad - 2\frac{\left[(\mathbf{x}^\top\mathbf{B}\mathbf{x})\mathbf{A} - (\mathbf{x}^\top\mathbf{A}\mathbf{x})\mathbf{B} + 2(\mathbf{A}\mathbf{x}\mathbf{x}^\top\mathbf{B} - \mathbf{B}\mathbf{x}\mathbf{x}^\top\mathbf{A})\right]\mathbf{x}\boldsymbol{v}^\top(2\mathbf{B})}{(\mathbf{x}^\top\mathbf{B}\mathbf{x})^3},
\end{aligned}$$

whose norm is bounded by $36\|\mathbf{A}\|\|\mathbf{B}\|^2/\lambda_{\min}^3(\mathbf{B})$, and

$$\begin{aligned}
&\nabla\left[\frac{\left[(\mathbf{x}^\top\mathbf{B}\mathbf{x})\mathbf{A} - (\mathbf{x}^\top\mathbf{A}\mathbf{x})\mathbf{B}\right]\mathbf{x}\mathbf{x}^\top\mathbf{B}}{(\mathbf{x}^\top\mathbf{B}\mathbf{x})^3}\boldsymbol{v}\right] = \nabla\left[\frac{(\mathbf{x}^\top\mathbf{B}\mathbf{v})\left[(\mathbf{x}^\top\mathbf{B}\mathbf{x})\mathbf{A}\mathbf{x} - (\mathbf{x}^\top\mathbf{A}\mathbf{x})\mathbf{B}\mathbf{x}\right]}{(\mathbf{x}^\top\mathbf{B}\mathbf{x})^3}\right] \\
&= \frac{\left[(\mathbf{x}^\top\mathbf{B}\mathbf{x})\mathbf{A} - (\mathbf{x}^\top\mathbf{A}\mathbf{x})\mathbf{B}\right]\mathbf{x}\boldsymbol{v}^\top\mathbf{B}}{(\mathbf{x}^\top\mathbf{B}\mathbf{x})^3} + \frac{(\mathbf{x}^\top\mathbf{B}\mathbf{v})\left[(\mathbf{x}^\top\mathbf{B}\mathbf{x})\mathbf{A} - (\mathbf{x}^\top\mathbf{A}\mathbf{x})\mathbf{B} + 2(\mathbf{A}\mathbf{x}\mathbf{x}^\top\mathbf{B} - \mathbf{B}\mathbf{x}\mathbf{x}^\top\mathbf{A})\right]}{(\mathbf{x}^\top\mathbf{B}\mathbf{x})^3} \\
&\quad - 3\left[\frac{(\mathbf{x}^\top\mathbf{B}\mathbf{v})\left[(\mathbf{x}^\top\mathbf{B}\mathbf{x})\mathbf{A} - (\mathbf{x}^\top\mathbf{A}\mathbf{x})\mathbf{B}\right]\mathbf{x}\mathbf{x}^\top(2\mathbf{B})}{(\mathbf{x}^\top\mathbf{B}\mathbf{x})^4}\right],
\end{aligned}$$

whose norm is thus bounded by $20\|\mathbf{A}\|\|\mathbf{B}\|^3/\lambda_{\min}^4(\mathbf{B})$. Again by mean value theorem we have

$$\|\mathrm{I}\| \le \frac{36\|\mathbf{A}\|\|\mathbf{B}\|^2}{\lambda_{\min}^3(\mathbf{B})}\|\mathbf{x} - \mathbf{y}\|,$$

and

$$\|\mathrm{II}\| \leq \frac{20\|\mathbf{A}\|\|\mathbf{B}\|^3}{\lambda_{\min}^4(\mathbf{B})}\|\mathbf{x}-\mathbf{y}\|,$$

so

$$\|\mathcal{H}(\mathbf{x})\boldsymbol{v} - \mathcal{H}(\mathbf{y})\boldsymbol{v}\| \leq \frac{56\|\mathbf{A}\|\|\mathbf{B}\|^3}{\lambda_{\min}^4(\mathbf{B})}\|\mathbf{x}-\mathbf{y}\|,$$

which concludes (67) via the definition of operator norm.

Lastly to conclude (68), we utilize the properties of projection matrices, $\|P_{\mathcal{T}(\mathbf{z})}\| \leq 1$, $\|P_{\mathcal{T}(\mathbf{z}_1)} - P_{\mathcal{T}(\mathbf{z}_2)}\| \leq \|\mathbf{z}_1 - \mathbf{z}_2\|$ and hence from matrix operator theory

$$\begin{aligned}
&\left\|P_{\mathcal{T}(\mathbf{z})}^\top \mathcal{H}(\mathbf{z}) P_{\mathcal{T}(\mathbf{z})} - P_{\mathcal{T}(\mathbf{z}')}^\top \mathcal{H}(\mathbf{z}') P_{\mathcal{T}(\mathbf{z}')}\right\| \\
&\leq \left\|P_{\mathcal{T}(\mathbf{z})}^\top \mathcal{H}(\mathbf{z}) P_{\mathcal{T}(\mathbf{z})} - P_{\mathcal{T}(\mathbf{z})}^\top \mathcal{H}(\mathbf{z}) P_{\mathcal{T}(\mathbf{z}')}\right\| \\
&+ \left\|P_{\mathcal{T}(\mathbf{z})}^\top \mathcal{H}(\mathbf{z}) P_{\mathcal{T}(\mathbf{z}')} - P_{\mathcal{T}(\mathbf{z})}^\top \mathcal{H}(\mathbf{z}') P_{\mathcal{T}(\mathbf{z}')}\right\| + \left\|P_{\mathcal{T}(\mathbf{z})}^\top \mathcal{H}(\mathbf{z}') P_{\mathcal{T}(\mathbf{z}')} - P_{\mathcal{T}(\mathbf{z}')}^\top \mathcal{H}(\mathbf{z}') P_{\mathcal{T}(\mathbf{z}')}\right\| \\
&\leq \|P_{\mathcal{T}(\mathbf{z})}^\top\|\|\mathcal{H}(\mathbf{z})\|\|P_{\mathcal{T}(\mathbf{z})} - P_{\mathcal{T}(\mathbf{z}')}\| \\
&\quad + \|P_{\mathcal{T}(\mathbf{z})}^\top\| \, \|\mathcal{H}(\mathbf{z}) - \mathcal{H}(\mathbf{z}')\| \, \|P_{\mathcal{T}(\mathbf{z}')}\| + \left\|(P_{\mathcal{T}(\mathbf{z})} - P_{\mathcal{T}(\mathbf{z}')})^\top\right\| \|\mathcal{H}(\mathbf{z}')\|\|P_{\mathcal{T}(\mathbf{z}')}\| \\
&\leq L_G\|\mathbf{z}-\mathbf{z}'\| + L_H\|\mathbf{z}-\mathbf{z}'\| + L_G\|\mathbf{z}-\mathbf{z}'\| \\
&= (2L_G + L_H)\|\mathbf{z}-\mathbf{z}'\|.
\end{aligned}$$

We complete our proof.

■

We now come to explore what the small gradient condition $\|g(\mathbf{x})\| \leq \gamma$, where $g(\cdot)$ is defined in (64), means for a point $\mathbf{x}$ in the GEV Problem. We first analyze the case where $\mathbf{B}$ is the identity matrix, which reduces to the classical Eigenvector Problem. Define for convenience

$$\gamma_1 \equiv \left(\frac{\|\mathbf{B}\|}{\lambda_{\min}(\mathbf{B})}\right)^{1/2} \frac{\gamma}{2\lambda_{\text{gap}}}. \tag{69}$$

Lemma 32 *When $\mathbf{B} = \mathbf{I}$, we have under $\|\boldsymbol{w}\| = 1$, an arbitrary constant $\gamma_1 \in (0, 1/2)$ and*

$$\left\|\Lambda\boldsymbol{w} - \frac{\boldsymbol{w}^\top\Lambda\boldsymbol{w}}{\boldsymbol{w}^\top\boldsymbol{w}}\boldsymbol{w}\right\| \leq \lambda_{gap}\gamma_1,$$

and for some $j = 1, \ldots, d$ (for consistency we define $\lambda_0 = \lambda_1$ and $\lambda_{d+1} = \lambda_d$)

$$\frac{\boldsymbol{w}^\top\Lambda\boldsymbol{w}}{\boldsymbol{w}^\top\boldsymbol{w}} \in \left[\frac{\lambda_{j-1}+\lambda_j}{2}, \frac{\lambda_j+\lambda_{j+1}}{2}\right],$$

together imply

$$(\mathbf{e}_j^\top\boldsymbol{w})^2 \geq 1 - 4\gamma_1^2.$$

Proof Denote till the rest of this proof $w_i = \mathbf{e}_i^\top\boldsymbol{w}$. Note we have by

$$\begin{aligned}
\lambda_{\text{gap}}^2\gamma_1^2 &\geq \left\|\Lambda\boldsymbol{w} - \frac{\boldsymbol{w}^\top\Lambda\boldsymbol{w}}{\boldsymbol{w}^\top\boldsymbol{w}}\boldsymbol{w}\right\|^2 = \sum_{i=1}^d \left(\lambda_i - \frac{\boldsymbol{w}^\top\Lambda\boldsymbol{w}}{\boldsymbol{w}^\top\boldsymbol{w}}\right)^2 w_i^2 \\
&\geq \sum_{i=1}^{j-1}\left(\lambda_i - \frac{\boldsymbol{w}^\top\Lambda\boldsymbol{w}}{\boldsymbol{w}^\top\boldsymbol{w}}\right)^2 w_i^2 + \sum_{i=j+1}^{d}\left(\lambda_i - \frac{\boldsymbol{w}^\top\Lambda\boldsymbol{w}}{\boldsymbol{w}^\top\boldsymbol{w}}\right)^2 w_i^2 \\
&\geq \sum_{i=1}^{j-1}\left(\lambda_i - \frac{\lambda_{j-1}+\lambda_j}{2}\right)^2 w_i^2 + \sum_{i=j+1}^{d}\left(\lambda_i - \frac{\lambda_j+\lambda_{j+1}}{2}\right)^2 w_i^2 \\
&\geq \left(\frac{\lambda_j - \lambda_{j-1}}{2}\right)^2 \sum_{i=1}^{j-1} w_i^2 + \left(\frac{\lambda_{j+1}-\lambda_j}{2}\right)^2 \sum_{i=j+1}^{d} w_i^2 \geq \frac{\lambda_{\text{gap}}^2}{4}\left(1 - w_j^2\right).
\end{aligned}$$

This implies the lemma immediately. ■

To study the case of general $\mathbf{B}$, we first introduce an auxiliary lemma.

Lemma 33 *Given two norms $\|\cdot\|_1, \|\cdot\|_2$ that are equivalent: there are constants $C_L, C_U > 0$ such that for every nonzero vector $\boldsymbol{v}$, $C_L\|\boldsymbol{v}\|_2 \le \|\boldsymbol{v}\|_1 \le C_U\|\boldsymbol{v}\|_2$. Then for two given nonzero vectors $\boldsymbol{w}_1, \boldsymbol{w}_2$, we have*

$$\left\|\frac{\boldsymbol{w}_1}{\|\boldsymbol{w}_1\|_1} - \frac{\boldsymbol{w}_2}{\|\boldsymbol{w}_2\|_1}\right\|_1 \le 2C_L^{-1}C_U \left\|\frac{\boldsymbol{w}_1}{\|\boldsymbol{w}_1\|_2} - \frac{\boldsymbol{w}_2}{\|\boldsymbol{w}_2\|_2}\right\|_2$$

Proof Without loss of generality we set $\|\boldsymbol{w}_1\|_2 = 1 = \|\boldsymbol{w}_2\|_2$. Then using triangle inequality we have

$$\begin{aligned}
LHS &= \left\|\frac{\boldsymbol{w}_1}{\|\boldsymbol{w}_1\|_1} - \frac{\boldsymbol{w}_2}{\|\boldsymbol{w}_2\|_1}\right\|_1 = \frac{\|\|\boldsymbol{w}_2\|_1\boldsymbol{w}_1 - \|\boldsymbol{w}_1\|_1\boldsymbol{w}_2\|_1}{\|\boldsymbol{w}_1\|_1\|\boldsymbol{w}_2\|_1} \\
&= \frac{\|\|\boldsymbol{w}_2\|_1\boldsymbol{w}_1 - \|\boldsymbol{w}_1\|_1\boldsymbol{w}_1 + \|\boldsymbol{w}_1\|_1\boldsymbol{w}_1 - \|\boldsymbol{w}_1\|_1\boldsymbol{w}_2\|_1}{\|\boldsymbol{w}_1\|_1\|\boldsymbol{w}_2\|_1} \\
&\le \frac{\left|\|\boldsymbol{w}_2\|_1 - \|\boldsymbol{w}_1\|_1\right| \|\boldsymbol{w}_1\|_1 + \|\boldsymbol{w}_1\|_1 \|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_1}{\|\boldsymbol{w}_1\|_1\|\boldsymbol{w}_2\|_1} \\
&\le \frac{2\|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_1\|\boldsymbol{w}_1\|_1}{\|\boldsymbol{w}_1\|_1\|\boldsymbol{w}_2\|_1} = 2\|\boldsymbol{w}_2\|_1^{-1} \cdot \|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_1 \le 2C_L^{-1}\|\boldsymbol{w}_2\|_2^{-1} \cdot C_U\|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_2 = RHS.
\end{aligned}$$

■

We conclude the following lemma.

Lemma 34 *We have for $\mathbf{x} \in \mathcal{S}^{d-1}$,*

$$\gamma \in \left(0, \left(\frac{\|\mathbf{B}\|}{\lambda_{\min}(\mathbf{B})}\right)^{-1/2} \lambda_{gap}\right), \tag{70}$$

and $\|g(\mathbf{x})\| \le \gamma$ implies that there exists at least one $j = 1, \ldots, d$ such that

$$(\mathbf{e}_j^\top \mathbf{B}^{1/2}\mathbf{x})^2 \ge (1 - 4\gamma_1^2)\|\mathbf{B}^{1/2}\mathbf{x}\|^2. \tag{71}$$

Furthermore, we have that there exists at least one $j = 1, \ldots, d$ such that

$$\min\left(\|\mathbf{x} - \boldsymbol{v}_j\|, \|\mathbf{x} + \boldsymbol{v}_j\|\right) \le 4\sqrt{2}\left(\frac{\|\mathbf{B}\|}{\lambda_{\min}(\mathbf{B})}\right)^{1/2} \cdot \gamma_1. \tag{72}$$

Proof Since $\mathbf{B}$ is positive definite, letting in (32) $w = \mathbf{B}^{1/2}\mathbf{x}/\|\mathbf{B}^{1/2}\mathbf{x}\|$, we have $\|\boldsymbol{w}\| = 1$ and recall that $\mathbf{A} = \mathbf{B}^{1/2}\Lambda\mathbf{B}^{1/2}$

$$\begin{aligned}
\left\|\Lambda\boldsymbol{w} - \frac{\boldsymbol{w}^\top\Lambda\boldsymbol{w}}{\boldsymbol{w}^\top\boldsymbol{w}}\boldsymbol{w}\right\| &= \|\mathbf{B}^{1/2}\mathbf{x}\| \left\|\mathbf{B}^{-1/2}\left(\frac{\mathbf{B}^{1/2}\Lambda\mathbf{B}^{1/2}}{\mathbf{x}^\top\mathbf{B}\mathbf{x}}\mathbf{x} - \frac{\mathbf{x}^\top\mathbf{B}^{1/2}\Lambda\mathbf{B}^{1/2}\mathbf{x}}{(\mathbf{x}^\top\mathbf{B}\mathbf{x})^2}\mathbf{B}\mathbf{x}\right)\right\| \\
&\le \left(\frac{\|\mathbf{B}\|}{\lambda_{\min}(\mathbf{B})}\right)^{1/2} \left\|\frac{(\mathbf{x}^\top\mathbf{B}\mathbf{x})\mathbf{A} - (\mathbf{x}^\top\mathbf{A}\mathbf{x})\mathbf{B}}{(\mathbf{x}^\top\mathbf{B}\mathbf{x})^2}\mathbf{x}\right\| \le \left(\frac{\|\mathbf{B}\|}{\lambda_{\min}(\mathbf{B})}\right)^{1/2} \frac{\gamma}{2} = \lambda_{\text{gap}}\gamma_1.
\end{aligned}$$

Note (73) gives $\gamma_1 \in (0, 1/2)$, and hence applying Lemma 32 gives the following: there is at least one $j = 1, \ldots, d$ such that $(\mathbf{e}_j^\top\boldsymbol{w})^2 \ge 1 - 4\gamma_1^2$. Translating this back in terms of $\mathbf{x}$ concludes (71).

To conclude (74) we note (71) gives if $\langle\mathbf{z}_1, \mathbf{z}_2\rangle_B \equiv \mathbf{z}_1^\top\mathbf{B}\mathbf{z}_2$ and $\|\mathbf{z}\|_\mathbf{B} \equiv \langle z, z\rangle_B^{1/2}$:

$$\left\langle \mathbf{e}_j^\top\mathbf{B}^{-1/2}, \frac{\mathbf{x}}{\|\mathbf{x}\|_\mathbf{B}}\right\rangle_B^2 \ge 1 - 4\gamma_1^2,$$

so

$$\left\|\mathbf{e}_j^\top \mathbf{B}^{-1/2} \pm \frac{\mathbf{x}}{\|\mathbf{x}\|_{\mathbf{B}}}\right\|_{\mathbf{B}}^2 = \left\|\mathbf{e}_j^\top \mathbf{B}^{-1/2}\right\|_{\mathbf{B}}^2 + \left\|\frac{\mathbf{x}}{\|\mathbf{x}\|_{\mathbf{B}}}\right\|_{\mathbf{B}}^2 \pm 2\left\langle \mathbf{e}_j^\top \mathbf{B}^{-1/2}, \frac{\mathbf{x}}{\|\mathbf{x}\|_{\mathbf{B}}}\right\rangle_{\mathbf{B}}$$
$$= 2 \pm 2\left\langle \mathbf{e}_j^\top \mathbf{B}^{-1/2}, \frac{\mathbf{x}}{\|\mathbf{x}\|_{\mathbf{B}}}\right\rangle_{\mathbf{B}},$$

and hence using $1-\sqrt{1-t} \le t$ for $t \in [0,1]$

$$\min\left\|\mathbf{e}_j^\top \mathbf{B}^{-1/2} \pm \frac{\mathbf{x}}{\|\mathbf{x}\|_{\mathbf{B}}}\right\|_{\mathbf{B}}^2 = 2 - 2\left|\left\langle \mathbf{e}_j^\top \mathbf{B}^{-1/2}, \frac{\mathbf{x}}{\|\mathbf{x}\|_{\mathbf{B}}}\right\rangle_{\mathbf{B}}\right| = 2 - 2\sqrt{1-4\gamma_1^2} \le 8\gamma_1^2.$$

Using this and applying Lemma 33 with $\|\cdot\|_1 = \|\cdot\|$ and $\|\cdot\|_2 = \|\cdot\|_{\mathbf{B}}$ we have $\lambda_{\max}^{-1/2}(\mathbf{B})\|\boldsymbol{v}\|_{\mathbf{B}} \le \|\boldsymbol{v}\| \le \lambda_{\min}^{-1/2}(\mathbf{B})\|\boldsymbol{v}\|_{\mathbf{B}}$ and hence for two given nonzero vectors (in the Euclidean norm) $\mathbf{x}$ and $\mp\boldsymbol{v}_j = \mp\|\mathbf{e}_j^\top \mathbf{B}^{-1/2}\|^{-1}\mathbf{e}_j^\top \mathbf{B}^{-1/2}$

$$\min\|\mathbf{x} \pm \boldsymbol{v}_j\| \le 2\left(\frac{\|\mathbf{B}\|}{\lambda_{\min}(\mathbf{B})}\right)^{1/2} \cdot \min\left\|\mathbf{e}_j^\top \mathbf{B}^{-1/2} \pm \frac{\mathbf{x}}{\|\mathbf{x}\|_{\mathbf{B}}}\right\|_{\mathbf{B}} \le 4\sqrt{2}\left(\frac{\|\mathbf{B}\|}{\lambda_{\min}(\mathbf{B})}\right)^{1/2} \cdot \gamma_1.$$

■

Now we finish the proof of Lemma 12.

Proof [Proof of Lemma 12]

(i) We have $\mathbf{A}\mathbf{x} = (\mathbf{x}^\top\mathbf{A}\mathbf{x}/\mathbf{x}^\top\mathbf{B}\mathbf{x})\mathbf{B}\mathbf{x}$ if and only if $g(\mathbf{x}) = 0$. For $\Lambda = \mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2}$ being WLOG diagonal, one can see that for $j = 2, \ldots, d$ and $\boldsymbol{v}_j$ on the unit sphere with $\mathbf{A}\boldsymbol{v}_j = \lambda_j\mathbf{B}\boldsymbol{v}_j$,

$$\mathcal{H}(\boldsymbol{v}_j) = -2 \cdot \frac{(\boldsymbol{v}_j^\top\mathbf{B}\boldsymbol{v}_j)\mathbf{A} - (\boldsymbol{v}_j^\top\mathbf{A}\boldsymbol{v}_j)\mathbf{B}}{(\boldsymbol{v}_j^\top\mathbf{B}\boldsymbol{v}_j)^2} = -2 \cdot \frac{\mathbf{A} - \lambda_j\mathbf{B}}{\boldsymbol{v}_j^\top\mathbf{B}\boldsymbol{v}_j}.$$

Thus

$$\begin{aligned}
&(\boldsymbol{v}_1 - c\boldsymbol{v}_j)^\top\mathcal{H}(\boldsymbol{v}_j)(\boldsymbol{v}_1 - c\boldsymbol{v}_j) \\
&= -2 \cdot \frac{(\boldsymbol{v}_1 - c\boldsymbol{v}_j)^\top(\mathbf{A} - \lambda_j\mathbf{B})(\boldsymbol{v}_1 - c\boldsymbol{v}_j)}{\boldsymbol{v}_j^\top\mathbf{B}\boldsymbol{v}_j} \\
&= -2 \cdot \frac{\boldsymbol{v}_1^\top(\mathbf{A} - \lambda_j\mathbf{B})\boldsymbol{v}_1}{\boldsymbol{v}_j^\top\mathbf{B}\boldsymbol{v}_j} + 4c \cdot \frac{\boldsymbol{v}_j^\top(\mathbf{A} - \lambda_j\mathbf{B})\boldsymbol{v}_1}{\boldsymbol{v}_j^\top\mathbf{B}\boldsymbol{v}_j} - 2c^2 \cdot \frac{\boldsymbol{v}_j^\top(\mathbf{A} - \lambda_j\mathbf{B})\boldsymbol{v}_j}{\boldsymbol{v}_j^\top\mathbf{B}\boldsymbol{v}_j} \\
&= -2 \cdot \frac{\boldsymbol{v}_1^\top(\mathbf{A} - \lambda_j\mathbf{B})\boldsymbol{v}_1}{\boldsymbol{v}_j^\top\mathbf{B}\boldsymbol{v}_j} \\
&= -2(\lambda_1 - \lambda_j) \cdot \frac{\boldsymbol{v}_1^\top\mathbf{B}\boldsymbol{v}_1}{\boldsymbol{v}_j^\top\mathbf{B}\boldsymbol{v}_j} \le -2(\lambda_1 - \lambda_2) \cdot \frac{\lambda_{\min}(\mathbf{B})}{\|\mathbf{B}\|}.
\end{aligned}$$

In the display above, we use the fact that $\boldsymbol{v}_j^\top(\mathbf{A} - \lambda_j\mathbf{B})\boldsymbol{v}_1 = (\lambda_1 - \lambda_j)\boldsymbol{v}_j^\top\mathbf{B}\boldsymbol{v}_1 = 0$ and $\boldsymbol{v}_j^\top(\mathbf{A} - \lambda_j\mathbf{B})\boldsymbol{v}_j = (\lambda_j - \lambda_j)\boldsymbol{v}_j^\top\mathbf{B}\boldsymbol{v}_j = 0$. By picking $c = \boldsymbol{v}_j^\top\boldsymbol{v}_1$ such that $P_{\mathcal{T}(\boldsymbol{v}_j)}\boldsymbol{v}_1 = \boldsymbol{v}_1 - (\boldsymbol{v}_j^\top\boldsymbol{v}_1)\boldsymbol{v}_j = \boldsymbol{v}_1 - c\boldsymbol{v}_j$, we conclude

$$\|P_{\mathcal{T}(\boldsymbol{v}_j)}\boldsymbol{v}_1\| = \sqrt{1 + (\boldsymbol{v}_j^\top\boldsymbol{v}_1)^2 - 2(\boldsymbol{v}_j^\top\boldsymbol{v}_1)^2} = \sqrt{1 - (\boldsymbol{v}_j^\top\boldsymbol{v}_1)^2} \in (0, 1],$$

(since $\boldsymbol{v}_1 \neq \pm\boldsymbol{v}_j$ otherwise $0 = \boldsymbol{v}_j^\top\mathbf{B}\boldsymbol{v}_1 = \pm\boldsymbol{v}_1^\top\mathbf{B}\boldsymbol{v}_1$ which leads to $\boldsymbol{v}_1 = 0$ due to the positive definiteness of $\mathbf{B}$.) and hence from the above two displays

$$\boldsymbol{v}_1^\top\left[P_{\mathcal{T}(\boldsymbol{v}_j)}^\top\mathcal{H}(\boldsymbol{v}_j)P_{\mathcal{T}(\boldsymbol{v}_j)}\right]\boldsymbol{v}_1 \le -2(\lambda_1 - \lambda_2) \cdot \frac{\lambda_{\min}(\mathbf{B})}{\|\mathbf{B}\|}\|P_{\mathcal{T}(\boldsymbol{v}_j)}(\boldsymbol{v}_1)\|^2.$$

(ii) To conclude points that are close to $P_{\mathcal{T}(\boldsymbol{v}_j)}\boldsymbol{v}_1$, Lemma 34 gives for $\mathbf{x} \in \mathcal{S}^{d-1}$,

$$\gamma \in \left(0, \left(\frac{\|\mathbf{B}\|}{\lambda_{\min}(\mathbf{B})}\right)^{-1/2} \lambda_{\text{gap}}\right), \tag{73}$$

and $\|g(\mathbf{x})\| \leq \gamma$ implies that there exists at least one $j = 1, \ldots, d$ such that

$$\min \|\mathbf{x} \pm \boldsymbol{v}_j\| \leq 4\sqrt{2}\left(\frac{\|\mathbf{B}\|}{\lambda_{\min}(\mathbf{B})}\right)^{1/2} \cdot \gamma_1. \tag{74}$$

Without loss of generality we suppose the minus sign in the above display is taken, so $\min \|\mathbf{x} - \boldsymbol{v}_j\| \leq 4\sqrt{2}\left(\frac{\|\mathbf{B}\|}{\lambda_{\min}(\mathbf{B})}\right)^{1/2} \cdot \gamma_1$. Then given the definition of γ_1 in (69) we have from Lemma 31 that

$$\begin{aligned}
&\boldsymbol{v}_1^\top \left[P_{\mathcal{T}(\mathbf{x})}^\top \mathcal{H}(\mathbf{x}) P_{\mathcal{T}(\mathbf{x})}\right] \boldsymbol{v}_1 \\
&\leq \boldsymbol{v}_1^\top \left[P_{\mathcal{T}(\boldsymbol{v}_j)}^\top \mathcal{H}(\boldsymbol{v}_j) P_{\mathcal{T}(\boldsymbol{v}_j)}\right] \boldsymbol{v}_1 + \left\|P_{\mathcal{T}(\boldsymbol{v}_j)}^\top \mathcal{H}(\boldsymbol{v}_j) P_{\mathcal{T}(\boldsymbol{v}_j)} - P_{\mathcal{T}(\mathbf{x})}^\top \mathcal{H}(\mathbf{x}) P_{\mathcal{T}(\mathbf{x})}\right\| \\
&\leq -2(\lambda_1 - \lambda_2) \cdot \frac{\lambda_{\min}(\mathbf{B})}{\|\mathbf{B}\|} + (2L_G + L_H)\|\mathbf{x} - \boldsymbol{v}_j\| \\
&\leq -(\lambda_1 - \lambda_2) \cdot \frac{\lambda_{\min}(\mathbf{B})}{\|\mathbf{B}\|} \leq -(\lambda_1 - \lambda_2) \cdot \frac{\lambda_{\min}(\mathbf{B})}{\|\mathbf{B}\|} \|P_{\mathcal{T}(\mathbf{x})}\boldsymbol{v}_1\|^2,
\end{aligned}$$

as long as (combined with (74))

$$4\sqrt{2}(2L_G + L_H)\left(\frac{\|\mathbf{B}\|}{\lambda_{\min}(\mathbf{B})}\right)^{1/2} \cdot \gamma_1 \leq (\lambda_1 - \lambda_2) \cdot \frac{\lambda_{\min}(\mathbf{B})}{\|\mathbf{B}\|},$$

where we applied $\|P_{\mathcal{T}(\mathbf{x})}\boldsymbol{v}_1\| \leq 1$. This completes the proof of Lemma combining with the definition of β in (36).

■

E DEFERRED PROOFS OF §B.1

We collect the deferred proofs from §B.1.

E.1 PROOF OF LEMMA 13

Proof [Proof of Lemma 13] 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 the Markov inequality and Assumption 2 (with law of total expectation applied). Taking a union bound,

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

■

E.2 PROOF OF LEMMA 14

Proof [Proof of Lemma 14] For all $\boldsymbol{u}, \boldsymbol{v} \in \mathcal{S}^{d-1}$, we have

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

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 $\|\mathcal{H}(\boldsymbol{v})\| \le B_H \equiv L_F + L_K$ within $\{\boldsymbol{v} : \|\boldsymbol{v}\| \le 1, \|\boldsymbol{v} - \boldsymbol{v}^*\| \le \delta\}$. ■

E.3 PROOF OF LEMMA 15

Proof [Proof of Lemma 15] We have by a Taylor series expansion that for any $y \in \mathbb{R}$ satisfying $|y| \le 1/2$

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

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

$$\begin{aligned}
&\left|\|\boldsymbol{v}_{t-1} - \eta\Gamma(\boldsymbol{v}_{t-1};\boldsymbol{\zeta}_t)\|^{-1} - 1 - \eta\boldsymbol{v}_{t-1}^\top\Gamma(\boldsymbol{v}_{t-1};\boldsymbol{\zeta}_t)\right| \\
&\le \left|\left(1 - 2\eta\boldsymbol{v}_{t-1}^\top\Gamma(\boldsymbol{v}_{t-1};\boldsymbol{\zeta}_t) + \eta^2\|\Gamma(\boldsymbol{v}_{t-1};\boldsymbol{\zeta}_t)\|^2\right)^{-1/2} - 1 - \eta\boldsymbol{v}_{t-1}^\top\Gamma(\boldsymbol{v}_{t-1};\boldsymbol{\zeta}_t)\right| \\
&\le \left|(1-y)^{-1/2} - 1 - \frac{y}{2}\right| + \frac{\eta^2\|\Gamma(\boldsymbol{v}_{t-1};\boldsymbol{\zeta}_t)\|^2}{2} \\
&\le \frac{3y^2}{4} + \frac{1}{2}\eta^2 M^2 \le \frac{3}{4}\cdot\frac{121}{25}\eta^2 M^2 + \frac{1}{2}\eta^2 M^2 \le 5\eta^2 M^2.
\end{aligned} \tag{75}$$

By defining

$$\boldsymbol{\xi}_t = (\mathbf{I} - \boldsymbol{v}_{t-1}\boldsymbol{v}_{t-1}^\top)(\Gamma(\boldsymbol{v}_{t-1};\boldsymbol{\zeta}_t) - D(\boldsymbol{v}_{t-1})\nabla F(\boldsymbol{v}_{t-1})), \tag{76}$$

and

$$\begin{aligned}
\boldsymbol{Q}_t = \eta^{-2}\cdot&\left(\|\boldsymbol{v}_{t-1} - \eta\Gamma(\boldsymbol{v}_{t-1};\boldsymbol{\zeta}_t)\|^{-1} - 1 - \eta\boldsymbol{v}_{t-1}^\top\Gamma(\boldsymbol{v}_{t-1};\boldsymbol{\zeta}_t)\right)(\boldsymbol{v}_{t-1} - \eta\Gamma(\boldsymbol{v}_{t-1};\boldsymbol{\zeta}_t)) \\
&- (\boldsymbol{v}_{t-1}^\top\Gamma(\boldsymbol{v}_{t-1};\boldsymbol{\zeta}_t))\Gamma(\boldsymbol{v}_{t-1};\boldsymbol{\zeta}_t),
\end{aligned} \tag{77}$$

the update formula (9) is equivalent to

$$\boldsymbol{v}_t = \boldsymbol{v}_{t-1} - \eta D(\boldsymbol{v}_{t-1})g(\boldsymbol{v}_{t-1}) + \eta\boldsymbol{\xi}_t + \eta^2\boldsymbol{Q}_t. \tag{78}$$

Using (75), we have

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

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

$$\boldsymbol{R}_t = D(\mathcal{H}_* + \mathcal{N}_*)(\boldsymbol{v}_{t-1} - \boldsymbol{v}^*) - D(\boldsymbol{v}_{t-1})g(\boldsymbol{v}_{t-1}), \tag{79}$$

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 $\|\mathbf{I} - \boldsymbol{v}_{t-1}\boldsymbol{v}_{t-1}^\top\| \leq 1$, and hence from Assumption 2 and Lemma 27 we know

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

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

Finally, we apply the mean-value theorem using (42) and $g(\boldsymbol{v}^*) = 0$ to obtain

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

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 15. ■

E.4 PROOF OF LEMMA 16

Proof [Proof of Lemma 16] 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

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

by the 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 $\boldsymbol{v} = \boldsymbol{v}_t$ gives

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

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

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

and hence

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

(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

$$\|(\mathbf{I} - \eta D\mathcal{M}_*)^{1/2}\boldsymbol{u}\| \leq (1 - \eta D\mu)^{1/2}\|\boldsymbol{u}\|. \tag{80}$$

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 (80) and obtain for each $t \geq 0$

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

■

E.5 PROOF OF LEMMA 17

Proof [Proof of Lemma 17] By left multiplying (44) in Lemma 15 by $(\mathbf{I} - \boldsymbol{v}^*\boldsymbol{v}^{*\top})$ and noticing $(\mathbf{I} - \boldsymbol{v}^*\boldsymbol{v}^{*\top})\mathcal{N}_* = 0$, we obtain

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

We have the decomposition

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

where $(\mathbf{I} - \boldsymbol{v}^*\boldsymbol{v}^{*\top})\mathcal{H}_*\boldsymbol{\Delta}_t = \mathcal{M}_*\boldsymbol{\Delta}_t$, and based on Lemma 16 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

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

which proves (49). The rest of Lemma 17 can be easily verified in steps similar to the proof of Lemma 15.

■

E.6 PROOF OF LEMMA 18

Proof [Proof of Lemma 18] For $t = 0$ the lemma holds by definition. In general if it holds for $t - 1$ then from the definitions in (50) 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 (51) we obtain (52), which concludes our lemma. ■

E.7 PROOF OF LEMMA 19

Proof [Proof of Lemma 19] For any fixed $t \ge 0$, we have the following:

(i) For the first term on the right hand of (52) which we repeat here

$$\begin{aligned}\overline{\boldsymbol{\Delta}}_t = (\mathbf{I} - \eta D\mathcal{M}_*)^t\,\boldsymbol{\Delta}_0 + \eta\sum_{s=1}^{t}(\mathbf{I} - \eta D\mathcal{M}_*)^{t-s}\,\boldsymbol{\chi}_s \\ + \eta\sum_{s=1}^{t}(\mathbf{I} - \eta D\mathcal{M}_*)^{t-s}\,\widetilde{\boldsymbol{S}}_s + \eta^2\sum_{s=1}^{t}(\mathbf{I} - \eta D\mathcal{M}_*)^{t-s}\,\widetilde{\boldsymbol{P}}_s,\end{aligned} \tag{(52)}$$

since $\boldsymbol{\chi}_s \in \mathcal{T}(\boldsymbol{v}^*)$, (48) in Lemma 16 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

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

Modifying the results in Fan et al. [2012] provides a concentration inequality for α-sub-Weibull random vectors, which gives[3]

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

(ii) For the second term on the right-hand side of (52), by applying (48) in Lemma 16 and using Lemma 17, given $\|\boldsymbol{v}_{s-1}-\boldsymbol{v}^*\|\leq r$ for all $s=1,\ldots,t$ we have,

$$\left\|\eta\sum_{s=1}^{t}(\mathbf{I}-\eta D\mathcal{M}_*)^{t-s}\,\widetilde{\boldsymbol{S}}_s\right\| \leq \eta\sum_{s=1}^{t}(1-\eta D\mu)^{t-s}\cdot\rho r^2 \leq \frac{\rho r^2}{D\mu}. \tag{81}$$

(iii) For the third term on the right-hand side of (52), from Lemma 17 we know $\|\widetilde{\boldsymbol{P}}_t\|\leq 7M^2$ and

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

where we use the definition of M in (40).

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

E.8 PROOF OF LEMMA 20

Proof [Proof of Lemma 20] From the given assumptions, under scaling condition (17), we have

$$r = 2\max\left\{\|\boldsymbol{\Delta}_0\|,\ \frac{2^7G_\alpha\mathcal{V}}{\sqrt{D\mu}}\log^{\frac{\alpha+2}{2\alpha}}\epsilon^{-1}\cdot\eta^{1/2}\right\} \leq \frac{D\mu}{16\rho}.$$

We let event $\mathscr{J}$ be (53) holding for each $t\in[0,T]$, i.e.

$$\left\|\overline{\boldsymbol{\Delta}}_t-(\mathbf{I}-\eta D\mathcal{M}_*)^t\boldsymbol{\Delta}_0\right\| \leq \frac{8G_\alpha\mathcal{V}}{\sqrt{D\mu}}\log^{\frac{\alpha+2}{2\alpha}}\epsilon^{-1}\cdot\eta^{1/2}+\frac{\rho r^2}{D\mu}+\frac{7\mathcal{V}^2}{D\mu}\log^{\frac{2}{\alpha}}\epsilon^{-1}\cdot\eta.$$

Then on event $\mathscr{J}$, under scaling condition (17), because $\|\boldsymbol{\Delta}_0\|\leq\frac{r}{2}$, for each $t\in[0,T]$ we have

$$\|\overline{\boldsymbol{\Delta}}_t\| \leq \|\boldsymbol{\Delta}_0\|+\frac{16G_\alpha\mathcal{V}}{\sqrt{D\mu}}\log^{\frac{\alpha+2}{2\alpha}}\epsilon^{-1}\cdot\eta^{1/2}+\frac{\rho r^2}{D\mu} \leq \frac{r}{2}+\frac{r}{16}+\frac{r}{16}\leq r.$$

Applying Lemma 19 and taking a union bound gives

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

Furthermore, using (48) in Lemma 16 and definition of T_η^* in (16), if $T_\eta^*\in[0,T]$, on event $\mathscr{J}$ we have at time T_η^*

$$\|\overline{\boldsymbol{\Delta}}_{T_\eta^*}\| \leq \|(\mathbf{I}-\eta D\mathcal{M}_*)^{T_\eta^*}\boldsymbol{\Delta}_0\|+\frac{16G_\alpha\mathcal{V}}{\sqrt{D\mu}}\log^{\frac{\alpha+2}{2\alpha}}\epsilon^{-1}\cdot\eta^{1/2}+\frac{\rho r^2}{D\mu} \leq \frac{r}{8}+\frac{r}{16}+\frac{r}{16}\leq\frac{r}{4}.$$

In Lemma 18 we have shown that, on the event $(T<\mathcal{T}_M)$, we have $\overline{\boldsymbol{\Delta}}_t=\boldsymbol{\Delta}_t$. In Lemma 13, we have proved $\mathbb{P}(T<\mathcal{T}_M)\geq 1-2T\epsilon$. Together with Lemma 19, we take an intersection and obtain

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

At this point we have proved all elements in Lemma 20. ■

[3] A similar concentration inequality method for the scalar case is adopted by Li and Jordan [2021].