Journal of Machine Learning Research 24 (2023) 1-67 Submitted 11/21; Revised 12/22; Published 2/23

Online Optimization over Riemannian Manifolds*

Xi Wang WANGXI14.UCAS@GMAIL.COM
Zhipeng Tu TUZHIPENG@AMSS.AC.CN
Academy of Mathematics and Systems Science
Chinese Academy of Sciences
Beijing 100190, P. R. China,
and Australian Center for Robotics
School of Aerospace, Mechanical and Mechatronic Engineering
The University of Sydney
NSW 2006, Australia

Yiguang Hong YGHONG@ISS.AC.CN
Shanghai Research Institute for Intelligent Autonomous Systems
Tongji University
Shanghai 201210, P. R. China

Yingyi Wu WUYY@UCAS.AC.CN
Department of Mathematics
University of Chinese Academy of Sciences
Beijing 100040, P. R. China

Guodong Shi† GUODONG.SHI@SYDNEY.EDU.AU
Australian Center for Robotics
School of Aerospace, Mechanical and Mechatronic Engineering
The University of Sydney
NSW 2006, Australia

Editor: Silvia Villa

Abstract

Online optimization has witnessed a massive surge of research attention in recent years. In this paper, we propose online gradient descent and online bandit algorithms over Riemannian manifolds in full information and bandit feedback settings respectively, for both geodesically convex and strongly geodesically convex functions. We establish a series of upper bounds on the regrets for the proposed algorithms over Hadamard manifolds. We also find a universal lower bound for achievable regret on Hadamard manifolds. Our analysis shows how time horizon, dimension, and sectional curvature bounds have impact on the regret bounds. When the manifold permits positive sectional curvature, we prove similar regret bound can be established by handling non-constrictive project maps. In addition, numerical studies on problems defined on symmetric positive definite matrix manifold, hyperbolic spaces, and Grassmann manifolds are provided to validate our theoretical findings, using synthetic and real-world data.

∗. A preliminary version of the paper is scheduled for presentation at NeurIPS-2021 (Wang et al., 2021).
†. Correspondence author (G. Shi, Ross Street Building, Darlington NSW 2006, Sydney, Australia; +61-02-8627 8037).

Keywords: Online optimization, Riemannian manifolds, Riemannian optimization, Gradient estimation, Fréchet mean

1. Introduction

The *online optimization* has been widely studied in the past decades in online routing, spam filtering, and machine learning (Agmon, 1954; Hazan, 2016; Arnold et al., 2019). Without a prior knowledge of loss functions, an online convex optimization algorithm predicts solutions before loss functions are revealed.

In this paper, we consider the following Riemannian online convex optimization (R-OCO) problem,

$$\min_{x_t \in \mathcal{K} \subset \mathcal{M}} \mathbf{f}_t(x_t), t = 1, 2, \ldots, T, \tag{1}$$

where $\mathcal{M}$ is a complete Riemannian manifold equipped with a Riemannian metric g and $\mathcal{K}$ is a geodesically convex (g-convex) subset of $\mathcal{M}$. Here, $\{\mathbf{f}_t\}_{t=1,2,\ldots,T}$ is a sequence of unknown loss functions and every $\mathbf{f}_t$ is a geodesically convex (g-convex) function with sufficient smoothness. The R-OCO problem (1) extends the online convex optimization in Euclidean spaces with potential applications in machine learning, such as online principal component analysis (PCA), dictionary learning, and neural networks (Lee and Kriegman, 2005; Feng et al., 2013; Hu et al., 2020).

The R-OCO problem (1) can be understood as a learning process of T rounds. At each round $t = 1, 2, 3, \ldots, T$, an online learner chooses a strategy x_t from the g-convex subset $\mathcal{K}$. Later or simultaneously, the adversary (or nature) produces a g-convex loss function $\mathbf{f}_t : \mathcal{K} \to \mathbb{R}$ of which the learner has no prior knowledge. Finally, the learner receives the feedback and suffers the loss $\mathbf{f}_t(x_t)$. Generally, there are two types of information feedback. One is the full information feedback, where the entire function $\mathbf{f}_t$ is revealed to the learner; the other is the bandit feedback, where only the value $\mathbf{f}_t(x_t)$ is revealed. The goal of the R-OCO is to minimize the *regret*, defined as

$$\mathsf{Reg}(T) = \sum_{t=1}^{T} \mathbf{f}_t(x_t) - \min_{x \in \mathcal{K}} \sum_{t=1}^{T} \mathbf{f}_t(x),$$

which measures the difference between the cost by $\{x_t\}_{t=1,\ldots,T}$ and the best-fixed point chosen in hindsight. An algorithm is termed no-regret (Srinivas et al., 2010), if the regret of the algorithm goes sublinearly with the time horizon T.

For carrying out optimization over a manifold, some classical methods treat the manifold as a subset of an ambient Euclidean space and employ Euclidean constrained optimization techniques. For instance, Nie et al. (2016) presented an online PCA algorithm, where the variables were updated in an embedding Euclidean space and then projected onto a manifold. However, in practical applications, the dimension of an embedding Euclidean space can be too high (e.g., the Grassmann manifold, Boumal and Absil, 2015), and the projection can be expensive to compute (e.g., the manifold of symmetric positive definite (SPD) matrices, Zhang et al., 2018). An alternative approach termed *Riemannian optimization* makes use of intrinsic geometry of manifolds so that it can optimize directly on manifolds

as an unconstrained problem, and thus avoiding high dimension embedding and high computing cost for the projection. Furthermore, this viewpoint has shown benefits from the g-convexity, by which a nonconvex optimization problem can be converted into a g-convex one (Allen-Zhu et al., 2018). Consequently, it is important to take a Riemannian approach in our problem (1).

Although there were many existing algorithms for offline manifold optimization problems (Absil et al., 2009; Ring and Wirth, 2012; Ahn and Sra, 2020), very few results were obtained about the Riemannian online optimization problem. Tupker et al. (2021) proposed an online algorithm for estimating hidden Markov chains on Hadamard homogeneous spaces; Becigneul et al. (2019) analyzed Riemannian adaptive methods on product manifolds in the regret sense. More recently, Maass et al. (2022) studied a zeroth-order online optimization problem on Hadamard manifolds and achieved no-regret bound with a sublinear assumption.

Contribution This paper aims to design no-regret algorithms for the R-OCO problem in both full information feedback and bandit feedback. The contribution of this paper is summarized as follows:

- We propose a Riemannian online gradient descent algorithm (R-OGD) for the R-OCO problem in the full information feedback, and then establish the regret bounds on Hadamard manifolds for g-convex and strongly g-convex functions. In addition, we present a universal lower regret bound which matches the regret bound achieved by R-OGD in g-convex setting.

- We introduce a Riemannian bandit algorithm (R-BAN) and construct regret bounds for g-convex and strongly g-convex functions with the one-point bandit feedback. We also proposed a Riemannian two-point bandit algorithm (R-2-BAN) with the two-point bandit feedback, of which regret bounds can be improved to resemble the bounds in full information cases. Moreover, we develop a key technique to analyze the derivative of a local integration on homogeneous manifolds, which can be applied to estimate gradients in Riemannian optimization and beyond.

- We generalize the R-OGD, R-BAN and R-2-BAN algorithms to non-Hadamard manifolds. We overcome the challenge of non-constrictive projection maps and derive regret bounds of the same order in time horizon compared to those in Hadamard cases.

The established lower and upper bounds on the achievable bounds of the R-OCO match their counterparts for Euclidean online convex optimization (Zinkevich, 2003; Hazan et al., 2006; Flaxman et al., 2005; Abernethy et al., 2008; Agarwal et al., 2010). Please see Table 1 for the detail.

Some preliminary results of the paper are scheduled for presentation at NeurIPS-2021 (Wang et al., 2021). Compared to the conference version, we have expanded the theoretical study considerably into two-point bandit algorithm and regret analysis for non-Hadamard manifolds, and presented a comprehensive set of numerical tests on both synthetic and real-world data.

Related Work The Euclidean online convex optimization was introduced by Zinkevich (2003). Inspired by the gradient descent method, Zinkevich (2003) proposed the online gradient descent algorithm (OGD) of which the regret bound was proven to be $\mathcal{O}\big(\sqrt{T}\big)$. Then Hazan et al. (2006) proceeded with the study of the OGD algorithm and established a regret bound $\mathcal{O}\big(\log T\big)$ for strongly convex functions. In addition, Abernethy et al. (2008) gave a universal lower bound of $\mathcal{O}\big(\sqrt{T}\big)$ for online algorithms, which indicated that the bounds in Zinkevich (2003) and Hazan et al. (2006) are essentially optimal. In the bandit setting, Flaxman et al. (2005) provided a detailed exposition of a one-point bandit algorithm. By modifying the gradient in the OGD algorithm to a randomized estimator, the regret bounds attained $\mathcal{O}\big(nT^{\frac{3}{4}}\big)$ and $\mathcal{O}\big(n^{\frac{2}{3}}(1+\log T)^{\frac{1}{3}}T^{\frac{2}{3}}\big)$ for convex loss functions and strongly convex loss functions, respectively. By extending the one-point bandit algorithm, Agarwal et al. (2010) developed a multi-point bandit algorithm and presented regret bounds $\mathcal{O}\big(n\sqrt{T}\big)$ and $\mathcal{O}\big(n^2(1+\log T)\big)$ for convex and strongly convex loss functions. The Riemannian online algorithms proposed in this paper in the full information feedback and the bandit feedback settings are extensions of the Euclidean online algorithms to Riemannian manifolds.

Riemannian optimization has drawn much research attention in the past decades. Many basic algorithms in Euclidean spaces such as the gradient descent method, Newton's method, and trust-region methods have been adapted into a Riemannian setting (see Fiori, 2005; Absil et al., 2009; Ahn and Sra, 2020; Ring and Wirth, 2012; Koudounas and Fiori, 2020). Some research of Riemannian stochastic optimization (R-SO) was intended to deal with time-varying optimization problems (Bonnabel, 2013; Zhang and Sra, 2016; Zhang et al., 2018; Tupker et al., 2021). Among them, Zhang and Sra (2016) provided the first global complexity analysis for the R-SGD algorithm on geodesically convex problems over Hadamard manifolds, and Tupker et al. (2021) proposed an online algorithm to deal with hidden Markov chains on Hadamard homogeneous spaces. When loss functions are arrived in batch, R-SO methods are actually to minimize the average regret in the case of knowing the prior distribution of loss functions. In this case, the R-SO can be viewed as a kind of R-OCO problems and R-OCO algorithms can handle broader settings without prior knowledge.

The results about the R-OCO problem are fairly limited. Antonakopoulos et al. (2020) proposed regularized online optimization methods via a Riemann–Lipschitz continuity condition, which focused on convex functions and vector addition from an ambient Euclidean space. In the full information setting, Becigneul et al. (2019) proposed the Riemannian versions of ADAGRAD and ADAM algorithms, which depended on a product manifold structure. In addition, Becigneul et al. (2019) constructed $\mathcal{O}\big(\sqrt{T}\big)$ regret bounds of both ADAGARD and ADAM algorithms for g-convex functions. When the form of losses is not available, Maass et al. (2022) proposed a Riemannian online zeroth-order (R-OZO) algorithm for strongly g-convex functions. The R-OZO generated a random Gaussian vector u_t in an ambient embedding Euclidean space, and then used a two-point difference to present a descent along the projection of u_t on the tangent space of the manifold. For g-strongly convex functions on Hadamard manifolds, Maass et al. (2022) derived asymptotic tracking error and a $\mathcal{O}\big(\sqrt{T}+V_T\big)$ dynamic regret bound of the R-OZO, where V_T is the accumulated distance between two consecutive minimizers. In contrast, the regret bounds established for our online gradient-based/bandit Riemannian optimization algorithms are sublinear for any time, matching those for Euclidean online optimization.

A detailed comparison of our results with the existing works is summarized in Table 1.

Feedback setting		G-convex	Strongly g-convex
Full information	Our work	$\mathcal{O}\big(\zeta^{\frac{1}{2}}\sqrt{T}\big)$	$\mathcal{O}\big(\zeta \log T\big)$
	Previous Work	$\mathcal{O}\big(\sqrt{T}\big)$ (Product space) (Becigneul et al., 2019)	–
	Euclidean	$\mathcal{O}\big(\sqrt{T}\big)$ (Zinkevich, 2003)	$\mathcal{O}\big(\log T\big)$ (Hazan et al., 2006)
One-point bandit	Our work	$\mathcal{O}\big(n\zeta^{\frac{1}{2}}T^{\frac{3}{4}}\big)$	$\mathcal{O}\big(n^{\frac{2}{3}}\zeta(1+\log T)^{\frac{1}{3}}T^{\frac{2}{3}}\big)$
	Previous Work	–	–
	Euclidean	$\mathcal{O}\big(nT^{\frac{3}{4}}\big)$ (Flaxman et al., 2005)	$\mathcal{O}\big(n^{\frac{2}{3}}(1+\log T)^{\frac{1}{3}}T^{\frac{2}{3}}\big)$ (Flaxman et al., 2005)
Two-point bandit	Our work	$\mathcal{O}\big(n\zeta^{\frac{1}{2}}\sqrt{T}\big)$	$\mathcal{O}\big(n^2\zeta(1+\log T)\big)$
	Previous Work	–	$\mathcal{O}\big(\sqrt{T}+V_T\big)$ (Dynamic regret) (Maass et al., 2022)
	Euclidean	$\mathcal{O}\big(n\sqrt{T}\big)$ (Agarwal et al., 2010)	$\mathcal{O}\big(n^2(1+\log T)\big)$ (Agarwal et al., 2010)
Universal lower bound	Our work	$\Omega(\sqrt{T})$	–
	Previous Work	–	–
	Euclidean	$\Omega(\sqrt{T})$ (Agarwal et al., 2010)	$\Omega(\log T)$ (Agarwal et al., 2010)

Table 1: Comparison of regret among our work, previous Riemannian online optimization, and corresponding results in Euclidean spaces. T: the time horizon; D: the diameter of the feasible set; n: the dimension of the manifold; ζ: a constant related to the sectional curvature bound κ; V_T: the accumulated distance between two consecutive minimizers.

2. Preliminaries

In this section, we present a brief review on the Riemannian manifold and introduce basic functions classes for Riemannian optimization. We refer readers to the following textbooks and tutorial papers (do Carmo, 1992; Chern et al., 1999; Berestovskii and Nikonorov, 2020; Ghomi and Spruck, 2019; Fiori, 2021) for more details.

Riemannian manifolds An n-dimensional *manifold* $\mathcal{M}$ is a topological space locally diffeomorphic to the vector space $\mathbb{R}^n$. The *tangent space* $T_x\mathcal{M}$ is a linearization of the manifold $\mathcal{M}$ at a point x. A *Riemannian manifold* is a smooth manifold $\mathcal{M}$ equipped with a metric tensor g (called Riemannian metric), which defines an inner product

$$g_x : T_x\mathcal{M} \times T_x\mathcal{M} \to \mathbb{R}$$
$$g_x(X, Y) = \langle X, Y \rangle_x$$

in every tangent space $T_x\mathcal{M}$ of $x \in \mathcal{M}$. The Riemannian metric g gives us a way to measure the length of curves, bringing a metric space structure to $\mathcal{M}$ with distance function

$$d(x, y) = \inf_{\gamma}\{\text{Length}(\gamma)|\gamma \text{ is a curve connecting } x \text{ and } y\}.$$

A curve is a *geodesic* if it locally minimizes the length, which is an analog of a straight line in Euclidean spaces. On Riemannian manifolds, a geodesic is uniquely determined by its starting point and initial tangent vector. In this way, the *exponential map* $\exp_x : T_x\mathcal{M} \to \mathcal{M}$ is defined by mapping a vector $X \in T_p\mathcal{M}$ to $\gamma(1) \in \mathcal{M}$ for the geodesic γ such that $\gamma(0) = x$ and $\dot{\gamma}(0) = X$. A set $\mathcal{K}$ is termed *geodesically convex* (g-convex) if, for any points $x, y \in \mathcal{K}$, there admits a geodesic $\gamma \subset \mathcal{K}$ connecting x and y. Moreover, if the γ is unique, the set $\mathcal{K}$ is termed *uniquely geodesically convex* (uniquely g-convex). It is shown that the exponential map $\exp_x$ is locally a diffeomorphism and consequently has an inverse $\exp_x^{-1}(\cdot)$ on a uniquely g-convex set.

Curvature reflects the geometry of manifolds. We focus on *sectional curvature*, which is the Gauss curvature of a two-dimensional submanifold. Following Zhang and Sra (2016), we mainly consider the *Hadamard manifold*, which is a simply connected and complete manifold with non-positive sectional curvature. The Cartan-Hadamard theorem (Berger, 2009) shows that the Hadamard manifold is uniquely g-convex so that the exponential map $\exp_x$ has an global inverse $\exp_x^{-1}(\cdot)$ on Hadamard manifolds. In this way, the distance $d(x, y)$ can be expressed as $\|\exp_x^{-1}(y)\|_x$.

Isometries of Riemannian manifolds have been widely studied in differential geometry (Berger, 2009; Berestovskii and Nikonorov, 2020). An isometry $\phi : \mathcal{M} \to \mathcal{M}$ is a diffeomorphism preserving distance, i.e., $d(x, y) = d(\phi(x), \phi(y))$ for all $x, y \in \mathcal{M}$. It is remarked that all isometries of a Riemannian manifold form a Lie group G. A Riemannian manifold is a *homogeneous manifold* if the group of isometries G acts on $\mathcal{M}$ transitively, i.e., for any points $x, y \in \mathcal{M}$ there exists an isometry such that $\phi(x) = y$. A Riemannian manifold is a *symmetric manifold* if for any $x \in \mathcal{M}$, there exists a symmetry $s_x \in G$ such that x is an isolated fixed point of s_x.

Vector fields and Their flows A *vector field* X is a map that assigns every point $x \in \mathcal{M}$ to a tangent vector $X(x) \in T_x\mathcal{M}$. Let $\mathfrak{X}(M)$ denote the set of all vector fields. A vector field X can be also viewed as a differential operator over smooth functions on $\mathcal{M}$, i.e., the operation $X(\mathbf{f})$ gives a function on $\mathcal{M}$, defined as

$$X(\mathbf{f})(x) = \lim_{t\to 0} \frac{1}{t}(\mathbf{f}(\xi(t)) - \mathbf{f}(x)),$$

where ξ is a curve that starts at x with the tangent vector $X(x)$.

The *Levi-Civita connection* ∇ is an analogue of the differential operator over vector fields in Euclidean spaces and uniquely determined by properties

$$\begin{cases} X\langle Y, Z\rangle = \langle \nabla_X Y, Z\rangle + \langle Y, \nabla_X Z\rangle \\ \nabla_X Y - \nabla_Y X = XY - YX \end{cases}$$

for all $X, Y, Z \in \mathfrak{X}(\mathcal{M})$.

The infinitesimal variation of a geodesic is described by the Jacobi field. A vector field η along a geodesic γ is a Jacobi field if it satisfies the Jacobi equation

$$\nabla_{\dot\gamma}\nabla_{\dot\gamma}\eta + R(\dot\gamma,\eta)\dot\gamma = 0.$$

A vector η is a Killing field if it satisfies for all $X, Y \in \mathfrak{X}(\mathcal{M})$

$$\langle \nabla_X \eta, Y\rangle + \langle X, \nabla_Y \eta\rangle = 0.$$

We follow the same idea in Euclidean spaces to define the flow of a vector field. Suppose that $\mathcal{M}$ is a smooth manifold and $X \in \mathfrak{X}(\mathcal{M})$. Let there be a smooth map $\phi : \mathbb{R}\times\mathcal{M} \to \mathcal{M}$. Denote $\phi_t(p) = \phi(t,p)$, for any $(t,p) \in \mathbb{R}\times\mathcal{M}$, such that the following conditions are satisfied:

1) $\phi_0(p) = p$;

2) $\phi_s \circ \phi_t = \phi_{s+t}$ for any real numbers s, t;

3) $X(p) = \frac{\partial \phi_t(p)}{\partial t}|_{t=0}$.

Then we call ϕ_t the flow (or the one-parameter group of diffeomorphism) of X, and term X the infinitesimal transformation of ϕ_t.

Function Classes A function $\mathbf{f} : \mathcal{K} \to \mathbb{R}$ is called *geodesically convex* (or g-convex) if for any geodesic $\gamma : [0,1] \to \mathcal{M}$,

$$\mathbf{f}(\gamma(t)) \le (1-t)\mathbf{f}(\gamma(0)) + t\mathbf{f}(\gamma(1)).$$

The g-convexity has some equivalent conditions. When $\mathbf{f}$ is differentiable, which means that there exists a *gradient* vector field $\nabla\mathbf{f}$ such that $\langle \nabla\mathbf{f}(x), X\rangle = X(\mathbf{f})(x)$ for every vector field $X \in \mathfrak{X}(\mathcal{M})$, the g-convexity is equivalent to the following condition

$$\mathbf{f}(y) \ge \mathbf{f}(x) + \langle \nabla\mathbf{f}(x), \exp_x^{-1}(y)\rangle, \forall x, y \in \mathcal{M}.$$

Furthermore, if $\mathbf{f}$ is twice differentiable, the g-convexity is equivalent to

$$\nabla^2\mathbf{f}(X,X) := \langle \nabla_X \nabla\mathbf{f}, X\rangle = X(X(\mathbf{f})) - \nabla_X X(\mathbf{f}) \ge 0$$

for any $X \in \mathfrak{X}(\mathcal{M})$.

A differentiable function $\mathbf{f} : \mathcal{M} \to \mathbb{R}$ is *geodesically μ-strongly convex* (or μ-strongly g-convex) if there exists a constant $\mu > 0$ such that for any $x, y \in \mathcal{M}$, there holds

$$\mathbf{f}(y) \ge \mathbf{f}(x) + \langle \nabla\mathbf{f}(x), \exp_x^{-1}(y)\rangle + \frac{\mu}{2} d^2(x,y).$$

We term a function to be *geodesically L-Lipschitz* (or g-L-Lipschitz) if there exists a constant $L > 0$ such that, for any $x, y \in \mathcal{M}$,

$$|\mathbf{f}(y) - \mathbf{f}(x)| \le L \cdot d(x,y),$$

which is equivalent to

$$\|\nabla\mathbf{f}(x)\| \le L, \forall x \in \mathcal{M},$$

if $\mathbf{f}$ is differentiable.

Algorithm 1: Riemannian Online Gradient Descent Algorithm (R-OGD)

Input: Manifold $\mathcal{M}$, time T, step sizes (or schedule) $\{\alpha_t\}$
Output: $\{x_t\}_{t=1,\dots,T}$
for $t = 1$ *to* T **do**
Play x_t and observe the function $\mathbf{f}_t$;
Update x_{t+1} with

$$\begin{cases} \tilde{x}_{t+1} = \exp_{x_t}(-\alpha_t \nabla \mathbf{f}_t(x_t)) \\ x_{t+1} = \mathscr{P}_{\mathcal{K}}(\tilde{x}_{t+1}), \end{cases}$$

where $\mathscr{P}_{\mathcal{K}}$ is the Riemannian projection mapping of x onto $\mathcal{K}$, i.e.,

$$\mathscr{P}_{\mathcal{K}}(x) := \arg\min_{y \in \mathcal{K}} d(x, y);$$

Return x_{t+1}, and suffer the loss $\mathbf{f}_t(x_t)$;
end

3. R-OCO with Full Information Feedback

This section is devoted to the study of the R-OCO problem in the full information feedback. We first propose our R-OGD algorithm and then analyze the upper regret bounds of the R-OGD for both g-convex and strongly g-convex functions. In addition, a universal lower regret bound in the g-convex case is presented to illustrate that the regret bound of the R-OGD algorithm is tight up to a constant.

3.1 Riemannian Online Gradient Algorithm

In the full information setting, we consider the following assumptions, which were standard in the literature of Euclidean online convex optimization and Riemannian optimization (Zinkevich, 2003; Ahn and Sra, 2020; Huang et al., 2015).

Assumption 1 *There exists a* $x^\star \in \mathcal{M}$ *such that* $x^\star = \arg\min \sum_{t=1}^{T} \mathbf{f}_t(x)$.

Assumption 2 $(\mathcal{M}, g)$ *is a Hadamard manifold with the sectional curvature lower bounded by a constant* κ.

Note that the Hadamard manifold plays an important role in Riemannian geometry (see Ghomi and Spruck, 2019). Some well-known spaces, such as the Euclidean space $\mathbb{R}^n$, the hyperbolic space H^n, and the manifold of SPD matrices, are all Hadamard manifolds (Ahn and Sra, 2020; Huang et al., 2015).

Assumption 3 *The set* $\mathcal{K}$ *is a bounded and g-convex set with diameter* D, *i.e.,*

$$d(x, y) \leq D, \forall x, y \in \mathcal{K}.$$

It is worth emphasizing that the Cartan-Hadamard theorem indicates that any g-convex set on Hadamard manifolds is uniquely g-convex. Consequently, we can define the inverse exponential map $\exp_x^{-1}(\cdot)$ for any point $x \in \mathcal{K}$ (do Carmo, 1992).

Assumption 4 *For any $t = 1, \ldots, T$, $\mathbf{f}_t$ is differentiable and g-L-Lipschitz.*

We now propose our Riemannian online gradient descent algorithm (R-OGD) in Algorithm 1, where the exponential map replaces the vector addition in the Euclidean online gradient descent (Zinkevich, 2003).

3.2 Regret Upper Bounds

In Theorems 5 and 6 we present upper regret bounds of the R-OGD algorithm for g-convex loss functions and strongly g-convex functions, respectively. Take

$$\zeta(\kappa, d) = \begin{cases} \frac{\sqrt{|\kappa|}\cdot d}{\tanh\left(\sqrt{|\kappa|}\cdot d\right)}, & \kappa < 0 \\ 1, & \kappa \geq 0. \end{cases}$$

By direct observation, ζ decreases with respect to κ, and increases with respect to d.

Theorem 5 (Convex Case) *Suppose that Assumptions 1-4 hold, and $\mathbf{f}_t$ is g-convex for any $t = 1, \ldots, T$. Then the R-OGD algorithm with step sizes $\{\alpha_t = \frac{D}{L\sqrt{\zeta(\kappa,D)t}}\}$ guarantees the following regret bound for all $T \geq 1$:*

$$\mathsf{Reg}(T) \leq \frac{3}{2} DL\sqrt{\zeta(\kappa, D)\cdot T}.$$

Theorem 6 (Strongly-convex Case) *Suppose that Assumptions 1-4 hold, and that $\mathbf{f}_t$ is μ-strongly g-convex for any $t = 1, \ldots, T$. Then the R-OGD algorithm with step sizes $\{\alpha_t = \frac{1}{\mu t}\}$ guarantees the following regret bound for all $T \geq 1$:*

$$\mathsf{Reg}(T) \leq \frac{L^2\zeta(\kappa, D)}{2\mu}(1 + \log T).$$

The proofs of Theorems 5 and 6 are in Appendix B. A major challenge in proving Theorems 5 and 6 is that there is no vector space structure on Riemannian manifolds. Thanks to the trigonometric distance bound proposed in Zhang and Sra (2016), we manage to obtain the regrets $\mathcal{O}\big(\sqrt{T}\big)$ and $\mathcal{O}\big(\log T\big)$ for g-convex and strongly g-convex loss functions, respectively. By gradually moving κ to zero, the results recover the regret bounds of Euclidean gradient descent of Zinkevich (2003) and Hazan et al. (2006).

Theorems 5 and 6 also reveal the influence of curvature on the regret bounds. Since $\zeta(\kappa, d)$ is an increasing function of κ, the upper regret bounds in the R-OGD algorithm are greater than those in Euclidean spaces and the increase of κ raises the upper regret bounds. Therefore, a proper Riemannian metric should be chosen in the optimization to avert high sectional curvature bounds.

3.3 Regret Lower Bound

This section is intended to answer the question of whether there exists an algorithm that attains a tighter regret bound than $\mathcal{O}\big(\sqrt{T}\big)$ for g-convex functions. Theorem 7 provides a negative answer.

Theorem 7 *Suppose that Assumptions 1-4 hold. Then for any Hadamard manifold $\mathcal{M}$, the Riemannian online convex optimization incurs the regret $\Omega(DL\sqrt{T})$ for any possible strategy in the worst case.*

The proof of Theorem 7 is in Appendix C. The result illustrates that, as in Euclidean spaces, the regret of a Riemannian online convex algorithm can not be less than $\Omega(\sqrt{T})$ in the worst case. Moreover, Theorem 7 shows that the regret of the R-OGD algorithm in Theorem 5 is tight up to a constant.

4. R-OCO with One-Point Bandit Feedback

In this section, we consider the Riemannian online convex optimization with the one-point bandit feedback. We first present the Riemannian bandit algorithm (R-BAN) on Hadamard homogeneous manifolds and then analyze (expected) regret bounds for the R-BAN. Here and subsequently, we denote by $\mathcal{B}_\delta(x)$ the ball centered at x with radius δ and by $\mathcal{S}_\delta(x)$ the sphere centered at x with radius δ.

4.1 Riemannian Bandit Algorithm

In the bandit setting, Assumptions 2-4 are slightly modified as follows.

Assumption 8 *$\mathcal{M}$ is an n-dimensional homogeneous Hadamard manifold with the sectional curvature lower bounded by a constant κ.*

The homogeneous Hadamard manifold has been widely studied in differential geometry (Berestovskii and Nikonorov, 2020; Berger, 2009). The homogeneity has received much attention in machine learning (Tang et al., 2020; Tupker et al., 2021; Bronstein et al., 2021). It has been seen that many manifolds often considered in Riemannian optimization, such as the Euclidean space $\mathbb{R}^n$, the Hyperbolic space H^n, and the manifold of SPD matrices, are Hadamard homogeneous manifolds. Note that on homogeneous manifolds, the volume and surface area of a ball are only related to the radius but not to the center of the ball. Thus, we denote by V_δ the volume of $\mathcal{B}_\delta(x)$ and $\mathcal{S}_\delta$ as the surface area of $\mathcal{S}_\delta(x)$ for all points x over the homogeneous manifold $\mathcal{M}$.

Assumption 9 *There exists an interior point $p \in \mathcal{K}$ such that the set $\mathcal{K}$ contains a ball with radius r centered at p, and $\mathcal{K}$ is also contained in a ball with radius D, i.e.,*

$$\mathcal{B}_r(p) \subset \mathcal{K} \subset \mathcal{B}_D(p).$$

Assumption 10 *For any $t = 1, \ldots, T$, $\mathbf{f}_t$ is differentiable, g-L-Lipschitz and the function value of $\mathbf{f}_t$ is bounded by C.*

Inspired by the Euclidean bandit algorithm, we replace the gradient $\nabla \mathbf{f}_t(x_t)$ with a randomized estimator g_t and propose our R-BAN in Algorithm 2 over Hadamard homogeneous manifolds.

Algorithm 2: Riemannian Bandit Algorithm (R-BAN)

Input: Manifold $\mathcal{M}$, time T, step sizes (or schedule) α_t, parameters δ, τ.
Output: Sequence $\{x_t\}_{t=1,\dots,T}$
for $t = 1$ *to* T **do**
Choose x_t uniformly from $\mathcal{S}_\delta(y_t)$ and play x_t;
Observe the loss $\mathbf{f}_t(x_t)$ and compute

$$\mathbf{g}_t = \mathbf{f}_t(x_t)\frac{\exp_{y_t}^{-1}(x_t)}{\|\exp_{y_t}^{-1}(x_t)\|};$$

Update y_{t+1} with

$$\begin{cases} \tilde{y}_{t+1} = \exp_{y_t}(-\alpha_t \mathbf{g}_t) \\ y_{t+1} = \mathscr{P}_{(1-\tau)\mathcal{K}}\big(\mathscr{P}_{\mathcal{K}}(\tilde{y}_{t+1})\big), \end{cases}$$

where the symbols $\mathscr{P}_{\mathcal{K}}$ and $\mathscr{P}_{(1-\tau)\mathcal{K}}$ represent the projection mappings onto the feasible set $\mathcal{K}$ and the shrinking set $(1-\tau)\mathcal{K} = \{\exp_p((1-\tau)u) | u = \exp_p^{-1}(x), x \in \mathcal{K}\}$, respectively.
Return x_t and suffer the loss $\mathbf{f}_t(x_t)$;
end

4.2 Challenge from Geometry

Since Algorithm 2 is an extension of the Euclidean bandit algorithm by Flaxman et al. (2005), it is worth reviewing the analysis in the work by Flaxman et al. (2005). In the Euclidean setting, we uniformly choose x_t on the $\mathcal{S}_\delta(y_t)$ and update y_t by the rule

$$\begin{cases} \mathbf{g}_t^E = \mathbf{f}(x_t)\frac{x_t - y_t}{\|x_t - y_t\|}, \\ y_{t+1} = \mathscr{P}_{(1-\tau)\mathcal{K}}(y_t - \alpha \mathbf{g}_t^E). \end{cases} \tag{2}$$

The basic idea for the analysis is to introduce the *smoothed loss function* (Flaxman et al., 2005)

$$\hat{\mathbf{f}}_t^E(x) = \mathbb{E}_{u \in \mathcal{B}_\delta(x)}[\mathbf{f}_t(u)] = \frac{1}{V_\delta}\int_{\mathcal{B}_\delta(x)} \mathbf{f}_t(u)du,$$

where $\hat{\mathbf{f}}_t^E$ is a convex approximation of $\mathbf{f}_t$ when δ is small. It is shown that $\frac{n}{\delta}\mathbf{g}_t^E$ is an unbiased estimator of the gradient $\nabla\hat{\mathbf{f}}_t^E(y_t)$, hence the bandit algorithm is actually an expected gradient descent method (Flaxman et al., 2005) with the loss function $\hat{\mathbf{f}}_t^E$. In this way, an Euclidean regret bound of the bandit algorithm is established by Flaxman et al. (2005).

Back to the Riemannian case, we attempt to generalize the analysis of Flaxman et al. (2005) in parallel by defining the "Riemannian version" of the smoothed loss function,

$$\hat{\mathbf{f}}_t(x) = \mathbb{E}_{u \in \mathcal{B}_\delta(x)}[\mathbf{f}_t(u)] = \frac{1}{V_\delta}\int_{\mathcal{B}_\delta(x)} \mathbf{f}_t(u)\omega,$$

where ω is the volume element with the respect to the metric g. Analyzing the smoothed loss function in the Riemannian space, however, is fundamentally challenging due to the following difficulties.

The gradient is hard to estimate. Estimating the gradient of $\hat{\mathbf{f}}_t$ is quite different from that in Euclidean spaces, due to absence of the commutativity of the derivative operator ∇ and the integration operator $\int_{\mathcal{B}_\delta(y_t)}$. In Euclidean spaces, the derivative operator ∇ commutes with the integration operator $\int_{\mathcal{B}_\delta(y_t)}$. Accordingly, for the Euclidean smoothed loss function $\hat{\mathbf{f}}_t^E$, there holds

$$\nabla\hat{\mathbf{f}}_t^E(y_t) = \frac{1}{V_\delta}\nabla\int_{\mathcal{B}_\delta(y_t)}\mathbf{f}_t(u)du = \frac{1}{V_\delta}\int_{\mathcal{B}_\delta(y_t)}\nabla\mathbf{f}_t(u)du, \tag{3}$$

which implies $\frac{n}{\delta}E[\mathbf{g}_t^E] = \nabla\hat{\mathbf{f}}_t^E(y_t)$. However, on Riemannian manifolds the derivative operator ∇ does not commute with the integration operator $\int_{\mathcal{B}_\delta(y_t)}$. Consequently, the equation (3) fails to hold for functions on Riemannian manifolds.

The convexity may be lost. Another essential challenge for regret analysis is the convexity of $\hat{\mathbf{f}}_t$. In Euclidean spaces, one can easily conclude the convexity of $\hat{\mathbf{f}}_t^E$. However, the convexity may not hold for Riemannian manifolds. Through calculation, the Hessian of $\hat{\mathbf{f}}_t$ on a Riemannian manifold is

$$\nabla^2(\hat{\mathbf{f}}_t)(X,X) = \frac{1}{V_\delta}\int_{\mathcal{B}_\delta(x)}(\nabla^2(\mathbf{f}_t)(\eta,\eta)(u) + \langle\nabla_\eta\eta, \nabla\mathbf{f}_t(u)\rangle)\omega,$$

where η is a Killing field with $\eta(x) = X$. Since the quadratic form $\nabla^2(\mathbf{f}_t)(\eta,\eta)(x) + \langle\nabla_\eta\eta, \nabla\mathbf{f}_t(x)\rangle$ can be negative at some $\eta \in T_p\mathcal{M}$, the g-convexity of $\hat{\mathbf{f}}_t$ is violated for some small δ.

4.3 Gradient Bound and Approximate G-convexity

We first propose a key technique to analyze the derivative of local integration by introducing the Killing vector field. With the help of this technique, we manage to estimate the gradient of $\hat{\mathbf{f}}_t = \frac{1}{V_\delta}\int_{\mathcal{B}_\delta(x)}\mathbf{f}_t(u)\omega$ in Lemma 11.

Lemma 11 *Suppose $\mathcal{M}$ is a homogeneous Hadamard manifold, $\mathbf{f}$ is a C^1 function on $\mathcal{M}$ with bound C, and $x \in \mathcal{M}$. Denote*

$$\mathbf{g}(u) = \mathbf{f}(u)\frac{\exp_x^{-1}(u)}{\|\exp_x^{-1}(u)\|}.$$

Then for a fixed $\delta > 0$, the following statements hold.

(i) If u is uniformly chosen from $\mathcal{S}_\delta(x)$, then $\frac{S_\delta}{V_\delta}\mathbf{g}(u)$ is an unbiased estimator of $\nabla\hat{\mathbf{f}}(x)$, i.e.,

$$\mathbb{E}_{u\in\mathcal{S}_\delta(x)}\Big[\frac{S_\delta}{V_\delta}\mathbf{g}(u)\Big|x\Big] = \frac{1}{V_\delta}\int_{\mathcal{S}_\delta(x)}\mathbf{f}(u)\frac{\exp_x^{-1}(u)}{\|\exp_x^{-1}(u)\|}\omega_{\mathcal{S}_\delta(x)} = \nabla\hat{\mathbf{f}}(x), \quad \forall x \in \mathcal{M};$$

(ii) If the sectional curvature of $\mathcal{M}$ is bounded lower by κ, then the estimator $\frac{S_\delta}{V_\delta}\mathbf{g}(u)$ is bounded, i.e.,

$$\mathbb{E}_{u\in\mathcal{S}_\delta(x)}\Big[\Big\|\frac{S_\delta}{V_\delta}\mathbf{g}\Big\|\Big|x\Big] \le \frac{S_\delta}{V_\delta}C \le (\frac{n}{\delta}+n|\kappa'|\delta)C,$$

where $\kappa' = \min\{\kappa, 0\}$.

The proof of Lemma 11 can be found in Appendix D. The first part of the lemma establishes a gradient estimator of $\hat{\mathbf{f}}_t(x)$, while the second part gives us an easy-to-compute bound of the gradient. In the proof, we develop a technique that transforms a derivation of an integration on $\mathcal{B}_\delta(x)$ to an integration of a derivative of corresponding Killing vector field on $\mathcal{B}_\delta(x)$, i.e.,

$$X\big(\int_{\mathcal{B}_\delta(x)}\mathbf{f}(u)\omega\big) = \int_{\mathcal{B}_\delta(x)}\eta(\mathbf{f}(u))\omega, \tag{4}$$

where η is a Killing field such that $\eta(x) = X$. This technique does not rely on the curvature and other specific manifold structures.

We also notice that although the function $\hat{\mathbf{f}}_t$ may not be g-convex, it is very close to be g-convex. We introduce the following definition.

Definition 12 *A function $\mathbf{f}: \mathcal{K}\subset\mathcal{M}\to\mathbb{R}$ is called to be*

(i) λ-sub g-convex if there exists a constant $\lambda\ge 0$ such that for any $x, y\in\mathcal{M}$

$$\mathbf{f}(y) - \mathbf{f}(x) - \langle\nabla\mathbf{f}(x), \exp_x^{-1}(y)\rangle \ge -\lambda.$$

(ii) μ-strongly λ-sub g-convex if there exist two constants $\lambda, \mu \ge 0$ such that for any $x, y\in\mathcal{M}$

$$\mathbf{f}(y) - \mathbf{f}(x) - \langle\nabla\mathbf{f}(x), \exp_x^{-1}(y)\rangle - \frac{\mu}{2}d^2(x,y) \ge -\lambda.$$

Lemma 13 *Suppose that $(\mathcal{M}, g)$ is a Hadamard homogeneous manifold. If $\mathcal{K}$ is a g-convex and bounded set of $\mathcal{M}$, then there exists a constant $\rho\ge 0$ depending only on the set $\mathcal{K}$ such that, the following statements hold.*

(i) For any g-convex and g-L-Lipschitz function $\mathbf{f}$, the smoothed function $\hat{\mathbf{f}}$ is $2\rho\delta L$-sub g-convex.

(ii) For any μ-strongly g-convex and g-L-Lipschitz function $\mathbf{f}$, the smoothed function $\hat{\mathbf{f}}$ is μ-strongly $2(\rho\delta L + \mu D\delta)$-sub g-convex.

The proof of Lemma 13 is in Appendix D. It is worth mentioning that the constant ρ describes how close a smoothed function is to being g-convex. Notice that

$$\rho = \sup_{x,y,u\in\mathcal{K}} |\frac{1}{\sqrt{G}}\frac{\partial}{\partial x_i}\big(\sqrt{G}\exp_u^{-1}\phi(u)\big)^i| \quad s.t. \quad \phi(x) = y$$

does not depend on the function $\hat{\mathbf{f}}_t$, and the time T. Moreover, for a given manifold $\mathcal{M}$, once the set $\mathcal{K}$ is fixed and the explicit expression of ϕ is given, we can compute the constant ρ as a finite number. We briefly list the bound of ϕ in the following two types of manifolds.

(i) Let manifold $\mathcal{M}$ be a Euclidean space. We can find the isometry $\phi(z) = z + y - x$. Hence we can conclude that $\rho = 0$ and $\hat{\mathbf{f}}$ is convex, which coincides with the result in Euclidean spaces.

(ii) Let $\mathcal{M}$ be a 2-dimensional Poincaré disk. Then the isometry ϕ from x to y has the closed form of

$$\phi = \phi_x \circ \phi_y,$$

where $\phi_x(z) = \frac{x-z}{1-\bar{x}z}$ and $\phi_y(z) = \frac{y-z}{1-\bar{y}z}$. Therefore, if $\mathcal{K}$ has a diameter D, we can figure out a bound of ρ in

$$\rho \leq 16\frac{1+\tanh(D/2)}{1-\tanh(D/2)}\left(\frac{1}{1-\tanh(2D)^2} + \frac{D}{\tanh(D/2)}\right),$$

which implies that ρ may grow exponentially with respect to D.

Although the value of ρ is generally difficult to calculate, our algorithm analysis and parameter selection do not depend on the specific value of ρ (see Theorems 14 and 15).

4.4 Regret Bounds

With the above effort, we now carry out the analysis of the expected regret bounds of Algorithm 2. Denote $B = \frac{n}{\delta} + n|\kappa|$.

Theorem 14 (Convex Cases) *Suppose that Assumptions 1, 8, 9 and 10 hold, and $\mathbf{f}_t$ is g-convex for any $t = 1, \ldots, T$. If we take $\delta = T^{-\frac{1}{4}}$, $\theta = \frac{\sqrt{\kappa}(D+r)}{\sinh\sqrt{\kappa}(D+r)}$, $\tau = \frac{\delta}{r\theta}$, and $\alpha_t = \frac{D}{C\sqrt{\zeta(\kappa,D)T}}$, then the expected regret of Algorithm 2 is upper bounded by*

$$\mathbb{E}[\mathsf{Reg}(T)] \leq n|\kappa|DC\sqrt{\zeta(\kappa,D)}T^{\frac{1}{4}} + \Big(nD\sqrt{\zeta(\kappa,D)} + \frac{2D^2}{r\theta} + (3 + \frac{D}{r\theta} + 2\rho)L\Big)T^{\frac{3}{4}}.$$

Theorem 15 (Strongly Convex Cases) *Suppose that Assumptions 1, 8, 9 and 10 hold, and $\mathbf{f}_t$ is μ-strongly g-convex for any $t = 1, \ldots, T$. If we take $\delta = \sqrt[3]{\frac{n^2C^2(1+\log T)}{T}}$, $\theta = \frac{\sqrt{\kappa}(D+r)}{\sinh\sqrt{\kappa}(D+r)}$, $\tau = \frac{\delta}{r\theta}$, and $\alpha_t = \frac{B}{\mu t}$, then the expected regret of Algorithm 2 is upper bounded by*

$$\begin{aligned}\mathbb{E}[\mathsf{Reg}(T)] \leq\ & \frac{2n^{\frac{8}{3}}C^{\frac{8}{3}}D\kappa^2\zeta(\kappa,D)}{\mu} \\ & + n^{\frac{2}{3}}C^{\frac{2}{3}}\Big(\frac{\zeta(\kappa,D)}{\mu} + 3L + \frac{2D^2}{r\theta} + \frac{DL}{r\theta} + 2\rho L + 2D\mu\Big)(1+\log T)^{\frac{1}{3}}T^{\frac{2}{3}}.\end{aligned}$$

The proofs of Theorems 14 and 15 can be seen in Appendix E. Theorems 14 and 15 show that the regrets of the Riemannian bandit algorithm achieve $\mathcal{O}\big(T^{\frac{3}{4}}\big)$ and $\mathcal{O}\big(T^{\frac{2}{3}}\big)$ for g-convex loss functions and strongly g-convex functions on homogeneous Hadamard manifolds, which are same as the regret bounds in Euclidean spaces (Flaxman et al., 2005).

We also note that, different from bandit algorithms in the Euclidean space, Theorems 14 and 15 introduce the parameter θ in the selection of $\tau = \frac{\delta}{r\theta}$ to ensure that the ball

$B_\delta(y_t) = B_{\theta\cdot\tau r}(y_t)$ always remains within the feasible set $\mathcal{K}$, thereby ensuring the feasibility of the algorithm (see Lemma 45 for details). Because the value of θ in Theorems 14 and 15 is chosen to ensure the feasibility for all possible subsets $\mathcal{K}$, we may find that the chosen value of θ is too small and conservative in practical situations. This may lead to an over-shrinkage of the set $(1-\tau)\mathcal{K}$. However, we would like to point out that, for a specific subset $\mathcal{K}$, as long as the value of θ satisfies the feasibility requirement of $B_\delta(y_t)$, the same regrets result as Theorems 14 and 15 can be obtained. This means that we can choose a much larger value of θ specifically tailored to the subset $\mathcal{K}$ to achieve better practical application results.

5. R-OCO with Two Point Bandit Feedback

In this section, we consider the Riemannian online convex optimization with the two-point bandit feedback. We first propose a Riemannian two-point bandit algorithm (R-2-BAN), which estimates the gradient of $\mathbf{f}_t(y_t)$ with two queries of values around y_t, and then analyze regret bounds for g-convex and strongly g-convex functions.

5.1 Riemannian Two-point Bandit Algorithm

In this subsection, we propose our R-2-BAN algorithm in Algorithm 3 with an additional assumption.

Assumption 16 *$\mathcal{M}$ is an n-dimensional symmetric Hadamard manifold with the sectional curvature lower bounded by a constant κ.*

The assumption of symmetry is important in the two-point bandit feedback setting. From symmetry, it is shown that for any $y \in \mathcal{M}$, the "*minus*" map

$$-x := \exp_y(-\exp_y^{-1}(x))$$

defines a isometry in $\mathcal{M}$. The result implies that the uniform distribution on the geodesic sphere $\mathcal{S}_\delta(y)$ is symmetric, which is key insight in the Euclidean two bandit algorithm (Agarwal et al., 2010), as well as in our R-2-BAN analysis. The symmetric Hadamard manifold is widely studied in differential geometry (Berestovskii and Nikonorov, 2020; Berger, 2009). Moreover, many manifolds with great practical value are symmetric manifolds, such as the Euclidean space $\mathbb{R}^n$, the Hyperbolic space H^n, and the manifold of SPD matrices.

5.2 Regret Bound

This subsection studies the regret of the R-2-BAN, which is defined as

$$\mathsf{Reg}(T) = \sum_{t=1}^{T} \frac{1}{2}(\mathbf{f}_t(x_{t,1}) + \mathbf{f}_t(x_{t,2})) - \min_{x\in\mathcal{K}} \sum_{t=1}^{T} \mathbf{f}_t(x)$$

Since the uniform distribution on geodesic spheres is symmetric, we derive that the two-point estimator $\frac{S_\delta}{V_\delta}\tilde{\mathbf{g}}$ is also unbiased and bounded gradient estimator of the smoothed function $\hat{\mathbf{f}}(x) = \frac{1}{V_\delta}\int_{\mathcal{B}_\delta(x)} \mathbf{f}\omega$.

Algorithm 3: Riemannian Two-point Bandit Algorithm (R-2-BAN)

Input: Manifold $\mathcal{M}$, time T, step size (or schedule) α_t, parameter δ, τ.
Output: Sequence $\{x_t\}_{t=1,\dots,T}$
for $t = 1$ *to* T *do* **do**
Pick $x_{t,1}$ uniformly from $\mathcal{S}_\delta(y_t)$ and set $x_{t,2} = -x_{t,1}$;
Play $x_{t,1}$ and $x_{t,2}$, then observe $\mathbf{f}_t(x_{t,1})$ and $\mathbf{f}_t(x_{t,2})$;
Compute

$$\tilde{\mathbf{g}}_t = \frac{1}{2}\big(\mathbf{f}_t(x_{t,1}) - \mathbf{f}_t(x_{t,2})\big)\frac{\exp_{y_t}^{-1}(x_{t,1})}{\|\exp_{y_t}^{-1}(x_{t,1})\|}$$

Update y_{t+1} with

$$\begin{cases} \tilde{y}_{t+1} = \exp_{y_t}(-\alpha_t \tilde{\mathbf{g}}_t) \\ y_{t+1} = \mathscr{P}_{(1-\tau)\mathcal{K}}\big(\mathscr{P}_{\mathcal{K}}(\tilde{y}_{t+1})\big), \end{cases}$$

where the symbols $\mathscr{P}_{\mathcal{K}}$ and $\mathscr{P}_{(1-\tau)\mathcal{K}}$ represent the projection mappings onto the feasible set $\mathcal{K}$ and the shrinking set $(1-\tau)\mathcal{K} = \{\exp_p((1-\tau)u) | u = \exp_p^{-1}(x), x \in \mathcal{K}\}$, respectively.
Return x_t and suffer the loss $\frac{1}{2}\big(\mathbf{f}_t(x_{t,1}) + \mathbf{f}_t(x_{t,2})\big)$;
end

Lemma 17 *Suppose $\mathcal{M}$ is a symmertic Hadamard manifold and $\mathbf{f}$ is a g-L-Lipschitz function on $\mathcal{M}$. Then for a fixed $\delta > 0$ the gradient estimator*

$$\frac{S_\delta}{V_\delta}\tilde{\mathbf{g}} = \frac{S_\delta}{2V_\delta}\big(\mathbf{f}(u) - \mathbf{f}(-u)\big)\frac{\exp_x^{-1}(u)}{\|\exp_x^{-1}(u)\|}$$

satisfies the following.

(i) $\mathbb{E}_{u\in\mathcal{S}_\delta(x)}\left[\frac{S_\delta}{V_\delta}\tilde{\mathbf{g}}\middle| x\right] = \nabla\hat{\mathbf{f}}(x), \quad \forall x \in \mathcal{M};$

(ii) $\mathbb{E}_{u\in\mathcal{S}_\delta(x)}\left[\|\frac{S_\delta}{V_\delta}\tilde{\mathbf{g}}\|\middle| x\right] \le \frac{S_\delta}{V_\delta}\delta L \le nL(1+\kappa\delta^2).$

Then we carry out the regret analysis in Theorem 18 and Theorem 19. Notice that $B = \frac{n}{\delta} + n|\kappa|\delta$ and ρ is the constant in Lemma 13 that only depends on $\mathcal{K}$.

Theorem 18 (Convex Cases) *Suppose that Assumptions 1, 8, 9, 10 and 16 hold, and $\mathbf{f}_t$ is g-convex for any $t = 1, \dots, T$. If we take $\delta = \frac{1}{\sqrt{T}}$, $\theta = \frac{\sqrt{\kappa}(D+r)}{\sinh\sqrt{\kappa}(D+r)}$, $\tau = \frac{\delta}{r\theta}$, and $\alpha_t = \frac{n}{\delta L\sqrt{\zeta(\kappa,D)T}}$, then the expected regret of Algorithm 3 is upper bounded by*

$$\mathbb{E}[\mathsf{Reg}(T)] \le n\kappa DL\sqrt{\zeta(\kappa,D)}\frac{1}{\sqrt{T}} + \Big(nDL\sqrt{\zeta(\kappa,D)} + \frac{2D^2}{r\theta} + (3 + \frac{D}{r\theta} + 2\rho)L\Big)\sqrt{T}.$$

Theorem 19 (Strongly Convex Cases) *Suppose that Assumptions 1, 8, 9, 10 and 16 hold, and $\mathbf{f}_t$ is μ-strongly g-convex for any $t = 1, \dots, T$. If we take $\delta = \frac{1+\log T}{T}$, $\theta =$*

$\frac{\sqrt{\kappa}(D+r)}{\sinh\sqrt{\kappa}(D+r)}$, $\tau = \frac{\delta}{r\theta}$, *and* $\alpha_t = \frac{B}{\mu t}$, *then the expected regret of Algorithm 3 is upper bounded by*

$$\mathbb{E}[\mathsf{Reg}(T)] \leq \frac{3n^2C^2\kappa^2\zeta(\kappa,D)}{2\mu} + \Big(\frac{\zeta(\kappa,D)n^2L^2}{\mu} + \frac{2D^2}{r\theta} + (3+\frac{D}{r\theta}+2\rho)L + 2\mu D\Big)(1+\log T).$$

The proofs of Theorems 18 and 19 can be seen in Appendix F. Theorems 18 and 19 show that regrets of the Riemannian two-point bandit algorithm achieve $\mathcal{O}\big(\sqrt{T}\big)$ and $\mathcal{O}\big(\log T\big)$ for g-convex and strongly g-convex functions on symmetric Hadamard manifolds, which improves the regret bounds of the Riemannian one point bandit algorithm. Such improvement is consistent with the results in Euclidean spaces (Agarwal et al., 2010).

6. Generalizing to Non-Hadamard Cases

In this section, we generalize the R-OGD, R-BAN and R-2-BAN algorithms to a manifold $\mathcal{M}$ with sectional curvature lower bounded by κ and upper bounded by $K > 0$. However, positive sectional curvature can make projection maps no longer constrictive so that a straightforward generalization may fail to be no-regret. We demonstrate how the non-Hadamard structure affects the property of projection maps and how non-constrictive projection maps cause trouble in regret analysis. Furthermore, we try to address the difficulty withour assuming the invariance condition adopted in the previous work (Ahn and Sra, 2020; Alimisis et al., 2021), and then construct no-regret bounds of Algorithm 1, 2 and 3.

6.1 Impact of Positive Curvature in Regret Analysis

The projection map can be non-constrictive for non-Hadamard manifolds, since the positive sectional curvature affects the g-convexity of the norm of Jacobi fields. The following example provides an illustration.

Example 1 *Suppose that $\mathcal{M}$ is a manifold with positive sectional curvature, and $\mathcal{K}$ is a geodesic $\xi(s) : [0,1] \to \mathcal{M}$. Let $U(s)$ be a parallel vector field on ξ that is normal to ξ. Then we consider the following variation*

$$\gamma_s(t) : [0,1]\times[0,\delta] \to \mathcal{M}, \quad (s,t) \to \exp_{\xi(s)}(tU(s)).$$

For a sufficiently small δ_1 we have,

$$\mathscr{P}_{\mathcal{K}}(\gamma_s(t)) = \xi(s), \quad \forall t \leq \delta_1.$$

In addition, we notice that derivatives of the length of the $s-$curve, namely $L(t)$, are characterized by the first and second variation formula (do Carmo, 1992),

$$L'(0) = \langle U(b), \xi(b)\rangle - \langle U(a), \xi(a)\rangle = 0$$
$$L''(0) = \int_0^1 |\nabla_\gamma U|^2 - R(U,\dot\xi,U,\dot\xi)ds = \int_0^1 -R(U,\dot\xi,U,\dot\xi)ds$$

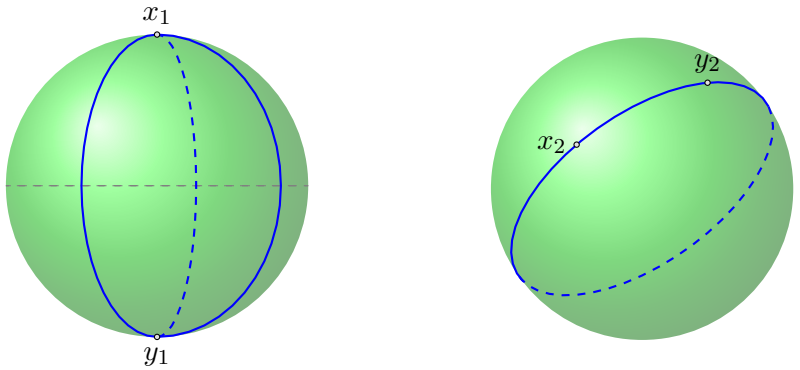
Because the sectional curvature is positive, we found $L''(0) < 0$, which means that $L(0)$ is a local maximum. As a result, we can take a sufficiently small $\delta_2 > 0$ such that

$$L(t) < L(0) = d(\xi(a), \xi(b)), \quad t \leq \delta_2.$$

Take $\delta_3 = \min\{\delta_1, \delta_2\}$, $p = \gamma_0(\delta_3)$ and $q = \gamma_1(\delta_3)$ we have

$$d(p, q) \leq L(\delta_3) < L(0) = d(\xi(a), \xi(b)) = d(\mathscr{P}_{\mathcal{K}}(p), \mathscr{P}_{\mathcal{K}}(q)),$$

which indicates the non-expansiveness of the projection map $\mathscr{P}_{\mathcal{K}}$ fails to hold. ■



(a) Minimal geodesics are not unique (b) Geodesics are loops

Figure 1: Examples in $\mathbb{S}^2$

One may try to sidestep the projection map by assuming compactness of $\mathcal{M}$ and $\mathcal{K} = \mathcal{M}$. However, these assumptions do not work as g-convexity on non-Hadamard manifolds only holds locally. On one hand, positive sectional curvature admits conjugate points, where the connecting geodesic is not unique, leading $\mathcal{K}$ no longer to be uniquely g-convex. On the other hand, global g-convex functions may not exist on compact manifolds. From the study by Yau (1974), the existence of nontrivial global g-convex functions implies infinity of volume. As a result, there is no global g-convex function apart from constant functions on compact Riemannian manifolds.

We take examples on the sphere $\mathbb{S}^2$ to illustrate the above points.

Example 2 *(i) Let x_1, y_1 be the north pole and the south pole of the sphere $\mathbb{S}^2$ (see Figure 1 (a)). Then every arc connecting x and y is a geodesic. Therefore $\mathbb{S}^2$ is not uniquely g-convex.*

(ii) Suppose that $\mathbf{f}$ is a g-convex function. Since for any x_2, y_2 in $\mathbb{S}^2$, the geodesic is the great circle connecting x_2, y_2 (see Figure 1 (b)), we can choose a geodesic loop which starts at x_2 and ends at x_2. By g-convexity, we have

$$\mathbf{f}(y_2) \leq (1-t)\mathbf{f}(x_2) + t\mathbf{f}(x_2) = \mathbf{f}(x_2).$$

Choosing the geodesic loop that starts at y_2, we can obtain $\mathbf{f}(x_2) \leq \mathbf{f}(y_2)$. Therefore there must hold $\mathbf{f}(x_2) = \mathbf{f}(y_2)$, which indicates that $\mathbf{f}$ is actually a constant.

Our regret analysis of Algorithms 1, 2 and 3 on Hadamard manifolds greatly depends on non-expansiveness of projection maps. During the analysis, we use Lemma 35 to bound the loss $\mathbf{f}_t(x_t) - \mathbf{f}_t(x^*)$ with

$$\begin{aligned}\mathbf{f}_t(x_t) - \mathbf{f}_t(x^*) &\leq \langle -\nabla \mathbf{f}_t(x_t), \exp_{x_t}^{-1}(x^*)\rangle \\ &\leq \frac{1}{2\alpha_t}(d^2(x_t, x^*) - d^2(\exp_{x_t}(-\alpha_t \nabla \mathbf{f}_t(x_t), x^*)) + \frac{1}{2}\zeta(\kappa, d(x_t, x))\alpha_t \|\nabla \mathbf{f}_t(x_t)\|^2.\end{aligned} \tag{5}$$

Denote the intermediate point $\tilde{x}_{t+1} = \exp_{x_t}(-\alpha_t \nabla \mathbf{f}_t(x_t))$. Applying non-expansiveness of the projection map $\mathscr{P}_{\mathcal{K}}$, we have

$$d(x_{t+1}, x^*) = d(\mathscr{P}_{\mathcal{K}}(\tilde{x}_{t+1}), x^*) \leq d(\exp_{x_t}(\tilde{x}_{t+1}, x^*)). \tag{6}$$

Combing (5) and (6) we get

$$\mathbf{f}_t(x_t) - \mathbf{f}_t(x^*) \leq \frac{1}{2\alpha_t}(d^2(x_t, x^*) - d^2(x_{t+1}, x^*)) + \frac{1}{2}\zeta(\kappa, d(x_t, x))\alpha_t \|\nabla \mathbf{f}_t(x_t)\|^2, \tag{7}$$

which is a key step to get sublinear regrets, as we can rearrange the summation and cancel term by term. Unfortunately, when it turns to non-Hadamard manifolds, equations (6) and (7) no longer hold. Thus the loss $\mathbf{f}_t(x_t) - \mathbf{f}_t(x^*)$ can be only bounded with

$$\begin{aligned}\mathbf{f}_t(x_t) - \mathbf{f}_t(x^*) &\leq \frac{1}{2\alpha_t}(d^2(x_t, x^*) - d^2(\tilde{x}_{t+1}, x^*)) + \frac{1}{2}\zeta(\kappa, d(x_t, x))\alpha_t \|\nabla \mathbf{f}_t(x_t)\|^2, \\ &\leq \frac{1}{2\alpha_t}(d^2(x_t, x^*) - d^2(x_{t+1}, x^2)) + \frac{1}{2}\zeta(\kappa, d(x_t, x))\alpha_t \|\nabla \mathbf{f}_t(x_t)\|^2 \\ &+ \frac{1}{2\alpha_t}(d^2(x_{t+1}, x^*) - d^2(\exp_{x_t}(\tilde{x}_{t+1}, x^*)),\end{aligned}$$

and there is an additional projection error term

$$\sum_{t=1}^{T} \frac{1}{2\alpha_t}(d^2(x_{t+1}, x^*) - d^2(\exp_{x_t}(\tilde{x}_{t+1}, x^*)) \tag{8}$$

in the final regret, which is not clear to be sublinear.

6.2 Dealing with Projection Error and Regret Analysis

An approach in recent research to deal with non-constrictive projection maps on non-Hadamard manifolds is to omit projections by an *invariant condition*. The invariant condition assumes that all iterations of the algorithm remain in the unique g-convex feasible set $\mathcal{K}$ (Zhang and Sra, 2018; Ahn and Sra, 2020; Alimisis et al., 2021). In this section, we try to remove the invariant condition assumption by developing analysis on the projection error term (8) with control of the step size α_t.

First, we require the feasible set $\mathcal{K}$ to be bounded with the respect to the positive curvature bound K.

Assumption 20 *The diameter D of the g-convex subset $\mathcal{K}$ is less than $\frac{\pi}{2\sqrt{K}}$.*

Assumption 20 is a standard condition in the literature (e.g., Zhang and Sra, 2018; Ahn and Sra, 2020; Alimisis et al., 2021). From the conjugate point theorem (Lemma 31), Assumption 20 guarantees that there is no pair of conjugate points on $\mathcal{K}$, thus $\mathcal{K}$ is uniquely g-convex and the inverse exponential map $\exp_x^{-1}(\cdot)$ can be defined throughout $\mathcal{K}$ (even in the area that slightly deviates from $\mathcal{K}$). Moreover, the Hessian comparison theorem (Lemma 33) shows that when the diameter of $\mathcal{K}$ is greater than $\frac{\pi}{\sqrt{K}}$, the subset $\mathcal{K}$ may be "infinitely curved", i.e., the Hessian of the distance function $d(x,\cdot)$ reaches the infinity. Consequently, Assumption 20 is essential to establish theoretical regret bounds.

In practical applications, the feasible set $\mathcal{K}$ depends on the prior knowledge, physical constraints, or even artificial tuning. Assumption 20 indicates that, in order to have guaranteed regret bounds, the choice of $\mathcal{K}$ should rely on the curvature bound K for non-Hadamard cases. Besides, our experiments (see Subsection 7.3) indicate that a feasible set $\mathcal{K}$ with a much larger diameter than $\frac{\pi}{2\sqrt{K}}$ does not inherently impact the performance of the proposed algorithms.

Under Assumption 20, projection error term (8) can be fixed by Lemma 21. Let we denote

$$\sigma(K,d)=\begin{cases}0, & \frac{\pi}{2\sqrt{K}}<d\leq\frac{\pi}{\sqrt{K}} \text{ or } K\leq 0\\ -\sqrt{K}d\cot(\sqrt{K}d), & d\leq\frac{\pi}{2\sqrt{K}} \text{ and } K>0.\end{cases}$$

Lemma 21 *Suppose $\mathcal{K}\subset\mathcal{M}$ with the radius $D<\frac{\pi}{2\sqrt{K}}$. Assume that the iteration is as follows with $\|\alpha_t g_t\|\leq D$,*

$$\begin{cases}\tilde{x}_{t+1}=\exp_{x_t}(\alpha_t g_t)\\ x_{t+1}=\mathscr{P}_{\mathcal{K}}(\tilde{x}_{t+1})\end{cases}$$

Then it holds that

$$\sum_{t=1}^{T}\frac{1}{2\alpha_t}(d^2(x_{t+1},x^*)-d^2(\tilde{x}_{t+1},x^*))\leq\sigma(K,2D)\sum_{t=1}^{T}\frac{1}{2}\alpha_t\|\mathbf{g}_t\|^2 \tag{9}$$

Now we look back on the R-OGD (Algorithm 1), the R-BAN (Algorithm 2) and the R-2-BAN (Algorithm 3). For g-convex cases, the condition (9) is generally fulfilled as

$$\begin{cases}\|\alpha_t\nabla\mathbf{f}(x_t)\|\leq\frac{D}{L\sqrt{\zeta(\kappa,D)T}}L\leq D, & \text{(R-OGD)}\\ \|\alpha_t\mathbf{g}_t\|\leq\frac{D}{C\sqrt{\zeta(\kappa,D)T}}C\leq D, & \text{(R-BAN)}\\ \|\alpha_t\tilde{\mathbf{g}}_t\|\leq\frac{D}{\delta L\sqrt{\zeta(\kappa,D)T}}(\frac{1}{2}2\delta L)\leq D. & \text{(R-2-BAN)}\end{cases}$$

Furthermore, the summation $\sum_{t=1}^{T}\alpha_t\|\mathbf{g}_t\|^2$ is $\mathcal{O}\big(\sqrt{T}\big)$ for g-convex cases in Algorithms 1, 2 and 3. Consequently, sublinear regret bounds of Algorithms 1, 2 and 3 can be achieved.

On the other hand, the condition (9) does not hold for strongly g-convex cases, since the strong convexity coefficient μ can be arbitrary. In these cases, we may set a sufficiently large constant c_0 to the step size $\alpha_t=\frac{1}{\mu(t+c_0)}$ (or $\alpha_t=\frac{B}{\mu(t+c_0)}$ in bandit settings) such that

$\|\alpha_t \nabla \mathbf{f}(x_t)\|$ (or $\|\alpha_t \mathbf{g}_t\|$, $\|\alpha_t \tilde{\mathbf{g}}_t\|$) is small enough and thus ensure our Algorithms 1, 2 and 3 continue to be no-regret.

In the following, we formally state our results of Algorithms 1, 2 and 3 over non-Hadamard cases under Assumptions 20, along with a modified lemma for the constants in sub g-convexity. The analysis is quite similar to that in Hadamard cases, so we only present the proof of strong g-convex cases in Appendix G.

Lemma 22 (modification of Lemma 13) *Suppose that $(\mathcal{M}, g)$ is a Riemannian manifold whose sectional curvature is bounded above by K and below by κ. Let $\mathcal{K}$ be a g-convex set of $\mathcal{M}$ with diameter D. Denote $\kappa' = \min\{\kappa, 0\}$ and $\iota = \frac{2\mathbf{s}(\kappa', D)}{\mathbf{s}(K, D)}$ (see (10)). Then there exists a constant $\rho \geq 0$ depending only on the set $\mathcal{K}$ such that, the following statements hold.*

(i) For any g-convex and g-L-Lipschitz function $\mathbf{f}$, the smoothed function $\hat{\mathbf{f}}$ is $\big(2\rho\delta L + (n + n|\kappa'|\delta^2)\pi^2 \iota L\delta\big)$-sub g-convex.

(ii) For any μ-strongly g-convex and g-L-Lipschitz function $\mathbf{f}$, the smoothed function $\hat{\mathbf{f}}$ is μ-strongly $\big(2\rho\delta L + 2\mu D\delta + (n + n|\kappa'|\delta^2)\pi^2 \iota L\delta\big)$-sub g-convex.

Theorem 23 *Suppose that the sectional curvature $\mathcal{M}$ is lower bounded by κ and upper bounded by K. Assume that the previous assumptions for the R-OGD, R-BAN, R-2-BAN and Assumption 20 hold. Denote $\kappa' = \min\{\kappa, 0\}$, $\iota = \frac{2\mathbf{s}(\kappa', D)}{\mathbf{s}(K, D)}$ and $\theta = \frac{\mathbf{s}(\kappa', D+r)}{\mathbf{s}(K, D+r)}$. Then for g-convex loss functions on non-Hadamard manifolds, the following statements hold.*

(i) Setting step size $\alpha_t = \frac{D}{L\sqrt{\zeta(\kappa, D)t}}$, the R-OGD algorithm achieves regret

$$\mathsf{Reg}(T) \leq \frac{3}{2} DL\sqrt{\zeta(\kappa, D)T} + \frac{DL\sigma(K, 2D)}{2\sqrt{\zeta(\kappa, D)}}\sqrt{T}.$$

(ii) Setting $\tau = \frac{\delta}{r\theta}$, $\alpha_t = \frac{D}{C\sqrt{\zeta(\kappa, D)T}}$ and $\delta = T^{-\frac{1}{4}}$, the R-BAN algorithm achieves regret

$$\begin{aligned}\mathbb{E}[\mathsf{Reg}(T)] \leq \Big(n|\kappa'|DC\big(\sqrt{(\zeta(\kappa, D)} + \frac{\sigma(K, 2D)}{2\sqrt{\zeta(\kappa, D)}}\big)\big) + n|\kappa'|\pi^2 \iota L\Big) T^{\frac{1}{4}} \\ + \Big(nDC\sqrt{(\zeta(\kappa, D)} + \frac{\sigma(K, 2D)}{2\sqrt{\zeta(\kappa, D)}} + 3L \\ + \frac{DL + 2D^2}{r\theta} + 2\rho L + n\pi^2 \iota L\Big) T^{\frac{3}{4}}.\end{aligned}$$

(iii) Setting $\tau = \frac{\delta}{r\theta}$, $\alpha_t = \frac{D}{\delta L\sqrt{\zeta(\kappa, D)T}}$ and $\delta = T^{-\frac{1}{2}}$, the R-2-BAN algorithm achieves regret

$$\begin{aligned}\mathbb{E}[\mathsf{Reg}(T)] \leq (n|\kappa'|DL(\sqrt{(\zeta(\kappa, D)} + \frac{\sigma(K, 2D)}{2\sqrt{\zeta(\kappa, D)}}) + n|\kappa'|\pi^2 \iota L) T^{-\frac{1}{2}} \\ + \Big(nDL(\sqrt{(\zeta(\kappa, D)} + \frac{\sigma(K, 2D)}{2\sqrt{\zeta(\kappa, D)}}) + 3L \\ + \frac{DL + 2D^2}{r\theta} + 2\rho L + n\pi^2 \iota L\Big)\sqrt{T}.\end{aligned}$$

Theorem 24 *Suppose that the sectional curvature $\mathcal{M}$ is lower bounded by κ and upper bounded by K. Assume that the previous assumptions for the R-OGD, R-BAN, R-2-BAN, and Assumptions 20 hold. Denote $\kappa' = \min\{\kappa, 0\}$, $B = \frac{n}{\delta} + n|\kappa'|\delta$, $\iota = \frac{2\mathbf{s}(\kappa', D)}{\mathbf{s}(K, D)}$ an and $\theta = \frac{\mathbf{s}(\kappa', D+r)}{\mathbf{s}(K, D+r)}$. Then for μ-strongly g-convex loss fucntions on non-Hadamard manifolds, the following statements hold.*

(i) Setting $c_0 \geq \frac{L}{\mu D}$ and step size $\alpha_t = \frac{1}{\mu(t+c_0)}$, the R-OGD algorithm achieves regret

$$\mathsf{Reg}(T) \leq \frac{D^2 \mu c_0}{2} + \frac{1}{2}(\zeta(\kappa, D) + \sigma(K, 2D))G^2\big(1 + \log(T + c_0)\big).$$

(ii) Setting $c_0 \geq \frac{BC}{\mu D}$, $\tau = \frac{\delta}{r\theta}$, $\alpha_t = \frac{B}{\mu(t+c_0)}$ and $\delta = \sqrt[3]{\frac{nC(1+\log(T+c_0))}{T}}$, the R-BAN algorithm achieves regret

$$\begin{aligned}\mathbb{E}[\mathsf{Reg}(T)] \leq{}& \frac{D^2\mu c_0}{2} + \frac{4(c_0+1)}{\mu}(\zeta(\kappa, D) + \sigma(K, 2D))C^{\frac{8}{3}} n^{\frac{8}{3}} \kappa'^2 D \\ &+ \Bigg(\big(\frac{\zeta(\kappa, D) + \sigma(K, 2D)}{\mu} + |\kappa'| D^2 \pi^2 \iota\big) n^{\frac{4}{3}} C^{\frac{4}{3}} \\ &\qquad + \big(n\pi^2\iota + 2\rho L + 3L + \frac{DL + 2D^2}{r\theta} + 2\mu D\big) n^{\frac{1}{3}} C^{\frac{1}{3}} \Bigg) (1 + \log(T + c_0))^{\frac{1}{3}} T^{\frac{2}{3}}.\end{aligned}$$

(iii) Setting $c_0 \geq \frac{B\delta L}{\mu D}$, $\tau = \frac{\delta}{r\theta}$, $\alpha_t = \frac{B}{\mu(t+c_0)}$ and $\delta = \frac{1+\log(T+c_0)}{T}$, the R-2-BAN algorithm achieves regret

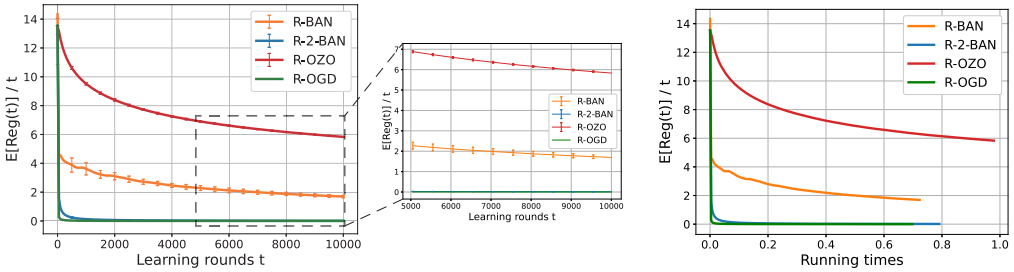
$$\begin{aligned}\mathbb{E}[\mathsf{Reg}(T)] \leq{}& \frac{D^2\mu c_0}{2} + 3(c_0+1)n^2\kappa'^2 L^2 \frac{\zeta(\kappa, D) + \sigma(K, 2D)}{2\mu} + \frac{3(c_0+1)n|\kappa'|\pi^2\iota L}{2} \\ &+ \big(\frac{\zeta(\kappa, D) + \sigma(K, 2D) n^2 L^2}{\mu} + 3L + \frac{DL + 2D^2}{r\theta} + 2\rho L + 2\mu D\big)(1 + \log(T + c_0)).\end{aligned}$$

In Theorem 24, a regret bound of $\mathcal{O}\big(n^{\frac{4}{3}}(1 + \log T)^{\frac{1}{3}} T^{\frac{2}{3}}\big)$ has been established on the R-BAN algorithm for strongly g-convex functions. In contrast, for online optimization over Hadamard manifolds or in Euclidean spaces, such regret bound with strongly g-convex losses is of the order $\mathcal{O}\big(n^{\frac{2}{3}}(1 + \log T)^{\frac{1}{3}} T^{\frac{2}{3}}\big)$.

7. Numerical Experiment

In this section, we validate the findings of the proposed R-OGD, R-BAN and R-2-BAN algorithms over a number of tasks. We also compare our algorithms with the Riemannian online zeroth optimization (R-OZO) algorithm by Maass et al. (2022). The code is built with the help of the Pymanopt package (Townsend et al., 2016) and all experiments are performed in Python 3.8 on a 3.4 GHz AMD Ryzen5 machine with 16GB RAM. For reproduction of the results, all the source codes are accessible online[1].

1. `https://github.com/RiemannianOCO/experiments`



(a) (Expected) Average regrets vs learning rounds (b) Average regrets vs running time

Figure 2: Algorithm performance on hyperbolic Fréchet mean problem

7.1 Fréchet Mean on the Hyperbolic Space

The Fréchet mean problem is known as finding the Riemannian centroid of a set of points on a manifold. The Fréchet mean problem has many applications, such as diffusion tensor magnetic resonance imaging (DT-MRI) (Cheng et al., 2012; Rathi et al., 2007), radar signal processing (Lapuyade-Lahorgue and Barbaresco, 2008), and batch normalization (Brooks et al., 2019). In the following, we study an online version of the Fréchet mean problem, which attempts to average a set of N time-variant points in a hyperbolic space.

Denote by $\langle \cdot, \cdot \rangle_M$ the Minkowski dot product

$$\langle x, y \rangle_M = \sum_{i=1}^{n} x_i y_i - x_{n+1} y_{n+1}.$$

The hyperbolic space can be modeled as

$$H^n = \{x \in \mathbb{R}^{n+1} | \langle x, x \rangle_M = -1\},$$

with the metric $g_x(u, v) = \langle u, v \rangle_M$. A hyperbolic space has constant curvature -1 and thus is a Hadamard manifold (Lee, 2018). Given points $\{A_{t,1}, A_{t,2}, \ldots, A_{t,N}\}$ in a hyperbolic space, the loss function $\mathbf{f}_t$ of the online Fréchet mean problem is

$$\mathbf{f}_t(x_t) = \frac{1}{2N} \sum_{i=1}^{N} d^2(x_t, A_{t,i}) = \frac{1}{2N} \sum_{i=1}^{N} \cosh^{-1}(-\langle x_t, A_{t,i} \rangle_M)^2$$

The loss function $\mathbf{f}_t$, yet is not convex in the Euclidean view, is 1-strongly g-convex (da Silva Alves et al., 2021) so that we can apply Algorithms 1, 2 and 3 to the problem.

We consider the online Fréchet mean problem where $[n, N, T] = [100, 10, 10000]$. The first n indices of $A_{t,i}$ are generated by an Gaussian distribution with the covariance matrix $\text{diag}(\sqrt{n}, \ldots, \sqrt{n})$ and the last index is calculated by the equation $\langle A_{t,i}, A_{t,i} \rangle_M = -1$. We examine Algorithms 1, 2 and 3 for strongly g-convex cases with $\mu = 1$. Additionally, we discuss the choice of δ, α_t and τ in the R-BAN and the R-2-BAN algorithms as follows.

- Since the value of a point grows exponentially with its length in hyperbolic spaces, a large step size α_t in the R-BAN and the R-2-BAN algorithms may cause numerical

overflows. Consequently, we set $\alpha_t = \frac{B}{\mu(t+C_0)}$ for some $C_0 \in \mathbb{N}$, which means to start the optimization at the time C_0. We take $C_0 = 2125$ for the R-BAN algorithm and $C_0 = 170$ for the R-2-BAN algorithm in the experiment, in case of the numerical stability.

- For the R-BAN algorithm, the theoretical δ turns out too conservative and not practical. In the experiment, we set $\delta = 7\sqrt[3]{\frac{1+\log T}{T}}$ instead.
- In this experiment, the feasible set $\mathcal{K}$ is a geodesic ball. Thus, we find that $\theta = 1$ is sufficient to guarantee feasibility. As a result, we opted to use $\tau = \frac{\delta}{r}$ in our experiment.

Figure 2 shows the performance of the average regret $\frac{\mathsf{Reg}(T)}{T}$ versus the number of learning rounds and the running time. The expected average regrets of the R-BAN and the R-2-BAN are performed in the average of 100 random runs with error bars. Figure 2 indicates that the regrets of all three algorithms go sublinearly with the number of learning rounds T. As seen, the one-point bandit algorithm performs most poorly among the three algorithms, while the two-bandit algorithm achieves a comparable regret bound with the R-OGD algorithm in the full information feedback setting, which is consistent with our theoretical findings and also matches the empirical performance in Euclidean spaces (Lei et al., 2020).

We also compare our bandit algorithms (Algorithms 2 and 3) with the Riemannian online zeroth algorithm (R-OZO) by Maass et al. (2022). We observe that in this case our R-BAN and R-2-BAN algorithms perform better regrets with less or equal information, since the R-OZO needs function values of two points.

7.2 Operator Scaling on SPD Matrices

The operator scaling problem is an example of g-convex but not strongly g-convex Riemannian optimization, which is defined on the manifold of SPD matrices

$$\{X \in \mathbb{R}^{n\times n} | X^T = X, X \succ 0\},$$

with the metric

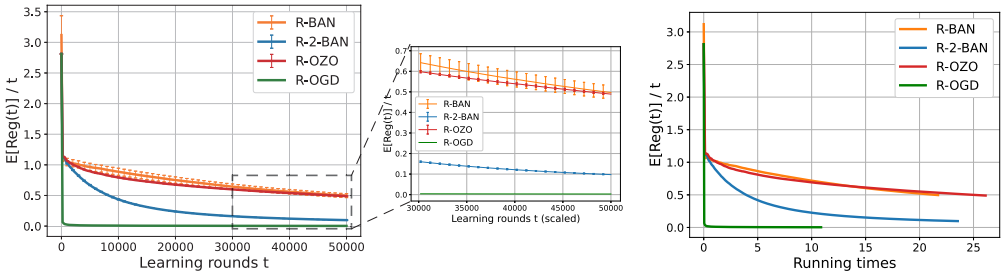
$$g_X(U, V) = \mathrm{Tr}(X^{-1}UX^{-1}V).$$

Given a tuple of $n \times n$ matrices $(A_1, A_2, \ldots, A_N)$, the operator scaling attempts to find $X, Y \in \mathbb{R}^{n\times n}$ such that $\hat{A}_i = Y^{-1}A_iX$ is doubly stochastic for all $i = 1, 2, 3, \ldots, N$. The operator scaling problem has drawn abundant interest in many areas, such as computing non-commutative rank (Ivanov et al., 2017) and computing Brascamp-Lieb constants (Garg et al., 2018). In this subsection, we study an online form of the operator scaling problem, which is to find X_i and Y_i for time-varying matrices $(A_{t,1}, \ldots, A_{t,N})$. The problem can be formulated in terms of minimizing the log capacity of the operator $T_t(X) = \sum_{i=1}^N A_{t,i}XA_{t,i}^T$, that is

$$\mathbf{f}_t(X_t) = \log\det(T(X_t)) - \log\det(X_t), t = 1, 2, \ldots, T.$$

The loss function $\mathbf{f}_t$ is g-convex, but not strongly g-convex.

We test Algorithms 1, 2, 3 and the R-OZO for the case $[n, N, T] = [5, 2, 50000]$ by taking $D = 5$, $L = 2$, and $C = 7$. For the R-BAN algorithm, we set the parameter $\delta = 0.67$, which is 10 times as the theoretical value and also set $\tau = \frac{\delta}{r}$. In addition, since the R-OZO



(a) (Expected) average regrets vs learning rounds (b) Average regrets vs running time

Figure 3: Algorithm performance on operator scaling problem

is designed for σ-strongly g-convex functions and requires a nonzero σ to set up the step size, we take $\sigma = 0.001$ in the R-OZO. The remaining parameters of the R-OZO are set as $\tilde{L} = 1$ (smooth coefficient) and $V = 9$. Figure 3 again shows that the (expected) regrets of Algorithms 1, 2 and 3 are sublinear with T. The R-2-BAN reaches a comparable bound under the full information setting in the online operator scaling problem and the R-OZO presents a regret bound comparable to that of the R-BAN. The above results showcase the applicability for our Algorithms 1, 2 and 3 in g-convex settings.

7.3 Principal Component Analysis on Grassmann Manifolds

At last, we test the effectiveness of Algorithms 1, 2 and 3 on manifolds with positive curvature. An important instance is the principal component analysis (PCA) on Grassmann manifolds. Given a set of data points $\{A_1, \ldots, A_N\}$ in $\mathbb{R}^n$, the PCA problem is to learn an orthogonal projector $X \in \mathbb{R}^{n\times r}$ that minimizes the sum of the squared residual errors between the projected data points and the original data points, which is a significant dimensionality reduction issue when handling high-dimensional data in the real world (Anzai, 2012).

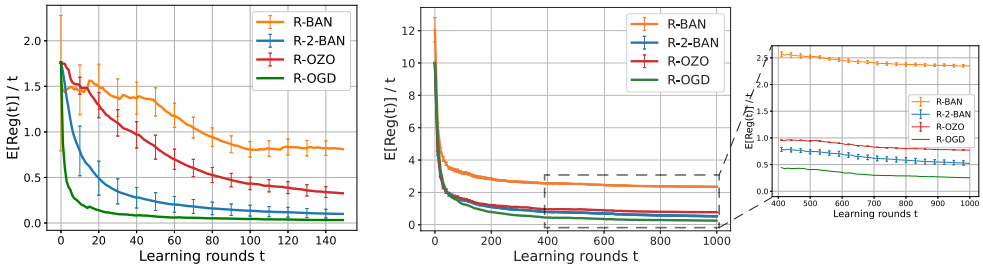
In this subsection, we consider an online PCA problem, where the loss at the time t is

$$\mathbf{f}_t(X_t) = \frac{1}{2N}\sum_{i=1}^{N} \|A_{t,i} - X_t X_t^T A_{t,i}\|_2^2,$$

where $\{A_{t,1}, \ldots A_{t,N}\}$ is a batch of N data points, and X_t is a point on the Stiefel manifold $\mathrm{St}(d, n)$, which is formed by $d \times n$ matrices with orthogonal columns.

data set	class	sample	feature	experimental parameters
iris	3	4	150	$[n, N, T, d, \mu, \theta] = [4, 1, 150, 2, 2, 1]$
egg-eye-state	2	14	14980	$[n, N, T, d, \mu, \theta] = [14, 1, 14000, 3, 0.1, 1]$
waveform-5000	3	40	5000	$[n, N, T, d, \mu, \theta] = [40, 5, 1000, 10, 5, 1]$

Table 2: Descriptions and settings of testing data sets for online PCA



(a) Iris (b) Waveform

Figure 4: Algorithm performance on the iris and waveform data set.

Since the group action $X \to XY$ does not change the value of $\mathbf{f}_t(x_t)$ for any orthogonal matrices $Y \in \mathrm{O(d)}$, we view $\mathbf{f}_t$ as a function on the quotient Grassmann manifold

$$\mathrm{Gr(d,n)} = \mathrm{St(d,n)/O(d)}$$

with the metric $g_X(U,V) = Tr(U^T V)$. Then the online problem is equivalent to find

$$\min \mathbf{f}_t(X_t) = -\frac{1}{2N}\sum_{i=1}^{N} A_{t,i}^T X_t X_t^T A_{t,i}, \quad X_t \in \mathrm{Gr}(d,n).$$

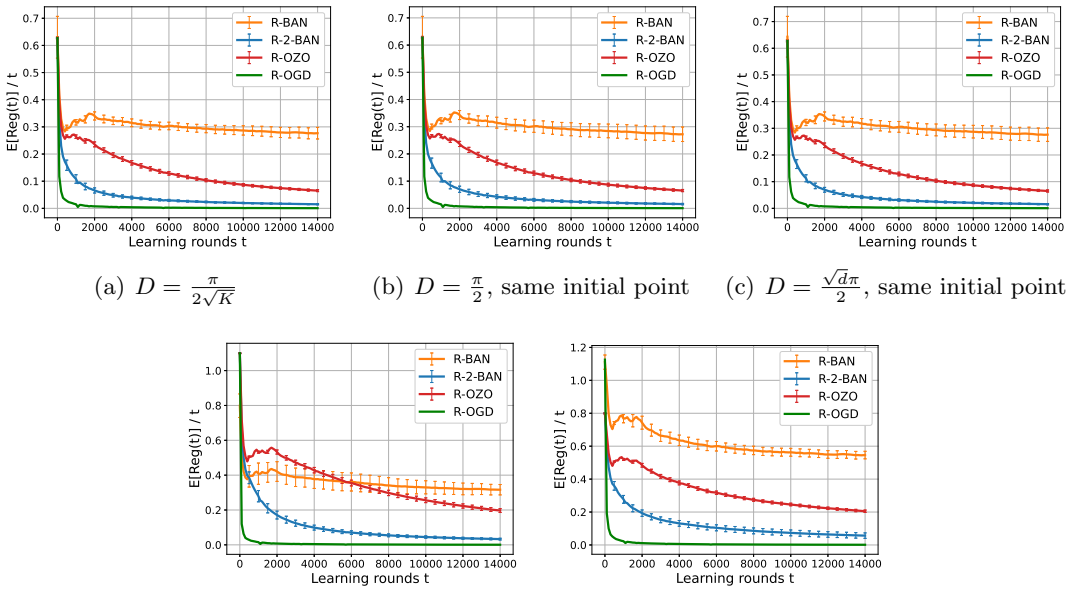
The Grassmann manifold is homogeneous and symmetric, and the sectional curvature of Grassmann manifolds takes value in $[0,2]$. Consequently, we can apply Algorithms 1, 2 and 3 to the online PCA problem.

We examine Algorithms 1, 2, 3 and the R-OZO on three real-world data sets from the openml database[2], including iris, eeg-eye-state and waveform-5000. All the data sets are normalized, and the eeg-eye-state data set is randomly shuffled and excludes outliers.

Figure 4 and Figure 5(a) show the average regret $\frac{\mathrm{Reg}(T)}{T}$ in the three real-world data sets. All the R-BAN, R-2-BAN and R-OZO are conducted for 100 random runs and ploted with error bars. In the R-BAN and R-2-BAN algorithms, the step size α_t and δ are taken referring to the theoretical values. For the R-OZO algorithm, we set $\sigma = \mu$, $\tilde{L} = 1$ (smooth coefficient) and $V = 2$. As shown, the R-2-BAN performs comparably with the R-OGD, and the performance of the R-OZO is between that of the R-BAN algorithm and that of the R-2-BAN algorithm. Besides, All of our algorithms achieve sublinear regret. The results in PCA analysis demonstrate the effectiveness of our algorithms in non-Hadamard cases.

At last, we test the eeg-eye-state data set for the situation when the diameter $D \geq \frac{\pi}{2\sqrt{K}}$. In particular, we test $D = \frac{\pi}{2}$ (the injectivity radius of the Grassmann manifold) in Figures 5(b) and (d), and $D = \sqrt{d}\frac{\pi}{2}$ (the diameter of the Grassmann manifold) in Figures 5(c) and (e). The initial points in Figures 5(b) and (c) are as same as that in Figure 5(c), and are extremely close to the boarder of the feasible set in Figures 5 (d) and (e). Figures 5(a), (b), and (c) show that, when the initial point does not change, diameter D does not influence the (expected) regrets of our Algorithms 1, 2 and 3. Furthermore, Figures 5(d) and (e) show that Algorithms 1, 2 and 3 can still achieve no-regret even when Assumption 20 does

2. https://www.openml.org/



(a) $D = \frac{\pi}{2\sqrt{K}}$ (b) $D = \frac{\pi}{2}$, same initial point (c) $D = \frac{\sqrt{d}\pi}{2}$, same initial point

(d) $D = \frac{\pi}{2}$, close to the border (e) $D = \frac{\sqrt{d}\pi}{2}$, close to the border

Figure 5: Algorithm performance on the eeg-eye-state data set

not hold. The above results give evidence for our applicability of our Algorithms 1, 2 and 3 in practical scenarios.

8. Conclusion

We considered an online optimization problem on Riemannian manifolds in the full information, one-point bandit, and two-point bandit feedback settings, which extended the Euclidean counterpart. The upper regret bounds of the R-OGD, R-BAN, and R-2-BAN algorithm, together with a universal lower regret bound were established with the influence of curvature clearly indicated. All of the regret bounds were consistent with their Euclidean counterpart.

An interesting direction moving forward is to take retraction into consideration. A retraction map is a cheap approximation of the exponential map on manifolds and is a sensible choice in many real scenarios. In future work, we intend to design Riemannian online optimization methods with the retraction map, so that resulting algorithms can be more effective in large-scale optimization problems.

Acknowledgments

The authors would like to express their gratitude to the action editor Silvia Villa, and the anonymous reviewers for their constructive comments and valuable suggestions, which greatly improved this work. The authors also would like to thank Zihao Hu from the School of Computer Science at the Georgia Institude of Technology for his helpful comments and

discussions on the feasibility and projection error of the Riemannian bandit algorithms, which played a significant role in enhancing the mathematical rigor of this work. This work is supported in part by Shanghai Municipal Science and Technology Major Project under Grant 2021SHZDZX0100 and the National Natural Science Foundation of China under Grant 61733018, and in part by Australian Research Council under Grants DP190103615, LP210200473, and DP230101014.

Appendix A. Basic Definitions and Technical Lemmas

In this section, we recall some basic definitions and results from Riemannian geometry, which are useful in our analysis.

Definition 25 (Divergence) *For a vector field X, the divergence* $\mathrm{Div}(X)$ *is the trace of the operator ∇X. More precisely, if $\{e_1, \ldots, e_n\}$ is a normal orthogonal basis of the tangent space $T_x\mathcal{M}$, the divergence of X at x can be expressed as*

$$Div(X)(x) = \sum_{i=1}^{n} \langle \nabla_{e_i} X(x), e_i \rangle.$$

Lemma 26 (Berestovskii and Nikonorov, 2020, Prop. 3.1.6.) *If η is a Killing vector field on $\mathcal{M}$, then it satisfies the following.*

(i) For every vector field X, $\langle \nabla_X \eta, X \rangle = 0$. As a corollary, the divergence $\mathrm{Div}(\eta) \equiv 0$.

(ii) For every geodesic γ, $\eta|_\gamma$ is a Jacobi field.

Lemma 27 (do Carmo, 1992, Prop. 3.6) *If η is a Jacobi vector field along a geodesic $\gamma : [0, 1] \to \mathbb{R}$, denote $\eta(t) = \eta(\gamma(t))$, then*

$$\langle \eta(t), \dot{\gamma}(t) \rangle = \langle \eta(0), \dot{\gamma}(t) \rangle + t \langle \nabla_{\dot{\gamma}} \eta(0), \dot{\gamma}(t) \rangle$$

for all $t \in [0, 1]$.

Lemma 28 (Nomizu, 1960) *Let $\mathcal{M}$ be a simply connected complete Riemannian homogeneous manifold. Then for every $x \in M$ and every $X \in T_x\mathcal{M}$, there exists a Killing vector field η such that $\eta(x) = X$. The flow of η exists and consists of a one-parameter group of isometries.*

Lemma 29 (Divergence theorem, Lee, 2018, 2-22) *Let $\mathcal{M}$ be a Riemannian manifold $\mathcal{M}$ with the volume form ω, $\mathcal{K} \subset \mathcal{M}$ with the boundary $\partial\mathcal{K}$, and $\vec{n}$ be the (outer) unit normal vector field of $\partial\mathcal{K}$. Then, for any vector field X and any differentiable function $\mathbf{f}$,*

$$\int_{\mathcal{K}} X(\mathbf{f})(u)\omega = \int_{\partial\mathcal{K}} \mathbf{f}(u) \langle X, \vec{n} \rangle \omega_{\partial\mathcal{K}} - \int_{\mathcal{K}} \mathrm{Div}(X)\mathbf{f}(u)\omega,$$

where $\omega_{\partial\mathcal{K}}$ is the volume form of $\partial\mathcal{K}$ induced by ω.

Lemma 30 (Jacobi Field Comparison, Lee, 2018, Thm. 11.9) *Suppose that $\mathcal{M}$ is a Riemannian manifold, $\gamma : [0, b] \to \mathcal{M}$ is a unit-speed geodesic segment without conjugate points, and J is a Jacobi field along γ such that $J(0) = 0$. Denote*

$$\mathbf{s}(\kappa, t) = \begin{cases} t, & \text{if } \kappa = 0; \\ \frac{1}{\sqrt{\kappa}} \sin(\sqrt{\kappa} t), & \text{if } \kappa > 0; \\ \frac{1}{\sqrt{-\kappa}} \sinh(\sqrt{-\kappa} t), & \text{if } \kappa < 0. \end{cases} \tag{10}$$

(i) If the sectional curvature of $\mathcal{M}$ is bounded above by a constant $K > 0$, then

$$\|J(t)\| \geq \mathbf{s}(K,t)\|\dot{J}(0)\|$$

for all $t \in [0, b_1]$, where $b_1 = \min\{b, \frac{\pi}{\sqrt{\kappa}}\}$.

(ii) If the sectional curvature of $\mathcal{M}$ is bounded below by a constant $\kappa < 0$, then

$$\|J(t)\| \leq \mathbf{s}(\kappa',t)\|\dot{J}(0)\|$$

for all $t \in [0, b]$, where $\kappa' = \min\{\kappa, 0\}$.

Lemma 31 (Conjugate Theorem, Lee, 2018, Thm. 11.12) *Suppose $\mathcal{M}$ is a Riemannian manifold whose sectional curvature is bounded above by K.*

(i) If $K \leq 0$, then a point of $\mathcal{M}$ has no conjugate points along any geodesic.

(ii) If $K \geq 0$, then there are no conjugate point along any geodesic segment shorter that $\frac{\pi}{\sqrt{K}}$.

From the Morse index theorem Lee, 2018, Thm. 10.18, the absence of conjugate points is equivalent to the unique g-convexity, we have the following corollary.

Corollary 32 *Suppose $(\mathcal{M}, g)$ is a Riemannian manifold whose sectional curvature is bounded by K.*

(i) If $K \leq 0$, then any g-convex set in $\mathcal{M}$ is uniquely g-convex.

(ii) If $K \geq 0$, then any g-convex set in $\mathcal{M}$ with diameter less than $\frac{\pi}{\sqrt{K}}$ is uniquely g-convex.

Lemma 33 (Hessian Comparison Theorem, Lee, 2018, Thm. 11.7) *Suppose $\mathcal{M}$ is a Riemannian manifold and $x \in \mathcal{M}$. Denote $\rho_x(y) = d(x,y)$ and*

$$\mathbf{c}(\kappa,t) = \begin{cases} \frac{1}{t}, & \kappa = 0; \\ \frac{1}{\sqrt{\kappa}}\cot(\sqrt{\kappa}t), & \kappa > 0; \\ \frac{1}{\sqrt{-\kappa}}\coth(\sqrt{\kappa}t), & \kappa < 0. \end{cases}$$

(i) If the sectional curvature of $\mathcal{M}$ is bounded above by a constant $K > 0$, then the following inequality holds in $U := \{y \in \mathcal{M} | d(x,y) < \frac{\pi}{\sqrt{K}}\}$

$$\nabla^2 \rho_x(y) \succeq \mathbf{c}(K, \rho_x(y)) Id.$$

(ii) If the section curvature of $\mathcal{M}$ is bounded below by a constant $\kappa \leq 0$, then the following inequality holds in all of $\mathcal{M}$

$$\nabla^2 \rho_x(y) \preceq \mathbf{c}(\kappa, \rho_x(y)) Id.$$

Lemma 34 (Volume comparison theorem, Lee, 2018, Thm. 11.19) *Let $\mathcal{M}$ denote an n-dimensional Riemannian manifold with sectional curvature lower bounded by κ. Given $p \in \mathcal{M}$, we denote by V_r the volume of the ball of radius r about p, and $V_{r,\kappa}$ as the volume of a ball of radius r on n-dim constant-curvature model spaces with curvature κ, that is, the sphere $S^n(\frac{1}{\sqrt{\kappa}})$, the Euclidean space $\mathbb{R}^n$, or the hyperbolic space $H^n(\frac{1}{\sqrt{-\kappa}})$ when κ is positive, zero, or negative. Then the function*

$$g(r) = \frac{V_r}{V_{r,\kappa}}$$

is non-increasing.

Lemma 35 (Zhang and Sra, 2016, Lemma 5) *Let a, b, c be the sides (side lengths) of a geodesic triangle on a Riemannian manifold with sectional curvature lower bounded by κ. Let A be the angle between sides b and c. Then*

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

Lemma 36 (Bacák, 2014) *Let $(\mathcal{M}, g)$ be a Hadamard manifold. Let $\mathcal{K}$ be a closed g-convex set. Then the mapping $\mathscr{P}_{\mathcal{K}}(x)$ is single-valued and nonexpansive, that is, we have for every $x, y \in \mathcal{M}$*

$$d(\mathscr{P}_{\mathcal{K}}(x), \mathscr{P}_{\mathcal{K}}(y)) \leq d(x, y).$$

Lemma 37 (Berestovskii and Nikonorov, 2020) *For a given point x on a Riemannian symmetric manifold M, the symmetry s_x reverses every geodesic through the point x. Moreover, the derivative map ds_x at x is $-Id_{T_x\mathcal{M}}$.*

Lemma 38 (Berestovskii and Nikonorov, 2020) *A Riemannian symmetric manifold is homogeneous.*

Appendix B. Proofs of Theorems 5 and 6

In this appendix, we prove Theorems 5 and 6 in Subsections B.1 and B.2, respectively.

B.1 Proof of Theorem 5

By the g-convexity, we have

$$\mathbf{f}_t(x_t) - \mathbf{f}_t(x^*) \leq \langle -\nabla \mathbf{f}_t(x_t), \exp_{x_t}^{-1}(x^*)\rangle. \tag{11}$$

Denote $\tilde{x}_{t+1} = \exp_{x_t}(-\alpha_t \nabla \mathbf{f}_t(x_t))$. Recalling Lemma 35 in the geodesic triangle $\triangle x_t \tilde{x}_{t+1} x^*$ gives that

$$\langle -\alpha_t \nabla \mathbf{f}_t(x_t), \exp_{x_t}^{-1}(x^*)\rangle \leq \frac{1}{2}(d^2(x_t, x^*) - d^2(\tilde{x}_{t+1}, x^*)) + \frac{1}{2}\zeta(\kappa, d(x_t, x^*))\|\alpha_t \nabla \mathbf{f}_t(x_t)\|^2.$$

Since $x_{t+1} = \mathscr{P}_{\mathcal{K}}(\exp_{x_t}(-\alpha_t g_t)) = \mathscr{P}_{\mathcal{K}}(\tilde{x}_{t+1})$, applying Lemma 36 we have

$$d(x_{t+1}, x^*) \leq d(\tilde{x}_{t+1}, x^*).$$

Therefore,

$$\langle -\alpha_t \nabla \mathbf{f}_t(x_t), \exp_{x_t}^{-1}(x^*)\rangle \leq \frac{1}{2}(d^2(x_t, x^*) - d^2(x_{t+1}, x^*)) + \frac{1}{2}\zeta(\kappa, d(x_t, x^*))\|\alpha_t \nabla \mathbf{f}_t(x_t)\|^2. \tag{12}$$

Combining (11) and (12), we get

$$\mathbf{f}_t(x_t) - \mathbf{f}_t(x^*) \leq \frac{1}{2\alpha_t}(d^2(x_t, x^*) - d^2(x_{t+1}, x^*)) + \frac{1}{2}\zeta\big(\kappa, d(x_t, x^*)\big)\alpha_t\|\nabla \mathbf{f}_t(x_t)\|^2.$$

With the Lipschitz constant L, we have

$$\mathbf{f}_t(x_t) - \mathbf{f}_t(x^*) \leq \frac{1}{2\alpha_t}(d^2(x_t, x^*) - d^2(x_{t+1}, x^*)) + \frac{1}{2}\zeta\big(\kappa, d(x_t, x^*)\big)L^2\alpha_t. \tag{13}$$

Summing (13) from 1 to T, we obtain

$$\begin{aligned}
\mathsf{Reg}(T) &= \sum_{t=1}^{T} \mathbf{f}_t(x_t) - \sum_{t=1}^{T} \mathbf{f}_t(x^\star) \\
&\leq \sum_{t=1}^{T} \frac{1}{2\alpha_t}\big(d^2(x_t, x^\star) - d^2(x_{t+1}, x^\star)\big) + \sum_{t=1}^{T} \frac{1}{2}\zeta\big(\kappa, d(x_t, x^*)\big)L^2\alpha_t \\
&= \sum_{t=2}^{T} d^2(x_t, x^\star)(\frac{1}{2\alpha_t} - \frac{1}{2\alpha_{t-1}}) + \frac{1}{2\alpha_1}d^2(x_1, x^\star) + \frac{1}{2}L^2\sum_{t=1}^{T}\zeta\big(\kappa, d(x_t, x^*)\big)\alpha_t.
\end{aligned}$$

Since the set $\mathcal{K}$ has diameter D, it follows immediately that $d(x_t, x^*) \leq D$ and

$$\zeta\big(\kappa, d(x_t, x^*)\big) \leq \zeta(\kappa, D)$$

for every $t = 1, 2, \ldots, T$, which implies

$$\begin{aligned}
\mathsf{Reg}(T) &\leq D^2\sum_{t=2}^{T}(\frac{1}{2\alpha_t} - \frac{1}{2\alpha_{t-1}}) + D^2\frac{1}{2\alpha_1} + \frac{1}{2}\zeta(\kappa, D)L^2\sum_{t=1}^{T}\alpha_t \\
&= D^2\frac{1}{2\alpha_t} + \frac{1}{2}\zeta(\kappa, D)L^2\sum_{t=1}^{T}\alpha_t,
\end{aligned}$$

Setting $\alpha_t = \frac{D}{L\sqrt{\zeta(\kappa,D)t}}$, we get

$$\begin{aligned}
\mathsf{Reg}(T) &\leq \frac{DL\sqrt{\zeta(\kappa, D)}}{2}\sqrt{T} + \frac{1}{2}\zeta(\kappa, D)L^2\sum_{t=1}^{T}\alpha_t \\
&\leq \frac{DL\sqrt{\zeta(\kappa, D)}}{2}\sqrt{T} + \frac{1}{2}\zeta(\kappa, D)L^2\frac{2D}{L\sqrt{\zeta(\kappa, D)}}\sqrt{T} \\
&= \frac{3}{2}DL\sqrt{\zeta(\kappa, D)T}.
\end{aligned}$$

The second inequality is based on the inequality $\sum_{t=1}^{T}\frac{1}{\sqrt{t}} \leq 2\sqrt{T}$, and then we complete our proof for Theorem 5. ■

B.2 Proof of Theorem 6

By the strong g-convexity, we have

$$\mathbf{f}_t(x_t) - \mathbf{f}_t(x^*) \leq \langle -\nabla \mathbf{f}_t(x_t), \exp_{x_t}^{-1}(x^*) \rangle - \frac{\mu}{2} d^2(x_t, x^*).$$

With the help of Lemma 35, Lemma 36 and the Lipschitz constant L, we have

$$\mathbf{f}_t(x_t) - \mathbf{f}_t(x^*) \leq \frac{1}{2\alpha_t}\big(d^2(x_t, x^*) - d^2(x_{t+1}, x^*)\big) + \frac{1}{2}\zeta\big(\kappa, d(x_t, x^*)\big) L^2 \alpha_t - \frac{\mu}{2} d^2(x_t, x^*). \tag{14}$$

Summing (14) from 1 to T, we obtain

$$\begin{aligned}
\mathsf{Reg}(T) &= \sum_{t=1}^{T} \mathbf{f}_t(x_t) - \sum_{t=1}^{T} \mathbf{f}_t(x^\star) \\
&\leq \sum_{t=1}^{T} \frac{1}{2\alpha_t}\big(d^2(x_t, x^\star) - d^2(x_{t+1}, x^\star)\big) + \sum_{t=1}^{T} \frac{1}{2}\zeta(\kappa, d(x_t, x^*)) L^2 \alpha_t - \sum_{t=1}^{T} \frac{\mu}{2} d^2(x_t, x^*) \\
&= \sum_{t=2}^{T} d^2(x_t, x^\star)\Big(\frac{1}{2\alpha_t} - \frac{1}{2\alpha_{t-1}} - \frac{\mu}{2}\Big) + d^2(x_1, x^\star)\Big(\frac{1}{2\alpha_1} - \frac{\mu}{2}\Big) + \frac{1}{2} L^2 \sum_{t=1}^{T} \zeta(\kappa, d(x_t, x^*)) \alpha_t.
\end{aligned}$$

Substituting $d(x_t, x^*) \leq D$ and $\zeta(\kappa, d(x_t, x^*)) \leq \zeta(\kappa, D)$ for $t = 1, 2, \ldots, T$, we obtain

$$\begin{aligned}
\mathsf{Reg}(T) &\leq \sum_{t=2}^{T} D^2\Big(\frac{1}{2\alpha_t} - \frac{1}{2\alpha_{t-1}} - \frac{\mu}{2}\Big) + D^2\Big(\frac{1}{2\alpha_1} - \frac{\mu}{2}\Big) + \frac{1}{2} L^2 \sum_{t=1}^{T} \zeta(\kappa, d(x_t, x^*)) \alpha_t \\
&= D^2\Big(\frac{1}{2\alpha_1} - \frac{\mu}{2}\Big) + \frac{1}{2}\zeta(\kappa, D) L^2 \sum_{t=1}^{T} \alpha_t.
\end{aligned}$$

Setting $\alpha_t = \frac{1}{\mu t}$, we get

$$\mathsf{Reg}(T) \leq 0 + \frac{1}{2}\zeta(\kappa, D) L^2 \sum_{t=1}^{T} \alpha_t \leq \frac{\zeta(\kappa, D) L^2}{2\mu}(1 + \log T).$$

We completed the proof. ■

Appendix C. Proof of Theorem 7

In this appendix, we first introduce an instance of Riemannian online convex optimization called the Riemannian online Busemann optimization (ROBO) and then prove Theorem 7 by analyzing the worst-case regret of the ROBO problem.

C.1 Riemannian Online Busemann Optimization

We first introduce the definition of Busemann functions (Ballmann, 2012), which are used to study the large-scale geometry of Hadamard manifolds.

Definition 39 (Ballmann, 2012) *Let $\mathcal{M}$ be a Hadamard manifold and $\gamma : [0, \infty)$ be a geodesic ray on $\mathcal{M}$ with $\|\dot{\gamma}(0)\| = 1$. Then the Busemann function with γ is defined as*

$$\mathbf{f}_\gamma(x) = \lim_{t\to\infty} \big(d(x, \gamma(t)) - t\big).$$

Here are some properties of Busemann functions.

Lemma 40 (Ballmann, 2012) *If $\mathbf{f}_\gamma$ is a Busemann function, then the following properties hold.*

(i) $\mathbf{f}_\gamma$ *is g-convex;*

(ii) $\nabla \mathbf{f}_\gamma(\gamma(t)) = -\dot{\gamma}(t)$ *for every* $t \in [0, \infty)$*;*

(iii) $\|\nabla \mathbf{f}_\gamma(x)\| \leq 1$ *for every* $x \in \mathcal{M}$*.*

Next, we introduce some notations. Let $D, L > 0$ be two constants, $\mathcal{M}$ be a Hadamard manifold, $p \in \mathcal{M}$ and $\gamma : \mathbb{R} \to \mathcal{M}$ be a geodesic with $\|\dot{\gamma}(0)\| = 1$ and $\gamma(0) = p$. Then we consider an instance of R-OCO problem termed *Riemannian online Busemann optimization* (ROBO) on $\mathcal{M}$, where the convex set $\mathcal{K}$ is the ball centered p with radius D, i.e.,

$$\mathcal{K} = \{x \in \mathcal{M} | d(x, p) \leq D\},$$

and the loss function $\mathbf{f}_t$ is randomly and uniformly chosen in the set

$$\{L\mathbf{f}_+, L\mathbf{f}_-\}.$$

Here, $\mathbf{f}_+$ and $\mathbf{f}_-$ are Busemann functions related to the geodesic rays $\gamma_+(t) = \gamma(t)$ and $\gamma_-(t) = \gamma(-t)$. The regret of the ROBO problem is

$$\mathsf{Reg}(T) = \sum_{t=1}^{T} \mathbf{f}_t(x_t) - \min_{x\in\mathcal{K}} \sum_{t=1}^{T} \mathbf{f}_t(x).$$

In the last part of the section, we propose a lemma on the minimum of $\sum_{t=1}^{T} \mathbf{f}_t(x)$.

Lemma 41 *The minimum of $a\mathbf{f}_+(t) + b\mathbf{f}_-(t), (a, b \in \mathbb{N})$ in $\mathcal{K}$ is $-|a - b|D$.*

Proof By the convexity of $\mathbf{f}_+$ and $\mathbf{f}_-$, we have

$$a\mathbf{f}_+(x) + b\mathbf{f}_-(x) \geq a\mathbf{f}_+(p) + b\mathbf{f}_-(p) + \langle a\nabla \mathbf{f}_+(p) + b\nabla \mathbf{f}_-(p), \exp_p^{-1}(x)\rangle, \forall x \in \mathcal{K}.$$

Because $\nabla \mathbf{f}_+(p) = \dot{\gamma}(0)$, $\nabla \mathbf{f}_-(p) = -\dot{\gamma}(0)$ and $\mathbf{f}_\pm(p) = 0$, we have

$$a\mathbf{f}_+(x) + b\mathbf{f}_-(x) \geq \langle -(a - b)\dot{\gamma}(0), \exp_p^{-1}(x)\rangle, \forall x \in \mathcal{K}.$$

Moreover, since $\|\dot{\gamma}(0)\| = 1$ and $\|\exp_p^{-1}(x)\| = d(x, p) \leq D$, we have

$$\min_{x\in\mathcal{K}} a\mathbf{f}_+(x) + b\mathbf{f}_-(x) \geq \min_{x\in\mathcal{K}} \langle -(a - b)\dot{\gamma}(0), \exp_p^{-1}(x)\rangle \geq -|a - b|D. \tag{15}$$

However, we see that $a\mathbf{f}_+(\gamma(D)) + b\mathbf{f}_-(\gamma(D)) = (b - a)D$, and $a\mathbf{f}_+(\gamma(-D)) + b\mathbf{f}_-(\gamma(-D)) = (a - b)D$, which imply that

$$\min_{x\in\mathcal{K}} a\mathbf{f}_+(x) + b\mathbf{f}_-(x) \leq \min\Big\{(b - a)D, (a - b)D\Big\} = -|a - b|D. \tag{16}$$

Following from (15) and (16), we complete our proof. ■

C.2 Proof of Theorem 7

We begin our proof with an analysis of the worst-case regret of the ROBO problem. In the ROBO, the expectation of the regret on loss functions $\{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_t\}$ is

$$\begin{aligned}\mathbb{E}_{\mathbf{f}_1,\ldots,\mathbf{f}_t}[\mathsf{Reg}(T)] =& \mathbb{E}_{\mathbf{f}_1,\ldots,\mathbf{f}_t}[\sum_{t=1}^{T}\mathbf{f}_t(x_t) - \min_{x\in\mathcal{K}}\sum_{t=1}^{T}\mathbf{f}_t(x)] \\ =& \mathbb{E}_{\mathbf{f}_1,\ldots,\mathbf{f}_t}[\sum_{t=1}^{T}\mathbf{f}_t(x_t)] - \mathbb{E}_{\mathbf{f}_1,\ldots,\mathbf{f}_t}[\min_{x\in\mathcal{K}}\sum_{t=1}^{T}\mathbf{f}_t(x)]. \end{aligned} \tag{17}$$

Since $\mathbf{f}_t$ is uniformly and independently chosen in $\{\mathbf{f}_+, \mathbf{f}_-\}$, we can get

$$\begin{aligned}\mathbb{E}_{\mathbf{f}_1,\ldots,\mathbf{f}_t}[\sum_{t=1}^{T}\mathbf{f}_t(x_t)] &= \sum_{t=1}^{T}\mathbb{E}_{\mathbf{f}_t}[\mathbf{f}_t(x_t)] \\ &= \sum_{t=1}^{T}\frac{1}{2}(L\mathbf{f}_+(x_t) + L\mathbf{f}_-(x_t)) \\ &\geq \frac{LT}{2}\min_{x\in\mathcal{K}}(\mathbf{f}_+(x) + \mathbf{f}_-(x)).\end{aligned}$$

From Lemma 41,

$$\mathbb{E}_{\mathbf{f}_1,\ldots,\mathbf{f}_t}[\sum_{t=1}^{T}\mathbf{f}_t(x_t)] \geq 0. \tag{18}$$

Putting (18) into (17), we obtain

$$\mathbb{E}_{\mathbf{f}_1,\ldots,\mathbf{f}_t}[\mathsf{Reg}(T)] \geq -\mathbb{E}_{\mathbf{f}_1,\ldots,\mathbf{f}_t}[\min_{x\in\mathcal{K}}\sum_{t=1}^{T}\mathbf{f}_t(x)].$$

By Lemma 41,

$$\begin{aligned}\mathbb{E}_{\mathbf{f}_1,\ldots,\mathbf{f}_t}[\mathsf{Reg}(T)] &\geq -\mathbb{E}_{\mathbf{f}_1,\ldots,\mathbf{f}_t}[\min_{x\in\mathcal{K}}\sum_{t=1}^{T}\mathbf{f}_t(x)] \\ &= -\mathbb{E}_{\mathbf{f}_1,\ldots,\mathbf{f}_t}\Big[-DL\Big|\sum_{\mathbf{f}_t=L\mathbf{f}_+}1 - \sum_{\mathbf{f}_t=L\mathbf{f}_-}1\Big|\Big] \\ &= \mathbb{E}_{\epsilon_1,\ldots,\epsilon_T}\Big[DL\Big|\sum_{\epsilon_t=1}1 + \sum_{\epsilon_t=-1}-1\Big|\Big] \\ &= \mathbb{E}_{\epsilon_1,\ldots,\epsilon_T}\Big[DL\Big|\sum_{t=1}^{T}\epsilon_t\Big|\Big],\end{aligned}$$

where ϵ_t are i.i.d Rademacher variables $\epsilon_t = \pm 1$ with probability $1/2$. From the Khinchine's inequality (Cesa-Bianchi and Lugosi, 2006), we finally get

$$\mathbb{E}_{\mathbf{f}_1,\ldots,\mathbf{f}_t}[\mathsf{Reg}(T)] \geq \frac{DL}{\sqrt{2}}E_{\epsilon_1,\ldots,\epsilon_T}\Big[\sum_{t=1}^{T}\epsilon_t^2\Big] = \frac{DL}{\sqrt{2}}\sqrt{T}, \tag{19}$$

which indicates that no matter how we choose strategies in the ROBO, there is a sequence of functions $\{\mathbf{f}_1, \ldots, \mathbf{f}_t\} \in \{L\mathbf{f}_+, L\mathbf{f}_-\}^T$ to make the regret not less than $\frac{DL}{\sqrt{2}}\sqrt{T}$. Considering that the diameter of the set $\mathcal{K}$ is $2D$ and the Lipschitz constant of $\{L\mathbf{f}_+, L\mathbf{f}_-\}$ is L, we complete our proof. ■

Appendix D. Proofs of Lemmas 11 and 13

In this appendix, we prove Lemmas 11 and 13 in Subsections D.1 and D.2, respectively.

D.1 Proof of Lemma 13

We initially examine the first part of the lemma. Take a vector $X \in M_x$ arbitrarily. From Lemma 28, we can find a Killing vector field η on $\mathcal{M}$ such that $\eta(x) = X$. The flow of η consists of a one-parameter group of isometries $\{\phi_t\}_{t\in\mathbb{R}}$. Then the directional derivative of $\hat{\mathbf{f}}$ along X can be written as

$$X(\hat{\mathbf{f}}(x)) = \lim_{t\to 0} \frac{\hat{\mathbf{f}}(\phi_t(x)) - \hat{\mathbf{f}}(x)}{t} = \frac{1}{V_\delta} \lim_{t\to 0} \frac{1}{t} \Big(\int_{\mathcal{B}_\delta(\phi_t(x))} \mathbf{f}(u)\omega - \int_{\mathcal{B}_\delta(x)} \mathbf{f}(u)\omega \Big). \tag{20}$$

Since ϕ_t is an isometry that preserves the distance, $\phi_t(\mathcal{B}_\delta(x)) = \mathcal{B}_\delta(\phi_t(x))$. By the substitution rule of integration (Chern et al., 1999), we have

$$\int_{\mathcal{B}_\delta(\phi_t(x))} \mathbf{f}(u)\omega = \int_{\mathcal{B}_\delta(x)} \mathbf{f}(\phi_t(u))\phi_t^*(\omega). \tag{21}$$

Because ϕ_t preserves the metric g, it preserves the volume form, i.e., $\phi_t^*(\omega) = \omega$, which gives

$$\int_{\mathcal{B}_\delta(\phi_t(x))} \mathbf{f}(u)\omega = \int_{\mathcal{B}_\delta(x)} \mathbf{f}(\phi_t(u))\omega. \tag{22}$$

Combining equations (20) and (22) together, we have

$$\begin{aligned} X(\hat{\mathbf{f}}(x)) &= \frac{1}{V_\delta} \int_{\mathcal{B}_\delta(x)} \lim_{t\to 0} \frac{\mathbf{f}(\phi_t(u)) - \mathbf{f}(u)}{t} \omega \\ &= \frac{1}{V_\delta} \int_{\mathcal{B}_\delta(x)} \frac{\partial \phi_t(p)}{\partial t}|_{t=0}(\mathbf{f})\omega. \end{aligned} \tag{23}$$

By definition of the flow, $\frac{\partial \phi_t(u)}{\partial t}|_{t=0} = \eta(u)$. Hence, we can rewrite (23) as

$$X(\hat{\mathbf{f}}(x)) = \frac{1}{V_\delta} \int_{\mathcal{B}_\delta(x)} \eta(\mathbf{f})\omega.$$

According to Lemma 29, we have

$$\begin{aligned} X(\hat{\mathbf{f}}(x)) &= \frac{1}{V_\delta} \int_{\mathcal{S}_\delta(x)} \mathbf{f}(u)\langle \eta(u), \vec{n}(u)\rangle \omega_{\mathcal{S}_\delta(x)} - \frac{1}{V_\delta} \int_{\mathcal{B}_\delta(x)} \mathrm{Div}(\eta)(u)\mathbf{f}(u)\omega \\ &= \frac{1}{V_\delta} \int_{\mathcal{S}_\delta(x)} \mathbf{f}(u)\langle \eta(u), \vec{n}(u)\rangle \omega_{\mathcal{S}_\delta(x)} \end{aligned} \tag{24}$$

where $\omega_{\mathcal{S}_\delta(x)}$ is the volume form of $\mathcal{S}_\delta(x)$ induced by ω and $\vec{n}$ is the (outer) unit normal vector field of $\mathcal{S}_\delta(x)$. The last equation is due to $\mathrm{Div}(\eta) \equiv 0$ stated in Lemma 26.

Then we need to compute $\langle \eta, \vec{n} \rangle$ for each point $u \in \mathcal{S}_\delta(x)$. Since geodesics start at the center x are vertical to the sphere $\mathcal{S}_\delta(x)$, the outer normal vector $\vec{n}(u)$ can be written as $\frac{\dot{\gamma}_u(1)}{\|\dot{\gamma}_u(1)\|}$ for the geodesic γ_u such that $\gamma_u(0) = x$ and $\gamma_u(1) = u$. Therefore, we can write $\langle \eta(u), \vec{n}(u) \rangle$ as

$$\langle \eta(u), \vec{n}(u) \rangle = \frac{1}{\|\dot{\gamma}_u(1)\|} \langle \eta(\gamma_u(1)), \dot{\gamma}_u(1) \rangle.$$

Since η is Killing, by Lemma 26, $\eta(\gamma_u(t))$ is Jacobi. By Lemma 27, we have

$$\begin{aligned}
\langle \eta(u), \vec{n}(u) \rangle &= \frac{1}{\|\dot{\gamma}_u(1)\|} \langle \eta(\gamma_u(1)), \dot{\gamma}_u(1) \rangle \\
&= \frac{1}{\|\dot{\gamma}_u(1)\|} \Big(\langle \eta(\gamma_u(0)), \dot{\gamma}_u(0) \rangle + 1 \langle \nabla_{\dot{\gamma}_u} \eta(\gamma_u(0)), \dot{\gamma}_u(0) \rangle \Big), \\
&= \frac{1}{\|\dot{\gamma}_u(0)\|} \Big(\langle \eta(\gamma_u(0)), \dot{\gamma}_u(0) \rangle + 0 \Big). \qquad (25)
\end{aligned}$$

Applying $\eta(\gamma_u(0)) = \eta(x) = X$ and $\dot{\gamma}_u(0) = \exp_x^{-1}(u)$ to (25) yields

$$\langle \eta(u), \vec{n}(u) \rangle = \frac{\langle X, \exp_x^{-1}(u) \rangle}{\| \exp_x^{-1}(u) \|}. \qquad (26)$$

Substituting (26) to (24), we have

$$X(\hat{\mathbf{f}}(x)) = \frac{1}{V_\delta} \int_{\mathcal{S}_\delta(x)} \mathbf{f}(u) \frac{\langle X, \exp_x^{-1}(u) \rangle}{\| \exp_x^{-1}(u) \|} \omega_{\mathcal{S}_\delta(x)} = \langle \frac{1}{V_\delta} \int_{\mathcal{S}_\delta(x)} \mathbf{f}(u) \frac{\exp_x^{-1}(u)}{\| \exp_x^{-1}(u) \|} \omega_{\mathcal{S}_\delta(x)}, X \rangle.$$

Because the directional derivative $X(\hat{\mathbf{f}}(x))$ coincides with the term $\langle \nabla \hat{f}(x), X \rangle$, we can obtain

$$\langle \nabla \hat{\mathbf{f}}(x), X \rangle = \langle \frac{1}{V_\delta} \int_{\mathcal{S}_\delta(x)} \mathbf{f}(u) \frac{\exp_x^{-1}(u)}{\| \exp_x^{-1}(u) \|} \omega_{\mathcal{S}_\delta(x)}, X \rangle.$$

For the arbitrariness of the vector field X, we conclude that

$$\nabla \hat{\mathbf{f}}(x) = \frac{1}{V_\delta} \int_{\mathcal{S}_\delta(x)} \mathbf{f}(u) \frac{\exp_x^{-1}(u)}{\| \exp_x^{-1}(u) \|} \omega_{\mathcal{S}_\delta(x)} = \frac{S_\delta}{V_\delta} E_{u \in \mathcal{S}_\delta(x)} \Big[\mathbf{f}(u) \frac{\exp_x^{-1}(u)}{\| \exp_x^{-1}(u) \|} \Big],$$

which completes the proof of the first part.

Then we examine the second part of the lemma. From the first part, it is clear to see

$$\mathbb{E}[\| \frac{S_\delta}{V_\delta} \mathbf{g} \|] \leq \frac{S_\delta}{V_\delta} C.$$

Since the sectional curvature of $\mathcal{M}$ is lower bounded by κ', from Lemma 34, the function $g(r) = \frac{V_r}{V_{r,\kappa'}}$ is non-increasing and so does $\log g(r)$. Therefore,

$$\frac{d}{dr} \log(g(r)) = \frac{d}{dr} \log V_r - \frac{d}{dr} \log V_{r,\kappa'} \leq 0.$$

Since deriving the volume of a ball along the radius gives the surface area of its sphere, we can write

$$\frac{d}{dr}\log(g(r)) = \frac{S_r}{V_r} - \frac{S_{r,\kappa'}}{V_{r,\kappa'}} \leq 0, \tag{27}$$

where S_r and $S_{r,\kappa'}$ are the surface area of the balls in $\mathcal{M}$ and the corresponding constant curvature space, respectively.

Setting $r = \delta$ in (27), we get $\frac{S_\delta}{V_\delta} \leq \frac{S_{\delta,\kappa'}}{V_{\delta,\kappa'}}$. From calculation, it shows that

$$\frac{S_\delta}{V_\delta} \leq \frac{S_{\delta,\kappa'}}{V_{\delta,\kappa'}} = \begin{cases} \frac{n}{\delta}, & \kappa' = 0 \\ \frac{\sinh^{n-1}(\sqrt{|\kappa'|}\delta)}{\int_0^\delta \sinh^{n-1}(\sqrt{|\kappa'|}t)dt}, & \kappa' = \kappa < 0 \end{cases}$$

So we have completed the proof for the case $\kappa' = 0$. Then we focus on the case that $\kappa' < 0$. By a change of variable $u = \sinh t$, we find

$$\int_0^\delta \sinh^{n-1}(\sqrt{|\kappa'|}t)dt = |\kappa'|^{-1/2}\int_0^{\sinh(\sqrt{|\kappa'|}\delta)} u^{n-1}(1+u^2)^{-1/2}du$$

Integration by parts gives

$$\begin{aligned}\int_0^\delta \sinh^{n-1}(\sqrt{|\kappa'|}t)dt &= \frac{\sinh^n(\sqrt{|\kappa'|}\delta)}{n\sqrt{|\kappa'|}\cosh(\sqrt{|\kappa'|}\delta)} + |\kappa'|^{-1/2}\int_0^{\sinh(\sqrt{|\kappa'|}\delta)} \frac{1}{n}u^{n+1}(1+u^2)^{-3/2}du \\ &\geq \frac{\sinh^n(\sqrt{|\kappa'|}\delta)}{n\sqrt{|\kappa'|}\cosh(\sqrt{|\kappa'|}\delta)}.\end{aligned}$$

Putting it into the expression of $\frac{S_\delta}{V_\delta}$, we get

$$\frac{S_\delta}{V_\delta} \leq n\sqrt{|\kappa'|}\coth(\sqrt{|\kappa'|}\delta).$$

Applying the inequality $\coth(x) < x + 1/x$, we have

$$\frac{S_\delta}{V_\delta} \leq \frac{n}{\delta} + n|\kappa'|\delta, \quad \forall\delta > 0.$$

Hence, for every $\delta > 0$,

$$\mathbb{E}[\|\frac{S_\delta}{V_\delta}\mathbf{g}\|] \leq \frac{S_\delta}{V_\delta}C \leq C\big(\frac{n}{\delta} + n|\kappa'|\delta\big),$$

which completes our proof. ■

D.2 Proof of Lemma 11

First we examine (i). Without loss of generality, we assume $\mathbf{f}(x) = 0$. By the homogeneity of the manifold $\mathcal{M}$, we can find an isometry ϕ such that $\phi(x) = y$. Denote the vector field $V(u) = \exp_u^{-1}(\phi(u))$. Clearly, we obtain

$$\begin{aligned}\hat{\mathbf{f}}(y) - \hat{\mathbf{f}}(x) &= \frac{1}{V_\delta}\Big(\int_{\mathcal{B}_\delta(y)} \mathbf{f}(u)\omega - \int_{\mathcal{B}_\delta(x)} \mathbf{f}(u)\omega\Big) \\ &= \frac{1}{V_\delta}\Big(\int_{\mathcal{B}_\delta(\phi(x))} \mathbf{f}(u)\omega - \int_{\mathcal{B}_\delta(x)} \mathbf{f}(u)\omega\Big).\end{aligned}$$

With the method shown in (21) and (22), we obtain

$$\hat{\mathbf{f}}(y) - \hat{\mathbf{f}}(x) = \frac{1}{V_\delta}\int_{\mathcal{B}_\delta(x)} \mathbf{f}(\phi(u)) - \mathbf{f}(u)\omega.$$

By the g-convexity of $\mathbf{f}$,

$$\begin{aligned}\hat{\mathbf{f}}(y) - \hat{\mathbf{f}}(x) &= \frac{1}{V_\delta}\int_{\mathcal{B}_\delta(x)} \mathbf{f}(\phi(u)) - \mathbf{f}(u)\omega \\ &\geq \frac{1}{V_\delta}\int_{\mathcal{B}_\delta(x)} \langle \nabla \mathbf{f}(u), \exp_u^{-1}(\phi(u))\rangle\omega \\ &= \frac{1}{V_\delta}\int_{\mathcal{B}_\delta(x)} \langle \nabla \mathbf{f}(u), V(u)\rangle\omega \\ &= \frac{1}{V_\delta}\int_{\mathcal{B}_\delta(x)} V(\mathbf{f}(u))\omega. \end{aligned} \tag{28}$$

By Lemma 29, we have

$$\int_{\mathcal{B}_\delta(x)} V(\mathbf{f}(u))\omega = \int_{\mathcal{S}_\delta(x)} \mathbf{f}(u)\langle V(u), \vec{n}(u)\rangle\omega_{\mathcal{S}_\delta(x)} - \int_{\mathcal{B}_\delta(x)} \mathrm{Div}(V)\mathbf{f}(u)\omega. \tag{29}$$

Hence, we rewrite (28) as

$$\hat{\mathbf{f}}(y) - \hat{\mathbf{f}}(x) \geq \frac{1}{V_\delta}\Big(\int_{\mathcal{S}_\delta(x)} \mathbf{f}(u)\langle V(u), \vec{n}(u)\rangle\omega_{\mathcal{S}_\delta(x)} - \int_{\mathcal{B}_\delta(x)} \mathrm{Div}(V)\mathbf{f}(u)\omega\Big).$$

In Lemma 11, we have already shown that

$$\begin{aligned}\langle \nabla\hat{\mathbf{f}}(x), \exp_x^{-1}(y)\rangle &= \langle \frac{1}{V_\delta}\int_{\mathcal{S}_\delta(x)} \mathbf{f}(u)\frac{\exp_x^{-1}(u)}{\|\exp_x^{-1}(u)\|}\omega_{\mathcal{S}_\delta(x)}, \exp_x^{-1}(y)\rangle \\ &= \langle \frac{1}{V_\delta}\int_{\mathcal{S}_\delta(x)} \mathbf{f}(u)\frac{\exp_x^{-1}(u)}{\|\exp_x^{-1}(u)\|}\omega_{\mathcal{S}_\delta(x)}, V(x)\rangle. \end{aligned} \tag{30}$$

Denote by $\vec{m}(u)$ the vector $\frac{\exp_x^{-1}(u)}{\|\exp_x^{-1}(u)\|}$. Combining (29) and (30) gives

$$\begin{aligned}\hat{\mathbf{f}}(y) - \hat{\mathbf{f}}(x) - \langle \nabla\hat{\mathbf{f}}(x), \exp_x^{-1}(y)\rangle \geq \frac{1}{V_\delta}\Big(\int_{\mathcal{S}_\delta(x)} \mathbf{f}(u)\Big(\langle V(u), \vec{n}(u)\rangle - \langle V(x), \vec{m}(u)\rangle\Big)\omega_{\mathcal{S}_\delta(x)}\Big) \\ - \frac{1}{V_\delta}\Big(\int_{\mathcal{B}_\delta(x)} \mathrm{Div}(V)\mathbf{f}(u)\omega\Big). \end{aligned} \tag{31}$$

Here we claim that

$$\langle V(u), \vec{n}(u)\rangle - \langle V(x), \vec{m}(u)\rangle \leq 0, \quad \forall u \in \mathcal{S}_\delta(x). \tag{$*$}$$

This claim requires many geometric details that deviates our attention from the proof, and we will prove it afterwards. If the claim $(*)$ holds, then with the g-L-Lipschitzness of $\mathbf{f}$ and the condition $\mathbf{f}(x) = 0$, we have

$$\begin{aligned}
&\int_{\mathcal{S}_\delta(x)} \mathbf{f}(u)\Big(\langle V(u), \vec{n}(u)\rangle - \langle V(x), \vec{m}(u)\rangle\Big)\omega_{\mathcal{S}_\delta(x)} \\
&\qquad \geq \int_{\mathcal{S}_\delta(x)} \delta L\Big(\langle V(u), \vec{n}(u)\rangle - \langle V(x), \vec{m}(u)\rangle\Big)\omega_{\mathcal{S}_\delta(x)} \\
&\qquad = \Big(\int_{\mathcal{S}_\delta(x)} \delta L\langle V(u), \vec{n}(u)\rangle\omega_{\mathcal{S}_\delta(x)}\Big) - \Big(\int_{\mathcal{S}_\delta(x)} \delta L\langle V(x), \vec{m}(u)\rangle\omega_{\mathcal{S}_\delta(x)}\Big).
\end{aligned} \tag{32}$$

By Lemma 11, $\frac{1}{V_\delta}\int_{\mathcal{S}_\delta(x)} \delta L\langle V(x), \vec{m}(u)\rangle\omega_{\mathcal{S}_\delta(x)}$ in (32) is the gradient of the function

$$\hat{\mathbf{g}}(x) := \frac{1}{V_\delta}\Big(\int_{\mathcal{B}_\delta(x)} \delta L \cdot \omega\Big) \equiv \delta L,$$

and then

$$\frac{1}{V_\delta}\int_{\mathcal{S}_\delta(x)} \delta L\langle V(x), \vec{m}(u)\rangle\omega_{\mathcal{S}_\delta(x)} = 0. \tag{33}$$

Combining (31)-(33), we have

$$\begin{aligned}
\hat{\mathbf{f}}(y) - \hat{\mathbf{f}}(x) - \langle\nabla\hat{\mathbf{f}}(x), \exp_x^{-1}(y)\rangle \geq\ & \frac{1}{V_\delta}\Big(\int_{\mathcal{S}_\delta(x)} \delta L\langle V(u), \vec{n}(u)\rangle\omega_{\mathcal{S}_\delta(x)}\Big) \\
& - \frac{1}{V_\delta}\Big(\int_{\mathcal{B}_\delta(x)} \mathrm{Div}(V)\mathbf{f}(u)\omega\Big).
\end{aligned}$$

Applying Lemma 29 again, we obtain

$$\begin{aligned}
\frac{1}{V_\delta}\int_{\mathcal{S}_\delta(x)} \delta L\langle V(u), \vec{n}(u)\rangle\omega_{\mathcal{S}_\delta(x)} &= \frac{1}{V_\delta}\int_{\mathcal{B}_\delta(x)} V(\delta L)\omega - \frac{1}{V_\delta}\Big(\int_{\mathcal{B}_\delta(x)} \mathrm{Div}(V)\delta L\omega\Big) \\
&= -\frac{1}{V_\delta}\Big(\int_{\mathcal{B}_\delta(x)} \mathrm{Div}(V)\delta L\omega\Big).
\end{aligned}$$

Therefore, there holds

$$\begin{aligned}
\hat{\mathbf{f}}(y) - \hat{\mathbf{f}}(x) - \langle\nabla\hat{\mathbf{f}}(x), \exp_x^{-1}(y)\rangle &\geq -\frac{1}{V_\delta}\Big(\int_{\mathcal{B}_\delta(x)} \mathrm{Div}(V)(\mathbf{f}(x) + \delta L)\omega\Big) \\
&\geq -2\delta L \sup_{u\in\mathcal{B}_\delta(x)} |\mathrm{Div}(V(u))|.
\end{aligned} \tag{34}$$

Note that $V(u) = \exp_u^{-1}(\phi(u))$ is continuous on p and ϕ, and ϕ is continuous on x and y. Thus, $|\mathrm{Div}(V(u))|$ is a continuous function of $(x, y, u) \in \bar{\mathcal{K}} \times \bar{\mathcal{K}} \times \bar{\mathcal{K}}$. Denote

$$\rho = \sup_{(x,y,u)\in\bar{\mathcal{K}}\times\bar{\mathcal{K}}\times\bar{\mathcal{K}}} |\mathrm{Div}(V(u))|.$$

Since the boundedness of $\mathcal{K}$ set yields the compactness of $\bar{\mathcal{K}} \times \bar{\mathcal{K}} \times \bar{\mathcal{K}}$, we have $\rho < \infty$. Putting ρ into (34) establishes the desired result.

Then we begin to prove ii). From the strong g-convexity of $\mathbf{f}$,

$$\hat{\mathbf{f}}(y) - \hat{\mathbf{f}}(x) = \frac{1}{V_\delta} \int_{\mathcal{B}_\delta(x)} V(\mathbf{f}(u)) + \frac{\mu}{2} d^2(u, \phi(u)) \omega.$$

Thus it remains to prove that

$$\frac{1}{V_\delta} \int_{\mathcal{B}_\delta(x)} \frac{\mu}{2} d^2(u, \phi(u)) \omega \geq \frac{\mu}{2} d^2(x, y) - 2\mu D,$$

which is obvious from the fact

$$\begin{aligned} |d^2(u, \phi(u)) - d^2(x, y)| &= (d(u, \phi(u)) + d(x, y))|(d(u, \phi(u)) - d(x, y))| \\ &\leq 2D \cdot 2\delta = 4D\delta. \end{aligned}$$

■

D.3 Proof of the Claim $(*)$

Fix $u \in \mathcal{S}_\delta(x)$ and denote by $\xi_u(s) = \exp_x(s\vec{m}(u))$ the geodesic with the initial tangent vector $\vec{m}(u)$. Consider the following rectangle map

$$\begin{aligned} \Gamma_u : [0,1] \times [0, \delta] &\to \mathcal{M} \\ (t, s) &\to \exp_{\xi_u(s)}(tV(\xi_u(s))). \end{aligned}$$

Set $T(t,s) = \frac{\partial \Gamma_u}{\partial t}(t,s)$ and $S(t,s) = \frac{\partial \Gamma_u}{\partial s}(t,s)$. For a fixed t, the length of the curve $\gamma_t(s) = \Gamma_u(t,s), (0 \leq s \leq \delta)$ is defined as

$$l_u(t) = \int_0^\delta \sqrt{\langle S(t,s), S(t,s) \rangle} ds.$$

The first variation formula (see Lee, 2018, Theorem 6.3) gives,

$$l_u'(0) = \langle T(0,\delta), S(0,\delta) \rangle - \langle T(0,0), S(0,0) \rangle.$$

Because $T(0,s) = V(\xi_u(s))$, for all $s \in [0,\delta]$, $S(0,0) = \vec{m}(u)$ and $S(0,\delta) = \vec{n}(u)$, we have

$$l_u'(0) = \langle V(u), \vec{n}(u) \rangle - \langle V(x), \vec{m}(u) \rangle.$$

To prove $(*)$, it is sufficient to show that $l_u'(0) \leq 0$. Let us focus on the second derivative of the function $l_u(t)$, that is,

$$\begin{aligned} l_u''(t) &= \frac{d^2}{dt^2} \int_0^\delta \sqrt{\langle S(t,s), S(t,s) \rangle} ds \\ &= \int_0^\delta \frac{d^2}{dt^2} \sqrt{\langle S(t,s), S(t,s) \rangle} ds \\ &= \int_0^\delta \frac{d}{dt} (\frac{1}{\|S\|} \langle \nabla_T S, S \rangle) ds \\ &= \int_0^\delta -\frac{1}{\|S\|^3} \langle \nabla_T S, S \rangle^2 + \frac{1}{\|S\|} \langle \nabla_T S, \nabla_T S \rangle + \frac{1}{\|S\|} \langle \nabla_T \nabla_T S, S \rangle ds. \end{aligned} \tag{35}$$

For every fixed s, the curve $\gamma_s(t) = \Gamma_u(t, s)$ is a geodesic, hence S is the variation field of the geodesic $\gamma_s(t)$ and becomes a Jacobi field. Putting the Jacobi equation into (35), we have

$$l_u''(t) = \int_0^\delta -\frac{1}{\|S\|^3}\langle \nabla_T S, S\rangle^2 + \frac{1}{\|S\|}\langle \nabla_T S, \nabla_T S\rangle + \frac{1}{\|S\|} - R(T, S, S, T)ds.$$

By the Cauchy–Schwarz inequality, $-\langle \nabla_T S, S\rangle^2 \geq -\|S\|^2\|\nabla_T S\|^2$, which yields

$$\begin{aligned} l_u''(t) &\geq \int_0^\delta -\frac{1}{\|S\|^3} - \|S\|^2\|\nabla_T S\|^2 + \frac{1}{\|S\|}\langle \nabla_T S, \nabla_T S\rangle + \frac{1}{\|S\|} - R(T, S, S, T)ds \\ &\geq \int_0^\delta \frac{1}{\|S\|} - R(T, S, S, T)ds. \end{aligned}$$

From the definition of sectional curvature, $R(T, S, S, T) = K(\Pi)|T \wedge S|^2$, where $K(\Pi)$ is the sectional curvature of the two-dimensional submanifold spanned by T and S. Under the assumption that $\mathcal{M}$ has nonpostive sectional curvature, we get

$$l_u''(t) \geq \int_0^\delta \frac{1}{\|S\|} - R(T, S, S, T)ds \geq 0,$$

which means that $l_u(t)$ is convex in $[0, 1]$.

Let us look back on the function $l_u(t)$. Note that the 0-curve is

$$\gamma_s(0) = \xi(s),$$

and the 1-curve is

$$\gamma_s(1) = \exp_{\xi(s)}(V(\xi(s))) = \exp_{\xi(s)}(\exp^{-1}_{\xi(s)}(\phi(\xi(s))) = \phi(\xi(s)).$$

Since the mapping ϕ is an isometry, the length of $\xi(s)$ is equal to the length of $\phi(\xi(s))$. As a result, there holds

$$l_u(0) = l_u(1). \tag{36}$$

The convexity of l_u immediately leads that

$$l_u'(0) \leq 0,$$

which proves the claim $(*)$. ■

Appendix E. Proofs of Theorems 14 and 15

Before the proof, we propose some lemmas about the expected online gradient descent for λ-sub g-convex functions.

Lemma 42 (Sub-convex Cases) *Suppose that $\mathcal{S}$ is a subset of a g-convex set $\mathcal{K} \subseteq \mathcal{M}$ with diameter D, and $\{\mathbf{f}_t\}_{t=1,2,\ldots,T}$ be a series of λ_1-sub g-convex smooth functions. If the sequence $\{x_t\}_{t=1,2,\ldots,T}$ is generated by*

$$x_{t+1} = \mathscr{H}\Big(\mathscr{P}_{\mathcal{K}}\big(\exp_{x_t}(-\alpha g_t)\big)\Big), \tag{37}$$

where the random vector g_t satisfies that $\mathbb{E}[\mathbf{g}_t|x_t] = \nabla \mathbf{f}_t(x_t)$ and $\mathbb{E}[\|\mathbf{g}_t\|] \leq G$, and for the operator $\mathscr{H} : \mathcal{K} \to \mathcal{S}$, there exist a constant $\lambda_2 \geq 0$ satisfies

$$d^2(\mathscr{H}(x), y) \leq d^2(x, y) + \lambda_2, \forall x \in \mathcal{K}, \forall y \in \mathcal{S}.$$

Then with $\alpha = \frac{D}{G\sqrt{\zeta(\kappa,D)T}}$, we have

$$\mathbb{E}\Big[\sum_{t=1}^{T} \mathbf{f}_t(x_t)\Big] - \min_{x\in\mathcal{S}} \sum_{t=1}^{T} \mathbf{f}_t(x) \leq DG\sqrt{\zeta(\kappa, D)T} + \lambda_1 T + \lambda_2 T.$$

Proof Let $x^* = \arg\min_{x\in\mathcal{S}} \sum_{t=1}^{T} \mathbf{f}_t(x)$. From λ_1-sub g-convexity, the difference between $\mathbf{f}_t(x_t)$ and $\mathbf{f}_t(x^*)$ is bounded by

$$\begin{aligned}
\mathbf{f}_t(x_t) - \mathbf{f}_t(x^*) &\leq \langle \nabla \mathbf{f}_t(x_t), \exp_{x_t}^{-1}(x^*)\rangle + \lambda_1 \\
&= \langle \mathbb{E}[\mathbf{g}_t|x_t], \exp_{x_t}^{-1}(x^*)\rangle + \lambda_1 \\
&= \mathbb{E}\Big[\langle g_t, \exp_{x_t}^{-1}(x^*)\rangle | x_t\Big] + \lambda_1.
\end{aligned}$$

Taking the expectation on both sides yields

$$\mathbb{E}[\mathbf{f}_t(x_t) - \mathbf{f}_t(x^*)] \leq \mathbb{E}\Big[\langle g_t, \exp_{x_t}^{-1}(x^*)\rangle\Big] + \lambda_1.$$

From Lemma 35 and 36,

$$\begin{aligned}
\mathbb{E}[\mathbf{f}_t(x_t) - \mathbf{f}_t(x^*)] \leq \mathbb{E}\Big[\frac{1}{2\alpha}(d^2(x_t, x^*) - d^2(\mathscr{P}_{\mathcal{K}}(\exp_{x_t}(-\alpha_t g_t)), x^*))\Big] \\
+ \mathbb{E}\Big[\frac{1}{2}\zeta(\kappa, d(x_t, x^*))\alpha\|g_t\|^2\Big] + \lambda_1.
\end{aligned}$$

Combining with the condition $\mathbb{E}[\|\mathbf{g}_t\|] \leq G$ and $d^2(x_t, x^*) \leq D$, we have

$$\begin{aligned}
\mathbb{E}[\mathbf{f}_t(x_t) - \mathbf{f}_t(x^*)] &\leq \mathbb{E}\Big[\frac{1}{2\alpha}(d^2(x_t, x^*) - d^2(\mathscr{P}_{\mathcal{K}}(\exp_{x_t}(-\alpha g_t), x^*)))\Big] + \frac{1}{2}\zeta(\kappa, D)\alpha G^2 + \lambda_1 \\
&\leq \mathbb{E}\Big[\frac{1}{2\alpha}(d^2(x_t, x^*) - d^2\Big(\mathscr{H}\Big(\mathscr{P}_{\mathcal{K}}\big(\exp_{x_t}(-\alpha_t g_t)\big)\Big), x^*\Big)\Big] \\
&\quad + \frac{1}{2}\zeta(\kappa, D)\alpha G^2 + \lambda_1 + \lambda_2. \\
&= \mathbb{E}\Big[\frac{1}{2\alpha}(d^2(x_t, x^*) - d^2(x_{t+1}, x^*))\Big] + \frac{1}{2}\zeta(\kappa, D)\alpha G^2 + \lambda_1 + \lambda_2. \qquad (38)
\end{aligned}$$

Summing (38) from 1 to T, we have

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{E}[\mathbf{f}_t(x_t) - \mathbf{f}_t(x^*)] &\leq \sum_{t=1}^{T} \mathbb{E}\Big[\frac{1}{2\alpha}(d^2(x_t, x^*) - d^2(x_{t+1}, x^*))\Big] + \frac{1}{2}\zeta(\kappa, D)\alpha G^2 T + \lambda_1 T + \lambda_2 T \\
&\leq \mathbb{E}\Big[\frac{1}{2\alpha} d^2(x_1, x^*)\Big] + \frac{1}{2}\zeta(\kappa, D)\alpha G^2 T + \lambda_1 T + \lambda_2 T \\
&\leq \frac{D^2}{2\alpha} + \frac{1}{2}\zeta(\kappa, D)\alpha G^2 T + \lambda_1 T + \lambda_2 T. \qquad (39)
\end{aligned}$$

Putting $\alpha = \frac{D}{G\sqrt{\zeta(\kappa,D)T}}$ in (39), we complete our proof. ■

Lemma 43 (Strongly Sub-convex Cases) *Suppose that $\mathcal{S}$ is a subset of a g-convex set $\mathcal{K} \subseteq \mathcal{M}$ with diameter D, and $\{\mathbf{f}_t\}_{t=1,2,\ldots,T}$ be a series of μ-strongly λ_1-sub g-convex smooth functions. If the sequence $\{x_t\}_{t=1,2,\ldots,T}$ is generated by*

$$x_{t+1} = \mathscr{H}\Big(\mathscr{P}_{\mathcal{K}}\big(\exp_{x_t}(-\alpha g_t)\big)\Big),$$

where the random vector g_t satisfies that $\mathbb{E}[\mathbf{g}_t|x_t] = \nabla \mathbf{f}_t(x_t)$, $\mathbb{E}[\|\mathbf{g}_t\||x_t] \leq G$,and the operator $\mathscr{H}$ satisfies (37). Then with constant $\nu \geq 1$ and $\alpha_t = \frac{\nu}{\mu t}$ we have that

$$\mathbb{E}\Big[\sum_{t=1}^{T} \mathbf{f}_t(x_t)\Big] - \min_{x\in\mathcal{S}} \sum_{t=1}^{T} \mathbf{f}_t(x) \leq \frac{1}{2}\zeta(\kappa, D)\nu G^2(1+\log T) + \lambda_1 T + \lambda_2 T.$$

Proof Let $x^* = \arg\min_{x\in\mathcal{S}} \sum_{t=1}^{T} \mathbf{f}_t(x)$. From μ-strongly λ_1-sub g-convexity, the difference between $\mathbf{f}_t(x_t)$ and $\mathbf{f}_t(x^*)$ is bounded by

$$\begin{aligned}
\mathbf{f}_t(x_t) - \mathbf{f}_t(x^*) &\leq \langle \nabla \mathbf{f}_t(x_t), \exp_{x_t}^{-1}(x^*)\rangle - \frac{\mu}{2} d^2(x_t, x^*) + \lambda_1 \\
&= \langle \mathbb{E}[\mathbf{g}_t|x_t], \exp_{x_t}^{-1}(x^*)\rangle - \frac{\mu}{2} d^2(x_t, x^*) + \lambda_1 \\
&= \mathbb{E}\Big[\langle g_t, \exp_{x_t}^{-1}(x^*)\rangle | x_t\Big] - \frac{\mu}{2} d^2(x_t, x^*) + \lambda_1.
\end{aligned}$$

Taking the expectation on both sides yields

$$\mathbb{E}[\mathbf{f}_t(x_t) - \mathbf{f}_t(x^*)] \leq \mathbb{E}\Big[\langle g_t, \exp_{x_t}^{-1}(x^*)\rangle - \frac{\mu}{2} d^2(x_t, x^*)\Big] + \lambda_1.$$

From Lemma 35,

$$\begin{aligned}
\mathbb{E}[\mathbf{f}_t(x_t) - \mathbf{f}_t(x^*)] &\leq \mathbb{E}\Big[\frac{1}{2\alpha_t}(d^2(x_t, x^*) - d^2(\mathscr{P}_{\mathcal{K}}(\exp_{x_t}(-\alpha_t g_t))), x^*)) - \frac{\mu}{2} d^2(x_t, x^*)\Big] \\
&\quad + \mathbb{E}\Big[\frac{1}{2}\zeta(\kappa, d(x_t, x^*))\alpha_t \|g_t\|^2\Big] + \lambda_1. \\
&\leq \mathbb{E}\Big[\frac{1}{2\alpha_t}(d^2(x_t, x^*) - d^2(x_{t+1}, x^*)) - \frac{\mu}{2} d^2(x_t, x^*)\Big] \\
&\quad + \mathbb{E}\Big[\frac{1}{2}\zeta(\kappa, d(x_t, x^*))\alpha_t \|g_t\|^2\Big] + \lambda_1 + \lambda_2
\end{aligned}$$

Combining with the condition $\mathbb{E}[\|\mathbf{g}_t\|] \leq G$ and $d^2(x_t, x^*) \leq D$, we have

$$\begin{aligned}
\mathbb{E}[\mathbf{f}_t(x_t) - \mathbf{f}_t(x^*)] &\leq \mathbb{E}\Big[\frac{1}{2\alpha_t}(d^2(x_t, x^*) - d^2(x_{t+1}, x^*)) - \frac{\mu}{2} d^2(x_t, x^*)\Big] \\
&\quad + \frac{1}{2}\zeta(\kappa, D)\alpha_t G^2 + \lambda_1 + \lambda_2. \qquad (40)
\end{aligned}$$

Summing (40) from 1 to T, we have

$$\sum_{t=1}^{T}\mathbb{E}[\mathbf{f}_t(x_t)-\mathbf{f}_t(x^*)] \le \sum_{t=1}^{T}\mathbb{E}\Big[d^2(x_t,x^*)(\frac{1}{2\alpha_t}-\frac{1}{2\alpha_{t-1}}-\frac{\mu}{2})\Big] + \frac{1}{2}\zeta(\kappa,D)G^2\sum_{t=1}^{T}\alpha_t+\lambda_1 T+\lambda_2 T. \tag{41}$$

Since $\alpha_t=\frac{\nu}{\mu t}$ and $\nu\ge 1$,

$$\frac{1}{2\alpha_t}-\frac{1}{2\alpha_{t-1}}-\frac{\mu}{2}=\frac{\mu}{2}(\frac{1}{\nu}-1)\le 0.$$

Putting it into (41), we find

$$\begin{aligned}\sum_{t=1}^{T}\mathbb{E}[\mathbf{f}_t(x_t)-\mathbf{f}_t(x^*)] &\le \frac{1}{2}\zeta(\kappa,D)G^2\nu\sum_{t=1}^{T}\frac{1}{\mu t}+\lambda_1 T+\lambda_2 T.\\ &\le \frac{1}{2\mu}\zeta(\kappa,D)\nu G^2(1+\log T)+\lambda_1 T+\lambda_2 T,\end{aligned}$$

which completes our proof. ■

Lemma 44 relates the gap between regret of $\hat{\mathbf{f}}_t$ and the real regret of $\mathbf{f}_t$.

Lemma 44 *Suppose all $\mathbf{f}_t$ are g-L-Lipschitz. The (expected) regret of the R-BAN algorithm satisfies*

$$\mathbb{E}[\mathsf{Reg}(T)]\le \mathbb{E}\Big[\sum_{t=1}^{T}(\hat{\mathbf{f}}_t(y_t))\Big]-\min_{x\in(1-\tau)\mathcal{K}}\sum_{t=1}^{T}\hat{\mathbf{f}}_t(x)+3\delta LT+\tau DL.T$$

Proof Denote by x_τ^* the minimizer of the problem $\min_{x\in(1-\tau)\mathcal{K}}\sum_{t=1}^{T}\mathbf{f}_t(x)$. The expectation can be reformulated as follows

$$\begin{aligned}\mathbb{E}[\mathsf{Reg}(T)] &= \sum_{t=1}^{T}\mathbb{E}\Big[\mathbf{f}_t(x_t)-\mathbf{f}_t(x^*)\Big]\\ &= \mathbb{E}\Big[\sum_{t=1}^{T}(\mathbf{f}_t(x_t)-\mathbf{f}_t(y_t))\Big]+\mathbb{E}\Big[\sum_{t=1}^{T}(\mathbf{f}_t(y_t)-\hat{\mathbf{f}}_t(y_t))\Big]+\mathbb{E}\Big[\sum_{t=1}^{T}(\hat{\mathbf{f}}_t(y_t)-\hat{\mathbf{f}}_t(x_\tau^*))\Big]\\ &\quad+\mathbb{E}\Big[\sum_{t=1}^{T}(\hat{\mathbf{f}}_t(x_\tau^*)-\mathbf{f}_t(x_\tau^*))\Big]+\mathbb{E}\Big[\sum_{t=1}^{T}(\mathbf{f}_t(x_\tau^*)-\mathbf{f}_t(x^*))\Big].\end{aligned}$$

The Lipschitz condition leads to

$$\begin{cases}\mathbf{f}_t(x_t)-\mathbf{f}_t(y_t)\le\delta L\\ \mathbf{f}_t(y_t)-\hat{\mathbf{f}}_t(y_t)\le\delta L\\ \hat{\mathbf{f}}_t(x_\tau^*)-\mathbf{f}_t(x_\tau^*)\le\delta L,\end{cases}$$

which gives us

$$\mathbb{E}[\mathsf{Reg}(T)] \leq \mathbb{E}\Big[\sum_{t=1}^{T}(\hat{\mathbf{f}}_t(y_t) - \hat{\mathbf{f}}_t(x_\tau^*))\Big] + \mathbb{E}\Big[\sum_{t=1}^{T}(\mathbf{f}_t(x_\tau^*) - \mathbf{f}_t(x^*))\Big] + 3\delta LT. \tag{42}$$

Next, we notice that $(1-\tau)\mathcal{K} = \{\exp_p((1-\tau)u) | u = \exp_p^{-1}(x) \in \mathcal{K}\}$, It is easy to check that

$$\begin{aligned}\min_{x\in(1-\tau)\mathcal{K}} \sum_{t=1}^{T} \mathbf{f}_t(x) &= \min_{x\in\mathcal{K}} \sum_{t=1}^{T} \mathbf{f}_t\Big(\exp_p((1-\tau)\exp_p^{-1}(x))\Big) \\ &\leq \min_{x\in\mathcal{K}}(\sum_{t=1}^{T} \mathbf{f}_t(x) + \tau DL) \\ &\leq \tau DLT + \min_{x\in\mathcal{K}} \sum_{t=1}^{T} \mathbf{f}_t(x).\end{aligned}$$

This forces

$$\sum_{t=1}^{T}(\mathbf{f}_t(x_\tau^*) - \mathbf{f}_t(x^*)) \leq \tau DLT. \tag{43}$$

Combining (42) and (43) together, we can conclude

$$\mathbb{E}[\mathsf{Reg}(T)] \leq \mathbb{E}\Big[\sum_{t=1}^{T}(\hat{\mathbf{f}}_t(y_t) - \hat{\mathbf{f}}_t(x_\tau^*))\Big] + 3\delta LT + \tau DLT.$$

which is the desired result. ■

Lemma 45 guarantees the feasibility of the proposed bandit algorithms.

Lemma 45 *Let $\mathcal{M}$ be a Riemannian manifold with sectional curvature bounded below by a constant $\kappa \leq 0$ and above by a constant $K \geq 0$. Suppose there exists a g-convex set $\mathcal{K} \subseteq \mathcal{M}$, a point $p \in \mathcal{K}$, and two constants $0 < r \leq D$ such that $B_r(p) \subseteq \mathcal{K} \subseteq B_D(p)$, where $D \leq \frac{\pi}{2\sqrt{K}}$ if $K > 0$. Denote $\theta = \frac{\mathbf{s}(K,D+r)}{\mathbf{s}(\kappa,D+r)} \leq 1$. Then, for every $y \in (1-\tau)\mathcal{K}$, the geodesic ball $B_{\theta\cdot\tau r}(y)$ lies in $\mathcal{K}$.*

Proof Let $x \in \mathcal{K}$ be arbitrary and $y = \exp_p\big((1-\tau)\exp_p^{-1}(x)\big)$. For any $v \in T_y\mathcal{M}$ with $\|v\| = 1$, let $z = \exp_y\big(\theta\tau r \cdot v\big)$. We will show that $z \in \mathcal{K}$.

Denote $\xi(s)$ to be the geodesic with $\xi(0) = y$ and $\xi(1) = z$. Then we can define a rectangle map

$$\Gamma : [0, 1/\tau] \times [0, 1] \to \mathcal{K}; \qquad (t, s) \mapsto \exp_x\big(t \exp_x^{-1} \xi(s)\big).$$

Now, we have a vector field $v(t,s) = \frac{\partial\Gamma}{\partial s}(s,t)$ over the rectangle. The vector field $v(t,s)$ is a variation field of the geodesic $\gamma_s(t) = \exp_x\big(t \exp_x^{-1} \xi(s)\big)$ with $v(0,s) = 0$. Thus, $v(s,t)$ is an initial zero Jacobi field and we can apply Theorem 30 at $t = 1$. Since the sectional

curvature of $\mathcal{M}$ is upper bounded by K, we can use the normalization of the geodesic $\gamma_s(t)$ to get

$$\theta\tau r = \|v(1,s)\| \geq \mathbf{s}\big(K, d(x,\xi(s))\big)\|\dot{v}(0,s)\|. \tag{44}$$

Next, we apply Theorem 30 to $v(s,t)$ at $t=\frac{1}{\tau}$. Since the sectional curvature of $\mathcal{M}$ is lower bounded by κ, we can use the normalization of the geodesic $\gamma_s(t)$ to get

$$\|v(1/\tau,s)\| \leq \mathbf{s}(\kappa, \frac{d(x,\xi(s))}{\tau})\|\dot{v}(0,s)\| \tag{45}$$

Combining (44) and (45), we obtain

$$\|v(1/\tau,s)\| \leq \frac{\mathbf{s}(\kappa, \frac{d(x,\xi(s))}{\tau})}{\mathbf{s}(K, d(x,\xi(s)))}\tau\theta r.$$

By concavity of the function $\sin(x)$ in $[0,\pi]$, we can get

$$\sin(\sqrt{K}d(x,\xi(s))) \geq \tau\sin(\sqrt{K}\frac{d(x,\xi(s))}{\tau}),$$

which derives

$$\frac{1}{\tau}\mathbf{s}(K, d(x,\xi(s))) \geq \mathbf{s}(K, \frac{d(x,\xi(s))}{\tau})$$

and

$$\|v(1/\tau,s)\| \leq \frac{\mathbf{s}(\kappa, \frac{d(x,\xi(s))}{\tau})}{\mathbf{s}(K, d(x,\xi(s)))}\tau\theta r \leq \frac{\mathbf{s}(\kappa, \frac{d(x,\xi(s))}{\tau})}{\mathbf{s}(K, \frac{d(x,\xi(s))}{\tau})}\theta r$$

Since $d(x,\xi(s)) \leq d(x,y)+d(y,z) \leq \tau(D+r)$, we have

$$\|v(1/\tau,s)\| \leq \frac{\mathbf{s}(\kappa, D+r)}{\mathbf{s}(K, D+r)}\theta r = \frac{\mathbf{s}(\kappa, D+r)}{\mathbf{s}(K, D+r)}\frac{\mathbf{s}(K, D+r)}{\mathbf{s}(\kappa, D+r)}r = r.$$

Hence, the length of the curve $c(s) = \Gamma(1/\tau, s)$ is bounded by

$$l(c(s)) = \int_0^1 \|v(1/\tau,s)\|ds \leq \int_0^1 r ds = r.$$

Notice that $p = \Gamma(1/\tau, 0)$ and denote $w = \Gamma(1/\tau, 1)$, we have $d(p,w) \leq l(c(s)) = r$, which means that $w \in \mathcal{B}_r(p) \subset \mathcal{K}$.

Therefore, for the geodesic $\gamma_1(t)$, we have $\gamma_1(0) = x \in \mathcal{K}$ and $\gamma_1(1/\tau) = w \in \mathcal{K}$. Thus, by g-convexity, we have $z = \gamma_1(1) \in \mathcal{K}$, which completes our proof. ■

Now we carry out the proofs of Theorems 14 and 15.

E.1 Proof of Theorem 14

First, we focus on the update rule of y_t, that is

$$\begin{aligned} y_{t+1} &= \mathscr{P}_{(1-\tau)\mathcal{K}} \circ \mathscr{P}_{\mathcal{K}}\bigg(\exp_{y_t}(\alpha_t \mathbf{g}_t)\bigg) \\ &= \mathscr{P}_{(1-\tau)\mathcal{K}} \circ \mathscr{P}_{\mathcal{K}}\bigg(\exp_{y_t}\big(\frac{D}{C\sqrt{\zeta(\kappa,D)T}}\mathbf{g}_t\big)\bigg) \\ &= \mathscr{P}_{(1-\tau)\mathcal{K}} \circ \mathscr{P}_{\mathcal{K}}\bigg(\exp_{y_t}\big(\frac{D}{\frac{S_\delta}{V_\delta}C\sqrt{\zeta(\kappa,D)T}}(\frac{S_\delta}{V_\delta}\mathbf{g}_t)\big)\bigg). \end{aligned}$$

Since

$$\mathbf{g}_t = \mathbf{f}_t(x_t)\frac{\exp_{y_t}^{-1}(x_t)}{\|\exp_{y_t}^{-1}(x_t)\|},$$

from Lemma 11 we obtain $\mathbb{E}\Big[\frac{S_\delta}{V_\delta}\mathbf{g}_t\Big|y_t\Big] = \nabla\hat{\mathbf{f}}(y_t)$ and $\mathbb{E}\Big[\|\frac{S_\delta}{V_\delta}\mathbf{g}_t\|\Big] \leq \frac{S_\delta}{V_\delta}C$. Moreover, according to Lemma 13, $\hat{\mathbf{f}}_t$ is $2\delta\rho L$-sub g-convex. Also, for all $x \in \mathcal{K}$ and $y \in (1-\tau)\mathcal{K}$, we have

$$\begin{aligned} d^2(\mathscr{P}_{(1-\tau)\mathcal{K}}(x)), y) - d^2(x,y) &\leq 2D \cdot d(\mathscr{P}_{(1-\tau)\mathcal{K}}(x)), x) \\ &\leq 2D \cdot d(\exp_p((1-\tau)\exp_p^{-1}(x)), x) \\ &\leq (2D)(\tau D) = 2\tau D^2. \end{aligned} \tag{46}$$

Thus, the update rule in Algorithm 2 is exactly the expected gradient descent in Lemma 42 with parameters $\mathcal{S} = (1-\tau)\mathcal{K}$, $G = \frac{S_\delta}{V_\delta}C$, $\lambda_1 = 2\delta\rho L$ and $\lambda_2 = 2\tau D^2$. We can get

$$\mathbb{E}\Big[\sum_{t=1}^{T}\mathbf{f}_t(y_t)\Big] - \min_{x\in(1-\tau)\mathcal{K}}\sum_{t=1}^{T}\mathbf{f}_t(x) \leq \frac{S_\delta}{V_\delta}DC\sqrt{\zeta(\kappa,D)T} + 2\delta\rho LT + 2\tau D^2 T. \tag{47}$$

From the inequality $\frac{S_\delta}{V_\delta} \leq \frac{n}{\delta} + n|\kappa|\delta$ in Lemma 11, we have

$$\mathbb{E}\Big[\sum_{t=1}^{T}\mathbf{f}_t(y_t)\Big] - \min_{x\in(1-\tau)\mathcal{K}}\sum_{t=1}^{T}\mathbf{f}_t(x) \leq (\frac{n}{\delta} + n|\kappa|\delta)DC\sqrt{\zeta(\kappa,D)T} + 2\delta\rho LT + 2\tau D^2 T. \tag{48}$$

Applying Lemma 44 in (48), we have that

$$\mathbb{E}[\mathsf{Reg}(T)] \leq (\frac{n}{\delta} + n|\kappa|\delta)DC\sqrt{\zeta(\kappa,D)T} + 3\delta LT + \tau DLT + 2\delta\rho LT + 2\tau D^2 T. \tag{49}$$

Finally, taking $\tau = \frac{\delta}{r\theta}$ and $\delta = T^{-\frac{1}{4}}$ in (49), we get

$$\begin{aligned} \mathbb{E}[\mathsf{Reg}(T)] &\leq \frac{n}{\delta}DC\sqrt{\zeta(\kappa,D)T} + n|\kappa|\delta DC\sqrt{\zeta(\kappa,D)T} + 3\delta LT + \tau DLT + 2\delta\rho LT + 2\tau D^2 T \\ &\leq nDC\sqrt{\zeta(\kappa,D)}T^{\frac{3}{4}} + n|\kappa|DC\sqrt{\zeta(\kappa,D)}T^{\frac{1}{4}} + (3L + \frac{DL}{r\theta} + \frac{2D^2}{r\theta} + 2\rho L)T^{\frac{3}{4}} \\ &\leq n|\kappa|DC\sqrt{\zeta(\kappa,D)}T^{\frac{1}{4}} + \Big(nDC\sqrt{\zeta(\kappa,D)} + 3L + \frac{DL}{r\theta} + \frac{2D^2}{r\theta} + 2\rho L\Big)T^{\frac{3}{4}}, \end{aligned}$$

which completes our proof. ■

E.2 Proof of Theorem 15

As we have carried out in the proof of Theorem 14, we focus on the update rule of y_t, that is

$$\begin{aligned} y_{t+1} &= \mathscr{P}_{(1-\tau)\mathcal{K}} \circ \mathscr{P}_{\mathcal{K}}\Big(\exp_{y_t}\big(\alpha_t \mathbf{g}_t\big)\Big) \\ &= \mathscr{P}_{(1-\tau)\mathcal{K}} \circ \mathscr{P}_{\mathcal{K}}\Big(\exp_{y_t}\big(\frac{B}{\mu t}\mathbf{g}_t\big)\Big) \\ &= \mathscr{P}_{(1-\tau)\mathcal{K}} \circ \mathscr{P}_{\mathcal{K}}\Big(\exp_{y_t}\big(\frac{B}{\frac{S_\delta}{V_\delta}\mu t}\frac{S_\delta}{V_\delta}\mathbf{g}_t\big)\Big). \end{aligned}$$

Since

$$\mathbf{g}_t = \mathbf{f}(x_t)\frac{\exp_{y_t}^{-1}(x_t)}{\|\exp_{y_t}^{-1}(x_t)\|},$$

from Lemma 11 we obtain $\mathbb{E}\Big[\frac{S_\delta}{V_\delta}\mathbf{g}_t\Big|y_t\Big] = \nabla\hat{\mathbf{f}}(y_t)$ and $\mathbb{E}\Big[\|\frac{S_\delta}{V_\delta}\mathbf{g}_t\|\Big] \leq \frac{S_\delta}{V_\delta}C$. Additionally, according to Lemma 13, $\hat{\mathbf{f}}_t$ is μ-strongly $(2\rho L + 2D\mu)\delta$-sub g-convex. Thus, the update rule in Algorithm 2 is exactly the expected gradient descent in Lemma 43 with parameters $\mathcal{S} = (1-\tau)\mathcal{K}$, $G = \frac{S_\delta}{V_\delta}C$, $\lambda_1 = (2\rho L + 2D\mu)\delta$, $\lambda_2 = 2\tau D^2$ and $\nu = \frac{BV_\delta}{S_\delta}$. We can get,

$$\begin{aligned} \mathbb{E}\Big[\sum_{t=1}^{T}\mathbf{f}_t(y_t)\Big] - \min_{x\in(1-\tau)\mathcal{K}}\sum_{t=1}^{T}\mathbf{f}_t(x) &\leq \frac{1}{2\mu}\zeta(\kappa,D)(\frac{S_\delta}{V_\delta})^2C^2\frac{BV_\delta}{S_\delta}(1+\log T) \\ &\quad + (2\rho L + 2D\mu)\delta T + 2\tau D^2 T \\ &\leq \frac{1}{2\mu}B\zeta(\kappa,D)\frac{S_\delta}{V_\delta}C^2(1+\log T) \\ &\quad + (2\rho L + 2D\mu)\delta T + 2\tau D^2 T. \end{aligned}$$

From the inequality $\frac{S_\delta}{V_\delta} \leq \frac{n}{\delta} + n|\kappa|\delta = B$ in Lemma 11, we have

$$\begin{aligned} \mathbb{E}\Big[\sum_{t=1}^{T}\mathbf{f}_t(y_t)\Big] - \min_{x\in(1-\tau)\mathcal{K}}\sum_{t=1}^{T}\mathbf{f}_t(x) &\leq \frac{1}{2\mu}(\frac{n}{\delta} + n|\kappa|\delta)^2\zeta(\kappa,D)C^2(1+\log T) \\ &\quad + (2\rho L + 2D\mu)\delta T + 2\tau D^2 T \\ &\leq \frac{n^2}{\mu}(\frac{1}{\delta^2} + \kappa^2\delta^2)\zeta(\kappa,D)C^2(1+\log T) \\ &\quad + (2\rho L + 2D\mu)\delta T + 2\tau D^2 T. \end{aligned}$$

Applying Lemma 44, we have that

$$\begin{aligned} \mathbb{E}[R(T)] \leq &\frac{1}{\mu}(\frac{n^2}{\delta^2} + n^2\kappa^2\delta^2)\zeta(\kappa,D)C^2(1+\log T) + (2\rho L + 2D\mu)\delta T \\ &+ 3\delta LT + \tau DLT + 2\tau D^2 T. \end{aligned} \tag{50}$$

Taking $\tau = \frac{\delta}{r\theta}$ and $\delta = \sqrt[3]{\frac{n^2C^2(1+\log T)}{T}}$ in (50), we get

$$\begin{aligned}
\mathbb{E}[\mathsf{Reg}(T)] &\le \frac{n^{\frac{2}{3}}C^{\frac{2}{3}}\zeta(\kappa,D)}{\mu}(1+\log T)^{\frac{1}{3}}T^{\frac{2}{3}} + \frac{n^2C^2\delta^2\kappa^2\zeta(\kappa,D)}{\mu}(1+\log T) \\
&\quad + n^{\frac{2}{3}}C^{\frac{2}{3}}(3L + \frac{DL}{r\theta} + \frac{2D^2}{r\theta} + 2\rho L + 2D\mu)(1+\log T)^{\frac{1}{3}}T^{\frac{2}{3}} \\
&\le \frac{n^{\frac{2}{3}}C^{\frac{2}{3}}\zeta(\kappa,D)}{\mu}(1+\log T)^{\frac{1}{3}}T^{\frac{2}{3}} + \frac{n^{\frac{8}{3}}C^{\frac{8}{3}}D\kappa^2\zeta(\kappa,D)}{\mu}(1+\log T)^{\frac{4}{3}}T^{-\frac{1}{3}} \\
&\quad + n^{\frac{2}{3}}C^{\frac{2}{3}}(3L + \frac{DL}{r\theta} + \frac{2D^2}{r\theta} + 2\rho L + 2D\mu)(1+\log T)^{\frac{1}{3}}T^{\frac{2}{3}} \\
&\le \frac{2n^{\frac{8}{3}}C^{\frac{8}{3}}D\kappa^2\zeta(\kappa,D)}{\mu} \\
&\quad + n^{\frac{2}{3}}C^{\frac{2}{3}}\big(\frac{\zeta(\kappa,D)}{\mu} + 3L + \frac{DL}{r\theta} + \frac{2D^2}{r\theta} + 2D\mu\big)(1+\log T)^{\frac{1}{3}}T^{\frac{2}{3}},
\end{aligned}$$

where the last inequality is due to $\max_{T\ge 1}\sqrt[3]{\frac{(1+\log T)^4}{T}} = \frac{4}{e}\sqrt[3]{4} \le 2$.Then we completes our proof. ■

Appendix F. Proofs of Theorems 18 and 19

In this section, we present proof of the regret bounds of the R-2-BAN algorithm.

F.1 Proof of Lemma 17

We first examine (i). By Lemma 38, $\mathcal{M}$ is homogeneous. Thus from Lemma 11 we know

$$\begin{aligned}
\nabla\hat{\mathbf{f}}(x) &= \frac{S_\delta}{V_\delta}\mathbb{E}_{u\in S_\delta(x)}\Big[\mathbf{f}(u)\frac{\exp_x^{-1}(u)}{\|\exp_x^{-1}(u)\|}\Big] \\
&= \frac{1}{V_\delta}\int_{S_\delta(x)}\mathbf{f}(u)\frac{\exp_x^{-1}(u)}{\|\exp_x^{-1}(u)\|}\omega_{S_\delta(x)}. \qquad (51)
\end{aligned}$$

From Lemma 37, we notice that the symmetry s_x maps a random variable $u \in \mathcal{M}$ to

$$s_x(u) = \exp(-\exp_x^{-1}(u)) = -u.$$

Substituting $s_x(u)$ with u in (51) yields

$$\begin{aligned}
\nabla\hat{\mathbf{f}}(x) &= \frac{1}{V_\delta}\int_{S_\delta(x)}\mathbf{f}(s_x(u))\frac{\exp_x^{-1}(s_x(u))}{\|\exp_x^{-1}(s_x(u))\|}s_x^*\omega_{S_\delta(x)} \\
&= \frac{1}{V_\delta}\int_{S_\delta(x)}\mathbf{f}(-u)\frac{-\exp_x^{-1}(u)}{\|-\exp_x^{-1}(u)\|}s_x^*\omega_{S_\delta(x)}.
\end{aligned}$$

Since s_x is a isometry, we have $s_x^*\omega_{S_\delta(x)} = \omega_{S_\delta(x)}$. Therefore,

$$\begin{aligned}
\nabla\hat{\mathbf{f}}(x) &= \frac{1}{V_\delta}\int_{S_\delta(x)} -\mathbf{f}(-u)\frac{\exp_x^{-1}(u)}{\|\exp_x^{-1}(u)\|}\omega_{S_\delta(x)} \\
&= \frac{S_\delta}{V_\delta}\mathbb{E}_{u\in S_\delta(x)}\Big[-\mathbf{f}(-u)\frac{\exp_x^{-1}(u)}{\|\exp_x^{-1}(u)\|}\Big]. \qquad (52)
\end{aligned}$$

Combining (51) and (52), we have

$$\begin{aligned}2\nabla\hat{\mathbf{f}}(x) &= \frac{S_\delta}{V_\delta}\mathbb{E}_{u\in S_\delta(x)}\Big[\mathbf{f}(u)\frac{\exp_x^{-1}(u)}{\|\exp_x^{-1}(u)\|}\Big] + \frac{S_\delta}{V_\delta}\mathbb{E}_{u\in S_\delta(x)}\Big[-\mathbf{f}(-u)\frac{\exp_x^{-1}(u)}{\|\exp_x^{-1}(u)\|}\Big]\\ &= \frac{S_\delta}{V_\delta}\mathbb{E}_{u\in S_\delta(x)}\Big[(\mathbf{f}(u)-\mathbf{f}(-u))\frac{\exp_x^{-1}(u)}{\|\exp_x^{-1}(u)\|}\Big]\\ &= \frac{S_\delta}{V_\delta}\mathbb{E}_{u\in S_\delta(x)}\big[\tilde{\mathbf{g}}\big],\end{aligned}$$

which completes the proof of (i).

Then we prove (ii). Notice that

$$\mathbb{E}_{u\in S_\delta(x)}[\frac{S_\delta}{V_\delta}\|\tilde{g}\|] \le \frac{S_\delta}{2V_\delta}|\mathbf{f}(\exp_x u)-\mathbf{f}(\exp_x -u)| \le \frac{S_\delta}{2V_\delta}2L\delta. \tag{53}$$

It follows from Lemma 11 that

$$\frac{S_\delta}{V_\delta} \le \frac{n}{\delta} + n|\kappa|\delta. \tag{54}$$

Putting (53) and (54) together we get

$$\mathbb{E}_{u\in S_\delta(x)}[\frac{S_\delta}{V_\delta}\|\tilde{\mathbf{g}}\|] \le \frac{1}{2}(\frac{n}{\delta}+n|\kappa|\delta)2L\delta = nL(1+|\kappa|\delta^2),$$

which completes our proof. ■

F.2 Gap between Regrets

As in the one-point bandit case, we also conclude the gap between the regret of $\hat{\mathbf{f}}_t$ and the real regret of $\mathbf{f}_t$ in Lemma 46 for the two-point bandit case.

Lemma 46 *Suppose all $\mathbf{f}_t$ are g-L-Lipschitz. The (expected) regret of the R-BAN algorithm satisfies*

$$\mathbb{E}[\mathsf{Reg}(T)] \le \mathbb{E}\Big[\sum_{t=1}^{T}(\hat{\mathbf{f}}_t(y_t))\Big] - \min_{x\in(1-\tau)\mathcal{K}}\sum_{t=1}^{T}\hat{\mathbf{f}}_t(x) + 3\delta LT + \tau DLT.$$

Proof Denote by x_τ^* the minimizer of the problem $\min_{x\in(1-\tau)\mathcal{K}}\sum_{t=1}^{T}\mathbf{f}_t(x)$. The expectation can be reformulated as follows

$$\begin{aligned}\mathbb{E}[\mathsf{Reg}(T)] &= \sum_{t=1}^{T}\mathbb{E}\Big[\frac{1}{2}(\mathbf{f}_t(x_{t,1})+\mathbf{f}_t(x_{t,2}))-\mathbf{f}_t(x^*)\Big]\\ &= \mathbb{E}\Big[\sum_{t=1}^{T}(\frac{1}{2}(\mathbf{f}_t(x_{t,1})+\mathbf{f}_t(x_{t,2}))-\mathbf{f}_t(y_t))\Big] + \mathbb{E}\Big[\sum_{t=1}^{T}(\mathbf{f}_t(y_t)-\hat{\mathbf{f}}_t(y_t))\Big]\\ &+ \mathbb{E}\Big[\sum_{t=1}^{T}(\hat{\mathbf{f}}_t(y_t)-\hat{\mathbf{f}}_t(x_\tau^*))\Big] + \mathbb{E}\Big[\sum_{t=1}^{T}(\hat{\mathbf{f}}_t(x_\tau^*)-\mathbf{f}_t(x_\tau^*))\Big] + \mathbb{E}\Big[\sum_{t=1}^{T}(\mathbf{f}_t(x_\tau^*)-\mathbf{f}_t(x^*))\Big].\end{aligned}$$

The Lipschitz condition leads to

$$\begin{cases} \mathbf{f}_t(x_{t,1}) - \mathbf{f}_t(y_t) \le \delta L \\ \mathbf{f}_t(x_{t,2}) - \mathbf{f}_t(y_t) \le \delta L \\ \mathbf{f}_t(y_t) - \hat{\mathbf{f}}_t(y_t) \le \delta L \\ \hat{\mathbf{f}}_t(x_\tau^*) - \mathbf{f}_t(x_\tau^*) \le \delta L. \end{cases}$$

In addition, Lemma 44 shows that

$$\sum_{t=1}^{T} (\mathbf{f}_t(x_\tau^*) - \mathbf{f}_t(x^*)) \le \tau DLT.$$

Thus, we can conclude

$$\mathbb{E}[\mathsf{Reg}(T)] \le \mathbb{E}\Big[\sum_{t=1}^{T} (\hat{\mathbf{f}}_t(y_t) - \hat{\mathbf{f}}_t(x_\tau^*))\Big] + 3\delta LT + \tau DLT.$$

which is the desired result. ■

F.3 Proof of Theorem 18

Denote $\tilde{\mathbf{g}}_t = \frac{S_\delta}{2V_\delta}\big(\mathbf{f}_t(\exp_{x_t} u_t) - \mathbf{f}_t(\exp_{x_t}(-u_t))\big)\frac{u_t}{\|u_t\|}$. The update rule of y_t is

$$\begin{aligned} y_{t+1} &= \mathscr{P}_{(1-\tau)\mathcal{K}} \circ \mathscr{P}_{\mathcal{K}}\Big(\exp_{y_t}(\alpha_t \tilde{\mathbf{g}}_t)\Big) \\ &= \mathscr{P}_{(1-\tau)\mathcal{K}} \circ \mathscr{P}_{\mathcal{K}}\Big(\exp_{y_t}(\frac{D}{\delta L\sqrt{\zeta(\kappa, D)T}}\tilde{\mathbf{g}}_t)\Big) \\ &= \mathscr{P}_{(1-\tau)\mathcal{K}} \circ \mathscr{P}_{\mathcal{K}}\Big(\exp_{y_t}(\frac{D}{\frac{S_\delta}{V_\delta}\delta L\sqrt{\zeta(\kappa, D)T}}\frac{S_\delta}{V_\delta}\tilde{\mathbf{g}}_t)\Big). \end{aligned}$$

Since

$$\tilde{\mathbf{g}}_t = \frac{1}{2}\big(\mathbf{f}_t(x_{t,1}) - \mathbf{f}_t(x_{t,2})\big)\frac{\exp_{y_t}^{-1}(x_{t,1})}{\|\exp_{y_t}^{-1}(x_{t,1})\|},$$

from Lemma 17 we obtain $\mathbb{E}\Big[\frac{S_\delta}{V_\delta}\tilde{\mathbf{g}}_t\Big|y_t\Big] = \nabla\hat{\mathbf{f}}(y_t)$ and $\mathbb{E}\Big[\|\frac{S_\delta}{V_\delta}\tilde{\mathbf{g}}_t\|\Big] \le \frac{S_\delta}{V_\delta}\delta L$. Additionally, according to Lemma 13, $\hat{\mathbf{f}}_t$ is $2\delta\rho L$-sub g-convex. Thus, the update rule in Algorithm 2 is exactly the expected gradient descent in Lemma 42 with parameters $\mathcal{S} = (1-\tau)\mathcal{K}$, $G = \frac{S_\delta}{V_\delta}\delta L$, $\lambda_1 = 2\delta\rho L$ and $\lambda_2 = 2\tau D^2$. We can get

$$\mathbb{E}\Big[\sum_{t=1}^{T} \mathbf{f}_t(y_t)\Big] - \min_{x\in(1-\tau)\mathcal{K}} \sum_{t=1}^{T} \mathbf{f}_t(x) \le \frac{S_\delta}{V_\delta} D\delta L\sqrt{\zeta(\kappa, D)T} + 2\delta\rho LT + 2\tau D^2 T. \tag{55}$$

From the inequality $\frac{S_\delta}{V_\delta} \le \frac{n}{\delta} + n|\kappa|\delta$ in Lemma 11, we have

$$\mathbb{E}\Big[\sum_{t=1}^{T} \mathbf{f}_t(y_t)\Big] - \min_{x\in(1-\tau)\mathcal{K}} \sum_{t=1}^{T} \mathbf{f}_t(x) \le (\frac{n}{\delta} + n|\kappa|\delta) D\delta L\sqrt{\zeta(\kappa, D)T} + 2\delta\rho LT + 2\tau D^2 T. \tag{56}$$

Applying Lemma 44 in (56), we have that

$$\mathbb{E}[\mathsf{Reg}(T)] \leq (\frac{n}{\delta} + n|\kappa|\delta)D\delta L\sqrt{\zeta(\kappa, D)T} + 3\delta LT + \tau DLT + 2\delta\rho LT + 2\tau D^2 T. \tag{57}$$

Finally, taking $\tau = \frac{\delta}{r\theta}$ and $\delta = \frac{1}{\sqrt{T}}$ in (57), we get

$$\begin{aligned}
\mathbb{E}[\mathsf{Reg}(T)] &\leq nDL\sqrt{\zeta(\kappa, D)T} + n|\kappa|\delta^2 DL\sqrt{\zeta(\kappa, D)T} + 3\delta LT + \tau DLT + 2\delta\rho LT + 2\tau D^2 T \\
&\leq nDL\sqrt{\zeta(\kappa, D)T} + n|\kappa|DL\sqrt{\zeta(\kappa, D)}\frac{1}{\sqrt{T}} + (3L + \frac{DL}{r\theta} + \frac{2D^2}{r\theta} + 2\rho L)\sqrt{T} \\
&\leq n|\kappa|DL\sqrt{\zeta(\kappa, D)}\frac{1}{\sqrt{T}} + \Big(nDL\sqrt{\zeta(\kappa, D)} + 3L + \frac{DL}{r\theta} + \frac{2D^2}{r\theta} + 2\rho L\Big)\sqrt{T},
\end{aligned}$$

which completes our proof. ■

F.4 Proof of Theorem 19

As we do in the proof of Theorem 14, we focus on the update rule of y_t, that is

$$\begin{aligned}
y_{t+1} &= \mathscr{P}_{(1-\tau)\mathcal{K}} \circ \mathscr{P}_{\mathcal{K}}\Big(\exp_{y_t}\big(\alpha_t \tilde{\mathbf{g}}_t\big)\Big) \\
&= \mathscr{P}_{(1-\tau)\mathcal{K}} \circ \mathscr{P}_{\mathcal{K}}\Big(\exp_{y_t}\big(\frac{B}{\mu t}\tilde{\mathbf{g}}_t\big)\Big) \\
&= \mathscr{P}_{(1-\tau)\mathcal{K}} \circ \mathscr{P}_{\mathcal{K}}\Big(\exp_{y_t}\big(\frac{B}{\frac{S_\delta}{V_\delta}\mu t}\frac{S_\delta}{V_\delta}\tilde{\mathbf{g}}_t\big)\Big).
\end{aligned}$$

Since

$$\tilde{\mathbf{g}}_t = \frac{S_\delta}{2V_\delta}\big(\mathbf{f}(\exp_{x_t} u_t) - \mathbf{f}_t(\exp_{x_t}(-u_t))\big)\frac{u_t}{\|u_t\|},$$

from Lemma 17 we obtain $\mathbb{E}\Big[\frac{S_\delta}{V_\delta}\mathbf{g}_t\Big|y_t\Big] = \nabla\hat{\mathbf{f}}(y_t)$ and $\mathbb{E}\Big[\|\frac{S_\delta}{V_\delta}\mathbf{g}_t\|\Big] \leq \frac{S_\delta}{V_\delta}\delta L$. Additionally, according to Lemma 13, $\hat{\mathbf{f}}_t$ is μ-strongly $(2\rho L + 2D\mu)\delta$-sub g-convex. Thus, the update rule in Algorithm 2 is exactly the expected gradient descent in Lemma 43 with parameters $S = (1-\tau)\mathcal{K}$, $G = \frac{S_\delta}{V_\delta}\delta L$, $\lambda_1 = (2\rho L + 2D\mu)\delta$, $\lambda_2 = 2\tau D^2$ and $\nu = \frac{BV_\delta}{S_\delta}$. We can get,

$$\begin{aligned}
\mathbb{E}\Big[\sum_{t=1}^{T}\mathbf{f}_t(y_t)\Big] - \min_{x\in(1-\tau)\mathcal{K}}\sum_{t=1}^{T}\mathbf{f}_t(x) &\leq \frac{1}{2\mu}\zeta(\kappa, D)(\frac{S_\delta}{V_\delta})^2\delta^2L^2\frac{BV_\delta}{S_\delta}(1+\log T) \\
&\quad + (2\rho L + 2D\mu)\delta T + 2\tau D^2 T \\
&\leq \frac{1}{2\mu}B\zeta(\kappa, D)\frac{S_\delta}{V_\delta}\delta^2L^2(1+\log T) \\
&\quad + (2\rho L + 2D\mu)\delta T + 2\tau D^2 T.
\end{aligned}$$

From the inequality $\frac{S_\delta}{V_\delta} \leq \frac{n}{\delta} + n|\kappa|\delta = B$ in Lemma 11, we have

$$\begin{aligned}\mathbb{E}\Big[\sum_{t=1}^{T} \mathbf{f}_t(y_t)\Big] - \min_{x\in(1-\tau)\mathcal{K}} \sum_{t=1}^{T} \mathbf{f}_t(x) &\leq \frac{1}{2\mu}(n + n|\kappa|\delta^2)^2\zeta(\kappa, D)L^2(1+\log T) \\ &\quad + (2\rho L + 2D\mu)\delta T + 2\tau D^2 T \\ &\leq \frac{n^2}{\mu}(1+\kappa^2\delta^4)\zeta(\kappa, D)L^2(1+\log T) \\ &\quad + (2\rho L + 2D\mu)\delta T + 2\tau D^2 T.\end{aligned}$$

Applying Lemma 44, we have that

$$\begin{aligned}\mathbb{E}[\mathsf{Reg}(T)] \leq &\frac{n^2}{\mu}(1+\kappa^2\delta^4)\zeta(\kappa, D)L^2(1+\log T) + 3\delta LT + \tau DLT \\ &+ (2\rho L + 2D\mu)\delta T + 2\tau D^2 T.\end{aligned}$$

Taking $\tau = \frac{\delta}{r\theta}$ and $\delta = \frac{1+\log T}{T}$, we get

$$\begin{aligned}\mathbb{E}[\mathsf{Reg}(T)] &\leq \frac{n^2L^2\zeta(\kappa, D)}{\mu}(1+\log T) + \frac{n^2C^2\kappa^2\zeta(\kappa, D)}{\mu}\frac{(1+\log T)^5}{T^4} \\ &\quad + (3L + \frac{DL}{r\theta} + \frac{2D^2}{r\theta} + 2\rho L + 2D\mu)(1+\log T) \\ &\leq \frac{3n^2C^2\kappa^2\zeta(\kappa, D)}{2\mu} + \Big(\frac{\zeta(\kappa, D)n^2L^2}{\mu} \\ &\qquad + 3L + \frac{DL + 2D^2}{r\theta} + 2\rho L + 2D\mu\Big)(1+\log T).\end{aligned}$$

The last inequality is due to $\max_{T\geq 1}\frac{(1+\log T)^5}{T^4} = \sqrt[3]{\frac{3125}{1024e}} \leq \frac{3}{2}$. Then we completes our proof. ■

Appendix G. Proof of Statements in Section 6

G.1 Proof of Lemma 21

Lemma 47 (Walter, 1974) *Set $\mathcal{K} \subset \mathcal{M}$, then for any $y \in \mathcal{M} \setminus \mathcal{K}$,*

$$\langle \exp^{-1}_{\mathscr{P}_{\mathcal{K}}(y)}(y), \exp^{-1}_{\mathscr{P}_{\mathcal{K}}(y)}(x)\rangle \leq 0, \quad \forall x \in \mathcal{K}.$$

Lemma 48 *Suppose $(\mathcal{M}, g)$ is a Riemannian manifold whose sectional curvature is bounded above by $K > 0$ and $x \in \mathcal{K}$. Then for any $y \in \mathcal{M}$, if there exists a geodesic γ lying in the geodesic ball $B_D(x)$ for some $D \leq \frac{\pi}{\sqrt{K}}$ that connects y and $\mathscr{P}_{\mathcal{K}}(y)$, then,*

$$d^2(x, \mathscr{P}_{\mathcal{K}}(y)) - d^2(x, y) \leq \sigma(\kappa, D)d^2(y, \mathscr{P}_{\mathcal{K}}(y)).$$

Proof Since $\gamma \subset B_D(x) \subset B_{\frac{\pi}{\sqrt{K}}}(x)$, we can apply Hessian comparison theorem (Theorem 33) to $\rho(p) = d(x, p)$, which implies

$$\nabla^2\rho(p) \succeq \sqrt{K}\cot(\rho(p)\sqrt{K}) \geq \sqrt{K}\cot(\sqrt{K}D)Id, \quad \forall p \in \gamma.$$

Then for $\rho^2(p) = d^2(x,p)$, we have

$$\nabla^2\rho^2(p) = \rho\nabla^2\rho + \nabla\rho\nabla\rho^T \succeq \begin{cases} 0, & \text{if } D \geq \frac{\pi}{2\sqrt{K}}, \\ \sqrt{K}D\cot(\sqrt{K}D)Id, & \text{if } \frac{\pi}{2\sqrt{K}} < D \leq \frac{\pi}{\sqrt{K}}, \end{cases} \quad \forall p \in \gamma,$$

which means that

$$\nabla^2\rho^2(p) \succeq -\sigma(K,D)Id, \quad \forall p \in \gamma.$$

By the mean value theorem, there exist a $q \in \gamma$ such that

$$\begin{aligned} \rho^2(y) - \rho^2(\mathscr{P}_{\mathcal{K}}(y)) &= \langle \nabla\rho^2(\mathscr{P}_{\mathcal{K}}(y)), \exp^{-1}_{\mathscr{P}_{\mathcal{K}}(y)}(y)\rangle + \nabla^2\rho(q)(\exp^{-1}_{\mathscr{P}_{\mathcal{K}}(y)}(y), \exp^{-1}_{\mathscr{P}_{\mathcal{K}}(y)}(y)) \\ &\geq -\langle \exp^{-1}_{\mathscr{P}_{\mathcal{K}}(y)}(x), \exp^{-1}_{\mathscr{P}_{\mathcal{K}}(y)}(y)\rangle - \sigma(\kappa,D)d^2(y,\mathscr{P}_{\mathcal{K}}(y)). \end{aligned} \tag{58}$$

Lemma (47) yields

$$\langle \exp^{-1}_{\mathscr{P}_{\mathcal{K}}(y)}(x), \exp^{-1}_{\mathscr{P}_{\mathcal{K}}(y)}(y)\rangle \leq 0.$$

Hence,

$$\rho^2(y) - \rho^2(\mathscr{P}_{\mathcal{K}}(y)) = d^2(x,y) - d^2(x,\mathscr{P}_{\mathcal{K}}(y)) \geq -\sigma(\kappa,D)d^2(y,\mathscr{P}_{\mathcal{K}}(y)),$$

which completes our proof. ■

We now carry out our proof of Lemma 21.

Proof Since $x_{t+1} = \mathscr{P}_{\mathcal{K}}(\tilde{x}_{t+1}) \in \mathcal{K}$, we have

$$d(x_{t+1}, x^*) \leq D < \frac{\pi}{2\sqrt{K}}. \tag{59}$$

Also, we have

$$d(\tilde{x}_{t+1}, x_{t+1}) = \|\alpha_t g_t\| \leq D < \frac{\pi}{2\sqrt{K}}, \tag{60}$$

which implies there is an unique geodesic γ connecting $\tilde{x}_{t+1}$ and x_{t+1} by Lemma 31. For any point $s \in \gamma$, we can find

$$\begin{aligned} d(x^*, s) &\leq d(x^*, x_{t+1}) + d(x_{t+1}, s) \\ &\leq d(x^*, x_{t+1}) + d(x_{t+1}, \tilde{x}_{t+1}) \\ &\leq 2D < \frac{D}{\sqrt{K}}. \end{aligned}$$

So γ is contained in the geodesic ball $B_{2D}(x^*) \subset B_{\frac{D}{\sqrt{K}}}(x^*)$, which satisfies the condition in Lemma 48. It give us

$$\frac{1}{2\alpha_t}\big(d^2(x_{t+1},x^*) - d^2(\tilde{x}_{t+1},x^*)\big) \leq \frac{1}{2\alpha_t}\big(\sigma(\kappa,2D)d^2(x_{t+1},\tilde{x}_{t+1})\big) \tag{61}$$

$$\leq \frac{1}{2\alpha_t}\big(\sigma(\kappa,2D)d^2(x_t,\tilde{x}_{t+1})\big) \tag{62}$$

$$\leq \sigma(\kappa,2D)\alpha_t\|g_t\|^2. \tag{63}$$

Summing t from 1 to T, we complete our proof. ■

G.2 Proof of Lemma 22

The second part is directly from the first part of lemma. So we focus on the first part of the lemma. By the proof of Lemma 13, we have

$$\hat{\mathbf{f}}(y)-\hat{\mathbf{f}}(x)-\langle\nabla\hat{\mathbf{f}}(x),\exp_x^{-1}(y)\rangle\geq\frac{1}{V_\delta}\Big(\int_{\mathcal{S}_\delta(x)}f(u)\Big(\langle V(u),\vec{n}(u)\rangle-\langle V(x),\vec{m}(u)\rangle\Big)\omega_{\mathcal{S}_\delta(x)}\Big)\\-\frac{1}{V_\delta}\Big(\int_{\mathcal{B}_\delta(x)}Div(V)f(u)\omega\Big).$$

Here we change $(*)$ with the claim $(**)$

$$\langle V(u),\vec{n}(u)\rangle-\langle V(x),\vec{m}(u)\rangle-\frac{1}{2}\pi^2\iota\delta\leq 0,\quad\forall u\in\mathcal{S}_\delta(x).\tag{$**$}$$

This claim requires many geometric details that deviates our attention from the proof, and we will prove it afterwards.

If the claim $(**)$ holds, we have

$$\int_{\mathcal{S}_\delta(x)}f(u)\Big(\langle V(u),\vec{n}(u)\rangle-\langle V(x),\vec{m}(u)\rangle\Big)\omega_{\mathcal{S}_\delta(x)}\\\geq\int_{\mathcal{S}_\delta(x)}f(u)\Big(\langle V(u),\vec{n}(u)\rangle-\langle V(x),\vec{m}(u)\rangle-\frac{1}{2}\pi^2\iota\delta\Big)+\big(\mathbf{f}(u)\frac{1}{2}\pi^2\iota\delta\big)\omega_{\mathcal{S}_\delta(x)}.\tag{64}$$

Then with the g-L-Lipschitz of $\mathbf{f}$ and the condition $\mathbf{f}(x)=0$, we have $|\mathbf{f}(u)|\leq\delta L$, thus

$$\int_{\mathcal{S}_\delta(x)}f(u)\Big(\langle V(u),\vec{n}(u)\rangle-\langle V(x),\vec{m}(u)\rangle\Big)\omega_{\mathcal{S}_\delta(x)}\tag{65}$$

$$\geq\int_{\mathcal{S}_\delta(x)}\delta L\Big(\langle V(u),\vec{n}(u)\rangle-\langle V(x),\vec{m}(u)\rangle-\frac{1}{2}\pi^2\iota\delta\Big)-\big(\delta L\frac{1}{2}\pi^2\iota\delta\big)\omega_{\mathcal{S}_\delta(x)}\tag{66}$$

$$=\int_{\mathcal{S}_\delta(x)}\delta L\langle V(u),\vec{n}(u)\rangle\omega_{\mathcal{S}_\delta(x)}-\int_{\mathcal{S}_\delta(x)}\delta L\langle V(x),\vec{m}(u)\rangle\omega_{\mathcal{S}_\delta(x)}-\pi^2\iota\delta^2LS_\delta.\tag{67}$$

Continuing to analyze the first two terms with the method in the proof of Lemma 13, we have

$$\begin{aligned}\hat{\mathbf{f}}(y)-\hat{\mathbf{f}}(x)-\langle\nabla\hat{\mathbf{f}}(x),\exp_x^{-1}(y)\rangle&\geq-\frac{1}{V_\delta}\Big(\int_{\mathcal{B}_\delta(x)}Div(V)(\mathbf{f}(x)+\delta L)\omega\Big)-\frac{S_\delta}{V_\delta}\pi^2\iota L\delta^2\\&\geq-2\delta L\sup_{u\in\mathcal{B}_\delta(x)}|Div(V(u))|-(\frac{n}{\delta}+n|\kappa'|\delta)\pi^2\iota L\delta^2\\&=-2\delta L\sup_{u\in\mathcal{B}_\delta(x)}|Div(V(u))|-(n+n|\kappa'|\delta^2)\pi^2\iota L\delta.\end{aligned}$$

Again, setting

$$\rho=\sup_{(x,y,u)\in\bar{\mathcal{K}}\times\bar{\mathcal{K}}\times\bar{\mathcal{K}}}|Div(V(u))|<\infty,$$

we establish the $\big(2\rho\delta L+(n+n|\kappa'|\delta^2)\pi^2\iota L\delta\big)$-sub g-convexity.

For proving (ii), It is sufficient to show

$$\frac{1}{V_\delta}\int_{\mathcal{B}_\delta(x)}\frac{\mu}{2}d^2(u,\phi(u))\omega \geq \frac{\mu}{2}d^2(x,y) - 4\delta D,$$

which is obvious from the fact $d^2(u,\phi(u)) - d^2(x,y) = (d(u,\phi(u)) + d(x,y))(d(u,\phi(u)) - d(x,y)) \leq 2D \cdot 2\delta = 4D\delta$. ■

G.3 Proof of the Claim $(**)$

We use the symbol as same as those in the proof of the Claim $(*)$. Note that the rectangle map Γ is defined by

$$\begin{aligned}\Gamma_u : [0,1]\times[0,\delta] &\to \mathcal{M}\\ (t,s) &\to \exp_{\xi_u(s)}(tV(\xi_u(s))).\end{aligned}$$

and the length of the s-curve $l_u(t)$ is described by

$$l_u(t) = \int_0^\delta \sqrt{\langle S(t,s), S(t,s)\rangle}ds.$$

We have

$$l_u'(0) = \langle V(u), \vec{n}(u)\rangle - \langle V(x), \vec{m}(u)\rangle,$$

and

$$l_u''(t) \geq \int_0^\delta \frac{1}{\|S\|} - R(T,S,S,T)ds.$$

The following analysis is quite different because the section curvature of $\mathcal{M}$ is no longer non positive, but bounded by $K > 0$. As a result, we have

$$\begin{aligned}l_u''(t) &\geq \int_0^\delta \frac{1}{\|S\|} - R(T,S,S,T)ds\\ &\geq \int_0^\delta \frac{1}{\|S\|} - K\|T\|^2\|S\|^2 ds\\ &= \int_0^\delta -K\|T\|^2\|S\|dS\\ &\geq -KD^2\int_0^\delta \|S\|dS. \end{aligned} \tag{68}$$

Then we try to estimate the norm $\|S(t,s)\|$ in $[0,1]\times[0,\delta]$. We separate the Jacobi field

$$S(t,s) = S_0(t,s) + S_1(t,s).$$

over the geodesic $\gamma_s(t)$ for every $t \in [0,1]$, where $S_i(t,s)$ is also a Jacobi field with condition $S_i(i,s) = S(i,s)$ and $S_i(1-i,s) = 0$. We estimate $\|S_i(t,s)\|$ with Theorem 30. W.l.o.g., we set $i = 0$.

Since all sectional curvature of $\mathcal{M}$ is bounded below by κ', we have

$$\|S_0(t,s)\| \leq \mathbf{s}(\kappa', t\|T(t,s)\|)\|\dot{S}_0(0,s)\|. \tag{69}$$

Also, since all sectional curvature of $\mathcal{M}$ is bounded above by K, we have

$$1 = \|S_0(t,s)\| \geq \mathbf{s}(K, t\|T(t,s)\|)\|\dot{S}_0(0,s)\|. \tag{70}$$

Putting (69) and (70) together, we have

$$\|S_0(t,s)\| \leq \frac{\mathbf{s}(\kappa', t\|T(t,s)\|)}{\mathbf{s}(K, t\|T(t,s)\|)}.$$

With the condition $t\|T(t,s)\| \leq \|T(t,s)\| \leq D$, we have

$$\|S_0(t,s)\| \leq \frac{\mathbf{s}(\kappa', D)}{\mathbf{s}(K, D)} = \frac{1}{2}\iota.$$

Then we finally get

$$\|S(t,s)\| = \|S_0(t,s)\| + \|S_1(t,s)\| \leq \frac{1}{2}\iota + \frac{1}{2}\iota = \iota. \tag{71}$$

Putting (71) into (68), we have

$$\begin{aligned} l_u''(t) &\geq -KD^2 \int_0^\delta \|S\| dS && (72)\\ &\geq -KD^2\iota\delta && (73)\\ &\geq -\pi^2\iota\delta. && (74) \end{aligned}$$

Hence, by the mean value theorem there exists a $t \in (0,1)$ such that

$$\begin{aligned} l_u(1) &= l_u(0) + l_u'(0) + \frac{1}{2}l_u''(t) && (75)\\ &\geq l_u(0) + l_u'(0) - \frac{1}{2}\pi^2\iota\delta. && (76) \end{aligned}$$

With the equality (36), we have

$$l_u'(0) \leq \frac{1}{2}\pi^2\iota\delta, \tag{77}$$

which complete our proof of claim $(**)$. ■

G.4 Proof of Theorem 24

Lemma 49 *Suppose that $\mathcal{S}$ is subset of a unique g-convex set $\mathcal{K} \subseteq \mathcal{M}$ with the diameter $D \leq \frac{\pi}{2\sqrt{K}}$, and $\{\mathbf{f}_t\}_{t=1,2,\ldots,T}$ is a series of λ_1-sub μ-strongly g-convex functions. Let the sequence be generated by*

$$x_{t+1} = \mathscr{H}(\mathscr{P}_{\mathcal{K}}(\exp_{x_t}(-\alpha_t g_t))), \tag{78}$$

where the random vector g_t satisfies that $\mathbb{E}[\mathbf{g}_t|x_t] = \nabla \mathbf{f}_t(x_t)$, $\|\mathbf{g}_t\| \leq G$ and the operator $\mathscr{H}$ satisfies (37)*. For constants $\nu \geq 1$ and $c_0 \geq \frac{\nu G}{\mu D}$, if the step size $\alpha_t = \frac{\nu}{\mu(t+c_0)}$, then*

$$\mathbb{E}[\sum_{i=1}^{T} \mathbf{f}_t(x_t)] - \min_{x\in\mathcal{S}} \sum_{i=1}^{T} \mathbf{f}_t(x) \leq \frac{D^2\mu c_0}{2} + \lambda_1 T + \lambda_2 T + \frac{1}{2}(\zeta(\kappa, D) + \sigma(K, 2D))\nu G^2(1 + \log(T + c_0))).$$

Proof By Lemma 43, we have

$$\mathbb{E}[\mathbf{f}_t(x_t) - \mathbf{f}_t(x^*)] \leq \mathbb{E}\Big[\langle g_t, \exp_{x_t}^{-1}(x^*)\rangle - \frac{\mu}{2} d^2(x_t, x^*)\Big] + \lambda_1.$$

Denote $\tilde{x}_{t+1} = \exp_{x_t}(-\alpha_t g_t)$. Recalling Lemma 35 in the geodesic triangle $\triangle x_t \tilde{x}_{t+1} x^*$ gives that

$$\langle -\alpha_t g_t, \exp_{x_t}^{-1}(x^*)\rangle \leq \frac{1}{2}(d^2(x_t, x^*) - d^2(\tilde{x}_{t+1}, x^*)) + \frac{1}{2}\zeta(\kappa, d(x_t, x))\|\alpha_t g_t\|^2. \tag{79}$$

Combining (11) and (12), we get

$$\begin{aligned}
\mathbb{E}[\mathbf{f}_t(x_t) - \mathbf{f}_t(x^*)] &\leq \frac{1}{2\alpha_t}(d^2(x_t, x^*) - d^2(\tilde{x}_{t+1}, x^*)) + \frac{1}{2}\zeta\big(\kappa, d(x_t, x^*)\big)\alpha_t\|g_t\|^2 \\
&\quad - \frac{\mu}{2} d^2(x_t, x^*) + \lambda_1 \\
&\leq \frac{1}{2\alpha_t}(d^2(x_t, x^*) - d^2(\mathscr{P}_{\mathcal{K}}(\tilde{x}_{t+1}), x^*)) \\
&\quad + \frac{1}{2\alpha_t}(d^2(\mathscr{P}_{\mathcal{K}}(\tilde{x}_{t+1}), x^*) - d^2(\tilde{x}_{t+1}, x^*)) \\
&\quad + \frac{1}{2}\zeta\big(\kappa, d(x_t, x^*)\big)\alpha_t G^2 - \frac{\mu}{2} d^2(x_t, x^*) + \lambda_1 \\
&\leq \frac{1}{2\alpha_t}(d^2(x_t, x^*) - d^2(x_{t+1}, x^*)) + \frac{1}{2\alpha_t}(d^2((\mathscr{P}_{\mathcal{K}}(\tilde{x}_{t+1}), x^*) - d^2(\tilde{x}_{t+1}, x^*)) \\
&\quad + \frac{1}{2}\zeta\big(\kappa, d(x_t, x^*)\big)\alpha_t G^2 - \frac{\mu}{2} d^2(x_t, x^*) + \lambda_1 + \lambda_2.
\end{aligned} \tag{80}$$

Summing (80) from 1 to T, we obtain

$$\begin{aligned}
\mathbb{E}[\sum_{i=1}^{T} \mathbf{f}_t(x_t)] - \min_{x\in\mathcal{S}} \sum_{i=1}^{T} \mathbf{f}_t(x) &\leq \sum_{t=1}^{T}(d^2(x_t, x^*)\frac{1}{2\alpha_t} - \frac{1}{2\alpha_{t-1}} - \frac{\mu}{2}) + \sum_{t=1}^{T}\frac{1}{2}\zeta\big(\kappa, d(x_t, x^*)\big)G^2\alpha_t + \lambda_1 T + \lambda_2 T \\
&\quad + \sum_{t=1}^{T}\frac{1}{2\alpha_t}(d^2(\mathscr{P}_{\mathcal{K}}(\tilde{x}_{t+1}), x^*) - d^2(\tilde{x}_{t+1}, x^*)) \\
&\leq \sum_{t=1}^{T} d^2(x_t, x^*)(\frac{1}{2\alpha_t} - \frac{1}{2\alpha_{t-1}} - \frac{\mu}{2}) + \frac{1}{2}\zeta(\kappa, D)G^2\sum_{t=1}^{T}\alpha_t \\
&\quad + \sum_{t=1}^{T}\frac{1}{2\alpha_t}(d^2(\mathscr{P}_{\mathcal{K}}(\tilde{x}_{t+1}), x^*) - d^2(\tilde{x}_{t+1}, x^*)) + \lambda_1 T + \lambda_2 T.
\end{aligned}$$

By Lemma 21, we have

$$\sum_{t=1}^{T}\frac{1}{2\alpha_t}(d^2(\mathscr{P}_{\mathcal{K}}(\tilde{x}_{t+1}),x^*)-d^2(\tilde{x}_{t+1},x^*))\leq\sigma(K,2D)\sum_{t=1}^{T}\frac{1}{2}\alpha_t\|\mathbf{g}_t\|^2$$
$$\leq\frac{1}{2}\sigma(K,2D)G^2\sum_{t=1}^{T}\alpha_t$$

Thus

$$\mathbb{E}[\sum_{i=1}^{T}\mathbf{f}_t(x_t)]-\min_{x\in\mathcal{S}}\sum_{i=1}^{T}\mathbf{f}_t(x)\leq\sum_{t=1}^{T}(d^2(x_t,x^*)(\frac{1}{2\alpha_t}-\frac{1}{2\alpha_{t-1}}-\frac{\mu}{2})$$
$$+\frac{1}{2\mu}(\zeta(\kappa,D)+\sigma(K,2D))G^2\sum_{t=1}^{T}\alpha_t+\lambda_1T+\lambda_2T.$$

Taking $\alpha_t=\frac{\nu}{\mu(t+c_0)}$, we have

$$\mathbb{E}[\sum_{i=1}^{T}\mathbf{f}_t(x_t)]-\min_{x\in\mathcal{S}}\sum_{i=1}^{T}\mathbf{f}_t(x)\leq d^2(x_1,x^*)(\frac{\mu(c_0+1)}{2\nu}-\frac{\mu}{2})$$
$$+\frac{1}{2}(\zeta(\kappa,D)+\sigma(K,2D))\nu G^2\big(1+\log(T+c_0)\big)+\lambda_1T+\lambda_2T$$
$$\leq\frac{D^2\mu c_0}{2}+\lambda_1T+\lambda_2T$$
$$+\frac{1}{2\mu}(\zeta(\kappa,D)+\sigma(K,2D))\nu G^2\big(1+\log(T+c_0)\big),$$

which proves Lemma 49. ■

We are now ready to prove Theorem 24.

Proof (i) R-OGD algorithm: The update rule in the R-OGD algorithm is actually (78) with $\mathcal{S}=\mathcal{K}$, $G=L$, $\nu=1$ and $\lambda_1=\lambda_2=0$. By Lemma 49, the regret bound of the R-OGD is

$$\mathsf{Reg}(T)=\mathbb{E}[\sum_{i=1}^{T}\mathbf{f}_t(x_t)]-\min_{x\in\mathcal{K}}\sum_{i=1}^{T}\mathbf{f}_t(x)$$
$$\leq\frac{D^2\mu c_0}{2}+\frac{1}{2}(\zeta(\kappa,D)+\sigma(K,2D))G^2\big(1+\log(T+c_0)\big).$$

which completes the proof of the R-OGD algorithm.

(ii) R-BAN algorithm: It follows from Lemma 44 that

$$\mathbb{E}[\mathsf{Reg}(T)]\leq\mathbb{E}\Big[\sum_{t=1}^{T}(\hat{\mathbf{f}}_t(y_t)-\hat{\mathbf{f}}_t(x_\tau^*))\Big]+3\delta LT+\tau DLT.\tag{81}$$

As shown in the proof of Theorem 14, the update rule of y_t is actually (78) with parameter $\mathcal{S} = (1-\tau)\mathcal{K}$, $\nu = \frac{BV_\delta}{\mathcal{S}_\delta}$,$G = \frac{S_\delta}{V_\delta}C$, $\lambda_1 = \big(2\rho\delta L + 2D\mu\delta + (n+n|\kappa'|\delta^2)\pi^2\iota L\delta\big)$, and $\lambda_2 = 2\tau D^2$. By Lemma 49,

$$\begin{aligned}
\mathbb{E}\Big[\sum_{t=1}^{T}(\hat{\mathbf{f}}_t(y_t) - \hat{\mathbf{f}}_t(x_\tau^*))\Big] &\le \frac{D^2\mu c_0}{2} + \frac{1}{2\mu}(\zeta(\kappa, D) + \sigma(K, 2D))\frac{BV_\delta}{\mathcal{S}_\delta}(\frac{S_\delta}{V_\delta})^2C^2(1+\log(T+c_0)) \\
&\quad + 2\delta\rho LT + 2D\mu\delta T + 2\tau D^2T + (n+n|\kappa'|\delta^2)\pi^2\iota L\delta T \\
&\le \frac{D^2\mu c_0}{2} + \frac{1}{2\mu}(\zeta(\kappa, D) + \sigma(K, 2D))B^2C^2(1+\log(T+c_0)) \\
&\quad + 2\delta\rho LT + 2D\mu\delta T + 2\tau D^2T + (n+n|\kappa'|\delta^2)\pi^2\iota L\delta T \\
&\le \frac{D^2\mu c_0}{2} + 2\tau D^2T + 2\delta\rho LT + 2D\mu\delta T + (n+n|\kappa'|\delta^2)\pi^2\iota L\delta T \\
&\quad + \frac{1}{\mu}(\zeta(\kappa, D) + \sigma(K, 2D))(\frac{n^2}{\delta^2} + n^2\kappa'^2\delta^2)C^2(1+\log(T+c_0)).
\end{aligned} \tag{82}$$

Combining (81) and (82), we have

$$\begin{aligned}
\mathbb{E}[\mathsf{Reg}(T)] \le \frac{D^2\mu c_0}{2} &+ \frac{1}{\mu}(\zeta(\kappa, D) + \sigma(K, 2D))(\frac{n^2}{\delta^2} + n^2\kappa'^2\delta^2)C^2(1+\log(T+c_0)) \\
&+ 2\delta\rho LT + 2D\mu\delta T + (n+n|\kappa'|\delta^2)\pi^2\iota L\delta T + 3\delta LT + \tau DLT + 2\tau D^2T.
\end{aligned}$$

Taking $\tau = \frac{\delta}{r\theta}$ and $\delta = \sqrt[3]{\frac{nC(1+\log(T+c_0))}{T}}$ in (49), we get

$$\begin{aligned}
\mathbb{E}[\mathsf{Reg}(T)] &\le \frac{D^2\mu c_0}{2} + \frac{1}{\mu}(\zeta(\kappa, D) + \sigma(K, 2D))n^{\frac{4}{3}}C^{\frac{4}{3}}(1+\log(T+c_0))^{\frac{1}{3}}T^{\frac{2}{3}} \\
&\quad + \frac{1}{\mu}(\zeta(\kappa, D) + \sigma(K, 2D))C^2n^2\kappa'^2\delta^2(1+\log(T+c_0)) \\
&\quad + \big(2\rho L + 3L + \frac{DL}{r\theta} + \frac{2D^2}{r\theta} + 2D\mu\big)n^{\frac{1}{3}}C^{\frac{1}{3}}(1+\log(T+c_0))^{\frac{1}{3}}T^{\frac{2}{3}} \\
&\quad + (n\pi^2\iota L + n|\kappa'|\delta^2\pi^2\iota L)n^{\frac{1}{3}}C^{\frac{1}{3}}(1+\log(T+c_0))^{\frac{1}{3}}T^{\frac{2}{3}} \\
&\le \frac{D^2\mu c_0}{2} + \frac{1}{\mu}(\zeta(\kappa, D) + \sigma(K, 2D))n^{\frac{4}{3}}C^{\frac{4}{3}}(1+\log(T+c_0))^{\frac{1}{3}}T^{\frac{2}{3}} \\
&\quad + \frac{1}{\mu}(\zeta(\kappa, D) + \sigma(K, 2D))C^{\frac{8}{3}}n^{\frac{8}{3}}\kappa'^2D(1+\log(T+c_0))^{\frac{4}{3}}T^{-\frac{1}{3}} \\
&\quad + \big(2\rho L + 3L + \frac{DL}{r\theta} + \frac{2D^2}{r\theta} + 2D\mu\big)n^{\frac{1}{3}}C^{\frac{1}{3}}(1+\log(T+c_0))^{\frac{1}{3}}T^{\frac{2}{3}} \\
&\quad + (n\pi^2\iota L + n|\kappa'|\delta^2\pi^2\iota L)n^{\frac{1}{3}}C^{\frac{1}{3}}(1+\log(T+c_0))^{\frac{1}{3}}T^{\frac{2}{3}}.
\end{aligned}$$

Then, we have,

$$\begin{aligned}\mathbb{E}[\mathsf{Reg}(T)] \leq & \frac{D^2\mu c_0}{2} + \frac{4(c_0+1)}{\mu}(\zeta(\kappa,D)+\sigma(K,2D))C^{\frac{8}{3}}n^{\frac{8}{3}}\kappa'^2 D \\ & + \frac{1}{\mu}(\zeta(\kappa,D)+\sigma(K,2D))n^{\frac{4}{3}}C^{\frac{4}{3}}(1+\log(T+c_0))^{\frac{1}{3}}T^{\frac{2}{3}} \\ & + \big(2\rho L + 3L + \frac{DL}{r\theta} + \frac{2D^2}{r\theta} + 2D\mu\big)n^{\frac{1}{3}}C^{\frac{1}{3}}(1+\log(T+c_0))^{\frac{1}{3}}T^{\frac{2}{3}} \\ & + (n^{\frac{4}{3}}C^{\frac{1}{3}}\pi^2\iota L + 2n^{\frac{4}{3}}C^{\frac{4}{3}}|\kappa'|D^2\pi^2\iota)(1+\log(T+c_0))^{\frac{1}{3}}T^{\frac{2}{3}}.\end{aligned}$$

Finally, we have

$$\begin{aligned}\mathbb{E}[\mathsf{Reg}(T)] = & \frac{D^2\mu c_0}{2} + \frac{4(c_0+1)}{\mu}(\zeta(\kappa,D)+\sigma(K,2D))C^{\frac{8}{3}}n^{\frac{8}{3}}\kappa'^2 D \\ & + \Bigg(\big(\frac{1}{\mu}(\zeta(\kappa,D)+\sigma(K,2D)) + |\kappa'|D^2\pi^2\iota\big)n^{\frac{4}{3}}C^{\frac{4}{3}} + \\ & + \big(n\pi^2\iota + 2\rho L + 3L + \frac{DL}{r\theta} + \frac{2D^2}{r\theta} + 2D\mu\big)n^{\frac{1}{3}}C^{\frac{1}{3}}L\Bigg)(1+\log(T+c_0))^{\frac{1}{3}}T^{\frac{2}{3}}.\end{aligned}$$

The last inequality is due to $\delta < D$, $\delta L < 2C$ and

$$\max_{T\geq 1}\sqrt[3]{\frac{(1+\log(T+c_0))^4}{T}} = \max_{T\geq 1}\sqrt[3]{\frac{(1+\log(T+c_0))^4}{T+c_0}}\frac{T+c_0}{T} \leq 4(c_0+1).$$

Then we completes the proof of the R-BAN algorithm.

(iii) R-2-BAN algorithm: Notice that the update rule of y_t is (78) with parameter $\mathcal{S} = (1-\tau)\mathcal{K}$, $\nu = \frac{BV_\delta}{\mathcal{S}_\delta}$, $G = \frac{S_\delta}{V_\delta}\delta L$, $\lambda_1 = \big(2\rho\delta L + 2D\mu\delta + (n+n|\kappa'|\delta^2)\pi^2\iota L\delta\big)$, and $\lambda_2 = 2\tau D^2$. Hence,

$$\begin{aligned}\mathbb{E}[\mathsf{Reg}(T)] &\leq \mathbb{E}\Big[\sum_{t=1}^{T}(\hat{\mathbf{f}}_t(y_t) - \hat{\mathbf{f}}_t(x_\tau^*))\Big] + 3\delta LT + \tau DLT \\ &\leq \frac{D^2\mu c_0}{2} + \frac{1}{2\mu}\big(\zeta(\kappa,D)+\sigma(K,2D)\big)B^2\delta^2L^2(1+\log(T+c_0)) \\ &\quad + 3\delta LT + \tau DLT + 2\delta\rho LT + 2D\mu\delta T + 2\tau D^2T + ((n+n|\kappa'|\delta^2)\pi^2\iota L\delta)T, \\ &\leq \frac{D^2\mu c_0}{2} + \frac{1}{\mu}\big(\zeta(\kappa,D)+\sigma(K,2D)\big)(n^2+n^2\kappa'^2\delta^4)L^2(1+\log(T+c_0)) \\ &\quad + 3\delta LT + \tau DLT + 2\delta\rho LT + 2D\mu\delta T + 2\tau D^2T + ((n+n|\kappa'|\delta^2)\pi^2\iota L\delta)T.\end{aligned}$$

Taking $\tau = \frac{\delta}{r\theta}$ and $\delta = \frac{1+\log(T+c_0)}{T}$, we get

$$\begin{aligned}
\mathbb{E}[\mathsf{Reg}(T)] \leq{}& \frac{D^2\mu c_0}{2} + n|\kappa'|\pi^2\iota L\frac{(1+\log(T+c_0))^3}{T^2} \\
&+ \big(\frac{1}{\mu}(\zeta(\kappa,D)+\sigma(K,2D)\big)n^2{\kappa'}^2L^2\frac{(1+\log(T+c_0))^5}{T^4} \\
&+ \Big(\frac{1}{\mu}(\zeta(\kappa,D)+\sigma(K,2D))n^2L^2 \\
&\qquad + 3L + \frac{2D^2+DL}{r\theta} + 2\rho L + 2D\mu\Big)(1+\log(T+c_0)) \\
\leq{}& \frac{D^2\mu c_0}{2} + \big(\frac{1}{\mu}(\zeta(\kappa,D)+\sigma(K,2D)\big)\frac{3(c_0+1)n^2{\kappa'}^2L^2}{2} + \frac{3(c_0+1)n|\kappa'|\pi^2\iota L}{2} \\
&+ \Big(\frac{1}{\mu}(\zeta(\kappa,D)+\sigma(K,2D))n^2L^2 \\
&\qquad + 3L + \frac{DL+2D^2}{r\theta} + 2\rho L + 2D\mu\Big)(1+\log(T+c_0)). \qquad (83)
\end{aligned}$$

The last inequality is due to $\max_{T\leq 1}\frac{(1+\log(T+c_0))^5}{T^4}$ and $\max_{T\leq 1}\frac{(1+\log(T+c_0))^3}{T^2}$ are less than $\frac{3}{2}(c_0+1)$. Then we complete our proof. ■

References

Jacob Abernethy, Peter Bartlett, Alexander Rakhlin, and Ambuj Tewari. Optimal strategies and minimax lower bounds for online convex game. In *Proceedings of the 21st Annual Conference on Learning Theory*, pages 415–423, 2008.

Pierre-Antoine Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2009.

Alekh Agarwal, Ofer Dekel, and Lin Xiao. Optimal algorithms for online convex optimization with multi-point bandit feedback. In *Proceedings of the 23rd Annual Conference on Learning Theory*, pages 28–40, 2010.

Shmuel Agmon. The relaxation method for linear inequalities. *Canadian Journal of Mathematics*, 6:382–392, 1954.

Kwangjun Ahn and Suvrit Sra. From Nesterov's estimate sequence to Riemannian acceleration. In *Proceedings of the Annual 33rd Conference on Learning Theory*, pages 84–118, 2020.

Foivos Alimisis, Antonio Orvieto, Gary Bécigneul, and Aurelien Lucchi. Momentum improves optimization on Riemannian manifolds. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistic*, pages 1351–1359, 2021.

Zeyuan Allen-Zhu, Ankit Garg, Yuanzhi Li, Rafael Oliveira, and Avi Wigderson. Operator scaling via geodesically convex optimization, invariant theory and polynomial identity

testing. In *Proceedings of the 50th Annual ACM Symposium on Theory of Computing*, pages 172–181, 2018.

Kimon Antonakopoulos, Elena V. Belmega, and Panayotis Mertikopoulos. Online and stochastic optimization beyond Lipschitz continuity: a Riemannian approach. In *Proceedings of the 8th International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=rkxZyaNtwB`.

Yuichiro Anzai. *Pattern Recognition and Machine Learning*. Elsevier, 2012.

Sébastien Arnold, Pierre-Antoine Manzagol, Reza B. Harikandeh, Ioannis Mitliagkas, and Nicolas Le Roux. Reducing the variance in online optimization by transporting past gradients. In *Proceedings of the 32nd Advances in Neural Information Processing Systems*, pages 5392–5403, 2019.

Miroslav Bacák. *Convex Analysis and Optimization in Hadamard Spaces*. Walter de Gruyter GmbH & Co KG, 2014.

Werner Ballmann. *Lectures on Spaces of Nonpositive Curvature*. Birkhäuser, 2012.

Gary Becigneul and Octavian-Eugen Ganea. Riemannian adaptive optimization methods. In *Proceedings of the 7th International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=r1eiqi09K7`.

Valerii Berestovskii and Yurii Nikonorov. *Riemannian Manifolds and Homogeneous Geodesics*. Springer, 2020.

Marcel Berger. *Geometry I*. Springer Science & Business Media, 2009.

Silvere Bonnabel. Stochastic gradient descent on Riemannian manifolds. *IEEE Transactions on Automatic Control*, 58(9):2217–2229, 2013.

Nicolas Boumal and Pierre-Antoine Absil. Low-rank matrix completion via preconditioned optimization on the Grassmann manifold. *Linear Algebra and its Applications*, 475:200–239, 2015.

Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

Daniel Brooks, Olivier Schwander, Frederic Barbaresco, Jean-Yves Schneider, and Matthieu Cord. Riemannian batch normalization for SPD neural networks. In *Proceedings of the 32nd Advances in Neural Information Processing Systems*, 2019.

Manfredo Perdigao do Carmo. *Riemannian Geometry*. Birkhäuser, 1992.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

Guang Cheng, Hesamoddin Salehian, and Baba C. Vemuri. Efficient recursive algorithms for computing the mean diffusion tensor and applications to DTI segmentation. In *Proceedings of the 12th European Conference on Computer Vision*, pages 390–401, 2012.

Shiing-Shen Chern, Weihuan Chen, and Kai Shue Lam. *Lectures on Differential Geometry*. World Scientific, 1999.

Charlan Dellon da Silva Alves, Paulo Roberto Oliveira, and Ronaldo Malheiros Gregório. l_α Riemannian weighted centers of mass applied to compose an image filter to diffusion tensor imaging. *Applied Mathematics and Computation*, 390:125603, 2021.

Jiashi Feng, Huan Xu, and Shuicheng Yan. Online robust PCA via stochastic optimization. In *Proceedings of the 27th Advances in Neural Information Processing Systems*, pages 404–412, 2013.

Abraham D. Flaxman, Adam Tauman Kalai, and Brendan McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proceedings of the 16th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 385–394, 2005.

Simone Fiori. Manifold calculus in system theory and control—fundamentals and first-order systems. *Symmetry*, 13(11):2092, 2021.

Simone Fiori. Quasi-geodesic neural learning algorithms over the orthogonal group: a tutorial. *Journal of Machine Learning Research*, 6(26):743–781, 2005.

Ankit Garg, Leonid Gurvits, Rafael Oliveira, and Avi Wigderson. Algorithmic and optimization aspects of Brascamp-Lieb inequalities, via operator scaling. *Geometric and Functional Analysis*, 28(1):100–145, 2018.

Mohammad Ghomi and Joel Spruck. Total curvature and the isoperimetric inequality in cartan-hadamard manifolds. *The Journal of Geometric Analysis*, 32(2):50, 2022.

Elad Hazan. *Introduction to online convex optimization*. MIT Press, 2022.

Elad Hazan, Adam Kalai, Satyen Kale, and Amit Agarwal. Logarithmic regret algorithms for online convex optimization. In *Proceedings of the 19th Annual Conference on Learning Theory*, pages 499–513, 2006.

Jiang Hu, Xin Liu, Zai-Wen Wen, and Ya-Xiang Yuan. A brief introduction to manifold optimization. *Journal of the Operations Research Society of China*, 8(2):199–248, 2020.

Wen Huang, Pierre-Antoine Absil, and Kyle A. Gallivan. A Riemannian symmetric rank-one trust-region method. *Mathematical Programming*, 150(2):179–216, 2015.

Sergey K. Ivanov, Anatoly M. Kamchatnov, Thibault Congy, and Nicolas Pavloff. Solution of the Riemann problem for polarization waves in a two-component Bose-Einstein condensate. *Physical Review E*, 96(6):062202, 2017.

Alkis Koudounas and Simone Fiori. Gradient-based learning methods extended to smooth manifolds applied to automated clustering. *Journal of Artificial Intelligence Research*, 68:777–816, 2020.

Jerome Lapuyade-Lahorgue and Frederic Barbaresco. Radar detection using Siegel distance between autoregressive processes, application to HF and X-band radar. In *Proceedings of 2008 IEEE Radar Conference*, pages 1–6, 2008.

John M. Lee. *Introduction to Riemannian Manifolds*. Springer, 2018.

Kuang-Chih Lee and David Kriegman. Online learning of probabilistic appearance manifolds for video-based recognition and tracking. In *Proceedings of 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 852–859, 2005.

Jinlong Lei, Peng Yi, Yiguang Hong, Jie Chen, and Guodong Shi. Online convex optimization over Erdos-Renyi random networks. In *Proceedings of the 33rd Advances in Neural Information Processing Systems*, pages 15591–15601, 2020.

Alejandro I. Maass, Chris Manzie, Dragan Nesic, Jonathan H. Manton, and Iman Shames. Tracking and regret bounds for online zeroth-order Euclidean and Riemannian optimization. *SIAM Journal on Optimization*, 32(2):445–469, 2022.

Jiazhong Nie, Wojciech Kotłowski, and Manfred K. Warmuth. Online PCA with optimal regret. *Journal of Machine Learning Research*, 17(173):1–49, 2016.

Katsumi Nomizu. On local and global existence of Killing vector fields. *Annals of Mathematics*, 72(1):105–120, 1960.

Yogesh Rathi, Allen Tannenbaum, and Oleg Michailovich. Segmenting images on the tensor manifold. In *Proceedings of 2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

Wolfgang Ring and Benedikt Wirth. Optimization methods on Riemannian manifolds and their application to shape space. *SIAM Journal on Optimization*, 22(2):596–627, 2012.

Niranjan Srinivas, Andreas Krause, Sham Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: no regret and experimental design. In *Proceedings of the 27th International Conference on Machine Learning*, page 1015–1022, 2010.

Hao Tang, Zhiao Huang, Jiayuan Gu, Bao-Liang Lu, and Hao Su. Towards scale-invariant graph-related problem solving by iterative homogeneous GNNs. In *Proceedings of the 33rd Advances in Neural Information Processing Systems*, pages 15811–15822, 2020.

James Townsend, Niklas Koep, and Sebastian Weichwald. Pymanopt: a python toolbox for optimization on manifolds using automatic differentiation. *Journal of Machine Learning Research*, 17(137):1–5, 2016.

Quinten Tupker, Salem Said, and Cyrus Mostajeran. Online learning of Riemannian hidden Markov models in homogeneous hadamard spaces. In *Proceedings of the 5th International Conference on Geometric Science of Information*, pages 37–44, 2021.

Rolf Walter. On the metric projection onto convex sets in Riemannian spaces. *Archiv der Mathematik*, 25(1):91–98, 1974.

Xi Wang, Zhipeng Tu, Yiguang Hong, Yingyi Wu, and Guodong Shi. No-regret online learning over Riemannian manifolds. In *Proceedings of the 34th Advances in Neural Information Processing Systems*, pages 28323–28335, 2021.

Shing-Tung Yau. Non-existence of continuous convex functions on certain Riemannian manifolds. *Mathematische Annalen*, 207(4):269–270, 1974.

Hongyi Zhang and Suvrit Sra. First-order methods for geodesically convex optimization. In *Proceedings of the 29th Annual Conference on Learning Theory*, pages 1617–1638, 2016.

Hongyi Zhang and Suvrit Sra. An estimate sequence for geodesically convex optimization. In *Proceedings of the 31st Annual Conference on Learning Theory*, pages 1703–1723, 2018.

Jingzhao Zhang, Hongyi Zhang, and Suvrit Sra. R-spider: A fast Riemannian stochastic optimization algorithm with curvature independent rate. *arXiv preprint arXiv:1811.04194*, 2018.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on International Conference on Machine Learning*, pages 928–935, 2003.