Nonconvex Stochastic Scaled-Gradient Descent and Generalized Eigenvector Problems

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 $\boldsymbol{v}$. 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

Submitted to the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023). **To be used for reviewing only**.

$\widetilde{\nabla} F(\boldsymbol{v}_{t-1}; \boldsymbol{\zeta}_t)$ 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 Arora et al. [2012], the major stumbling block in applying SGD to this problem lies in obtaining an unbiased stochastic sample of the gradient (6), due to the fact that the objective function takes a fractional form of two expectations. In our approach we circumvent this issue by simply replacing the denominator on the right-hand side of (6) by the constant 1 and using 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 $\widetilde{\mathbf{B}}$ is taken as $\mathbf{I}$, (7) essentially reproduces Oja's online PCA algorithm [Oja, 1982] with an incurred $O(\eta^2)$ 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):

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

To proceed, we take $\widetilde{\mathbf{A}}$ and $\widetilde{\mathbf{B}}'$ as mutually independent and unbiased stochastic samples of $\mathbf{A}$ and $\mathbf{B}$ respectively. It can be easily seen that the expectation of (8) is a scaled gradient of the generalized Rayleigh quotient, where the scaling is the factor $(\boldsymbol{v}^\top \mathbf{B} \boldsymbol{v})^2/2$. This approach, which has been referred to as *double stochastic sampling* in the setting of kernel methods [Dai et al., 2014, 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 vector-vector inner products and is hence efficient.

Our contributions relative to previous work on nonconvex stochastic optimization as are follows. First, we propose a novel algorithm—the stochastic scaled-gradient descent (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. §A presents our two-phase convergence results for arbitrary initialization. §4 investigates the asymptotic property of our algorithm. §5 uses the example of Canonical Correlation Analysis to demonstrate the practical computation and experimental performance of our algorithm. §B presents the proofs of our theoretical analysis. §6 summarizes the entire paper. Limited by space we relegate to Appendix all 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 $\mathcal{M}$.*

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

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

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

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 $\{\boldsymbol{v} : \|\boldsymbol{v}\| \leq 1, \|\boldsymbol{v} - \boldsymbol{v}^*\| \leq \delta\}$, 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 $\boldsymbol{v} \in \{\boldsymbol{v} : \|\boldsymbol{v}\| \leq 1, \|\boldsymbol{v} - \boldsymbol{v}^*\| \leq \delta\}$, 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 $\Gamma(\boldsymbol{v}_{t-1}; \boldsymbol{\zeta}_t), t \geq 1$ 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 $\mathbb{E} \exp\left(\frac{\|\Gamma(\boldsymbol{v};\boldsymbol{\zeta})\|^\alpha}{\mathcal{V}^\alpha}\right) \leq 2$, 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 projection 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,

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

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

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

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*

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

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

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

such that on event $\mathcal{H}_2$ the iterates generated by the SSGD algorithm satisfy for all $t \in [K_{\eta,\epsilon}T_\eta^*, T]$:

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

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

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

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

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

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

Proposition 3 studies $\mathbf{\Delta}_t$ 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 $\{\mathbf{\Delta}_t\}$ for $K_{\eta,\epsilon}$ rounds, initializing each round with the output $\mathbf{\Delta}_{T_\eta^*}$ 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

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

where the last inequality is due to initialization condition (14). 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

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

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 $T\ge K_{\eta,\epsilon}T_\eta^*$. 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*

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

there exists an event $\mathcal{H}_4$ with $\mathbb{P}(\mathcal{H}_4)\ge 1-\left(14+8\left(\frac{3}{\alpha}\right)^{\frac{2}{\alpha}}\log^{-\frac{\alpha+2}{\alpha}}\epsilon^{-1}\right)T\epsilon$, such that on the event $\mathcal{H}_4$ the iterates generated by the SSGD algorithm satisfy

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

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

4 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 $\mathcal{H}_*\equiv\mathcal{H}(\boldsymbol{v}^*),\mathbf{\Sigma}_*\equiv\mathbf{\Sigma}(\boldsymbol{v}^*)$ 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 $\boldsymbol{u}^\top\mathcal{M}_*\boldsymbol{u}\ge\mu\|\boldsymbol{u}\|^2$ 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 $\lambda_1\ge\ldots\ge\lambda_{d-1}>0$ with minimum positive eigenvalue $\lambda_{d-1}\ge\mu$. We take the inverse of all positive eigenvalues and define the following matrix

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

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:

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

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 3 *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 $\boldsymbol{\zeta}$*

$$\mathbb{E}\left\|\Gamma(\boldsymbol{v};\boldsymbol{\zeta}) - \Gamma(\boldsymbol{v}';\boldsymbol{\zeta})\right\|^2 \leq L_S^2\|\boldsymbol{v} - \boldsymbol{v}'\|^2. \tag{24}$$

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

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

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

we obtain Gaussian convergence in distribution:

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

We relegate the proof details of Theorem 5 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.

[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 Stochastic Scaled-Gradient Descent (Noisy)

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 $\boldsymbol{v}_T$

5 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.

Proposition 6 *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 6 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, $1 - \text{align}(\boldsymbol{v}, \boldsymbol{v}^*) \gtrsim d/T$, in terms of a new measure of error:

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

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 $1 - \text{align}(\boldsymbol{v}, \boldsymbol{v}^*) \asymp 1 - \boldsymbol{v}^\top \boldsymbol{v}^2 = \|\boldsymbol{v} - \boldsymbol{v}^*\|^2/2$ 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 6, 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.

5.1 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

$$\boldsymbol{\Sigma}_{\boldsymbol{XX}} = 3\mathbf{I}_{d_1} + \mathbf{A}_1, \quad \boldsymbol{\Sigma}_{\boldsymbol{YY}} = 3\mathbf{I}_{d_2} + \mathbf{A}_2, \tag{27}$$

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 $\boldsymbol{\Sigma}_{\boldsymbol{XX}}^{-\frac{1}{2}} \boldsymbol{\Sigma}_{\boldsymbol{XY}} \boldsymbol{\Sigma}_{\boldsymbol{YY}}^{-\frac{1}{2}}$ is significantly large, in particular, no less than 0.5, we set

$$\boldsymbol{\Sigma}_{\boldsymbol{XY}} = \mathbf{A}_3 + \boldsymbol{\Sigma}_{\boldsymbol{XX}}^{1/2} \mathbf{U}\, \text{diag}(0.5, \mathbf{O}) \mathbf{V}^\top \boldsymbol{\Sigma}_{\boldsymbol{YY}}^{1/2}. \tag{28}$$

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 $\boldsymbol{\Sigma}_{\boldsymbol{XX}}^{1/2} \mathbf{A}_3 \boldsymbol{\Sigma}_{\boldsymbol{YY}}^{1/2} = \mathbf{UDV}^\top$, 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:

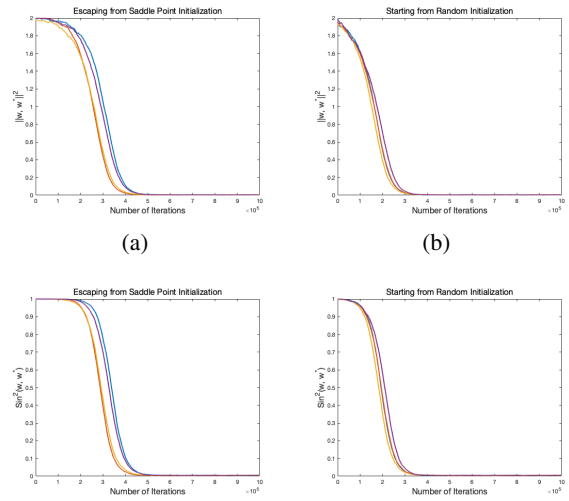


Figure 1: Comparison between saddle point initialization and random initialization

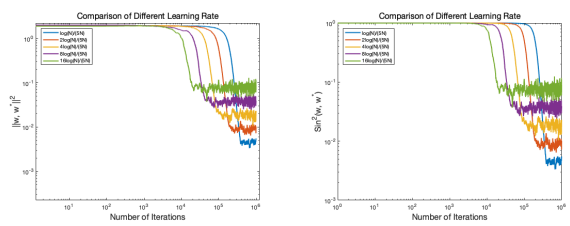
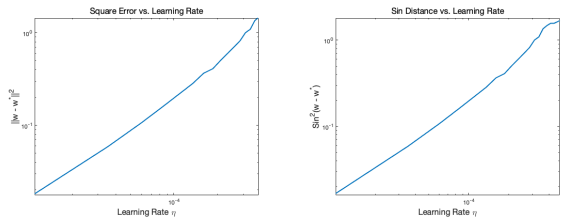


Figure 2: Log-log plot regarding the convergence with respect to a range of step sizes η. Figure 2(a) illustrates the squared errors in terms of squared distance to optimality $\|\boldsymbol{v} - \boldsymbol{v}^*\|^2$, and Figure 2(b) does so in terms of $\sin^2(\boldsymbol{v}, \boldsymbol{v}^*)$

Saddle-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 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 $\|\boldsymbol{v}_t - \boldsymbol{v}^*\|^2$.



(a) (b)

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

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 9. 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.

6 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.

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: Simple & efficient algorithm for streaming generalized eigenvector computation. In *Advances in Neural Information Processing Systems*, pages 7016–7025, 2018.

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

Kamalika Chaudhuri, Sham M Kakade, Karen Livescu, and Karthik Sridharan. Multi-view clustering via canonical correlation analysis. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 129–136. ACM, 2009.

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, Quansheng Liu, et al. 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.