Yokohama Publishers
ISSN 1349-8169 ONLINE JOURNAL

Pacific Journal of Optimization
Vol. 20, No. 4, pp. 743–767, 2024

CONVERGENCE OF RIEMANNIAN STOCHASTIC GRADIENT DESCENT ON HADAMARD MANIFOLD

Hiroyuki Sakai* and Hideaki Iiduka

Abstract: Riemannian stochastic gradient descent (RSGD) is the most basic Riemannian stochastic optimization algorithm and is used in many applications of machine learning. This study presents novel convergence analyses of RSGD on a Hadamard manifold that incorporate the mini-batch strategy used in deep learning and overcome several problems with the previous analyses. Four types of convergence analysis are described for both constant and diminishing step sizes. The number of steps needed for RSGD convergence is shown to be a convex monotone decreasing function of the batch size. Application of RSGD with several batch sizes to a Riemannian stochastic optimization problem on a symmetric positive-definite manifold theoretically shows that increasing the batch size improves RSGD performance. A numerical evaluation of the relationship between batch size and RSGD performance provides evidence supporting the theoretical results.

Key words: *Riemannian optimization, Hadamard manifolds, RSGD, critical batch size*

Mathematics Subject Classification: *65K05, 90C26, 57R35*

1 Introduction

Riemannian optimization has attracted a great deal of attention [1, 9, 22] in the field of machine learning. In this paper, we consider a Hadamard manifold, which is a complete Riemannian manifold whose sectional curvatures are less than or equal to zero. From the viewpoints of application and convergence analysis, optimization problems on a Hadamard manifold are very important. This is because, from the Cartan-Hadamard theorem, there exists an inverse map of the exponential map. Hyperbolic spaces (represented, for example, by the Poincaré ball model and Poincaré half-plane model) and symmetric positive definite (SPD) manifolds are examples of Hadamard manifolds with many applications of machine learning.

Optimization problems on SPD manifolds have a variety of applications in computer vision and machine learning. In particular, visual representations often rely on SPD manifolds, such as the kernel matrix, the covariance descriptor [27], and the diffusion tensor image [8]. Optimization problems on SPD manifolds are especially important in medical imaging [4, 8]. Furthermore, optimization problems in hyperbolic spaces are important in natural language processing. Nickel and Kiela [19] proposed using Poincaré embedding, an example of a Riemannian optimization problem in hyperbolic space. More specifically, they

*Corresponding author

 DOI: https://doi.org/10.61208/pjo-2024-005

proposed embedding hierarchical representations of symbolic data (e.g., text, graph data) into the Poincaré ball model or Poincaré half-plane model of hyperbolic space.

Riemannian stochastic optimization algorithms are used for solving these optimization problems on a Riemannian manifold. Various such algorithms have been developed by extending gradient-based optimization algorithms in Euclidean space. Bonnabel proposed Riemannian stochastic gradient descent (RSGD), which is the most basic Riemannian stochastic optimization algorithm [3]. Sato, Kasai, and Mishra proposed the Riemannian stochastic variance reduced gradient (RSVRG) algorithm with a retraction and vector transport [25]. Moreover, they gave a convergence analysis of RSVRG under certain reasonable assumptions. In general, the RSVRG algorithm converges to an optimal solution faster than RSGD; however, the full gradient needs to be calculated every few steps with RSVRG.

Adaptive optimization algorithms such as AdaGrad [6], Adadelta [28], Adam [16], and AMSGrad [21] are widely used for training deep neural networks in Euclidean space. However, they cannot be easily extended to general Riemannian manifolds due to the absence of a canonical coordinate system. Special measures must therefore be considered when extending them to Riemannian manifolds. Kasai, Jawanpuria, and Mishra proposed generalizing adaptive stochastic gradient algorithms to Riemannian matrix manifolds by adapting the row and column subspaces of the gradients [15]. Bécigneul and Ganea proposed the Riemannian AMSGrad (RAMSGrad) algorithm [2]; however, RAMSGrad is defined only on the *product* of Riemannian manifolds by regarding each component of the product Riemannian manifold as a coordinate component in Euclidean space.

Bonnabel presented two types of RSGD convergence analysis on a Hadamard manifold [3], but both of them are based on unrealistic assumptions, and use only diminishing step sizes. The first type is based on the assumption that the sequence generated by RSGD is contained in a compact set of a Hadamard manifold M. Since it is difficult to predict the complete sequence, this assumption should be removed. The second type is based on an unrealistic assumption regarding the step-size selection. Specifically, a function ($v : M \to \mathbb{R}$) determined by a Riemannian optimization problem must be computed, and a diminishing step size (α_k divided by $v(x_k)$, where x_k is a k-th approximation defined by RSGD) must be used. This is not a realistic assumption because the step size is determined by the Riemannian optimization problem to be solved and must be adapted *manually*.

In this paper, we improve the RSGD convergence analysis on a Hadamard manifold in accordance with the points mentioned above and present four types of convergence analysis for constant and diminishing step sizes (see Section 3). First, we consider the case in which an objective function $f : M \to \mathbb{R}$ is L-smooth (Definition 3.1). Theorems 3.4 and 3.5 are for convergence analyses with the L-smooth assumption for constant and diminishing step sizes, respectively. Since calculating the constant L in the definition of L-smooth is often difficult, we also present convergence analyses for the function $f : M \to \mathbb{R}$ *not* L-smooth. Theorems 3.8 and 3.9 support convergence analyses without the L-smooth assumption for constant and diminishing step sizes, respectively. Table 1 summarizes the existing and proposed analyses.

Moreover, we show that the number of steps K needed for an ε-approximation of RSGD is convex and monotone decreasing with respect to the batch size b. We also show that stochastic first-order oracle (SFO) complexity [24] (defined as Kb) is convex with respect to b and that there exists a critical batch size such that SFO complexity is minimized (see Section 3.4).

Our contributions are summarized as follows. First, we provide improved theoretical analyses of RSGD. The convergence analyses with constant learning rates are novel contributions. The convergence analyses with diminishing learning rates are with and without the L-smooth assumption. Second, we analyze how the number of steps and the SFO complexity vary with the choice of batch size. Finally, we perform numerical experiments to support our theoretical findings.

This paper is organized as follows. Section 2 reviews the fundamentals of Riemannian geometry and Riemannian optimization. Section 3 presents four novel convergence analyses of RSGD on a Hadamard manifold. Section 4 experimentally evaluates the performance of RSGD by solving the Riemannian centroid problem on an SPD manifold for several batch sizes and evaluates the relationship between the number of steps K and batch size b. Section 5 concludes the paper.

Table 1: Assumptions used in existing convergence analyses given by [3] and our novel convergence analyses (Theorems 3.4–3.9). In all cases, M is a Hadamard manifold, and $f : M \to \mathbb{R}$ is a smooth function.

Theorem	Assumptions		
	Step size α_k	Function f	Sequence x_k
[3] Theorem 2	Diminishing	–	$(x_k)_{k=0}^{\infty} \subset C$ C: compact set
[3] Theorem 3	Determined by problem	–	–
Theorem 3.4	Constant depending on L	L-smooth	–
Theorem 3.5	Diminishing not depending on L	L-smooth	–
Theorem 3.8	Constant	–	$(x_k)_{k=0}^{\infty} \subset C$ C: bounded set
Theorem 3.9	Diminishing	–	$(x_k)_{k=0}^{\infty} \subset C$ C: bounded set

2 Mathematical Preliminaries

Let $\mathbb{R}$ be the set of real numbers and $\mathbb{N}$ be the set of natural numbers (i.e., positive integers). We denote $[n] := \{1, 2, \cdots, n\}$ $(n \in \mathbb{N})$, $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$ and $\mathbb{R}_{++} := \{x \in \mathbb{R} : x > 0\}$. Let M be a Riemannian manifold and T_xM be a tangent space at $x \in M$. An exponential map at $x \in M$, written as $\mathrm{Exp}_x : T_xM \to M$, is a mapping from T_xM to M with the requirement that a vector $\xi_x \in T_xM$ is mapped to the point $y := \mathrm{Exp}_x(\xi_x) \in M$ such that there exists a geodesic $c : [0, 1] \to M$ that satisfies $c(0) = x$, $c(1) = y$, and $\dot{c}(0) = \xi_x$, where $\dot{c}$ is the derivative of c [23]. Let $\langle \cdot, \cdot \rangle_x$ be a Riemannian metric at $x \in M$ and $\|\cdot\|_x$ be the norm defined by the Riemann metric at $x \in M$. Let $d(\cdot, \cdot) : M \times M \to \mathbb{R}_{++} \cup \{0\}$ be the distance function on M. A complete simply connected Riemannian manifold of a nonpositive sectional curvature is called a Hadamard manifold.

2.1 Riemannian stochastic optimization problem

We define a Riemannian stochastic optimization problem and two standard conditions. Given a data point z in data domain Z, a Riemannian stochastic optimization problem provides a smooth loss function, $\ell(\cdot; z) : M \to \mathbb{R}$. We minimize the expected loss $f : M \to \mathbb{R}$

defined by

$$f(x) := \mathbb{E}_{z\sim\mathcal{D}}\left[\ell(x;z)\right] = \mathbb{E}\left[\ell_\xi(x)\right], \tag{2.1}$$

where $\mathcal{D}$ is a probability distribution over Z, ξ denotes a random variable with distribution function P, and $\mathbb{E}[\cdot]$ denotes the expectation taken with respect to ξ. We assume that an SFO exists such that, for a given $x \in M$, it returns the stochastic gradient $\mathsf{G}_\xi(x)$ of function f defined by (2.1), where a random variable ξ is supported on Ξ independently of x. The standard conditions (C1) and (C2) are assumed in the discussion hereafter:

C1 Let $(x_k)_{k=0}^{\infty} \subset M$ be the sequence generated by the algorithm. For each iteration k,

$$\mathbb{E}_{\xi_k}\left[\mathsf{G}_{\xi_k}(x_k)\right] = \operatorname{grad} f(x_k),$$

where $\xi_0, \xi_1, \cdots$ are independent samples, and the random variable ξ_k is independent of $(x_i)_{i=0}^{k}$. There exists a nonnegative constant σ^2 such that

$$\mathbb{E}_{\xi_k}\left[\|\mathsf{G}_{\xi_k}(x_k) - \operatorname{grad} f(x_k)\|_{x_k}^2\right] \le \sigma^2.$$

C2 For each iteration $k \in \mathbb{N}_0$, the optimizer samples a batch B_k of size b independently of k and estimates the full gradient $\operatorname{grad} f$ as

$$\operatorname{grad} f_{B_k}(x_k) := \frac{1}{b}\sum_{i\in[b]} \mathsf{G}_{\xi_{k,i}}(x_k),$$

where $\xi_{k,i}$ is a random variable generated by the i-th sampling in the k-th iteration.

$\mathbb{E}_{\xi_k}[\cdot|x_k]$ represents the conditional expectation with respect to the random variable ξ_k given x_k. We denote $\mathbb{E}_{\xi_k}[\cdot]$ and $\mathbb{E}_{\xi_k}[\cdot|x_k]$ simply as $\mathbb{E}[\cdot]$ and $\mathbb{E}[\cdot|x_k]$, respectively. From (C1) and (C2), we immediately have

$$\mathbb{E}\left[\operatorname{grad} f_{B_k}(x_k) \mid x_k\right] = \operatorname{grad} f(x_k), \tag{2.2}$$

$$\mathbb{E}\left[\|\operatorname{grad} f_{B_k}(x_k) - \operatorname{grad} f(x_k)\|_{x_k}^2 \;\middle|\; x_k\right] \le \frac{\sigma^2}{b}. \tag{2.3}$$

Bonnabel [3] proposed RSGD for solving Riemannian optimization problems. In this paper, we use RSGD with a variable batch, as shown in Algorithm 1.

Algorithm 1 Riemannian stochastic gradient descent [3].

Require: Initial point $x_0 \in M$, step sizes $(\alpha_k)_{k=0}^{\infty} \subset \mathbb{R}_{++}$, batch size $b \in \mathbb{N}$.
Ensure: Sequence $(x_k)_{k=0}^{\infty} \subset M$.

```
k ← 0.
loop
    η_k := −grad f_{B_k}(x_k) = −b^{−1} Σ_{i∈[b]} G_{ξ_{k,i}}(x_k).
    x_{k+1} := Exp_{x_k}(α_k η_k).
    k ← k + 1.
end loop
```

$\boxed{2.2}$ Useful Lemma

The following lemma plays an important role in our discussion of the convergence of Riemannian stochastic gradient descent on a Hadamard manifold in Section 3.

Lemma 2.1. *Suppose that (C1) and (C2) and M define a Riemannian manifold and that $f : M \to \mathbb{R}$ is a smooth function on M. Then, the sequence $(x_k)_{k=0}^{\infty} \subset M$ generated by Algorithm 1 satisfies*

$$\mathbb{E}\left[\|\operatorname{grad} f_{B_k}(x_k)\|_{x_k}^2 \;\middle|\; x_k\right] \leq \frac{\sigma^2}{b} + \|\operatorname{grad} f(x_k)\|_{x_k}^2$$

for all $k \in \mathbb{N}_0$.

Proof. Using (2.2) and (2.3), we obtain

$$\begin{aligned}
\mathbb{E}\left[\|\operatorname{grad} f_{B_k}(x_k)\|_{x_k}^2 \;\middle|\; x_k\right] &= \mathbb{E}\left[\|\operatorname{grad} f_{B_k}(x_k) - \operatorname{grad} f(x_k) + \operatorname{grad} f(x_k)\|_{x_k}^2 \;\middle|\; x_k\right] \\
&= \mathbb{E}\left[\|\operatorname{grad} f_{B_k}(x_k) - \operatorname{grad} f(x_k)\|_{x_k}^2 \;\middle|\; x_k\right] \\
&\quad + 2\mathbb{E}\left[\langle \operatorname{grad} f_{B_k}(x_k) - \operatorname{grad} f(x_k), \operatorname{grad} f(x_k)\rangle_{x_k} \;\middle|\; x_k\right] \\
&\quad + \mathbb{E}\left[\|\operatorname{grad} f(x_k)\|_{x_k}^2 \;\middle|\; x_k\right] \\
&\leq \frac{\sigma^2}{b} + \|\operatorname{grad} f(x_k)\|_{x_k}^2
\end{aligned}$$

for all $k \in \mathbb{N}_0$. ☐

Lemma 2.2 is useful in showing convergence of the limit inferior.

Lemma 2.2. *The sequences $(\alpha_k)_{k=0}^{\infty} \subset \mathbb{R}_{++}$ and $(\beta_k)_{k=0}^{\infty} \subset \mathbb{R}$ such that*

$$\sum_{k=0}^{+\infty} \alpha_k = +\infty, \quad \sum_{k=0}^{+\infty} \alpha_k \beta_k < +\infty$$

satisfy

$$\liminf_{k \to +\infty} \beta_k \leq 0.$$

Zhang and Sra developed the following lemma [30].

Lemma 2.3. *Let a, b, and c be the side lengths of a geodesic triangle in a Riemannian manifold with a sectional curvature lower bounded by κ, and let θ be the angle between sides b and c. Then,*

$$a^2 \leq \zeta(\kappa, c) b^2 + c^2 - 2bc\cos(\theta),$$

where $\zeta : \mathbb{R} \times \mathbb{R}_{++} \to \mathbb{R}_{++}$ is defined as

$$\zeta(\kappa, c) := \frac{\sqrt{|\kappa|}c}{\tanh(\sqrt{|\kappa|}\, c)}.$$

3 Convergence of Riemannian Stochastic Gradient Descent on Hadamard Manifold

In this section, we describe four types of convergence analysis of Algorithm 1 on a Hadamard manifold.

3.1 Convergence of Riemannian Stochastic Gradient Descent with L-smoothness

First, we describe the convergence of Algorithm 1 when L-smoothness is assumed. We start by defining the L-smoothness of a smooth function [11, 30].

Definition 3.1 (smoothness)**.** Let M be a Hadamard manifold and $f : M \to \mathbb{R}$ be a smooth function on M. For a positive number $L \in \mathbb{R}_{++}$, f is said to be geodesically L-smooth if for any $x, y \in M$,

$$\left\|\operatorname{grad} f(x) - \Gamma_y^x(\operatorname{grad} f(y))\right\|_x \leq L \left\|\operatorname{Exp}_x^{-1}(y)\right\|_x,$$

where $\Gamma_y^x : T_yM \to T_xM$ is the parallel transport from y to x.

We state the following lemma giving the necessary conditions for L-smooth [30]. Lemma 3.2 plays an important role in convergence analysis with L-smoothness.

Lemma 3.2. *Let M be a Hadamard manifold and $f : M \to \mathbb{R}$ be a smooth function on M. If f is geodesically L-smooth, it follows that for all $x, y \in M$,*

$$f(y) \leq f(x) + \left\langle \operatorname{grad} f(x), \operatorname{Exp}_x^{-1}(y)\right\rangle_x + \frac{L}{2}\left\|\operatorname{Exp}_x^{-1}(y)\right\|_x.$$

To show the main result of this section (i.e., Theorems 3.4 and 3.5), we present the following lemma, which plays a central role.

Lemma 3.3. *Let M be a Hadamard manifold and $f : M \to \mathbb{R}$ be a smooth function. We assume that f is geodesically L-smooth and bounded below by $f_\star \in \mathbb{R}$. Then, the sequence $(x_k)_{k=0}^\infty \subset M$ generated by Algorithm 1 satisfies*

$$\sum_{k=0}^{K-1} \alpha_k \left(1 - \frac{L\alpha_k}{2}\right) \mathbb{E}\left[\|\operatorname{grad} f(x_k)\|_{x_k}^2\right] \leq f(x_0) - f_\star + \frac{L\sigma^2}{2b}\sum_{k=0}^{K-1}\alpha_k^2$$

for all $K \in \mathbb{N}$.

Proof. From the L-smoothness of f and $x_{k+1} = \operatorname{Exp}_{x_k}(\alpha_k \eta_k)$, we have

$$f(x_{k+1}) \leq f(x_k) - \alpha_k \left\langle \operatorname{grad} f(x_k), \operatorname{grad} f_{B_k}(x_k)\right\rangle_{x_k} + \frac{L\alpha_k^2}{2}\|\operatorname{grad} f_{B_k}(x_k)\|_{x_k}^2 \tag{3.1}$$

for all $k \in \mathbb{N}_0$. From (2.2), we obtain

$$\begin{aligned}\mathbb{E}\left[\left\langle \operatorname{grad} f(x_k), \operatorname{grad} f_{B_k}(x_k)\right\rangle_{x_k} \mid x_k\right] &= \left\langle \operatorname{grad} f(x_k), \mathbb{E}\left[\operatorname{grad} f_{B_k}(x_k) \mid x_k\right]\right\rangle_{x_k} \\ &= \|\operatorname{grad} f(x_k)\|_{x_k}^2 \end{aligned}\tag{3.2}$$

for all $k \in \mathbb{N}_0$. Hence, by taking $\mathbb{E}[\cdot|x_k]$ of both sides of (3.1), we obtain

$$\begin{aligned}\mathbb{E}\left[f(x_{k+1}) \mid x_k\right] &\leq \mathbb{E}\left[f(x_k) \mid x_k\right] - \alpha_k \mathbb{E}\left[\langle \operatorname{grad} f(x_k), \operatorname{grad} f_{B_k}(x_k)\rangle_{x_k} \;\middle|\; x_k\right] \\ &\quad + \frac{L\alpha_k^2}{2}\mathbb{E}\left[\|\operatorname{grad} f_{B_k}(x_k)\|_{x_k}^2 \;\middle|\; x_k\right] \\ &\leq f(x_k) - \alpha_k \|\operatorname{grad} f(x_k)\|^2 + \frac{L\alpha_k^2}{2}\left(\frac{\sigma^2}{b} + \|\operatorname{grad} f(x_k)\|_{x_k}^2\right) \\ &= f(x_k) + \left(\frac{L\alpha_k}{2} - 1\right)\alpha_k \|\operatorname{grad} f(x_k)\|_{x_k}^2 + \frac{L\sigma^2\alpha_k^2}{2b}\end{aligned}$$

for all $k \in \mathbb{N}_0$, where the second inequality comes from (3.2) and Lemma 2.1. Moreover, by taking $\mathbb{E}[\cdot]$ of both sides, we obtain

$$\mathbb{E}\left[f(x_{k+1})\right] \leq \mathbb{E}\left[f(x_k)\right] + \left(\frac{L\alpha_k}{2} - 1\right)\alpha_k \mathbb{E}\left[\|\operatorname{grad} f(x_k)\|_{x_k}^2\right] + \frac{L\sigma^2\alpha_k^2}{2b}$$

for all $k \in \mathbb{N}_0$. By summing up the above inequalities from $k = 0$ to $k = K - 1$, we obtain

$$\begin{aligned}\sum_{k=0}^{K-1}\left(1 - \frac{L\alpha_k}{2}\right)\alpha_k \mathbb{E}\left[\|\operatorname{grad} f(x_k)\|_{x_k}^2\right] &\leq f(x_0) - \mathbb{E}[f(x_K)] + \frac{L\sigma^2}{2b}\sum_{k=0}^{K-1}\alpha_k^2 \\ &\leq f(x_0) - f_\star + \frac{L\sigma^2}{2b}\sum_{k=0}^{K-1}\alpha_k^2\end{aligned}$$

for all $K \in \mathbb{N}$. This completes the proof. ☐

We can use Lemma 3.3 to conduct a convergence analysis for a constant step size under the assumption of L-smoothness.

Theorem 3.4. *Let M be a Hadamard manifold and $f : M \to \mathbb{R}$ be a smooth function. We assume that f is geodesically L-smooth and bounded below by $f_\star \in \mathbb{R}$. If a constant step size $\alpha_k := \alpha$ ($k \in \mathbb{N}_0$) satisfies $0 < \alpha < 2/L$, the sequence $(x_k)_{k=0}^\infty \subset M$ generated by Algorithm 1 satisfies,*

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\|\operatorname{grad} f(x_k)\|_{x_k}^2\right] \leq \frac{C_1}{K} + \frac{C_2\sigma^2}{b}, \tag{3.3}$$

for some $C_1, C_2 \in \mathbb{R}_{++}$ and for all $K \in \mathbb{N}$.

Proof. From Lemma 3.3, we have

$$\sum_{k=0}^{K-1}\alpha\left(1 - \frac{L\alpha}{2}\right)\mathbb{E}\left[\|\operatorname{grad} f(x_k)\|_{x_k}^2\right] \leq f(x_0) - f_\star + \frac{L\sigma^2}{2b}\sum_{k=0}^{K-1}\alpha^2$$

for all $K \in \mathbb{N}$. Moreover, from $0 < \alpha < 2/L$, we have

$$0 < \frac{L\alpha}{2} < 1,$$

which implies that

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\|\operatorname{grad} f(x_k)\|_{x_k}^2\right] \leq \underbrace{\frac{2(f(x_0) - f_\star)}{(2 - L\alpha)\alpha}}_{C_1}\cdot\frac{1}{K} + \underbrace{\frac{L\alpha}{(2 - L\alpha)}}_{C_2}\cdot\frac{\sigma^2}{b},$$

for all $K \in \mathbb{N}$. This completes the proof. ☐

Moreover, we can use Lemma 3.3 to conduct a convergence analysis with a diminishing step size under the assumption of L-smoothness.

Theorem 3.5. *Let M be a Hadamard manifold and $f : M \to \mathbb{R}$ be a smooth function. We assume that f is geodesically L-smooth and bounded below by $f_\star \in \mathbb{R}$ and use a diminishing step size $(\alpha_k)_{k=0}^{\infty} \subset \mathbb{R}_{++}$ that satisfies*

$$\sum_{k=0}^{+\infty} \alpha_k = +\infty, \quad \sum_{k=0}^{+\infty} \alpha_k^2 < +\infty. \tag{3.4}$$

Then, the sequence $(x_k)_{k=0}^{\infty} \subset M$ generated by Algorithm 1 satisfies

$$\liminf_{k\to+\infty} \mathbb{E}\left[\|\operatorname{grad} f(x_k)\|_{x_k}\right] = 0. \tag{3.5}$$

If the diminishing step size $(\alpha_k)_{k=0}^{\infty} \subset (0,1)$ is monotonically decreasing, then for all $K \in \mathbb{N}$,

$$\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\left[\|\operatorname{grad} f(x_k)\|_{x_k}^2\right] \le \left(C_1 + \frac{C_2\sigma^2}{b}\sum_{k=0}^{K-1}\alpha_k^2\right)\frac{1}{K\alpha_{K-1}} \tag{3.6}$$

for some $C_1, C_2 \in \mathbb{R}_{++}$.

Proof. From (3.4), we obtain

$$\sum_{k=0}^{+\infty} \alpha_k \left(1 - \frac{L\alpha_k}{2}\right) = +\infty. \tag{3.7}$$

In addition, from (3.4), $(\alpha_k)_{k=0}^{\infty}$ satisfies $\alpha_k \to 0$ $(k \to +\infty)$. This implies that there exists a natural number $k_0 \in \mathbb{N}_0$ such that, for all $k \in \mathbb{N}_0$, if $k \ge k_0$, then $0 < \alpha_k < 2/L$. Therefore, we obtain

$$0 < 1 - \frac{L\alpha_k}{2} < 1,$$

which, together with Lemma 3.3, means that

$$\begin{aligned}
&\sum_{k=k_0}^{K-1} \alpha_k \left(1 - \frac{L\alpha_k}{2}\right) \mathbb{E}\left[\|\operatorname{grad} f(x_k)\|_{x_k}^2\right] \\
&\le f(x_0) - f_\star + \frac{L\sigma^2}{2b}\sum_{k=0}^{K-1}\alpha_k^2 - \sum_{k=0}^{k_0-1}\alpha_k\left(1 - \frac{L\alpha_k}{2}\right)\mathbb{E}\left[\|\operatorname{grad} f(x_k)\|_{x_k}^2\right]
\end{aligned} \tag{3.8}$$

for all $K \ge k_0 + 1$. This implies

$$\sum_{k=0}^{+\infty} \alpha_k \left(1 - \frac{L\alpha_k}{2}\right) \mathbb{E}\left[\|\operatorname{grad} f(x_k)\|_{x_k}^2\right] < +\infty. \tag{3.9}$$

By applying Lemma 2.2 with (3.7) and (3.9), we have

$$\liminf_{k\to+\infty} \mathbb{E}\left[\|\operatorname{grad} f(x_k)\|_{x_k}^2\right] \le 0,$$

In this case, $(\alpha_k)_{k=0}^{\infty} \subset (0,1)$ need not satisfy the conditions (3.4).

which, together with the convexity of $\|\cdot\|_{x_k}$, means that

$$0 \leq \liminf_{k\to+\infty} \left(\mathbb{E}\left[\|\operatorname{grad} f(x_k)\|_{x_k}\right]\right)^2 \leq \liminf_{k\to+\infty} \mathbb{E}\left[\|\operatorname{grad} f(x_k)\|_{x_k}^2\right] \leq 0.$$

This implies

$$\liminf_{k\to+\infty} \mathbb{E}\left[\|\operatorname{grad} f(x_k)\|_{x_k}\right] = 0$$

and ensures that (3.5) follows from the above discussion.

Furthermore, we show that (3.6) follows from the above discussion. From the monotonicity of $(\alpha_k)_{k=0}^{\infty} \subset (0,1)$ and (3.8), we have

$$\begin{aligned}
&\alpha_{K-1}\left(1 - \frac{L\alpha_{k_0}}{2}\right)\sum_{k=k_0}^{K-1} \mathbb{E}\left[\|\operatorname{grad} f(x_k)\|_{x_k}^2\right] \\
&\leq f(x_0) - f_\star + \frac{L\sigma^2}{2b}\sum_{k=0}^{K-1}\alpha_k^2 + \sum_{k=0}^{k_0-1}\alpha_k\left(\frac{L\alpha_k}{2} - 1\right)\mathbb{E}\left[\|\operatorname{grad} f(x_k)\|_{x_k}^2\right] \\
&\leq f(x_0) - f_\star + \frac{L\sigma^2}{2b}\sum_{k=0}^{K-1}\alpha_k^2 + \sum_{k=0}^{k_0-1}L\alpha_k^2\mathbb{E}\left[\|\operatorname{grad} f(x_k)\|_{x_k}^2\right]
\end{aligned}$$

for all $K \in \mathbb{N}$. Hence, for all $K \in \mathbb{N}$,

$$\begin{aligned}
\sum_{k=k_0}^{K-1} \mathbb{E}\left[\|\operatorname{grad} f(x_k)\|_{x_k}^2\right] \leq{}& \frac{2\left(f(x_0) - f_\star + \sum_{k=0}^{k_0-1} L\alpha_k^2\mathbb{E}\left[\|\operatorname{grad} f(x_k)\|_{x_k}^2\right]\right)}{(2 - L\alpha_{k_0})\alpha_{K-1}} \\
&+ \frac{L\sigma^2}{b(2 - L\alpha_{k_0})\alpha_{K-1}}\sum_{k=0}^{K-1}\alpha_k^2,
\end{aligned}$$

which, together with $\alpha_k \in (0,1)$ $(k \in \mathbb{N}_0)$, gives us

$$\begin{aligned}
&\frac{1}{K}\sum_{k=k_0}^{K-1} \mathbb{E}\left[\|\operatorname{grad} f(x_k)\|_{x_k}^2\right] \\
&\leq \underbrace{\left\{\frac{2\left(f(x_0) - f_\star + \sum_{k=0}^{k_0-1} L\alpha_k^2\mathbb{E}\left[\|\operatorname{grad} f(x_k)\|_{x_k}^2\right]\right)}{2 - L\alpha_{k_0}} + \sum_{k=0}^{k_0-1}\mathbb{E}\left[\|\operatorname{grad} f(x_k)\|_{x_k}^2\right]\right\}}_{C_1}\frac{1}{K\alpha_{K-1}} \\
&\quad + \underbrace{\frac{L}{2 - L\alpha_{k_0}}}_{C_2} \cdot \frac{\sigma^2}{bK\alpha_{K-1}}\sum_{k=0}^{K-1}\alpha_k^2
\end{aligned}$$

for all $K \in \mathbb{N}$. This completes the proof. □

$\boxed{3.2}$ Convergence of Riemannian Stochastic Gradient Descent without L-smoothness

Next, we describe the convergence of Algorithm 1 without the assumption of L-smoothness. We start by reconsidering the definition of convergence of Algorithm 1. Németh [18] developed the variational inequality problem on a Hadamard manifold. Motivated by the variational inequality problem in Euclidean space [12,13], we undertook the following proposition.

Proposition 3.6. *Let M be a Riemannian manifold and $f : M \to \mathbb{R}$ be a smooth function. Then, a stationary point $x \in M$ of f satisfies*

$$\operatorname{grad} f(x) = 0 \Leftrightarrow \forall y \in V_x,\ \left\langle \operatorname{grad} f(x), -\operatorname{Exp}_x^{-1}(y) \right\rangle_x \leq 0,$$

where V_x is a neighborhood of $x \in M$ such that $\operatorname{Exp}_x^{-1} : V_x \to T_x M$ is defined.

Proof. If $x \in M$ satisfies $\operatorname{grad} f(x) = 0$, we have

$$\left\langle \operatorname{grad} f(x), -\operatorname{Exp}_x^{-1}(y) \right\rangle_x = \left\langle 0_x, -\operatorname{Exp}_x^{-1}(y) \right\rangle_x \leq 0$$

for all $y \in V_x$. We assume that $x \in M$ satisfies $\left\langle \operatorname{grad} f(x), -\operatorname{Exp}_x^{-1}(y) \right\rangle_x \leq 0$ for all $y \in M$. Let $y := \operatorname{Exp}_x(-\varepsilon \operatorname{grad} f(x))$, for which we choose a sufficiently small $\varepsilon > 0$ such that $y \in V_x$. We then have

$$\begin{aligned}
\left\langle \operatorname{grad} f(x), -\operatorname{Exp}_x^{-1}(y) \right\rangle_x &= \left\langle \operatorname{grad} f(x), -\operatorname{Exp}_x^{-1}(\operatorname{Exp}_x(-\varepsilon \operatorname{grad} f(x))) \right\rangle_x \\
&= \langle \operatorname{grad} f(x), \varepsilon \operatorname{grad} f(x) \rangle_x \\
&= \varepsilon \left\| \operatorname{grad} f(x) \right\|_x^2 \leq 0.
\end{aligned}$$

This implies that $\operatorname{grad} f(x) = 0$ and completes the proof. $\square$

Note that, for a Hadamard manifold, $V_x = M$. From Proposition 3.6, we use the performance measure of the sequence $(x_k)_{k=0}^{\infty}$,

$$V_k(x) := \mathbb{E}\left[\left\langle \operatorname{grad} f(x_k), -\operatorname{Exp}_{x_k}^{-1}(x) \right\rangle_{x_k} \right],$$

for all $x \in M$. In practice, we can use $V_k(x)$ for showing the convergence of Algorithm 1 in Theorems 3.8 and 3.9. To show the main result of this section (i.e., Theorems 3.8 and 3.9), we need the following lemma.

Lemma 3.7. *Let M be a Hadamard manifold with a sectional curvature lower bounded by κ and let $f : M \to \mathbb{R}$ be a smooth function. Then, the sequence $(x_k)_{k=0}^{\infty} \subset M$ generated by Algorithm 1 satisfies for all $K \in \mathbb{N}$ and $x \in M$,*

$$\begin{aligned}
\sum_{k=0}^{K-1} \alpha_k V_k(x) \leq{}& \frac{\zeta(\kappa, D(x))}{2} \sum_{k=0}^{K-1} \alpha_k^2 \left(\frac{\sigma^2}{b} + \mathbb{E}\left[\left\| \operatorname{grad} f(x_k) \right\|_{x_k}^2 \right] \right) \\
&+ \frac{1}{2} \left(\mathbb{E}\left[\left\| \operatorname{Exp}_{x_0}^{-1}(x) \right\|_{x_0}^2 \right] - \mathbb{E}\left[\left\| \operatorname{Exp}_{x_K}^{-1}(x) \right\|_{x_K}^2 \right] \right),
\end{aligned}$$

where $\zeta : \mathbb{R}_{++} \times \mathbb{R}_{++} \to \mathbb{R}_{++}$ is defined as in Lemma 2.3.

Proof. For arbitrary $x \in M$, we consider a geodesic triangle consisting of three points, x_k, x_{k+1}, and x. Let the length of each side be a, b, and c, respectively, such that

$$\begin{cases} a := d(x_{k+1}, x) \\ b := d(x_k, x_{k+1}) = \alpha_k \left\| \operatorname{grad} f_{B_k}(x_k) \right\|_{x_k} \\ c := d(x_k, x). \end{cases} \tag{3.10}$$

Let $\theta \in \mathbb{R}$ be the angle between sides b and c. It then follows that

$$\cos(\theta) := \frac{\left\langle \operatorname{Exp}_{x_k}^{-1}(x_{k+1}), \operatorname{Exp}_{x_k}^{-1}(x) \right\rangle_{x_k}}{\left\| \operatorname{Exp}_{x_k}^{-1}(x_{k+1}) \right\|_{x_k} \left\| \operatorname{Exp}_{x_k}^{-1}(x) \right\|_{x_k}}.$$

From Lemma 2.3 with (3.10), we have

$$\begin{aligned}\left\|\mathrm{Exp}_{x_{k+1}}^{-1}(x)\right\|_{x_{k+1}}^2 &\leq \alpha_k^2 \zeta\left(\kappa, D(x)\right) \left\|\mathrm{grad}\, f_{B_k}(x_k)\right\|_{x_k}^2 \\ &\quad - 2\alpha_k \left\langle \mathrm{grad}\, f_{B_k}(x_k), -\mathrm{Exp}_{x_k}^{-1}(x)\right\rangle_{x_k} + \left\|\mathrm{Exp}_{x_k}^{-1}(x)\right\|_{x_k}^2 .\end{aligned}$$

By taking $\mathbb{E}[\cdot|x_k]$ of both sides of this inequality, we obtain

$$\begin{aligned}\mathbb{E}\left[\left\|\mathrm{Exp}_{x_{k+1}}^{-1}(x)\right\|_{x_{k+1}}^2 \,\middle|\, x_k\right] &\leq \alpha_k^2 \zeta\left(\kappa, D(x)\right) \mathbb{E}\left[\left\|\mathrm{grad}\, f_{B_k}(x_k)\right\|_{x_k}^2 \,\middle|\, x_k\right] \\ &\quad - 2\alpha_k \mathbb{E}\left[\left\langle \mathrm{grad}\, f_{B_k}(x_k), -\mathrm{Exp}_{x_k}^{-1}(x)\right\rangle_{x_k} \,\middle|\, x_k\right] \\ &\quad + \mathbb{E}\left[\left\|\mathrm{Exp}_{x_k}^{-1}(x)\right\|_{x_k}^2 \,\middle|\, x_k\right] \\ &\leq \alpha_k^2 \zeta\left(\kappa, D(x)\right) \left(\frac{\sigma^2}{b} + \left\|\mathrm{grad}\, f(x_k)\right\|_{x_k}^2\right) \\ &\quad - 2\alpha_k \left\langle \mathrm{grad}\, f(x_k), -\mathrm{Exp}_{x_k}^{-1}(x)\right\rangle_{x_k} \\ &\quad + \left\|\mathrm{Exp}_{x_k}^{-1}(x)\right\|_{x_k}^2\end{aligned}$$

for all $k \in \mathbb{N}_0$, where the second inequality comes from Lemma 2.1. Furthermore, by taking $\mathbb{E}[\cdot]$ of both sides, we obtain

$$\begin{aligned}\mathbb{E}\left[\left\|\mathrm{Exp}_{x_{k+1}}^{-1}(x)\right\|_{x_{k+1}}^2\right] &\leq \alpha_k^2 \zeta\left(\kappa, D(x)\right) \left(\frac{\sigma^2}{b} + \mathbb{E}\left[\left\|\mathrm{grad}\, f(x_k)\right\|_{x_k}^2\right]\right) \\ &\quad - 2\alpha_k V_k(x) + \mathbb{E}\left[\left\|\mathrm{Exp}_{x_k}^{-1}(x)\right\|_{x_k}^2\right]\end{aligned} \tag{3.11}$$

for all $k \in \mathbb{N}_0$. Hence,

$$\begin{aligned}\alpha_k V_k(x) &\leq \frac{1}{2}\left(\mathbb{E}\left[\left\|\mathrm{Exp}_{x_k}^{-1}(x)\right\|_{x_k}^2\right] - \mathbb{E}\left[\left\|\mathrm{Exp}_{x_{k+1}}^{-1}(x)\right\|_{x_{k+1}}^2\right]\right) \\ &\quad + \frac{\alpha_k^2 \zeta\left(\kappa, D(x)\right)}{2}\left(\frac{\sigma^2}{b} + \mathbb{E}\left[\left\|\mathrm{grad}\, f(x_k)\right\|_{x_k}^2\right]\right)\end{aligned}$$

for all $k \in \mathbb{N}_0$. By summing up the above inequalities from $k = 0$ to $k = K - 1$ ($K \in \mathbb{N}$), we obtain

$$\begin{aligned}\sum_{k=0}^{K-1} \alpha_k V_k(x) &\leq \frac{\zeta\left(\kappa, D(x)\right)}{2} \sum_{k=0}^{K-1} \alpha_k^2 \left(\frac{\sigma^2}{b} + \mathbb{E}\left[\left\|\mathrm{grad}\, f(x_k)\right\|_{x_k}^2\right]\right) \\ &\quad + \frac{1}{2}\left(\mathbb{E}\left[\left\|\mathrm{Exp}_{x_0}^{-1}(x)\right\|_{x_0}^2\right] - \mathbb{E}\left[\left\|\mathrm{Exp}_{x_K}^{-1}(x)\right\|_{x_K}^2\right]\right)\end{aligned}$$

for all $K \in \mathbb{N}$. This completes the proof. □

Here, we make the following assumptions:

Assumption 3.1. Let M be a Hadamard manifold, $f : M \to \mathbb{R}$ be a smooth function, and $(x_k)_{k=0}^{\infty} \subset M$ be a sequence generated by Algorithm 1.

(A1) We assume that there exists a positive number $G \in \mathbb{R}_{++}$ such that

$$\mathbb{E}\left[\|\operatorname{grad} f(x_k)\|_{x_k}\right] \leq G < +\infty$$

for all $k \in \mathbb{N}_0$.

(A2) We define $D : M \to \mathbb{R}$ as

$$D(x) := \sup\left\{\mathbb{E}\left[d(x_k, x)\right] \in \mathbb{R}_{++} : k \in \mathbb{N}_0\right\}$$

and assume that $D(x) < +\infty$ for all $x \in M$.

We can use Lemma 3.7 to conduct convergence analysis for a constant step size.

Theorem 3.8. *Suppose Assumption 3.1, and let M be a Hadamard manifold with a sectional curvature lower bounded by κ and $f : M \to \mathbb{R}$ be a smooth function. If we use a constant step size $\alpha_k := \alpha > 0$ ($k \in \mathbb{N}_0$), the sequence $(x_k)_{k=0}^{\infty} \subset M$ generated by Algorithm 1 satisfies*

$$\liminf_{k \to +\infty} V_k(x) \leq \left(\frac{\sigma^2}{b} + G^2\right) \alpha C \tag{3.12}$$

for some $C \in \mathbb{R}_{++}$. Moreover, for all $K \in \mathbb{N}$ and $x \in M$,

$$\frac{1}{K} \sum_{k=0}^{K-1} V_k(x) \leq \left(\frac{\sigma^2}{b} + G^2\right) \alpha C_1 + \frac{C_2}{K} \tag{3.13}$$

for some $C_1, C_2 \in \mathbb{R}_{++}$.

Proof. If $\operatorname{grad} f(x_{k_0}) = 0$ for some $k_0 \in \mathbb{N}_0$, then (3.12) follows. Thus, it is sufficient to prove (3.12) only when $\operatorname{grad} f(x_k) \neq 0$ for all $k \in \mathbb{N}_0$. Suppose that there exists a positive number $\varepsilon \in \mathbb{R}_{++}$ such that

$$\liminf_{k \to +\infty} V_k(x) > \frac{\zeta(\kappa, D(x))}{2} \left(\frac{\sigma^2}{b} + G^2\right) \alpha + \varepsilon \tag{3.14}$$

for all $x \in M$. Furthermore, from the definition of the limit inferior, there exists $k_0 \in \mathbb{N}_0$ such that, for all $k \geq k_0$,

$$\liminf_{k \to +\infty} V_k(x) - \frac{\varepsilon}{2} < V_k(x),$$

from which, together with (3.14), we obtain

$$V_k(x) > \frac{\zeta(\kappa, D(x))}{2} \left(\frac{\sigma^2}{b} + G^2\right) \alpha + \frac{\varepsilon}{2}$$

for all $k \geq k_0$. Here, from (3.11) and $\alpha_k = \alpha$ $(k \in \mathbb{N}_0)$, then for all $k \geq k_0$,

$$\begin{aligned}\mathbb{E}\left[\left\|\mathrm{Exp}_{x_{k+1}}^{-1}(x)\right\|_{x_{k+1}}^2\right] &\leq \alpha^2 \zeta(\kappa, D(x)) \left(\frac{\sigma^2}{b} + \mathbb{E}\left[\|\mathrm{grad}\, f(x_k)\|_{x_k}^2\right]\right) \\ &\quad - 2\alpha V_k(x) + \mathbb{E}\left[\left\|\mathrm{Exp}_{x_k}^{-1}(x)\right\|_{x_k}^2\right], \\ &< \mathbb{E}\left[\left\|\mathrm{Exp}_{x_k}^{-1}(x)\right\|_{x_k}^2\right], \\ &\quad + \alpha^2 \zeta(\kappa, D(x)) \left(\frac{\sigma^2}{b} + G^2\right) \\ &\quad - 2\alpha \left\{\frac{\zeta(\kappa, D(x))}{2}\left(\frac{\sigma^2}{b} + G^2\right)\alpha + \frac{\varepsilon}{2}\right\} \\ &= \mathbb{E}\left[\left\|\mathrm{Exp}_{x_k}^{-1}(x)\right\|_{x_k}^2\right] - \alpha\varepsilon.\end{aligned}$$

Hence,

$$0 \leq \mathbb{E}\left[\left\|\mathrm{Exp}_{x_{k+1}}^{-1}(x)\right\|_{x_{k+1}}^2\right] < \mathbb{E}\left[\left\|\mathrm{Exp}_{x_{k_0}}^{-1}(x)\right\|_{x_{k_0}}^2\right] - \alpha\varepsilon(k + 1 - k_0). \tag{3.15}$$

When k diverges to $+\infty$, the right side of (3.15) diverges to $-\infty$. By contradiction, we have

$$\liminf_{k \to +\infty} V_k(x) \leq \underbrace{\frac{\zeta(\kappa, D(x))}{2}}_{C} \left(\frac{\sigma^2}{b} + G^2\right)\alpha.$$

Next, we show the upper bound of $(1/K)\sum_{k=0}^{K-1} V_k(x)$ such as expressed by (3.13). From Lemma 3.7 with $\alpha_k = \alpha$ $(k \in \mathbb{N}_0)$, we have that, for all $K \in \mathbb{N}$,

$$\begin{aligned}\sum_{k=0}^{K-1} \alpha V_k(x) &\leq \frac{\zeta(\kappa, D(x))}{2} \sum_{k=0}^{K-1} \alpha^2 \left(\frac{\sigma^2}{b} + \mathbb{E}\left[\|\mathrm{grad}\, f(x_k)\|_{x_k}^2\right]\right) \\ &\quad + \frac{1}{2}\left(\mathbb{E}\left[\left\|\mathrm{Exp}_{x_0}^{-1}(x)\right\|_{x_0}^2\right] - \mathbb{E}\left[\left\|\mathrm{Exp}_{x_K}^{-1}(x)\right\|_{x_K}^2\right]\right) \\ &\leq \frac{\zeta(\kappa, D(x))}{2}\left(\frac{\sigma^2}{b} + G^2\right) K\alpha^2 + \frac{D(x)}{2},\end{aligned}$$

which implies

$$\frac{1}{K}\sum_{k=0}^{K-1} V_k(x) \leq \underbrace{\frac{\zeta(\kappa, D(x))}{2}}_{C_1}\left(\frac{\sigma^2}{b} + G^2\right)\alpha + \underbrace{\frac{D(x)}{2\alpha}}_{C_2} \cdot \frac{1}{K}$$

for all $K \in \mathbb{N}$. This completes the proof. □

Now let us use Lemma 3.7 to conduct a convergence analysis for a diminishing step size.

Theorem 3.9. *Suppose Assumption 3.1, and let M be a Hadamard manifold with sectional curvature lower bounded by κ and $f : M \to \mathbb{R}$ be a smooth function. If we use a diminishing step size $(\alpha_k)_{k=0}^{\infty} \subset \mathbb{R}_{++}$ such as (3.4), the sequence $(x_k)_{k=0}^{\infty} \subset M$ generated by Algorithm 1 satisfies*

$$\liminf_{k \to +\infty} V_k(x) \leq 0. \tag{3.16}$$

If the diminishing step size $(\alpha_k)_{k=0}^{\infty} \subset \mathbb{R}_{++}$ is monotone decreasing, then for all $K \in \mathbb{N}$ and

In this case, $(\alpha_k)_{k=0}^{\infty} \subset \mathbb{R}_{++}$ need not satisfy the conditions (3.4).

$x \in M$,

$$\frac{1}{K}\sum_{k=0}^{K-1} V_k(x) \leq \left(\frac{\sigma^2}{b} + G^2\right)\frac{C_1}{K}\sum_{k=0}^{K-1}\alpha_k + \frac{C_2}{\alpha_{K-1}K} \tag{3.17}$$

for some $C_1, C_2 \in \mathbb{R}_{++}$.

Proof. Using Lemma 3.7, for all $K \in \mathbb{N}$ and $x \in M$, we obtain

$$\sum_{k=0}^{K-1}\alpha_k V_k(x) \leq \frac{\zeta(\kappa, D(x))}{2}\left(\frac{\sigma^2}{b} + G^2\right)\sum_{k=0}^{K-1}\alpha_k^2 + \frac{1}{2}\mathbb{E}\left[\left\|\mathrm{Exp}_{x_0}^{-1}(x)\right\|_{x_0}^2\right],$$

from which, together with $\sum_{k=0}^{+\infty}\alpha_k < +\infty$, we obtain

$$\sum_{k=0}^{K-1}\alpha_k V_k(x) < +\infty \tag{3.18}$$

for all $K \in \mathbb{N}$. From Lemma 2.2 with $\sum_{k=0}^{+\infty}\alpha_k = +\infty$ and (3.18), we have that

$$\liminf_{k\to+\infty} V_k(x) \leq 0.$$

(3.16) follows immediately.

Next, we show an upper bound of $(1/K)\sum_{k=0}^{K-1}V_k(x)$ such as (3.17). From Lemma 3.7, we obtain

$$\begin{aligned}\frac{1}{K}\sum_{k=0}^{K-1}V_k(x) &\leq \frac{\zeta(\kappa, D(x))}{2K}\left(\frac{\sigma^2}{b} + G^2\right)\sum_{k=0}^{K-1}\alpha_k \\ &\quad + \frac{1}{2K}\underbrace{\sum_{k=0}^{K-1}\frac{\mathbb{E}\left[\left\|\mathrm{Exp}_{x_k}^{-1}(x)\right\|_{x_k}^2\right] - \mathbb{E}\left[\left\|\mathrm{Exp}_{x_{k+1}}^{-1}(x)\right\|_{x_{k+1}}^2\right]}{\alpha_k}}_{X_K(x)}\end{aligned}$$

for all $K \in \mathbb{N}$. Hence,

$$\begin{aligned}X_K(x) &:= \frac{\mathbb{E}\left[\left\|\mathrm{Exp}_{x_0}^{-1}(x)\right\|_{x_0}^2\right]}{\alpha_0} + \sum_{k=1}^{K-1}\left(\frac{\mathbb{E}\left[\left\|\mathrm{Exp}_{x_k}^{-1}(x)\right\|_{x_k}^2\right]}{\alpha_k} - \frac{\mathbb{E}\left[\left\|\mathrm{Exp}_{x_k}^{-1}(x)\right\|_{x_k}^2\right]}{\alpha_{k-1}}\right) \\ &\quad - \frac{\mathbb{E}\left[\left\|\mathrm{Exp}_{x_K}^{-1}(x)\right\|_{x_K}^2\right]}{\alpha_{K-1}} \\ &\leq \frac{D(x)}{\alpha_0} + D(x)\sum_{k=0}^{K-1}\left(\frac{1}{\alpha_k} - \frac{1}{\alpha_{k-1}}\right) \\ &= \frac{D(x)}{\alpha_{K-1}},\end{aligned}$$

where the second inequality comes from the monotonicity of $(\alpha_k)_{k=0}^{\infty}$. Therefore, it follows that

$$\frac{1}{K}\sum_{k=0}^{K-1}V_k(x) \leq \underbrace{\frac{\zeta(\kappa, D(x))}{2}}_{C_1}\left(\frac{\sigma^2}{b} + G^2\right)\frac{1}{K}\sum_{k=0}^{K-1}\alpha_k + \underbrace{\frac{D(x)}{2}}_{C_2}\cdot\frac{1}{\alpha_{K-1}K}$$

for all $K \in \mathbb{N}$ and $x \in M$. This completes the proof. $\square$

3.3 Convergence Rate of Practical Step Sizes

Here, we calculate the convergence rate for practical step sizes.

First, we consider the constant step size defined as $\alpha_k := \alpha \in \mathbb{R}_{++}$. From Theorems 3.4 and 3.8, we immediately obtain the convergence rate for the constant step size. Hence, the convergence rates with and without L-smoothness are

$$\mathcal{O}\left(\frac{1}{K} + \frac{\sigma^2}{b}\right), \text{ and } \mathcal{O}\left(\frac{1}{K}\right) + \left(\frac{\sigma^2}{b} + G^2\right)\alpha,$$

respectively.

Next, we consider the diminishing step size defined as $\alpha_k := 1/\sqrt{k+1}$. Substituting $\alpha_k := 1/\sqrt{k+1}$ for the right side of (3.6), we obtain

$$\left(C_1 + \frac{C_2\sigma^2}{b}\sum_{k=0}^{K-1}\alpha_k^2\right)\frac{1}{K\alpha_{K-1}} \leq \left(C_1 + \frac{C_2\sigma^2}{b}(1+\log K)\right)\frac{1}{\sqrt{K}},$$

where

$$\sum_{k=0}^{K-1}\frac{1}{K+1} \leq 1 + \int_1^K \frac{dt}{t} = 1 + \log K.$$

Substituting $\alpha_k := 1/\sqrt{k+1}$ for the right side of (3.17), we obtain

$$\left(\frac{\sigma^2}{b} + G^2\right)\frac{C_1}{K}\sum_{k=0}^{K-1}\alpha_k + \frac{C_2}{\alpha_{K-1}K} \leq C_1\left(\frac{\sigma^2}{b} + G^2\right)\left(\frac{2}{\sqrt{K}} - \frac{1}{K}\right) + \frac{C_2}{\sqrt{K}}, \tag{3.19}$$

where

$$\frac{1}{K}\sum_{k=0}^{K-1}\frac{1}{\sqrt{K+1}} \leq \frac{1}{K}\left(1 + \int_1^K \frac{dt}{\sqrt{t}}\right) = \frac{2}{\sqrt{K}} - \frac{1}{K}.$$

Therefore, the convergence rates of $\alpha_k := 1/\sqrt{k+1}$ with and without L-smoothness are

$$\mathcal{O}\left(\frac{\log K}{\sqrt{K}}\right), \text{ and } \mathcal{O}\left(\left(1 + \frac{\sigma^2}{b}\right)\frac{1}{\sqrt{K}}\right),$$

respectively.

Finally, we consider the diminishing step size defined as $\alpha_k := \alpha\gamma^{p_k}$, where $\alpha, \gamma \in (0,1)$, $n, T \in \mathbb{N}$ and

$$p_k := \min\left\{\max\left\{m \in \mathbb{N}_0 : m \leq \frac{k}{T}\right\}, n\right\}.$$

This step size is explicitly represented as

$$\begin{aligned}&\underbrace{\alpha, \alpha, \cdots, \alpha}_{T}, \underbrace{\alpha\gamma, \alpha\gamma, \cdots, \alpha\gamma}_{T}, \underbrace{\alpha\gamma^2, \alpha\gamma^2, \cdots, \alpha\gamma^2}_{T},\\ &\quad \cdots, \underbrace{\alpha\gamma^{n-1}, \alpha\gamma^{n-1}, \cdots, \alpha\gamma^{n-1}}_{T}, \alpha\gamma^n, \alpha\gamma^n, \cdots,\end{aligned}$$

which means that $0 < \alpha\gamma^n \leq \alpha_k \leq \alpha$ for all $k \in \mathbb{N}_0$. Substituting $\alpha_k := \alpha\gamma^{p_k}$ for the right side of (3.6), we obtain

$$\left(C_1 + \frac{C_2\sigma^2}{b}\sum_{k=0}^{K-1}\alpha_k^2\right)\frac{1}{K\alpha_{K-1}} \leq \frac{C_1}{\alpha\gamma^n K} + \frac{C_2\alpha\sigma^2}{\gamma^n b}.$$

Substituting $\alpha_k := \alpha\gamma^{p_k}$ for the right side of (3.17), we obtain

$$\left(\frac{\sigma^2}{b} + G^2\right)\frac{C_1}{K}\sum_{k=0}^{K-1}\alpha_k + \frac{C_2}{\alpha_{K-1}K} \leq \left(\frac{\sigma^2}{b} + G^2\right)C_1\alpha + \frac{C_2}{\alpha\gamma^n K}. \tag{3.20}$$

Therefore, the convergence rates of $\alpha_k := \alpha\gamma^{p_k}$ with and without L-smoothness are

$$\mathcal{O}\left(\frac{1}{K} + \frac{\sigma^2}{b}\right), \text{ and } \mathcal{O}\left(\frac{1}{K}\right) + \mathcal{O}\left(\left(\frac{\sigma^2}{b} + G^2\right)\alpha\right),$$

respectively.

We summarize the convergence rates of practical step sizes in Table 2 It shows that increasing the batch size improves RSGD performance and that the constant α in learning rates should be sufficiently small.

Table 2: Convergence rates of three practical step sizes with and without assumption of L-smoothness ($\gamma \in (0,1)$, $G, \alpha > 0$, $\sigma^2 \geq 0$, and b is batch size).

Step size α_k	Convergence rate	
	with L-smooth	without L-smooth
$\alpha_k = \alpha$ (Constant)	$\mathcal{O}\left(\frac{1}{K} + \frac{\sigma^2}{b}\right)$	$\mathcal{O}\left(\frac{1}{K}\right) + \left(\frac{\sigma^2}{b} + G^2\right)\alpha$
$\alpha_k = 1/\sqrt{k+1}$ (Diminishing)	$\mathcal{O}\left(\frac{\log K}{\sqrt{K}}\right)$	$\mathcal{O}\left(\left(1 + \frac{\sigma^2}{b}\right)\frac{1}{\sqrt{K}}\right)$
$\alpha_k = \alpha\gamma^k$ (Diminishing)	$\mathcal{O}\left(\frac{1}{K} + \frac{\sigma^2}{b}\right)$	$\mathcal{O}\left(\frac{1}{K}\right) + \mathcal{O}\left(\left(\frac{\sigma^2}{b} + G^2\right)\alpha\right)$

3.4 Existence of a Critical Batch Size

Motivated by the work of Zhang and others [12,26,29], we will use the SFO complexity as the performance measure for a Riemannian stochastic optimizer. In particular, we will define SFO complexity as Kb, where K is the number of steps needed for solving the problem, and b is the batch size used in Algorithm 1. Furthermore, we will let $b^\star$ be the critical batch size for which Kb is minimized.

Our analyses (i.e., Theorems 3.8 and 3.9) give the number of steps K needed to satisfy an ε-approximation, which is defined as

$$\frac{1}{K}\sum_{k=0}^{K-1} V_k(x) \leq \varepsilon.$$

Theorem 3.10. *Suppose Assumption 3.1 is true, and let M be a Hadamard manifold with sectional curvature lower bounded by κ and $f : M \to \mathbb{R}$ be a smooth function. Then, the numbers of steps K needed to satisfy an ε-approximation for $\alpha_k = \alpha$, $\alpha_k = 1/\sqrt{k+1}$, and $\alpha_k = \alpha\gamma^{p_k}$ are respectively*

$$\begin{aligned} K &= \frac{C_2 b}{\varepsilon b - (\sigma^2 + G^2 b)\alpha C_1}, \quad b > \frac{(\sigma^2 + G^2 b)\alpha C_1}{\varepsilon}, \\ K &= \left(\frac{2C_1\sigma^2 + (2C_1 G^2 + C_2) b}{\varepsilon b} \right)^2 \\ K &= \frac{C_2 b \alpha^{-1} \gamma^{-n}}{\varepsilon b - (\sigma^2 + G^2 b)\alpha C_1}, \quad b > \frac{(\sigma^2 + G^2 b)\alpha C_1}{\varepsilon}. \end{aligned}$$

Here, K is convex and monotone decreasing with respect to b. SFO complexity Kb is convex with respect to b, and there exist critical batch sizes $b^\star$ for $\alpha_k = \alpha$, $\alpha_k = 1/\sqrt{k+1}$, and $\alpha_k = \alpha\gamma^{p_k}$

$$\begin{aligned} b^\star &= \frac{2C_2\sigma^2\alpha}{\varepsilon - G^2\alpha C_1}, \\ b^\star &= \exp\left\{ \left(\frac{2C_1 G^2 + C_2}{2C_1\sigma^2} \right)^2 \right\}, \\ b^\star &= \frac{2C_2\sigma^2\alpha}{\varepsilon - G^2\alpha C_1}. \end{aligned}$$

Proof. First, let us consider the case of $\alpha_k = \alpha$. From the upper bound of (3.13), we have

$$\left(\frac{\sigma^2}{b} + G^2 \right) \alpha C_1 + \frac{C_2}{K} = \varepsilon,$$

which implies

$$K = \frac{C_2 b}{\varepsilon b - (\sigma^2 + G^2 b)\alpha C_1}, \quad b > \frac{(\sigma^2 + G^2 b)\alpha C_1}{\varepsilon}.$$

Since

$$\begin{aligned} \frac{dK}{db} &= -\frac{C_1 C_2 \alpha \sigma^2}{\{\varepsilon b - (\sigma^2 + G^2 b)\alpha C_1\}^2} \leq 0, \\ \frac{d^2K}{db^2} &= \frac{2C_1 C_2 \alpha \sigma^2 (\varepsilon - G^2 \alpha C_1)}{\{\varepsilon b - (\sigma^2 + G^2 b)\alpha C_1\}^3} \geq 0, \end{aligned}$$

where the second inequality comes from $\varepsilon = (\sigma^2/b + G^2)\alpha C_1 + C_2/K > G^2\alpha C_1$, and K is convex and monotone decreasing with respect to the batch size b. Moreover, the SFO complexity, defined as

$$Kb := \frac{C_2 b^2}{\varepsilon b - (\sigma^2 + G^2 b)\alpha C_1},$$

is convex with respect to b since

$$\frac{d^2(Kb)}{db^2} = \frac{2C_1^2 C_2 \alpha^2 \sigma^4}{\{\varepsilon b - (\sigma^2 + G^2 b)\alpha C_1\}^3} \geq 0.$$

As well, from

$$\frac{d(Kb)}{db} = \frac{C_2 b(\varepsilon b - bG^2\alpha C_1 - 2\sigma^2\alpha C_1)}{\{\varepsilon b - (\sigma^2 + G^2 b)\alpha C_1\}^2},$$

$d(Kb)/db = 0$ if and only if $b = 2C_2\sigma^2\alpha/(\varepsilon - G^2\alpha C_1)$. Therefore, SFO complexity Kb is minimized at

$$b^\star := \frac{2C_2\sigma^2\alpha}{\varepsilon - G^2\alpha C_1},$$

which is the critical batch size.

Next, let us consider the case of $\alpha_k = 1/\sqrt{k+1}$. From (3.19), we have

$$\frac{2C_1}{\sqrt{K}}\left(\frac{\sigma^2}{b} + G^2\right) + \frac{C_2}{\sqrt{K}} = \varepsilon,$$

which implies

$$K = \left(\frac{2C_1\sigma^2 + (2C_1G^2 + C_2)b}{\varepsilon b}\right)^2.$$

Moreover, from

$$\frac{d(Kb)}{db} = -\left(\frac{2C_1\sigma^2}{\varepsilon}\right)^2 \log b + \left(\frac{2C_1G^2 + C_2}{\varepsilon}\right)^2,$$

$d(Kb)/db = 0$ if and only if $\log b = \{(2C_1G^2 + C_2)/(2C_1\sigma^2)\}^2$. Therefore, SFO complexity Kb is minimized at

$$b^\star := \exp\left\{\left(\frac{2C_1G^2 + C_2}{2C_1\sigma^2}\right)^2\right\},$$

which again is the critical batch size.

Finally, let us consider the case of $\alpha_k = \alpha\gamma^{p_k}$. From (3.20), we have

$$\left(\frac{\sigma^2}{b} + G^2\right)C_1\alpha + \frac{C_2}{\alpha\gamma^n K} = \varepsilon,$$

which implies

$$K = \frac{1}{\alpha\gamma^n}\cdot\frac{C_2 b}{\varepsilon b - (\sigma^2 + G^2 b)\alpha C_1}, \quad b > \frac{(\sigma^2 + G^2 b)\alpha C_1}{\varepsilon}.$$

Since this shows that we can multiply the result of $\alpha = \alpha$ by $\alpha^{-1}\gamma^{-n} > 0$, the proof follows immediately from the above discussion. □

Table 3 shows the relationship between the number of steps K and batch size b for each step size.

4 Numerical Experiments

The experiments were run on a MacBook Air (2020) laptop with a 1.8 GHz Intel Core i5 CPU, 8 GB 1600 MHz DDR3 memory, and the Monterey operating system (version 12.2). The algorithms were written in Python 3.10.7 using the PyTorch 1.13.1 package and the Matplotlib 3.6.2 package. The code is available at `https://github.com/iiduka-researches/rsgd-kylberg.git`.

Table 3: Relationship between number of steps K needed for ε-approximation and batch size b ($\gamma \in (0,1)$, $G, \alpha, C_1, C_2, \varepsilon > 0$, and $\sigma^2 \geq 0$).

Step size α_k	Relationship between K and b	Lower bound of b
$\alpha_k = \alpha$ (Constant)	$K = \dfrac{C_2 b}{\varepsilon b - (\sigma^2 + G^2 b)\alpha C_1}$	$\dfrac{(\sigma^2 + G^2 b)\alpha C_1}{\varepsilon}$
$\alpha_k = 1/\sqrt{k+1}$ (Diminishing)	$K = \left(\dfrac{2C_1\sigma^2 + (2C_1 G^2 + C_2)b}{\varepsilon b}\right)^2$	–
$\alpha_k = \alpha\gamma^k$ (Diminishing)	$K = \dfrac{C_2 b \alpha^{-1}\gamma^{-n}}{\varepsilon b - (\sigma^2 + G^2 b)\alpha C_1}$	$\dfrac{(\sigma^2 + G^2 b)\alpha C_1}{\varepsilon}$

4.1 Geometry of Symmetric Positive Definite Manifold

The set of $d \times d$ SPD matrices

$$\mathcal{S}_{++}^d := \{P \in \mathbb{R}^{d\times d} : P^\top = P, \forall x \in \mathbb{R}^d - \{0\}, \, x^\top P x > 0\}$$

endowed with the affine-invariant metric,

$$\langle X_P, Y_P \rangle_P := \mathrm{tr}(X_P^\top P^{-1} Y_P P^{-1}),$$

where $P \in \mathcal{S}_{++}^d$ and $X_P, Y_P \in T_P\mathcal{S}_{++}^d$, is a $d(d+1)/2$ dimensional Hadamard manifold (i.e., sectional curvatures of $\mathcal{S}_{++}^d$ are less than or equal to zero). This Riemannian manifold is called an "SPD manifold with an affine-invariant Riemannian metric" [7,9,20,25]. Criscitiello and Boumal [5] showed that the sectional curvatures of $\mathcal{S}_{++}^d$ are at least $-1/2$.

The exponential map $\mathrm{Exp}_P : T_P\mathcal{S}_{++}^d \to \mathcal{S}_{++}^d$ at a point $P \in \mathcal{S}_{++}^d$ was computed using

$$\mathrm{Exp}_P(X_P) = P^{\frac{1}{2}} \exp\left(P^{-\frac{1}{2}} X_P P^{-\frac{1}{2}}\right) P^{\frac{1}{2}},$$

where $X_P \in T_P\mathcal{S}_{++}^d$ [7].

4.2 Riemannian centroid problem on SPD manifold

We considered the Riemannian centroid problem of a set of SPD matrices $\{A_i\}_{i=0}^n$, which is frequently used in computer vision problems such as visual object categorization and pose categorization [14]. The loss function can be expressed as

$$f(M) := \frac{1}{N}\sum_{i=0}^{N} \left\| \log\left(A_i^{-\frac{1}{2}} M A_i^{-\frac{1}{2}}\right)\right\|_F^2,$$

where $\|\cdot\|_F$ is the Frobenius norm.

We took preprocessing steps similar to ones used elsewhere [10] and used the Kylberg dataset [17], which contains 28 texture classes of different natural and human-made surfaces. Each class has 160 unique samples imaged with and without rotation. The original images were scaled to 128×128 pixels and covariance descriptors were generated from 1024 4×4 non-overlapping pixel grids. The feature vector at each pixel was represented as

$$x_{u,v} = \left[I_{u,v}, \left|\frac{\partial I}{\partial u}\right|, \left|\frac{\partial I}{\partial v}\right|, \left|\frac{\partial^2 I}{\partial u^2}\right|, \left|\frac{\partial^2 I}{\partial v^2}\right| \right],$$

where $I_{u,v}$ is the intensity value.

We evaluated Algorithm 1 for several batch sizes by solving the Riemannian centroid problem on an SPD manifold on the Kylberg dataset. We used three learning rates: $\alpha_k = \alpha$ (constant), $\alpha_k = 1/\sqrt{k+1}$ (diminishing1), and $\alpha_k = \alpha\gamma^{p_k}$ (diminishing2). We used $\alpha = 5 \times 10^{-4}$, $\gamma = 0.5$, and $n = 10$. We defined T as the number of steps until all the elements in the data set had been used once. Numerical experiments had been performed for all batch sizes between 2^4 and 2^9.

4.3 Numerical Results

Figures 1 and 2 plot the number of steps K needed by Algorithm 1 to achieve $f(x_k) < \varepsilon$ versus the batch size b. The results are for $\varepsilon = 1/2$ and $\varepsilon = 1/4$, respectively. It can be seen that the number of steps K is monotone decreasing and convex with respect to the batch size, which is in support of the discussion in Section 3.4. Figures 3 and 4 plot SFO complexity Kb for the number of steps K needed to satisfy $f(x_k) < \varepsilon$ versus the batch size. Tresults are for $\varepsilon = 1/2$ and $\varepsilon = 1/4$, respectively. It is clear that the SFO complexity is convex with respect to b, in support of the discussion in Section 3.4.

Figure 3 shows that if $\varepsilon = 1/2$, the critical batch sizes for the constant, diminishing1, and diminishing2 learning rates are 247, 205, and 247, respectively. Figure 4 shows that if $\varepsilon = 1/4$, the critical batch sizes are 2^8, 233 and 2^8, respectively. These results support Theorem 3.10, which implies that the constant and diminishing2 learning rates have the same critical batch size and that it decreases as ε is increased. As indicated by Theorem 3.10, the critical batch size of diminishing1 for $\varepsilon = 1/2$ is almost the same as that of diminishing1 for $\varepsilon = 1/4$.

Table 4 shows the number of calculations of the objective function before an ε-approximation for each step size (i.e., $b = 2^4, 2^5, 2^6, 2^7, 2^8, 2^9$). In particular, its shows the ε-approximation for two cases $\varepsilon = 1/2$ and $\varepsilon = 1/4$. Overall, it shows that our theory and numerical experiments are consistent.

5 Conclusion

Our novel convergence analyses of Riemannian stochastic gradient descent on a Hadamard manifold, which incorporate the concept of mini-batch learning, overcome several problems with the previous analyses. We analyzed the relationship between batch size and the number of steps and demonstrated the existence of a critical batch size. In practice, the number of steps for ε-approximation is monotone decreasing and convex with respect to batch size. Moreover, stochastic first-order oracle complexity is convex with respect to batch size, and there exists a critical batch size that minimizes this complexity. Numerical experiments in which we solved the Riemannian centroid problem on a symmetric positive definite manifold were performed using several batch sizes to verify the results of theoretical analysis. With a constant step size, as ε decreases, the critical batch size increases. With a diminishing step size ($\alpha_k = \alpha\gamma^k$), the critical batch size matches that for the constant step size. Therefore, the experiments give numerical evidence in support of the theoretical analysis.

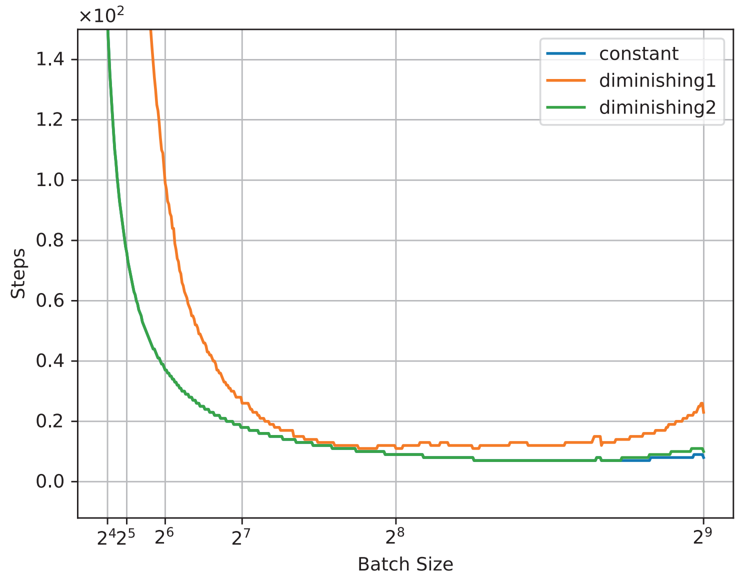

Figure 1: Number of steps K of Algorithm 1 versus batch size b when $\varepsilon = 1/2$.

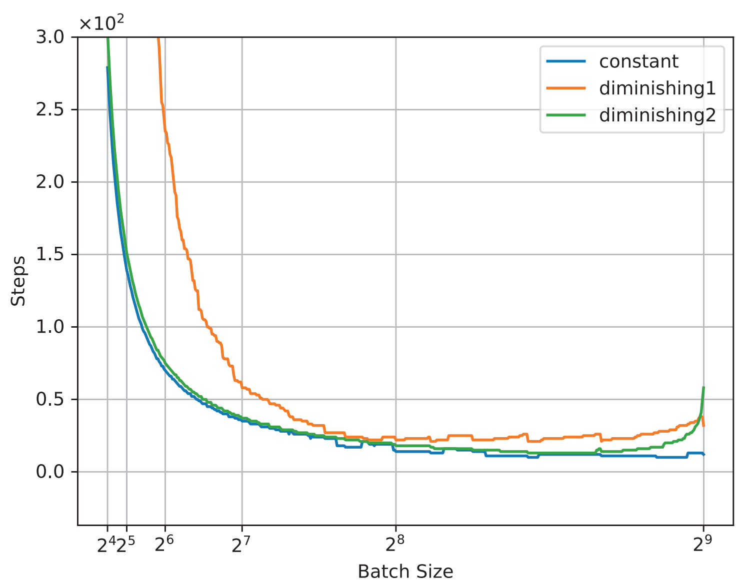

Figure 2: Number of steps K of Algorithm 1 versus batch size b when $\varepsilon = 1/4$.

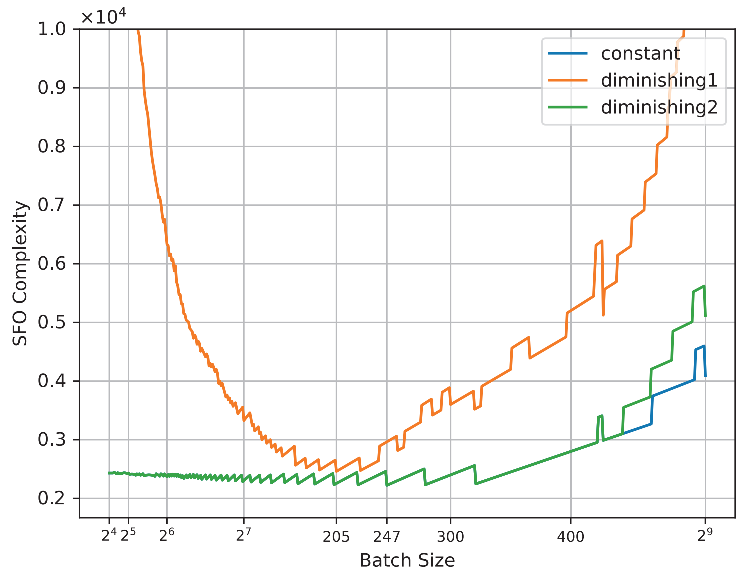

Figure 3: SFO complexity Kb of Algorithm 1 versus batch size b when $\varepsilon = 1/2$.

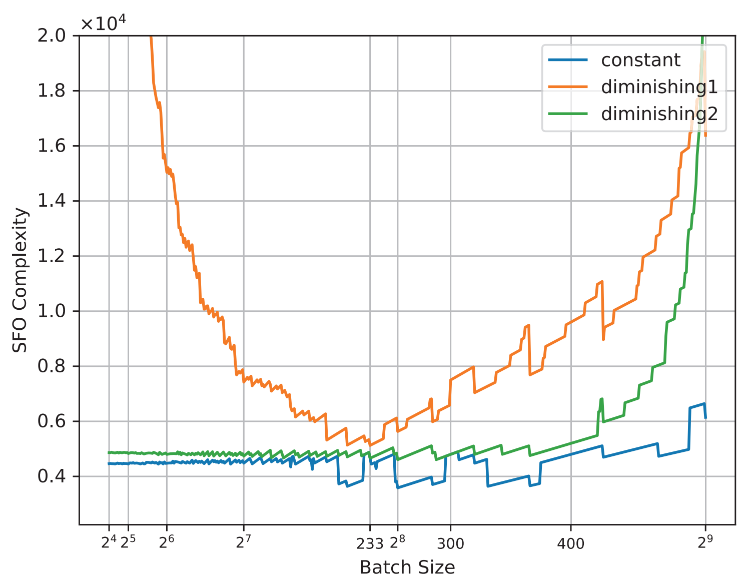

Figure 4: SFO complexity Kb of Algorithm 1 versus batch size b when $\varepsilon = 1/4$.

Table 4: Number of calculations of the objective function value before an ε-approximation is achieved for different batch sizes and ε.

		constant	diminshing1	diminishing2
$\varepsilon = 1/2$	$b = 2^4$	152	-	152
	$b = 2^5$	76	388	76
	$b = 2^6$	37	99	37
	$b = 2^7$	18	26	18
	$b = 2^8$	9	11	9
	$b = 2^9$	8	23	10
$\varepsilon = 1/4$	$b = 2^4$	279	-	304
	$b = 2^5$	139	948	151
	$b = 2^6$	70	235	75
	$b = 2^7$	35	58	37
	$b = 2^8$	14	22	18
	$b = 2^9$	12	32	58

References

[1] P-A. Absil, R. Mahony and R. Sepulchre, *Optimization Algorithms on Matrix Manifolds*, Princeton University Press, 2008.

[2] G. Bécigneul and O-E. Ganea, Riemannian adaptive optimization methods, in: *Proceedings of The International Conference on Learning Representations*, 2019.

[3] S. Bonnabel, Stochastic gradient descent on Riemannian manifolds, *IEEE Trans. Automat. Control.* 58 (2013) 2217–2229.

[4] A. Cherian, P. Stanitsas, M. Harandi, V. Morellas and N. Papanikolopoulos, Learning discriminative $\alpha\beta$-divergences for positive definite matrices, in: *2017 IEEE International Conference on Computer Vision*, 2017, pp. 4280–4289.

[5] C. Criscitiello and N. Boumal, An accelerated first-order method for non-convex optimization on manifolds, *Found. Comput. Math.* 23 (2022) 1–77.

[6] J. Duchi, E. Hazan and Y. Singer, Adaptive subgradient methods for online learning and stochastic optimization, *J. Mach. Learn. Res.* 12 (2011) 2121–2159

[7] R. Ferreira, J. Xavier, J.P. Costeira and V. Barroso, Newton method for Riemannian centroid computation in naturally reductive homogeneous spaces, in: *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, Toulouse, France, 2006, pp. III–III.

[8] P.T. Fletcher and S. Joshi, Riemannian geometry for the statistical analysis of diffusion tensor data, *Signal Process.* 87 (2007) 250–262.

[9] Z. Gao, Y. Wu, Y. Jia and M. Harandi, Learning to Optimize on SPD Manifolds, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7700–7709.

[10] M. Harandi, M. Salzmann and F. Porikli, Bregman divergences for infinite dimensional covariance matrices, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1003–1010.

[11] W. Huang, K.A. Gallivan and P-A. Absil, A Broyden class of quasi-Newton methods for Riemannian optimization, *SIAM J. Optim.* 25 (2015) 1660–1685.

[12] H. Iiduka, Critical Bach Size Minimizes Stochastic First-Order Oracle Complexity of Deep Learning Optimizer using Hyperparameters Close to One, *arXiv preprint arXiv:2208.09814.* (2022).

[13] H. Iiduka, Theoretical analysis of Adam using hyperparameters close to one without Lipschitz smoothness, *Numer. Algorithms.* 95 (2024) 383-421.

[14] S. Jayasumana, R. Hartley, M. Salzmann, H. Li and M. Harandi, Kernel methods on Riemannian manifolds with Gaussian RBF kernels, *IEEE Trans. Pattern Anal. Mach. Intell.* 37 (2015) 2464–2477.

[15] H. Kasai, P. Jawanpuria and B. Mishra, Riemannian adaptive stochastic gradient algorithms on matrix manifolds, in: *International Conference on Machine Learning*, 2019, pp. 3262–3271.

[16] D.P. Kingma and J. Ba, Adam: A method for stochastic optimization, in: *Proceedings of The International Conference on Learning Representations*, 2015, pp. 1–15.

[17] G. Kylberg, The Kylberg Texture Dataset v. 1.0 External report (Blue series), No. 35, Centre for Image Analysis, Swedish University of Agricultural Sciences and Uppsala University, September 2011.

[18] S.Z. Németh, Variational inequalities on Hadamard manifolds, *Nonlinear Anal.* 52 (2003) 1491–1498.

[19] M. Nickel and D. Kiela, Poincaré embeddings for learning hierarchical representations, in: *Advances in Neural Information Processing Systems*, 30, 2017.

[20] X. Pennec, P. Fillard and N. Ayache, A Riemannian framework for tensor computing *Int. J. Comput. Vis.* 66 (2006) 41–66.

[21] S.J. Reddi, S. Kale and S. Kumar, in: *Proceedings of The International Conference on Learning Representations*, 2018, pp. 1–23.

[22] H. Sakai and H. Iiduka, Riemannian Adaptive Optimization Algorithm and its Application to Natural Language Processing, *IEEE Trans. Cybern.* 52 (2022) 7328–7339.

[23] T. Sakai, *Riemannian Geometry*, American Mathematical Society, 1996.

[24] N. Sato and H. Iiduka, Existence and Estimation of Critical Batch Size for Training Generative Adversarial Networks with Two Time-Scale Update Rule, *Proceedings of the 40th International Conference on Machine Learning.* 202 (2023) 30080–30104.

[25] H. Sato, H. Kasai and B. Mishra, Riemannian stochastic variance reduced gradient algorithm with retraction and vector transport, *SIAM J. Optim.* 29 (2019) 1444–1472.

[26] C.J. Shallue, J. Lee, J. Antognini, J. Sohl-Dickstein, R. Frostig, and G.E. Dahl, Measuring the effects of data parallelism on neural network training, *J. Mach. Learn. Res.* 20 (2019) 1–49.

[27] O. Tuzel, F. Porikli, and P. Meer, Region covariance: A fast descriptor for detection and classification, in: *European Conference on Computer Vision*, 2006, pp. 589–600.

[28] M.D. Zeiler, Adadelta: An adaptive learning rate method, *arXiv preprint arXiv:1212.5701*. 2012.

[29] G. Zhang, L. Li, Z. Nado, J. Martens, S. Sachdeva, G. Dahl, C. Shallue and R. B. Grosse, Which algorithmic choices matter at which batch sizes? insights from a noisy quadratic model, *Advances in Neural Information Processing Systems*, 34, 2019, pp. 8196–8207.

[30] H. Zhang and S. Sra, First-order methods for geodesically convex optimization, in: *Conference on Learning Theory*, 2016, pp. 1617–1638.

Manuscript received 16 September 2023
revised 27 December 2023
accepted for publication 7 January 2024

Hiroyuki Sakai
Department of Computer Science, Meiji University,
1-1-1 Higashimita, Tama-ku, Kawasaki-shi, Kanagawa, 214-8571, Japan
E-mail address: sakai0815@cs.meiji.ac.jp

Hideaki Iiduka
Department of Computer Science, Meiji University,
1-1-1 Higashimita, Tama-ku, Kawasaki-shi, Kanagawa, 214-8571, Japan
E-mail address: iiduka@cs.meiji.ac.jp