Journal of Machine Learning Research 24 (2023) 1-52 Submitted 5/26; Revised -/–; Published -/–

Riemannian Optimistic Algorithms

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

Deming Yuan DMYUAN@NJUPT.EDU.CN
School of Automation
Nanjing University of Science and Technology
Jiangsu 210023, P. R. China

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

Zihao Hu ZIHAOHU@GATECH.EDU
College of Computing
Georgia Institute of Technology
Atlanta,GA 30339, The United States

Lei Wang LEI.WANGZJU@ZJU.EDU.CN
College of Control Science and Engineering
Zhejiang University
Zhejiang 310058, 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: My editor

Abstract

In this paper, we consider Riemannian online convex optimization with dynamic regret. First, we propose two novel algorithms, namely the Riemannian Online Optimistic Gradient Descent (R-OOGD) and the Riemannian Adaptive Online Optimistic Gradient Descent (R-AOOGD), which combine the advantages of classical optimistic algorithms with the rich geometric properties of Riemannian manifolds. We analyze the dynamic regrets of the R-OOGD and R-AOOGD in terms of regularity of the sequence of cost functions and

*. Correspondence author (G. Shi, Ross Street Building, Darlington NSW 2006, Sydney, Australia; +61-02-8627 8037).

comparators. Next, we apply the R-OOGD to Riemannian zero-sum games, leading to the Riemannian Optimistic Gradient Descent Ascent algorithm (R-OGDA). We analyze the average iterate and best-iterate of the R-OGDA in seeking Nash equilibrium for a two-player, zero-sum, g-convex-concave games. We also prove the last-iterate convergence of the R-OGDA for g-strongly convex-strongly concave problems. Our theoretical analysis shows that all proposed algorithms achieve results in regret and convergence that match their counterparts in Euclidean spaces. Finally, we conduct several experiments to verify our theoretical findings.

Keywords: Riemannian Manifold, Online Learning, Zero-sum Games

1 Introduction

Online optimization has become increasingly important in recent decades, as it aims to optimize a sequence of decision variables in real-time, despite uncertainty and limited feedback. The online optimization has numerous applications in fields such as machine learning, signal imaging, and control systems (Agmon, 1954; Hazan, 2022; Arnold et al., 2019).

The decision variables in online learning may be defined on Riemannian manifolds. Modeling signals on Riemannian manifolds can enhance data representation capabilities (Liu et al., 2019) and reduce problem dimension (Li et al., 2021; Hu et al., 2020). Moreover, Riemannian optimization benefits from the property of *geodesic convexity* (g-convexity) (Allen-Zhu et al., 2018), which permits conversion of Euclidean non-convex optimization problems into g-convex ones by appropriately choosing the Riemannian metric on the manifold. In this paper, we focus on the *Riemannian online convex optimization* (R-OCO) problem, defined as:

$$\min_{x_t \in \mathcal{K} \subset \mathcal{M}} \mathbf{f}_t(x_t), \tag{1}$$

where a learner plays against an adversary or nature. In each round $t \in 1, 2, \ldots, T$, the learner selects an action x_t from a geodesically convex (g-convex) subset $\mathcal{K}$. The adversary or nature then produces a geodesically convex (g-convex) function $\mathbf{f}_t$ defined on $\mathcal{K}$ for which the learner has no prior knowledge. Finally, the learner receives feedback on $\mathbf{f}_t$ and incurs a corresponding loss $\mathbf{f}_t(x_t)$. The problem (1) of Riemannian online convex optimization (R-OCO) is an extension of the classic online convex optimization in Euclidean spaces, with potential applications in machine learning, including robotic control, medical imaging, and neural networks (Lee and Kriegman, 2005; Fan et al., 2020; Shin and Oh, 2022).

In the context of the R-OCO problem, one important metric is the *dynamic regret* (Jadbabaie et al., 2015), which measures the difference in cumulative loss between an online optimization algorithm and a sequence of comparators $\{u_1, \ldots, u_T\}$, that is:

$$\mathsf{Reg}_{\mathsf{D}}(u_1, \ldots, u_T) = \sum_{t=1}^{T} \mathbf{f}_t(x_t) - \sum_{t=1}^{T} \mathbf{f}_t(u_t).$$

Compared to the well-known *static regret* (Zinkevich, 2003)

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

the dynamic regret provides a more comprehensive evaluation of online algorithms, as it takes into account the adjustments and adaptations of the environment at each time step.

While it has been demonstrated that online convex optimization algorithms may result in $\Omega(T)$ dynamic regret bound in the worst case, it is also possible to bound dynamic regrets related to quantities that reflect the regularity of the problem (Jadbabaie et al., 2015; Zhang et al., 2018; Zhao et al., 2020), such as the *path-length*

$$P_T := \sum_{t=2}^{T} d(u_t, u_{t-1}),$$

the *gradient variation*

$$V_T := \sum_{t=2}^{T} \sup_{x \in \mathcal{K}} \|\nabla \mathbf{f}_t(x) - \nabla \mathbf{f}_{t-1}(x)\|^2,$$

and the *comparator loss*

$$F_T := \sum_{t=1}^{T} \mathbf{f}_t(u_t).$$

If the comparators u_t and the cost function $\mathbf{f}_t$ adapts slowly, the dynamic regret can be greatly reduced.

In Euclidean online convex optimization (OCO), the Online Optimistic Gradient Descent algorithm (OOGD)(Jadbabaie et al., 2015; Zhao et al., 2020) is a noteworthy example that certifies dynamic regret bounds based on path-length and gradient variation. The OOGD has been extensively studied and shown to achieve near-optimal dynamic regret bounds for a wide range of convex optimization problems. For instance, Jadbabaie et al. (2015) presented a dynamic regret bound of $\mathcal{O}(P_T\sqrt{1+V_T})$ for the vanilla OOGD, and then Zhao et al. (2020) introduced a meta-expert structured OOGD that achieved a dynamic regret bound of $\mathcal{O}(\sqrt{(1+P_T+V_T)(1+P_T)})$. Moreover, the OOGD algorithm has also been applied to solve zero-sum games (Mokhtari et al., 2020b; Wei et al., 2021; Gorbunov et al., 2022), which may be used to study generative adversarial networks (Mertikopoulos et al., 2019).

Despite its success in Euclidean space, the extension of OOGD to Riemannian manifolds for dynamic regret has received limited attention. Some studies have examined the static regret of the R-OCO problem. Becigneul and Ganea (2019) studied the static regret of Riemannian adaptive methods, which required a product manifold structure. Wang et al. (2023) investigated the static regrets of the Riemannian online gradient descent algorithm and Riemannian online bandit methods.

With dynamic regrets, Maass et al. (2022) explored a zeroth-order dynamic regret bound for strongly g-convex and strongly g-smooth functions on Hadamard manifolds. In recent works, Hu et al. (2023) proposed several algorithms for adaptive dynamic regret bounds, incorporating gradient variation bounds, small loss bounds, and best-of-world bounds. In particular, Hu et al. (2023) introduced the gradient variation regret bounds via a Riemannian extragradient algorithm framework. The extragradient framework, similar to optimistic algorithms, typically requires two gradients per iteration. One gradient is computed at the

current decision point x_t, while the other gradient is obtained by extrapolating the current gradient to a midpoint y_t. However, it is worth noting that optimistic algorithms can operate solely based on the strategy point x_t. Thus, developing a Riemannian version of OOGD is crucial for advancing the field of online optimization on manifolds.

Contribution Motivated by above, this paper aims to design optimistic online methods for Riemannian online optimization and derive dynamic regret bounds with respect to the regularity of the problem. The contribution of this paper is summarized as follows:

- We propose the Riemannian online optimistic gradient descent algorithm (R-OOGD), which uses only the gradient of the strategy point. We establish an $\mathcal{O}(P_T\sqrt{1+V_T})$ dynamic regret bound for g-convex losses.
- We introduce the meta-expert framework in Hu et al. (2023) to the R-OOGD algorithm and propose the Riemannian online adaptive optimistic gradient descent (R-AOOGD) algorithm. We then establish a dynamic regret bound of $\mathcal{O}(\sqrt{(1+V_T+P_T)(1+P_T)})$ of the R-AOOGD on g-convex losses.
- We apply the R-OOGD algorithm to two-player zero-sum games on Riemannian manifolds and obtain the Riemannian Optimistic Gradient Descent Ascent algorithm (R-OGDA) for Nash equilibrium seeking in Riemannian zero-sum games. We prove $\mathcal{O}(\frac{1}{T})$ average-iterate and $\mathcal{O}(\frac{1}{\sqrt{T}})$ best-iterate convergence of the R-OGDA for g-convex-concave games. Moreover, we prove linear last-iterate convergence for g-strongly convex-strongly concave games.

The established regret bounds and convergence rates in our paper match the works in Euclidean space (Jadbabaie et al., 2015; Zhao et al., 2020; Mokhtari et al., 2020a,b). We briefly list them in Table 1.

2 Related Work

In this section, we will provide a brief review of previous work on online convex optimization and zero-sum games in both Euclidean spaces and Riemannian manifolds.

2.1 Online Convex Optimization

Euclidean OCO The concept of online optimization and static regret was introduced by Zinkevich (2003). Zinkevich (2003) also proposed the online gradient descent (OGD) method and constructed an $\mathcal{O}(\sqrt{T})$ regret bound on convex functions. Later, Hazan et al. (2006) demonstrated that the OGD method achieves an $\mathcal{O}(\log T)$ regret bound on strongly convex functions. Abernethy et al. (2008) proved the universal lower bounds for online algorithms to be $\Omega(\sqrt{T})$ and $\Omega(\log T)$ for convex and strongly convex functions respectively, which illustrated that the bounds of OGD are tight.

Aside from regret bounds related to the time horizon T, several papers explored the idea of exploiting regularity in the online optimization problem to achieve better regret bounds. For example, in addressing the online optimization problem where the loss functions have a small deviation, Chiang et al. (2012) proposed an online version of the extragradient algorithm, which achieved a regret bound of $\mathcal{O}(\sqrt{1+V_T})$ with respect to gradient variation

Algorithm	Problem setting	Result
R-OOGD	R-OCO, g-convex	$\mathcal{O}(\frac{\zeta_0}{\sqrt{\sigma_0}} P_T \sqrt{1+V_T})$
	Euclidean, convex	$\mathcal{O}(P_T\sqrt{1+V_T})$ (Jadbabaie et al., 2015)
R-AOOGD	R-OCO, g-convex	$\mathcal{O}(\frac{\zeta_0}{\sqrt{\sigma_0}}\sqrt{(1+V_T+P_T)(1+P_T)})$
	Euclidean, convex	$\mathcal{O}(\sqrt{(1+V_T+P_T)(1+P_T)})$ (Zhao et al., 2020)
R-OGDA	RZS, g-convex-concave, average-iterate	$\mathcal{O}(\frac{\zeta_1}{\sigma_1 T})$
	Euclidean convex-concave average-iterate	$\mathcal{O}(\frac{1}{T})$ (Mokhtari et al., 2020b)
	RZS, g-convex-concave best-iterate	$\mathcal{O}(\frac{1}{\sqrt{\sigma_1 T}})$
	Euclidean convex-concave best-iterate	$\mathcal{O}(\frac{1}{\sqrt{T}})$ (Chavdarova et al., 2021)
	RZS, g-SCSC, last-iterate	linear
	Euclidean, SCSC, last-iterate	linear (Mokhtari et al., 2020a)

Table 1: Comparison of Algorithms in our work and corresponding Euclidean algorithms. SCSC denotes strongly convex-strongly concave. The constants $\zeta_0, \zeta_1, \sigma_0, \sigma_1$ are related to Riemannian curvature and diameter bounds.

$V_T := \sum_{t=2}^{T} \sup_{x\in\mathcal{K}} \|\nabla \mathbf{f}_t(x) - \nabla \mathbf{f}_{t-1}(x)\|^2$ on convex and smooth functions. The online extragradient algorithm required constructing two gradients per iteration. To improve upon the extragradient algorithm, Rakhlin and Sridharan (2013) proposed an online optimistic gradient descent (OOGD) method. The OOGD algorithm was also proved to achieve $\mathcal{O}(\sqrt{1+V_T})$ regret bounds by requiring only one gradient per iteration, which is suitable for one-point gradient feedback.

For the dynamic regret, Zinkevich (2003) established a dynamic regret bound of $\mathcal{O}((1+P_T)\sqrt{T})$ for the OGD algorithm, and Jadbabaie et al. (2015) derived a dynamic regret bound of $\mathcal{O}(P_T\sqrt{1+V_T})$ for the OOGD algorithm. However, there is still a gap between these results and the universal lower dynamic regret bound of $\Omega(\sqrt{(1+P_T)T})$ mentioned by Zhang et al. (2018). To address this issue, Zhang et al. (2018) introduced a meta-expert framework to the OGD algorithm, which led to an optimal dynamic regret bound of $\mathcal{O}(\sqrt{(1+P_T)T})$. Building upon the work of Zhang et al. (2018), Zhao et al. (2020) applied the meta-expert framework to the OOGD algorithm using a novel Online Optimistic Hedge technique in the meta-algorithm and obtained a gradient-variation bound $\mathcal{O}(\sqrt{(1+V_T+P_T)(1+P_T)})$, a small-loss bound $\mathcal{O}(\sqrt{(1+F_T+P_T)(1+P_T)})$, and a best-of-both-worlds bound $\mathcal{O}(\sqrt{(1+\min(V_T,F_T)+P_T)(1+P_T)})$ in Euclidean space. In this paper, the proposed R-OOGD algorithm and the R-AOOGD algorithm extend the results

of the above Euclidean online optimistic algorithms to Riemannian manifolds beyond the restriction of linear structure.

Riemannian OCO Riemannian optimization has garnered significant interest from researchers in the past few decades (Zhang and Sra, 2016; Ahn and Sra, 2020; Becigneul and Ganea, 2019). In the context of Riemannian online convex optimization, Antonakopoulos et al. (2020) introduced a regularized method that leverages the Riemann–Lipschitz continuity condition, which specifically targeted convex functions in an ambient Euclidean space. Furthermore, Becigneul and Ganea (2019) provided regret analysis for Riemannian versions of the Adagrad and Adam algorithms, which relied on a product manifold structure. Later on, Maass et al. (2022) proposed a Riemannian online zeroth-order algorithm and analyzed the dynamic regret bound of $\mathcal{O}(\sqrt{T}+P_T^*)$ in the setting of g-strongly convex and g-smooth functions on Hadamard manifolds. Here, P_T^* represents the length of the path between the optimal solutions $P_T^* := \sum_{t=2}^{T} d(x_t^*, x_{t+1}^*)$. Wang et al. (2023) proposed Riemannian online gradient methods with a sublinear static regret of $\mathcal{O}(\sqrt{T})$ in the full information setting, $\mathcal{O}(T^{2/3})$ in the one-point bandit information setting, and $\mathcal{O}(\sqrt{T})$ in the two-point bandit information setting for g-convex functions. In this paper, the dynamic regret bound of the proposed R-OOGD algorithm holds for geodesically convex functions on general manifolds and achieves a better static regret $\mathcal{O}(\sqrt{1+V_T})$ than the above mentioned results.

A recent breakthrough in understanding dynamic regret for Riemannian OCO was made by Hu et al. (2023). Hu et al. (2023) first established a lower bound $\mathcal{O}(\sqrt{(1+P_T)T})$ for minimax dynamic regrets for the Riemannian OCO. Then, Hu et al. (2023) utilized the Fréchet mean and proposed a Riemannian meta-expert structure into the R-OGD, effectively achieving the universal lower bound for minimax dynamic regret. Additionally, Hu et al. (2023) enhanced the meta-algorithm by incorporating Optimistic Hedge methods into Riemannian manifolds, which lead to a gradient-variation regret bound $\mathcal{O}(\sqrt{(1+V_T+P_T)(1+P_T)})$, a small-loss regret bound $\mathcal{O}(\sqrt{(1+F_T+P_T)(1+P_T)})$, and a best-of-both-worlds regret bound $\mathcal{O}(\sqrt{(1+\min(V_T,F_T)+P_T)(1+P_T)}+\min(V_T,F_T)\log T)$. Moreover, Hu et al. (2023) also extended the regret bounds to the constrained setting using improper learning techniques. In order to obtain gradient-related dynamic regret bounds, Hu et al. (2023) proposed the RADRv algorithm, which utilizes an extragradient-type algorithm (referred to as R-OCEG) as expert algorithm. The main insight of the R-OCEG algorithm is the extrapolation step, where the gradient of the current strategy point is used to extrapolate and obtain a midpoint. The current strategy point is then updated based on the gradient of the extrapolated point. As a result, the R-OCEG algorithm requires two gradient information in each iteration. In contrast, our R-OOGD algorithm does not rely on gradient extrapolation. Instead, the R-OOGD combines the gradient at the current decision point with parallel transported gradients from past decision points to perform update. The approach that parallel transporting past gradients allows our R-OOGD algorithm to achieve the same dynamic regret bound of $\mathcal{O}(P_T\sqrt{1+V_T})$ as the R-OCEG, while requiring only one round of gradient computations each turn. Furthermore, compared to the RADRv, our R-AOOGD algorithm, which employs the R-OOGD as the expert algorithm, achieves the same dynamic regret bound of $\mathcal{O}(\sqrt{(1+V_T+P_T)(1+P_T)})$ while utilizing half of the gradient information.

2.2 Optimistic algorithm in zero-sum games

ODGA in Euclidean zero-sum games The Extragradient (EG) and Optimistic Gradient Descent Ascent (OGDA) methods have been extensively researched in the field of finding Nash equilibrium of Euclidean zero-sum games since the work of Korpelevich (1976). In the unconstrained convex-concave setting, both the EG and the OGDA methods exhibited $\mathcal{O}(1/T)$ average convergence rates, e.g., (e.g., Mokhtari et al., 2020b). Chavdarova et al. (2021) and Gorbunov et al. (2022) proved $\mathcal{O}(1/\sqrt{T})$ best-iterate convergence rates. In the constrained convex-concave setting, a more recent study by Cai et al. (2022) demonstrated an $\mathcal{O}(1/\sqrt{T})$ last-iterate convergence rate of the OGDA. For the strongly convex-strongly concave setting, Mokhtari et al. (2020a) demonstrated linear last-iterate convergence for both the OGDA and EG methods. Furthermore, Wei et al. (2021); Gorbunov et al. (2022) extended this linear convergence rate to the constrained setting. In this paper, we propose a Riemannian extension of the OGDA method which preserves the average iterate and best-iterate convergence rate for g-convex-concave games, and linear last-iterate for g-strongly convex-strongly concave games.

Riemannian zero-sum games Algorithms for coomputing the Nash equilibria (NE) in Riemannian zero-sum (RZS) games on Riemannian manifolds have been developed in the work of Li et al. (2009); Wang et al. (2010); Ferreira et al. (2005). Huang and Gao (2023) introduced a Riemannian gradient descent ascent algorithm for RZS games where the second variable y lied in a Euclidean space. Zhang et al. (2022) presented the Riemannian Corrected Extragradient algorithm (RCEG), which computes two gradients in one iteration and achieves $\mathcal{O}(\frac{1}{T})$ average-iteration convergence. Furthermore, for non-smooth functions, Jordan et al. (2022) presented $\mathcal{O}(\frac{1}{T})$ average-iterate convergence in the g-convex-concave setting and linear last-iterate convergence in the g-strongly-convex strongly-concave setting using Riemannian gradient descent ascent. Jordan et al. (2022) also constructed linear last-iterate convergence of the RCEG in the strongly g-convex setting. Han et al. (2022) introduced Riemannian Hamiltonian methods (RHM) that use the second-order Riemannian Hessian operator and established linear last-iterate convergence when the gradient norm of the payoff function satisfies the Riemannian Polyak–Łojasiewicz (PL) condition. In contrast, our ROGDA algorithm uses first-order information only once in each iteration and achieves the same average-iterate and last-iterate convergence. Moreover, our ROGDA algorithm shows the first best-iterate convergence result for g-convex-concave setting.

3 Preliminaries

Riemannian geometry A *Riemannian manifold* $(\mathcal{M}, g)$ is a manifold $\mathcal{M}$ with a point-varying Riemannian metric g. The Riemannian metric g induces an inner product $\langle u, v\rangle_x = g_x(u, v)$ on every tangent space $T_x\mathcal{M}$. Via the inner product, notions of geometry can be brought onto Riemannian manifolds. For example, the norm of $u \in T_x\mathcal{M}$ is defined as $\|u\| = \sqrt{\langle u, u\rangle_x}$, the angle between $u, v \in T_x\mathcal{M}$ is $\arccos \frac{\langle u,v\rangle_x}{\|u\|\|v\|}$, and the length of a curve $\gamma : [0, 1] \to \mathcal{M}$ is defined as $\int_0^1 \|\dot\gamma(t)\|dt$.

A Riemannian manifold $\mathcal{M}$ also enjoys a metric space structure with distance $d(x, y)$, which is the minimum of the lengths of the curves connecting x and y. A curve γ is called a *geodesic* if it locally reaches the minimum length. An *exponential map* $\exp_x$ acts as a vector

addition on Riemannian manifolds, which maps a tangent vector $v \in T_x\mathcal{M}$ to the endpoint $\gamma(1)$ of a geodesic γ with the initial tangent vector v. *Parallel transport* Γ_γ along the curve γ translates vectors from one tangent space to another while preserving the inner product, i.e., $\langle u, v\rangle = \langle \Gamma_\gamma u, \Gamma_\gamma v\rangle$. In particular, we denote Γ_x^y as the parallel transport along the geodesic between x and y. The parallel transport determines the covariant derivative of the vector field X along the vector field Y, which is defined as $\nabla_X Y(x) = \lim_{t\to 0} \frac{1}{t}\big(\Gamma_\gamma X(\gamma(t)) - X(x)\big)$, where $\dot\gamma(0) = Y(x)$.

One of the most important notions in Riemannian geometry is the curvature tensor, defined as $R(X, Y, W, Z) := \langle \nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z - \nabla_{[X,Y]} Z, W\rangle$. The *sectional curvature* is defined as $\frac{R(X,Y,X,Y)}{|X|^2|Y|^2 - \langle X,Y\rangle^2}$ and characterizes the non-flatness of a 2-dimensional Riemannian submanifold. On manifolds with non-positive sectional curvature, geodesics at one point x spread away from each other so that the inverse exponential map $\exp_x^{-1}$ can be defined globally, while on manifolds with positive sectional curvature K, geodesics at one point gather with each other, making the inverse exponential map well-defined only in a neighborhood of x with diameter less than $\frac{\pi}{\sqrt{K}}$. On the domain where the inverse exponential map is well-defined, we can write the distance function as $d(x, y) = \|\exp_x^{-1} y\|$.

Function classes We introduce some function classes on Riemannian manifolds for the further analysis. First, we introduce the concept of geodesic convexity on Riemannian manifolds. A function $\mathbf{f}$ is considered to be *geodesically convex* (or g-convex) on $\mathcal{K}$, if for any x and y belonging to $\mathcal{K}$, it satisfies

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

A function $\mathbf{f}$ is said to be μ-strongly geodesically convex (or μ-strongly g-convex) on $\mathcal{K}$, if for any x and y belonging to the manifold $\mathcal{M}$, the following inequality holds

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

Strong g-convexity also implies that

$$\langle -\nabla \mathbf{f}(x), \exp_x^{-1} x^*\rangle \geq \frac{\mu}{2} d^2(x, x^*), \quad \forall x \in \mathcal{M},$$

where x^* is the global minimizer. Furthermore, a function $\mathbf{f} : \mathcal{K} \to \mathbb{R}$ is called *geodesically concave* (g-concave) if $-\mathbf{f}$ is g-convex and a function $\mathbf{f} : \mathcal{M} \to \mathbb{R}$ is called *μ-strongly geodesically concave* (μ-strongly g-concave) if $-\mathbf{f}$ is μ-strongly g-convex.

We now define Lipschitz functions and smooth functions on Riemannian manifolds. We define a function $\mathbf{f} : \mathcal{M} \to \mathbb{R}$ as *geodesically G-Lipschitz* (or g-G-Lipschitz) if there exists a constant $G > 0$ such that, for any $x, x' \in \mathcal{M}$, the inequality $|\mathbf{f}(x) - \mathbf{f}(x')| \leq G \cdot d(x, x')$ holds. In the case of differentiable functions, this condition is equivalent to $\|\nabla \mathbf{f}(x)\| \leq G$ for all $x \in \mathcal{M}$. Similarly, a function $\mathbf{f} : \mathcal{M} \to \mathbb{R}$ is referred to as *geodesically L-smooth* if the gradient of f satisfies the g-L-Lipschitz property, meaning that for any $x, x' \in \mathcal{M}$, we have $\|\nabla \mathbf{f}(x) - \Gamma_{x'}^x \nabla \mathbf{f}(x')\| \leq L \cdot d(x, x')$.

We now shift our focus to bivariate functions within the context of Riemannian zero-sum games. We call a bivariate function $\mathbf{f}(x, y) : \mathcal{M} \times \mathcal{N} \to \mathbb{R}$ *g-convex-concave* (or *μ-g-strongly convex-strongly concave*), if for every $(x, y) \in \mathcal{M} \times \mathcal{N}$, $\mathbf{f}(\cdot, y) : \mathcal{M} \to \mathbb{R}$ is g-convex (or μ-strongly g-convex) and $\mathbf{f}(x, \cdot) : \mathcal{N} \to \mathbb{R}$ is g-concave (or μ-strongly g-concave).

4 Riemannian Online Optimization with Dynamic Regret

In this section, we first introduce the Riemannian online optimistic gradient descent algorithm (R-OOGD) and the Riemannian adaptive online optimistic gradient descent algorithm (R-AOOGD), which aims to improve the dynamic regret bound by averaging N R-OOGD algorithms with different step sizes. Then, we analyze the dynamic regret bounds of the R-OOGD and the R-AOOGD under g-convex functions.

4.1 Riemannian Online Optimistic Gradient Descent Method

The proposed R-OOGD algorithm is described in Algorithm 1. At each iteration t, the algorithm collects the gradient $\nabla \mathbf{f}_t(x_t)$ and combines it with additional momentum

$$\nabla \mathbf{f}_t(x_t) - \Gamma_{x_t-1}^{x_t} \nabla \mathbf{f}_{t-1}(x_{t-1}).$$

This combined gradient is then used in a gradient descent step via the exponential map $\exp_{x_t}$. The R-OOGD algorithm extends the Euclidean optimistic framework (Mokhtari et al., 2020a) to Riemannian manifolds.

Algorithm 1 Riemannian Optimistic Gradient Descent Algorithm (R-OOGD)

Require: Manifold $\mathcal{M}$, step size η
  Initialize $x_{-1} = x_0 = x_1 \in \mathcal{M}$.
  for $t = 1$ to $T - 1$ **do**
    Play x_t and receive $\nabla \mathbf{f}_t(x_t)$.
    Update $x_{t+1} = \exp_{x_t}(-2\eta \nabla \mathbf{f}_t(x_t) + \eta \Gamma_{x_{t-1}}^{x_t} \nabla \mathbf{f}_{t-1}(x_{t-1}))$
  end for
Ensure: Sequence $(x_t)_{t=1}^T$.

Next, inspired by Hu et al. (2023), we introduce a meta-expert framework to the R-OOGD. We propose the Riemannian adaptive online optimistic gradient descent algorithm (R-AOOGD) in Algorithms 2 and 3.

4.2 Dynamic Regret Analysis

In order to analyze the regret bounds of the R-OOGD and the R-AOOGD, we impose some assumptions, which are standard in the literature of online learning and Riemannian optimization (Antonakopoulos et al., 2020; Mokhtari et al., 2020b,a; Ahn and Sra, 2020; Alimisis et al., 2021).

Assumption 1. *The function $\mathbf{f}_t$ is g-convex, g-G-Lipschitz, and g-L-smooth over the set $\mathcal{K} \subset \mathcal{M}$.*

The following two assumptions focus on the geometry of the manifolds $\mathcal{M}$.

Assumption 2. *All sectional curvatures of $\mathcal{M}$ are bounded below by a constant κ and bounded above by a constant K.*

Assumption 3. *The diameter of the feasible set $\mathcal{K}$ is bounded by D_0. If $K > 0$, the diameter D_0 is less than $\frac{\pi}{2\sqrt{K}}$.*

Algorithm 2 R-AOOGD: Meta Algorithm

Require: Manifold $\mathcal{M}$, learning rate β, step size pool $\mathcal{H} = \eta_i; i = 1, 2, \ldots, N$ and parameters K, α.

1: Initialize $x_0 \in \mathcal{M}$. Set initial weights $w_{0,1} = w_{0,2} = \cdots = w_{0,N} = 1/N$.
2: **for** $t = 1$ to T **do**
3: Receive $x_{i,t}$ from N expert algorithms with step size η_i.
4: Set $\bar{x}_t = \arg\min_x w_{t-1,i} d^2(x, x_{t,i})$
5: Update $w_{t,i} \propto e^{\left(-\beta(\sum_{j=1}^{t-1} l_{i,j} + m_{i,t})\right)}$ by

$$\begin{cases} l_{i,t} = \langle \nabla \mathbf{f}_t(x_t), \exp_{x_t}^{-1} x_{i,t} \rangle \\ m_{i,t} = \langle \nabla \mathbf{f}_t(\bar{x}_t), \exp_{\bar{x}_t}^{-1} x_{i,t} \rangle \end{cases}$$

6: Set $x_t = \arg\min_x w_{t,i} d^2(x, x_{t,i})$.
7: **end for**
8: **return** $\{x_t\}_{t=1}^T$.

Algorithm 3 R-AOOGD: Expert Algorithm

Require: Manifold $\mathcal{M}$, feasible set $\mathcal{K}$ and step size η_i from the pool $\mathcal{H}$.

Initialize $x_{i,-1} = x_{i,0} = x_{i,1} \in \mathcal{K}$.
for $t = 1$ to $T - 1$ **do**
Send $x_{i,t}$ to the meta algorithm.
Update $x_{i,t+1} = \exp_{x_{i,t}}(-2\eta_i \nabla \mathbf{f}_t(x_{i,t}) + \eta_i \Gamma_{xi,t-1}^{x_{i,t}} \nabla \mathbf{f}_{t-1}(x_{i,t-1}))$.
end for

Assuming boundedness of the feasible set is a fundamental setting in online optimization algorithms (Zinkevich, 2003; Abernethy et al., 2008; Zhao et al., 2020), while the additional constraint $D_0 \leq \frac{\pi}{2\sqrt{K}}$ is also a common condition adopted in the literature of Riemannian optimization on positively curved manifolds (Zhang and Sra, 2016; Alimisis et al., 2021; Zhang et al., 2022). Assumption 3 ensures that there are no conjugate points on $\mathcal{K}$ according to the conjugate point theorem (Lee, 2018), guaranteeing that the inverse exponential map $\exp_x^{-1}(\cdot)$ can be defined throughout $\mathcal{K}$. Additionally, the Hessian comparison theorem (Lee, 2018) indicates that when the diameter of $\mathcal{K}$ is greater than $\frac{\pi}{\sqrt{K}}$, the subset $\mathcal{K}$ may be *"infinitely curved,"* meaning that the Hessian of the distance function $d(x, \cdot)$ may blow up. As in Riemannian optimization, we usually bound the loss by the g-convexity, i.e.,

$$\mathbf{f}_t(x_t) - \mathbf{f}_t(u_t) \leq \langle \nabla - \mathbf{f}_t(x_t), \exp_{x_t}^{-1} u_t \rangle = \langle \nabla \mathbf{f}_t(x_t), \nabla_{x_t}(\frac{1}{2} d^2(x_t, u_t)) \rangle,$$

and then use the Hessian of the distance function $d(x, u_t)$ for further analysis.

We also note that Assumption 3 does not necessarily affect the applicability of the proposed algorithms in practical problems. Our experiments (see Subsection 6.2) demonstrate that a feasible set $\mathcal{K}$ with a much larger diameter than $\frac{\pi}{2\sqrt{K}}$ (even the whole manifold $\mathcal{M}$) does not significantly affect the performance of our proposed algorithms.

We impose the following assumption on invariance of the set $\mathcal{K}$ during the execution of Algorithm 1. The same assumption has also been used in the work by Ahn and Sra

(2020); Alimisis et al. (2021); Zhang et al. (2022). In addition, and we do not observe the assumption to be violated in our experiments.

Assumption 4. *All iterations of Algorithms 1 lie in the set $\mathcal{K}$.*

Now we begin to analyze regret bounds. Our analysis heavily relies on comparison inequalities (Zhang and Sra, 2016; Alimisis et al., 2021), which enable us to quantify the distortion by nonlinear structure with respect to the curvature bound κ, K and domain diameter D_0. Specially, we can bound the minimum and maximum distortion rates by two parameters

$$\sigma(K,D)=\begin{cases}\frac{\sqrt{K}D}{\tan(\sqrt{K}D)} & K>0;\\ 1 & K\leq 0,\end{cases}\quad\text{and}\quad \zeta(\kappa,D)=\begin{cases}\frac{\sqrt{-\kappa}D}{\tanh(\sqrt{-\kappa}D)} & \kappa<0;\\ 1 & \kappa\geq 0.\end{cases}$$

After that, we can establish regret bounds for the R-OOGD algorithm.

4.2.1 Dynamic Regret for R-OOGD

Theorem 1. *Let $\sigma_0=\sigma(K,D_0)$ and $\zeta_0=\zeta(\kappa,D_0)$. Suppose that Assumptions 1-4 hold. Then, for the sequence $\{x_t\}_{t=1}^T$ generated by Algorithm 1 with step size $\eta\leq\frac{\sigma_0}{4\zeta_0 L}$, the following inequality holds for an arbitrary sequence $\{u_t\}_{t=1}^T$:*

$$\mathsf{Reg}_{\mathsf{D}}(u_1,\ldots,u_T)\leq\frac{D_0^2+2D_0P_T}{\eta}+\eta\frac{4\zeta_0^2}{\sigma_0}(G^2+V_T),$$

where $P_T=\sum_{t=2}^T d(u_t,u_{t-1})$ is the path-length, and $V_T=\sum_{t=2}^T\sup_{x\in\mathcal{K}}\|\nabla\mathbf{f}_t(x)-\nabla\mathbf{f}_{t-1}(x)\|^2$ is the gradient variation. If V_T is known beforehand, the dynamic regret bound can be improved to $\mathcal{O}(P_T\sqrt{1+V_T})$.

Corollary 2. *Suppose that Assumptions 1-4 hold. Algorithm 1 incurs the static regret with the optimal step size $\eta^*=\min(\sqrt{\frac{D_0^2\sigma_0}{4\zeta_0^2(G^2+V_T)}},\frac{\sigma_0}{4\zeta_0 L})$,*

$$\mathsf{Reg}_{\mathsf{S}}(T)\leq\mathcal{O}\big(\zeta_0\sqrt{\frac{1}{\sigma_0}(1+V_T)}\big).$$

The proofs of Theorem 1 and Corollary 2 are shown in Appendix B. In Theorem 1 and Corollary 2, we provide an $\mathcal{O}(P_T\sqrt{1+V_T})$ dynamic regret bound and an $\mathcal{O}(\sqrt{1+V_T})$ static regret bound for the R-OOGD, respectively. These bounds recover the corresponding work on online optimization in Euclidean space (Jadbabaie et al., 2015; Zhang et al., 2018; Zhao et al., 2020). Compared to the extragradient-structured R-OCEG algorithm by Hu et al. (2023) in the unconstrained setting, our method only requires one gradient information per iteration, making it less computationally demanding and more applicable to the one-point gradient feedback online learning model.

4.2.2 R-OOGD with Parallel Transport

Conversion of past gradients plays a vital role in designing Riemannian optimistic algorithms. Apart from the parallel transport method utilized in our R-OOGD algorithm, there

is another methodology known as the correction term to transport past graident in Riemannian optimization (Zhang et al., 2022; Hu et al., 2023). The correction term involves changing the base point of the exponential map to avoid parallel transport. Specifically, when considering two points a and b, along with a tangent vector $v \in T_a\mathcal{M}$, the correction term employs $\exp_a(v + \exp_a^{-1}(b))$ instead of the direct parallel transport $\exp_b(\Gamma_a^b v)$. If our R-OOGD algorithm incorporates the correction term methodology, the algorithm will be expressed as follows,

$$\begin{cases} x_{t+1} = \exp_{x_t}(-2\eta\nabla \mathbf{f}_t(x_t) + \exp_{x_t}^{-1} \hat{x}_t) \\ \hat{x}_{t+1} = \exp_{x_t}(-\eta\nabla \mathbf{f}_t(x_t) + \exp_{x_t}^{-1} \hat{x}_t). \end{cases} \tag{2}$$

However, incorporating the correction term into our R-OOGD algorithm does not yield an efficient online algorithm and is unable to guarantee a sublinear static regret. This is because the correction term introduces distortion in the inner product, and these distortions grow unboundedly over time. To be more precise, in order to analyze the static regret bound, we need to estimate the distortion in the inner product between the correction term and the real gradient term

$$A_t := \langle \exp_{x_t}^{-1} \hat{x}_t - \Gamma_{x_{t-1}}^{x_t} \eta\nabla \mathbf{f}_{t-1}(x_{t-1}), \exp_{x_t}^{-1} x \rangle$$

at each iteration. However, as demonstrated in Appendix C, the distortion bound A_t depends on the previous distortion A_{t-1}, which is governed by the recursive formula:

$$A_t \leq \mathcal{O}\Big((5\eta G + 2A_{t-1})^2(3\eta G + A_{t-1})\Big). \tag{3}$$

Therefore, the distortion accumulates iteratively and may eventually blow up. In contrast, our parallel transport method preserves the inner product, guaranteeing that the distortion term remains zero. Consequently, it is crucial to emphasize that parallel transporting past gradients is essential to ensure the effectiveness of the R-OOGD algorithm.

4.2.3 Dynamic Regret for R-AOOGD

Now we begin to analyze the the dynamic bound of the R-AOOGD algorithm.

Theorem 3. *Suppose that Assumptions 1-4 hold. Set* $\mathcal{H} = \{\eta_i = 2^{i-1}\sqrt{\frac{\sigma_0 D_0^2}{16\zeta_0^2 G^2 T}}\}$, $\beta = \min\Big(\frac{1}{\sqrt{12D_0^4L^2 + D_0^2G^2\zeta_0^2}}, \sqrt{\frac{2+\ln N}{3D_0^2(V_T+G^2)}}\Big)$ *and* $N = \Big\lceil \frac{1}{2}\log_2\big(\frac{\sigma_0 G^2 T}{D_0^2 L^2}\big)\Big\rceil + 1$. *The R-AOOGD algorithm (Algorithms 2 and 3) incurs the dynamic regret*

$$\mathsf{Reg}_{\mathsf{D}}(u_1, \ldots, u_T) \leq \mathcal{O}\Big(\frac{\zeta_0}{\sqrt{\sigma_0}}\sqrt{(1 + V_T + P_T)(1 + P_T)}\Big).$$

The proof of Theorem 3 can be found in Appendix B. In Theorem 3, we establish an $\mathcal{O}\Big(\sqrt{(1 + V_T + P_T)(1 + P_T)}\Big)$ regret bound for our meta-expert algorithm R-AOOGD which aligns with the findings in online optimization in Euclidean space (Zhao et al., 2020). When considering the RADRv algorithm (Hu et al., 2023), which is another Riemannian online meta-expert algorithm, we observe certain differences compared to our R-AOOGD

algorithm. In the non-projection case, the RADRv algorithm at the meta level requires $N = \left\lceil \frac{1}{2}\log_2\left(\frac{4\sigma_0 G^2 T}{D_0^2 L^2}\right)\right\rceil + 1$ expert algorithms, which is equal to or more than the number of experts required by our R-AOOGD algorithm. Additionally, at the expert level, the RADRv requires two gradients per iteration, while our R-AOOGD algorithm only requires one gradient. As a result, our R-AOOGD algorithm achieves the same order of regret bound of $\mathcal{O}\left(\sqrt{(1+V_T+P_T)(1+P_T)}\right)$ while utilizing only half the number of gradients.

5 Application: Nash Equilibrium Seeking in Riemannian Zero-sum Games

In this section, we apply our R-OOGD to Riemannian zero-sum (RZS) games, where both players update their actions based on the R-OOGD dynamics. We then analyze the convergence rates of the resulting Riemannian Optimistic Gradient Descent Ascent method (R-OGDA). Specifically, we study the average-iterate convergence rate and the best-iterate convergence rate for g-convex-concave games, as well as the last-iterate convergence rate for g-strongly convex-strongly concave games.

5.1 Formulation of Riemannian Zero-sum Games

Riemannian zero-sum (RZS) games involve a competitive scenario between two players, denoted as `X` and `Y`, i.e.,

$$\min_{x\in\mathcal{M}}\max_{y\in\mathcal{N}} \mathbf{f}(x,y), \tag{4}$$

where player-`X` tries to find a strategy x from a Riemannian manifold $\mathcal{M}$ to minimize the payoff function $\mathbf{f}(x,y)$, while player-`Y` tries find a strategy y from a Riemannian manifold $\mathcal{N}$ to maximize the payoff function $\mathbf{f}(x,y)$.

A key concept in the RZS games is *Nash Equilibrium* (NE). In the RZS game, an action pair $(x^*, y^*) \in \mathcal{M}\times\mathcal{N}$ is an NE if no player can improve individual payoff by only deviating his own action, i.e.,

$$\max_{y\in\mathcal{N}} \mathbf{f}(x^*,y) = \mathbf{f}(x^*,y^*) = \min_{x\in\mathcal{M}} \mathbf{f}(x,y^*).$$

Computing NEs in the Riemannian zero-sum games has direct applications in several learning tasks including minimum balanced cut, robust geometry-aware PCA, and robust Wasserstein barycenters (Khuzani and Li, 2017; Horev et al., 2016; Lin et al., 2020; Zhang et al., 2022).

5.2 Riemannian Optimistic Gradient Descent Ascent Algorithm

We propose the Riemannian Optimistic Gradient Descent Ascent Algorithm (R-OGDA) as follows.

Algorithm 4 Riemannian Optimistic Gradient Descent Ascent

Require: $(\mathcal{M},\mathcal{N})$, step size η,

Initialize $(x_{-1}, y_{-1}) = (x_0, y_0) \in \mathcal{M} \times \mathcal{N}$.

for $t = 2$ to $T-1$ **do**

Update $x_{t+1} = \exp_{x_t}(-2\eta\nabla_x \mathbf{f}(x_t, y_t) + \eta\Gamma_{x_{t-1}}^{x_t}\nabla_x \mathbf{f}(x_{t-1}, y_{t-1}))$

Update $y_{t+1} = \exp_{y_t}(2\eta\nabla_y \mathbf{f}(x_t, y_t) - \eta\Gamma_{y_{t-1}}^{y_t}\nabla_y \mathbf{f}(x_{t-1}, y_{t-1}))$

Average from geodesic

$$\begin{cases} \bar{x}_{t+1} = \exp_{\bar{x}_t}(\frac{1}{(t+1)}\exp_{\bar{x}_t}^{-1} x_{t+1}) \\ \bar{y}_{t+1} = \exp_{\bar{y}_t}(\frac{1}{(t+1)}\exp_{\bar{y}_t}^{-1} y_{t+1}) \end{cases} \tag{5}$$

end for

Ensure: Sequence $(x_t, y_t)_{t=1}^T$, average $(\bar{x}_T, \bar{y}_T)$.

5.3 Average-Iterate Analysis

We first analyze the convergence rate of averaged iterate $(\bar{x}_T, \bar{y}_T)$ by adpoting the following assumptions. To ease the notation, we denote $z_t = (x_t, y_t)$, $z^* = (x^*, y^*)$, and $\mathbf{F}(z_t) = [\nabla_x \mathbf{f}(z_t), -\nabla_y \mathbf{f}(z_t)]$.

Assumption 5. *The payoff function f is g-G-Lipschitz and g-convex-concave on the manifold $\mathcal{M} \times \mathcal{N}$.*

The g-convexity-concavity is helpful in analyzing Nash equilibriums. For any NE point $z^* = (x^*, y^*)$, the g-convexity-concavity of $\mathbf{f}$ implies that

$$\langle -\mathbf{F}(z), \exp_z^{-1} z^* \rangle \geq 0$$

holds for all $z \in \mathcal{M}$. Moreover, if $\mathbf{f}$ is μ-g-strongly-convex strongly-concave, it further holds that for all $z \in \mathcal{M}$,

$$\langle -\mathbf{F}(z), \exp_z^{-1} z^* \rangle \geq \frac{\mu}{2} d^2(z, z^*). \tag{6}$$

Assumption 6. *The payoff function f is g-L-smooth on $\mathcal{M} \times \mathcal{N}$, i.e., there is a constant $L > 0$ such that*

$$\|\nabla \mathbf{f}(x, y) - \Gamma_{(x',y')}^{(x,y)} \nabla \mathbf{f}(x', y')\|^2 \leq L^2(d^2(x, x') + d^2(y, y')), \quad \forall (x, y), (x', y') \in \mathcal{M} \times \mathcal{N}.$$

Assumption 7. *There is a Nash equilibrium $z^* = (x^*, y^*)$ in the Riemannian zero-sum game* (4).

Assumption 7 can be directly derived from the g-strongly convex-strongly concave property of $\mathbf{f}$ in certain cases. In the case where the manifold $\mathcal{M}$ is compact or the level set $\{(x, y) \mid -\infty \leq \mathbf{f}(x, y) \leq \infty\}$ is bounded, we can rely on the Riemannian analog of Sion's minimax theorem (Zhang et al., 2022) to establish the existence of a Nash equilibrium. Furthermore, if we assume g-strong-convexity strong-concavity of the function $\mathbf{f}$, by (6), Assumption 7 guarantees the uniqueness of the resulting Nash equilibrium z^*.

Assumption 8. *All sectional curvatures of $\mathcal{M}$ and $\mathcal{N}$ are bounded below by a constant κ and bounded above by a constant K.*

We first show that it is possible to drop the boundedness assumption (Assumption 3) in the iteration of x_t and y_t in the Riemannian zero-sum (RZS) games (4), if prior knowledge of the distance $d(z_0, z^*)$ is available. The following Lemma 4 ensures that the iteration of x_t and y_t is in a bounded set $\mathcal{K}$.

Lemma 4. *Suppose Assumptions 5-8 hold. Let the step size η satisfying $\eta \leq \min(\frac{\sigma_1}{\zeta_1 L}, \frac{D_1}{3G})$. Let $d(z_0, z^*) \leq D_1 < \frac{\pi}{6\sqrt{K}}$, then by denoting $\zeta_1 = \zeta(\kappa, 3D_1)$ and $\sigma_1 = \zeta(K, 3D_1)$, we have $d(z_t, z^*) \leq 2D_1$ for all iterations t.*

Then we derive the average-iterate convergence rate for g-convex-concave RZS games in the following Theorem 5.

Theorem 5. *Under Assumptions 5-8 and the condition in Lemma 4, the averaged iterate $(\bar{x}_T, \bar{y}_T)$ of Algorithm 4 with the step size $\eta \leq \frac{\sigma_1}{2\zeta_1 L}$ satisfies:*

$$\max_{y \in \mathcal{N}} \mathbf{f}(\bar{x}_T, y) - \min_{x \in \mathcal{M}} \mathbf{f}(x, \bar{y}_T) \leq \frac{D_1^2 L + \frac{D_1^2}{2\eta}}{T}$$

The proof of Theorem 5 is in Appendix D. Theorem 5 demonstrates that the averaged iterate $(\bar{x}_T, \bar{y}_T)$ is an $\mathcal{O}(\frac{1}{T})$-NE in g-convex-concave RZS games. The result extends the corresponding results in Euclidean spaces (Mokhtari et al., 2020a) and recovers the convergence rate of RCEG (Zhang et al., 2022) by acquiring only one gradient in each iteration.

5.4 Last-Iterate/Best-Iterate Analysis

In this section, we focus on the last/best-iterate convergence of our R-OGDA algorithm. Dealing with the convergence of the R-OGDA algorithm in terms of the last-iterate or best-iterate is quite challenging. In Euclidean spaces, optimistic gradient descent/ascent algorithms benefit from the perspective of "extrapolation from the past" (Gidel et al., 2019). This means that by defining the immediate sequence

$$\hat{z}_{t+1}^E = z_t - \eta(\mathbf{F}(z_t) + \mathbf{F}(z_{t-1})), \tag{*}$$

the relationship holds Gidel et al. (2019):

$$\eta \mathbf{F}(z_t) = \hat{z}_t^E - \hat{z}_{t+1}^E. \tag{**}$$

However, when we define the Riemannian counterparts of (*) $\hat{z}_{t+1} = \exp_{z_t}(\eta(-\mathbf{F}(z_t) + \mathbf{F}(z_{t-1})))$, it fails to hold that $\exp_{\hat{z}_{t+1}}^{-1} \hat{z}_t = \eta \Gamma_{z_t}^{\hat{z}_{t+1}} \mathbf{F}(z_t)$, which is the Riemannian version of (**). In addition, measuring the distortion between vector $\exp_{\hat{z}_{t+1}}^{-1} \hat{z}_t$ and $\eta \Gamma_{z_t}^{\hat{z}_{t+1}} \mathbf{F}(z_t)$ is challenging.

One natural idea is to parallel transport the tangent vectors $\exp_{z_t}^{-1} z_{t-1}$, $\exp_{z_t}^{-1} \hat{z}_t$ and $\exp_{z_{t+1}}^{-1} \hat{z}_{t+1}$ to the tangent space at $\hat{z}_{t+1}$, and then estimate the distortion between G_t and $\Gamma_{z_t}^{\hat{z}t+1} \mathbf{F}(z_t)$ using their vector sum. A commonly used technique in manifold settings, the

comparison inequalities (Zhang and Sra, 2016; Alimisis et al., 2021), follows the same idea to estimate the Hessian-type distortion, i.e.,

$$\sigma(K,D)\|\exp_a^{-1} b\| \leq \|\exp_a^{-1} b - \Gamma_c^a \exp_c^{-1} b\| \leq \zeta(k,D)\|\exp_a^{-1} b\|.$$

The crux of these inequalities is to use the gradient squared distance $\nabla d(b,x) = -\exp_b^{-1} x$ on the manifold, and then estimate the error by exploiting the eigenvalues of the Hessian matrix of the distance function. Therefore, the comparison inequalities require the compared tangent vectors to share the same endpoint, which is not applicable in our case.

Furthermore, we have observed that the distortion analysis between G_t and $\Gamma_{z_t}^{\hat{z}t+1}\mathbf{F}(z_t)$ brings in extra errors beyond the Hessian-type distortions. Parallel transporting vectors in the two distinct tangent spaces to a third tangent space can lead to additional distortion. Specifically, when dealing with three points a, b, c, and vectors $v \in T_a\mathcal{M}$ and $w \in T_b\mathcal{M}$, we have

$$\|\Gamma_a^c v - \Gamma_b^c w\| \neq \|\Gamma_a^b v - w\|.$$

In particular, it holds

$$\|\Gamma_a^c v - \Gamma_b^c w\| = \|\Gamma_a^b v - w\| + \|\Gamma_b^a \Gamma_c^b \Gamma_a^c v - w\|.$$

On Riemannian manifolds, the latter term $\|\Gamma_b^a \Gamma_c^b \Gamma_a^c v - w\|$ does not vanish due to the presence of the holonomy effects, which indicates that parallel transport around closed loops fails to preserve the geometric data being transported. Therefore, estimating the distortion between G_t and $\Gamma_{z_t}^{\hat{z}t+1}\mathbf{F}(z_t)$ is faced with the change of holonomy distortion and becomes highly nontrivial.

To address the aforementioned difficulties, we propose a key technique inspired by the famous Gauss-Bonnet theorem (Chern et al., 1999), and estimate the holonomy distortion by the step size η and sectional curvature $K_m = \max\{|K|, |\kappa|\}$.

Lemma 6. *Suppose that* $\mathbf{f}$ *is g-L-smooth. If the step size* $\eta \leq \frac{1}{20L}$*, them* $G_{t+1} = \exp_{\hat{z}_{t+1}}^{-1} \hat{z}_t$ *satisfies the following*

(i) $\|G_{t+1}\|^2 - \|\eta\mathbf{F}(z_t)\|^2 \leq 64K_m^2\eta^6\|\mathbf{F}(z_{t-1})\|^6$*;*

(ii) $\|\Gamma_{\hat{z}_{t+1}}^{z_t} G_{t+1} - \eta\mathbf{F}(z_t)\| \leq 104K_m\eta^3\|\mathbf{F}(z_{t-1})\|^3$.

According to Lemma 6, the holonomy distortion turns out to be $\mathcal{O}(K_m\eta^3)$, which enables us to obtain the convergence in Theorems 7 and 8. In the following theorems, we denote $\Upsilon = \frac{1}{5}\sigma_1 L + \frac{28}{5}(\zeta_1 - \sigma_1)L + 104(2D_1 + \frac{1}{5})K_m G + 8\sigma_1 K_m G$.

Theorem 7. *Suppose Assumptions 5-8 and conditions in Lemma 4 hold. With the step size* $\eta \leq \min\{\frac{1}{20L}, \frac{1}{8G}, \frac{\sigma_1}{2\Upsilon}\}$*. Algorithm 4 incurs the best-iterate convergence,*

$$\min_{t\leq T} \|\nabla\mathbf{f}(z_t)\| \leq \mathcal{O}(\frac{1}{\sqrt{T}}),$$

and moreover, we have $\lim_{t\to\infty} \|\nabla\mathbf{f}(z_t)\| = 0$.

Theorem 8. *Suppose Assumptions 5-8 and conditions in Lemma 4 hold and* $\mathbf{f}$ *is* μ*-g-strongly convex-strongly concave. Recall* z^* *is the unique NE of the game* (4)*. Algorithm 4 with the step size* $\eta \leq \min\{\frac{1}{20L}, \frac{1}{8G}, \frac{\sigma_1}{\Upsilon+4\mu+8\sigma_1\mu}\}$ *incurs*

$$d^2(z_t, z^*) \leq \Big(\frac{1}{1+\eta\mu/2}\Big)^t \big(2(1+\frac{1}{\sigma_1})d^2(z_1, z^*)\big).$$

The proofs for Theorems 7 and 8 can be found in Appendix F. Theorem 7 demonstrates the $\mathcal{O}(\frac{1}{\sqrt{T}})$ convergence rate for the best iterate of g-convex-concave games, which is the first proven result in Riemannian NE seeking algorithms for g-convex-concave games. Moreover, Theorem 8 establishes a linear convergence rate for the last-iterate of the ROGDA algorithm, matching those in the RCEG algorithm and the second-order RHM algorithm (Han et al., 2022; Jordan et al., 2022) in the g-strongly convex-strongly concave setting, while requiring only one first-order information in each iteration.

6 Numerical Experiments

In this section, we presents several numerical experiments to validate our theoretical findings regarding Riemannian online optimization problems and Riemannian zero-sum games. We conduct experiments on both synthetic and real-world datasets, and compare the performance of our proposed algorithm with state-of-the-art methods in the literature. We implement our algorithm using the Pymanopt package Boumal et al. (2014) and conduct all experiments in Python 3.8 on a machine with an AMD Ryzen5 processor clocked at 3.4 GHz and 16GB RAM. To ensure reproducibility of our results, we provide access to all source codes online[1].

6.1 Online Fréchet Mean in the Hyperbolic Space

The Fréchet mean problem, also known as finding the Riemannian centroid of a set of points on a manifold, has numerous applications in various fields, including diffusion tensor magnetic resonance imaging (DT-MRI) (Cheng et al., 2012) and hyperbolic neural network (Liu et al., 2019). We focus on the online version of the Fréchet mean problem, which aims to compute the average of N time-variant points in a hyperbolic space. Hyperbolic space is a Riemannian manifold with constant negative sectional curvature -1, defined as

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

where the Minkowski dot product

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

defines the metric $\langle x, y\rangle_p = \langle x, y\rangle_M$. The loss function $\mathbf{f}_t$ of the online Fréchet mean problem is given by

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

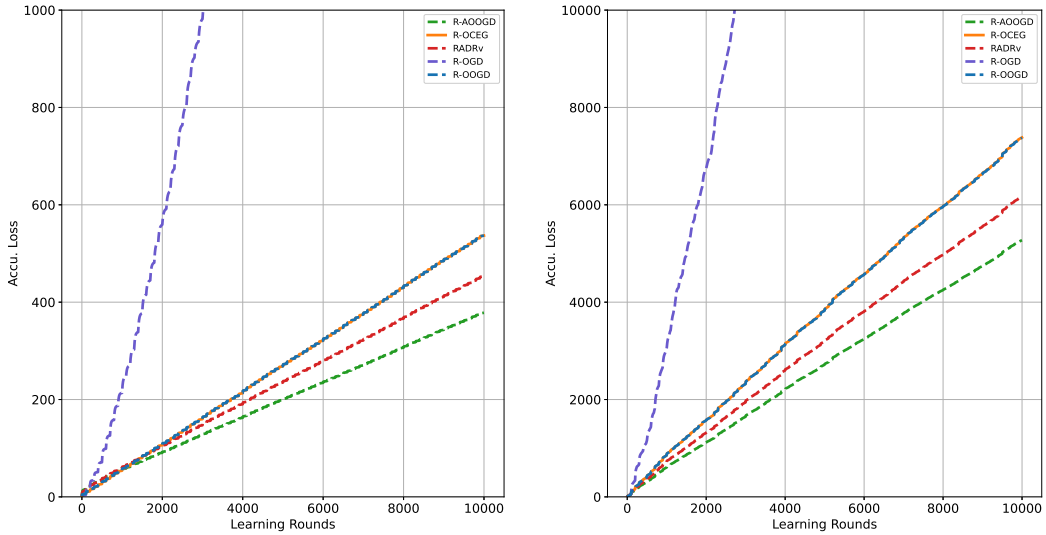
1. `https://github.com/RiemannianOCO/DynamicReg`

where $\{A_{t,1}, A_{t,2}, \ldots, A_{t,N}\}$ are the time-variant points in the hyperbolic space.

Experimental setting We test the online Fréchet mean on synthetic datasets generated as follows: In each iteration t, we randomly sample the point $A_{i,t}$ from a ball centered at a point P_t with a radius of $c = 1$. The choice of P_t is in following two ways to simulate non-stationary environments. 1) P_t remains fixed in between S time steps. After every S rounds, we re-select P_t randomly within a bounded set of diameter $D = 1$ to simulate abrupt changes in the environment. 2) P_t shifts a small distance of 0.1 between each time step, and is re-selected after S round, to simulate a slowly evolving environment. For our experiment, we set $T = 10000$, $n = 100$, $d = 20$, $D = 1, \kappa = 1$, and $L = \zeta(\kappa, D)$.

In addition, we compare our R-OOGD and R-AOOGD algorithms to other contenders in Riemannian online optimization. Specifically, we compare R-OOGD with the Riemannian online gradient descent (ROGD) algorithm (Wang et al., 2023) and the EG-type expert algorithm R-OCEG (Hu et al., 2023). We also compare our meta-expert algorithm R-AOOGD with the meta-expert RADRv algorithm (Hu et al., 2023).

Result We examine the performance in terms of cumulative loss and present the result in Fig 6.1. First, we can see that OGD suffers from a high cumulative loss throughout the horizon. Conversely, our methods, as well as R-OCEG and RADRv, demonstrate satisfactory performance in terms of dynamic regret in both situations. As meta-expert algorithms, our R-AOOGD slightly outperforms the RARDv. Our R-OOGD algorithm performs comparably to RARDv-exp, but with only half number of the gradient information required. These facts validate the effectiveness of our algorithms and demonstrate our advantage in situations where gradient computation is time-consuming.



(a) Stationary environment with abrupt changes

(b) Slowly evolving environment

Figure 1: Algorithm performance on hyperbolic Fréchet mean problems

6.2 Online Geodesic Regression

Geodesic Regression is a statistical model that generalizes the Euclidean least-square regression by modeling data as

$$y = \exp_{\hat{y}_t}(\epsilon), \quad \hat{y}_t = \exp_p(xV),$$

where $x \in \mathbb{R}$ is the feature, $y \in \mathcal{M}$ is the manifold-valued label, V is a tangent vector at P, and ϵ is a Gaussian-distributed error. Geodesic Regression has many applications in medical imaging, object recognition, and linear system identification (Yang and Hospedales, 2016; Shin and Oh, 2022; Hong et al., 2014).

We can also consider an online form of geodesic regression, where the model is trained sequentially. For each data point x_t, y_t, the online geodesic regression model minimizes the loss function

$$\mathbf{f}_t(p_t, V_t) = \frac{1}{2}d^2(\hat{y}_t, y_t) = \frac{1}{2}d^2(\exp_{p_t}(x_t V_t), y_t),$$

where $(p_t, V_t) \in T\mathcal{M}$ lies in the tangent bundle $T\mathcal{M}$ with the Sasaki metric (Muralidharan and Fletcher, 2012),

$$\left\langle (p_1', V_1'), (p_2', V_2') \right\rangle = \langle p_1', p_2' \rangle_{\mathcal{M}} + \langle V_1', V_2' \rangle_{T_P\mathcal{M}}.$$

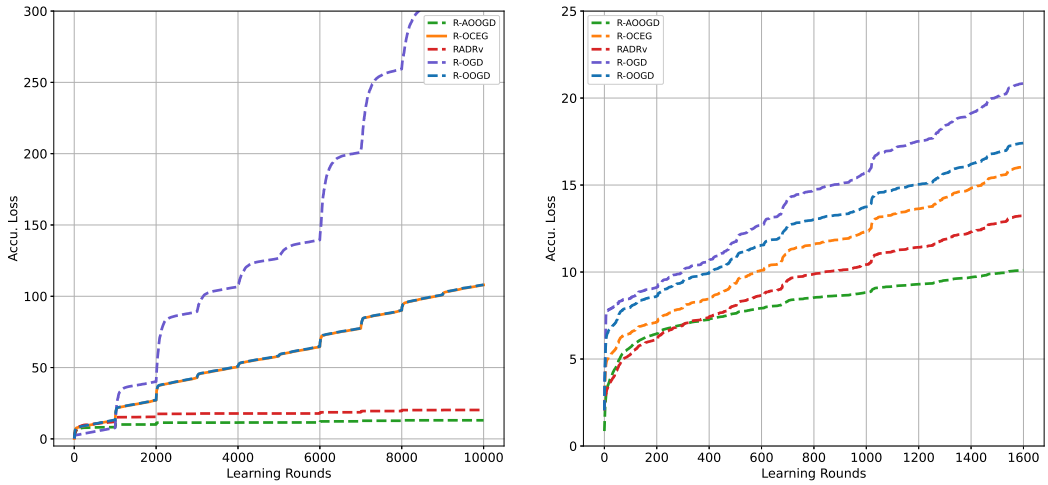
Experimental setting We now conduct our experiments on both synthetic and real-world dataset. The synthetic dataset is generated on a five-dimensional sphere $\mathbb{S}^5$. At each round, the feature x_t is uniformly sampled from $[0, 1]$, and the label y_t is obtained as $y_t = \exp_{\hat{y}_t}(\epsilon_t)$, where $\hat{y}_t = \exp_{p_t}(x_t V_t)$ and ϵ_t is a random tangent vector with norm chosen uniformly from $[0, 0.1]$. Similar to the previous experiment, we fix p_t and V_t for S rounds and randomly select new p_t and V_t from the half sphere after every S rounds.

The real-world dataset used in this experiment is the *corpus callosum dataset* from the Alzheimer's disease neuroimaging initiative, which was provided by Cornea et al. (2017) and can also be accessed online [2]. The dataset includes information about 408 subjecst as well as shape of the subjects' corpus callosum obtained from mid-sagittal slices of magnetic resonance images (MRI). The shape of an corpus callosum is described as a cloud point matrix $y_t \in \mathbb{R}^{50\times 2}$. During the experiment, we aimed to analyze the relationship between the age of the subjects and the shape of the corpus callosum by utilizing geodesic regression. To achieve this, we preprocessed the shape information into a Grassmann manifold $\mathcal{G}r(50, 2)$ by computing the left-singular vectors of each y_t as reported in a previous study by Hong et al. (2014). Then, we divided the 320 data points into a training set and others to a testing set. Additionally, we duplicated training data points 5 times for efficient learning rounds.

Results Figure 6.2 shows the accuracy and loss versus learning round for our algorithm. Additionally, we provide performance results on the test set of the corpus callosum shapes in Figure 6.2, which confirms the effectiveness of our R-OOGD algorithm and R-AOOGD algorithm in solving real-world problems. Our algorithm performs similarly or better than other state-of-the-art methods on both synthetic and real-world datasets, while requiring fewer gradient information. This finding aligns with our theoretical results.

2. `http://www.bios.unc.edu/research/bias/software.html`

It is noteworthy that we did not impose any special requirements on the boundedness of the training set in the real-world dataset. Therefore, our algorithm iterates over the entire Grassmann manifold with a diameter of $\frac{\sqrt{2}}{2}\pi > \frac{\pi}{2\sqrt{K}} = \frac{\pi}{2\sqrt{2}}$. This demonstrates the applicability of our algorithm when the diameter $D \geq \frac{\pi}{2\sqrt{K}}$.



(a) Synthetic data

(b) Real-world data

Figure 2: Algorithm performance on geodesic regression

6.3 Quadratic Geodesic Lodget Game

We now validate our theoretical finding in Riemannian zero-sum games. Consider the following toy-example RZS game

$$\min_{X\in\mathcal{S}_d^{++}} \max_{Y\in\mathcal{S}_d^{++}} c_1(\log\det(X))^2 + c_2\log\det(X)\log\det(Y) - c_1(\log\det(Y))^2, \tag{7}$$

where X and Y take values on the symmetric positive definite (SPD) matrix manifold

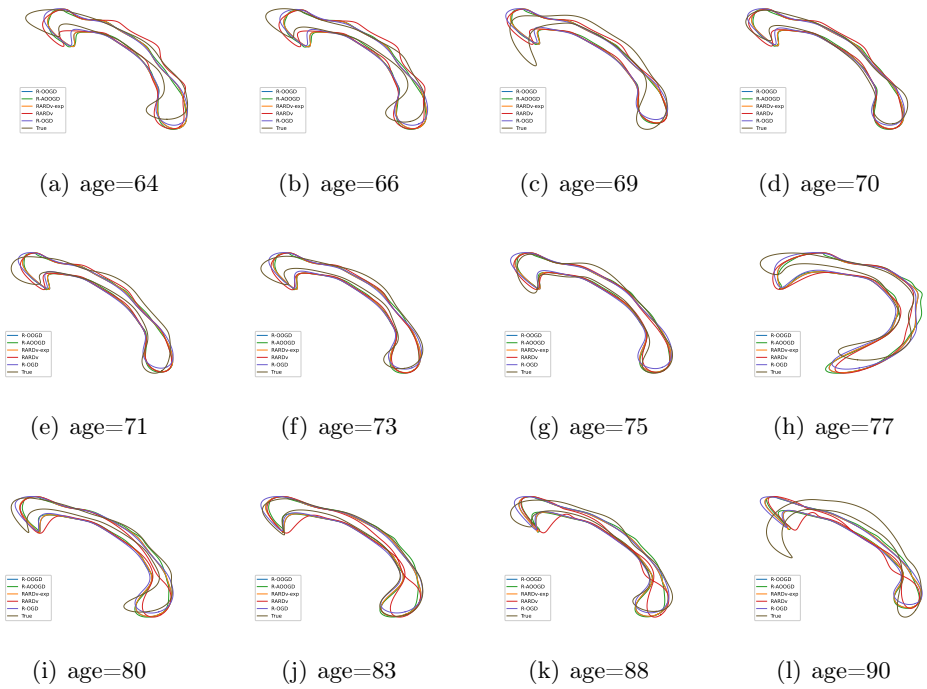
$$S_d^{++} := \{X \in \mathbb{R}^{d\times d}; X^T = X, X \succ 0\}$$

with affine-invariant metric

$$\langle U, V\rangle_x = tr(X^{-1}UX^{-1}V).$$

Since the logdet function is geodesic linear on S_d^{++} (Han et al., 2022), the quadratic game (7) is g-convex-concave with the g-strong convexity coefficient c_1. The NEs of the game (7) are (X^*, Y^*) where $\det(X^*) = \det(Y^*) = 1$.

Experimental setting We test the R-OGDA in the case when $d = 30$, $c_2 = 1$, and $c_1 \in \{0, 0.1, 1\}$. We check the R-OGDA with the step size $\eta = 0.5$ for $c_1 \in \{0, 0.1\}$ and $\eta = 0.2$ for $c_1 = 1$. We also compare our algorithm with the second-order Riemannian Hamitonian method (RHM) (Han et al., 2022), the Riemannian corrected extragradient method (RCEG) (Zhang et al., 2022), and the Riemannian gradient descent ascent algorithm (R-GDA) (Jordan et al., 2022) with the best-tuned step size.

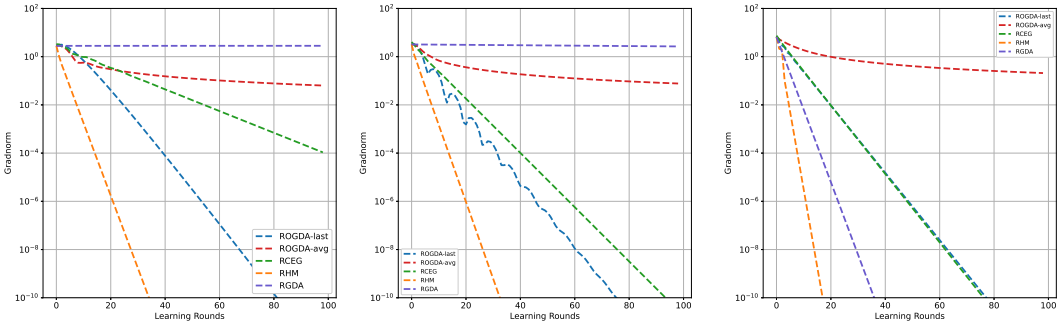


(a) age=64 (b) age=66 (c) age=69 (d) age=70

(e) age=71 (f) age=73 (g) age=75 (h) age=77

(i) age=80 (j) age=83 (k) age=88 (l) age=90

Figure 3: Algorithm performance on testing set of corpus callosum shapes



(a) $c_1 = 0, c_2 = 1$ (b) $c_1 = 0.1, c_2 = 1$ (c) $c_1 = 1, c_2 = 1$

Figure 4: Quadratic geodesic lodget games

Results We plot the norm of gradient $\|\nabla f(x_t, y_t)\|$ versus learning round t in Figure 4. From the results, all the algorithm converge to the NE. Among them, the second-order algorithm RHM performs the best. For first-order algorithms, our R-OGDA algorithm shows good performance in both g-convex-concave and g-strongly convex-strongly concave scenarios. In comparison, R-GDA performs well in g-strongly convex-strongly concave scenarios but does not converge in linear cases (i.e., $c = 0$). Similarly, RCEG shows slower convergence in linear scenarios and gradually improves with increasing g-strong convexity of the function, showing comparable convergence to our R-OGDA when $c = 1$.

6.4 Robust Geometry-Aware PCA

Robust geometry-aware principal component analysis (PCA) (Horev et al., 2016) is a dimensional reduction tool for SPD matrices. Given a set of SPD matrices $\{A_i\}_{i=1}^n$, the geometry-aware PCA aims to find a geometry mean A with the maximal data variance, which can be formulated as finding the NE in the following RZS game

$$\min_{A \in \mathcal{S}_d^{++}} \max_{X \in \mathbb{S}^d} X^T A X + \frac{\alpha}{n} \sum_{i}^{n} \| \log(A_i^{-\frac{1}{2}} A A_i^{-\frac{1}{2}}) \|_F, \tag{8}$$

where $\mathbb{S}^d$ is the d-dimensional unit sphere with the canonical metric $\langle U, V \rangle = U^T V$. The game (8) is g-strongly convex to A, but not g-concave to X. Hence, the game (8) is more challenging for our R-OGDA algorithm.

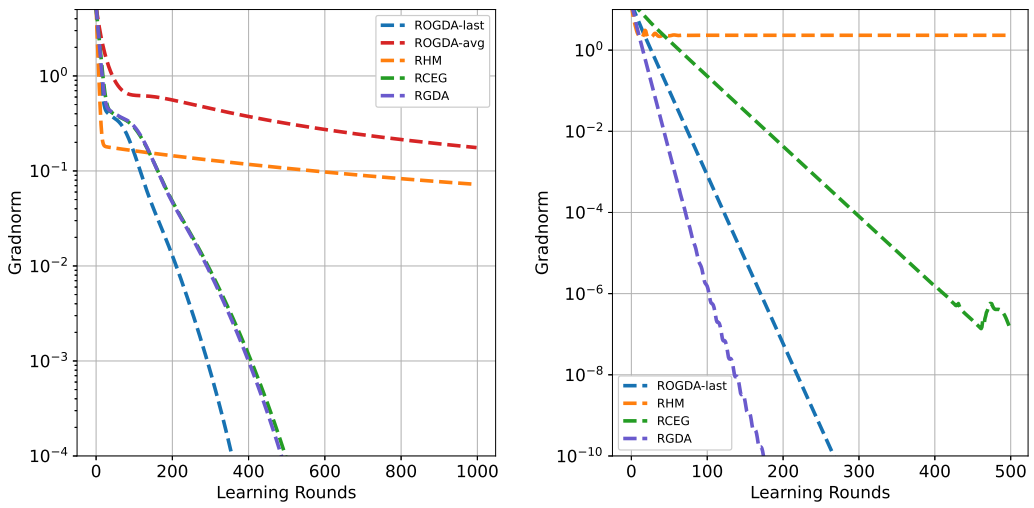
Experimental setting We run the experiment with a synthetic dataset $\{A_i\}_{i=1}^n$ and a real-world BCI dataset. The synthetic dataset is generated under conditions similar to those described in prior studies by Zhang et al. (2022) and Han et al. (2022), where the eigenvalues of A_i are in $[0.2, 4.5]$. The real-world dataset the *BCI competition dataset IV*[3] (Schlögl et al., 2005). The BCI competition dataset IV collected EEG signals from 59 channels electrodes of 5 subjects who executed left-hand, right-hand, foot and tongue movements. To preprocess the dataset, we follow the procedure outlined in Horev et al. (2016), which involves selecting 200 trials in the time interval from 0.5s to 2.5s, applying a band-pass filter to remove frequencies outside the 8–15Hz range and extracting covariant matrix into a 6×6 matrix. For numerical stability, we divide the covariant matrix by 200. For the synthetic dataset, we set the number of samples $n = 40$, the dimension $d = 50$, the regularization parameter $\alpha = 1$, and the step size $\eta = 0.07$. For the competition BCI dataset IV, we set $n = 200$, $d = 6$, $\alpha = 1$ and $\eta = 0.05$.

Results As shown in Figure 5, on the synthetic datase, the R-OGDA has the fastest convergence rate. For the real world dataset, R-GDA performs the best and the R-OGDA outperforms the RHM and RCEG. The performance in the g-convex-nonconcave RZS games illustrates the potential value of the R-OGDA.

7 Conclusion

In this paper, we have investigated the dynamic regret of Riemannian online optimization. We proposed R-OOGD and established a dynamic regret bound of $\mathcal{O}(P_T\sqrt{1 + V_T})$.

3. `https://www.bbci.de/competition/iv/download/`



(a) synthetic dataset (b) real-world dataset

Figure 5: Robust geometric-aware PCA

Additionally, we introduced R-AOOGD, a meta-expertalgorithm that averages R-OOGD algorithms with different step sizes, which further improved the dynamic regret bound to $\mathcal{O}(\sqrt{(1+P_T+V_T)(1+P_T)})$. We also applied the ROGD algorithm to Riemannian zero-sum games and achieved convergence rates of $\mathcal{O}(\frac{1}{T})$ for average-iterate, $\mathcal{O}(\frac{1}{\sqrt{T}})$ for best-iterate, and $\mathcal{O}(e^{-\rho t})$ for last-iterate, for smooth g-convex-concave and g-strongly convex-strongly concave games. Our results demonstrate the impact of Riemannian geometry on algorithm performance, and all our regret bounds and convergence rates match the corresponding Euclidean results.

One possible future direction of our work is to consider Riemannian dynamic regret in the bandit feedback setting, where the learner only receives the function value $\mathbf{f}_t(x_t)$ instead of the gradient $\nabla \mathbf{f}_t(x_t)$. We plan to design Riemannian optimistic bandit algorithms by incorporating Riemannian optimization techniques into existing bandit optimization methods. This will make the resulting algorithms more effective in settings where obtaining gradients is difficult or even impossible, such as reinforcement learning on Riemannian manifolds.

Acknowledgments and Disclosure of Funding

The authors would like to extend their heartfelt appreciation to Guanpu Chen from the School of Electrical Engineering and Computer Science at the KTH Royal Institute of Technology for his insightful discussions on Riemannian zero-sum games. This work is supported in part by the National Key Research and Development Program of China under No.2022YFA1004700 and Shanghai Municipal Science and Technology Major Project under No.2021SHZDZX0100, and in part by Australian Research Council under Grants DP190103615, LP210200473, and DP230101014. Additionally, Xi Wang would like to ac-

knowledge the financial support provided by the China Scholarships Council for visiting University of Sydney.

Appendix A. Definitions and Technical Lemmas

In the appendix, we introduce some prerequisite about Riemannian geometry for further analysis. In the rest of the appendix, we denote $\mathcal{X}(\mathcal{M})$ as the set of vector fields on the Riemannian manifold $\mathcal{M}$ and $C^\infty(\mathcal{M})$ as the class of infinitely differentiable functions on $\mathcal{M}$. A vector field X is equivalent to an operator on $C^\infty(\mathcal{M})$ via the directional derivative $X(\mathbf{f}) := \lim_{t\to 0} \frac{1}{t}(\mathbf{f}(\gamma(t)) - \mathbf{f}(t))$, where $\mathbf{f} \in C^\infty(\mathcal{M})$ and γ is a curve such that $\gamma(0) = p$ and $\dot\gamma(0) = X(p)$.

First, we recall some definitions on Riemannian manifolds.

Definition 9 (Lee, 2018)**.** *The gradient of a function* $\mathbf{f}$ *at the point* x *is defined as the tangent vector* $\nabla\mathbf{f}(x) \in T_x\mathcal{M}$ *such that* $\langle X(x), \nabla\mathbf{f}(x)\rangle = X(\mathbf{f})(x), \forall X \in \mathcal{X}(\mathcal{M})$.

Definition 10 (Lee, 2018)**.** *The hessian of a function* f *at the point* x *is defined as the bilinear operator* $\nabla^2\mathbf{f}(x) : T_x\mathcal{M} \times T_x\mathcal{M} \to \mathbb{R}$ *such that*

$$\nabla^2\mathbf{f}(X(x), Y(x))(x) = \langle \nabla_X \nabla\mathbf{f}(x), Y(x)\rangle, \forall X, Y \in \mathcal{X}(\mathcal{M}). \tag{9}$$

Definition 11 (Lee, 2018)**.** *A vector field* J *along a geodesic* $\gamma : [0,1] \to \mathcal{M}$ *is a Jacobi field if it satisfies:*

$$\langle \nabla_{\dot\gamma}\nabla_{\dot\gamma} J, W\rangle + R(\dot\gamma, J, \dot\gamma, W) = 0, \quad \forall W \in \mathcal{X}(\mathcal{M}).$$

Then we recall some properties of the covariant derivative $\nabla_X Y$, the curvature tensor $R(X, Y, W, Z)$ and the Jacobi field.

Lemma 12 (Carmo, 1992)**.** *The covariant derivative* $\nabla : \mathcal{X}(\mathcal{M}) \times \mathcal{X}(\mathcal{M}) \to \mathcal{X}(\mathcal{M})$ *satisfies the following properties:*

(i) $Z\langle X, Y\rangle = \langle \nabla_Z X, Y\rangle + \langle X, \nabla_Z Y\rangle, \quad \forall X, Y, Z \in \mathcal{X}(\mathcal{M})$;

(ii) $\nabla_X Y - \nabla_Y X = [X, Y]$, *where* $[X, Y] = XY - YX$ *is also a vector field on* $\mathcal{M}$.

Lemma 13 (Carmo, 1992)**.** *If* $\gamma : [0,1] \to \mathcal{M}$ *is a curve on* M *and* $z \in T_{\gamma(0)}\mathcal{M}$. *If we extend* z *to a vector field* $Z(t) \in T_{\gamma(t)}\mathcal{M}$ *on* γ *along the parallel transport* $\Gamma_{\gamma(t)}$, *then we have* $\nabla_{\dot\gamma(t)} Z(t) = 0$, $\forall t \in [0,1]$.

Lemma 14 (Andrews and Hopper, 2010)**.** *Denote the curvature tensor*

$$\langle \nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z - \nabla_{[X,Y]} Z, W\rangle = R(X, Y, W, Z),$$

then the following statement hold.

(i) $R(X, Y, W, Z)$ *is multilinear over* $C^\infty(\mathcal{M})$ *i.e.,*

$$R(f_1 X, f_2 Y, f_3 W, f_4 Z) = f_1 f_2 f_3 f_4 R(X, Y, W, Z), \quad \forall f_1, f_2, f_3, f_4 \in C^\infty(\mathcal{M}).$$

(ii) The curvature tensor is determined by the its values on $2-$*dimensional surfaces* $\mathcal{K}(X,Y) = R(X,Y,X,Y)$*, i.e.,*

$$\begin{aligned} 6R(X,Y,W,Z) = &\mathcal{K}(X+W,Y+Z) - \mathcal{K}(X,Y+Z) + \mathcal{K}(W,X+Z) \\ &+ \mathcal{K}(Y+W,X+Z) - \mathcal{K}(Y,X+Z) + \mathcal{K}(W,X+Z) \\ &- \mathcal{K}(X+W,Y) + \mathcal{K}(X,Y) + \mathcal{K}(W,Y) \\ &- \mathcal{K}(X+W,Z) + \mathcal{K}(X,Z) + \mathcal{K}(W,Z) \\ &+ \mathcal{K}(Y+W,X) + \mathcal{K}(Y,X) - \mathcal{K}(W,X) \\ &+ \mathcal{K}(Y+W,Z) - \mathcal{K}(Y,Z) - \mathcal{K}(W,Z). \end{aligned}$$

Lemma 15 (Jacobi comparison theorem,Lee, 2018)**.** *Suppose* $\mathcal{M}$ *is a Riemannian manifold and* $\gamma : [0,b] \to \mathcal{M}$ *is a geodesic with* $\|\dot{\gamma}(0)\| = 1$*.* J *is a Jacobi field along* γ*.*

(i) If all sectional curvatures of $\mathcal{M}$ *are upper bounded by a constant* K*, then denote*

$$\mathbf{S}(K,t) = \begin{cases} \frac{\sin(\sqrt{K}t)}{\sqrt{K}} & K > 0; \\ t & K \leq 0. \end{cases}$$

Then,

$$\|J(\gamma(t))\| \geq \mathbf{S}(K,t)\|\nabla_{\dot{\gamma}} J(\gamma(0))\|$$

for all $t \in [0, b_1]$*, where* $b_1 = b$ *if* $K \leq 0$ *and* $b_1 = \min(\frac{\pi}{\sqrt{K}}, b)$ *if* $K \geq 0$*.*

(ii) If all sectional curvatures of $\mathcal{M}$ *are lower bounded by a constant* κ*, then denote*

$$\mathbf{s}(\kappa,t) = \begin{cases} \frac{\sinh(\sqrt{-\kappa}t)}{\sqrt{-\kappa}} & \kappa < 0; \\ t & \kappa \geq 0. \end{cases}$$

Then,

$$\|J(\gamma(t))\| \leq \mathbf{s}(\kappa,t)\|\nabla_{\dot{\gamma}} J(\gamma(0))\|$$

for all $t \in [0,b]$*.*

Furthermore, we consider two comparison inequalities, which served as the law of consine over Riemannian manifolds.

Lemma 16 (Zhang and Sra, 2016)**.** *Let* $\mathcal{M}$ *be a Riemannian manifold with all sectional curvatures lower bounded by* κ*. Denote*

$$\zeta(\kappa, D) = \begin{cases} \frac{\sqrt{-\kappa}D}{\tanh(\sqrt{-\kappa}D)} & \kappa < 0; \\ 1 & \kappa \geq 0. \end{cases}$$

Then for a geodesic triangle $\triangle ABC$*, we have*

$$2\langle \exp_A^{-1} C, \exp_A^{-1} B\rangle \leq d^2(A,B) + \zeta(\kappa, d(A,B)) d^2(A,C) - d^2(B,C).$$

Lemma 17 (Alimisis et al., 2021). *Let $\mathcal{M}$ be a Riemannian manifold with all sectional curvatures upper bounded by K. Denote*

$$\sigma(K,D)=\begin{cases}\frac{\sqrt{K}D}{\tan(\sqrt{K}D)} & K>0;\\ 1 & K\leq 0,\end{cases} \quad \textit{and} \quad D(K)=\begin{cases}\infty & K\leq 0;\\ \frac{\pi}{2\sqrt{K}} & K>0.\end{cases}$$

If a geodesic triangle $\triangle ABC$ has diameter less than $D(K)$, then we have

$$2\langle \exp_A^{-1} C, \exp_A^{-1} B\rangle \geq d^2(A,B)+\sigma(K,d(A,B))d^2(A,C)-d^2(B,C).$$

In addtion, Ahn and Sra (2020) demonstrate bi-Lipschitzness of the exponential map on positive curved space.

Lemma 18 (Ahn and Sra, 2020). *Let A,B,C be points on Riemannian manifold $\mathcal{M}$ with sectional curvatures upper bounded by $K\geq 0$. If $d(A,C)\leq \frac{\pi}{2\sqrt{K}}$, then*

$$\|\exp_A^{-1} B-\exp_A^{-1} C\|\leq \sqrt{(1+2Kd^2(A,B))}d(B,C)\leq \frac{2}{\sqrt{\sigma(K,d(A,B))}}d(B,C).$$

Finally, we introduce a technique that bounds the metric distortion by parallel transport.

Lemma 19 (Alimisis et al., 2021). *Let $\mathcal{M}$ be a Riemannian manifold with sectional curvature lower bounded by κ and upper bounded by K. If a geodesic triangle $\triangle ABC$ admits a diameter less than $D(K)$, then there exists a point p in the edge AC such that*

$$\|\exp_A^{-1} B-\Gamma_C^A \exp_C^{-1} B\|=-\Gamma_p^A\big(\nabla^2(-\frac{1}{2}d^2(C,p))\big)\Gamma_A^p \exp_A^{-1} C,$$

where $\big(\nabla^2(-\frac{1}{2}d^2(C,p))\big)$ is the hessian of the function $-\frac{1}{2}d^2(C,\cdot)$ at point p.

We denote $H_{A,p}^c$ as the operator $-\Gamma_p^A\big(\nabla^2(-\frac{1}{2}d^2(C,p))\big)\Gamma_A^p$. From the hessian comparison theorem (Lee, 2018; Alimisis et al., 2020), we know that all the eigenvalues of $H_{A,p}^c$ are in the range $[\sigma(K,d(C,p)),\zeta(\kappa,d(C,p))]$. Since

$$\max\big\{\zeta(\kappa,d(C,p))-1,1-\sigma(K,d(C,p))\big\}\leq \max(|\kappa|,|K|)d^2(C,p),$$

we have the following corollary.

Corollary 20. *Let $\mathcal{M}$ be a Riemannian manifold with sectional curvature lower bounded by κ and upper bounded by K. If a geodesic triangle $\triangle ABC$ has diameter $D\leq D(K)$, then there exists a point p lying in the edge AC such that*

(i) $\|\exp_A^{-1} B-\Gamma_C^A \exp_C^{-1} B\|\leq \zeta(\kappa,D)\|\exp_A^{-1} C\|$;

(ii) $\|\exp_A^{-1} B-\Gamma_C^A \exp_C^{-1} B-\exp_A^{-1} C\|\leq d^2(C,p)\max\{|\kappa|,|K|\}\|\exp_A^{-1} C\|$.

Appendix B. Proofs in Section 4

Proof of Theorem 1 To ease the notation, we define $v_t = \sup_{x\in\mathcal{K}} \|\mathbf{f}_t(x) - \mathbf{f}_{t-1}(x)\|^2$ and $\nabla_t = \nabla\mathbf{f}_t(x_t)$. By the g-convexity of $\mathbf{f}_{t+1}$, we have

$$\mathbf{f}_{t+1}(x_{t+1}) - \mathbf{f}_{t+1}(u_{t+1}) \leq \langle \exp_{x_{t+1}}^{-1} u_{t+1}, -\nabla_{t+1}\rangle. \tag{10}$$

Applying Lemma 17 in the geodesic triangle $\triangle x_{t+1}u_{t+1}x_t$ we have,

$$2\langle \exp_{x_{t+1}}^{-1} u_{t+1}, \exp_{x_{t+1}}^{-1} x_t \geq d^2(x_{t+1}, u_{t+1}) - d^2(x_t, u_{t+1}) + \sigma_0 d^2(x_{t+1}, x_t).$$

Notice from Algorithm 1, we have $x_{t+1} = \exp_{x_t}(-2\eta\nabla_t + \eta\Gamma_{x_{t-1}}^{x_t}\nabla_{t-1})$. This give us

$$\exp_{x_{t+1}}^{-1} x_t = \eta(\Gamma_{x_t}^{x_{t+1}}(2\nabla_t - \Gamma_{x_{t-1}}^{x_t}\nabla_{t-1})).$$

Thus, we have

$$\begin{aligned} 0 \leq \langle \exp_{x_{t+1}}^{-1} u_{t+1}, &\Gamma_{x_t}^{x_{t+1}}(2\nabla_t - \Gamma_{x_{t-1}}^{x_t}\nabla_{t-1})\rangle - \\ &\frac{1}{2\eta}(d^2(x_{t+1}, u_{t+1}) - d^2(x_t, u_{t+1})) - \frac{\sigma_0}{2\eta} d^2(x_{t+1}, x_t). \end{aligned} \tag{11}$$

Combining (10) and (11), we have

$$\begin{aligned} \mathbf{f}_{t+1}(x_{t+1}) - \mathbf{f}_{t+1}(u_{t+1}) &\leq \langle \exp_{x_{t+1}}^{-1} u_{t+1}, -\nabla_{t+1}\rangle + \langle \exp_{x_{t+1}}^{-1} u_{t+1}, \Gamma_{x_t}^{x_{t+1}}(2\nabla_t - \Gamma_{x_{t-1}}^{x_t}\nabla_{t-1})\rangle \\ &\quad - \frac{1}{2\eta}(d^2(x_{t+1}, u_{t+1}) - d^2(x_t, u_{t+1})) - \frac{\sigma_0}{2\eta} d^2(x_{t+1}, x_t) \\ &= \langle \exp_{x_{t+1}}^{-1} u_{t+1}, -\nabla_{t+1} + \Gamma_{x_t}^{x_{t+1}}\nabla_t\rangle \\ &\quad - \langle \exp_{x_{t+1}}^{-1} u_{t+1}, \Gamma_{x_t}^{x_{t+1}}(\nabla_t - \Gamma_{x_{t-1}}^{x_t}\nabla_{t-1})\rangle \\ &\quad - \frac{1}{2\eta}(d^2(x_{t+1}, u_{t+1}) - d^2(x_t, u_{t+1})) - \frac{\sigma_0}{2\eta} d^2(x_{t+1}, x_t). \end{aligned} \tag{12}$$

Considering the term $-\langle \exp_{x_{t+1}}^{-1} u_{t+1}, \Gamma_{x_t}^{x_{t+1}}(\nabla_t - \Gamma_{x_{t-1}}^{x_t}\nabla_{t-1})\rangle - \frac{\sigma}{2\eta} d^2(x_t, x_{t+1})$, we have

$$\begin{aligned} &- \langle \exp_{x_{t+1}}^{-1} u_{t+1}, \Gamma_{x_t}^{x_{t+1}}(\nabla_t - \Gamma_{x_{t-1}}^{x_t}\nabla_{t-1})\rangle - \frac{\sigma}{2\eta} d^2(x_t, x_{t+1}) \\ &\leq \langle - \exp_{x_{t+1}}^{-1} u_{t+1} + \Gamma_{x_t}^{x_{t+1}} \exp_{x_t}^{-1} u_{t+1}, \Gamma_{x_t}^{x_{t+1}}(\nabla_t - \Gamma_{x_{t-1}}^{x_t}\nabla_{t-1})\rangle \\ &\quad + \langle - \exp_{x_t}^{-1} u_{t+1} + \exp_{x_t}^{-1} u_t, \nabla_t - \Gamma_{x_{t-1}}^{x_t}\nabla_{t-1}\rangle \\ &\quad - \langle \exp_{x_t}^{-1} u_t, \nabla_t - \Gamma_{x_{t-1}}^{x_t}\nabla_{t-1}\rangle - \frac{\sigma_0}{2\eta} d^2(x_{t+1}, x_t). \end{aligned} \tag{13}$$

By Lemma 20, we have

$$\begin{aligned} &\langle - \exp_{x_{t+1}}^{-1} u_{t+1} + \Gamma_{x_t}^{x_{t+1}} \exp_{x_t}^{-1} u_{t+1}, \Gamma_{x_t}^{x_{t+1}}(\nabla_t - \Gamma_{x_{t-1}}^{x_t}\nabla_{t-1})\rangle \\ &\leq \zeta_0 d(x_{t+1}, x_t)\|\nabla_t - \Gamma_{x_{t-1}}^{x_t}\nabla_{t-1}\| \\ &\leq \frac{\zeta_0^2}{\sigma_0}\eta\|\nabla_t - \Gamma_{x_{t-1}}^{x_t}\nabla_{t-1}\|^2 + \frac{\sigma_0}{4\eta} d^2(x_t, x_{t+1}). \end{aligned} \tag{14}$$

The latter inequality is due to the Young's inequality $\langle a, b\rangle \leq \frac{\alpha\|a\|^2}{2} + \frac{\|b\|^2}{2\alpha}$. Also, By Lemma 18, we have

$$\begin{aligned}
&\langle -\exp_{x_t}^{-1} u_{t+1} + \exp_{x_t}^{-1} u_t, \nabla_t - \Gamma_{x_{t-1}}^{x_t}\nabla_{t-1}\rangle \\
&\leq \frac{2}{\sqrt{\sigma}_0} d(u_t, u_{t+1})\|\nabla_t - \Gamma_{x_{t-1}}^{x_t}\nabla_{t-1}\| \\
&\leq \frac{1}{\sigma_0}\eta\|\nabla_t - \Gamma_{x_{t-1}}^{x_t}\nabla_{t-1}\|^2 + \frac{d^2(u_{t+1}, u_t)}{\eta} \\
&\leq \frac{1}{\sigma_0}\eta\|\nabla_t - \Gamma_{x_{t-1}}^{x_t}\nabla_{t-1}\|^2 + \frac{d(u_{t+1}, u_t)D}{\eta}. \qquad (15)
\end{aligned}$$

Since $\mathbf{f}_t$ is g-L-smooth, we have

$$\begin{aligned}
\|\nabla_t - \Gamma_{x_{t-1}}^{x_t}\nabla_{t-1}\|^2 &\leq 2\|\nabla\mathbf{f}_t(x_t) - \nabla\mathbf{f}_{t-1}(x_t)\|^2 + 2\|\nabla\mathbf{f}_{t-1}(x_t) - \Gamma_{x_{t-1}}^{x_t}\nabla_{t-1}\|^2 \\
&\leq 2v_t + 2L^2 d^2(x_t, x_{t-1}). \qquad (16)
\end{aligned}$$

Putting (13)-(16) together, we have

$$\begin{aligned}
\mathbf{f}_{t+1}(x_{t+1}) - \mathbf{f}_{t+1}(u_{t+1}) \leq &\langle \exp_{x_{t+1}}^{-1} u_{t+1}, -\nabla_{t+1} + \Gamma_{x_t}^{x_{t+1}}\nabla_t\rangle - \langle \exp_{x_t}^{-1} u_t, -\nabla_t + \Gamma_{x_{t-1}}^{x_t}\nabla_{t-1}\rangle \\
&+ \frac{1}{2\eta}(d^2(x_t, u_{t+1}) - d^2(x_{t+1}, u_{t+1})) \\
&+ \frac{2(1+\zeta_0^2)}{\sigma_0}\eta v_t + \frac{2L^2(1+\zeta_0^2)}{\sigma_0}\eta d(x_t, x_{t-1}) \\
&+ \frac{d(u_t, u_{t+1})D}{\eta} - \frac{\sigma_0}{4\eta}d^2(x_t, x_{t+1}). \qquad (17)
\end{aligned}$$

Summing (17) from $t = 0$ to $T-1$ together, we have

$$\begin{aligned}
\mathsf{Reg}_{\mathsf{D}}(u_1, u_2, \ldots, u_T) &= \sum_{t=0}^{T-1} \mathbf{f}_{t+1}(x_{t+1}) - \mathbf{f}_{t+1}(u_{t+1}) \\
&\leq \sum_{t=0}^{T-1} \langle \exp_{x_{t+1}}^{-1} u_{t+1}, -\nabla_{t+1} + \Gamma_{x_t}^{x_{t+1}}\nabla_t\rangle \\
&\quad - \sum_{t=0}^{T-1} \langle \exp_{x_t}^{-1} u_t, -\nabla_t + \Gamma_{x_{t-1}}^{x_t}\nabla_{t-1}\rangle \\
&\quad + \sum_{t=0}^{T-1} \frac{1}{2\eta} d^2(x_t, u_{t+1}) - \sum_{t=0}^{T-1} \frac{1}{2\eta} d^2(x_{t+1}, u_{t+1}) \\
&\quad + \sum_{t=0}^{T-1} \frac{2L^2(1+\zeta_0^2)}{\sigma_0}\eta d(x_t, x_{t-1}) - \sum_{t=0}^{T-1} \frac{\sigma_0}{4\eta} d^2(x_t, x_{t+1}).
\end{aligned}$$

Rearranging the summation, we can obtain

$$\begin{aligned}
\mathsf{Reg}_{\mathsf{D}}(u_1, u_2, \ldots, u_T) &\leq \frac{2(1+\zeta_0^2)}{\sigma_0}\eta V_T + \frac{DP_T}{\eta} \\
&\leq \sum_{t=1}^{T}\langle \exp_{x_t}^{-1} u_t, -\nabla_t + \Gamma_{x_{t-1}}^{x_t}\nabla_{t-1}\rangle - \sum_{t=0}^{T-1}\langle \exp_{x_t}^{-1} u_t, -\nabla_t + \Gamma_{x_{t-1}}^{x_t}\nabla_{t-1}\rangle \\
&\quad + \sum_{t=0}^{T-1}\frac{1}{2\eta}d^2(x_t, u_{t+1}) - \sum_{t=1}^{T}\frac{1}{2\eta}d^2(x_t, u_t) \\
&\quad + \sum_{t=0}^{T-1}\frac{4L^2\zeta_0^2}{\sigma_0}\eta d(x_t, x_{t-1}) - \sum_{t=1}^{T}\frac{\sigma_0}{4\eta}d^2(x_t, x_{t-1}) \\
&\quad + \frac{4\zeta_0^2}{\sigma_0}\eta V_T + \frac{DP_T}{\eta}.
\end{aligned}$$

Since $\eta \leq \frac{\sigma_0}{4\zeta_0 L}$, we have $\frac{4L^2\zeta_0^2\eta}{\sigma_0} \leq \frac{\sigma_0}{4\eta}$. Therefore, we have

$$\begin{aligned}
\mathsf{Reg}_{\mathsf{D}}(u_1, u_2, \ldots, u_T) &= \sum_{t=0}^{T-1}\mathbf{f}_{t+1}(x_{t+1}) - \mathbf{f}_{t+1}(u_{t+1}) \\
&\leq \langle \exp_{x_T}^{-1} u_T, -\nabla_T + \Gamma_{x_{T-1}}^{x_T}\nabla_{T-1}\rangle - \langle \exp_{x_0}^{-1} u_0, -\nabla_0 + \Gamma_{x_{-1}}^{x_0}\nabla_{-1}\rangle \\
&\quad + \frac{1}{2\eta}d^2(x_0, u_1) - \frac{1}{2\eta}d^2(x_T, u_T) \\
&\quad + \frac{4L^2\zeta_0^2}{\sigma_0}\eta d(x_0, x_{-1}) - \frac{\sigma_0}{4\eta}d^2(x_T, x_{T-1}) \\
&\quad + \frac{4\zeta_0^2}{\sigma_0}\eta V_T + \frac{2DP_T}{\eta}.
\end{aligned}$$

As $x_{-1} = x_0 = x_1$, we can see that $\nabla_0 = \nabla_{-1}$ and $d(x_0, x_{-1}) = 0$. In this way, we can see that

$$\begin{aligned}
\mathsf{Reg}_{\mathsf{D}}(u_1, u_2, \ldots, u_T) &= \sum_{t=0}^{T-1}\mathbf{f}_{t+1}(x_{t+1}) - \mathbf{f}_{t+1}(u_{t+1}) \\
&\leq \langle \exp_{x_T}^{-1} u_T, -\nabla_T + \Gamma_{x_{T-1}}^{x_T}\nabla_{T-1}\rangle + \frac{1}{2\eta}d^2(x_0, u_1) + \frac{4\zeta_0^2}{\sigma_0}\eta V_T + \frac{2DP_T}{\eta} \\
&\leq 2DG + \frac{D^2}{2\eta} + \frac{4\zeta_0^2}{\sigma_0}\eta V_T + \frac{2DP_T}{\eta} \\
&\leq \frac{D^2 + 2DP_T}{\eta} + \frac{4\zeta_0^2}{\sigma_0}\eta(V_T + G^2),
\end{aligned}$$

which completes our proof. ■

Proof of Theorem 3 We follow the idea to treat the dynamic regret by the meta-regret and expert-regret as in the work by Hu et al. (2023). For any $i \leq N$, it holds that

$$\sum_{t=1}^{T} \mathbf{f}_t(x_t) - \mathbf{f}_t(u_t) = \underbrace{\sum_{t=1}^{T} \mathbf{f}_t(x_t) - \mathbf{f}_t(x_{i,t})}_{\texttt{meta-regret}} + \underbrace{\sum_{t=1}^{T} \mathbf{f}_t(x_{t,i}) - \mathbf{f}_t(u_t)}_{\texttt{expert-regret}}.$$

Based on Theorem 2 of Hu et al. (2023), we obtain

$$\begin{aligned}
\texttt{meta-regret} &= \sum_{t=1}^{T} \mathbf{f}_t(x_t) - \mathbf{f}_t(x_{i,t}) \\
&\leq \frac{2+\ln N}{\beta} + 3D_0^2\beta(V_T + G^2) + \\
&\quad + \sum_{t=2}^{T} \big(3\beta(D_0^4L^2 + D_0^2G^2\zeta_0^2) - \frac{1}{4\beta}\big)\|w_t - w_{t-1}\|_1^2 \\
&\leq \max\Big(2\sqrt{3D_0^2(V_T+G^2)(2+\ln N)}, 2(2+\ln N)\sqrt{12(D_0^4L^2 + D_0^2G^2\zeta_0^2)}\Big).
\end{aligned}$$

Moreover, according to the dynamic regret in Theorem 1, for any index i with step size η_i, we have:

$$\begin{aligned}
\texttt{expert-regret} &\leq \sum_{t=1}^{T} \mathbf{f}_t(x_{t,i}) - \mathbf{f}_t(u_t) \\
&\leq \frac{D_0^2 + 2D_0P_T}{\eta_i} + \frac{4\zeta_0^2}{\sigma_0}\eta_i(V_T + G^2).
\end{aligned}$$

Our step size pool $\mathcal{H} = \Big\{\eta_i = 2^{i-1}\sqrt{\frac{D_0^2}{16\zeta_0^2G^2T}}\Big\}$ ensures that

$$\begin{cases}
\min_{\mathcal{H}} = \sqrt{\frac{D_0^2}{16\zeta_0^2G^2T}} \leq \sqrt{\frac{D_0^2+2D_0P_T}{4\zeta_0^2(G^2+V_T)}}, \\
\max_{\mathcal{H}} \geq \frac{\sigma_0}{\zeta_0 L}.
\end{cases}$$

So for the optimal step size $\eta^* = \min(\sqrt{\frac{D_0^2+P_T}{4\zeta_0^2(G^2+V_T)}}, \frac{\sigma_0}{\zeta_0 L})$, there exists $i^* \leq N$ such that $\eta_{i^*} \leq \eta^* \leq 2\eta_{i^*}$. Taking $i = i^*$, we have

$$\begin{aligned}
\texttt{expert}-\texttt{regret} &\leq \sum_{t=1}^{T} \mathbf{f}_t(x_{t,i}) - \mathbf{f}_t(u_t) \\
&\leq \frac{D_0^2 + 2D_0P_T}{\eta_{i^*}} + \frac{4\zeta_0^2}{\sigma_0}\eta_{i^*}(V_T + G^2) \\
&\leq \frac{4\zeta_0^2}{\sigma_0}(V_T + G^2)\sqrt{\frac{D_0^2 + 2D_0P_T}{4\zeta_0^2(G^2 + V_T)}} \\
&\quad + (D_0^2 + 2D_0P_T)(\sqrt{\frac{4\zeta_0^2(G^2 + V_T)}{D_0^2 + 2D_0P_T}} + \frac{4\zeta_0 L}{\sigma_0}) \\
&\leq \frac{4\zeta_0}{\sqrt{\sigma_0}}\sqrt{(D_0^2 + 2D_0P_T)((V_T + G^2))} + (D_0^2 + 2D_0P_T)\frac{4\zeta_0 L}{\sigma_0}.
\end{aligned}$$

Combining meta-regret and expert-regret together, we finally get

$$\begin{aligned}
\sum_{t=1}^{T} \mathbf{f}_t(x_t) - \mathbf{f}_t(u_t) &\leq \max\left(2\sqrt{3D_0^2(V_T + G^2)(2 + \ln N)}, 2(2 + \ln N)\sqrt{12(D_0^4L^2 + D_0^2G^2\zeta_0^2)}\right) \\
&\quad + \frac{4\zeta_0}{\sqrt{\sigma_0}}\sqrt{(D_0^2 + 2D_0P_T)((V_T + G^2))} + (D_0^2 + 2D_0P_T)\frac{4\zeta_0 L}{\sigma_0} \\
&= \mathcal{O}(\sqrt{((V_T + 1)\ln N)\ln N}) + \frac{\zeta_0}{\sqrt{\sigma_0}}bo(\sqrt{(1 + P_T + V_T)(1 + P_T)}) \\
&= \frac{\zeta_0}{\sqrt{\sigma_0}}\mathcal{O}(\sqrt{(1 + P_T + V_T)(1 + P_T)}),
\end{aligned}$$

which completes our proof. ■

Appendix C. Computing the Corrected R-OOGD

We first recall the corrected version of ROGD

$$\begin{cases} x_{t+1} = \exp_{x_t}(-2\eta\nabla_t + \exp_{x_t}^{-1}\hat{x}_t) \\ \hat{x}_t = \exp_{x_{t-1}}(-\eta\nabla_{t-1} + \exp_{x_{t-1}}^{-1}\hat{x}_{t-1}). \end{cases}$$

To analyse the regret bound of the corrected version of R-OOGD, we follow (11) and get

$$0 \leq \langle \exp_{x_{t+1}}^{-1} x, \Gamma_{x_{t+1}}^{x_t}(-2\eta\nabla_t + \exp_{x_t}^{-1}\hat{x}_t)\rangle + \frac{1}{2\eta}(d^2(x_t, x) - d^2(x_{t+1}, x)) - \frac{\sigma_0}{2\eta}d^2(x_{t+1}, x_t).$$

Thus, by g-convexity we have,

$$\begin{aligned}
\mathbf{f}_{t+1}(x_{t+1}) - \mathbf{f}_{t+1}(x) &\leq \langle \exp_{x_{t+1}}^{-1} x, -\nabla_{t+1} \rangle \\
&\leq \langle \exp_{x_{t+1}}^{-1} x, -\nabla_{t+1} + \Gamma_{x_t}^{x_{t+1}}(-2\eta\nabla_t + \exp_{x_t}^{-1} \hat{x}_t) \rangle \\
&\quad + \frac{1}{2\eta}(d^2(x_t, x) - d^2(x_{t+1}, x)) + \frac{\sigma_0}{2\eta} d^2(x_{t+1}, x_t) \\
&= \langle \exp_{x_{t+1}}^{-1} x, \Gamma_{x_t}^{x_{t+1}}(\nabla_t - \Gamma_{x_{t-1}}^{x_t} \nabla_{t-1}) \rangle \\
&\quad - \langle \exp_{x_{t+1}}^{-1} x, \nabla_{t+1} - \Gamma_{x_t}^{x_{t+1}} \nabla_t \rangle \\
&\quad + \frac{1}{2\eta}(d^2(x_t, x) - d^2(x_{t+1}, x)) - \frac{\sigma_0}{2\eta} d^2(x_{t+1}, x_t) \\
&\quad + \langle \exp_{x_{t+1}}^{-1} x, -\Gamma_{x_t}^{x_{t+1}} \big(\Gamma_{x_{t-1}}^{x_t} \eta \nabla_{t-1} + \exp_{x_t}^{-1} \hat{x}_t \big) \rangle. \qquad (18)
\end{aligned}$$

The expression (18) follows from the proof of Theorem 1 except the term

$$\langle \exp_{x_{t+1}}^{-1} x, -\Gamma_{x_t}^{x_{t+1}} \big(\Gamma_{x_{t-1}}^{x_t} \eta \nabla_{t-1} + \exp_{x_t}^{-1} \hat{x}_t \big) \rangle.$$

Applying Corollary 20 in the geodesic triangle $\triangle x_{t-1} x_t \hat{x}_t$, we have

$$\begin{aligned}
&\| \exp_{x_{t-1}}^{-1} \hat{x}_t - \exp_{x_{t-1}}^{-1} x_t - \Gamma_{x_t}^{x_{t-1}} \exp_{x_t}^{-1} \hat{x}_t \| \\
&= \| 2\eta\nabla_{t-1} - \exp_{x_{t-1}}^{-1} \hat{x}_{t-1} - \eta\nabla_{t-1} + \exp_{x_{t-1}}^{-1} \hat{x}_{t-1} - \Gamma_{x_t}^{x_{t-1}} \exp_{x_t}^{-1} \hat{x}_t \| \\
&\leq d^2(\hat{x}_t, p) K_m \| \exp_{x_{t-1}}^{-1} x_t \|,
\end{aligned}$$

where p lies in the geodesic $x_{t-1} x_t$. Since

$$\begin{aligned}
d(\hat{x}_t, p) &\leq d(\hat{x}_t, x_{t-1}) + d(x_{t-1}, x_t) \\
&\leq \| -2\eta\nabla_{t-1} + \eta\nabla_{t-2} - \eta\nabla_{t-2} + \exp_{x_{t-1}}^{-1} \hat{x}_{t-1} \| \\
&\quad + \| -\eta\nabla_{t-1} + \eta\nabla_{t-2} - \eta\nabla_{t-2} + \exp_{x_{t-1}}^{-1} \hat{x}_{t-1} \| \\
&\leq 5\eta G + 2A_{t-1},
\end{aligned}$$

and

$$\| \exp_{x_{t-1}}^{-1} x_t \| \leq 3\eta G + A_{t-1},$$

we have:

$$A_t \leq K_m (5\eta G + 2A_{t-1})^2 (3\eta G + A_{t-1}), \qquad (19)$$

indicating that the distortion in iteration $t-1$ evolves and accumulates in the distortion in iteration t. In the worst case scenario where $\eta G = 0.1$ and $K_m = 1$, the equality always holds. We observe that $A_t \to \infty$, which implies that the corrected ROGD fails to achieve sublinear static regret.

Appendix D. Proof of Theorem 5

We first prove Lemma 4.

Proof of Lemma 4 We prove by induction on $z_t = (x_t, y_t)$. The base case $d(z_0, z^*) \leq D_1 \leq 2D_1$ is straightforwardly hold. Then we assume that there exists a K_0 such that all $d(z_t, z^*) \leq 2D_1$ holds for all $t \leq K_0$. Then we carry out induction step on $K_0 + 1$. The distance $d(z_{K_0+1}, z^*)$ can be first bounded as

$$d(z_{K_0+1}, z^*) \leq d(z_{K_0}, z^*) + d(z_{K_0}, z_{K_0+1}) \leq 2D_1 + 3\eta G \leq 3D_1.$$

So, by setting $D_0 = 3D_1$, $\mathbf{f}_t(x) = \mathbf{f}(x, y_t)$ and plugging in $u_1 = u_2 = \cdots = u_n = x^*$ in the analysis of Theorem 1, the R-OOGD for player-$\mathtt{X}$ holds for all $t \leq K_0$

$$\begin{aligned}
&\langle \exp_{x_{t+1}}^{-1} x^*, -\nabla_x \mathbf{f}(x_{t+1}, y_{t+1}) \rangle \\
&\leq \langle \exp_{x_{t+1}}^{-1} x^*, \big(\nabla_x \mathbf{f}(x_{t+1}, y_{t+1}) - \Gamma_{x_t}^{x_{t+1}} \nabla_y \mathbf{f}(x_t, y_t) - \big(\nabla_x \mathbf{f}(x_t, y_t)\big) - \Gamma_{x_{t-1}}^{x_t} \mathbf{f}(x_{t-1}, y_{t-1})\big) \rangle \\
&\quad + \frac{1}{2\eta}(d^2(x_t, x^*) - d^2(x_{t+1}, x^*)) - \frac{\sigma_1}{2\eta} d^2(x_t, x_{t+1}) \\
&\leq \langle \exp_{x_{t+1}}^{-1} x^*, \nabla_x \mathbf{f}(x_{t+1}, y_{t+1}) - \Gamma_{x_t}^{x_{t+1}} \mathbf{f}(x_t, y_t) \rangle \\
&\quad - \langle \exp_{x_t}^{-1} x^*, \nabla_x \mathbf{f}(x_t, y_t) - \Gamma_{x_{t-1}}^{x_t} \mathbf{f}(x_{t-1}, y_{t-1}) \rangle + \frac{1}{2\eta}(d^2(x_t, x^*) - d^2(x_{t+1}, x^*)) \\
&\quad - \frac{\sigma_1}{2\eta} d^2(x_t, x_{t+1}) + \zeta_1 \|\nabla_x \mathbf{f}(x_t, y_t) - \Gamma_{x_{t-1}}^{x_t} \mathbf{f}(x_{t-1}, y_{t-1})\| d(x_t, x_{t+1}).
\end{aligned} \tag{20}$$

Similarly, the player-$\mathtt{Y}$ holds for all $t \leq K_0$

$$\begin{aligned}
&\langle \exp_{y_{t+1}}^{-1} y^*, \nabla_y \mathbf{f}(x_{t+1}, y_{t+1}) \rangle \\
&\leq \langle \exp_{y_{t+1}}^{-1} y^*, \nabla_y \mathbf{f}(x_{t+1}, y_{t+1}) - \Gamma_{y_t}^{y_{t+1}} \nabla_y \mathbf{f}(x_t, y_t) \rangle \\
&\quad - \langle \exp_{y_t}^{-1} y^*, \nabla_y \mathbf{f}(x_t, y_t) - \Gamma_{y_{t-1}}^{y_t} \nabla_y \mathbf{f}(x_{t-1}, y_{t-1}) \rangle + \frac{1}{2\eta}(d^2(y_t, y^*) - d^2(y_{t+1}, y^*)) \\
&\quad - \frac{\sigma_1}{2\eta} d^2(y_t, y_{t+1}) + \zeta_1 \|\nabla_y \mathbf{f}(x_t, y_t) - \Gamma_{y_{t-1}}^{y_t} \nabla_y \mathbf{f}(x_{t-1}, y_{t-1})\| d(y_t, y_{t+1}).
\end{aligned} \tag{21}$$

Adding (20) and (21) together, for all $t \leq K_0 + 1$, we have

$$\begin{aligned}
0 &\leq \langle \exp_{z_{t+1}}^{-1} z^*, -\mathbf{F}(z_{t+1}) \rangle \\
&\leq \langle \exp_{z_{t+1}}^{-1} z^*, -\mathbf{F}(z_{t+1}) + \Gamma_{z_t}^{z_{t+1}} \mathbf{F}(z_t) \rangle - \langle \exp_{z_t}^{-1} z^*, -\mathbf{F}(z_t) + \Gamma_{z_{t-1}}^{z_t} \mathbf{F}(z_{t-1}) \rangle \\
&\quad + \frac{1}{2\eta}(d^2(x_t, x^*) + d^2(y_t, y^*)) - \frac{1}{2\eta}(d^2(x_{t+1}, x^*) + d^2(y_{t+1}, y^*)) \\
&\quad + \zeta_1 \|\nabla_y \mathbf{f}(x_t, y_t) - \Gamma_{y_{t-1}}^{y_t} \nabla_y \mathbf{f}(x_{t-1}, y_{t-1})\| d(y_t, y_{t+1}) \\
&\quad + \zeta_1 \|\nabla_x \mathbf{f}(x_t, y_t) - \Gamma_{x_{t-1}}^{x_t} \mathbf{f}(x_{t-1}, y_{t-1})\| d(x_t, x_{t+1}) \\
&\quad - \frac{\sigma_1}{2\eta} d^2(x_t, x_{t+1}) - \frac{\sigma_1}{2\eta} d^2(y_t, y_{t+1}).
\end{aligned} \tag{22}$$

Taking Young's inequality with

$$\begin{cases}
a = (d(x_t, x_{t+1}), d(y_t, y_{t+1})) \\
b = (\|\nabla_x \mathbf{f}(x_t, y_t) - \Gamma_{x_{t-1}}^{x_t} \mathbf{f}(x_{t-1}, y_{t-1})\|, \|\nabla_y \mathbf{f}(x_t, y_t) - \Gamma_{y_{t-1}}^{y_t} \nabla_y \mathbf{f}(x_{t-1}, y_{t-1})\|), \\
\alpha = L,
\end{cases}$$

we have

$$\begin{aligned} a \cdot b &\leq \frac{L}{2}(d^2(x_t, x_{t+1}) + d^2(y_t, y_{t+1})) \\ &\quad + \frac{1}{2L}(\|\nabla_x \mathbf{f}(x_t, y_t) - \Gamma_{x_{t-1}}^{x_t} \mathbf{f}(x_{t-1}, y_{t-1})\|^2 + \|\nabla_y \mathbf{f}(x_t, y_t) - \Gamma_{y_{t-1}}^{y_t} \nabla_y \mathbf{f}(x_{t-1}, y_{t-1})\|^2). \\ &\leq \frac{L}{2} d^2(z_t, z_{t+1}) + \frac{1}{2L} L^2 d^2(z_{t-1}, z_t) = \frac{L}{2}(d^2(z_t, z_{t+1}) + d^2(z_{t-1}, z_t)). \end{aligned} \tag{23}$$

Plugging (23) into (22) yields

$$\begin{aligned} 0 &\leq \langle \exp_{z_{t+1}}^{-1} z^*, -\mathbf{F}(z_{t+1}) \rangle \\ &\leq \langle \exp_{z_{t+1}}^{-1} z^*, -\mathbf{F}(z_{t+1}) + \Gamma_{z_t}^{z_{t+1}} \mathbf{F}(z_t) \rangle - \langle \exp_{z_t}^{-1} z^*, -\mathbf{F}(z_t) + \Gamma_{z_{t-1}}^{z_t} \mathbf{F}(z_{t-1}) \rangle \\ &\quad + \frac{1}{2\eta}(d^2(z_t, z^*) - d^2(z_{t+1}, z^*)) + \frac{\zeta_1 L}{2}(d^2(z_t, z_{t+1}) + d^2(z_{t-1}, z_t)) - \frac{\sigma_1}{2\eta} d^2(z_t, z_{t+1}). \end{aligned} \tag{24}$$

Since $\eta \leq \frac{\zeta_1}{2\sigma_1 L}$, we have $-\frac{\sigma_1}{2\eta} + \frac{\zeta_1 L}{2} \leq -\frac{\zeta_1 L}{2}$, which gives us

$$\begin{aligned} 0 &\leq \langle \exp_{z_{t+1}}^{-1} z^*, -\mathbf{F}(z_{t+1}) \rangle \\ &\leq \langle \exp_{z_{t+1}}^{-1} z^*, -\mathbf{F}(z_{t+1}) + \Gamma_{z_t}^{z_{t+1}} \mathbf{F}(z_t) \rangle - \langle \exp_{z_t}^{-1} z^*, -\mathbf{F}(z_t) + \Gamma_{z_{t-1}}^{z_t} \mathbf{F}(z_{t-1}) \rangle \\ &\quad + \frac{1}{2\eta}(d^2(z_t, z^*) - d^2(z_{t+1}, z^*)) + \frac{\zeta_1 L}{2}(-d^2(z_t, z_{t+1}) + d^2(z_{t-1}, z_t)). \end{aligned} \tag{25}$$

By summing (22) from $t = 0$ to K_0, we observe that

$$\begin{aligned} 0 &\leq \sum_{t=0}^{K_0} \langle \exp_{z_{t+1}}^{-1} z^*, -\mathbf{F}(z_{t+1}) \rangle \\ &\leq \langle \exp_{z_{K_0+1}}^{-1} z^*, -\mathbf{F}(z_{K_0+1}) + \Gamma_{z_{K_0}}^{z_{K_0+1}} \mathbf{F}(z_{K_0}) \rangle - \langle \exp_{z_0}^{-1} z^*, -\mathbf{F}(z_{-1}) + \Gamma_{z_{-1}}^{z_0} \mathbf{F}(z_{-1}) \rangle \\ &\quad + \frac{1}{2\eta} d^2(z_0, z^*) - \frac{1}{2\eta} d^2(z_{K_0+1}, z^*) - \frac{\zeta_1 L}{2} d^2(z_{K_0}, z_{K_0+1}) + \frac{\zeta_1 L}{2} d^2(z_0, z_{-1}). \end{aligned}$$

Furthermore, based on the fact that $z_0 = z_{-1}$, we can express that

$$\begin{aligned} 0 &\leq \sum_{t=0}^{K_0} \langle \exp_{z_{t+1}}^{-1} z^*, -\mathbf{F}(z_{t+1}) \rangle \\ &\leq \langle \exp_{z_{K_0+1}}^{-1} z^*, -\mathbf{F}(z_{K_0+1}) + \Gamma_{z_{K_0}}^{z_{K_0+1}} \mathbf{F}(z_{K_0}) \rangle \\ &\quad + \frac{1}{2\eta} d^2(z_0, z^*) - \frac{1}{2\eta} d^2(z_{K_0+1}, z^*) - \frac{\zeta_1 L}{2} d^2(z_{K_0}, z_{K_0+1}). \end{aligned}$$

Using g-L-smoothness again, we have

$$\begin{aligned}
0 &\leq \sum_{t=0}^{K_0} \langle \exp_{z_{t+1}}^{-1} z^*, -\mathbf{F}(z_{t+1}) \rangle \\
&\leq L d(z_{K_0+1}, z^*) d(z_{K_0+1}, z_{K_0}) + \frac{1}{2\eta} d^2(z_0, z^*) - \frac{1}{2\eta} d^2(z_{K_0+1}, z^*) - \frac{\zeta_1 L}{2} d^2(z_{K_0}, z_{K_0+1}) \\
&\leq \frac{L}{2}(d^2(z_{K_0+1}, z^*) + d^2(z_{K_0+1}, z_{K_0})) + \frac{1}{2\eta} d^2(z_0, z^*) \\
&\quad - \frac{1}{2\eta} d^2(z_{K_0+1}, z^*) - \frac{\zeta_1 L}{2} d^2(z_{K_0}, z_{K_0+1}) \\
&\leq \frac{L}{2} d^2(z_{K_0+1}, z^*) + \frac{1}{2\eta} d^2(z_0, z^*) - \frac{1}{2\eta} d^2(z_{K_0+1}, z^*).
\end{aligned}$$

The last inequality is due to $\zeta_1 \geq 1$, and the inequality gives us

$$d^2(z_{K_0+1}, z^*) \leq \frac{1}{1 - \eta L} d^2(z_0, z^*) \leq \frac{1}{1 - \frac{\sigma_1}{2\zeta_1}} d^2(z_0, z^*) \leq 2 d^2(z_0, z^*) = 2 D_1^2.$$

Thus, $d^2(z_t, z^*) \leq 2D_1^2$ holds for $t = K_0 + 1$, By mathematical induction, the claim $d^2(z_t, z^*) \leq 2D_1$ holds for all $t \geq 0$, which completes our proof. ■

Now we shift our focus on proof of Theorem 5.

Proof of Theorem 5 By the g-convexity-concavity, for any $(x, y) \in \mathcal{M} \times \mathcal{N}$, there holds

$$\begin{cases} \sum_{t=1}^T \mathbf{f}(x_t, y) - \sum_{t=1}^T \mathbf{f}(x_t, y_t) \leq \sum_{t=1}^T \langle \nabla_y \mathbf{f}(x_t, y_t), \exp_{y_t}^{-1} y \rangle \\ \sum_{t=1}^T \mathbf{f}(x_t, y_t) - \sum_{t=1}^T \mathbf{f}(x, y_t) \leq \sum_{t=1}^T \langle -\nabla_x \mathbf{f}(x_t, y_t), \exp_{x_t}^{-1} x \rangle. \end{cases}$$

In this way, there holds

$$\begin{aligned}
\sum_{t=1}^T \mathbf{f}(x_t, y) - \sum_{t=1}^T \mathbf{f}(x, y_t) &\leq \sum_{t=1}^T \langle \nabla_y \mathbf{f}(x_t, y_t), \exp_{y_t}^{-1} y \rangle + \sum_{t=1}^T \langle -\nabla_x \mathbf{f}(x_t, y_t), \exp_{x_t}^{-1} x \rangle \\
&= \sum_{t=1}^T \langle -\mathbf{F}(z_t), \exp_{z_t}^{-1} z \rangle.
\end{aligned}$$

From the proof of Lemma 4, we can obtain

$$\begin{aligned}
\sum_{t=1}^T \mathbf{f}(x_t, y) - \sum_{t=1}^T \mathbf{f}(x, y_t) &\leq \sum_{t=1}^T \langle -\mathbf{F}(z_t), \exp_{z_t}^{-1} z \rangle \\
&\leq \langle \exp_{z_T}^{-1} z, -\mathbf{F}(z_T) + \Gamma_{z_{T-1}}^{z_T} \mathbf{F}(z_{T-1}) \rangle + \frac{1}{2\eta} d^2(z_0, z^*) \\
&\quad - \frac{1}{2\eta} d^2(z_{K_0+1}, z^*) - \frac{\zeta_1 L}{2} d^2(z_{K_0}, z_{K_0+1}) \\
&\leq 2 D_1 G + \frac{1}{\eta} D_1^2.
\end{aligned}$$

To complete the proof, it remains to show that

$$\mathbf{f}(\bar{x}_T, y) - \mathbf{f}(x, \bar{y}_T) \leq \frac{1}{T}\sum_{t=1}^{T}\mathbf{f}(x_t, y) - \frac{1}{T}\sum_{t=1}^{T}\mathbf{f}(x, y_t),$$

which can be proved by induction

$$\begin{aligned}
\mathbf{f}(\bar{x}_T, y) &= \mathbf{f}(\exp_{\bar{x}_t}(\frac{1}{T}\exp^{-1}_{\bar{x}_{T-1}} x_T), y) \\
&\leq \frac{1}{T}\mathbf{f}(x_T, y) + \frac{T-1}{T}\mathbf{f}(\bar{x}_{T-1}, y) \\
&\leq \frac{1}{T}\mathbf{f}(x_T, y) + \frac{T-1}{T}\frac{1}{T-1}\mathbf{f}(x_{T-1}, y) + \frac{T-1}{T}\frac{T-2}{T-1}\mathbf{f}(\bar{x}_{T-2}, y) \\
&\leq \frac{1}{T}\mathbf{f}(x_T, y) + \frac{1}{T}\mathbf{f}(x_{T-1}, y) + \cdots + \frac{1}{T}\mathbf{f}(x_2, y) + \frac{1}{T}\mathbf{f}(\bar{x}_1, y) \\
&= \frac{1}{T}\sum_{t=1}^{T}\mathbf{f}(x_t, y),
\end{aligned}$$

and

$$\begin{aligned}
\mathbf{f}(x, \bar{y}_T) &= \mathbf{f}(x, \exp_{\bar{y}_t}(\frac{1}{T}\exp^{-1}_{\bar{y}_{T-1}} y_T)) \\
&\geq \frac{1}{T}\mathbf{f}(x, y_T) + \frac{T-1}{T}\mathbf{f}(x, \bar{y}_{T-1}) \\
&\geq \frac{1}{T}\mathbf{f}(x, y_T) + \frac{1}{T}\mathbf{f}(x, y_{T-1}) + \cdots + \frac{1}{T}\mathbf{f}(x, y_2) + \frac{1}{T}\mathbf{f}(x, \bar{y}_1) \\
&= \frac{1}{T}\sum_{t=1}^{T}\mathbf{f}(x, y_t).
\end{aligned}$$

Then Theorem 5 has been established. ■

Appendix E. Proof of Lemma 6

We first propose some lemmas that are useful in proving Lemma 6.

Lemma 21 (A variant of Gauss-Bonnet theorem, Lee, 2018; Chern et al., 1999)**.** *Suppose M is a manifold with sectional curvature in $[\kappa, K]$ and $\Xi(s,t) : [0,1] \times [0.1] \to \mathcal{M}$ is a rectangle map. Γ_γ is the parallel transport around the boundary curve γ that $\gamma = \Xi(t,0) \cup \Xi(1,s) \cup \Xi(t,1) \cup \Xi(0,s)$. Denote vector fields $S(\Xi(s,t)) = \Xi_* \frac{\partial}{\partial s}(s,t)$, $T(\Xi(s,t)) = \Xi_* \frac{\partial}{\partial t}(s,t)$ and $K_m = \max(|\kappa|, |K|)$. Then we have*

$$\|\Gamma_\gamma z - z\| \leq 12 K_m \|z\| \int_0^1 \int_0^1 \|T\|\|S\| ds dt, \forall z \in T_{\Xi(0,0)}\mathcal{M}.$$

Proof We first extend z to a vector field $Z(s_0, t_0) = Z(\Xi(s_0, t_0))$ by first parallel transporting z along the curve $\Xi(0,t), (0 \leq t \leq t_0)$ and then parallel transporting along the curve $\Xi(s, t_0), (0 \leq s \leq s_0)$. It shows that

$$\begin{cases} \nabla_S Z(s,t) = 0 \\ \nabla_T Z(0,t) = 0 \end{cases} \quad \forall (s,t) \in [0,1] \times [0,1].$$

For an arbitrary vector $w \in T_{\Xi(0,0)}\mathcal{M}$, we also extend it to $W(s_0, t_0)$ by first parallel transporting along the curve $\Xi(s,0), (0 \le s \le 1)$, then along the curve $\Xi(1,t), (0 \le t \le t_0)$, and along the curve $\Xi(s,t_0), (1 \ge s \ge s_0)$. We can also have

$$\begin{cases} \nabla_S W(s,t) = 0 \\ \nabla_T W(1,t) = 0 \end{cases} \quad \forall (s,t) \in [0,1] \times [0,1].$$

We denote two curves that $\xi_1 = \Xi(s,0), (0 \le s \le 1)$ and $\xi_2 = \Xi(s,0) \cup \Xi(1,t), (0 \le s,t \le 1)$. By the above notation, we find

$$\begin{aligned} \langle \Gamma_\gamma z - z, w \rangle &= \langle \Gamma_\gamma z, w \rangle - \langle z, w \rangle \\ &= \langle \Gamma_{\xi_2} \Gamma_\gamma z, \Gamma_{\xi_2} w \rangle - \langle \Gamma_{\xi_1} z, \Gamma_{\xi_2} w \rangle. \end{aligned}$$

From the way we extend Z and W, we know that $\Gamma_{\xi_2}\Gamma_\gamma z = Z(1,1)$, $\Gamma_{\xi_2} w = W(1,1)$, $\Gamma_{\xi_1} z = Z(1,0)$, and $\Gamma_{\xi_2} w = W(1,0)$, thus we have

$$\begin{aligned} \langle \Gamma_\gamma z - z, w \rangle &= \langle Z(1,1), W(1,1) \rangle - \langle Z(1,0) - W(1,0) \rangle \\ &= \int_0^1 \frac{\partial}{\partial t} \langle Z(1,t), W(1,t) \rangle dt \\ &= \int_0^1 T \langle Z(1,t), W(1,t) \rangle dt \\ &= \int_0^1 \langle \nabla_T Z(1,t), W(1,t) \rangle + \langle Z(1,t), \nabla_T W(1,t) \rangle dt. \end{aligned}$$

Due to the fact that $\nabla_T W(1,t) = 0$, we have

$$\begin{aligned} \langle \Gamma_\gamma z - z, w \rangle &= \int_0^1 \langle \nabla_T Z(1,t), W(1,t) \rangle dt \\ &= \int_0^1 \Big(\langle \nabla_T Z(0,t), W(0,t) \rangle + \int_0^1 \partial_s \langle \nabla_T Z(s,t), W(s,t) \rangle ds \Big) dt \\ &= \int_0^1 \int_0^1 \langle \frac{\partial}{\partial s} \langle \nabla_T Z(s,t), W(s,t) \rangle ds dt \\ &= \int_0^1 \int_0^1 S \langle \nabla_T Z(s,t), W(s,t) \rangle ds dt \\ &= \int_0^1 \int_0^1 \langle \nabla_S \nabla_T Z(s,t), W(s,t) \rangle + \langle \nabla_T Z(s,t), \nabla_S W(s,t) \rangle ds dt \\ &= \int_0^1 \int_0^1 \langle \nabla_S \nabla_T Z(s,t), W(s,t) \rangle ds dt. \end{aligned}$$

The last equality is from the fact that $\nabla_S W(s,t) = 0$. Since the curvature has the form

$$R(S,T,Z,W) = \langle \nabla_S \nabla_T Z(s,t), W(s,t) \rangle + \langle \nabla_T \nabla_S Z(s,t), W(s,t) \rangle + \langle \nabla_{[S,T]} Z(s,t), W(s,t) \rangle$$

and we have $\nabla_S Z(s,t) = 0$, $[S,T] = 0$, it holds that

$$\begin{aligned} \langle \Gamma_\gamma z - z, w \rangle &= \int_0^1 \int_0^1 R(S,T,Z,W) ds dt \\ &= \int_0^1 \int_0^1 R(\frac{S}{\|S\|}, \frac{T}{\|T\|}, \frac{Z}{\|Z\|}, \frac{W}{\|W\|}) \|S\| \|T\| \|W\| \|Z\| ds dt. \end{aligned}$$

By Lemma 14 and $R(X,Y,X,Y) \leq K_m\|X\|^2\|Y\|^2$, we have

$$R(\frac{S}{\|S\|},\frac{T}{\|T\|},\frac{Z}{\|Z\|},\frac{W}{\|W\|}) \leq 12K_m.$$

Hence,

$$\begin{aligned}\langle \Gamma_\gamma z - z, w\rangle &\leq 12K_m \int_0^1\int_0^1 \|S\|\|T\|\|W\|\|Z\|dsdt \\ &= 12K_m\|z\|\|w\|\int_0^1\int_0^1 \|S\|\|T\|dsdt,\end{aligned}$$

which completes our proof since w is arbitary. ■

Lemma 22. *Suppose M is a manifold with sectional curvature in $[\kappa, K]$ and $\gamma : [0,b] \to \mathcal{M}$ is a geodesic with $\|\dot\gamma(0)\| = 1$ ($b \leq \frac{\pi}{\sqrt{K}}$ if $K > 0$). If J is a Jacobi field along γ with $\|J(\gamma(0))\| = \alpha_1$ and $\|J(\gamma(b))\| = \alpha_2$, then we have*

$$\|J(\gamma(t))\| \leq \frac{\mathbf{s}(\kappa,b)}{\mathbf{S}(K,b)}(\alpha_1+\alpha_2).$$

Proof We split $J(\gamma(t)) = J_1(\gamma(t)) + J_2(\gamma(t))$, where J_1 is a Jacobi field such that $J_1(\gamma(0)) = 0$ and $J_1(\gamma(b)) = J(\gamma(b))$, and where J_2 is a Jacobi field such that $J_2(\gamma(b)) = 0$ and $J_2(\gamma(0)) = J(\gamma(0))$. Applying the Jacobi comparison theorem (Lemma 15), we have

$$\begin{cases}\mathbf{S}(K,b)\|\nabla_{\dot\gamma}J_1(0)\| \leq \|J_1(\gamma(b))\| = \alpha_2 \\ \mathbf{S}(K,b)\|\nabla_{\dot\gamma}J_2(b)\| \leq \|J_2(\gamma(0))\| = \alpha_1,\end{cases} \tag{26}$$

and

$$\begin{cases}\|J_1(\gamma(t))\| \leq \mathbf{s}(\kappa,t)\|\nabla_{\dot\gamma}J_1(0)\| \leq \mathbf{s}(\kappa,b)\|\nabla_{\dot\gamma}J_1(0)\| \\ \|J_2(\gamma(t))\| \leq \mathbf{s}(\kappa,(b-t))\|\nabla_{\dot\gamma}J_2(b)\| \leq \mathbf{s}(\kappa,b)\|\nabla_{\dot\gamma}J_2(b)\|.\end{cases} \tag{27}$$

Combining (26) and (27) we have

$$\begin{cases}\|J_1(\gamma(t))\| \leq \frac{\mathbf{s}(\kappa,b)}{\mathbf{S}(K,b)}\alpha_1 \\ \|J_2(\gamma(t))\| \leq \frac{\mathbf{s}(\kappa,b)}{\mathbf{S}(K,b)}\alpha_2,\end{cases}$$

which gives us

$$\|J(\gamma(t))\| \leq \|J_1(\gamma(t))\| + \|J_2(\gamma(t))\| \leq \frac{\mathbf{s}(\kappa,b)}{\mathbf{S}(K,b)}(\alpha_1+\alpha_2).$$

This completes the proof. ■

Lemma 23. *Denote $K_m = \max(|\kappa|,|K|)$. If $0 \leq b \leq \frac{1}{\sqrt{K_m}}$, then we have $\frac{\mathbf{s}(\kappa,b)}{\mathbf{S}(K,b)} \leq 3$.*

Proof It suffice to proof the lemma in the case $K \geq 0$ and $\kappa \leq 0$. We first consider the case where $K > 0$ and $\kappa < 0$, where

$$\frac{\mathbf{s}(\kappa, b)}{\mathbf{S}(K, b)} = \frac{\sqrt{K}\sinh(\sqrt{-k}b)}{\sqrt{-\kappa}\sin(\sqrt{K}b)}. \tag{28}$$

We show that $\cosh(ax) \leq (1 + a^2x^2)$ for $0 \leq x \leq \frac{1}{a}$. Let

$$f(x) = \cosh(ax) - 1 - a^2x^2.$$

We have

$$\begin{cases} f(0) = 0 \\ f'(x) = a\sinh(ax) - 2a^2x, f'(0) = 0 \\ f''(x) = a^2(\cosh(ax) - 2). \end{cases}$$

Since $\cosh(ax) - 2 \leq 0$ for $0 \leq x \leq \frac{1}{a}$, we have $f(x) \leq 0$ for $0 \leq x \leq \frac{1}{a}$.

Then we show that $\frac{\sinh(ax)}{\sin(cx)} \leq \frac{a}{c} + \frac{a(a^2+c^2)x^2}{c}$ for $0 \leq x \leq \frac{1}{a}$. By defining

$$g(x) = \sinh(ax) - (\frac{a}{c} + \frac{a(a^2 + c^2)x^2}{c})\sin(cx),$$

we have

$$\begin{cases} g(0) = 0 \\ g'(x) = a\cosh(ax) - (a + ax^2(a^2 + c^2))\cos cx - (\frac{2a(a^2+c^2)x}{c}))\sin cx \end{cases}$$

As $0 \leq x \leq \frac{1}{a}$, we have

$$\begin{aligned} g'(x) &\leq a(1 + a^2x^2) - (a + ax^2(a^2 + c^2))\cos cx - (\frac{2a(a^2 + c^2)x}{c}))\sin cx \\ &\leq a(1 + a^2x^2)(1 - \cos cx) - 2acx\sin cx \\ &\leq a(1 + a^2x^2)(1 - \cos^2 cx) - 2a\sin^2 cx \\ &= a(1 + a^2x^2)\sin^2 cx - 2a\sin^2 cx \\ &\leq a(-1 + a^2x^2)\sin^2 cx \\ &\leq 0 \end{aligned}$$

And thus, we have

$$\sinh(ax) - (\frac{a}{c} + \frac{a(a^2 + c^2)x^2}{c})\sin(cx) \leq 0, \quad 0 \leq x \leq \frac{1}{a}$$

which is equivalent to

$$\frac{\sinh(ax)}{\sin(cx)} \leq \frac{a}{c} + \frac{a(a^2 + c^2)x^2}{c}, \quad 0 \leq x \leq \frac{1}{a} \tag{29}$$

Putting (28) into (29) with $a = \sqrt{-\kappa}$ and $c = \sqrt{K}$, we have, for $0 \leq b \leq \frac{1}{\sqrt{K_m}}$

$$\begin{aligned}\frac{\mathbf{s}(\kappa, b)}{\mathbf{S}(K, b)} &= \frac{\sqrt{K}\sinh(\sqrt{-k}b)}{\sqrt{-\kappa}\sin(\sqrt{K}b)} \\ &\leq 1 + (\kappa + K)b^2. \\ &\leq 1 + 2K_m \frac{1}{K_m} \\ &\leq 3,\end{aligned}$$

which completes the proof for the where case $K > 0$ and $\kappa < 0$. Since $\frac{\mathbf{s}(\kappa,b)}{\mathbf{S}(K,b)}$ is continuous on K and κ, we have proved the case with $K = 0$ or $\kappa = 0$ by continuity. ■

Lemma 24. *In Algorithm 4, if $\eta \leq \frac{1}{8L}$, it satisfies*

$$\frac{1}{2} \leq \frac{\|\mathbf{F}(z_{t+1})\|}{\|\mathbf{F}(z_t)\|} \leq \frac{3}{2}. \tag{30}$$

The proof is as same as that the prior work by Chavdarova et al. (2021), so we omit the proof.

Then we begin our proof of Lemma 6.

Proof of Lemma 6 We first estimate the length $d(z_{t-1}, z_t)$ by,

$$\begin{aligned}d(z_{t-1}, z_t) &= \|2\eta\mathbf{F}(z_{t-1}) - \Gamma_{z_{t-2}}^{z_{t-1}}\eta(\mathbf{F}(z_{t-2}))\| \\ &\leq \eta\|\mathbf{F}(z_{t-1})\| + \eta\|\mathbf{F}(z_{t-1}) - \Gamma_{z_{t-2}}^{z_{t-1}}(\mathbf{F}(z_{t-2}))\| \\ &\leq \eta\|\mathbf{F}(z_{t-1})\| + \eta L d(z_{t-1}, z_{t-2}) \\ &= \eta\|\mathbf{F}(z_{t-1})\| + \eta L\|2\mathbf{F}(z_{t-2}) - \Gamma_{z_{t-3}}^{z_{t-2}}\mathbf{F}(z_{t-3})\| \\ &\leq \eta\|\mathbf{F}(z_{t-1})\| + \eta L(\|2\mathbf{F}(z_{t-2})\| + \|\mathbf{F}(z_{t-3})\|).\end{aligned}$$

By Lemma 24, we have $\|\mathbf{F}(z_{t-2})\| \leq 2\|\mathbf{F}(z_{t-1})\|$ and $\|\mathbf{F}(z_{t-3})\| \leq 4\|\mathbf{F}(z_{t-1})\|$. So, we have

$$d(z_{t-1}, z_t) \leq (1 + 8L\eta)\eta\|\mathbf{F}(z_{t-1})\|. \tag{31}$$

Similarly, $d(\hat{z}_{t+1}, z_t)$ and $d(\hat{z}_t, z_{t-1})$ can be bounded by

$$\begin{aligned}d(\hat{z}_{t+1}, z_t) &= \|\eta\mathbf{F}(z_t) - \Gamma_{z_{t-1}}^{z_t}\mathbf{F}(z_{t-1})\| \\ &\leq L\eta d(z_t, z_{t-1}) \\ &\leq L\eta^2\|2\eta\mathbf{F}(z_{t-1}) - \Gamma_{z_{t-2}}^{z_{t-1}}\eta(\mathbf{F}(z_{t-2}))\| \\ &\leq L\eta^2(2\|\mathbf{F}(z_{t-1})\| + \|\mathbf{F}(z_{t-2})\|) \\ &\leq 4L\eta^2\|\mathbf{F}(z_{t-1})\|,\end{aligned}$$

and

$$\begin{aligned}d(\hat{z}_t, z_{t-1}) &= \eta\|\mathbf{F}(z_{t-1}) - \Gamma_{z_{t-2}}^{z_{t-1}}(\mathbf{F}(z_{t-2}))\| \\ &\leq L\eta^2\|2\mathbf{F}(z_{t-2}) - \Gamma_{z_{t-3}}^{z_{t-2}}(\mathbf{F}(z_{t-3}))\| \\ &\leq 8L\eta^2\|\mathbf{F}(z_{t-1})\|.\end{aligned}$$

Consequently, the $l_t := d(z_{t-1}, z_t) + d(\hat{z}_{t+1}, z_t) + d(\hat{z}_t, z_{t-1})$ has the bound

$$\begin{aligned} l_t &\leq (1 + 8L\eta)\eta\|\mathbf{F}(z_{t-1})\| + 4L\eta^2\|\mathbf{F}(z_{t-1})\| + 8L\eta^2\|\mathbf{F}(z_{t-1})\| \\ &\leq (1 + 20L\eta)\eta\|\mathbf{F}(z_{t-1})\|. \end{aligned}$$

Since $\eta \leq \frac{1}{20L}$, we have

$$l_t \leq 2\eta\|\mathbf{F}(z_{t-1})\|. \tag{32}$$

Now we examine (i). Applying Lemma 20 in the geodesic triangle $\triangle z_{t-1}\hat{z}_{t+1}z_t$ yields

$$\exp_{z_{t-1}}^{-1} \hat{z}_{t+1} = H_{z_{t-1},p_1}^{\hat{z}_{t+1}} \exp_{z_{t-1}}^{-1} z_t + \Gamma_{z_t}^{z_{t-1}} \exp_{z_t}^{-1} \hat{z}_{t+1}, \tag{33}$$

where p_1 lies in the geodesic between z_{t-1} and z_t. Applying Lemma 20 in the geodesic triangle $\triangle z_{t-1}\hat{z}_{t+1}\hat{z}_t$ yields

$$\exp_{z_{t-1}}^{-1} \hat{z}_{t+1} = H_{z_{t-1},p_2}^{\hat{z}_{t+1}} \exp_{z_{t-1}}^{-1} \hat{z}_t + \Gamma_{\hat{z}_t}^{z_{t-1}} \exp_{\hat{z}_t}^{-1} \hat{z}_{t+1}. \tag{34}$$

where p_2 lies in the geodesic between z_{t-1} and $\hat{z}_t$.

Combing (33) and (34), we have

$$\begin{aligned} &H_{z_{t-1},p_2}^{\hat{z}_{t+1}}(-\eta\mathbf{F}(z_{t-1}) + \eta\Gamma_{z_{t-2}}^{z_{t-1}}\mathbf{F}(z_{t-2})) - \Gamma_{\hat{z}_t}^{z_{t-1}}\Gamma_{\hat{z}_{t+1}}^{\hat{z}_t} G_{t+1} \\ &\quad = H_{z_{t-1},p_1}^{\hat{z}_{t+1}}(-2\eta\mathbf{F}(z_{t-1}) + \eta\Gamma_{z_{t-2}}^{z_{t-1}}\mathbf{F}(z_{t-2})) + \Gamma_{z_t}^{z_{t-1}}(-\eta\mathbf{F}(z_t) + \Gamma_{z_{t-1}}^{z_t}\eta\mathbf{F}(z_{t-1})). \end{aligned} \tag{35}$$

Rearranging (35), we have

$$\begin{aligned} \Gamma_{\hat{z}_t}^{z_{t-1}}\Gamma_{\hat{z}_{t+1}}^{\hat{z}_t} G_{t+1} - \Gamma_{z_t}^{z_{t-1}}(\eta\mathbf{F}(z_t)) &= (H_{z_{t-1},p_2}^{\hat{z}_{t+1}} - Id)(-\eta\mathbf{F}(z_{t-1}) + \eta\Gamma_{z_{t-2}}^{z_{t-1}}\mathbf{F}(z_{t-2})) \\ &\quad - (H_{z_{t-1},p_1}^{\hat{z}_{t+1}} - Id)(-2\eta\mathbf{F}(z_{t-1}) + \eta\Gamma_{z_{t-2}}^{z_{t-1}}\mathbf{F}(z_{t-2})). \end{aligned}$$

By Corollary 20, we have

$$\|\Gamma_{\hat{z}_t}^{z_{t-1}}\Gamma_{\hat{z}_{t+1}}^{\hat{z}_t} G_{t+1} - \eta\Gamma_{z_t}^{z_{t-1}}\mathbf{F}(z_t)\| \leq K_m d^2(\hat{z}_{t+1}, p_1)d(z_{t-1}, \hat{z}_t) + K_m d^2(\hat{z}_{t+1}, p_2)d(z_{t-1}, z_t).$$

Since p_1 lies in the geodesic between z_{t-1} and z_t, we have $d(\hat{z}_{t+1}, p_1) \leq d(\hat{z}_{t+1}, z_t) + d(\hat{z}_t, z_{t-1}) \leq l_t$. Also, we have $d(\hat{z}_{t+1}, p_2) \leq l_t$. Then, we have

$$\begin{aligned} \|\Gamma_{\hat{z}_t}^{z_{t-1}}\Gamma_{\hat{z}_{t+1}}^{\hat{z}_t} G_{t+1} - \eta\Gamma_{z_t}^{z_{t-1}}\mathbf{F}(z_t)\| &\leq K_m l_t^2(d(z_{t-1}, \hat{z}_t) + d(z_{t-1}, z_t)) \\ &\leq K_m l_t^3 = K_m 8\eta^3\|\mathbf{F}(z_{t-1})\|^3. \end{aligned}$$

Then,

$$\begin{aligned} \|G_{t+1}\|^2 - \|\eta\mathbf{F}(z_t)\|^2 &= \|\Gamma_{\hat{z}_t}^{z_{t-1}}\Gamma_{\hat{z}_{t+1}}^{\hat{z}_t} G_{t+1}\|^2 - \|\eta\Gamma_{z_t}^{z_{t-1}}\mathbf{F}(z_t)\|^2 \\ &\leq \|\Gamma_{\hat{z}_t}^{z_{t-1}}\Gamma_{\hat{z}_{t+1}}^{\hat{z}_t} G_{t+1} - \eta\Gamma_{z_t}^{z_{t-1}}\mathbf{F}(z_t)\|^2 = 64K_m\eta^6\|\mathbf{F}(z_{t-1})\|^6. \end{aligned}$$

Next we examine (ii). It is shown that

$$\begin{aligned} \|\Gamma_{\hat{z}_{t+1}}^{z_t} G_{t+1} - \eta\mathbf{F}(z_t)\| &\leq \|\Gamma_{\hat{z}_{t+1}}^{z_t} G_{t+1} - \Gamma_{\hat{z}_{t+1}}^{z_t}\Gamma_{\hat{z}_t}^{\hat{z}_{t+1}}\Gamma_{z_{t-1}}^{\hat{z}_t}\Gamma_{z_t}^{z_{t-1}}\eta\mathbf{F}(z_t)\| \\ &\quad + \|\Gamma_{\hat{z}_{t+1}}^{z_t}\Gamma_{\hat{z}_t}^{\hat{z}_{t+1}}\Gamma_{z_{t-1}}^{\hat{z}_t}\Gamma_{z_t}^{z_{t-1}}\eta\mathbf{F}(z_t) - \eta\mathbf{F}(z_t)\| \\ &= \|\Gamma_{\hat{z}_t}^{z_{t-1}}\Gamma_{\hat{z}_{t+1}}^{\hat{z}_t} G_{t+1} - \eta\Gamma_{z_t}^{z_{t-1}}\mathbf{F}(z_t)\| \\ &\quad + \|\Gamma_{\hat{z}_{t+1}}^{z_t}\Gamma_{\hat{z}_t}^{\hat{z}_{t+1}}\Gamma_{z_{t-1}}^{\hat{z}_t}\Gamma_{z_t}^{z_{t-1}}\eta\mathbf{F}(z_t) - \eta\mathbf{F}(z_t)\| \\ &\leq K_m 8\eta^3\|\mathbf{F}(z_{t-1})\|^3 + \|\Gamma_{\hat{z}_{t+1}}^{z_t}\Gamma_{\hat{z}_t}^{\hat{z}_{t+1}}\Gamma_{z_{t-1}}^{\hat{z}_t}\Gamma_{z_t}^{z_{t-1}}\eta\mathbf{F}(z_t) - \eta\mathbf{F}(z_t)\|. \end{aligned}$$

We turn our focus on the geodesic rectangle $z_t z_{t-1} \hat{z}_t \hat{z}_{t+1}$. Denote $\gamma_1(s) : [0,1] \to \mathcal{M}$ as the geodesic from z_{t-1} to $\hat{z}_t$ and $\gamma_1(s) : [0,1] \to \mathcal{M}$ as the geodesic from z_t to $\hat{z}_{t+1}$. We define a rectangle map $\Xi : [0,1] \times [0,1] \to \mathcal{M}$ such that

$$\Xi(s,t) = \exp_{\gamma_1(s)}(t \exp^{-1}_{\gamma_1(s)} \gamma_2(s)).$$

The boundary curve of Ξ is the geodesic rectangle $z_t z_{t-1} \hat{z}_t \hat{z}_{t+1}$. Denote $S = \Xi_*(\frac{\partial}{\partial s})$ and $T = \Xi_*(\frac{\partial}{\partial t})$. From Lemma 21, we have

$$\|\Gamma^{z_t}_{\hat{z}_{t+1}} \Gamma^{\hat{z}_{t+1}}_{\hat{z}_t} \Gamma^{\hat{z}_t}_{z_{t-1}} \Gamma^{z_{t-1}}_{z_t} \eta \mathbf{F}(z_t) - \eta \mathbf{F}(z_t)\| \leq 12 K_m \eta \|\mathbf{F}(z_t)\| \int_0^1 \int_0^1 \|S\| \|T\| ds dt. \tag{36}$$

Notice that each t-curve is a geodesic from $\gamma_1(s)$ to $\gamma_2(s)$, we have

$$\begin{aligned}
\|T(s,t)\| &= (d(\gamma_1(s), \gamma_2(s))) \\
&\leq (d(\gamma_1(s), z_{t-1}) + d(z_{t-1}, z_t) + d(\gamma_2(s), z_t)) \\
&\leq (d(\hat{z}_t, z_{t-1}) + d(z_{t-1}, z_t) + d(\hat{z}_{t+1}, z_t)) \\
&\leq l_t = 2\eta \|\mathbf{F}(z_{t-1})\|.
\end{aligned} \tag{37}$$

Moreover, the vector field S is a Jacobi field along every t-curve with $\|S(s,0)\| = d(\hat{z}_t, z_{t-1})$ and $\|S(s,1)\| = d(z_t, \hat{z}_{t+1})$. By Lemma 22, we have

$$\|S(s,t)\| \leq \frac{\mathbf{s}(\kappa, \|T(s,t)\|)}{\mathbf{S}(K, \|T(s,t)\|)} (d(\hat{z}_t, z_{t-1}) + d(z_t, \hat{z}_{t+1})).$$

Since $\eta \leq \frac{1}{2\sqrt{K_m} G}$, we have $\|T(s,t)\| \leq 2\eta \|\mathbf{F}(z_{t-1})\| \leq \frac{1}{\sqrt{K_m}}$, and thus by Lemma 23

$$\begin{aligned}
\|S(s,t)\| &\leq \frac{\mathbf{s}(\kappa, \|T(s,t)\|)}{\mathbf{S}(K, \|T(s,t)\|)} (d(\hat{z}_t, z_{t-1}) + d(z_t, \hat{z}_{t+1})) \\
&\leq 3(d(\hat{z}_t, z_{t-1}) + d(z_t, \hat{z}_{t+1}))
\end{aligned}$$

With $d(\hat{z}_t, z_{t-1}) \leq 4L\eta^2 \|\mathbf{F}(z_{t-1})\|$, $d(z_t, \hat{z}_{t+1}) \leq 8L\eta^2 \|\mathbf{F}(z_{t-1})\|$ and $\eta \leq \frac{1}{20L}$, we have

$$\|S(s,t)\| \leq \frac{9}{5} \eta \|\mathbf{F}(z_{t-1})\| \leq 2\eta \|\mathbf{F}(z_{t-1})\| \tag{38}$$

Taking (37) and (38) in (36), we have

$$\begin{aligned}
\|\Gamma^{z_t}_{\hat{z}_{t+1}} \Gamma^{\hat{z}_{t+1}}_{\hat{z}_t} \Gamma^{\hat{z}_t}_{z_{t-1}} \Gamma^{z_{t-1}}_{z_t} \eta \mathbf{F}(z_t) - \eta \mathbf{F}(z_t)\| &\leq 12 K_m \eta \|\mathbf{F}(z_t)\| \cdot 4\eta^2 \|\mathbf{F}(z_{t-1})\|^2 \\
&\leq 12 K_m \eta 8 \eta^2 \|\mathbf{F}(z_{t-1})\|^3 = 96 K_m \eta^3 \|\mathbf{F}(z_{t-1})\|^3.
\end{aligned}$$

Thus, we have

$$\begin{aligned}
\|\Gamma^{z_t}_{\hat{z}_{t+1}} G_{t+1} - \eta \mathbf{F}(z_t)\| &\leq K_m 8 \eta^3 \|\mathbf{F}(z_{t-1})\|^3 + \|\Gamma^{z_t}_{\hat{z}_{t+1}} \Gamma^{\hat{z}_{t+1}}_{\hat{z}_t} \Gamma^{\hat{z}_t}_{z_{t-1}} \Gamma^{z_{t-1}}_{z_t} \eta \mathbf{F}(z_t) - \eta \mathbf{F}(z_t)\| \\
&\leq K_m 8 \eta^3 \|\mathbf{F}(z_{t-1})\|^3 + 96 K_m \eta^3 \|\mathbf{F}(z_{t-1})\|^3 \\
&= 104 K_m \eta^3 \|\mathbf{F}(z_{t-1})\|^3,
\end{aligned}$$

which completes our proof. ■

Appendix F. Proof of Theorems 7 and 8

Proof of Theorem 7 Define $\hat{z}_{t+1} = \exp_{z_t}(\eta \mathbf{F}(z_t) + \Gamma_{z_{t-1}}^{z_t} \mathbf{F}(z_{t-1}))$. We consider the following Lyapunov function

$$\begin{aligned}\phi_{t+1} &:= d(\hat{z}_{t+1}, z^*) + \sigma_1 d^2(z_{t+1}, \hat{z}_{t+2}) \\ &= d(\hat{z}_{t+1}, z^*) + \sigma_1 \eta^2 \|\nabla \mathbf{F}(z_{t+1}) - \Gamma_{z_t}^{z_{t+1}} \mathbf{F}(z_t)\|^2.\end{aligned}$$

The difference between ϕ_{t+2} and ϕ_{t+1} is

$$\begin{aligned}\phi_{t+2} - \phi_{t+1} = d^2(\hat{z}_{t+2}, z^*) - d^2(\hat{z}_{t+1}, z^*) + 2\sigma_1 \langle \eta \mathbf{F}(z_{t+1}), \Gamma_{z_t}^{z_{t+1}} \eta \mathbf{F}(z_t) \rangle \\ + \sigma_1 \eta^2 (\|\mathbf{F}(z_{t+2})\|^2 - \|\mathbf{F}(z_t)\|^2 - 2\langle \mathbf{F}(z_{t+2}), \Gamma_{z_{t+1}}^{z_{t+2}} \mathbf{F}(z_{t+1}) \rangle. \end{aligned} \tag{39}$$

In the geodesic $\triangle \hat{z}_{t+1} \hat{z}_{t+2} z^*$, Lemma 4 and Lemma 17 give

$$d^2(\hat{z}_{t+2}, z^*) - d^2(\hat{z}_{t+1}, z^*) \leq 2\langle G_{t+2}, \exp_{\hat{z}_{t+2}}^{-1} z^* \rangle - \sigma_1 \|G_{t+2}\|^2, \tag{40}$$

where $G_{t+2} = \exp_{\hat{z}_{t+2}}^{-1} \hat{z}_{t+1}$. Substituting (40) into (39), we have

$$\begin{aligned}\phi_{t+2} - \phi_{t+1} = 2\langle G_{t+2}, \exp_{\hat{z}_{t+2}}^{-1} z^* \rangle - \sigma_1 \|G_{t+2}\|^2 + 2\sigma_1 \langle \eta \mathbf{F}(z_{t+1}), \Gamma_{z_t}^{z_{t+1}} \eta \mathbf{F}(z_t) \rangle \\ + \sigma_1 \eta^2 (\|\mathbf{F}(z_{t+2})\|^2 - \|\mathbf{F}(z_t)\|^2 - 2\langle \mathbf{F}(z_{t+2}), \Gamma_{z_{t+1}}^{z_{t+2}} \mathbf{F}(z_{t+1}) \rangle).\end{aligned}$$

With a matter of algebraic calculations, we have

$$\begin{aligned}\phi_{t+2} - \phi_{t+1} \leq 2\langle G_{t+2}, \exp_{\hat{z}_{t+2}}^{-1} z^* \rangle + 2\sigma_1 \langle \eta \mathbf{F}(z_{t+1}), -\eta \mathbf{F}(z_{t+1}) + \eta \Gamma_{z_t}^{z_{t+1}} \mathbf{F}(z_t) \rangle \\ + \eta \sigma_1 \big(\|\mathbf{F}(z_{t+2})\|^2 - \|\mathbf{F}(z_t)\|^2 + \|\mathbf{F}(z_{t+1})\|^2 - 2\langle \eta \mathbf{F}(z_{t+2}), \eta \Gamma_{z_{t+1}}^{z_{t+2}} \mathbf{F}(z_{t+1}) \rangle \big) \\ + \sigma_1 (\|\eta \mathbf{F}(z_{t+1})\|^2 - \|G_{t+2}\|^2). \end{aligned} \tag{41}$$

We define

$$\begin{cases} A := 2\langle G_{t+2}, \exp_{\hat{z}_{t+2}}^{-1} z^* \rangle + 2\sigma_1 \langle \eta \mathbf{F}(z_{t+1}), -\eta \mathbf{F}(z_{t+1}) + \eta \Gamma_{z_t}^{z_{t+1}} \mathbf{F}(z_t) \rangle; \\ B := \eta \sigma_1 \big(\|\mathbf{F}(z_{t+2})\|^2 - \|\mathbf{F}(z_t)\|^2 + \|\mathbf{F}(z_{t+1})\|^2 - 2\langle \eta \mathbf{F}(z_{t+2}), \eta \Gamma_{z_{t+1}}^{z_{t+2}} \mathbf{F}(z_{t+1}) \rangle \big); \\ C := \sigma_1 (\|\eta \mathbf{F}(z_{t+1})\|^2 - \|G_{t+2}\|^2). \end{cases}$$

and analyze them term by term.

We rewrite A as

$$\begin{aligned} A =& 2\langle \Gamma_{\hat{z}_{t+2}}^{z_{t+1}} G_{t+2}, \Gamma_{\hat{z}_{t+2}}^{z_{t+1}} \exp_{\hat{z}_{t+2}}^{-1} z^* \rangle + 2\sigma_1 \langle \eta \mathbf{F}(z_{t+1}), -\eta \mathbf{F}(z_{t+1}) + \eta \Gamma_{z_t}^{z_{t+1}} \mathbf{F}(z_t) \rangle \\ =& 2\langle \Gamma_{\hat{z}_{t+2}}^{z_{t+1}} \exp_{\hat{z}_{t+2}}^{-1} z^*, \Gamma_{\hat{z}_{t+2}}^{z_{t+1}} G_{t+2} - \eta \mathbf{F}(z_{t+1}) \rangle \\ &+ 2\langle \Gamma_{\hat{z}_{t+2}}^{z_{t+1}} \exp_{\hat{z}_{t+2}}^{-1} z^* - \exp_{z_{t+1}}^{-1} z^*, \eta \mathbf{F}(z_{t+1}) \rangle \\ &+ 2\langle \exp_{z_{t+1}}^{-1} z^*, \eta \mathbf{F}(z_{t+1}) \rangle + 2\sigma_1 \langle \eta \mathbf{F}(z_{t+1}), -\eta \mathbf{F}(z_{t+1}) + \eta \Gamma_{z_t}^{z_{t+1}} \mathbf{F}(z_t) \rangle. \end{aligned} \tag{42}$$

According to Lemma 19, there exists a point p in the geodesic between z_{t+1} and $\hat{z}_{t+2}$ such that

$$\exp_{z_{t+1}}^{-1} z^* - \Gamma_{\hat{z}_{t+2}}^{z_{t+1}} \exp_{\hat{z}_{t+2}}^{-1} z^* = H_{z_{t+1}, p}^{z^*} (-\eta \mathbf{F}(z_{t+1}) + \eta \Gamma_{z_t}^{z_{t+1}} \mathbf{F}(z_t)). \tag{43}$$

Combining (42) and (43), we have

$$\begin{aligned}A = &2\langle\Gamma_{\hat{z}_{t+2}}^{\hat{z}_{t+1}}G_{t+2} - \eta\mathbf{F}(z_{t+1}), \Gamma_{\hat{z}_{t+2}}^{\hat{z}_{t+1}}\exp_{\hat{z}_{t+2}}^{-1} z^*\rangle \\ &+ 2\langle\eta\mathbf{F}(z_{t+1}), (-H_{z_{t+1},p}^{z^*} + \sigma_1 Id)[-\eta\mathbf{F}(z_{t+1})] + \eta\Gamma_{z_t}^{z_{t+1}}\mathbf{F}(z_t)\rangle \\ &+ 2\langle\eta\mathbf{F}(z_{t+1}), \exp_{z_{t+1}}^{-1} z^*\rangle.\end{aligned}$$

Since the eigenvalue of $-H_{z_{t+1},p}^{z^*}$ lies in $[-\zeta_1, -\sigma_1]$, we have

$$\begin{aligned}&\langle\eta\mathbf{F}(z_{t+1}), (-H_{z_{t+1},p}^{z^*} + \sigma_1) - \eta\mathbf{F}(z_{t+1}) + \eta\Gamma_{z_t}^{z_{t+1}}\mathbf{F}(z_t)\rangle \\ &\qquad\leq (\zeta_1 - \sigma_1)\|\eta\mathbf{F}(z_{t+1}) - \eta\Gamma_{z_t}^{z_{t+1}}\mathbf{F}(z_t)\|\|\eta\mathbf{F}(z_{t+1})\| \\ &\qquad\leq (\zeta_1 - \sigma_1)L\eta^2 d(z_{t+1}, z_t)\|\mathbf{F}(z_{t+1})\|.\end{aligned}$$

In Lemma 6,we know that $d(z_{t+1}, z_t) \leq \eta(1 + 8L\eta)\|\mathbf{F}(z_t)\|$. Putting them together, we have

$$\begin{aligned}&\langle\eta\mathbf{F}(z_{t+1}), (-H_{z_{t+1},p}^{z^*} + \sigma_1) - \eta\mathbf{F}(z_{t+1}) + \eta\Gamma_{z_t}^{z_{t+1}}\mathbf{F}(z_t)\rangle \\ &\qquad\leq (\zeta_1 - \sigma_1)L\eta^3(1 + 8L\eta)\|\mathbf{F}(z_t)\|\|\mathbf{F}(z_{t+1})\| \\ &\qquad\leq 2(\zeta_1 - \sigma_1)L\eta^3(1 + 8L\eta)\|\mathbf{F}(z_t)\|^2.\end{aligned}$$

Moreover, by Lemma 6, we have

$$\begin{aligned}\langle\Gamma_{\hat{z}_{t+2}}^{\hat{z}_{t+1}}G_{t+2} - \eta\mathbf{F}(z_{t+1}), \Gamma_{\hat{z}_{t+2}}^{\hat{z}_{t+1}}\exp_{\hat{z}_{t+2}}^{-1} z^*\rangle &\leq d(\hat{z}_{t+2}, z^*)104\eta^3\|\mathbf{F}(z_t)\|^3 \\ &\leq (D + 4L\eta^2\|\mathbf{F}(z_t)\|)104K_m\eta^3\|\mathbf{F}(z_t)\|^3.\end{aligned}$$

Hence, we have

$$\begin{aligned}A \leq &2(D + 4L\eta^2\|\mathbf{F}(z_t)\|)104\eta^3\|\mathbf{F}(z_t)\|^3 + 2(\zeta_1 - \sigma_1)L\eta^3(1 + 8L\eta)2\|\mathbf{F}(z_t)\|^2 \\ &+ 2\langle\eta\mathbf{F}(z_{t+1}), \exp_{z_{t+1}}^{-1} z^*\rangle.\end{aligned}$$

Since $\eta \leq \frac{1}{2G}$ and $\eta \leq \frac{1}{20L}$, we can find

$$A \leq 104K_m(2D + \frac{1}{5})\eta^3\|\mathbf{F}(z_t)\|^3 + \frac{28}{5}(\zeta_1 - \sigma_1)L\eta^3\|\mathbf{F}(z_t)\|^2 + 2\langle\eta\mathbf{F}(z_{t+1}), \exp_{z_{t+1}}^{-1} z^*\rangle. \tag{44}$$

Next we analyze B and C as follows.

$$\begin{aligned}B &= \eta^2\sigma_1(\|Fz_{t+2} - \Gamma_{t+1}^{t+2}\mathbf{F}(z_{t+1})\|^2 - \|\mathbf{F}(z_t)\|^2) \\ &\leq \eta^2\sigma_1(L^2 d^2(z_{t+2}, z_{t+1}) - \|\mathbf{F}(z_t)\|^2) \\ &\leq \eta^2\sigma_1(4L^2\eta^2\|\mathbf{F}(z_t)\|^2 - \|\mathbf{F}(z_t)\|^2) \\ &\leq \eta^2\sigma_1(\frac{1}{5}L\eta\|\mathbf{F}(z_t)\|^2 - \|\mathbf{F}(z_t)\|^2), \qquad (45) \\ C &\leq \sigma_1 64\eta^6\|\mathbf{F}(z_t)\|^6 K_m \leq 8\sigma_1 K_m\eta^3\|\mathbf{F}(z_t)\|^3. \qquad (46)\end{aligned}$$

Finally, taking (44), (45) and (46) in (41), we have

$$\begin{aligned}&\phi_{t+2}-\phi_{t+1}\leq\\&\eta^2\|\mathbf{F}(z_t)\|^2\Big(104K_m(2D+\frac{1}{5})\eta\|\mathbf{F}(z_t)\|+\frac{28}{5}(\zeta_1-\sigma_1)L\eta+\sigma_1K_m\eta\|\mathbf{F}(z_t)\|+\frac{\sigma_1L}{5}\eta-\sigma_1\Big)\\&+2\langle\eta\mathbf{F}(z_{t+1}),\exp_{z_{t+1}}^{-1}z^*\rangle.\end{aligned}\tag{47}$$

Because f is g-convex-concave, we have

$$2\langle\eta\mathbf{F}(z_{t+1}),\exp_{z_{t+1}}^{-1}z^*\rangle\leq 0,$$

By g-G-Lipschitz, we have

$$\phi_{t+2}-\phi_{t+1}\leq\eta^2\|\mathbf{F}(z_t)\|^2\Big((104K_m(2D+\frac{1}{5})G+\frac{28}{5}(\zeta_1-\sigma_1)L+\sigma_1K_mG+\frac{\sigma_1L}{5})\eta-\sigma_1\Big).$$

Since

$$\eta\leq\frac{\sigma_1}{2(104K_m(2D+\frac{1}{5})G+\frac{28}{5}(\zeta_1-\sigma_1)L+\sigma_1K_mG+\frac{\sigma_1L}{5})},$$

we have

$$\phi_{t+2}-\phi_{t+1}\leq-\frac{\sigma_1}{2}\eta^2\|\mathbf{F}(z_t)\|^2.\tag{48}$$

Summing (48) from $t=0,1,2\ldots$ yields

$$\sum_{t=1}^{\infty}\|\mathbf{F}(z_t)\|^2\leq\frac{2}{\sigma_1\eta^2}\phi_1\leq\frac{2D^2}{\sigma_1\eta^2},$$

which immediately indicates that

$$\min_{t\leq T}\|\nabla\mathbf{f}(z_t)\|\leq\frac{2D}{\eta\sqrt{\sigma_1T}},$$

and

$$\lim_{t\to\infty}\|\nabla\mathbf{f}(z_t)\|=0.$$

In this way, we have completed our proof. ■

Proof of Theorem 7 The g-strongly convexity-strongly concavity implies

$$2\langle\eta\mathbf{F}(z_{t+1}),\exp_{z_{t+1}}^{-1}z^*\rangle\leq-\mu\eta d^2(z_{t+1},z^*).$$

By Young's inequality, we have

$$\begin{aligned}2\langle\eta\mathbf{F}(z_{t+1}),\exp_{z_{t+1}}^{-1}z^*\rangle&\leq-\mu\eta d^2(z_{t+1},z^*)\\&\leq\mu\eta d^2(z_{t+1},\hat{z}_{t+2})-\frac{\mu\eta}{2}d^2(\hat{z}_{t+2},z^*).\end{aligned}\tag{49}$$

Putting (49) into (47) and $\Upsilon = 104K_m(2D + \frac{1}{5})G + \frac{28}{5}(\zeta_1 - \sigma_1)L + \sigma_1 K_m G + \frac{\sigma_1 L}{5}$, we write

$$\begin{aligned}\phi_{t+2} - \phi_{t+1} \leq & \eta^2\|\mathbf{F}(z_t)\|^2(\Upsilon\eta - \sigma_1) + \mu\eta d^2(z_{t+1}, \hat{z}_{t+2}) - \frac{\mu\eta}{2}d^2(\hat{z}_{t+2}, z^*) \\ = & \eta^2\|\mathbf{F}(z_t)\|^2(\Upsilon\eta - \sigma_1) + \mu\eta d^2(z_{t+1}, \hat{z}_{t+2}) - \frac{\mu\eta}{2}\phi_{t+2} \\ & + \frac{\mu\eta}{2}\sigma_1\eta^2\|\mathbf{F}(z_{t+2}) - \Gamma_{z_{t+1}}^{z_{t+2}}\mathbf{F}(z_{t+1})\|^2. \end{aligned} \tag{50}$$

From the proof Lemma 6, we have

$$\begin{cases} d(z_{t+1}, \hat{z}_{t+2}) \leq 2\eta\|\mathbf{F}(z_t)\| \\ \eta\|\mathbf{F}(z_{t+2}) - \Gamma_{z_{t+1}}^{z_{t+2}}\mathbf{F}(z_{t+1})\| = d(z_{t+2}, \hat{z}_{t+3}) \leq 2\eta\|\mathbf{F}(z_{t+1})\| \leq 4\eta\|\mathbf{F}(z_t)\|. \end{cases} \tag{51}$$

Taking (51) in (50), we have

$$(1 + \frac{\mu\eta}{2})\phi_{t+2} - \phi_{t+1} \leq \eta^2\|\mathbf{F}(z_t)\|^2(\Upsilon\eta + 4\mu\eta + 8\sigma_1\eta - \sigma_1).$$

Since

$$\eta \leq \frac{\sigma_1}{\Upsilon + 4\mu + 8\sigma\mu},$$

we have

$$\phi_{t+2} \leq \frac{1}{1 + \mu\eta/2}\phi_{t+1}, \tag{52}$$

which give us

$$\begin{cases} \sigma_1 d^2(z_t, \hat{z}_{t+1}) \leq \phi_t \leq (\frac{1}{1+\mu\eta/2})^t\phi_1 \\ d^2(\hat{z}_{t+1}, z^*) \leq \phi_{t+1} \leq (\frac{1}{1+\mu\eta/2})^{t+1}\phi_1. \end{cases}$$

Finally, we get

$$\begin{aligned} d^2(z_t, z^*) & \leq 2(d^2(z_t, \hat{z}_{t+1}) + d^2(\hat{z}_{t+1}, z^*)) \\ & \leq 2(1 + \frac{1}{\sigma_1})(\frac{1}{1 + \mu\eta/2})^t\phi_1 \\ & = 2(1 + \frac{1}{\sigma_1})(\frac{1}{1 + \mu\eta/2})^t d^2(z_1, z^*), \end{aligned}$$

which completes our proof.

References

Jacob D. Abernethy, Peter L. Bartlett, Alexander Rakhlin, and Ambuj Tewari. Optimal stragies and minimax lower bounds for online convex games. In *Annual Conference on Learning Theory*, pages 415–424. Omnipress, 2008.

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 *Conference on Learning Theory*, pages 84–118. PMLR, 2020.

Foivos Alimisis, Antonio Orvieto, Gary Bécigneul, and Aurelien Lucchi. A continuous-time perspective for modeling acceleration in Riemannian optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 1297–1307. PMLR, 2020.

Foivos Alimisis, Antonio Orvieto, Gary Becigneul, and Aurelien Lucchi. Momentum improves optimization on Riemannian manifolds. In *International Conference on Artificial Intelligence and Statistics*, pages 1351–1359. PMLR, 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 *ACM SIGACT Symposium on Theory of Computing*, pages 172–181, 2018.

Ben Andrews and Christopher Hopper. *The Ricci flow in Riemannian geometry: a complete proof of the differentiable 1/4-pinching sphere theorem.* springer, 2010.

Kimon Antonakopoulos, E. Veronica Belmega, and Panayotis Mertikopoulos. Online and stochastic optimization beyond Lipschitz continuity: A Riemannian approach. In *International Conference on Learning Representations*, 2020.

Sébastien Arnold, Pierre-Antoine Manzagol, Reza Babanezhad Harikandeh, Ioannis Mitliagkas, and Nicolas Le Roux. Reducing the variance in online optimization by transporting past gradients. In *Advances in Neural Information Processing Systems*, volume 32, pages 2260–2268. PMLR, 2019.

Gary Becigneul and Octavian-Eugen Ganea. Riemannian adaptive optimization methods. In *International Conference on Learning Representations*, pages 3262–3271, 2019.

Nicolas Boumal, Bamdev Mishra, P.-A. Absil, and Rodolphe Sepulchre. Manopt, a matlab toolbox for optimization on manifolds. *Journal of Machine Learning Research*, 15(42): 1455–1459, 2014.

Yang Cai, Argyris Oikonomou, and Weiqiang Zheng. Tight last-iterate convergence of the extragradient method for constrained monotone variational inequalities. *arXiv preprint arXiv:2204.09228*, 2022.

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

Tatjana Chavdarova, Michael I Jordan, and Manolis Zampetakis. Last-iterate convergence of saddle point optimizers via high-resolution differential equations. *arXiv preprint arXiv:2112.13826*, 2021.

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

Shiing-Shen Chern, Weihuan Chen, and Kai Shue Lam. *Lectures on differential geometry*, volume 1. World Scientific, 1999.

Chao-Kai Chiang, Tianbao Yang, Chia-Jung Lee, Mehrdad Mahdavi, Chi-Jen Lu, Rong Jin, and Shenghuo Zhu. Online optimization with gradual variations. In *Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learning Research*, pages 1–20. PMLR, 2012.

Emil Cornea, Hongtu Zhu, Peter Kim, and Joseph G Ibrahim. Regression models on Riemannian symmetric spaces. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 79(2):463, 2017.

Taosha Fan, Hanlin Wang, Michael Rubenstein, and Todd Murphey. Cpl-slam: Efficient and certifiably correct planar graph-based slam using the complex number representation. *IEEE Transactions on Robotics*, 36(6):1719–1737, 2020.

Orizon Pereira Ferreira, LR Pérez, and Sándor Zoltán Németh. Singularities of monotone vector fields and an extragradient-type algorithm. *Journal of Global Optimization*, 31(1): 133–151, 2005.

Gauthier Gidel, Hugo Berard, Gaëtan Vignoud, Pascal Vincent, and Simon Lacoste-Julien. A variational inequality perspective on generative adversarial networks. In *International Conference on Learning Representations*, 2019.

Eduard Gorbunov, Adrien Taylor, and Gauthier Gidel. Last-iterate convergence of optimistic gradient method for monotone variational inequalities. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 21858–21870. Curran Associates, Inc., 2022.

Andi Han, Bamdev Mishra, Pratik Jawanpuria, Pawan Kumar, and Junbin Gao. Riemannian Hamiltonian methods for min-max optimization on manifolds. *arXiv preprint arXiv:2204.11418*, 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 *International Conference on Computational Learning Theory*, pages 499–513. Springer, 2006.

Yi Hong, Roland Kwitt, Nikhil Singh, Brad Davis, Nuno Vasconcelos, and Marc Niethammer. Geodesic regression on the Grassmannian. In *European Conference on Computer Vision*, pages 632–646. Springer, 2014.

Inbal Horev, Florian Yger, and Masashi Sugiyama. Geometry-aware principal component analysis for symmetric positive definite matrices. In *Asian Conference on Machine Learning*, pages 1–16. PMLR, 2016.

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.

Zihao Hu, Guanghui Wang, and Jacob Abernethy. Minimizing dynamic regret on geodesic metric spaces. *arXiv preprint arXiv:2302.08652*, 2023.

Feihu Huang and Shangqian Gao. Gradient descent ascent for minimax problems on riemannian manifolds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Ali Jadbabaie, Alexander Rakhlin, Shahin Shahrampour, and Karthik Sridharan. Online Optimization : Competing with Dynamic Comparators. In *International Conference on Artificial Intelligence and Statistics*, pages 398–406, San Diego, California, USA, 2015. PMLR.

Michael Jordan, Tianyi Lin, and Emmanouil-Vasileios Vlatakis-Gkaragkounis. First-order algorithms for min-max optimization in geodesic metric spaces. In *Advances in Neural Information Processing Systems*, pages 6557–6574. Curran Associates, Inc., 2022.

Masoud Badiei Khuzani and Na Li. Stochastic primal-dual method on riemannian manifolds of bounded sectional curvature. In *International Conference on Machine Learning and Applications*, pages 133–140. IEEE, 2017.

Galina M Korpelevich. The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756, 1976.

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 *Computer Vision and Pattern Recognition*, volume 1, pages 852–859. IEEE, 2005.

Chong Li, Genaro López, and Victoria Martín-Márquez. Monotone vector fields and the proximal point algorithm on Hadamard manifolds. *Journal of the London Mathematical Society*, 79(3):663–683, 2009.

Xin Li, Xiang Zhong, Haidong Shao, Te Han, and Changqing Shen. Multi-sensor gearbox fault diagnosis by using feature-fusion covariance matrix and multi-Riemannian kernel ridge regression. *Reliability Engineering & System Safety*, 216:108018, 2021.

Tianyi Lin, Chenyou Fan, Nhat Ho, Marco Cuturi, and Michael Jordan. Projection robust Wasserstein distance and Riemannian optimization. *Advances in neural information processing systems*, 33:9383–9397, 2020.

Qi Liu, Maximilian Nickel, and Douwe Kiela. Hyperbolic graph neural networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

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.

Panayotis Mertikopoulos, Bruno Lecouat, Houssam Zenati, Chuan-Sheng Foo, Vijay Chandrasekhar, and Georgios Piliouras. Optimistic mirror descent in saddle-point problems: Going the extra(-gradient) mile. In *International Conference on Learning Representations*, 2019.

Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil. A unified analysis of extragradient and optimistic gradient methods for saddle point problems: Proximal point approach. In *International Conference on Artificial Intelligence and Statistics*, pages 1497–1507. PMLR, 2020a.

Aryan Mokhtari, Asuman E Ozdaglar, and Sarath Pattathil. Convergence rate of O(1/k) for optimistic gradient and extragradient methods in smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 30(4):3230–3251, 2020b.

Prasanna Muralidharan and P Thomas Fletcher. Sasaki metrics for analysis of longitudinal data on manifolds. In *2012 IEEE conference on computer vision and pattern recognition*, pages 1027–1034. IEEE, 2012.

Alexander Rakhlin and Karthik Sridharan. Online learning with predictable sequences. In *Conference on Learning Theory*, pages 993–1019. PMLR, 2013.

Alois Schlögl, Felix Lee, Horst Bischof, and Gert Pfurtscheller. Characterization of four-class motor imagery eeg data for the bci-competition 2005. *Journal of neural engineering*, 2 (4):L14, 2005.

Ha-Young Shin and Hee-Seok Oh. Robust geodesic regression. *International Journal of Computer Vision*, 130(2):478–503, 2022.

JH Wang, G López, Victoria Martín-Márquez, and Chong Li. Monotone and accretive vector fields on Riemannian manifolds. *Journal of optimization theory and applications*, 146(3): 691–708, 2010.

Xi Wang, Zhipeng Tu, Yiguang Hong, Yingyi Wu, and Guodong Shi. Online optimization over Riemannian manifolds. *Journal of Machine Learning Research*, 24(84):1–67, 2023.

Chen-Yu Wei, Chung-Wei Lee, Mengxiao Zhang, and Haipeng Luo. Last-iterate convergence of decentralized optimistic gradient descent/ascent in infinite-horizon competitive markov games. In *Conference on learning theory*, pages 4259–4299. PMLR, 2021.

Yongxin Yang and Timothy M Hospedales. Multivariate regression on the Grassmannian for predicting novel domains. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5071–5080, 2016.

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

Lijun Zhang, Shiyin Lu, and Zhi-Hua Zhou. Adaptive online learning in dynamic environments. *Advances in neural information processing systems*, 31, 2018.

Peiyuan Zhang, Jingzhao Zhang, and Suvrit Sra. Minimax in geodesic metric spaces: Sion's theorem and algorithms. *arXiv preprint arXiv:2202.06950*, 2022.

Peng Zhao, Yu-Jie Zhang, Lijun Zhang, and Zhi-Hua Zhou. Dynamic regret of convex and smooth functions. *Advances in Neural Information Processing Systems*, 33:12510–12520, 2020.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *International Conference on International Conference on Machine Learning*, page 928–935. AAAI Press, 2003.