# Extragradient Type Methods for Riemannian Variational Inequality Problems

**Zihao Hu** Georgia Tech
**Guanghui Wang** Georgia Tech
**Xi Wang** AMSS, China
**Andre Wibisono** Yale University
**Jacob Abernethy** Georgia Tech, Google Research
**Molei Tao** Georgia Tech

## Abstract

In this work, we consider monotone Riemannian Variational Inequality Problems (RVIPs), which encompass both Riemannian convex optimization and minimax optimization as particular cases. In Euclidean space, the last-iterates of both the extragradient (EG) and past extragradient (PEG) methods converge to the solution of monotone variational inequality problems at a rate of $O\left(\frac{1}{\sqrt{T}}\right)$ (Cai et al., 2022). However, analogous behavior on Riemannian manifolds remains open. To bridge this gap, we introduce the Riemannian extragradient (REG) and Riemannian past extragradient (RPEG) methods. We show that both exhibit $O\left(\frac{1}{\sqrt{T}}\right)$ last-iterate convergence and $O\left(\frac{1}{T}\right)$ average-iterate convergence, aligning with observations in the Euclidean case. These results are enabled by judiciously addressing the holonomy effect so that additional complications in Riemannian cases can be reduced and the Euclidean proof inspired by the performance estimation problem (PEP) technique or the sum-of-squares (SOS) technique can be applied again.

## 1 Introduction

Variational inequality problems (VIPs) (Kinderlehrer and Stampacchia, 2000; Facchinei and Pang, 2003) play a pivotal role in mathematical programming, encompassing areas such as convex optimization and minimax optimization. Specifically, for the Euclidean space under the unconstrained setting, the objective of a VIP is to find $\mathbf{z}^*$ satisfying:

$$\langle F(\mathbf{z}^*), \mathbf{z} - \mathbf{z}^* \rangle \geq 0 \qquad \forall \mathbf{z} \in \mathbb{R}^d,$$

where $F : \mathbb{R}^d \to \mathbb{R}^d$ represents an operator. In particular, by setting $\mathbf{z} = \mathbf{z}^* - \eta F(\mathbf{z}^*)$, the solution to an unconstrained VIP can be reduced to identifying the zeros of $F(\cdot)$. At first glance, it might seem intuitive to employ gradient descent (GD) given by $\mathbf{z}_{t+1} = \mathbf{z}_t - \eta F(\mathbf{z}_t)$ to solve a VIP. However, while GD shows convergence for convex optimization tasks, it can, unfortunately, diverge for even monotone VIPs[1], irrespective of the step-sizes chosen (Facchinei and Pang, 2003). Fortunately, sophisticated methods such as extragradient (EG) (Korpelevich, 1976):

$$\begin{aligned} \tilde{\mathbf{z}}_t &= \mathbf{z}_t - \eta F(\mathbf{z}_t) \\ \mathbf{z}_{t+1} &= \mathbf{z}_t - \eta F(\tilde{\mathbf{z}}_t), \end{aligned} \tag{1}$$

and past extragradient (PEG) (Popov, 1980):

$$\begin{aligned} \tilde{\mathbf{z}}_t &= \mathbf{z}_t - \eta F(\tilde{\mathbf{z}}_{t-1}) \\ \mathbf{z}_{t+1} &= \mathbf{z}_t - \eta F(\tilde{\mathbf{z}}_t). \end{aligned} \tag{2}$$

offer solutions to this hurdle. In the unconstrained domain, PEG is equivalent to the optimistic gradient descent ascent (OGDA) technique:

$$\tilde{\mathbf{z}}_{t+1} = \tilde{\mathbf{z}}_t - 2\eta F(\tilde{\mathbf{z}}_t) + \eta F(\tilde{\mathbf{z}}_{t-1}).$$

Since extragradient type methods are easy to implement, relatively scalable to dimension, and have demonstrated pleasant empirical performance, they have become the standard tools for addressing VIPs and saddle point problems over the past few decades (Tseng, 2000; Gidel et al., 2018; Hsieh et al., 2019). At present, we understand that for convex-concave saddle point problems, both EG and PEG achieve $O\left(\frac{1}{T}\right)$ *average-iterate* convergence (Nemirovski, 2004;

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238. 

[1] Monotone VIP means the operator $F$ is monotone: $\langle F(\mathbf{z}) - F(\mathbf{z}'), \mathbf{z} - \mathbf{z}' \rangle \geq 0$ holds for any $\mathbf{z}, \mathbf{z}' \in \mathbb{R}^d$.

Table 1: Comparison of our results and previous extragradient type methods on the Riemannian gsc-convex gsc-concave saddle point problems, where $\zeta$ and $\sigma$ are geometric constants arising from Riemannian cosine laws (Zhang and Sra, 2016; Alimisis et al., 2020). Notably, we achieve the first non-asymptotic last-iterate convergence for Riemannian extragradient type methods.

| Algorithm | Results |
|---|---|
| RCEG (Zhang et al., 2023) | average-iterate: $O\left(\sqrt{\frac{\zeta}{\sigma}}\cdot\frac{1}{T}\right)$, best-iterate: $O\left(\frac{\sqrt{\zeta}}{\sigma}\cdot\frac{1}{\sqrt{T}}\right)$ |
| ROGDA (Wang et al., 2023) | average-iterate: $O\left(\frac{\zeta}{\sigma}\cdot\frac{1}{T}\right)$, best-iterate: $O\left(\frac{\zeta}{\sqrt{\sigma^3}}\cdot\frac{1}{\sqrt{T}}\right)$ |
| REG (Theorem 2) | average-iterate: $O\left(\frac{\zeta}{\sigma}\cdot\frac{1}{T}\right)$, last-iterate: $O\left(\frac{\zeta}{\sqrt{\sigma^3}}\cdot\frac{1}{\sqrt{T}}\right)$ |
| RPEG (Theorem 4) | average-iterate: $O\left(\frac{\zeta}{\sigma}\cdot\frac{1}{T}\right)$, last-iterate: $O\left(\frac{\zeta}{\sqrt{\sigma^3}}\cdot\frac{1}{\sqrt{T}}\right)$ |

Mokhtari et al., 2020) and $O\left(\frac{1}{\sqrt{T}}\right)$ *last-iterate* convergence (Golowich et al., 2020b,a; Gorbunov et al., 2022a,b; Cai et al., 2022).[2] The last-iterate convergence, although slower than the average-iterate convergence, offers two distinct advantages: (i) the last-iterate convergence is an appropriate performance metric even for non-convex-concave games, a condition not met by the average-iterate convergence due to the absence of Jensen's inequality; (ii) in practical scenarios like GAN training, the last-iterate exhibits strong empirical results (Daskalakis et al., 2018; Chavdarova et al., 2019).

Meanwhile, in recent years, Riemannian convex optimization and minimax optimization have attracted considerable interest (Zhang and Sra, 2016; Alimisis et al., 2020; Ahn and Sra, 2020; Kim and Yang, 2022; Zhang et al., 2023; Jordan et al., 2022; Martínez-Rubio et al., 2023). However, Riemannian variational inequality problems (RVIPs), the generalized counterpart of Riemannian convex optimization and minimax optimization, remain relatively underexplored. In this paper, we consider extragradient type methods for the following RVIP: identify $\mathbf{z}^*$ such that

$$\left\langle F(\mathbf{z}^*), \operatorname{Exp}_{\mathbf{z}^*}^{-1}\mathbf{z}\right\rangle \geq 0, \qquad \forall \mathbf{z}\in\mathcal{M}.$$

Here, $\mathcal{M}$ denotes a $d$-dimensional Riemannian manifold and $F(\cdot)$ is a vector field defined on $\mathcal{M}$. For Riemannian convex-concave saddle point problems, Zhang et al. (2023) introduce the Riemannian corrected extragradient (RCEG), while Wang et al. (2023) propose the Riemannian optimistic gradient descent ascent (ROGDA), both achieving $O\left(\frac{1}{T}\right)$ average-iterate convergence. However, the non-asymptotic *last-iterate* convergence of these methods remains unexplored except for strongly geodesically convex-concave problems (Jordan et al., 2022; Wang et al., 2023). This naturally raises the question:

*Do there exist Riemannian analogs of EG and PEG that concurrently exhibit non-asymptotic average-iterate and last-iterate convergence behaviors?*

In this study, we confirm this question and outline the following contributions.

- We introduce Riemannian extragradient (REG) and Riemannian past extragradient (RPEG) as novel first-order methods tailored for monotone RVIPs.
- Both REG and RPEG, as detailed in Theorems 1 and 3, exhibit $O\left(\frac{1}{\sqrt{T}}\right)$ last-iterate convergence in the context of monotone RVIPs.
- In the realm of Riemannian minimax optimization, both REG and RPEG achieve $O\left(\frac{1}{\sqrt{T}}\right)$ last iterate convergence and $O\left(\frac{1}{T}\right)$ average-iterate convergence, as delineated in Table 1.

For completeness, we have included the proof of the best-iterate convergence for RCEG in Appendix 9.1. It is beneficial to compare the convergence rates of the algorithms listed in Table 1, focusing on the geometric constants $\zeta$ and $\sigma$. Given that $\zeta \geq \sigma$, the average-iterate convergence rate of RCEG algorithm comes with a more favorable geometric constant compared to our REG algorithm. Meanwhile, the geometric constants of ROGDA show similarities to those of our REG and RPEG algorithms. This similarity may arise because all those methods rely on bounding the holonomy distortion. An intriguing open question is how to enhance the last-iterate convergence of REG and RPEG by optimizing the curvature constants.

We now outline the primary technical challenges and our solutions. State-of-the-art proofs validating the last-iterate convergence of EG and PEG are generally inspired by either the performance estimation problem

[2]The average-iterate and the last-iterate consider the convergence of $\bar{\mathbf{z}}_T \coloneqq \frac{1}{T}\sum_{t=1}^{T}\tilde{\mathbf{z}}_t$ and $\mathbf{z}_T$, respectively.

(PEP) approach (Gorbunov et al., 2022a,b) or the sum-of-squares (SOS) technique (Cai et al., 2022). At their core, both methods cast the estimation of an optimization algorithm's convergence rate as an optimization problem, subsequently obtaining a numerical solution through sequential convex relaxation. This solution inherently offers insights into crafting a proof. However, applying the PEP and SOS techniques directly to the manifold setting proves difficult, largely because they intrinsically depend on the "interpolation condition" (Taylor et al., 2017), which is inherently tied to the Euclidean space, not geodesic metric spaces[3].

Yet, this challenge does not inherently restrict us from leveraging the insights gleaned from the PEP or SOS methods. In our study, we meticulously craft REG and RPEG based on these insights, demonstrating that validating the last-iterate convergence in manifold contexts can largely mirror the proofs in the Euclidean setting, provided that the *holonomy effect*[4] is handled with care. We posit that this novel approach not only resolves our immediate challenges but could potentially benefit a broader range of Riemannian optimization problems in the future.

## 2 Related Work

In this section, we briefly review prior research on extragradient type algorithms in the Euclidean space and Riemannian minimax optimization.

**Extragradient Type Methods in Euclidean Space.** Nemirovski (2004) demonstrate an $O\left(\frac{1}{T}\right)$ average-iterate convergence rate of EG with respect to the primal-dual gap. A comparable result for OGDA is documented by Mokhtari et al. (2020). Building on additional assumptions, specifically the Lipschitz continuity of the Jacobian of $F$, Golowich et al. (2020b,a) are the first to establish an $O\left(\frac{1}{\sqrt{T}}\right)$ last-iterate convergence rate for EG and PEG. This milestone is further developed by Gorbunov et al. (2022a), who bypass the aforementioned assumption and showcase an $O\left(\frac{1}{\sqrt{T}}\right)$ last-iterate convergence using the PEP technique. Various works such as Cai et al. (2022) and Gorbunov et al. (2022b) extend these results to constrained settings. One intriguing discrepancy that emerges is between the average-iterate convergence $O\left(\frac{1}{T}\right)$ and the last-iterate convergence $O\left(\frac{1}{\sqrt{T}}\right)$ rates. It raises the question: are there accelerated first-order methods that can achieve $O\left(\frac{1}{T}\right)$ last-iterate convergence? Drawing inspiration from the Halpern iteration (Lieder, 2021), Yoon and Ryu (2021) introduce the extra anchored gradient (EAG), which achieves $O\left(\frac{1}{T}\right)$ last-iterate convergence rate.

**Riemannian Minimax Optimization.** Zhang et al. (2023) propose RCEG, which achieves $O\left(\frac{1}{T}\right)$ average-iterate convergence for geodesically convex-concave problems. Our contributions, REG and RPEG, surpass this by attaining both $O\left(\frac{1}{\sqrt{T}}\right)$ last-iterate and $O\left(\frac{1}{T}\right)$ average-iterate convergence rates. Hu et al. (2023) consider how to make RCEG work in the online improper learning setting. Han et al. (2023) put forth the Riemannian Hamiltonian gradient descent, which exhibits linear convergence under the Riemannian Polyak-Lojasiewicz condition. However, this condition predominantly applies to strongly geodesically convex-concave problems. Other notable contributions include Jordan et al. (2022), who demonstrate linear last-iterate convergence of RCEG for strongly convex-concave settings, and Wang et al. (2023), who propose ROGDA. This method achieves $O\left(\frac{1}{T}\right)$ average-iterate convergence and linear last-iterate convergence in convex-concave and strongly convex-concave settings, respectively. It should be noted that while Wang et al. (2023) rely on quantifying the holonomy effect in geodesic quadrilaterals, our approach capitalizes on the analogous effect in geodesic triangles. As we will see later, this is a key observation that enables us to establish the last-iterate convergence within the Riemannian context. The Riemannian gradient descent ascent (RGDA) in the deterministic setting and the stochastic setting has been considered by Jordan et al. (2022); Huang and Gao (2023). By leveraging the idea of converting minimax optimization to sequential strongly convex optimization problems, Martínez-Rubio et al. (2023) introduce doubly-looped algorithms specifically designed for Hadamard manifolds. Furthermore, the algorithms proposed in Martínez-Rubio et al. (2023) demonstrate accelerated convergence rates and inherently address the constrained scenario. The quest for singly-looped Riemannian first-order minimax optimization algorithms featuring accelerated rates continues to be a compelling area of investigation. Cai et al. (2023) show a curvature-independent linear last-iterate convergence of Riemannian gradient descent for strongly monotone RVIPs. Our approach offers a slower $O\left(\frac{1}{\sqrt{T}}\right)$ last-iterate convergence rate, but applies to general monotone RVIPs.

## 3 Preliminaries

In this section, we provide an overview of Riemannian geometry, RVIPs, and Riemannian minimax optimiza-

[3]For a deeper exploration of the "geodesically convex interpolation", readers are referred to Criscitiello and Boumal (2023, Section 8).

[4]To put it simply, after a vector undergoes parallel transport along a geodesic loop, the end result differs from its original form, a phenomenon termed holonomy.

tion. We also introduce some assumptions that are necessary to establish our key results.

### 3.1 Riemannian Geometry

We outline the foundational concepts of Riemannian geometry here. For a more in-depth exposition, the reader is directed to Petersen (2006); Lee (2018). We consider a $d$-dimensional smooth manifold $\mathcal{M}$, a topological space where every point possesses an open neighborhood that can be smoothly mapped to an open subset in $\mathbb{R}^d$. For each point $\mathbf{x}$ on the manifold $\mathcal{M}$, there are $d$ directions (tangent vectors). Beginning at $\mathbf{x}$ and moving infinitesimally in any of these directions remains within $\mathcal{M}$. The tangent space at $\mathbf{x}$, symbolized as $T_{\mathbf{x}}\mathcal{M}$, is a vector space comprising all these tangent vectors. A Riemannian manifold is a smooth manifold equipped with a continuously differentiable Riemannian metric. For any point $\mathbf{x} \in \mathcal{M}$, this metric allows the calculation of the inner product $\langle \mathbf{u}, \mathbf{v} \rangle_{\mathbf{x}}$ and the magnitude $\|\mathbf{u}\|_{\mathbf{x}} = \sqrt{\langle \mathbf{u}, \mathbf{u} \rangle_{\mathbf{x}}}$ for tangent vectors $\mathbf{u}, \mathbf{v} \in T_{\mathbf{x}}\mathcal{M}$. We sometimes omit the reference point $\mathbf{x}$ when it is clear from the context.

A geodesic segment connecting two points $\mathbf{x}, \mathbf{y} \in \mathcal{M}$ is a constant-speed curve that locally minimizes the distance between $\mathbf{x}$ and $\mathbf{y}$, serving as a natural extension of line segments in Euclidean space. Formally, $\gamma(t) : [0,1] \to \mathcal{M}$ represents a geodesic segment, with $\gamma(0) = \mathbf{x}$, $\gamma(1) = \mathbf{y}$, and its initial velocity given by $\dot{\gamma}(0) = \mathbf{v} \in T_{\mathbf{x}}\mathcal{M}$. The exponential map, $\mathrm{Exp}_{\mathbf{x}}(\cdot)$, transitions from a tangent space to the manifold, while the inverse exponential map, $\mathrm{Exp}_{\mathbf{x}}^{-1}(\cdot)$, maps from the manifold to a tangent vector. For the aforementioned geodesic segment $\gamma(t)$, $\mathrm{Exp}_{\mathbf{x}}\mathbf{v} = \mathbf{y}$ and $\mathrm{Exp}_{\mathbf{x}}^{-1}\mathbf{y} = \mathbf{v}$ hold true. The Riemannian distance, $d(\mathbf{x}, \mathbf{y})$, quantifies the geodesic distance between $\mathbf{x}$ and $\mathbf{y}$. Given the constant speed of the geodesic, we have $d(\mathbf{x}, \mathbf{y}) = \|\mathrm{Exp}_{\mathbf{x}}^{-1}\mathbf{y}\|$. For two distinct points $\mathbf{x}$ and $\mathbf{y}$ and a tangent vector $\mathbf{u} \in T_{\mathbf{x}}\mathcal{M}$, the parallel transport operation, denoted by $\Gamma_{\mathbf{x}}^{\mathbf{y}}\mathbf{u}$, smoothly shifts $\mathbf{u}$ from being an element in $T_{\mathbf{x}}\mathcal{M}$ to being an element in $T_{\mathbf{y}}\mathcal{M}$ via the geodesic joining $\mathbf{x}$ and $\mathbf{y}$. This operation preserves both the inner product and the norm of the tangent vector.

The curvature of a Riemannian manifold is precisely defined by the Riemannian curvature tensor. For practicality, the sectional curvature is used more frequently in machine learning (Zhang and Sra, 2016; Ahn and Sra, 2020; Kim and Yang, 2022). The sectional curvature at any point $\mathbf{x} \in \mathcal{M}$ relies on all 2-planes in $T_{\mathbf{x}}\mathcal{M}$. On Riemannian manifolds with positive, zero, and negative sectional curvatures, geodesics that start out parallel will, respectively, converge, remain parallel, and diverge. A class of Riemannian manifolds of particular interest is the Hadamard manifolds, which are complete and simply connected spaces with non-positive curvature. There, any pair of distinct points are connected by a globally length-minimizing geodesic.

A *geodesically-convex* (gsc-convex) set contains all length-minimizing geodesics connecting two distinct points within the set. Let $\mathcal{N} \subseteq \mathcal{M}$ be a gsc-convex set. A function $f : \mathcal{N} \to \mathbb{R}$ is termed gsc-convex if and only if $f$ is convex when restricted to any geodesic with the minimum length connecting two distinct points in $\mathcal{N}$. Formally, for all geodesic paths $\gamma(t) \subseteq \mathcal{N}$,

$$f(\gamma(t)) \leq (1-t)f(\gamma(0)) + tf(\gamma(1)).$$

For differentiable functions, geodesic convexity can be represented as:

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathrm{Exp}_{\mathbf{x}}^{-1}\mathbf{y} \rangle, \quad \forall \mathbf{x}, \mathbf{y} \in \mathcal{N},$$

where $\nabla f(\mathbf{x}) \in T_{\mathbf{x}}\mathcal{M}$ is the Riemannian gradient. Consequently, the notion of geodesically-smooth (gsc-smooth) functions emerges. A function $f$ is termed $L$-gsc-smooth if

$$\|\nabla f(\mathbf{x}) - \Gamma_{\mathbf{y}}^{\mathbf{x}} \nabla f(\mathbf{y})\| \leq L \cdot d(\mathbf{x}, \mathbf{y}),$$

or equivalently, for all $\mathbf{x}, \mathbf{y} \in \mathcal{N}$:

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathrm{Exp}_{\mathbf{x}}^{-1}\mathbf{y} \rangle + \tfrac{L}{2} d(\mathbf{x}, \mathbf{y})^2.$$

The *holonomy* effect, originating from the curvature of the Riemannian manifold, captures the "turning" of a vector as it undergoes parallel transport around a geodesic loop. As depicted in Figure 1, beginning with a tangent vector $\mathbf{u} \in T_{\mathbf{x}}\mathcal{M}$, and translating $\mathbf{u}$ along geodesic segments $\overline{\mathbf{xy}}$, $\overline{\mathbf{yz}}$ and $\overline{\mathbf{zx}}$, upon returning to $\mathbf{x}$, the vector $\Gamma_{\mathbf{z}}^{\mathbf{x}}\Gamma_{\mathbf{y}}^{\mathbf{z}}\Gamma_{\mathbf{x}}^{\mathbf{y}}\mathbf{u}$, though still in $T_{\mathbf{x}}\mathcal{M}$, deviates in direction from the initial tangent vector $\mathbf{u}$. Although quantifying the holonomy effect for general geodesic loops is complex, this work demonstrates that approximating the holonomy effect on a geodesic triangle is sufficient to ascertain the last-iterate convergence of Riemannian extragradient type methods.

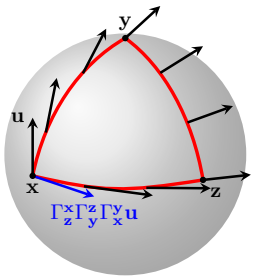


Figure 1: An illustration of the holonomy effect on a sphere.

### 3.2 RVIPs and Riemannian Minimax Optimization

Recall the definition of the Riemannian VIP is to seek a point $\mathbf{z}^*$ such that

$$\langle F(\mathbf{z}^*), \mathrm{Exp}_{\mathbf{z}^*}^{-1}\mathbf{z}\rangle \geq 0, \quad \forall \mathbf{z} \in \mathcal{M}, \tag{3}$$

which is equivalent to finding $\mathbf{z}^*$ for which $F(\mathbf{z}^*) = 0$ by substituting $\mathbf{z} = \mathrm{Exp}_{\mathbf{z}^*}(-\eta F(\mathbf{z}^*))$. We demonstrate that Riemannian convex optimization and minimax optimization are special cases of Riemannian VIPs in the sequel.

A Riemannian convex optimization problem is given by

$$\min_{\mathbf{z}\in\mathcal{M}} f(\mathbf{z}), \tag{4}$$

where $f(\mathbf{z})$ is a gsc-convex function on a Riemannian manifold $\mathcal{M}$. The corresponding RVIP is obtained by selecting $F(\mathbf{z}) = \nabla f(\mathbf{z})$. Riemannian convex optimization finds applications in operator scaling (Allen-Zhu et al., 2018), Gaussian mixture models (Hosseini and Sra, 2015), and the calculation of the Fréchet mean (Lou et al., 2020). Another additional remark is, most algorithms for convex optimization still work in general, nonconvex settings, and it is just that their quantitative convergence guarantees may not carry through. General Riemannian optimizations lead to even more applications, such as large scale eigenvalue/PCA/SVD problems Tao and Ohsawa (2020), generic improvement of transformer and approximation of optimal transport (Wasserstein) distance in high dimensions Kong et al. (2022).

A Riemannian minimax problem can be articulated as

$$\min_{\mathbf{x}\in\mathcal{M}_1}\max_{\mathbf{y}\in\mathcal{M}_2} f(\mathbf{x},\mathbf{y}), \tag{5}$$

where $\mathcal{M}_1$ and $\mathcal{M}_2$ are Riemannian manifolds and $f$ is gsc-convex in $\mathbf{x}$ and gsc-concave in $\mathbf{y}$. Examples in this category encompass Riemannian constrained convex optimization, robust geometry-aware PCA, and robust matrix Fréchet mean computation (Zhang et al., 2023; Jordan et al., 2022). Adopting $\mathbf{z} = \begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix}$ and $F(\mathbf{z}) = \begin{pmatrix}\nabla_{\mathbf{x}} f(\mathbf{x},\mathbf{y})\\ -\nabla_{\mathbf{y}} f(\mathbf{x},\mathbf{y})\end{pmatrix}$, it is evident that $\mathbf{z}^* = \begin{pmatrix}\mathbf{x}^*\\ \mathbf{y}^*\end{pmatrix}$, the RVIP solution, stands as a saddle point of $f(\mathbf{x},\mathbf{y})$.

For (unconstrained) RVIP which seeks a $\mathbf{z}^*$ such that $F(\mathbf{z}^*) = 0$, the norm $\|F(\mathbf{z}_t)\|$ serves as the convergence criterion, aligning with standard proofs of last-iterate convergence in the Euclidean domain (Golowich et al., 2020b; Gorbunov et al., 2022a; Cai et al., 2022). However, for constrained situations, the norm $\|F(\mathbf{z}_t)\|$ is unsuitable since the RVIP solution does not inherently ensure $F(\mathbf{z}^*) = 0$. In Euclidean space, an alternative concept known as the "tangent residual" has been shown to be effective in constrained cases, as detailed in (Cai et al., 2022). It would be intriguing to explore the applicability of this technique within the Riemannian context.

For Riemannian saddle point problems, a convergence measure analogous to $\|F(\mathbf{z}_t)\|$ is the Riemannian Hamiltonian:

**Definition 1** (Riemannian Hamiltonian)**.** *For a geodesically convex-concave objective $f$ defined on $\mathcal{M}_1 \times \mathcal{M}_2$, the Riemannian Hamiltonian at $(\mathbf{x},\mathbf{y})$ is*

$$Ham_f(\mathbf{x},\mathbf{y}) \coloneqq \|\nabla_{\mathbf{x}} f(\mathbf{x},\mathbf{y})\|^2 + \|\nabla_{\mathbf{y}} f(\mathbf{x},\mathbf{y})\|^2 = \|F(\mathbf{z})\|^2.$$

An alternative convergence metric introduces the Riemannian primal-dual gap:

**Definition 2** (Riemannian Primal-dual Gap)**.** *Assume $f : \mathcal{M}_1 \times \mathcal{M}_2 \to \mathbb{R}$ is geodesically convex-concave, and sets $\mathcal{X} \subseteq \mathcal{M}_1$ and $\mathcal{Y} \subseteq \mathcal{M}_2$ are gsc-convex and compact. The primal-dual gap on $\mathcal{X}\times\mathcal{Y}$ is defined as:*

$$Gap_f^{\mathcal{X}\times\mathcal{Y}}(\mathbf{x},\mathbf{y}) = \max_{\mathbf{y}'\in\mathcal{Y}} f(\mathbf{x},\mathbf{y}') - \min_{\mathbf{x}'\in\mathcal{X}} f(\mathbf{x}',\mathbf{y}).$$

This gap measures how the utility derived by one participant changes upon unilateral action modification. It is imperative to define the primal-dual gap on compact sets $\mathcal{X}\times\mathcal{Y}$ to ensure the quantity is bounded.

### 3.3 Assumptions

In this part, we provide some key assumptions that will be essential to our later results.

**Assumption 1.** *Let $\mathcal{M}$ be a $d$-dimensional complete and simply connected Riemannian manifold with sectional curvature lower bounded by $\kappa$ and upper bounded by $K$. Assume $F(\cdot)$ is a vector field on $\mathcal{M}$ and $F(\mathbf{z}) = 0$ admits a solution $\mathbf{z}^*$. We denote $D$ as an upper bound of $d(\mathbf{z}_0,\mathbf{z}^*)$. When $K > 0$, we require that $D \leq \frac{4\pi}{9\sqrt{K}}$.*

**Definition 3.** *Under Assumption 1, we define $K_m = \max\{|\kappa|,|K|\}$ and $\mathcal{D} = \{\mathbf{z} | d(\mathbf{z},\mathbf{z}^*) \leq \frac{6D}{5}\}$. We also define $\bar{\sigma} = \sigma\left(K, \frac{91D}{81}\right)$ and $\bar{\zeta} = \zeta\left(\kappa, \frac{7D}{5}\right)$, where $\sigma(K,\cdot)$ and $\zeta(\kappa,\cdot)$ are geometric constants defined in Lemmas 21 and 22 (Appendix 9.2), respectively.*

**Remark 1.** *Note that $D < \frac{\pi}{2\sqrt{K}}$ is required to guarantee the unique geodesic property (Cheeger et al., 1975, Theorem 5.14) and our condition $D \leq \frac{4\pi}{9\sqrt{K}}$ is substantially close to $D < \frac{\pi}{2\sqrt{K}}$.*

**Assumption 2.** *$F(\cdot)$ is monotone on $\mathcal{D}$:*

$$\left\langle \Gamma_{\mathbf{z}'}^{\mathbf{z}} F(\mathbf{z}') - F(\mathbf{z}), \mathrm{Exp}_{\mathbf{z}}^{-1}\mathbf{z}'\right\rangle \geq 0.$$

**Assumption 3.** *$F(\cdot)$ is $L$-Lipschitz on the manifold $\mathcal{M}$, which means*

$$\|F(\mathbf{z}) - \Gamma_{\mathbf{z}'}^{\mathbf{z}} F(\mathbf{z}')\| \leq L \cdot d(\mathbf{z},\mathbf{z}').$$

**Assumption 4.** *On the set $\mathcal{D}$, the norm of $F(\cdot)$ is bounded:*

$$\|F(\mathbf{z})\| \leq G.$$

**Remark 2.** *We utilize the fact that $F(\mathbf{z}^*) = 0$, which leads to*

$$\|F(\mathbf{z})\| = \|F(\mathbf{z}) - \Gamma_{\mathbf{z}^*}^{\mathbf{z}} F(\mathbf{z}^*)\| \leq L \cdot d(\mathbf{z}, \mathbf{z}^*).$$

*If the algorithm's trajectory remains bounded, it directly implies an upper bound on $\|F(\mathbf{z})\|$. Initiating with $\mathbf{z}_0$ where $d(\mathbf{z}_0, \mathbf{z}^*) \leq D$, it can be verified (see Corollaries 1 and 2 for details) that all iterates of both REG and RPEG are confined within*

$$\mathcal{D} = \left\{ \mathbf{z} \mid d(\mathbf{z}, \mathbf{z}^*) \leq \tfrac{6D}{5} \right\},$$

*allowing us to dispense with Assumption 4 without sacrificing generality.*

## 4 Riemannian Extragradient

In this part, we discuss the last-iterate and average-iterate convergence of the Riemannian extragradient method. The details of the proof for this section are deferred to Appendix 7. As a precursor, we revisit the proof in the Euclidean space.

### 4.1 Warm-up: the Euclidean Case

The *extragradient* (EG) method is pivotal for tackling saddle point and variational inequality problems. Given an operator $F : \mathbb{R}^d \to \mathbb{R}^d$, the EG algorithm updates as follows at each iteration:

$$\mathbf{z}_{t+1} = \mathbf{z}_t - \eta F(\tilde{\mathbf{z}}_t) \coloneqq \mathbf{z}_t - \eta F(\mathbf{z}_t - \eta F(\mathbf{z}_t)). \quad (6)$$

When $F(\cdot)$ is monotone and $L$-Lipschitz, as detailed in Assumptions 5 and 6, EG enjoys $O\left(\frac{1}{T}\right)$ average-iterate convergence (Nemirovski, 2004; Mokhtari et al., 2020) and $O\left(\frac{1}{\sqrt{T}}\right)$ last-iterate convergence rate (Golowich et al., 2020b; Gorbunov et al., 2022a; Cai et al., 2022).

**Assumption 5.** *$F(\cdot)$ is monotone, which means*

$$\langle F(\mathbf{z}) - F(\mathbf{z}'), \mathbf{z} - \mathbf{z}' \rangle \geq 0 \quad \forall \mathbf{z}, \mathbf{z}' \in \mathbb{R}^d.$$

**Assumption 6.** *$F(\cdot)$ is $L$-Lipschitz, which means*

$$\|F(\mathbf{z}) - F(\mathbf{z}')\| \leq L\|\mathbf{z} - \mathbf{z}'\| \quad \forall \mathbf{z}, \mathbf{z}' \in \mathbb{R}^d.$$

The analysis of the last-iterate convergence of EG contains two parts:

- (Best-iterate convergence) There exists $t' \in [T]$:

$$\|F(\mathbf{z}_{t'})\| \leq O\left(\tfrac{1}{\sqrt{T}}\right).$$

- (Non-increasing operator norm) For any $t \in [T]$,

$$\|F(\mathbf{z}_{t+1})\| \leq \|F(\mathbf{z}_t)\|.$$

While the best-iterate convergence of the extragradient method is well-established (Korpelevich, 1976; Nemirovski, 2004), the proof on the monotonicity of the operator norm is far from obvious. By introducing an additional assumption that the Jacobian of $F$ is $\Lambda$-Lipschitz, Golowich et al. (2020b) demonstrate a marginally weaker inequality:

$$\|F(\mathbf{z}_{t+1})\| \leq (1 + \epsilon_t)\|F(\mathbf{z}_t)\|.$$

where $\epsilon_t$ is small, thus establishing the $O\left(\frac{1}{\sqrt{T}}\right)$ last-iterate convergence. Subsequent studies by Gorbunov et al. (2022a); Cai et al. (2022) eliminate this additional assumption by resorting to the Performance Estimation Problem (PEP) and the Sum-of-Squares (SOS) technique. The following lemma lays the groundwork for these findings.

**Lemma 1.** *(Cai et al., 2022) Suppose*

$$\begin{aligned} &0 \leq \langle F(\mathbf{z}_t) - F(\mathbf{z}_{t+1}), F(\tilde{\mathbf{z}}_t) \rangle, \\ &\|F(\tilde{\mathbf{z}}_t) - F(\mathbf{z}_{t+1})\|^2 \leq L^2\eta^2 \|F(\tilde{\mathbf{z}}_t) - F(\mathbf{z}_t)\|^2 \end{aligned} \quad (7)$$

*hold, where $L^2\eta^2 \leq 1$, then $\|F(\mathbf{z}_{t+1})\| \leq \|F(\mathbf{z}_t)\|$.*

In Euclidean case, Equation (7) can be derived from Equation (6), Assumptions 5 and 6.

### 4.2 Riemannian Extragradient and Convergence Rates

A natural inquiry arises: what is the last-iterate convergence of the Riemannian extragradient for monotone RVIPs? In this section, we demonstrate that the answer remains $O\left(\frac{1}{\sqrt{T}}\right)$.

Intuitively, we aim to identify an analog of Equation (7) in the Riemannian setting. To this end, we propose Riemannian extragradient (REG):

$$\begin{aligned} \tilde{\mathbf{z}}_t &= \mathrm{Exp}_{\mathbf{z}_t}(-\eta F(\mathbf{z}_t)) \\ \mathbf{z}_{t+1} &= \mathrm{Exp}_{\mathbf{z}_t}(-\eta \Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_t} F(\tilde{\mathbf{z}}_t)), \end{aligned} \quad (8)$$

which employs the parallel transport to respect the Riemannian metric. In the high level, given REG's update rule, we can establish a Riemannian analog of Equation (7) and invoke Lemma 1 to show the operator norm $\|F(\mathbf{z}_t)\|$ is still non-increasing.

We first delve into the demonstration of the $O\left(\frac{1}{\sqrt{T}}\right)$ best-iterate convergence of REG. Since Assumptions 2 and 4 are only valid on $\mathcal{D}$ rather than the entire manifold $\mathcal{M}$, it is crucial to ensure that all iterates of REG

remain bounded. This fact is demonstrated with the aid of Lemma 2 and Corollary 1. Apart from establishing the boundedness of $\mathbf{z}_t$, Lemma 2 also suggests an $O\left(\frac{1}{\sqrt{T}}\right)$ best-iterate convergence of REG.

**Lemma 2.** *Under Assumptions 1, 3 and 4. For the iterates of REG as in Equation* (8)*,*

$$d(\mathbf{z}_t, \mathbf{z}^*) \leq D$$

*holds for any* $t \geq 0$ *and*

$$\eta \leq \min\left\{\frac{1}{\sqrt{8L^2+306K_mG^2}}, \frac{\bar{\sigma}}{56\sqrt{K_m}DL+8\bar{\zeta}L+\bar{\sigma}L}\right\}.$$

*Also, we can show there exists* $t' \in [T]$ *such that*

$$\|F(\mathbf{z}_{t'})\| = O\left(\frac{\bar{\zeta}}{\sqrt{\bar{\sigma}^3 T}}\right).$$

Therefore, under mild conditions, we can ensure that the trajectory of REG remains bounded.

**Corollary 1.** *Under Assumptions 1, 3 and 4. For the iterates of REG as in Equation* (8)*,* $\mathbf{z}_t, \tilde{\mathbf{z}}_t \in \mathcal{D}$ *holds for any* $t \geq 0$ *and*

$$\eta \leq \min\left\{\frac{1}{\sqrt{8L^2+306K_mG^2}}, \frac{\bar{\sigma}}{56\sqrt{K_m}DL+8\bar{\zeta}L+\bar{\sigma}L}\right\},$$

*where* $\mathcal{D}$ *is defined in Definition 3.*

To establish the last-iterate convergence of REG, we need to show the operator norm $\|F(\mathbf{z}_t)\|$ does not increase. The following lemma, which quantifies the holonomy effect on a geodesic triangle, plays a crucial role in achieving that objective.

**Lemma 3.** *(Informal) For a small geodesic triangle* $\triangle\mathbf{xyz}$ *on a Riemannian manifold* $\mathcal{M}$ *and* $\mathbf{u} \in T_{\mathbf{x}}\mathcal{M}$*, we have*

$$\|\Gamma_{\mathbf{z}}^{\mathbf{x}}\Gamma_{\mathbf{y}}^{\mathbf{z}}\Gamma_{\mathbf{x}}^{\mathbf{y}}\mathbf{u} - \mathbf{u}\| = O(1)\cdot\|\mathbf{u}\|\cdot d(\mathbf{y},\mathbf{z}).$$

**Remark 3.** *It is beneficial to recognize a more potent implication of Lemma 3:*

$$\|\Gamma_{\mathbf{z}}^{\mathbf{x}}\Gamma_{\mathbf{y}}^{\mathbf{z}}\Gamma_{\mathbf{x}}^{\mathbf{y}}\mathbf{u} - \mathbf{u}\| = O(1)\cdot\min\{d(\mathbf{x},\mathbf{y}), d(\mathbf{x},\mathbf{z}), d(\mathbf{y},\mathbf{z})\}$$

*holds for a small geodesic triangle, given that parallel transport preserves the norm of a tangent vector.*

A formal description and proof of Lemma 3 can be found in Appendix 9.3. For REG, with a small stepsize $\eta$, we demonstrate in the subsequent lemma that the distortion due to the holonomy effect is small and the operator norm does not increase.

**Lemma 4.** *Under Assumptions 1, 2, 3, 4. For the iterates of REG as in Equation* (8)*, we can show*

$$\|F(\mathbf{z}_{t+1})\| \leq \|F(\mathbf{z}_t)\|,$$

*by choosing*

$$\eta \leq \min\left\{\frac{1}{\sqrt{8L^2+306K_mG^2}}, \frac{\bar{\sigma}}{56\sqrt{K_m}DL+8\bar{\zeta}L+\bar{\sigma}L}\right\}.$$

Now we are ready to present the last-iterate convergence of REG.

**Theorem 1.** *Under Assumptions 1, 2, 3, 4. For the iterates of REG as in Equation* (8)*, we can choose*

$$\eta = \min\left\{\frac{1}{\sqrt{8L^2+306K_mG^2}}, \frac{\bar{\sigma}}{56\sqrt{K_m}DL+8\bar{\zeta}L+\bar{\sigma}L}\right\}$$

*to achieve* $O\left(\frac{1}{\sqrt{T}}\right)$ *last-iterate convergence for the monotone variational problem with an operator* $F$*.*

**Remark 4.** *It is instructive to compare the step-size from Theorem 1 against that of Gorbunov et al. (2022a). When we consider the Riemannian manifold as an Euclidean space, our required step-size is* $\eta \leq \frac{1}{8L}$*, whereas Gorbunov et al. (2022a) prescribes* $\eta \leq \frac{1}{\sqrt{2}L}$*. This disparity hints that the constants in our findings may not be the tightest.*

When we consider Riemannian saddle point problems, Theorem 2 shows REG attains $O\left(\frac{1}{\sqrt{T}}\right)$ last-iterate convergence and $O\left(\frac{1}{T}\right)$ average-iterate convergence, simultaneously.

**Theorem 2.** *Consider a Riemannian minimax optimization problem*

$$\min_{\mathbf{x}\in\mathcal{M}_1}\max_{\mathbf{y}\in\mathcal{M}_2} f(\mathbf{x},\mathbf{y})$$

*where* $f$ *is geodesically convex-concave. Let* $\mathcal{M} := \mathcal{M}_1 \times \mathcal{M}_2$*,* $\mathbf{z}^* := \binom{\mathbf{x}^*}{\mathbf{y}^*}$ *to be the saddle point, and* $\mathcal{X} = \mathcal{B}\left(\mathbf{x}^*, \frac{\sqrt{2}D}{2}\right)$*,* $\mathcal{Y} = \mathcal{B}\left(\mathbf{y}^*, \frac{\sqrt{2}D}{2}\right)$ *to be geodesic balls. Under Assumptions 1, 2, 3, 4, if we apply REG in Equation* (8) *with*

$$\eta = \min\left\{\frac{1}{\sqrt{8L^2+306K_mG^2}}, \frac{\bar{\sigma}}{56\sqrt{K_m}DL+8\bar{\zeta}L+\bar{\sigma}L}\right\},$$

*then*

$$\max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}_T,\mathbf{y}) - \min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x},\mathbf{y}_T) = O\left(\frac{\bar{\zeta}}{\sqrt{\bar{\sigma}^3 T}}\right),$$

*and*

$$\max_{\mathbf{y}\in\mathcal{Y}} f(\bar{\mathbf{x}}_T,\mathbf{y}) - \min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x},\bar{\mathbf{y}}_T) = O\left(\frac{\bar{\zeta}}{\sqrt{\bar{\sigma}^3 T}}\right),$$

*where* $\bar{\mathbf{x}}_T = \mathrm{Exp}_{\bar{\mathbf{x}}_{T-1}}\left(\frac{1}{T}\mathrm{Exp}^{-1}_{\bar{\mathbf{x}}_{T-1}}\tilde{\mathbf{x}}_T\right)$ *and* $\bar{\mathbf{y}}_T = \mathrm{Exp}_{\bar{\mathbf{y}}_{T-1}}\left(\frac{1}{T}\mathrm{Exp}^{-1}_{\bar{\mathbf{y}}_{T-1}}\tilde{\mathbf{y}}_T\right)$ *are the geodesic ergodic averages of* $\tilde{\mathbf{x}}_t$ *and* $\tilde{\mathbf{y}}_t$ *for* $t = 1,\ldots,T$*.*

Up to now, a natural question that might arise is whether the technique regarding holonomy distortion suffices to establish the last-iterate convergence for RCEG (Zhang et al., 2023) or ROGDA (Wang et al., 2023). We explored the possibility of establishing the last-iterate convergence of RCEG Appendix 7.7. The

issue appears to involve a new distortion term and we use an alternative update $\mathbf{z}_{t+1} = \mathrm{Exp}_{\mathbf{z}_t}(-\eta\Gamma^{\mathbf{z}_t}_{\tilde{\mathbf{z}}_t}F(\tilde{\mathbf{z}}_t))$ in REG to eliminate it. For ROGDA, note that the proof in Wang et al. (2023) heavily relies on the holonomy effect on a geodesic quadrilateral:

$$\begin{aligned}&\|\Gamma^{\mathbf{x}}_{\mathbf{w}}\Gamma^{\mathbf{w}}_{\mathbf{z}}\Gamma^{\mathbf{z}}_{\mathbf{y}}\Gamma^{\mathbf{y}}_{\mathbf{x}}\mathbf{u}-\mathbf{u}\|\\ =&O(1)\cdot(d(\mathbf{x},\mathbf{y})+d(\mathbf{y},\mathbf{z})+d(\mathbf{w},\mathbf{x}))\cdot(d(\mathbf{y},\mathbf{z})+d(\mathbf{w},\mathbf{x}))\end{aligned}$$

In our Lemma 24, we actually consider the case of $\mathbf{w} = \mathbf{x}$, so the quadrilateral degenerates to be a geodesic triangle. This aspect is crucial in the proof of our Lemma 4, where we aim to bound the holonomy effect using a specific geodesic edge $d(\mathbf{z}_{t+1}, \tilde{\mathbf{z}}_t)$, rather than summing two edges, mirroring the approach in Euclidean cases. Given these considerations, demonstrating the last-iterate convergence of RCEG and ROGDA presents non-trivial challenges.

## 5 Riemannian Past Extragradient

In the Euclidean space, the Past Extragradient (PEG) method, introduced by Popov (1980), performs iterates as in Equation (2). One of the advantages of PEG over EG is the halving of gradient queries in each iteration. The two inequalities presented in Gorbunov et al. (2022b, Lemma 3.1) are crucial for demonstrating the last-iterate convergence of PEG:

$$\begin{aligned}&0 \le \langle F(\mathbf{z}_t) - F(\mathbf{z}_{t+1}), F(\tilde{\mathbf{z}}_t)\rangle,\\ &\|F(\tilde{\mathbf{z}}_t) - F(\mathbf{z}_{t+1})\|^2 \le L^2\eta^2\|F(\tilde{\mathbf{z}}_t) - F(\tilde{\mathbf{z}}_{t-1})\|^2.\end{aligned} \tag{9}$$

And they can be deduced from Equation (2), Assumptions 5 and 6.

With insights gained from the Euclidean space, we introduce the Riemannian Past Extragradient (RPEG):

$$\begin{aligned}\tilde{\mathbf{z}}_t &= \mathrm{Exp}_{\mathbf{z}_t}\left(-\eta\Gamma^{\mathbf{z}_t}_{\tilde{\mathbf{z}}_{t-1}}F(\tilde{\mathbf{z}}_{t-1})\right)\\ \mathbf{z}_{t+1} &= \mathrm{Exp}_{\mathbf{z}_t}\left(-\eta\Gamma^{\mathbf{z}_t}_{\tilde{\mathbf{z}}_t}F(\tilde{\mathbf{z}}_t)\right)\end{aligned} \tag{10}$$

One can readily observe that RPEG is distinct from ROGDA (Wang et al., 2023):

$$\tilde{\mathbf{z}}_{t+1} = \mathrm{Exp}_{\tilde{\mathbf{z}}_t}\left(-2\eta F(\tilde{\mathbf{z}}_t) + \eta\Gamma^{\tilde{\mathbf{z}}_t}_{\tilde{\mathbf{z}}_{t-1}}F(\tilde{\mathbf{z}}_{t-1})\right),$$

even in the unconstrained scenario, owing to the nonlinear nature of the exponential map. This distinction becomes particularly noteworthy when contrasted with the Euclidean setting.

Following Gorbunov et al. (2022b), we use a Lyapunov analysis argument to show the last-iterate convergence of RPEG. Proof details of this part are deferred to Appendix 8 due to page limitations. We first establish the following lemma, which implies $d(\mathbf{z}_t, \mathbf{z}^*) \le D$ holds for any $t \ge 0$.

**Lemma 5.** *Under Assumptions 1, 3 and 4. For the iterates of RPEG in Equation* (10) *with*

$$\eta \le \min\left\{\frac{\bar{\sigma}}{141LD\sqrt{K_m}+32\bar{\zeta}L}, \frac{1}{\sqrt{648K_mG^2}}\right\},$$

$d(\mathbf{z}_t, \mathbf{z}^*) \le D$ *holds for any* $t \ge 0$.

Similar to REG, the trajectory of RPEG is also bounded due to the following corollary.

**Corollary 2.** *Under Assumptions 1, 3 and 4. For the iterates of RPEG as in Equation* (10), $\mathbf{z}_t, \tilde{\mathbf{z}}_t \in \mathcal{D}$ *holds for any* $t \ge 0$ *and*

$$\eta \le \min\left\{\frac{\bar{\sigma}}{141LD\sqrt{K_m}+32\bar{\zeta}L}, \frac{1}{\sqrt{648K_mG^2}}\right\},$$

*where* $\mathcal{D}$ *is defined in Definition 3.*

In Gorbunov et al. (2022b), it is discussed that for PEG, the function $\|F(\mathbf{z}_t)\|$ does not monotonically decrease with respect to $t$. However, they demonstrate that

$$\begin{aligned}&\|F(\mathbf{z}_{t+1})\|^2 + 2\|F(\mathbf{z}_{t+1}) - F(\tilde{\mathbf{z}}_t)\|^2\\ \le&\|F(\mathbf{z}_t)\|^2 + 2\|F(\mathbf{z}_t) - F(\tilde{\mathbf{z}}_{t-1})\|^2\end{aligned}$$

holds when the step-size $\eta$ is small. We introduce a Riemannian counterpart to this proposition under the help of Lemma 3.

**Lemma 6.** *With Assumptions 1, 2, 3 and 4, the iterates of RPEG satisfies*

$$\begin{aligned}&\|F(\mathbf{z}_{t+1})\|^2 + 2\|F(\mathbf{z}_{t+1}) - \Gamma^{\mathbf{z}_{t+1}}_{\tilde{\mathbf{z}}_t}F(\tilde{\mathbf{z}}_t)\|^2\\ \le&\|F(\mathbf{z}_t)\|^2 + 2\|F(\mathbf{z}_t) - \Gamma^{\mathbf{z}_t}_{\tilde{\mathbf{z}}_{t-1}}F(\tilde{\mathbf{z}}_{t-1})\|^2\\ &+ \rho\|\Gamma^{\mathbf{z}_t}_{\tilde{\mathbf{z}}_t}F(\tilde{\mathbf{z}}_t) - \Gamma^{\mathbf{z}_t}_{\tilde{\mathbf{z}}_{t-1}}F(\tilde{\mathbf{z}}_{t-1})\|^2\end{aligned}$$

*for any*

$$\eta \le \min\left\{\frac{\bar{\sigma}}{141LD\sqrt{K_m}+32\bar{\zeta}L}, \frac{1}{\sqrt{648K_mG^2}}\right\}.$$

*where* $\rho \coloneqq \left(\left(24L^2 + 432K_mG^2 + 48GL\sqrt{2K_m}\right)\eta^2 - \frac{2}{3}\right)$.

Now, we are able to establish a Lyapunov analysis for RPEG in Lemma 7.

**Lemma 7.** *We define*

$$\begin{aligned}\Phi_t \coloneqq& d(\mathbf{z}_t, \mathbf{z}^*)^2\\ &+ \lambda t\eta^2\left(\|F(\mathbf{z}_t)\|^2 + 2\|F(\mathbf{z}_t) - \Gamma^{\mathbf{z}_t}_{\tilde{\mathbf{z}}_{t-1}}F(\tilde{\mathbf{z}}_{t-1})\|^2\right)\end{aligned}$$

*with* $\lambda = \frac{\bar{\sigma}}{16}$. *Under Assumptions 1, 2, 3 and 4, the iterates of RPEG as in Equation* (10) *satisfies* $\Phi_{t+1} \le \Phi_t$ *for any*

$$\eta \le \min\left\{\frac{\bar{\sigma}}{152LD\sqrt{K_m}+35\bar{\zeta}L}, \frac{1}{\sqrt{36L^2+648K_mG^2+72\sqrt{2K_m}GL}}\right\}$$

The last-iterate convergence of RPEG is achieved by combining Lemmas 5, 6, 7, so the final step-size needs to satisfy all these requirements.

**Theorem 3.** *Under Assumptions 1, 2, 3, 4. For the iterates of RPEG as in Equation* (10)*, we can achieve* $O\left(\frac{\bar{\zeta}}{\sqrt{\bar{\sigma}^3 T}}\right)$ *last-iterate convergence for monotone variational problems by choosing*

$$\eta = \min\left\{\frac{\bar{\sigma}}{152LD\sqrt{K_m}+35\bar{\zeta}L}, \frac{1}{\sqrt{36L^2+648K_mG^2+72\sqrt{2K_m}GL}}\right\}$$

Similar to REG, we can show RPEG also achieves $O\left(\frac{1}{\sqrt{T}}\right)$ last-iterate convergence and $O\left(\frac{1}{T}\right)$ average-iterate convergence, when applied to Riemannian convex-concave saddle point problems.

**Theorem 4.** *Consider a Riemannian minimax optimization problem*

$$\min_{\mathbf{x}\in\mathcal{M}_1}\max_{\mathbf{y}\in\mathcal{M}_2} f(\mathbf{x},\mathbf{y})$$

*where $f$ is geodesically convex-concave. Let $\mathcal{M} := \mathcal{M}_1 \times \mathcal{M}_2$, $\mathbf{z}^* := \binom{\mathbf{x}^*}{\mathbf{y}^*}$ to be the saddle point, and $\mathcal{X} = \mathcal{B}\left(\mathbf{x}^*, \frac{\sqrt{2}D}{2}\right)$, $\mathcal{Y} = \mathcal{B}\left(\mathbf{y}^*, \frac{\sqrt{2}D}{2}\right)$ to be geodesic balls. Under Assumptions 1, 2, 3, 4, if we apply RPEG in Equation* (10) *with*

$$\eta \le \min\left\{\frac{\bar{\sigma}}{192LD\sqrt{2K_m}+35\bar{\zeta}L}, \frac{1}{\sqrt{36L^2+648K_mG^2+72\sqrt{2K_m}GL}}\right\}$$

*then*

$$\max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}_T,\mathbf{y}) - \min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x},\mathbf{y}_T) = O\left(\frac{\bar{\zeta}}{\sqrt{\bar{\sigma}^3 T}}\right),$$

*and*

$$\max_{\mathbf{y}\in\mathcal{Y}} f(\bar{\mathbf{x}}_T,\mathbf{y}) - \min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x},\bar{\mathbf{y}}_T) = O\left(\frac{\bar{\zeta}}{\bar{\sigma} T}\right),$$

*where $\bar{\mathbf{x}}_T = \mathrm{Exp}_{\bar{\mathbf{x}}_{T-1}}\left(\frac{1}{T}\mathrm{Exp}^{-1}_{\bar{\mathbf{x}}_{T-1}}\tilde{\mathbf{x}}_T\right)$ and $\bar{\mathbf{y}}_T = \mathrm{Exp}_{\bar{\mathbf{y}}_{T-1}}\left(\frac{1}{T}\mathrm{Exp}^{-1}_{\bar{\mathbf{y}}_{T-1}}\tilde{\mathbf{y}}_T\right)$ are the ergodic averages of $\tilde{\mathbf{x}}_t$ and $\tilde{\mathbf{y}}_t$ for $t = 1,\dots,T$.*

We wrap up this section with a comparison between our work and Martínez-Rubio et al. (2023). The work of Martínez-Rubio et al. (2023) can achieve a faster $O\left(\frac{1}{T}\right)$ last-iterate convergence rate for Riemannian gsc-convex gsc-concave problems, but there are some key distinctions: (i) their algorithm employs a double loop, while our algorithms are single looped and easier to implement, (ii) for gsc-convex gsc-concave optimization problems, they rely on the reduction from the strongly gsc-convex strongly gsc-concave case, and a predefined precision is required before starting the algorithm, while our algorithms are "anytime" and can find better solutions the longer they are running, and (iii) while our last-iterate rate of $O\left(\frac{1}{\sqrt{T}}\right)$ is slower, it aligns with the lower bound for $p$-SCLI algorithms in Euclidean space as established by Golowich et al. (2020b,a).

## 6 Conclusion

In this study, we introduce Riemannian adaptations of the extragradient and past extragradient methods. We establish $O\left(\frac{1}{T}\right)$ average-iterate convergence and $O\left(\frac{1}{\sqrt{T}}\right)$ last-iterate convergence for minimax optimization on Riemannian manifolds. A cornerstone of our approach is the realization that the proof for last-iterate convergence in the Riemannian setting can be significantly simplified by transitioning to the Euclidean domain, provided the holonomy effect is carefully bounded. Looking forward, we are keen to explore achieving last-iterate convergence in constrained scenarios and pursuing $O\left(\frac{1}{T}\right)$ accelerated rate through the use of single-loop first-order algorithms. Other interesting open problems include how to improve the curvature-dependence of the last-iterate convergence rate and investigate whether it is possible to replace the exponential map with computationally more efficient retractions.

## Acknowledgments

We would like to thank five anonymous referees for their constructive comments and suggestions. JA acknowledges the AI4OPT Institute for its funding as part of NSF Award 2112533 and thanks the NSF for its support through Award IIS-1910077. MT is thankful for partial support by NSF DMS-1847802, Cullen-Peck Scholarship, and GT-Emory Humanity.AI Award.

### References

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

Yang Cai, Argyris Oikonomou, and Weiqiang Zheng. Finite-time last-iterate convergence for learning in multi-player games. *Advances in Neural Information Processing Systems*, 35:33904–33919, 2022.

Yang Cai, Michael I Jordan, Tianyi Lin, Argyris Oikonomou, and Emmanouil-Vasileios Vlatakis-Gkaragkounis. Curvature-independent last-iterate convergence for games on riemannian manifolds. *arXiv preprint arXiv:2306.16617*, 2023.

Tatjana Chavdarova, Gauthier Gidel, Fran**c**cois Fleuret, and Simon Lacoste-Julien. Reducing noise in gan training with variance reduced extragradient. *Advances in Neural Information Processing Systems*, 32, 2019.

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.

Jeff Cheeger, David G Ebin, and David Gregory Ebin. *Comparison theorems in Riemannian geometry*, volume 9. North-Holland publishing company Amsterdam, 1975.

Christopher Criscitiello and Nicolas Boumal. Curvature and complexity: Better lower bounds for geodesically convex optimization. In Gergely Neu and Lorenzo Rosasco, editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 2969–3013. PMLR, 12–15 Jul 2023. URL `https://proceedings.mlr.press/v195/criscitiello23a.html`.

Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training gans with optimism. In *International Conference on Learning Representations*, 2018.

Francisco Facchinei and Jong-Shi Pang. *Finite-dimensional variational inequalities and complementarity problems*. Springer, 2003.

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*, 2018.

Noah Golowich, Sarath Pattathil, and Constantinos Daskalakis. Tight last-iterate convergence rates for no-regret learning in multi-player games. *In Advances in neural information processing systems*, 33:20766–20778, 2020a.

Noah Golowich, Sarath Pattathil, Constantinos Daskalakis, and Asuman Ozdaglar. Last iterate is slower than averaged iterate in smooth convex-concave saddle point problems. In *Conference on Learning Theory*, pages 1758–1784. PMLR, 2020b.

Eduard Gorbunov, Nicolas Loizou, and Gauthier Gidel. Extragradient method: O (1/k) last-iterate convergence for monotone variational inequalities and connections with cocoercivity. In *International Conference on Artificial Intelligence and Statistics*, pages 366–402. PMLR, 2022a.

Eduard Gorbunov, Adrien Taylor, and Gauthier Gidel. Last-iterate convergence of optimistic gradient method for monotone variational inequalities. *Advances in neural information processing systems*, 35: 21858–21870, 2022b.

Andi Han, Bamdev Mishra, Pratik Jawanpuria, Pawan Kumar, and Junbin Gao. Riemannian hamiltonian methods for min-max optimization on manifolds. *SIAM Journal on Optimization*, 33(3):1797–1827, 2023.

Reshad Hosseini and Suvrit Sra. Matrix manifold optimization for gaussian mixtures. *In Advances in Neural Information Processing Systems*, 28:910–918, 2015.

Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. On the convergence of single-call stochastic extra-gradient methods. *Advances in Neural Information Processing Systems*, 32, 2019.

Zihao Hu, Guanghui Wang, and Jacob D Abernethy. Minimizing dynamic regret on geodesic metric spaces. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 4336–4383. PMLR, 2023.

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

Michael Jordan, Tianyi Lin, and Emmanouil-Vasileios Vlatakis-Gkaragkounis. First-order algorithms for min-max optimization in geodesic metric spaces. *Advances in Neural Information Processing Systems*, 35:6557–6574, 2022.

Hermann Karcher. Riemannian center of mass and mollifier smoothing. *Communications on pure and applied mathematics*, 30(5):509–541, 1977.

Jungbin Kim and Insoon Yang. Accelerated gradient methods for geodesically convex optimization: Tractable algorithms and convergence analysis. In *International Conference on Machine Learning*, pages 11255–11282. PMLR, 2022.

David Kinderlehrer and Guido Stampacchia. *An introduction to variational inequalities and their applications*. SIAM, 2000.

Lingkai Kong, Yuqing Wang, and Molei Tao. Momentum stiefel optimizer, with applications to suitably-orthogonal attention, and optimal transport. In *International Conference on Learning Representations (ICLR)*, 2022.

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*, volume 2. Springer, 2018.

Felix Lieder. On the convergence rate of the halpern-iteration. *Optimization letters*, 15(2):405–418, 2021.

Aaron Lou, Isay Katsman, Qingxuan Jiang, Serge Belongie, Ser-Nam Lim, and Christopher De Sa. Differentiating through the fréchet mean. In *International Conference on Machine Learning*, pages 6393–6403. PMLR, 2020.

David Martínez-Rubio, Christophe Roux, Christopher Criscitiello, and Sebastian Pokutta. Accelerated methods for riemannian min-max optimization ensuring bounded geometric penalties. *arXiv preprint arXiv:2305.16186*, 2023.

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, 2020.

Arkadi Nemirovski. Prox-method with rate of convergence o (1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

Peter Petersen. *Riemannian geometry*, volume 171. Springer, 2006.

Leonid Denisovich Popov. A modification of the arrow-hurwicz method for search of saddle points. *Mathematical notes of the Academy of Sciences of the USSR*, 28:845–848, 1980.

Yue Sun, Nicolas Flammarion, and Maryam Fazel. Escaping from saddle points on riemannian manifolds. *In Advances in Neural Information Processing Systems*, 32:7276–7286, 2019.

Molei Tao and Tomoki Ohsawa. Variational optimization on lie groups, with examples of leading (generalized) eigenvalue problems. In *International Conference on Artificial Intelligence and Statistics*, pages 4269–4280. PMLR, 2020.

Adrien B Taylor, Julien M Hendrickx, and Fran**c**cois Glineur. Smooth strongly convex interpolation and exact worst-case performance of first-order methods. *Mathematical Programming*, 161:307–345, 2017.

Paul Tseng. A modified forward-backward splitting method for maximal monotone mappings. *SIAM Journal on Control and Optimization*, 38(2):431–446, 2000.

Xi Wang, Deming Yuan, Yiguang Hong, Zihao Hu, Lei Wang, and Guodong Shi. Riemannian optimistic algorithms. *arXiv preprint arXiv:2308.16004*, 2023.

TaeHo Yoon and Ernest K Ryu. Accelerated algorithms for smooth convex-concave minimax problems with o (1/k^ 2) rate on squared gradient norm. In *International Conference on Machine Learning*, pages 12098–12109. PMLR, 2021.

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

Peiyuan Zhang, Jingzhao Zhang, and Suvrit Sra. Sion's minimax theorem in geodesic metric spaces and a riemannian extragradient algorithm. *SIAM Journal on Optimization*, 33(4):2885–2908, 2023. doi: 10.1137/22M1505475. URL `https://doi.org/10.1137/22M1505475`.

## Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Not Applicable]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

# Extragradient Type Methods for Riemannian Variational Inequality Problems: Supplementary Materials

## 7 Omitted Proof for Section 4

### 7.1 Proof of Lemma 1

For simplicity, let $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$ be $F(\mathbf{z}_t)$, $F(\tilde{\mathbf{z}}_t)$ and $F(\mathbf{z}_{t+1})$ respectively. We indeed have

$$0 \leq \lambda \langle \mathbf{a} - \mathbf{c}, \mathbf{b} \rangle$$

and

$$\|\mathbf{b} - \mathbf{c}\|^2 \leq L^2\eta^2 \|\mathbf{b} - \mathbf{a}\|^2$$

for any $\lambda \geq 0$. Add the above two inequalities and rearrange, we obtain

$$L^2\eta^2\|\mathbf{a}\|^2 + (L^2\eta^2 - 1)\|\mathbf{b}\|^2 - \|\mathbf{c}\|^2 + (2-\lambda)\langle \mathbf{b}, \mathbf{c}\rangle + \left(\lambda - 2L^2\eta^2\right)\langle \mathbf{a}, \mathbf{b}\rangle \geq 0.$$

Since $L^2\eta^2 \leq 1$, there exists $\lambda \leq 2$ such that $\lambda - 2L^2\eta^2 \geq 0$. Applying Young's inequality on the cross product term, we have

$$\begin{aligned}
0 \leq & L^2\eta^2\|\mathbf{a}\|^2 + (L^2\eta^2 - 1)\|\mathbf{b}\|^2 - \|\mathbf{c}\|^2 + (2-\lambda)\langle \mathbf{b}, \mathbf{c}\rangle + \left(\lambda - 2L^2\eta^2\right)\langle \mathbf{a}, \mathbf{b}\rangle \\
\leq & L^2\eta^2\|\mathbf{a}\|^2 + (L^2\eta^2 - 1)\|\mathbf{b}\|^2 - \|\mathbf{c}\|^2 + (2-\lambda)\frac{\|\mathbf{b}\|^2 + \|\mathbf{c}\|^2}{2} + \left(\lambda - 2L^2\eta^2\right)\frac{\|\mathbf{a}\|^2 + \|\mathbf{b}\|^2}{2} \\
= & \frac{\lambda}{2}\|\mathbf{a}\|^2 - \frac{\lambda}{2}\|\mathbf{c}\|^2.
\end{aligned}$$

Hence, $\|\mathbf{c}\| \leq \|\mathbf{a}\|$, which means

$$\|F(\mathbf{z}_{t+1})\| \leq \|F(\mathbf{z}_t)\|$$

as asserted.

### 7.2 Auxillary Lemmas on the Iterates of REG

To demonstrate the $O\left(\frac{1}{\sqrt{T}}\right)$ best-iterate convergence of REG, the succeeding three lemmas prove to be instrumental in this regard. Lemma 8 indicates that $d(\tilde{\mathbf{z}}_t, \mathbf{z}_{t+1}) = O(\eta^2)$, while Lemma 9 elucidates the relationship between $\|F(\mathbf{z}_t)\|$ and $\|F(\tilde{\mathbf{z}}_t)\|$. Lemma 10 is a helpful lemma for proving that all iterates of REG remain bounded.

**Lemma 8.** *Under Assumptions 1, 3. For the iterates of REG in Equation* (8)*, suppose $\eta$ is chosen to ensure* $\max\{d(\mathbf{z}_t, \tilde{\mathbf{z}}_t), d(\mathbf{z}_t, \mathbf{z}_{t+1})\} \leq \frac{1}{\sqrt{K_m}}$*, then we have*

$$d(\tilde{\mathbf{z}}_t, \mathbf{z}_{t+1}) \leq 2L\eta^2\|F(\mathbf{z}_t)\|.$$

*Proof.* We have

$$\begin{aligned}
& d(\tilde{\mathbf{z}}_t, \mathbf{z}_{t+1}) \\
\leq & 2\|\mathrm{Exp}_{\mathbf{z}_t}^{-1}\tilde{\mathbf{z}}_t - \mathrm{Exp}_{\mathbf{z}_t}^{-1}\mathbf{z}_{t+1}\| = 2\eta \cdot \|F(\mathbf{z}_t) - \Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_t}F(\tilde{\mathbf{z}}_t)\| \\
\leq & 2\eta L d(\mathbf{z}_t, \tilde{\mathbf{z}}_t) = 2L\eta^2\|F(\mathbf{z}_t)\|.
\end{aligned}$$

where the first inequality is due to Lemma 17 and the second one is due to Assumption 3. □

**Lemma 9.** *Under Assumptions 1, 3. For the iterates of REG in Equation* (8)*, we have*

$$(1 - L\eta)\|F(\mathbf{z}_t)\| \leq \|F(\tilde{\mathbf{z}}_t)\| \leq (1 + L\eta)\|F(\mathbf{z}_t)\|.$$

*Proof.* Due to the triangle inequality,

$$\|F(\mathbf{z}_t)\| + \|\Gamma_{\mathbf{z}_t}^{\tilde{\mathbf{z}}_t} F(\mathbf{z}_t) - F(\tilde{\mathbf{z}}_t)\| \leq \|F(\tilde{\mathbf{z}}_t)\| = \|\Gamma_{\mathbf{z}_t}^{\tilde{\mathbf{z}}_t} F(\mathbf{z}_t) - \Gamma_{\mathbf{z}_t}^{\tilde{\mathbf{z}}_t} F(\mathbf{z}_t) + F(\tilde{\mathbf{z}}_t)\| \leq \|F(\mathbf{z}_t)\| + \|\Gamma_{\mathbf{z}_t}^{\tilde{\mathbf{z}}_t} F(\mathbf{z}_t) - F(\tilde{\mathbf{z}}_t)\|.$$

According to Assumption 3, we have

$$\|\Gamma_{\mathbf{z}_t}^{\tilde{\mathbf{z}}_t} F(\mathbf{z}_t) - F(\tilde{\mathbf{z}}_t)\| \leq L \cdot d(\mathbf{z}_t, \tilde{\mathbf{z}}_t) = \eta L \|F(\mathbf{z}_t)\|.$$

By combining the two aforementioned inequalities, the proof of the lemma is complete. □

**Lemma 10.** *Considering the iterates of REG as given in Equation* (8) *and with a step-size of* $\eta \leq \frac{1}{9L}$*, if we assume that* $d(\mathbf{z}_t, \mathbf{z}^*) \leq D$ *and Assumptions 1 and 3 are valid, then the following results hold:*

$$d(\mathbf{z}_{t+1}, \mathbf{z}^*) \leq \frac{91D}{81}$$

*and*

$$d(\tilde{\mathbf{z}}_t, \mathbf{z}^*) \leq \frac{10D}{9}.$$

*We can also obtain* $\mathbf{z}_t, \mathbf{z}_{t+1}, \tilde{\mathbf{z}}_t \in \mathcal{D}$ *holds where* $\mathcal{D}$ *is defined in Definition 3.*

*Proof.* This lemma can be proved via a combination of the triangle inequality, Lemma 9 and Assumption 3. For $d(\mathbf{z}_{t+1}, \mathbf{z}^*)$:

$$\begin{aligned} &d(\mathbf{z}_{t+1}, \mathbf{z}^*) \leq d(\mathbf{z}_t, \mathbf{z}^*) + d(\mathbf{z}_{t+1}, \mathbf{z}_t) \\ \leq &D + \eta\|F(\tilde{\mathbf{z}}_t)\| \leq D + \eta(1 + \eta L)\|F(\mathbf{z}_t)\| \\ \leq &D + \frac{10}{9}\eta DL \leq D + \frac{10}{9}\frac{1}{9L}DL = \frac{91D}{81}. \end{aligned}$$

We can bound $d(\tilde{\mathbf{z}}_t, \mathbf{z}^*)$ in a similar way:

$$\begin{aligned} &d(\tilde{\mathbf{z}}_t, \mathbf{z}^*) \leq d(\mathbf{z}_t, \mathbf{z}^*) + d(\tilde{\mathbf{z}}_t, \mathbf{z}_t) \\ \leq &D + \eta\|F(\mathbf{z}_t)\| \leq D + \eta DL \leq D + \frac{1}{9L}DL = \frac{10D}{9}. \end{aligned}$$

□

### 7.3 Proof of Lemma 2

*Proof.* To confirm that $d(\mathbf{z}_t, \mathbf{z}^*) \leq D$ for any $t \geq 0$, we employ an induction approach. The base case $d(\mathbf{z}_0, \mathbf{z}^*) \leq D$ is straightforward. Assume $d(\mathbf{z}_t, \mathbf{z}^*) \leq D$, we proceed to establish that $d(\mathbf{z}_{t+1}, \mathbf{z}^*) \leq D$.

By Riemannian cosine law Lemma 21, we have

$$d(\mathbf{z}_{t+1}, \mathbf{z}^*)^2 - d(\mathbf{z}_t, \mathbf{z}^*)^2 \leq 2\left\langle \mathrm{Exp}_{\mathbf{z}_{t+1}}^{-1} \mathbf{z}_t, \mathrm{Exp}_{\mathbf{z}_{t+1}}^{-1} \mathbf{z}^* \right\rangle - \sigma(K, d(\mathbf{z}_t, \mathbf{z}_{t+1}) + \min\{d(\mathbf{z}_t, \mathbf{z}^*), d(\mathbf{z}_{t+1}, \mathbf{z}^*)\}) \cdot d(\mathbf{z}_t, \mathbf{z}_{t+1})^2. \tag{11}$$

Based on the monotonicity of $\sigma(K, \cdot)$ and $\eta \leq \frac{\bar{\sigma}}{8\zeta L + \bar{\sigma} L} \leq \frac{1}{9L}$, we have

$$\begin{aligned} &- \sigma(K, d(\mathbf{z}_t, \mathbf{z}_{t+1}) + \min\{d(\mathbf{z}_t, \mathbf{z}^*), d(\mathbf{z}_{t+1}, \mathbf{z}^*)\}) \\ \leq &- \sigma(K, d(\mathbf{z}_t, \mathbf{z}^*) + d(\mathbf{z}_t, \mathbf{z}_{t+1})) \\ \leq &- \sigma\left(K, \frac{91D}{81}\right) = -\bar{\sigma}, \end{aligned} \tag{12}$$

The second inequality is based on the condition $d(\mathbf{z}_t, \mathbf{z}^*) + d(\mathbf{z}_{t+1}, \mathbf{z}_t) \leq \frac{91D}{81}$, as established in the proof of Lemma 10. Thus, we can deduce

$$d(\mathbf{z}_{t+1}, \mathbf{z}^*)^2 - d(\mathbf{z}_t, \mathbf{z}^*)^2 \leq 2\left\langle \mathrm{Exp}_{\mathbf{z}_{t+1}}^{-1}\mathbf{z}_t, \mathrm{Exp}_{\mathbf{z}_{t+1}}^{-1}\mathbf{z}^* \right\rangle - \bar{\sigma}\cdot d(\mathbf{z}_t, \mathbf{z}_{t+1})^2. \tag{13}$$

by combining the above two inequalities.

We establish an upper bound for $2\left\langle \mathrm{Exp}_{\mathbf{z}_{t+1}}^{-1}\mathbf{z}_t, \mathrm{Exp}_{\mathbf{z}_{t+1}}^{-1}\mathbf{z}^* \right\rangle$ as follows:

$$\begin{aligned}
&2\left\langle \mathrm{Exp}_{\mathbf{z}_{t+1}}^{-1}\mathbf{z}_t, \mathrm{Exp}_{\mathbf{z}_{t+1}}^{-1}\mathbf{z}^* \right\rangle = 2\eta\left\langle \Gamma_{\mathbf{z}_t}^{\mathbf{z}_{t+1}}\Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_t}F(\tilde{\mathbf{z}}_t), \mathrm{Exp}_{\mathbf{z}_{t+1}}^{-1}\mathbf{z}^* \right\rangle \\
=&2\eta\left\langle \Gamma_{\mathbf{z}_t}^{\mathbf{z}_{t+1}}\Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_t}F(\tilde{\mathbf{z}}_t) - \Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_{t+1}}F(\tilde{\mathbf{z}}_t), \mathrm{Exp}_{\mathbf{z}_{t+1}}^{-1}\mathbf{z}^* \right\rangle + 2\eta\left\langle \Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_{t+1}}F(\tilde{\mathbf{z}}_t), \mathrm{Exp}_{\mathbf{z}_{t+1}}^{-1}\mathbf{z}^* \right\rangle \\
\leq&2\eta\|\Gamma_{\mathbf{z}_t}^{\mathbf{z}_{t+1}}\Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_t}F(\tilde{\mathbf{z}}_t) - \Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_{t+1}}F(\tilde{\mathbf{z}}_t)\|\cdot d(\mathbf{z}_{t+1}, \mathbf{z}^*) + 2\eta\left\langle \Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_{t+1}}F(\tilde{\mathbf{z}}_t), \mathrm{Exp}_{\mathbf{z}_{t+1}}^{-1}\mathbf{z}^* \right\rangle
\end{aligned} \tag{14}$$

The term $\|\Gamma_{\mathbf{z}_t}^{\mathbf{z}_{t+1}}\Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_t}F(\tilde{\mathbf{z}}_t) - \Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_{t+1}}F(\tilde{\mathbf{z}}_t)\|$ corresponds to the geometric distortion due to the holonomy effect and can be addressed by Lemma 24. Specifically,

$$\begin{aligned}
&\|\Gamma_{\mathbf{z}_t}^{\mathbf{z}_{t+1}}\Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_t}F(\tilde{\mathbf{z}}_t) - \Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_{t+1}}F(\tilde{\mathbf{z}}_t)\| \\
\leq&36K_m\|F(\tilde{\mathbf{z}}_t)\|\cdot\min\{d(\mathbf{z}_{t+1}, \tilde{\mathbf{z}}_t) + d(\mathbf{z}_t, \mathbf{z}_{t+1}), d(\mathbf{z}_{t+1}, \tilde{\mathbf{z}}_t) + d(\mathbf{z}_t, \tilde{\mathbf{z}}_t)\}\cdot d(\mathbf{z}_{t+1}, \tilde{\mathbf{z}}_t) \\
\leq&36c\sqrt{K_m}\|F(\tilde{\mathbf{z}}_t)\|\cdot d(\tilde{\mathbf{z}}_t, \mathbf{z}_{t+1}) \\
\leq&72c\sqrt{K_m}L\eta^2\|F(\tilde{\mathbf{z}}_t)\|\cdot\|F(\mathbf{z}_t)\|
\end{aligned} \tag{15}$$

where the second inequality is due to

$$\begin{aligned}
&\min\{d(\mathbf{z}_{t+1}, \tilde{\mathbf{z}}_t) + d(\mathbf{z}_t, \mathbf{z}_{t+1}), d(\mathbf{z}_{t+1}, \tilde{\mathbf{z}}_t) + d(\mathbf{z}_t, \tilde{\mathbf{z}}_t)\} \\
\leq&d(\mathbf{z}_{t+1}, \mathbf{z}_t) + 2d(\mathbf{z}_t, \tilde{\mathbf{z}}_t) = \eta\|F(\tilde{\mathbf{z}}_t)\| + 2\eta\|F(\mathbf{z}_t)\| \\
\leq&3\eta G \leq \frac{3}{\sqrt{306K_m}} \leq \frac{1}{1152^{\frac{1}{4}}\sqrt{K_m}} \coloneqq \frac{c}{\sqrt{K_m}},
\end{aligned} \tag{16}$$

and the third one follows from Lemma 8. Note that the second inequality of Equation (16) is due to Lemma 10 and Assumption 4. Now it remains to bound $2\eta\left\langle \Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_{t+1}}F(\tilde{\mathbf{z}}_t), \mathrm{Exp}_{\mathbf{z}_{t+1}}^{-1}\mathbf{z}^* \right\rangle$:

$$\begin{aligned}
&2\eta\left\langle \Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_{t+1}}F(\tilde{\mathbf{z}}_t), \mathrm{Exp}_{\mathbf{z}_{t+1}}^{-1}\mathbf{z}^* \right\rangle = 2\eta\left\langle F(\tilde{\mathbf{z}}_t), \Gamma_{\mathbf{z}_{t+1}}^{\tilde{\mathbf{z}}_t}\mathrm{Exp}_{\mathbf{z}_{t+1}}^{-1}\mathbf{z}^* \right\rangle \\
=&2\eta\left\langle F(\tilde{\mathbf{z}}_t), \Gamma_{\mathbf{z}_{t+1}}^{\tilde{\mathbf{z}}_t}\mathrm{Exp}_{\mathbf{z}_{t+1}}^{-1}\mathbf{z}^* - \mathrm{Exp}_{\tilde{\mathbf{z}}_t}^{-1}\mathbf{z}^* \right\rangle + 2\eta\left\langle F(\tilde{\mathbf{z}}_t), \mathrm{Exp}_{\tilde{\mathbf{z}}_t}^{-1}\mathbf{z}^* \right\rangle \\
\leq&2\eta\bar{\zeta}\|F(\tilde{\mathbf{z}}_t)\|\cdot d(\tilde{\mathbf{z}}_t, \mathbf{z}_{t+1}) + 0 \\
\leq&4\eta\bar{\zeta}L\eta^2\|F(\tilde{\mathbf{z}}_t)\|\cdot\|F(\mathbf{z}_t)\| = 4\eta^3\bar{\zeta}L\|F(\mathbf{z}_t)\|\cdot\|F(\tilde{\mathbf{z}}_t)\|.
\end{aligned} \tag{17}$$

where the first inequality is by Lemma 22, the monotonicity of $\zeta(\kappa, \cdot)$ and

$$\begin{aligned}
&d(\tilde{\mathbf{z}}_t, \mathbf{z}_{t+1}) + \min\left\{d(\tilde{\mathbf{z}}_t, \mathbf{z}^*), d(\mathbf{z}_{t+1}, \mathbf{z}^*)\right\} \\
\leq&d(\tilde{\mathbf{z}}_t, \mathbf{z}_t) + d(\mathbf{z}_t, \mathbf{z}_{t+1}) + d(\mathbf{z}_t, \mathbf{z}^*) + d(\tilde{\mathbf{z}}_t, \mathbf{z}_t) \\
\leq&2\eta\|F(\mathbf{z}_t)\| + \eta\|F(\tilde{\mathbf{z}}_t)\| + D \\
\leq&\left(2\eta + \eta(1+\eta L)\right)\|F(\mathbf{z}_t)\| + D \\
\leq&\frac{28}{9}\eta\cdot DL + D \\
\leq&\frac{28D}{81} + D \leq \frac{7D}{5}.
\end{aligned} \tag{18}$$

And the second inequality of Equation (17) is due to Lemma 8. Combining Equations (13), (14), (15) and (17),

we have

$$\begin{aligned}
&d(\mathbf{z}_{t+1},\mathbf{z}^*)^2 - d(\mathbf{z}_t,\mathbf{z}^*)^2 \le 2\left\langle \mathrm{Exp}_{\mathbf{z}_{t+1}}^{-1}\mathbf{z}_t, \mathrm{Exp}_{\mathbf{z}_{t+1}}^{-1}\mathbf{z}^*\right\rangle - \bar{\sigma} d(\mathbf{z}_t,\mathbf{z}_{t+1})^2 \\
\le&\eta^3\left(144c\sqrt{K_m}\frac{91}{81}DL + 4\bar{\zeta}L\right)\|F(\mathbf{z}_t)\|\cdot\|F(\tilde{\mathbf{z}}_t)\| - \bar{\sigma}\eta^2\|F(\tilde{\mathbf{z}}_t)\|^2 \\
\le&\eta^2\left(\frac{(144c\sqrt{K_m}\frac{91}{81}DL + 4\bar{\zeta}L)\eta}{1-L\eta} - \bar{\sigma}\right)\|F(\tilde{\mathbf{z}}_t)\|^2 \\
\le&-\frac{\bar{\sigma}\eta^2}{2}\|F(\tilde{\mathbf{z}}_t)\|^2.
\end{aligned} \tag{19}$$

where we use Lemma 9, Lemma 10, the fact that $c = \frac{1}{1152^{\frac{1}{4}}}$ and $\eta \le \frac{\bar{\sigma}}{56\sqrt{K_m}DL+8\bar{\zeta}L+\bar{\sigma}L} \le \frac{\bar{\sigma}}{288c\sqrt{K_m}\frac{91}{81}DL+8\bar{\zeta}L+L\bar{\sigma}}$. By Equation (19), we know $d(\mathbf{z}_{t+1},\mathbf{z}^*) \le d(\mathbf{z}_t,\mathbf{z}^*) \le D$, and by induction, $d(\mathbf{z}_t,\mathbf{z}^*) \le D$ holds for any $t \ge 0$. Also, since $\eta \le \frac{\bar{\sigma}}{(8\bar{\zeta}+\bar{\sigma})L}$, we sum over Equation (19) from $t=1$ to $T$ to obtain:

$$\begin{aligned}
&\sum_{t=1}^T \|F(\mathbf{z}_t)\|^2 \le \sum_{t=1}^T \frac{1}{(1-L\eta)^2}\|F(\tilde{\mathbf{z}}_t)\|^2 \le \sum_{t=1}^T \frac{(8\bar{\zeta}+\bar{\sigma})^2}{(8\bar{\zeta})^2}\cdot\|F(\tilde{\mathbf{z}}_t)\|^2 \\
\le&\frac{2}{\bar{\sigma}\eta^2}\left(\frac{8\bar{\zeta}+\bar{\sigma}}{8\bar{\zeta}}\right)^2\cdot d(\mathbf{z}_0,\mathbf{z}^*)^2 = O\left(\frac{\bar{\zeta}^2}{\bar{\sigma}^3}\right).
\end{aligned}$$

Thus,

$$T\cdot\min_{t'\in[T]}\|F(\mathbf{z}_{t'})\|^2 \le \sum_{t=1}^T\|F(\mathbf{z}_t)\|^2 = O\left(\frac{\bar{\zeta}^2}{\bar{\sigma}^3}\right),$$

and there exists $t' \in [T]$, such that

$$\|F(\mathbf{z}_{t'})\| = O\left(\frac{\bar{\zeta}}{\sqrt{\bar{\sigma}^3 T}}\right).$$

□

### 7.4 Proof of Lemma 4

*Proof.* It is noteworthy that the chosen step-size $\eta$ satisfies the requirement of Corollary 1, thus $\mathbf{z}_t, \tilde{\mathbf{z}}_t \in \mathcal{D}$ for all $t \ge 0$. This observation is pivotal since Assumptions 2 and 4 are only applicable to $\mathcal{D}$.

We can directly show an analog of the first inequality in Equation (7) by

$$0 \le \left\langle \Gamma_{\mathbf{z}_{t+1}}^{\mathbf{z}_t}F(\mathbf{z}_{t+1}) - F(\mathbf{z}_t), \mathrm{Exp}_{\mathbf{z}_t}^{-1}\mathbf{z}_{t+1}\right\rangle = \left\langle F(\mathbf{z}_t) - \Gamma_{\mathbf{z}_{t+1}}^{\mathbf{z}_t}F(\mathbf{z}_{t+1}), \eta\Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_t}F(\tilde{\mathbf{z}}_t)\right\rangle, \tag{20}$$

where the inequality is due to Assumption 2, while the equality follows from Equation (8). Achieving an analog of the second inequality in Equation (7) is more complicated. We first show

$$\begin{aligned}
&\|\Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_t}F(\tilde{\mathbf{z}}_t) - \Gamma_{\mathbf{z}_{t+1}}^{\mathbf{z}_t}F(\mathbf{z}_{t+1})\|^2 \\
=&\|\Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_t}F(\tilde{\mathbf{z}}_t) - \Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_t}\Gamma_{\mathbf{z}_{t+1}}^{\tilde{\mathbf{z}}_t}F(\mathbf{z}_{t+1}) + \Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_t}\Gamma_{\mathbf{z}_{t+1}}^{\tilde{\mathbf{z}}_t}F(\mathbf{z}_{t+1}) - \Gamma_{\mathbf{z}_{t+1}}^{\mathbf{z}_t}F(\mathbf{z}_{t+1})\|^2 \\
\le&2\|\Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_t}F(\tilde{\mathbf{z}}_t) - \Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_t}\Gamma_{\mathbf{z}_{t+1}}^{\tilde{\mathbf{z}}_t}F(\mathbf{z}_{t+1})\|^2 + 2\|\Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_t}\Gamma_{\mathbf{z}_{t+1}}^{\tilde{\mathbf{z}}_t}F(\mathbf{z}_{t+1}) - \Gamma_{\mathbf{z}_{t+1}}^{\mathbf{z}_t}F(\mathbf{z}_{t+1})\|^2
\end{aligned} \tag{21}$$

where the inequality is due to $\|\mathbf{a}+\mathbf{b}\|^2 \le 2(\|\mathbf{a}\|^2 + \|\mathbf{b}\|^2)$.

Applying Lemma 24 with $\mathbf{x} = \mathbf{z}_t$, $\mathbf{y} = \mathbf{z}_{t+1}$, $\mathbf{z} = \tilde{\mathbf{z}}_t$ and $\Gamma_{\mathbf{x}}^{\mathbf{y}}\mathbf{u} = F(\mathbf{z}_{t+1})$, then we can easily verify

$$\begin{aligned}
\|\Gamma_{\mathbf{z}}^{\mathbf{x}}\Gamma_{\mathbf{y}}^{\mathbf{z}}\Gamma_{\mathbf{x}}^{\mathbf{y}}\mathbf{u} - \mathbf{u}\| =&\|\Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_t}\Gamma_{\mathbf{z}_{t+1}}^{\tilde{\mathbf{z}}_t}\Gamma_{\mathbf{z}_t}^{\mathbf{z}_{t+1}}\Gamma_{\mathbf{z}_{t+1}}^{\mathbf{z}_t}F(\mathbf{z}_{t+1}) - \Gamma_{\mathbf{z}_{t+1}}^{\mathbf{z}_t}F(\mathbf{z}_{t+1})\| \\
=&\|\Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_t}\Gamma_{\mathbf{z}_{t+1}}^{\tilde{\mathbf{z}}_t}F(\mathbf{z}_{t+1}) - \Gamma_{\mathbf{z}_{t+1}}^{\mathbf{z}_t}F(\mathbf{z}_{t+1})\|
\end{aligned}$$

and notice

$$3\eta G \leq \frac{3}{\sqrt{306K_m}} \leq \frac{1}{1152^{\frac{1}{4}}\sqrt{K_m}} \coloneqq \frac{c}{\sqrt{K_m}}$$

by $\eta \leq \frac{1}{\sqrt{8L^2+306K_mG^2}}$. Thus, by the triangle inequality,

$$\min\{d(\mathbf{z}_t,\mathbf{z}_{t+1})+d(\mathbf{z}_{t+1},\tilde{\mathbf{z}}_t), d(\mathbf{z}_t,\tilde{\mathbf{z}}_t)+d(\mathbf{z}_{t+1},\tilde{\mathbf{z}}_t)\} \leq 3\eta G \leq \frac{c}{\sqrt{K_m}},$$

and we have

$$\begin{aligned}
&\|\Gamma^{\mathbf{z}_t}_{\tilde{\mathbf{z}}_t}\Gamma^{\tilde{\mathbf{z}}_t}_{\mathbf{z}_{t+1}}F(\mathbf{z}_{t+1}) - \Gamma^{\mathbf{z}_t}_{\mathbf{z}_{t+1}}F(\mathbf{z}_{t+1})\| \\
\leq& 36K_m\|F(\mathbf{z}_{t+1})\| \cdot \min\{d(\mathbf{z}_t,\mathbf{z}_{t+1})+d(\mathbf{z}_{t+1},\tilde{\mathbf{z}}_t), d(\mathbf{z}_t,\tilde{\mathbf{z}}_t)+d(\mathbf{z}_{t+1},\tilde{\mathbf{z}}_t)\} \cdot d(\mathbf{z}_{t+1},\tilde{\mathbf{z}}_t) \\
\leq& 36\cdot c\sqrt{K_m}G \cdot d(\tilde{\mathbf{z}}_t,\mathbf{z}_{t+1}),
\end{aligned} \tag{22}$$

where we use Corollary 1 to show $\mathbf{z}_{t+1} \in \mathcal{D}$, and thus $\|F(\mathbf{z}_{t+1})\| \leq G$ by Assumption 4.

Combining Equations (21) and (22) yields

$$\begin{aligned}
&\|\Gamma^{\mathbf{z}_t}_{\tilde{\mathbf{z}}_t}F(\tilde{\mathbf{z}}_t) - \Gamma^{\mathbf{z}_t}_{\mathbf{z}_{t+1}}F(\mathbf{z}_{t+1})\|^2 \\
\overset{(1)}{\leq}& 2\|\Gamma^{\mathbf{z}_t}_{\tilde{\mathbf{z}}_t}F(\tilde{\mathbf{z}}_t) - \Gamma^{\mathbf{z}_t}_{\tilde{\mathbf{z}}_t}\Gamma^{\tilde{\mathbf{z}}_t}_{\mathbf{z}_{t+1}}F(\mathbf{z}_{t+1})\|^2 + 2\|\Gamma^{\mathbf{z}_t}_{\tilde{\mathbf{z}}_t}\Gamma^{\tilde{\mathbf{z}}_t}_{\mathbf{z}_{t+1}}F(\mathbf{z}_{t+1}) - \Gamma^{\mathbf{z}_t}_{\mathbf{z}_{t+1}}F(\mathbf{z}_{t+1})\|^2 \\
\overset{(2)}{\leq}& (2L^2+2592c^2K_mG^2)d(\tilde{\mathbf{z}}_t,\mathbf{z}_{t+1})^2 \\
\overset{(3)}{\leq}& (2L^2+2592c^2K_mG^2)\cdot 4\|\mathrm{Exp}^{-1}_{\mathbf{z}_t}\tilde{\mathbf{z}}_t - \mathrm{Exp}^{-1}_{\mathbf{z}_t}\mathbf{z}_{t+1}\|^2 \\
=& (8L^2+10368c^2K_mG^2)\eta^2\|F(\mathbf{z}_t) - \Gamma^{\mathbf{z}_t}_{\tilde{\mathbf{z}}_t}F(\tilde{\mathbf{z}}_t)\|^2 \\
\overset{(4)}{\leq}& (8L^2+306K_mG^2)\eta^2\|F(\mathbf{z}_t) - \Gamma^{\mathbf{z}_t}_{\tilde{\mathbf{z}}_t}F(\tilde{\mathbf{z}}_t)\|^2
\end{aligned} \tag{23}$$

where the second inequality is by Assumption 3, the third is by Lemma 17, and for the last inequality, we use the fact that $10368c^2 = \frac{10368}{\sqrt{1152}} \leq 306$.

Combining Equations (20) and (23), we have

$$\begin{aligned}
&0 \leq \left\langle F(\mathbf{z}_t) - \Gamma^{\mathbf{z}_t}_{\mathbf{z}_{t+1}}F(\mathbf{z}_{t+1}), \Gamma^{\mathbf{z}_t}_{\tilde{\mathbf{z}}_t}F(\tilde{\mathbf{z}}_t)\right\rangle \\
&\|\Gamma^{\mathbf{z}_t}_{\tilde{\mathbf{z}}_t}F(\tilde{\mathbf{z}}_t) - \Gamma^{\mathbf{z}_t}_{\mathbf{z}_{t+1}}F(\mathbf{z}_{t+1})\|^2 \leq (8L^2+306K_mG^2)\eta^2\|F(\mathbf{z}_t) - \Gamma^{\mathbf{z}_t}_{\tilde{\mathbf{z}}_t}F(\tilde{\mathbf{z}}_t)\|^2.
\end{aligned} \tag{24}$$

Since $\eta$ satisfying $(8L^2+306K_mG^2)\eta^2 \leq 1$, we can apply Lemma 1 with $\mathbf{a} = F(\mathbf{z}_t)$, $\mathbf{b} = \Gamma^{\mathbf{z}_t}_{\tilde{\mathbf{z}}_t}F(\tilde{\mathbf{z}}_t)$ and $\mathbf{c} = \Gamma^{\mathbf{z}_t}_{\mathbf{z}_{t+1}}F(\mathbf{z}_{t+1})$ to obtain

$$\|F(\mathbf{z}_{t+1})\| = \|\Gamma^{\mathbf{z}_t}_{\mathbf{z}_{t+1}}F(\mathbf{z}_{t+1})\| \leq \|F(\mathbf{z}_t)\|,$$

where the equality is due to the parallel transport preserves the vector norm. We note that $F(\mathbf{z}_t), \Gamma^{\mathbf{z}_t}_{\tilde{\mathbf{z}}_t}F(\tilde{\mathbf{z}}_t), \Gamma^{\mathbf{z}_t}_{\mathbf{z}_{t+1}}F(\mathbf{z}_{t+1})$ are tangent vectors in the same tangent space $T_{\mathbf{z}_t}\mathcal{M}$. □

### 7.5 Proof of Theorem 1

*Proof.* The theorem can be proved by directly combining Lemma 2 and Lemma 4. □

### 7.6 Proof of Theorem 2

To establish the average-iterate convergence, the following lemma proves beneficial.

**Lemma 11.** *Under Assumptions 1, 3 and 4. For the iterates of REG as in Equation* (8) *with*

$$\eta \leq \min\left\{\frac{1}{\sqrt{8L^2+306K_mG^2}}, \frac{\bar{\sigma}}{56\sqrt{K_m}DL + 8\bar{\zeta}L + \bar{\sigma}L}\right\},$$

*we have*

$$d(\mathbf{z}_{t+1},\mathbf{z})^2 - d(\mathbf{z}_t,\mathbf{z})^2 \leq 2\eta\left\langle F(\tilde{\mathbf{z}}_t), \mathrm{Exp}^{-1}_{\tilde{\mathbf{z}}_t}\mathbf{z}\right\rangle.$$

*holds for any $t \geq 0$ and $\mathbf{z} \in \mathcal{B}(\mathbf{z}^*, D)$, where $\mathcal{B}(\mathbf{z}^*, D)$ denotes the geodesic ball with center $\mathbf{z}^*$ and radius $D$.*

*Proof.* The proof is similar to that of Lemma 2. Combining Equations (13), (14), (15) and (17) but replacing $\mathbf{z}^*$ with $\mathbf{z}$, we have

$$\begin{aligned}
&d(\mathbf{z}_{t+1},\mathbf{z})^2 - d(\mathbf{z}_t,\mathbf{z})^2 \le 2\left\langle \mathrm{Exp}_{\mathbf{z}_{t+1}}^{-1}\mathbf{z}_t, \mathrm{Exp}_{\mathbf{z}_{t+1}}^{-1}\mathbf{z}\right\rangle - \bar{\sigma} d(\mathbf{z}_t,\mathbf{z}_{t+1})^2 \\
\le&\eta^3\left(\frac{144}{1152^{\frac{1}{4}}}\sqrt{K_m}\cdot d(\mathbf{z}_{t+1},\mathbf{z})L + 4\bar{\zeta}L\right)\|F(\mathbf{z}_t)\|\cdot\|F(\tilde{\mathbf{z}}_t)\| - \bar{\sigma}\eta^2\|F(\tilde{\mathbf{z}}_t)\|^2 + 2\eta\left\langle F(\tilde{\mathbf{z}}_t), \mathrm{Exp}_{\tilde{\mathbf{z}}_t}^{-1}\mathbf{z}\right\rangle \\
\le&\eta^3\left(\frac{144}{1152^{\frac{1}{4}}}\sqrt{K_m}\cdot 2DL + 4\bar{\zeta}L\right)\|F(\mathbf{z}_t)\|\cdot\|F(\tilde{\mathbf{z}}_t)\| - \bar{\sigma}\eta^2\|F(\tilde{\mathbf{z}}_t)\|^2 + 2\eta\left\langle F(\tilde{\mathbf{z}}_t), \mathrm{Exp}_{\tilde{\mathbf{z}}_t}^{-1}\mathbf{z}\right\rangle \\
\le&\frac{\eta^3}{1-L\eta}\left(\frac{144}{1152^{\frac{1}{4}}}\sqrt{K_m}\cdot 2DL + 4\bar{\zeta}L\right)\cdot\|F(\tilde{\mathbf{z}}_t)\|^2 - \bar{\sigma}\eta^2\|F(\tilde{\mathbf{z}}_t)\|^2 + 2\eta\left\langle F(\tilde{\mathbf{z}}_t), \mathrm{Exp}_{\tilde{\mathbf{z}}_t}^{-1}\mathbf{z}\right\rangle,
\end{aligned} \tag{25}$$

where the third inequality is due to Lemma 2 and $d(\mathbf{z}_{t+1},\mathbf{z}) \le d(\mathbf{z}_{t+1},\mathbf{z}^*) + d(\mathbf{z}^*,\mathbf{z}) \le 2D$, and the last inequality follows from Lemma 9. Now, we can pick up $\eta$ to ensure

$$\frac{\eta^3}{1-L\eta}\left(\frac{144}{1152^{\frac{1}{4}}}\sqrt{K_m}\cdot 2DL + 4\bar{\zeta}L\right)\cdot\|F(\tilde{\mathbf{z}}_t)\|^2 - \bar{\sigma}\eta^2\|F(\tilde{\mathbf{z}}_t)\|^2 \le 0,$$

which means:

$$\eta \le \frac{\bar{\sigma}}{\frac{288}{1152^{\frac{1}{4}}}\sqrt{K_m}DL + 4\bar{\zeta}L + \bar{\sigma}L} \approx \frac{\bar{\sigma}}{49.43\sqrt{K_m}DL + 4\bar{\zeta}L + \bar{\sigma}L}.$$

Therefore,

$$\eta \le \min\left\{\frac{1}{\sqrt{8L^2 + 306K_mG^2}}, \frac{\bar{\sigma}}{56\sqrt{K_m}DL + 8\bar{\zeta}L + \bar{\sigma}L}\right\} \le \frac{\bar{\sigma}}{\frac{288}{1152^{\frac{1}{4}}}\sqrt{K_m}DL + 4\bar{\zeta}L + \bar{\sigma}L},$$

which concludes the proof. □

Now we are able to provide the proof of Theorem 2.

*Proof.* We denote $\mathbf{z} = \binom{\mathbf{x}}{\mathbf{y}}$ and $\mathcal{Z} = \mathcal{B}(\mathbf{z}^*, D)$ for convenience. To show the last-iterate convergence,

$$\begin{aligned}
&\max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}_T,\mathbf{y}) - \min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x},\mathbf{y}_T) \le \max_{\mathbf{y}\in\mathcal{Y}}\left\langle \nabla_{\mathbf{y}} f(\mathbf{x}_T,\mathbf{y}_T), \mathrm{Exp}_{\mathbf{y}_T}^{-1}\mathbf{y}\right\rangle + \max_{\mathbf{x}\in\mathcal{X}}\left\langle \nabla_{\mathbf{x}} f(\mathbf{x}_T,\mathbf{y}_T), -\mathrm{Exp}_{\mathbf{x}_T}^{-1}\mathbf{x}\right\rangle \\
=&\max_{\mathbf{z}\in\mathcal{Z}}\left\langle -F(\mathbf{z}_T), \mathrm{Exp}_{\mathbf{z}_T}^{-1}\mathbf{z}\right\rangle \le \|F(\mathbf{z}_T)\|\cdot\max_{\mathbf{z}\in\mathcal{Z}} d(\mathbf{z}_T,\mathbf{z}) \le \|F(\mathbf{z}_T)\|\cdot\left(d(\mathbf{z}_T,\mathbf{z}^*) + \max_{\mathbf{z}\in\mathcal{Z}} d(\mathbf{z}^*,\mathbf{z})\right) \\
\le&\|F(\mathbf{z}_T)\|\cdot 2D = O\left(\frac{\bar{\zeta}}{\sqrt{\bar{\sigma}^3 T}}\right).
\end{aligned}$$

where the last equality is due to Theorem 1. For the average-iterate convergence,

$$\begin{aligned}
f(\bar{\mathbf{x}}_T,\mathbf{y}) - f(\mathbf{x},\bar{\mathbf{y}}_T) &\overset{(1)}{\le} \frac{1}{T}\left(\sum_{t=1}^T f(\tilde{\mathbf{x}}_t,\mathbf{y}) - \sum_{t=1}^T f(\mathbf{x},\tilde{\mathbf{y}}_t)\right) \\
&\overset{(2)}{\le} \frac{1}{T}\sum_{t=1}^T\left\langle -F(\tilde{\mathbf{z}}_t), \mathrm{Exp}_{\tilde{\mathbf{z}}_t}^{-1}\mathbf{z}\right\rangle \\
&\overset{(3)}{\le} \frac{1}{2\eta T}\sum_{t=1}^T\left(d(\mathbf{z}_t,\mathbf{z})^2 - d(\mathbf{z}_{t+1},\mathbf{z})^2\right) \\
&\overset{(4)}{\le} \frac{d(\mathbf{z}_0,\mathbf{z})^2}{2\eta T} \le \frac{(2D)^2}{2\eta T} = \frac{2D^2}{\eta T} = O\left(\frac{\bar{\zeta}}{\bar{\sigma}T}\right),
\end{aligned}$$

where the first inequality is due to a nested application of Jensen's inequality for gsc-convex functions:

$$f(\bar{\mathbf{x}}_t,\mathbf{y}) \le \frac{1}{t} f(\tilde{\mathbf{x}}_t,\mathbf{y}) + \frac{t-1}{t} f(\bar{\mathbf{x}}_{t-1},\mathbf{y}),$$

the second is by the gsc-convexity, and the third comes from Lemma 11.

□

### 7.7 Challenge for Establishing the Last-iterate Convergence of RCEG

We briefly touch upon our decision to omit a discussion on the last-iterate convergence of RCEG, as proposed by Zhang et al. (2023). This variant introduces a correction term to ensure metric compatibility:

$$\begin{aligned}\tilde{\mathbf{z}}_t &= \mathrm{Exp}_{\mathbf{z}_t}(-\eta F(\mathbf{z}_t))\\ \mathbf{z}_{t+1} &= \mathrm{Exp}_{\tilde{\mathbf{z}}_t}(-\eta F(\tilde{\mathbf{z}}_t) + \mathrm{Exp}^{-1}_{\tilde{\mathbf{z}}_t}(\mathbf{z}_t)).\end{aligned}$$

A keen reader may wonder about the behavior of RCEG when we aim to derive an equivalent of Equation (7). For RCEG, crafting a counterpart to the first inequality in Equation (7) appears challenging.

By Lemma 22, we have

$$\begin{aligned}0 \le& \left\langle F(\mathbf{z}_t) - \Gamma^{\mathbf{z}_t}_{\mathbf{z}_{t+1}} F(\mathbf{z}_{t+1}), -\mathrm{Exp}^{-1}_{\mathbf{z}_t}\mathbf{z}_{t+1}\right\rangle\\ =& \left\langle F(\mathbf{z}_t) - \Gamma^{\mathbf{z}_t}_{\mathbf{z}_{t+1}} F(\mathbf{z}_{t+1}), \Gamma^{\mathbf{z}_t}_{\tilde{\mathbf{z}}_t}(\mathrm{Exp}^{-1}_{\tilde{\mathbf{z}}_t}\mathbf{z}_t - \mathrm{Exp}^{-1}_{\tilde{\mathbf{z}}_t}\mathbf{z}_{t+1})\right\rangle\\ &+ \left\langle F(\mathbf{z}_t) - \Gamma^{\mathbf{z}_t}_{\mathbf{z}_{t+1}} F(\mathbf{z}_{t+1}), -\mathrm{Exp}^{-1}_{\mathbf{z}_t}\mathbf{z}_{t+1} - (\Gamma^{\mathbf{z}_t}_{\tilde{\mathbf{z}}_t}(\mathrm{Exp}^{-1}_{\tilde{\mathbf{z}}_t}\mathbf{z}_t - \mathrm{Exp}^{-1}_{\tilde{\mathbf{z}}_t}\mathbf{z}_{t+1}))\right\rangle\\ \le& \eta\left\langle F(\mathbf{z}_t) - \Gamma^{\mathbf{z}_t}_{\mathbf{z}_{t+1}} F(\mathbf{z}_{t+1}), \Gamma^{\mathbf{z}_t}_{\tilde{\mathbf{z}}_t} F(\tilde{\mathbf{z}}_t)\right\rangle + L\cdot\max\{\zeta(\kappa,\tau)-1, 1-\sigma(K,\tau)\}\cdot d(\mathbf{z}_t,\mathbf{z}_{t+1})\cdot d(\mathbf{z}_t,\tilde{\mathbf{z}}_t)\end{aligned}$$

where $\tau = d(\mathbf{z}_t,\tilde{\mathbf{z}}_t) + \min\{d(\mathbf{z}_t,\mathbf{z}_{t+1}), d(\tilde{\mathbf{z}}_t,\mathbf{z}_{t+1})\}$. We observe that the distortion term, proportional to $d(\mathbf{z}_t,\mathbf{z}_{t+1})\cdot d(\mathbf{z}_t,\tilde{\mathbf{z}}_t)$, poses challenges in establishing bounds. In contrast, for REG, the geometric distortion due to the holonomy effect can be handled with the aid of Lemma 24. Hence, in this study, we predominantly focus on REG, demonstrating that it indeed achieves $O\left(\frac{1}{\sqrt{T}}\right)$ last-iterate convergence.

## 8 Omitted Proof for Section 5

### 8.1 Auxillary Lemmas on the Iterates of RPEG

Lemma 12 confirms that $d(\tilde{\mathbf{z}}_t,\mathbf{z}_{t+1}) = O(\eta^2)$, whereas Lemmas 13 and 14 delineate the relationship between $\|F(\mathbf{z}_t)\|$, $\|F(\tilde{\mathbf{z}}_t)\|$ and $\|F(\tilde{\mathbf{z}}_{t+1})\|$. Lemma 15 demonstrates that if $\mathbf{z}_t$ is bounded, then $\mathbf{z}_{t+1},\tilde{\mathbf{z}}_t,\tilde{\mathbf{z}}_{t-1}$ are also bounded.

**Lemma 12.** *For the iterates of RPEG as in Equation* (10) *with* $\eta \le \frac{1}{G\sqrt{K_m}}$*, we have*

$$d(\tilde{\mathbf{z}}_t,\mathbf{z}_{t+1}) \le 2L\eta^2(2\|F(\tilde{\mathbf{z}}_{t-1})\| + \|F(\tilde{\mathbf{z}}_{t-2})\|).$$

*Proof.* We begin with Lemma 17,

$$\begin{aligned}& d(\tilde{\mathbf{z}}_t,\mathbf{z}_{t+1}) \le 2\|\mathrm{Exp}^{-1}_{\mathbf{z}_t}\tilde{\mathbf{z}}_t - \mathrm{Exp}^{-1}_{\mathbf{z}_t}\mathbf{z}_{t+1}\| = 2\eta\|\Gamma^{\mathbf{z}_t}_{\tilde{\mathbf{z}}_{t-1}}F(\tilde{\mathbf{z}}_{t-1}) - \Gamma^{\mathbf{z}_t}_{\tilde{\mathbf{z}}_t}F(\tilde{\mathbf{z}}_t)\|\\ =&2\eta\|\Gamma^{\mathbf{z}_t}_{\tilde{\mathbf{z}}_{t-1}}F(\tilde{\mathbf{z}}_{t-1}) - F(\mathbf{z}_t) + F(\mathbf{z}_t) - \Gamma^{\mathbf{z}_t}_{\tilde{\mathbf{z}}_t}F(\tilde{\mathbf{z}}_t)\|\\ \le&2\eta L\cdot d(\mathbf{z}_t,\tilde{\mathbf{z}}_{t-1}) + 2\eta L d(\mathbf{z}_t,\tilde{\mathbf{z}}_t)\\ \le&2\eta L(d(\mathbf{z}_t,\mathbf{z}_{t-1}) + d(\tilde{\mathbf{z}}_{t-1},\mathbf{z}_{t-1}) + d(\mathbf{z}_t,\tilde{\mathbf{z}}_t))\\ =&2\eta^2 L(\|F(\tilde{\mathbf{z}}_{t-1})\| + \|F(\tilde{\mathbf{z}}_{t-2})\| + \|F(\tilde{\mathbf{z}}_{t-1})\|)\\ =&2\eta^2 L(2\|F(\tilde{\mathbf{z}}_{t-1})\| + \|F(\tilde{\mathbf{z}}_{t-2})\|).\end{aligned}$$

□

**Lemma 13.** *(Chavdarova et al., 2021) Suppose* $\eta \le \frac{1}{8L}$*, then*

$$\frac{1}{2} \le \frac{\|F(\tilde{\mathbf{z}}_{t+1})\|}{\|F(\tilde{\mathbf{z}}_t)\|} \le \frac{3}{2}$$

*holds for RPEG in Equation* (10)*.*

In Chavdarova et al. (2021), Lemma 13 has been established for PEG. However, given that the primary argument hinges on the triangle inequality and the Lipschitz continuity of $F$, extending the proof to the manifold setting is straightforward.

**Lemma 14.** *Suppose $\eta \leq \frac{1}{8L}$, for the iterates of RPEG as in Equation (10), we have*

$$(1-2L\eta)\|F(\tilde{\mathbf{z}}_t)\| \leq \|F(\mathbf{z}_t)\| \leq (1+2L\eta)\|F(\tilde{\mathbf{z}}_t)\|.$$

*Proof.* The proof is immediate by the triangle inequality, Lemma 13 and Assumption 3. First, we have

$$\begin{aligned}\|F(\mathbf{z}_t)\| \leq& \|F(\tilde{\mathbf{z}}_t)\| + \|F(\mathbf{z}_t) - \Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_t} F(\tilde{\mathbf{z}}_t)\| \\ =& \|F(\tilde{\mathbf{z}}_t)\| + L \cdot d(\mathbf{z}_t, \tilde{\mathbf{z}}_t) = \|F(\tilde{\mathbf{z}}_t)\| + L\eta\|F(\tilde{\mathbf{z}}_{t-1})\| \\ \leq& (1+2\eta L)\|F(\tilde{\mathbf{z}}_t)\|.\end{aligned}$$

Next, we demonstrate that:

$$\begin{aligned}\|F(\mathbf{z}_t)\| \geq& \|F(\tilde{\mathbf{z}}_t)\| - \|F(\mathbf{z}_t) - \Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_t} F(\tilde{\mathbf{z}}_t)\| \\ =& \|F(\tilde{\mathbf{z}}_t)\| - L \cdot d(\mathbf{z}_t, \tilde{\mathbf{z}}_t) = \|F(\tilde{\mathbf{z}}_t)\| - L\eta\|F(\tilde{\mathbf{z}}_{t-1})\| \\ \geq& (1-2\eta L)\|F(\tilde{\mathbf{z}}_t)\|.\end{aligned}$$

□

**Lemma 15.** *For the iterates of RPEG as in Equation (10), with step-size $\eta \leq \frac{1}{32L}$, assume $d(\mathbf{z}_t, \mathbf{z}^*) \leq D$, Assumptions 1 and 3 holds. Then we have*

$$\begin{aligned}d(\mathbf{z}_{t+1}, \mathbf{z}^*) &\leq \frac{31D}{30} \\ d(\tilde{\mathbf{z}}_t, \mathbf{z}^*) &\leq \frac{16D}{15} \\ d(\tilde{\mathbf{z}}_{t-1}, \mathbf{z}^*) &\leq \frac{6D}{5}.\end{aligned}$$

*We can also obtain $\mathbf{z}_t, \mathbf{z}_{t+1}, \tilde{\mathbf{z}}_t, \tilde{\mathbf{z}}_{t-1} \in \mathcal{D}$ holds where $\mathcal{D}$ is defined in Definition 3.*

*Proof.* This lemma can be proved via a combination of the triangle inequality, Lemma 13, Lemma 14 and Assumption 3. For $d(\mathbf{z}_{t+1}, \mathbf{z}^*)$:

$$\begin{aligned}&d(\mathbf{z}_{t+1}, \mathbf{z}^*) \leq d(\mathbf{z}_t, \mathbf{z}^*) + d(\mathbf{z}_{t+1}, \mathbf{z}_t) \\ \leq& D + \eta\|F(\tilde{\mathbf{z}}_t)\| \leq D + \frac{\eta}{1-2L\eta}\|F(\mathbf{z}_t)\| \\ \leq& D + \frac{\eta}{1-2L\eta}DL \leq D + \frac{1}{30}D = \frac{31D}{30}.\end{aligned}$$

We can bound $d(\tilde{\mathbf{z}}_t, \mathbf{z}^*)$ in a similar way:

$$\begin{aligned}&d(\tilde{\mathbf{z}}_t, \mathbf{z}^*) \leq d(\mathbf{z}_t, \mathbf{z}^*) + d(\tilde{\mathbf{z}}_t, \mathbf{z}_t) \\ \leq& D + \eta\|F(\tilde{\mathbf{z}}_{t-1})\| \\ \leq& D + 2\eta\|F(\tilde{\mathbf{z}}_t)\| \\ \leq& D + \frac{2}{30}D = \frac{16D}{15}.\end{aligned}$$

The case of $d(\tilde{\mathbf{z}}_{t-1}, \mathbf{z}^*)$ is slightly more involved. First, by the triangle inequality,

$$d(\tilde{\mathbf{z}}_{t-1}, \mathbf{z}^*) \leq d(\tilde{\mathbf{z}}_{t-1}, \mathbf{z}_{t-1}) + d(\mathbf{z}_{t-1}, \mathbf{z}^*).$$

We bound both terms individually as:

$$\begin{aligned}d(\mathbf{z}_{t-1}, \mathbf{z}^*) &\leq d(\mathbf{z}_t, \mathbf{z}_{t-1}) + d(\mathbf{z}_t, \mathbf{z}^*) \leq \eta\|F(\tilde{\mathbf{z}}_{t-1})\| + D \leq \frac{16D}{15} \\ d(\tilde{\mathbf{z}}_{t-1}, \mathbf{z}_{t-1}) &= \eta\|F(\tilde{\mathbf{z}}_{t-2})\| \leq 4\eta\|F(\tilde{\mathbf{z}}_t)\| \leq \frac{4\eta}{1-2L\eta}DL \leq \frac{2D}{15}.\end{aligned}$$

Thus,

$$d(\tilde{\mathbf{z}}_{t-1}, \mathbf{z}^*) \leq d(\tilde{\mathbf{z}}_{t-1}, \mathbf{z}_{t-1}) + d(\mathbf{z}_{t-1}, \mathbf{z}^*) \leq \frac{2D}{15} + \frac{16D}{15} = \frac{6D}{5}.$$

□

### 8.2 Proof of Lemma 5

*Proof.* We again prove the lemma by induction. The case of $t = 0$ is obvious. Now, we assume $d(\mathbf{z}_t, \mathbf{z}^*) \leq D$ and we intend to show $d(\mathbf{z}_{t+1}, \mathbf{z}^*) \leq D$. We define $c \coloneqq \frac{1}{6\sqrt{2}}$ for convenience.

By Lemma 21,

$$d(\mathbf{z}_{t+1}, \mathbf{z}^*)^2 - d(\mathbf{z}_t, \mathbf{z}^*)^2 \leq 2\left\langle \mathrm{Exp}^{-1}_{\mathbf{z}_{t+1}} \mathbf{z}_t, \mathrm{Exp}^{-1}_{\mathbf{z}_{t+1}} \mathbf{z}^* \right\rangle - \sigma(K, d(\mathbf{z}_t, \mathbf{z}_{t+1}) + \min\left\{d(\mathbf{z}_t, \mathbf{z}^*), d(\mathbf{z}_{t+1}, \mathbf{z}^*)\right\}) \cdot d(\mathbf{z}_t, \mathbf{z}_{t+1})^2. \tag{26}$$

Based on the monotonicity of $\sigma(K, \cdot)$ and $\eta \leq \frac{\bar{\sigma}}{32\tilde{\zeta}L} \leq \frac{1}{32L}$, we have

$$\begin{aligned} & - \sigma(K, d(\mathbf{z}_t, \mathbf{z}_{t+1}) + \min\left\{d(\mathbf{z}_t, \mathbf{z}^*), d(\mathbf{z}_{t+1}, \mathbf{z}^*)\right\}) \\ \leq & - \sigma(K, d(\mathbf{z}_t, \mathbf{z}^*) + d(\mathbf{z}_t, \mathbf{z}_{t+1})) \\ \leq & - \sigma\left(K, \frac{31D}{30}\right) \leq -\bar{\sigma}, \end{aligned} \tag{27}$$

where $d(\mathbf{z}_t, \mathbf{z}^*) + d(\mathbf{z}_t, \mathbf{z}_{t+1}) \leq \frac{31D}{30}$ follows from the proof of Lemma 15. Thus, we can deduce

$$d(\mathbf{z}_{t+1}, \mathbf{z}^*)^2 - d(\mathbf{z}_t, \mathbf{z}^*)^2 \leq 2\left\langle \mathrm{Exp}^{-1}_{\mathbf{z}_{t+1}} \mathbf{z}_t, \mathrm{Exp}^{-1}_{\mathbf{z}_{t+1}} \mathbf{z}^* \right\rangle - \bar{\sigma} \cdot d(\mathbf{z}_t, \mathbf{z}_{t+1})^2. \tag{28}$$

We can decompose the term $2\left\langle \mathrm{Exp}^{-1}_{\mathbf{z}_{t+1}} \mathbf{z}_t, \mathrm{Exp}^{-1}_{\mathbf{z}_{t+1}} \mathbf{z}^* \right\rangle$ as:

$$\begin{aligned} & 2\left\langle \mathrm{Exp}^{-1}_{\mathbf{z}_{t+1}} \mathbf{z}_t, \mathrm{Exp}^{-1}_{\mathbf{z}_{t+1}} \mathbf{z}^* \right\rangle = 2\eta \left\langle \Gamma^{\mathbf{z}_{t+1}}_{\mathbf{z}_t} \Gamma^{\mathbf{z}_t}_{\tilde{\mathbf{z}}_t} F(\tilde{\mathbf{z}}_t), \mathrm{Exp}^{-1}_{\mathbf{z}_{t+1}} \mathbf{z}^* \right\rangle \\ = & 2\eta \left\langle \Gamma^{\mathbf{z}_{t+1}}_{\mathbf{z}_t} \Gamma^{\mathbf{z}_t}_{\tilde{\mathbf{z}}_t} F(\tilde{\mathbf{z}}_t) - \Gamma^{\mathbf{z}_{t+1}}_{\tilde{\mathbf{z}}_t} F(\tilde{\mathbf{z}}_t), \mathrm{Exp}^{-1}_{\mathbf{z}_{t+1}} \mathbf{z}^* \right\rangle + 2\eta \left\langle \Gamma^{\mathbf{z}_{t+1}}_{\tilde{\mathbf{z}}_t} F(\tilde{\mathbf{z}}_t), \mathrm{Exp}^{-1}_{\mathbf{z}_{t+1}} \mathbf{z}^* \right\rangle \end{aligned} \tag{29}$$

The first term on the RHS of Equation (29) corresponds to the holonomy effect and can be bounded by Lemma 24. To this end, we first compute

$$\begin{aligned} & \min\{d(\mathbf{z}_t, \mathbf{z}_{t+1}) + d(\mathbf{z}_{t+1}, \tilde{\mathbf{z}}_t), d(\mathbf{z}_t, \tilde{\mathbf{z}}_t) + d(\mathbf{z}_{t+1}, \tilde{\mathbf{z}}_t)\} \\ \leq & d(\mathbf{z}_t, \mathbf{z}_{t+1}) + d(\mathbf{z}_{t+1}, \tilde{\mathbf{z}}_t) \leq 2d(\mathbf{z}_t, \mathbf{z}_{t+1}) + d(\mathbf{z}_t, \tilde{\mathbf{z}}_t) = 2\eta\|F(\tilde{\mathbf{z}}_t)\| + \eta\|F(\tilde{\mathbf{z}}_{t-1})\| \leq 3\eta G \\ \leq & 3G \cdot \frac{1}{\sqrt{648 K_m G^2}} = \frac{1}{6\sqrt{2} \cdot \sqrt{K_m}} = \frac{c}{\sqrt{K_m}}. \end{aligned}$$

where $2\eta\|F(\tilde{\mathbf{z}}_t)\| + \eta\|F(\tilde{\mathbf{z}}_{t-1})\| \leq 3\eta G$ is due to Lemma 15 and Assumption 4.

Now we have

$$\begin{aligned} & 2\left\langle \mathrm{Exp}^{-1}_{\mathbf{z}_{t+1}} \mathbf{z}_t, \mathrm{Exp}^{-1}_{\mathbf{z}_{t+1}} \mathbf{z}^* \right\rangle \\ \overset{(1)}{\leq} & 36c\sqrt{K_m} \cdot \|F(\tilde{\mathbf{z}}_t)\| \cdot d(\tilde{\mathbf{z}}_t, \mathbf{z}_{t+1}) \cdot 2\eta d(\mathbf{z}^*, \mathbf{z}_{t+1}) + 2\eta \left\langle \Gamma^{\mathbf{z}_{t+1}}_{\tilde{\mathbf{z}}_t} F(\tilde{\mathbf{z}}_t), \mathrm{Exp}^{-1}_{\mathbf{z}_{t+1}} \mathbf{z}^* \right\rangle \\ \overset{(2)}{\leq} & 144cL\frac{31D}{30}\sqrt{K_m}\eta^3\|F(\tilde{\mathbf{z}}_t)\| \cdot (2\|F(\tilde{\mathbf{z}}_{t-1})\| + \|F(\tilde{\mathbf{z}}_{t-2})\|) + 2\eta \left\langle \Gamma^{\mathbf{z}_{t+1}}_{\tilde{\mathbf{z}}_t} F(\tilde{\mathbf{z}}_t), \mathrm{Exp}^{-1}_{\mathbf{z}_{t+1}} \mathbf{z}^* \right\rangle, \end{aligned} \tag{30}$$

where the first inequality follows from Lemma 24 and the second is a result of Lemma 12 and Lemma 15.

$$\begin{aligned} & \min\{d(\mathbf{z}_t, \mathbf{z}_{t+1}) + d(\mathbf{z}_{t+1}, \tilde{\mathbf{z}}_t), d(\mathbf{z}_t, \tilde{\mathbf{z}}_t) + d(\mathbf{z}_{t+1}, \tilde{\mathbf{z}}_t)\} \\ \leq & d(\mathbf{z}_t, \mathbf{z}_{t+1}) + d(\mathbf{z}_{t+1}, \tilde{\mathbf{z}}_t) \leq 2d(\mathbf{z}_t, \mathbf{z}_{t+1}) + d(\mathbf{z}_t, \tilde{\mathbf{z}}_t) = 2\eta\|F(\tilde{\mathbf{z}}_t)\| + \eta\|F(\tilde{\mathbf{z}}_{t-1})\| \leq 3\eta G \\ \leq & 3G \cdot \frac{1}{\sqrt{648 K_m G^2}} = \frac{1}{6\sqrt{2} \cdot \sqrt{K_m}} = \frac{c}{\sqrt{K_m}}. \end{aligned}$$

while the second is a result of Lemma 12. We also achieve an upper bound on $2\eta \left\langle \Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_{t+1}} F(\tilde{\mathbf{z}}_t), \operatorname{Exp}_{\mathbf{z}_{t+1}}^{-1} \mathbf{z}^* \right\rangle$ as follows:

$$\begin{aligned}
&2\eta \left\langle \Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_{t+1}} F(\tilde{\mathbf{z}}_t), \operatorname{Exp}_{\mathbf{z}_{t+1}}^{-1} \mathbf{z}^* \right\rangle = 2\eta \left\langle F(\tilde{\mathbf{z}}_t), \Gamma_{\mathbf{z}_{t+1}}^{\tilde{\mathbf{z}}_t} \operatorname{Exp}_{\mathbf{z}_{t+1}}^{-1} \mathbf{z}^* \right\rangle \\
=&2\eta \left\langle F(\tilde{\mathbf{z}}_t), \Gamma_{\mathbf{z}_{t+1}}^{\tilde{\mathbf{z}}_t} \operatorname{Exp}_{\mathbf{z}_{t+1}}^{-1} \mathbf{z}^* - \operatorname{Exp}_{\tilde{\mathbf{z}}_t}^{-1} \mathbf{z}^* \right\rangle + 2\eta \left\langle F(\tilde{\mathbf{z}}_t), \operatorname{Exp}_{\tilde{\mathbf{z}}_t}^{-1} \mathbf{z}^* \right\rangle \\
\leq&2\eta\bar{\zeta}\|F(\tilde{\mathbf{z}}_t)\| \cdot d(\tilde{\mathbf{z}}_t, \mathbf{z}_{t+1}) \\
\leq&4\bar{\zeta}L\eta^3\|F(\tilde{\mathbf{z}}_t)\| \cdot (2\|F(\tilde{\mathbf{z}}_{t-1})\| + \|F(\tilde{\mathbf{z}}_{t-2})\|),
\end{aligned} \tag{31}$$

where the inequality is due to Lemma 22 and Assumption 2, while the last equality is due to Lemma 12. The correctness of the geometric distortion $\bar{\zeta}$ can be verified in an analog way as Equation (18). More specifically,

$$\begin{aligned}
&d(\tilde{\mathbf{z}}_t, \mathbf{z}_{t+1}) + \min\left\{d(\tilde{\mathbf{z}}_t, \mathbf{z}^*), d(\mathbf{z}_{t+1}, \mathbf{z}^*)\right\} \\
\leq&d(\tilde{\mathbf{z}}_t, \mathbf{z}_{t+1}) + \frac{31D}{30} \\
\leq&d(\tilde{\mathbf{z}}_t, \mathbf{z}_t) + d(\mathbf{z}_t, \mathbf{z}_{t+1}) + \frac{31D}{30} \\
=&\eta\|F(\tilde{\mathbf{z}}_{t-1})\| + \eta\|F(\tilde{\mathbf{z}}_t)\| + \frac{31D}{30} \\
\leq&3\eta\|F(\tilde{\mathbf{z}}_t)\| + \frac{31D}{30} \\
\leq&\frac{3\eta}{1-2L\eta}\|F(\mathbf{z}_t)\| \leq \frac{3\eta}{1-2\eta L} \cdot DL + D \leq \frac{11D}{10} \leq \frac{7D}{5}.
\end{aligned}$$

Combining Equations (28), (30) and (31), we have

$$\begin{aligned}
&d(\mathbf{z}_{t+1}, \mathbf{z}^*)^2 - d(\mathbf{z}_t, \mathbf{z}^*)^2 \leq 2\left\langle \operatorname{Exp}_{\mathbf{z}_{t+1}}^{-1} \mathbf{z}_t, \operatorname{Exp}_{\mathbf{z}_{t+1}}^{-1} \mathbf{z}^* \right\rangle - \bar{\sigma} \cdot d(\mathbf{z}_t, \mathbf{z}_{t+1})^2 \\
\leq&(144cLD\sqrt{K_m} + 4\bar{\zeta}L)\eta^3 \cdot \|F(\tilde{\mathbf{z}}_t)\| \cdot (2\|F(\tilde{\mathbf{z}}_{t-1})\| + \|F(\tilde{\mathbf{z}}_{t-2})\|) - \bar{\sigma}\eta^2\|F(\tilde{\mathbf{z}}_t)\|^2 \\
\leq&8\left(144cL\frac{31D}{30}\sqrt{K_m} + 4\bar{\zeta}L\right)\eta^3\|F(\tilde{\mathbf{z}}_t)\|^2 - \bar{\sigma}\eta^2\|F(\tilde{\mathbf{z}}_t)\|^2,
\end{aligned} \tag{32}$$

where the third inequality follows from Lemma 13. Remind that $c = \frac{1}{6\sqrt{2}}$, we can guarantee the RHS of Equation (32) to be non-positive by choosing

$$\eta \leq \frac{\bar{\sigma}}{141LD\sqrt{K_m} + 32\bar{\zeta}L}.$$

Since

$$\eta \leq \min\left\{\frac{\bar{\sigma}}{141LD\sqrt{K_m} + 32\bar{\zeta}L}, \frac{1}{\sqrt{648K_mG^2}}\right\},$$

already satisfies this requirement, by induction, we know $d(\mathbf{z}_t, \mathbf{z}^*) \leq D$ holds for $t \geq 0$. □

### 8.3 Proof of Lemma 6

*Proof.* First, we note that the step-size $\eta$ already satisfies the requirement of Lemma 2, so we know $\mathbf{z}_t, \mathbf{z}_{t+1}, \tilde{\mathbf{z}}_t, \tilde{\mathbf{z}}_{t-1} \in \mathcal{D}$ holds for any $t \geq 0$ by combining Lemmas 10 and 2. This is important, because Assumptions 2 and 4 only hold on $\mathcal{D}$. By Assumption 2, we have

$$\begin{aligned}
0 \leq& \left\langle \Gamma_{\mathbf{z}_{t+1}}^{\mathbf{z}_t} F(\mathbf{z}_{t+1}) - F(\mathbf{z}_t), \operatorname{Exp}_{\mathbf{z}_t}^{-1} \mathbf{z}_{t+1} \right\rangle \\
=& -\eta \left\langle \Gamma_{\mathbf{z}_{t+1}}^{\mathbf{z}_t} F(\mathbf{z}_{t+1}) - F(\mathbf{z}_t), \Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_t} F(\tilde{\mathbf{z}}_t) \right\rangle,
\end{aligned}$$

which is an analog of the first inequality in Equation (9). To show the second inequality, by Assumption 3,

$$\|\Gamma_{\mathbf{z}_{t+1}}^{\tilde{\mathbf{z}}_t} F(\mathbf{z}_{t+1}) - F(\tilde{\mathbf{z}}_t)\|^2 \leq L^2 d(\mathbf{z}_{t+1}, \tilde{\mathbf{z}}_t)^2,$$

and our goal is

$$\|\Gamma_{\mathbf{z}_{t+1}}^{\mathbf{z}_t} F(\mathbf{z}_{t+1}) - \Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_t} F(\tilde{\mathbf{z}}_t)\|^2 \le O(1) \cdot d(\mathbf{z}_{t+1}, \tilde{\mathbf{z}}_t)^2.$$

First, we have

$$\begin{aligned}
&\|\Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_t} F(\tilde{\mathbf{z}}_t) - \Gamma_{\mathbf{z}_{t+1}}^{\mathbf{z}_t} F(\mathbf{z}_{t+1})\|^2 \\
=&\|\Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_t} F(\tilde{\mathbf{z}}_t) - \Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_t}\Gamma_{\mathbf{z}_{t+1}}^{\tilde{\mathbf{z}}_t} F(\mathbf{z}_{t+1}) + \Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_t}\Gamma_{\mathbf{z}_{t+1}}^{\tilde{\mathbf{z}}_t} F(\mathbf{z}_{t+1}) - \Gamma_{\mathbf{z}_{t+1}}^{\mathbf{z}_t} F(\mathbf{z}_{t+1})\|^2 \\
\le&2\|\Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_t} F(\tilde{\mathbf{z}}_t) - \Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_t}\Gamma_{\mathbf{z}_{t+1}}^{\tilde{\mathbf{z}}_t} F(\mathbf{z}_{t+1})\|^2 + 2\|\Gamma_{\mathbf{z}_t}^{\mathbf{z}_{t+1}}\Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_t}\Gamma_{\mathbf{z}_{t+1}}^{\tilde{\mathbf{z}}_t} F(\mathbf{z}_{t+1}) - F(\mathbf{z}_{t+1})\|^2
\end{aligned} \tag{33}$$

where the inequality follows from $\|\mathbf{a}+\mathbf{b}\|^2 \le 2(\|\mathbf{a}\|^2 + \|\mathbf{b}\|^2)$. Since

$$\begin{aligned}
&\min\{d(\mathbf{z}_t, \mathbf{z}_{t+1}) + d(\mathbf{z}_{t+1}, \tilde{\mathbf{z}}_t), d(\mathbf{z}_t, \tilde{\mathbf{z}}_t) + d(\mathbf{z}_{t+1}, \tilde{\mathbf{z}}_t)\} \\
\le&d(\mathbf{z}_t, \mathbf{z}_{t+1}) + d(\mathbf{z}_{t+1}, \tilde{\mathbf{z}}_t) \le 2d(\mathbf{z}_t, \mathbf{z}_{t+1}) + d(\mathbf{z}_t, \tilde{\mathbf{z}}_t) \le 3\eta G \\
\le&3G \cdot \frac{1}{\sqrt{648 K_m G^2}} = \frac{1}{6\sqrt{2}\cdot\sqrt{K_m}} \\
:=&\frac{c}{\sqrt{K_m}}
\end{aligned}$$

By Lemma 24,

$$\begin{aligned}
&\|\Gamma_{\mathbf{z}_t}^{\mathbf{z}_{t+1}}\Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_t}\Gamma_{\mathbf{z}_{t+1}}^{\tilde{\mathbf{z}}_t} F(\mathbf{z}_{t+1}) - F(\mathbf{z}_{t+1})\| \\
\le&36K_m\|F(\mathbf{z}_{t+1})\| \cdot \min\{d(\mathbf{z}_t, \mathbf{z}_{t+1}) + d(\mathbf{z}_{t+1}, \tilde{\mathbf{z}}_t), d(\mathbf{z}_t, \tilde{\mathbf{z}}_t) + d(\mathbf{z}_{t+1}, \tilde{\mathbf{z}}_t)\} \cdot d(\mathbf{z}_{t+1}, \tilde{\mathbf{z}}_t) \\
\le&36c\sqrt{K_m} G \cdot d(\tilde{\mathbf{z}}_t, \mathbf{z}_{t+1})
\end{aligned} \tag{34}$$

Combining Equations (33) and (34) yields

$$\begin{aligned}
&\|\Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_t} F(\tilde{\mathbf{z}}_t) - \Gamma_{\mathbf{z}_{t+1}}^{\mathbf{z}_t} F(\mathbf{z}_{t+1})\|^2 \\
\overset{(1)}{\le}&2\|\Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_t} F(\tilde{\mathbf{z}}_t) - \Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_t}\Gamma_{\mathbf{z}_{t+1}}^{\tilde{\mathbf{z}}_t} F(\mathbf{z}_{t+1})\|^2 + 2\|\Gamma_{\mathbf{z}_t}^{\mathbf{z}_{t+1}}\Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_t}\Gamma_{\mathbf{z}_{t+1}}^{\tilde{\mathbf{z}}_t} F(\mathbf{z}_{t+1}) - F(\mathbf{z}_{t+1})\|^2 \\
\overset{(2)}{\le}&(2L^2 + 2592c^2 K_m G^2) d(\tilde{\mathbf{z}}_t, \mathbf{z}_{t+1})^2 \\
\overset{(3)}{\le}&(2L^2 + 2592c^2 K_m G^2) \cdot 4\|\mathrm{Exp}_{\mathbf{z}_t}^{-1}\tilde{\mathbf{z}}_t - \mathrm{Exp}_{\mathbf{z}_t}^{-1}\mathbf{z}_{t+1}\|^2 \\
=&(8L^2 + 10368c^2 K_m G^2)\eta^2\|\Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_t} F(\tilde{\mathbf{z}}_t) - \Gamma_{\tilde{\mathbf{z}}_{t-1}}^{\mathbf{z}_t} F(\tilde{\mathbf{z}}_{t-1})\|^2,
\end{aligned} \tag{35}$$

where the first inequality is due to $\|\mathbf{a}+\mathbf{b}\|^2 \le 2(\|\mathbf{a}\|^2 + \|\mathbf{b}\|^2)$, the second is by Assumption 3 and Lemma 24, the third is due to Lemma 17. Thus, we have

$$0 \le \left\langle F(\mathbf{z}_t) - \Gamma_{\mathbf{z}_{t+1}}^{\mathbf{z}_t} F(\mathbf{z}_{t+1}), \Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_t} F(\tilde{\mathbf{z}}_t) \right\rangle \tag{36}$$

and

$$\|\Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_t} F(\tilde{\mathbf{z}}_t) - \Gamma_{\mathbf{z}_{t+1}}^{\mathbf{z}_t} F(\mathbf{z}_{t+1})\|^2 \le (8L^2 + 10368c^2 K_m G^2)\eta^2\|\Gamma_{\tilde{\mathbf{z}}_{t-1}}^{\mathbf{z}_t} F(\tilde{\mathbf{z}}_{t-1}) - \Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_t} F(\tilde{\mathbf{z}}_t)\|^2. \tag{37}$$

Adding two times of Equation (36) and three times of Equation (37) together yields

$$\begin{aligned}
&3\|\Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_t} F(\tilde{\mathbf{z}}_t) - \Gamma_{\mathbf{z}_{t+1}}^{\mathbf{z}_t} F(\mathbf{z}_{t+1})\|^2 \\
\le&2\left\langle F(\mathbf{z}_t) - \Gamma_{\mathbf{z}_{t+1}}^{\mathbf{z}_t} F(\mathbf{z}_{t+1}), \Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_t} F(\tilde{\mathbf{z}}_t) \right\rangle + 3(8L^2 + 10368c^2 K_m G^2)\eta^2\|\Gamma_{\tilde{\mathbf{z}}_{t-1}}^{\mathbf{z}_t} F(\tilde{\mathbf{z}}_{t-1}) - \Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_t} F(\tilde{\mathbf{z}}_t)\|^2 \\
=&\|F(\mathbf{z}_t)\|^2 - \|F(\mathbf{z}_t) - \Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_t} F(\tilde{\mathbf{z}}_t)\|^2 + \|\Gamma_{\mathbf{z}_{t+1}}^{\mathbf{z}_t} F(\mathbf{z}_{t+1}) - \Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_t} F(\tilde{\mathbf{z}}_t)\|^2 - \|F(\mathbf{z}_{t+1})\|^2 \\
&+ 3(8L^2 + 10368c^2 K_m G^2)\eta^2\|\Gamma_{\tilde{\mathbf{z}}_{t-1}}^{\mathbf{z}_t} F(\tilde{\mathbf{z}}_{t-1}) - \Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_t} F(\tilde{\mathbf{z}}_t)\|^2
\end{aligned}$$

where we use $2\langle \mathbf{a}, \mathbf{b}\rangle = \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 - \|\mathbf{a}-\mathbf{b}\|^2$ holds for $\mathbf{a}, \mathbf{b}$ in the same tangent space. Rearranging, we obtain

$$\begin{aligned}
&\|F(\mathbf{z}_{t+1})\|^2 + 2\|\Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_t} F(\tilde{\mathbf{z}}_t) - \Gamma_{\mathbf{z}_{t+1}}^{\mathbf{z}_t} F(\mathbf{z}_{t+1})\|^2 \\
\le&\|F(\mathbf{z}_t)\|^2 - \|F(\mathbf{z}_t) - \Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_t} F(\tilde{\mathbf{z}}_t)\|^2 + 3(8L^2 + 10368c^2 K_m G^2)\eta^2\|\Gamma_{\tilde{\mathbf{z}}_{t-1}}^{\mathbf{z}_t} F(\tilde{\mathbf{z}}_{t-1}) - \Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_t} F(\tilde{\mathbf{z}}_t)\|^2.
\end{aligned} \tag{38}$$

Applying $-\|\mathbf{a}-\mathbf{b}\|^2 \leq -\frac{1}{1+\alpha}\|\mathbf{a}\|^2 + \frac{1}{\alpha}\|\mathbf{b}\|^2$ with $\mathbf{a} = \Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_t} F(\tilde{\mathbf{z}}_t) - \Gamma_{\tilde{\mathbf{z}}_{t-1}}^{\mathbf{z}_t} F(\tilde{\mathbf{z}}_{t-1})$, $\mathbf{b} = F(\mathbf{z}_t) - \Gamma_{\tilde{\mathbf{z}}_{t-1}}^{\mathbf{z}_t} F(\tilde{\mathbf{z}}_{t-1})$ and $\alpha = \frac{1}{2}$, we have

$$-\|F(\mathbf{z}_t) - \Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_t} F(\tilde{\mathbf{z}}_t)\|^2 \leq -\frac{2}{3}\|\Gamma_{\tilde{\mathbf{z}}_{t-1}}^{\mathbf{z}_t} F(\tilde{\mathbf{z}}_{t-1}) - \Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_t} F(\tilde{\mathbf{z}}_t)\|^2 + 2\|F(\mathbf{z}_t) - \Gamma_{\tilde{\mathbf{z}}_{t-1}}^{\mathbf{z}_t} F(\tilde{\mathbf{z}}_{t-1})\|^2.$$

Plugging into Equation (38), rearranging, we have

$$\begin{aligned}\|F(\mathbf{z}_{t+1})\|^2 + 2\|\Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_t} F(\tilde{\mathbf{z}}_t) - \Gamma_{\mathbf{z}_{t+1}}^{\mathbf{z}_t} F(\mathbf{z}_{t+1})\|^2 \leq & \|F(\mathbf{z}_t)\|^2 + 2\|F(\mathbf{z}_t) - \Gamma_{\tilde{\mathbf{z}}_{t-1}}^{\mathbf{z}_t} F(\tilde{\mathbf{z}}_{t-1})\|^2 \\ & + 3\left(8L^2 + 10368c^2K_mG^2\eta^2 - \frac{2}{9}\right)\|\Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_t} F(\tilde{\mathbf{z}}_t) - \Gamma_{\tilde{\mathbf{z}}_{t-1}}^{\mathbf{z}_t} F(\tilde{\mathbf{z}}_{t-1})\|^2.\end{aligned} \tag{39}$$

We are nearing the completion of the proof, but a subtle issue arises. The left-hand side (LHS) of Equation (39) should feature $2\|F(\mathbf{z}_{t+1}) - \Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_{t+1}} F(\tilde{\mathbf{z}}_t)\|^2$ as opposed to $2\|\Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_t} F(\tilde{\mathbf{z}}_t) - \Gamma_{\mathbf{z}_{t+1}}^{\mathbf{z}_t} F(\mathbf{z}_{t+1})\|^2$ to better suit the subsequent Lyapunov analysis. Fortunately, this discrepancy can be rectified by taking a closer look at the holonomy effect and bounding it appropriately. To that end, we present the following calculations:

$$\begin{aligned}
&\|\Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_{t+1}} F(\tilde{\mathbf{z}}_t) - F(\mathbf{z}_{t+1})\|^2 = \|\Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_{t+1}} F(\tilde{\mathbf{z}}_t) - \Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_{t+1}}\Gamma_{\mathbf{z}_t}^{\tilde{\mathbf{z}}_t}\Gamma_{\mathbf{z}_{t+1}}^{\mathbf{z}_t} F(\mathbf{z}_{t+1})\|^2 \\
&\quad + \left\langle 2\Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_{t+1}} F(\tilde{\mathbf{z}}_t) - F(\mathbf{z}_{t+1}) - \Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_{t+1}}\Gamma_{\mathbf{z}_t}^{\tilde{\mathbf{z}}_t}\Gamma_{\mathbf{z}_{t+1}}^{\mathbf{z}_t} F(\mathbf{z}_{t+1}), \Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_{t+1}}\Gamma_{\mathbf{z}_t}^{\tilde{\mathbf{z}}_t}\Gamma_{\mathbf{z}_{t+1}}^{\mathbf{z}_t} F(\mathbf{z}_{t+1}) - F(\mathbf{z}_{t+1})\right\rangle \\
\leq & \|\Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_{t+1}} F(\tilde{\mathbf{z}}_t) - \Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_{t+1}}\Gamma_{\mathbf{z}_t}^{\tilde{\mathbf{z}}_t}\Gamma_{\mathbf{z}_{t+1}}^{\mathbf{z}_t} F(\mathbf{z}_{t+1})\|^2 + 2\left\langle \Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_{t+1}} F(\tilde{\mathbf{z}}_t) - F(\mathbf{z}_{t+1}), \Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_{t+1}}\Gamma_{\mathbf{z}_t}^{\tilde{\mathbf{z}}_t}\Gamma_{\mathbf{z}_{t+1}}^{\mathbf{z}_t} F(\mathbf{z}_{t+1}) - F(\mathbf{z}_{t+1})\right\rangle \\
= & \|\Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_t} F(\tilde{\mathbf{z}}_t) - \Gamma_{\mathbf{z}_{t+1}}^{\mathbf{z}_t} F(\mathbf{z}_{t+1})\|^2 + 2\left\langle \Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_{t+1}} F(\tilde{\mathbf{z}}_t) - F(\mathbf{z}_{t+1}), \Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_{t+1}}\Gamma_{\mathbf{z}_t}^{\tilde{\mathbf{z}}_t}\Gamma_{\mathbf{z}_{t+1}}^{\mathbf{z}_t} F(\mathbf{z}_{t+1}) - F(\mathbf{z}_{t+1})\right\rangle \\
\leq & \|\Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_t} F(\tilde{\mathbf{z}}_t) - \Gamma_{\mathbf{z}_{t+1}}^{\mathbf{z}_t} F(\mathbf{z}_{t+1})\|^2 + 2L \cdot d(\tilde{\mathbf{z}}_t, \mathbf{z}_{t+1}) \cdot \|F(\mathbf{z}_{t+1}) - \Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_{t+1}}\Gamma_{\mathbf{z}_t}^{\tilde{\mathbf{z}}_t}\Gamma_{\mathbf{z}_{t+1}}^{\mathbf{z}_t} F(\mathbf{z}_{t+1})\| \\
\leq & \|\Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_t} F(\tilde{\mathbf{z}}_t) - \Gamma_{\mathbf{z}_{t+1}}^{\mathbf{z}_t} F(\mathbf{z}_{t+1})\|^2 + 2L \cdot 36c\sqrt{K_m} G d(\tilde{\mathbf{z}}_t, \mathbf{z}_{t+1})^2 \\
\leq & \|\Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_t} F(\tilde{\mathbf{z}}_t) - \Gamma_{\mathbf{z}_{t+1}}^{\mathbf{z}_t} F(\mathbf{z}_{t+1})\|^2 + 2L \cdot 36c\sqrt{K_m} G \cdot 4\|\mathrm{Exp}_{\mathbf{z}_t}^{-1}\tilde{\mathbf{z}}_t - \mathrm{Exp}_{\mathbf{z}_t}^{-1}\mathbf{z}_{t+1}\|^2 \\
= & \|\Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_t} F(\tilde{\mathbf{z}}_t) - \Gamma_{\mathbf{z}_{t+1}}^{\mathbf{z}_t} F(\mathbf{z}_{t+1})\|^2 + 288cGL\sqrt{K_m}\eta^2\|\Gamma_{\tilde{\mathbf{z}}_{t-1}}^{\mathbf{z}_t} F(\tilde{\mathbf{z}}_{t-1}) - \Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_t} F(\tilde{\mathbf{z}}_t)\|^2
\end{aligned} \tag{40}$$

where the first equality and the first inequality follows from

$$\|\mathbf{a}-\mathbf{b}\|^2 = \|\mathbf{a}-\mathbf{c}\|^2 + \langle 2\mathbf{a}-\mathbf{b}-\mathbf{c}, \mathbf{c}-\mathbf{b}\rangle \leq \|\mathbf{a}-\mathbf{c}\|^2 + 2\langle \mathbf{a}-\mathbf{b}, \mathbf{c}-\mathbf{b}\rangle$$

holds for $\mathbf{a}, \mathbf{b}, \mathbf{c} \in T_{\mathbf{z}_{t+1}}\mathcal{M}$. Combining Equations (39) and (40), we get

$$\begin{aligned}
&\|F(\mathbf{z}_{t+1})\|^2 + 2\|F(\mathbf{z}_{t+1}) - \Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_{t+1}} F(\tilde{\mathbf{z}}_t)\|^2 \\
\leq & \|F(\mathbf{z}_t)\|^2 + 2\|F(\mathbf{z}_t) - \Gamma_{\tilde{\mathbf{z}}_{t-1}}^{\mathbf{z}_t} F(\tilde{\mathbf{z}}_{t-1})\|^2 \\
& + \left(\left(24L^2 + 31104c^2K_mG^2 + 576cGL\sqrt{K_m}\right)\eta^2 - \frac{2}{3}\right)\|\Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_t} F(\tilde{\mathbf{z}}_t) - \Gamma_{\tilde{\mathbf{z}}_{t-1}}^{\mathbf{z}_t} F(\tilde{\mathbf{z}}_{t-1})\|^2 \\
= & \|F(\mathbf{z}_t)\|^2 + 2\|F(\mathbf{z}_t) - \Gamma_{\tilde{\mathbf{z}}_{t-1}}^{\mathbf{z}_t} F(\tilde{\mathbf{z}}_{t-1})\|^2 \\
& + \left(\left(24L^2 + 432K_mG^2 + 48GL\sqrt{2K_m}\right)\eta^2 - \frac{2}{3}\right)\|\Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_t} F(\tilde{\mathbf{z}}_t) - \Gamma_{\tilde{\mathbf{z}}_{t-1}}^{\mathbf{z}_t} F(\tilde{\mathbf{z}}_{t-1})\|^2
\end{aligned}$$

where for the equality, we plug in $c = \frac{1}{6\sqrt{2}}$. □

### 8.4 Proof of Lemma 7

*Proof.* Similar to the proof of Lemma 5, we define $c \coloneqq \frac{1}{6\sqrt{2}}$. Since $\eta \leq \frac{1}{\sqrt{36L^2+648K_mG^2+72\sqrt{2K_m}GL}} \leq \frac{1}{\sqrt{648K_mG^2}}$, by Lemma 6, Lemma 21, the definition of $\Phi_t$, and an analog of Equation (27), we have

$$\begin{aligned}
\Phi_{t+1} - \Phi_t =& d(\mathbf{z}_{t+1}, \mathbf{z}^*)^2 + \lambda(t+1)\eta^2 \left(\|F(\mathbf{z}_{t+1})\|^2 + 2\|F(\mathbf{z}_{t+1}) - \Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_{t+1}} F(\tilde{\mathbf{z}}_t)\|^2\right) \\
&- \left(d(\mathbf{z}_t, \mathbf{z}^*)^2 + \lambda t\eta^2 \left(\|F(\mathbf{z}_t)\|^2 + 2\|F(\mathbf{z}_t) - \Gamma_{\tilde{\mathbf{z}}_{t-1}}^{\mathbf{z}_t} F(\tilde{\mathbf{z}}_{t-1})\|^2\right)\right) \\
\leq& 2\left\langle \mathrm{Exp}_{\mathbf{z}_{t+1}}^{-1} \mathbf{z}_t, \mathrm{Exp}_{\mathbf{z}_{t+1}}^{-1} \mathbf{z}^* \right\rangle - \bar{\sigma}\eta^2 \|F(\tilde{\mathbf{z}}_t)\|^2 \\
&+ \lambda\eta^2(\|F(\mathbf{z}_t)\|^2 + 2\|F(\mathbf{z}_t) - \Gamma_{\tilde{\mathbf{z}}_{t-1}}^{\mathbf{z}_t} F(\tilde{\mathbf{z}}_{t-1})\|^2) \\
&+ \lambda(t+1)\left(\left(24L^2 + 432K_mG^2 + 48GL\sqrt{2K_m}\right)\eta^2 - \frac{2}{3}\right) \|\Gamma_{\tilde{\mathbf{z}}_t}^{\mathbf{z}_t} F(\tilde{\mathbf{z}}_t) - \Gamma_{\tilde{\mathbf{z}}_{t-1}}^{\mathbf{z}_t} F(\tilde{\mathbf{z}}_{t-1})\|^2.
\end{aligned} \tag{41}$$

By $\eta \leq \frac{1}{\sqrt{36L^2+648K_mG^2+72\sqrt{2K_m}GL}}$, we have

$$\left(\left(24L^2 + 432K_mG^2 + 48GL\sqrt{2K_m}\right)\eta^2 - \frac{2}{3}\right) \leq 0,$$

thus, the last term on the RHS of Equation (41) vanishes. Note that since $\eta$ satisfies the requirement in Lemma 5, by Equation (32), we have

$$\begin{aligned}
& d(\mathbf{z}_{t+1}, \mathbf{z}^*)^2 - d(\mathbf{z}_t, \mathbf{z}^*)^2 \leq 2\left\langle \mathrm{Exp}_{\mathbf{z}_{t+1}}^{-1} \mathbf{z}_t, \mathrm{Exp}_{\mathbf{z}_{t+1}}^{-1} \mathbf{z}^* \right\rangle - \bar{\sigma} \cdot d(\mathbf{z}_t, \mathbf{z}_{t+1})^2 \\
\leq& (144cLD\sqrt{K_m} + 4\bar{\zeta}L)\eta^3 \cdot \|F(\tilde{\mathbf{z}}_t)\| \cdot (2\|F(\tilde{\mathbf{z}}_{t-1})\| + \|F(\tilde{\mathbf{z}}_{t-2})\|) - \bar{\sigma}\eta^2\|F(\tilde{\mathbf{z}}_t)\|^2 \\
\leq& 8\left(144cL\frac{31D}{30}\sqrt{K_m} + 4\bar{\zeta}L\right)\eta^3\|F(\tilde{\mathbf{z}}_t)\|^2 - \bar{\sigma}\eta^2\|F(\tilde{\mathbf{z}}_t)\|^2,
\end{aligned} \tag{42}$$

By $\eta \leq \frac{1}{32L}$ and Lemmas 13 and 14, we also have

$$\|F(\mathbf{z}_t)\| \leq (1 + 2L\eta)\|F(\tilde{\mathbf{z}}_t)\| \leq \frac{17}{16}\|F(\tilde{\mathbf{z}}_t)\|$$

and

$$\begin{aligned}
\|F(\mathbf{z}_t) - \Gamma_{\tilde{\mathbf{z}}_{t-1}}^{\mathbf{z}_t} F(\tilde{\mathbf{z}}_{t-1})\| \leq& Ld(\mathbf{z}_t, \tilde{\mathbf{z}}_{t-1}) \leq L(d(\mathbf{z}_t, \mathbf{z}_{t-1}) + d(\mathbf{z}_{t-1}, \tilde{\mathbf{z}}_{t-1}) \\
=& L\eta(\|F(\tilde{\mathbf{z}}_{t-1})\| + \|F(\tilde{\mathbf{z}}_{t-2})\|) \leq 6L\eta\|F(\tilde{\mathbf{z}}_t)\|.
\end{aligned}$$

Thus,

$$\begin{aligned}
& \lambda\eta^2\|F(\mathbf{z}_t)\|^2 + 2\lambda\eta^2\|F(\mathbf{z}_t) - \Gamma_{\tilde{\mathbf{z}}_{t-1}}^{\mathbf{z}_t} F(\tilde{\mathbf{z}}_{t-1})\|^2 \\
\leq& \frac{289}{256}\lambda\eta^2\|F(\tilde{\mathbf{z}}_t)\|^2 + 2\lambda\eta^2 \cdot 36L^2\eta^2\|F(\tilde{\mathbf{z}}_t)\|^2 \\
=& \left(\frac{289}{256}\lambda\eta^2 + 72\lambda L^2\eta^4\right)\|F(\tilde{\mathbf{z}}_t)\|^2.
\end{aligned} \tag{43}$$

Combining Equations (41), (42) and (43) and choosing $\lambda = \frac{\bar{\sigma}}{16}$, we have

$$\begin{aligned}
\Phi_{t+1} - \Phi_t \leq& 8\left(144cL\frac{31D}{30}\sqrt{K_m} + 4\bar{\zeta}L\right)\eta^3\|F(\tilde{\mathbf{z}}_t)\|^2 - \bar{\sigma}\eta^2\|F(\tilde{\mathbf{z}}_t)\|^2 + \left(\frac{289}{256}\lambda\eta^2 + 72\lambda L^2\eta^4\right)\|F(\tilde{\mathbf{z}}_t)\|^2 \\
\leq& 8\left(144cL\frac{31D}{30}\sqrt{K_m} + 4\bar{\zeta}L\right)\eta^3\|F(\tilde{\mathbf{z}}_t)\|^2 - \bar{\sigma}\eta^2\|F(\tilde{\mathbf{z}}_t)\|^2 + \left(\frac{\bar{\sigma}}{14}\eta^2 + \frac{9\bar{\sigma}}{64}L\eta^3\right)\|F(\tilde{\mathbf{z}}_t)\|^2 \\
\leq& \left(\left(141LD\sqrt{K_m} + 32\bar{\zeta}L + \frac{9\bar{\sigma}L}{64}\right)\eta^3 - \frac{13\bar{\sigma}\eta^2}{14}\right) \cdot \|F(\tilde{\mathbf{z}}_t)\|^2
\end{aligned}$$

where we recall $c = \frac{1}{6\sqrt{2}}$. Now, we find by taking

$$\eta = \frac{\bar{\sigma}}{152LD\sqrt{K_m} + 35\bar{\zeta}L},$$

$\Phi_{t+1} \leq \Phi_t$ holds. This requirement is always satisfied because

$$\eta \leq \min\left\{\frac{\bar{\sigma}}{152LD\sqrt{K_m} + 35\bar{\zeta}L}, \frac{1}{\sqrt{36L^2 + 648K_mG^2 + 72\sqrt{2K_m}GL}}\right\} \leq \frac{\bar{\sigma}}{152LD\sqrt{K_m} + 35\bar{\zeta}L}$$

always holds. □

### 8.5 Proof of Theorem 3

*Proof.* By Lemma 7, we have

$$\begin{aligned} &d(\mathbf{z}_T, \mathbf{z}^*)^2 + \frac{\bar{\sigma}}{16}T\eta^2\left(\|F(\mathbf{z}_T)\|^2 + 2\|F(\mathbf{z}_T) - \Gamma_{\tilde{\mathbf{z}}_{T-1}}^{\mathbf{z}_T}F(\tilde{\mathbf{z}}_{T-1})\|^2\right) \\ =&\Phi_T \leq \Phi_{T-1} \leq \cdots \leq \Phi_0 = d(\mathbf{z}_0, \mathbf{z}^*)^2 \leq D^2. \end{aligned}$$

Thus,

$$\|F(\mathbf{z}_T)\|^2 \leq D^2 \cdot \frac{16}{\bar{\sigma}T\eta^2} = O\left(\frac{\bar{\zeta}^2}{\bar{\sigma}^3T}\right)$$

and $\|F(\mathbf{z}_T)\| = O\left(\frac{\bar{\zeta}}{\sqrt{\bar{\sigma}^3T}}\right)$.

□

### 8.6 Proof of Theorem 4

Similar to REG, we need the following auxillary lemma to establish the average-iterate convergence of RPEG.

**Lemma 16.** *Under Assumptions 1, 3 and 4. For the iterates of RPEG as in Equation* (10) *with*

$$\eta \leq \min\left\{\frac{\bar{\sigma}}{192LD\sqrt{2K_m} + 32\bar{\zeta}L}, \frac{1}{\sqrt{648K_mG^2}}\right\},$$

*we have*

$$d(\mathbf{z}_{t+1}, \mathbf{z})^2 - d(\mathbf{z}_t, \mathbf{z})^2 \leq 2\eta\left\langle F(\tilde{\mathbf{z}}_t), \mathrm{Exp}_{\tilde{\mathbf{z}}_t}^{-1}\mathbf{z}\right\rangle.$$

*holds for any $t \geq 0$ and $\mathbf{z} \in \mathcal{B}(\mathbf{z}^*, D)$ where $\mathcal{B}(\mathbf{z}^*, D)$ denotes the geodesic ball with center $\mathbf{z}^*$ and radius $D$.*

*Proof.* The proof is similar to that of Lemma 5. Combining Equations (28), (30) and (31), and replacing $\mathbf{z}^*$ with $\mathbf{z}$, we have

$$\begin{aligned} &d(\mathbf{z}_{t+1}, \mathbf{z})^2 - d(\mathbf{z}_t, \mathbf{z})^2 \leq 2\left\langle \mathrm{Exp}_{\mathbf{z}_{t+1}}^{-1}\mathbf{z}_t, \mathrm{Exp}_{\mathbf{z}_{t+1}}^{-1}\mathbf{z}\right\rangle - \bar{\sigma}d(\mathbf{z}_t, \mathbf{z}_{t+1})^2 \\ \leq&8\left(\frac{144}{6\sqrt{2}}L \cdot d(\mathbf{z}_{t+1}, \mathbf{z})\sqrt{K_m} + 4\bar{\zeta}L\right)\eta^3\|F(\tilde{\mathbf{z}}_t)\|^2 - \bar{\sigma}\eta^2\|F(\tilde{\mathbf{z}}_t)\|^2 + 2\eta\left\langle F(\tilde{\mathbf{z}}_t), \mathrm{Exp}_{\tilde{\mathbf{z}}_t}^{-1}\mathbf{z}\right\rangle, \\ =&\left(96\sqrt{2}L \cdot d(\mathbf{z}_{t+1}, \mathbf{z})\sqrt{K_m} + 4\bar{\zeta}L\right)\eta^3\|F(\tilde{\mathbf{z}}_t)\|^2 - \bar{\sigma}\eta^2\|F(\tilde{\mathbf{z}}_t)\|^2 + 2\eta\left\langle F(\tilde{\mathbf{z}}_t), \mathrm{Exp}_{\tilde{\mathbf{z}}_t}^{-1}\mathbf{z}\right\rangle, \\ \leq&\left(192\sqrt{2}L \cdot D\sqrt{K_m} + 4\bar{\zeta}L\right)\eta^3\|F(\tilde{\mathbf{z}}_t)\|^2 - \bar{\sigma}\eta^2\|F(\tilde{\mathbf{z}}_t)\|^2 + 2\eta\left\langle F(\tilde{\mathbf{z}}_t), \mathrm{Exp}_{\tilde{\mathbf{z}}_t}^{-1}\mathbf{z}\right\rangle, \end{aligned} \tag{44}$$

where the last inequality is due to Lemma 5 and $d(\mathbf{z}_{t+1}, \mathbf{z}) \leq d(\mathbf{z}_{t+1}, \mathbf{z}^*) + d(\mathbf{z}^*, \mathbf{z}) \leq 2D$. It is straightforward to see, for any

$$\eta \leq \frac{\bar{\sigma}}{192LD\sqrt{2K_m} + 32\bar{\zeta}L},$$

we have

$$d(\mathbf{z}_{t+1}, \mathbf{z})^2 - d(\mathbf{z}_t, \mathbf{z})^2 \leq 2\eta\left\langle F(\tilde{\mathbf{z}}_t), \mathrm{Exp}_{\tilde{\mathbf{z}}_t}^{-1}\mathbf{z}\right\rangle,$$

which concludes the proof. □

Now, we can start to prove Theorem 4.

*Proof.* The proof closely parallels that of Theorem 2, but we provide details for the sake of completeness. Denote $\mathbf{z} = \binom{\mathbf{x}}{\mathbf{y}}$ and $\mathcal{Z} = \mathcal{B}(\mathbf{z}^*, D)$. We start with the last-iterate convergence,

$$\begin{aligned}
&\max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}_T, \mathbf{y}) - \min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}, \mathbf{y}_T) \le \max_{\mathbf{y}\in\mathcal{Y}} \left\langle \nabla_{\mathbf{y}} f(\mathbf{x}_T, \mathbf{y}_T), \mathrm{Exp}_{\mathbf{y}_T}^{-1}\mathbf{y} \right\rangle + \max_{\mathbf{x}\in\mathcal{X}} \left\langle \nabla_{\mathbf{x}} f(\mathbf{x}_T, \mathbf{y}_T), -\mathrm{Exp}_{\mathbf{x}_T}^{-1}\mathbf{x} \right\rangle \\
=&\max_{\mathbf{z}\in\mathcal{Z}} \left\langle -F(\mathbf{z}_T), \mathrm{Exp}_{\mathbf{z}_T}^{-1}\mathbf{z} \right\rangle \le \|F(\mathbf{z}_T)\| \cdot \max_{\mathbf{z}\in\mathcal{Z}} d(\mathbf{z}_T, \mathbf{z}) \le \|F(\mathbf{z}_T)\| \cdot \left( d(\mathbf{z}_T, \mathbf{z}^*) + \max_{\mathbf{z}\in\mathcal{Z}} d(\mathbf{z}^*, \mathbf{z}) \right) \\
\le&\|F(\mathbf{z}_T)\| \cdot 2D = O\left( \frac{\bar{\zeta}}{\sqrt{\bar{\sigma}^3 T}} \right).
\end{aligned}$$

where the last equality is due to Theorem 3. For the average-iterate convergence,

$$\begin{aligned}
f(\bar{\mathbf{x}}_T, \mathbf{y}) - f(\mathbf{x}, \bar{\mathbf{y}}_T) &\overset{(1)}{\le} \frac{1}{T} \left( \sum_{t=1}^{T} f(\tilde{\mathbf{x}}_t, \mathbf{y}) - \sum_{t=1}^{T} f(\mathbf{x}, \tilde{\mathbf{y}}_t) \right) \\
&\overset{(2)}{\le} \frac{1}{T} \sum_{t=1}^{T} \left\langle -F(\tilde{\mathbf{z}}_t), \mathrm{Exp}_{\tilde{\mathbf{z}}_t}^{-1}\mathbf{z} \right\rangle \\
&\overset{(3)}{\le} \frac{1}{2\eta T} \sum_{t=1}^{T} \left( d(\mathbf{z}_t, \mathbf{z})^2 - d(\mathbf{z}_{t+1}, \mathbf{z})^2 \right) \\
&\overset{(4)}{\le} \frac{d(\mathbf{z}_0, \mathbf{z})^2}{2\eta T} \le \frac{(2D)^2}{2\eta T} = \frac{2D^2}{\eta T} = O\left( \frac{\bar{\zeta}}{\bar{\sigma} T} \right),
\end{aligned}$$

where the first inequality comes from a nested application of Jensen's inequality for gsc-convex functions:

$$f(\bar{\mathbf{x}}_t, \mathbf{y}) \le \frac{1}{t} f(\tilde{\mathbf{x}}_t, \mathbf{y}) + \frac{t-1}{t} f(\bar{\mathbf{x}}_{t-1}, \mathbf{y}),$$

the second is by the gsc-convexity, and the third comes from Lemma 16. We also note that

$$\eta \le \min \left\{ \frac{\bar{\sigma}}{192LD\sqrt{2K_m} + 35\bar{\zeta}L}, \frac{1}{\sqrt{36L^2 + 648K_m G^2 + 72\sqrt{2K_m}GL}} \right\}$$

satisfies the requirement for $\eta$ as specified in Lemma 16. □

## 9 Technical Lemmas

### 9.1 Best-iterate Convergence of RCEG

For completeness, we provide the $O\left(\frac{1}{\sqrt{T}}\right)$ rate for the best-iterate convergence of RCEG as follows. The proof is inspired by Proposition 5 of Martínez-Rubio et al. (2023).

**Theorem 5.** *Consider a Riemannian manifold $\mathcal{M}$ with sectional curvature in $[\kappa, K]$, $D = d(\mathbf{z}_0, \mathbf{z}^*)$. If $K > 0$, we require that $D < \frac{2\pi}{2\sqrt{K}}$. Let $\bar{\zeta} = \zeta(\kappa, \frac{3D}{2})$ and $\bar{\sigma} = \sigma(K, \frac{3D}{2})$ be geometric constants defined in Lemma 22 and Lemma 21. With $\eta \le \sqrt{\frac{\bar{\sigma}}{4\bar{\zeta}L^2}}$, RCEG defined by*

$$\begin{aligned}
\tilde{\mathbf{z}}_t &= \mathrm{Exp}_{\mathbf{z}_t}(-\eta F(\mathbf{z}_t)) \\
\mathbf{z}_{t+1} &= \mathrm{Exp}_{\tilde{\mathbf{z}}_t}\left(-\eta F(\tilde{\mathbf{z}}_t) + \mathrm{Exp}_{\tilde{\mathbf{z}}_t}^{-1}\mathbf{z}_t\right).
\end{aligned}$$

*achieves $O\left(\frac{1}{\sqrt{T}}\right)$ best-iterate convergence for Riemannian variational inequality problems.*

*Proof.* We use mathematical induction to establish $d(\mathbf{z}_t, \mathbf{z}^*) \le D$ holds for any $t \ge 0$. The base case $t = 0$ is straightforward. Assuming that $d(\mathbf{z}_t, \mathbf{z}^*) \le D$ holds, we proceed to show that

$$\begin{aligned} d(\tilde{\mathbf{z}}_t, \mathbf{z}^*) &\leq d(\tilde{\mathbf{z}}_t, \mathbf{z}_t) + d(\mathbf{z}_t, \mathbf{z}^*) \\ &= \eta \|F(\mathbf{z}_t)\| + d(\mathbf{z}_t, \mathbf{z}^*) \\ &= \eta \|F(\mathbf{z}_t) - \Gamma_{\mathbf{z}^*}^{\mathbf{z}_t} F(\mathbf{z}^*)\| + d(\mathbf{z}_t, \mathbf{z}^*) \\ &\leq \eta L d(\mathbf{z}_t, \mathbf{z}^*) + d(\mathbf{z}_t, \mathbf{z}^*) \\ &= (1 + \eta L) d(\mathbf{z}_t, \mathbf{z}^*) \\ &\leq \frac{3}{2} d(\mathbf{z}_t, \mathbf{z}^*), \end{aligned} \tag{45}$$

where the second inequality is due to the L-Lipschitzness of $F$ and the third inequality follows from $\eta \leq \sqrt{\frac{\bar{\sigma}}{4\bar{\zeta}L^2}} \leq \frac{1}{2L}$.

Now by Equation (45), Lemma 23 and Lemma 21,

$$\begin{aligned} & 2\left\langle \mathrm{Exp}_{\tilde{\mathbf{z}}_t}^{-1} \mathbf{z}^*, \mathrm{Exp}_{\tilde{\mathbf{z}}_t}^{-1} \mathbf{z}_{t+1} - \mathrm{Exp}_{\tilde{\mathbf{z}}_t}^{-1} \mathbf{z}_t \right\rangle \\ \leq & \bar{\zeta} d(\tilde{\mathbf{z}}_t, \mathbf{z}_{t+1})^2 - \bar{\sigma} d(\tilde{\mathbf{z}}_t, \mathbf{z}_t)^2 + d(\mathbf{z}^*, \mathbf{z}_t)^2 - d(\mathbf{z}^*, \mathbf{z}_{t+1})^2 \\ = & \bar{\zeta}\eta^2 \|F(\tilde{\mathbf{z}}_t) - \Gamma_{\mathbf{z}_t}^{\tilde{\mathbf{z}}_t} F(\mathbf{z}_t)\|^2 - \bar{\sigma} d(\tilde{\mathbf{z}}_t, \mathbf{z}_t)^2 + d(\mathbf{z}^*, \mathbf{z}_t)^2 - d(\mathbf{z}^*, \mathbf{z}_{t+1})^2 \\ \leq & \left(\bar{\zeta}\eta^2 L^2 - \bar{\sigma}\right) d(\tilde{\mathbf{z}}_t, \mathbf{z}_t)^2 + d(\mathbf{z}^*, \mathbf{z}_t)^2 - d(\mathbf{z}^*, \mathbf{z}_{t+1})^2. \end{aligned} \tag{46}$$

On the other hand, we have

$$\begin{aligned} & \left\langle \mathrm{Exp}_{\tilde{\mathbf{z}}_t}^{-1} \mathbf{z}^*, \mathrm{Exp}_{\tilde{\mathbf{z}}_t}^{-1} \mathbf{z}_{t+1} - \mathrm{Exp}_{\tilde{\mathbf{z}}_t}^{-1} \mathbf{z}_t \right\rangle \\ = & \left\langle \mathrm{Exp}_{\tilde{\mathbf{z}}_t}^{-1} \mathbf{z}^*, -\eta F(\tilde{\mathbf{z}}_t) \right\rangle \\ = & \eta \left\langle \Gamma_{\mathbf{z}^*}^{\tilde{\mathbf{z}}_t} F(\mathbf{z}^*) - F(\tilde{\mathbf{z}}_t), \mathrm{Exp}_{\tilde{\mathbf{z}}_t}^{-1} \mathbf{z}^* \right\rangle \geq 0. \end{aligned} \tag{47}$$

Combining Equation (46) and Equation (47), we have

$$d(\mathbf{z}_t, \mathbf{z}^*)^2 \geq d(\mathbf{z}_{t+1}, \mathbf{z}^*)^2 + (\bar{\sigma} - \bar{\zeta}\eta^2 L^2) d(\tilde{\mathbf{z}}_t, \mathbf{z}_t)^2. \tag{48}$$

Given that $\bar{\zeta}\eta^2 L^2 \leq \bar{\sigma}$, it follows that $d(\mathbf{z}_{t+1}, \mathbf{z}^*) \leq d(\mathbf{z}_t, \mathbf{z}^*) \leq D$, which completes the induction step. Summing over Equation (48), we obtain

$$\begin{aligned} d(\mathbf{z}_0, \mathbf{z}^*)^2 \geq & (\bar{\sigma} - \bar{\zeta}\eta^2 L^2) \sum_{t=0}^{T-1} d(\tilde{\mathbf{z}}_t, \mathbf{z}_t)^2 \\ = & \eta^2 (\bar{\sigma} - \bar{\zeta}\eta^2 L^2) \sum_{t=0}^{T-1} \|F(\mathbf{z}_t)\|^2 \\ \geq & \eta^2 (\bar{\sigma} - \bar{\zeta}\eta^2 L^2) T \cdot \min_{t' \in [T]} \|F(\mathbf{z}'_t)\|^2, \end{aligned}$$

where the final inequality demonstrates that the best-iterate convergence rate of RCEG is $O\left(\frac{\sqrt{\bar{\zeta}}}{\bar{\sigma}\sqrt{T}}\right)$. □

### 9.2 Miscellaneous Technical Lemmas

**Lemma 17.** *For a Riemannian manifold $\mathcal{M}$ with sectional curvature in $[\kappa, K]$ and a geodesic triangle on $\mathcal{M}$ with vertices $\mathbf{x}, \mathbf{y}, \mathbf{z}$. If $K > 0$, we require the maximum side length is smaller than $\frac{\pi}{\sqrt{K}}$. If $\kappa < 0$, we assume*

$$\max\{d(\mathbf{x}, \mathbf{z}), d(\mathbf{y}, \mathbf{z})\} \leq \frac{1}{\sqrt{-\kappa}},$$

*then we have*

$$d(\mathbf{x}, \mathbf{y}) \leq 2 \cdot \|\mathrm{Exp}_{\mathbf{z}}^{-1} \mathbf{x} - \mathrm{Exp}_{\mathbf{z}}^{-1} \mathbf{y}\|.$$

*Proof.* By Proposition B.2 in Ahn and Sra (2020) and Rauch Comparison Theorem, we have

$$d(\mathbf{x},\mathbf{y}) \leq \begin{cases} \frac{\sinh(\sqrt{-\kappa}\max\{d(\mathbf{x},\mathbf{z}),d(\mathbf{y},\mathbf{z})\})}{\sqrt{-\kappa}\max\{d(\mathbf{x},\mathbf{z}),d(\mathbf{y},\mathbf{z})\}} \cdot \|\mathrm{Exp}_{\mathbf{z}}^{-1}\mathbf{x} - \mathrm{Exp}_{\mathbf{z}}^{-1}\mathbf{y}\| & \kappa < 0 \\ \|\mathrm{Exp}_{\mathbf{z}}^{-1}\mathbf{x} - \mathrm{Exp}_{\mathbf{z}}^{-1}\mathbf{y}\| & \kappa \geq 0. \end{cases}$$

Since $\frac{\sinh x}{x}$ is monotonically increasing with respect to $x$ and

$$\max\{d(\mathbf{x},\mathbf{z}), d(\mathbf{y},\mathbf{z})\} \leq \frac{1}{\sqrt{-\kappa}},$$

we have

$$\frac{\sinh(\sqrt{-\kappa}\max\{d(\mathbf{x},\mathbf{z}),d(\mathbf{y},\mathbf{z})\})}{\sqrt{-\kappa}\max\{d(\mathbf{x},\mathbf{z}),d(\mathbf{y},\mathbf{z})\}} \leq \frac{\sinh 1}{1} \leq 2,$$

which completes the proof. □

**Lemma 18.** *(Lee, 2018, Theorem 7.11), (Wang et al., 2023, Lemma 11) For a Riemannian manifold $\mathcal{M}$ with sectional curvature in $[\kappa, K]$, denote $\Lambda(s,t) : [0,1]\times[0,1] \to \mathcal{M}$ to be a rectangle map and $\boldsymbol{\ell}$ to be the geodesic loop connecting $\Lambda(0,0)$, $\Lambda(0,1)$, $\Lambda(1,1)$ and $\Lambda(1,0)$. We also define $K_m = \max(|\kappa|,|K|)$, vector fields $S(s,t) = \Lambda_\star\frac{\partial}{\partial s}(s,t)$ and $T(s,t) = \Lambda_\star\frac{\partial}{\partial t}(s,t)$. Then for any $\mathbf{u} \in T_{\Lambda(0,0)}\mathcal{M}$, we have*

$$\|\Gamma_{\boldsymbol{\ell}}\mathbf{u} - \mathbf{u}\| \leq 12K_m\|\mathbf{u}\| \cdot \int_0^1\int_0^1 \|T\| \cdot \|S\| dsdt,$$

*where $\Gamma_{\boldsymbol{\ell}}$ is the parallel transport along the geodesic loop $\boldsymbol{\ell}$.*

**Definition 4.** *We define*

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

*and*

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

**Lemma 19.** *(Lee, 2018) Let the sectional curvature of a Riemannian manifold $\mathcal{M}$ be in $[\kappa, K]$, $\gamma(t) : [0,s] \to \mathcal{M}$ be a geodesic with unit velocity , and $J$ be a Jacobi field along $\gamma(t)$. When $K > 0$, we assume the length of $\gamma(t)$ is smaller than $\frac{\pi}{\sqrt{K}}$. Then*

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

*where $\boldsymbol{s}(\kappa,t)$ and $\boldsymbol{S}(K,t)$ are defined in Definition 4.*

**Lemma 20.** *(Wang et al., 2023) We denote $K_m = \max(|\kappa|,|K|)$. For any $0 \leq t \leq \frac{1}{\sqrt{K_m}}$, we have $\frac{\boldsymbol{s}(\kappa,t)}{\boldsymbol{S}(K,t)} \leq 3$, where $\boldsymbol{s}(\kappa,t)$ and $\boldsymbol{S}(K,t)$ are defined in Definition 4.*

**Lemma 21.** *(Alimisis et al., 2020) Let $\mathcal{M}$ be a Riemannian manifold with sectional curvature upper bounded by $K$. Consider a geodesic triangle with side lengths $a, b, c$ such that*

$$b + \min\{a,c\} < \begin{cases} \frac{\pi}{\sqrt{K}} & K > 0 \\ \infty & K \leq 0. \end{cases}$$

*Then we have*

$$a^2 \geq \sigma(K, b+\min\{a,c\})b^2 + c^2 - 2bc\cos A$$

*where*

$$\sigma(K,\tau) \coloneqq \begin{cases} \sqrt{K}\tau\cot(\sqrt{K}\tau) & K > 0 \\ 1 & K \leq 0. \end{cases}$$

**Remark 5.** *Lemma 21 is indeed a variant of Corollary 2.1 of Alimisis et al. (2020). The original version therein states: given a geodesic triangle with vertices* $\mathbf{x}$, $\mathbf{y}$, *and* $\mathbf{z}$, *and corresponding edge lengths* $a$, $b$, *and* $c$. *There exists a point* $\mathbf{q} \in \overline{\mathbf{xz}}$ *such that*

$$a^2 \geq \sigma(K, d(\mathbf{y}, \mathbf{q}))b^2 + c^2 - 2bc\cos A.$$

*By the triangle inequality*

$$d(\mathbf{y}, \mathbf{q}) \leq \min\{d(\mathbf{x}, \mathbf{y}) + d(\mathbf{x}, \mathbf{z}), d(\mathbf{x}, \mathbf{z}) + d(\mathbf{y}, \mathbf{z})\} = b + \min\{a, c\}.$$

*Thus, by the monotonicity of* $\sigma(K, \cdot)$, *we have*

$$\begin{aligned} a^2 \geq& \sigma(K, d(\mathbf{y}, \mathbf{q}))b^2 + c^2 - 2bc\cos A \\ \geq& \sigma(K, b + \min\{a, c\})b^2 + c^2 - 2bc\cos A. \end{aligned}$$

**Lemma 22.** *(Alimisis et al., 2021) Assume* $\mathcal{M}$ *is a Riemannian manifold with sectional curvature in* $[\kappa, K]$. *For a geodesic triangle on* $\mathcal{M}$ *with vertices* $\mathbf{x}, \mathbf{y}, \mathbf{z}$ *such that*

$$\tau \coloneqq d(\mathbf{x}, \mathbf{z}) + \min\{d(\mathbf{x}, \mathbf{y}), d(\mathbf{y}, \mathbf{z})\} < \begin{cases} \frac{\pi}{\sqrt{K}} & K > 0 \\ \infty & K \leq 0. \end{cases}$$

*Then*

*1)*

$$\|\mathrm{Exp}_{\mathbf{x}}^{-1}\mathbf{y} - \Gamma_{\mathbf{z}}^{\mathbf{x}}\mathrm{Exp}_{\mathbf{z}}^{-1}\mathbf{y}\| \leq \zeta(\kappa, \tau) \cdot d(\mathbf{x}, \mathbf{z})$$

*2)*

$$\|\mathrm{Exp}_{\mathbf{x}}^{-1}\mathbf{y} - \Gamma_{\mathbf{z}}^{\mathbf{x}}\mathrm{Exp}_{\mathbf{z}}^{-1}\mathbf{y} - \mathrm{Exp}_{\mathbf{x}}^{-1}\mathbf{z}\| \leq \max\{\zeta(\kappa, \tau) - 1, 1 - \sigma(K, \tau)\} \cdot d(\mathbf{x}, \mathbf{z})$$

*where*

$$\zeta(\kappa, \tau) \coloneqq \begin{cases} \sqrt{-\kappa}\tau\coth(\sqrt{-\kappa}\tau) & \kappa < 0 \\ 1 & \kappa \geq 0, \end{cases}$$

*and* $\sigma(K, \tau)$ *is defined in Lemma 21.*

**Lemma 23.** *(Zhang and Sra, 2016, Lemma 5). Let* $\mathcal{M}$ *be a Riemannian manifold with sectional curvature lower bounded by* $\kappa$. *Consider a geodesic triangle fully lies within* $\mathcal{M}$ *with side lengths* $a, b, c$, *we have*

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

*where* $\zeta(\kappa, c)$ *is defined in Lemma 22.*

### 9.3 Bounding the Holonomy Effect on a Geodesic Triangle

**Lemma 24.** *For a Riemannian manifold* $\mathcal{M}$ *with sectional curvature in* $[\kappa, K]$ *and a geodesic triangle* $\triangle\mathbf{xyz}$ *on* $\mathcal{M}$, *we denote* $K_m = \max\{|\kappa|, |K|\}$. *Then as long as*

$$\min\{d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z}), d(\mathbf{x}, \mathbf{z}) + d(\mathbf{y}, \mathbf{z})\} \leq \frac{1}{\sqrt{K_m}},$$

*for any* $\mathbf{u} \in T_{\mathbf{x}}\mathcal{M}$, *we have*

$$\|\Gamma_{\mathbf{z}}^{\mathbf{x}}\Gamma_{\mathbf{y}}^{\mathbf{z}}\Gamma_{\mathbf{x}}^{\mathbf{y}}\mathbf{u} - \mathbf{u}\| \leq 36K_m\|\mathbf{u}\| \cdot \min\{d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z}), d(\mathbf{x}, \mathbf{z}) + d(\mathbf{y}, \mathbf{z})\} \cdot d(\mathbf{y}, \mathbf{z}).$$

*Proof.* We define $\Lambda(s, t) : [0, 1] \times [0, 1] \to \mathcal{M}$ to be a rectangle map such that

$$\Lambda(s, t) \coloneqq \mathrm{Exp}_{\gamma_1(s)}(t\mathrm{Exp}_{\gamma_1(s)}^{-1}\gamma_2(s))$$

where $\gamma_1(s)$ and $\gamma_2(s)$ are geodesics which satisfy $\gamma_1(0) = \Lambda(0,0) \coloneqq \mathbf{x}$, $\gamma_1(1) = \Lambda(1,0) \coloneqq \mathbf{w}$, $\gamma_2(0) = \Lambda(0,1) \coloneqq \mathbf{y}$ and $\gamma_2(1) = \Lambda(1,1) \coloneqq \mathbf{z}$. Also, we denote

$$S(s,t) = \Lambda_\star \frac{\partial}{\partial s}(s,t)$$
$$T(s,t) = \Lambda_\star \frac{\partial}{\partial t}(s,t).$$

We use $\boldsymbol{\ell}$ to denote the geodesic loop starts at $\mathbf{x}$ and consists of geodesic segments $\overline{\mathbf{xy}}$, $\overline{\mathbf{yz}}$, $\overline{\mathbf{zw}}$ and $\overline{\mathbf{wx}}$, then by Lemma 18, we have

$$\|\Gamma_{\boldsymbol{\ell}}\mathbf{u} - \mathbf{u}\| \leq 12K_m\|\mathbf{u}\| \cdot \int_0^1 \int_0^1 \|T\| \cdot \|S\| ds dt \tag{49}$$

By the triangle inequality, $T(s,t)$ simultaneously satisfies

$$\begin{aligned}\|T(s,t)\| =& d(\gamma_1(s), \gamma_2(s)) \\ \leq& d(\gamma_1(s), \mathbf{x}) + d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \gamma_2(s)) \\ \leq& d(\mathbf{w}, \mathbf{x}) + d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z})\end{aligned}$$

and

$$\begin{aligned}\|T(s,t)\| =& d(\gamma_1(s), \gamma_2(s)) \\ \leq& d(\gamma_1(s), \mathbf{w}) + d(\mathbf{w}, \mathbf{z}) + d(\mathbf{z}, \gamma_2(s)) \\ \leq& d(\mathbf{x}, \mathbf{w}) + d(\mathbf{w}, \mathbf{z}) + d(\mathbf{z}, \mathbf{y}),\end{aligned}$$

we have

$$\|T(s,t)\| \leq \min\{d(\mathbf{w}, \mathbf{x}) + d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z}), d(\mathbf{x}, \mathbf{w}) + d(\mathbf{w}, \mathbf{z}) + d(\mathbf{z}, \mathbf{y})\}$$

Note that $S$ is a Jacobi field with $\|S(s,0)\| = d(\mathbf{x}, \mathbf{w})$ and $\|S(s,1)\| = d(\mathbf{y}, \mathbf{z})$. Now as we set $\mathbf{w} = \mathbf{x}$,

$$\begin{aligned}&\|T(s,t)\| \leq \min\{d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z}), d(\mathbf{x}, \mathbf{z}) + d(\mathbf{y}, \mathbf{z})\} \\ &\|S(s,0)\| = 0 \\ &\|S(s,1)\| = d(\mathbf{y}, \mathbf{z}).\end{aligned} \tag{50}$$

We define a unit speed geodesic $\gamma(t)$ such that $\gamma(0) = \gamma_1(s)$ and $\gamma(b) = \gamma_2(s)$, where $b \coloneqq d(\gamma_1(s), \gamma_2(s))$. Then we find $J(\gamma(t)) \coloneqq S(s, t/b)$ is a Jacobi field associated with the geodesic $\gamma(t)$. By Lemma 19, for any $t \in [0, b]$, we have

$$\boldsymbol{S}(K, b)\|\nabla_{\dot\gamma} J(\gamma(0))\| \leq \|J(\gamma(b))\|$$

and

$$\|J(\gamma(t))\| \leq \boldsymbol{s}(K, t)\|\nabla_{\dot\gamma} J(\gamma(0))\| \leq \boldsymbol{s}(K, b)\|\nabla_{\dot\gamma} J(\gamma(0))\|.$$

Combining the above two inequalities yields

$$\|J(\gamma(t))\| \leq \frac{\boldsymbol{s}(\kappa, b)}{\boldsymbol{S}(K, b)}\|J(\gamma(b))\| = \frac{\boldsymbol{s}(\kappa, b)}{\boldsymbol{S}(K, b)}\|S(s,1)\| = \frac{\boldsymbol{s}(\kappa, b)}{\boldsymbol{S}(K, b)} d(\mathbf{y}, \mathbf{z})$$

holds for any $t \in [0, b]$, which is equivalent to

$$\left\| S\left(s, \frac{t}{b}\right) \right\| \leq \frac{\boldsymbol{s}(\kappa, b)}{\boldsymbol{S}(K, b)} d(\mathbf{y}, \mathbf{z}). \tag{51}$$

holds for any $t \in [0, b]$. By combining Equations (49) , (50) and (51), we find

$$\begin{aligned}\|\Gamma_{\mathbf{z}}^{\mathbf{x}}\Gamma_{\mathbf{y}}^{\mathbf{z}}\Gamma_{\mathbf{x}}^{\mathbf{y}}\mathbf{u} - \mathbf{u}\| \leq& 12K_m\|\mathbf{u}\| \cdot \min\{d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z}), d(\mathbf{x}, \mathbf{z}) + d(\mathbf{y}, \mathbf{z})\} \cdot \frac{\boldsymbol{s}(\kappa, \|T(s,t)\|)}{\boldsymbol{S}(K, \|T(s,t)\|)} d(\mathbf{y}, \mathbf{z}) \\ \leq& 36K_m\|\mathbf{u}\| \cdot \min\{d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z}), d(\mathbf{x}, \mathbf{z}) + d(\mathbf{y}, \mathbf{z})\} \cdot d(\mathbf{y}, \mathbf{z}),\end{aligned}$$

where in the second inequality, we apply

$$\|T(s,t)\| \leq \min\{d(\mathbf{x},\mathbf{y}) + d(\mathbf{y},\mathbf{z}), d(\mathbf{x},\mathbf{z}) + d(\mathbf{y},\mathbf{z})\} \leq \frac{1}{\sqrt{K_m}}$$

and Lemma 20. □

**Remark 6.** *In the literature (Karcher, 1977; Sun et al., 2019), a proposition similar to our Lemma 24 is presented as:*

$$\|\Gamma_{\mathbf{z}}^{\mathbf{x}}\Gamma_{\mathbf{y}}^{\mathbf{z}}\Gamma_{\mathbf{x}}^{\mathbf{y}}\mathbf{u} - \mathbf{u}\| \leq \tilde{C} \cdot d(\mathbf{x},\mathbf{y}) \cdot d(\mathbf{y},\mathbf{z})\|\mathbf{u}\|, \tag{52}$$

*which holds for some constant $\tilde{C}$ and for all $\mathbf{x},\mathbf{y},\mathbf{z} \in \mathcal{M}$. By comparison, Lemma 24 explicitly elucidates the dependence of $\tilde{C}$ on both the diameter of the geodesic triangle $\triangle\mathbf{xyz}$ and the Riemannian curvature.*