Inexact Riemannian Gradient Descent for Nonconvex Optimization: Theory, Algorithms, and Applications in Min-Max Problems

Chris Junchi Li[◇]

Department of Electrical Engineering and Computer Sciences[◇]
University of California, Berkeley

September 22, 2024

Abstract

This paper introduces a novel framework for *Inexact Riemannian Gradient Descent* (IRGD) methods, aimed at solving nonconvex optimization problems on Riemannian manifolds. Our framework generalizes inexact gradient conditions from Euclidean spaces to Riemannian settings, addressing challenges in gradient approximation for both deterministic and stochastic perturbations. We provide rigorous convergence guarantees under the *Kurdyka-Łojasiewicz* (KL) property, ensuring the stationarity of accumulation points and convergence of the gradient sequence. Additionally, we propose an application to Riemannian extragradient methods, showcasing the versatility of our approach in min-max optimization problems.

Keywords: Inexact Riemannian Gradient Descent; Nonconvex Optimization; Riemannian Manifolds; Min-Max Optimization; Extragradient Methods; Kurdyka-Łojasiewicz Property

1 Introduction

Optimization on Riemannian manifolds has gained significant traction due to its wide range of applications in machine learning, signal processing, and statistics, where non-Euclidean geometries naturally arise. Many optimization problems in these fields, such as low-rank matrix completion, principal component analysis (PCA), and robust optimization, involve constraints or structures that are better modeled on manifolds. Classical Euclidean methods, while effective in simpler settings, often fail to account for the geometric complexity inherent in such problems.

One prominent method for solving optimization tasks on Riemannian manifolds is Riemannian Gradient Descent (RGD), which extends classical gradient descent by iterating over tangent spaces of the manifold. However, in practical settings, obtaining the exact Riemannian gradient may not always be feasible due to derivative-free requirements or noisy gradient approximations. This challenge motivates the need for algorithms that can handle inexact gradient information.

In this work, we propose the Inexact Riemannian Gradient Descent (IRGD) framework, which extends classical inexact gradient methods to the Riemannian setting. Our approach incorporates both unnormalized and normalized inexact gradient conditions, which have been well studied in Euclidean optimization, and adapts them for addressing nonconvex optimization on Riemannian manifolds. This framework introduces a flexible mechanism for managing gradient approximation errors, making it robust to both deterministic and stochastic perturbations in gradient information.

Additionally, we explore the application of our IRGD framework to min-max optimization, a foundational tool in adversarial learning, robust optimization, and multi-agent systems. These problems frequently involve constrained optimization within non-Euclidean geometries, such as those represented by Riemannian manifolds. By leveraging the geometric properties of these spaces,

we extend traditional first-order methods like gradient descent ascent (GDA) and extragradient (EG) to the Riemannian setting, offering both theoretical guarantees and practical insights into their behavior in complex optimization landscapes.

Contributions. The main contributions of this paper are as follows:

- We propose a novel framework for Inexact Riemannian Gradient Descent (IRGD), including both unnormalized and normalized gradient conditions, with rigorous convergence analysis.
- We introduce the Riemannian extragradient method, showing its effectiveness under the IRGD framework, and provide theoretical guarantees of convergence for nonconvex optimization problems.
- We extend our framework to min-max optimization in geodesic metric spaces, offering insights into first-order algorithmic performance in non-Euclidean geometries.

Organization. The paper is organized as follows: Section 2 presents the Inexact Riemannian Gradient Descent (IRGD) framework and its convergence analysis. Section 3 introduces the Riemannian extragradient method and discusses its application in optimization problems. In Section 4, we extend our analysis to min-max optimization in geodesic metric spaces. Finally, Section 5 concludes the paper with directions for future research.

2 Inexact Riemannian Gradient Descent for Nonconvex Optimization

We solve the following Riemannian optimization problem:

$$\min_{x\in\mathcal{M}} f(x) \tag{1}$$

where $\mathcal{M}$ is a Riemannian submanifold embedded in $\mathbb{R}^n$, and $f : \mathcal{M} \to \mathbb{R}$ is a continuously differentiable ($\mathcal{C}^1$-smooth) and nonconvex function. One of the most classical and effective methods for solving (1) is the Riemannian gradient descent (RGD) algorithm [AMS08, Smi14]. Given an initial point $x_0 \in \mathcal{M}$, the iterative procedure of the RGD is designed as follows:

$$x_{k+1} = \mathcal{R}_{x_k}(-t_k \mathrm{grad} f(x_k)) \tag{2}$$

for all $k \in \mathbb{N}$, where $\mathcal{R}$ denotes the retraction operator on the manifold $\mathcal{M}$, $t_k \geq 0$ is the stepsize at the k-th iteration, and $\mathrm{grad} f$ is the Riemannian gradient of f, which will be defined in the next section. Due to its simplicity, the RGD method is widely applied to various optimization problems [ATV13, WCCL16, HLZ20, HG23]. However, in many practical scenarios, the exact gradient $\mathrm{grad} f(x)$ may not be accessible, or deterministic errors can occur in the gradient computations. This compels us to develop inexact gradient-based algorithms.

In the Euclidean setting, commonly used inexact gradient conditions are the unnormalized and normalized conditions [Ped04, Car91], defined as follows:

$$\|g - \nabla f(x)\| \leq \epsilon \tag{3}$$

$$\|g - \nabla f(x)\| \leq \nu \|\nabla f(x)\| \tag{4}$$

where g is an approximation of $\nabla f(x)$, and $\epsilon, \nu \geq 0$ are parameters. These two inexact conditions have broad applicability across various domains, including derivative-free optimization [CS18, BCCS22, KMT23a], finite difference approximation [CS18, PS20], and more. Recently, [KMT23b] proposed an inexact gradient method under condition (3) for solving nonconvex smooth problems and provided a convergence analysis. Additionally, an inexact algorithm for nonsmooth convex problems has also been proposed by [KMT24b]. Although inexact algorithms have been developed in Euclidean spaces, no algorithm has yet been proposed for addressing problem (1) under an inexact gradient oracle. The purpose of this paper is to generalize those two inexact conditions to Riemannian manifolds, propose a unified inexact gradient algorithmic framework under those conditions, and provide a comprehensive convergence analysis.

Motivated by the works of [KMT23b, KMT24b, KMT24a], we propose a unified algorithmic framework for solving problem (1) under inexact gradient conditions. The section is organized as follows. Sect. 2.1 covers the preliminaries of the IRGD method. In Sect. 2.2, we introduce our unified algorithmic framework and present the convergence analysis under inexact gradient conditions.

2.1 Preliminaries

The Riemannian concepts presented in this section are consistent with the established Riemannian optimization literature [AMS08]. Let $\mathcal{M} \subset \mathbb{R}^n$ be a differentiable embedded submanifold, equipped with a smoothly varying inner product $\langle \cdot, \cdot \rangle_x : T_x\mathcal{M} \times T_x\mathcal{M} \to \mathbb{R}$ defined on the tangent space $\mathcal{M}$ at each point x. The norm of a vector $\xi_x \in T_x\mathcal{M}$ is then given by $\|\xi_x\|_x := \sqrt{\langle \xi_x, \xi_x \rangle_x}$. The Riemannian gradient $\mathrm{grad} f(x) \in T_x\mathcal{M}$ of a smooth function f at a point $x \in \mathcal{M}$ is the unique tangent vector satisfying $\langle \mathrm{grad} f(x), \xi \rangle_x = Df(x)[\xi]$, $\forall \xi \in T_x\mathcal{M}$, where $Df(x)[\xi]$ represents the directional derivative along the direction ξ_x. As is well known, a point $\overline{x}$ is stationary point for a $\mathcal{C}^1$-smooth function $f \colon \mathcal{M} \to \mathbb{R}$ if $\mathrm{grad} f(\overline{x}) = 0$.

We next introduce the concepts of retraction and vector transport, which are defined as follows. A smooth mapping $R : T\mathcal{M} \to \mathcal{M}$ is called a retraction on a manifold $\mathcal{M}$ if its restriction at x, denoted as $\mathcal{R}_x : T_x\mathcal{M} \to \mathcal{M}$, satisfies

- $\mathcal{R}_x(0_x) = x$ for all $x \in \mathcal{M}$, where 0_x denotes the zero element of $T_x\mathcal{M}$;
- the differential of $\mathcal{R}_x$ at 0_x is an identity mapping on $T_x\mathcal{M}$, i.e, $D\mathcal{R}_x(0_x) = \mathrm{id}_{T_x\mathcal{M}}$.

A smooth mapping $T\mathcal{M} \oplus T\mathcal{M} \to T\mathcal{M} : (\eta_x, \xi_x) \mapsto \mathcal{T}_{\eta_x}(\xi_x)$ is called a vector transport on a manifold $\mathcal{M}$, if it satisfies

- $\mathcal{T}_{\eta_x}\xi_x \in T_{\mathcal{R}_x(\eta_x)}\mathcal{M}$ for all $x \in \mathcal{M}$, and for all $\eta_x, \xi_x \in T_x\mathcal{M}$;
- $\mathcal{T}_{0_x}\xi_x = \xi_x$ for all $\xi_x \in T_x\mathcal{M}$;
- $\mathcal{T}_{\eta_x}(a\xi_x + b\zeta_x) = a\mathcal{T}_{\eta_x}\xi_x + b\mathcal{T}_{\eta_x}\zeta_x$ for all $a, b \in \mathbb{R}$, and for all $\eta_x, \xi_x, \zeta_x \in T_x\mathcal{M}$.

A vector transport $\mathcal{T}$ is called isometric if it satisfies $\langle \mathcal{T}_{\eta_x}\xi_x, \mathcal{T}_{\eta_x}\zeta_x \rangle_{\mathcal{R}_x(\eta_x)} = \langle \xi_x, \zeta_x \rangle_x$, for any $\eta_x, \xi_x, \zeta_x \in T_x\mathcal{M}$. The adjoint operator $\mathcal{T}^\sharp$ of a vector transport $\mathcal{T}$ is defined such that $\langle \xi_y, \mathcal{T}_{\eta_x}\zeta_x \rangle_y = \langle \mathcal{T}^\sharp_{\eta_x}\xi_y, \zeta_x \rangle_y$ for all $\eta_x, \zeta_x \in T_x\mathcal{M}$ and $\xi_y \in T_y\mathcal{M}$, where $y = \mathcal{R}_x(\eta_x)$. The inverse operator $\mathcal{T}^{-1}$ is defined by the condition $\mathcal{T}^{-1}_{\eta_x}\mathcal{T}_{\eta_x} = \mathrm{id}$ for all $\eta_x \in T_x\mathcal{M}$, where id denotes the identity operator.

The global convergence and convergence rates of various nonconvex algorithms benefit from the Riemannian KL property [HW22, BNO11]. The following provides the definition.

Definition 1 (Riemannian Kurdyka-Łojasiewicz property)**.** A continuous smooth function $f : \mathcal{M} \to \mathbb{R}$ enjoys the Riemannian *KL property* at $\overline{x} \in \mathrm{grad} f$ if and only if there exist $\eta \in (0, \infty]$, a neighborhood $U \subset \mathcal{M}$ of $\overline{x}$, and a desingularizing concave continuous function $\varphi : [0, \delta) \to [0, \infty)$ such that:

(i) $\varphi(0) = 0$.

(ii) φ is $\mathcal{C}^1$-smooth on $(0, \delta)$.

(iii) $\varphi' > 0$ on $(0, \delta)$.

(iv) For all $x \in U$ with $f(\overline{x}) < f(x) < f(\overline{x}) + \delta$, we have

$$\varphi'(f(x) - f(\overline{x}))\|\mathrm{grad} f(x)\| \geq 1 \tag{5}$$

The following lemma establishes that if the Riemannian KL property holds at every single point within a compact set that shares the same function value, then there exists a single desingularizing function such that the Riemannian KL property holds uniformly for all points within that compact set. This generalization is also implicitly covered in the proofs presented in [Hos15, HW22].

Lemma 1. *Let $\overline{\Omega}$ be a compact set in $\mathcal{M}$ and let $h : \mathcal{M} \to (-\infty, \infty]$ be a continuous function. Assume that h is a constant on $\overline{\Omega}$ and satisfies the Riemannian KL property at each point of $\overline{\Omega}$. Then, there exists $\overline{\omega}, \delta > 0$ and a continuous concave function $\varphi : [0, \delta] \to [0, \infty)$ such that for all $\overline{u}$ in $\overline{\Omega}$ and all u in the following intersection:*

$$\{u \in \mathcal{M} : \inf_{v \in \overline{\Omega}} \|u - v\| < \overline{\omega}\} \cap \{u \in \mathcal{M} : h(\overline{u}) < h(u) < h(\overline{u}) + \delta\}$$

one has

$$\varphi'(h(u) - h(\overline{u}))\|\mathrm{grad} h(u)\| \geq 1$$

2.2 Inexact Riemannian gradient descent methods

In this section, we introduce a unified algorithmic framework for the IRGD method, featuring two inexact gradient conditions designed to solve problem (1). The IRGD techniques are defined as iterative optimization schemes, given by

$$x_{k+1} = \mathcal{R}_{x_k}(-t_k g_k) \tag{6}$$

where t_k is a diminishing stepsize, and g_k is an approximation of $\mathrm{grad} f(x_k)$ that satisfies

$$\|g_k - \mathrm{grad} f(x_k)\| \leq \epsilon_k \tag{7}$$

or

$$\|g_k - \mathrm{grad} f(x_k)\| \leq \nu \|\mathrm{grad} f(x_k)\| \tag{8}$$

where $\epsilon_k > 0$ and $\nu \geq 0$ is the relative error parameter. The main motivation for the construction algorithm is that by making the inexact gradient g_k, it avoids the situation where the gradient is not available, which ensures the convergence of the algorithm.

Next, we recall some basic stepsize selections for the iterative procedure (6). If the step size t_k satisfies the Armijo rule, it guarantees the nonincreasing property of the sequence $\{f(x_k)\}$. However, the Armijo stepsize may be very small, leading to many iterations with only minor changes

in the sequence. Although a constant step size can significantly simplify the iterative design, it does not generally guarantee the nonincreasing property of $\{f(x_k)\}$, resulting in inefficiency [Ber97, NW99]. This observation motivated our approach, which is based on the analysis of a diminishing stepsize given by

$$\sum_{k=1}^{\infty} t_k = \infty \ \text{ and } \ \sum_{k=1}^{\infty} t_k^2 < \infty \tag{9}$$

This step size also ensures that $t_k \downarrow 0$, which is satisfied for the cosine step size scheduler in each cycle [LH16], making it a more favorable approach.

Before giving our algorithms, the following key lemma provides a unified conclusion for the convergence analysis of our methods.

Lemma 2. *[BAC19b] Assume that there exist a constant $\kappa > 0$ such that*

$$\|\mathcal{R}_x(\eta) - x\| \leq \kappa \|\eta\|$$

for all $x \in \mathcal{M}$ and $\eta \in T_x\mathcal{M}$.

Lemma 3. *[KLMT24] Let $\{\alpha_k\}, \{\beta_k\}, \{\gamma_k\}$ be sequences of nonnegative numbers satisfying the conditions*

$$\alpha_{k+1} - \alpha_k \leq \beta_k \alpha_k + \gamma_k \ \text{ for sufficient large } \ k \in \mathbb{N} \tag{10}$$

$$\{\beta_k\} \text{ is bounded} \qquad \sum_{k=1}^{\infty} \beta_k = \infty, \ \sum_{k=1}^{\infty} \gamma_k < \infty, \ \text{ and } \ \sum_{k=1}^{\infty} \beta_k \alpha_k^2 < \infty \tag{11}$$

Then we have that $\alpha_k \to 0$ as $k \to \infty$.

Lemma 4. *[KMT22] Let $\{x_k\}$ and $\{\eta_k\}$ be sequences in $\mathcal{M}$ satisfying the condition*

$$\sum_{k=1}^{\infty} \|x_{k+1} - x_k\| \cdot \|\eta_k\| < \infty \tag{12}$$

If $\overline{x}$ is an accumulation point of the sequence $\{x_k\}$ and 0 is an accumulation points of the sequence $\{\eta_k\}$, then there exists an infinite set $J \subset \mathbb{N}$ such that we have

$$x_k \xrightarrow{J} \overline{x} \ \text{ and } \ \eta_k \xrightarrow{J} 0 \tag{13}$$

Remark. We use $\mathbb{N} := \{1, 2, \ldots\}$ signifies the collection of natural numbers. The symbol $x_k \xrightarrow{J} \overline{x}$ means that $x_k \to \overline{x}$ as $k \to \infty$ with $k \in J \subset \mathbb{N}$.

2.2.1 Unnormalized condition

In this subsection, we present a general framework for our novel IRGD method, focusing on cases where the inexact gradient satisfies the unnormalized condition given by (7). We also provide a detailed convergence analysis of the IRGD method.

The algorithm proposed in this work is inspired by the inexact gradient descent methods studied in Euclidean space by [KMT23b, KMT24b, KLMT24]. The latter methods consider relative errors in gradient calculation, while Algorithm 1 uses absolute errors. The following assumption presents the Riemannian setting of L-smooth functions.

Algorithm 1 Inexact Riemannian Gradient Descent (IRGD) Methods

1: Choose some initial point $x_0 \in \mathcal{M}$, sequence of errors $\{\epsilon_k\} \subset (0, \infty)$, and sequence of stepsizes $\{t_k\} \subset (0, \infty)$. For $k = 1, 2, \ldots$, do the following
2: Set $x_{k+1} = \mathcal{R}_{x_k}(-t_k g_k)$ with $\|g_k - \mathrm{grad} f(x_k)\| \leq \epsilon_k$, where $g_k \in T_{x_k}\mathcal{M}$

Assumption 1 (L-Retraction Smoothness)**.** *Assume that there exists a constant $L > 0$ such that*

$$f(\mathcal{R}_x(\eta)) \leq f(x) + \langle \mathrm{grad} f(x), \eta \rangle + \frac{L}{2}\|\eta\|^2$$

for all $x \in \mathcal{M}, \eta \in T_x\mathcal{M}$.

The following lemma immediately confirms that IRGD is a descent algorithm.

Lemma 5. *Suppose that Assumption 1 holds. Let $\{x_k\}$ be generated by Algorithm 1 with stepsizes and errors satisfying the conditions*

$$\sum_{k=1}^{\infty} t_k = \infty \qquad t_k \downarrow 0 \qquad \sum_{k=1}^{\infty} t_k \epsilon_k < \infty \qquad \limsup \epsilon_k < 2 \tag{14}$$

Then there exists $K > 0$ such that for $k > K$,

$$f(\mathcal{R}_{x_k}(-t_k g_k)) \leq f(x_k) - c_1 t_k \|\mathrm{grad} f(x_k)\|^2 + c_2 t_k \epsilon_k$$

Assumption 2. *Assume that the objective function f is bounded below, i.e., $\inf_{k\in\mathbb{N}} f(x_k) > -\infty$.*

To analyze the convergence properties of the IRGD method based on the Riemannian KL property, we require a result analogous to those in [HW22] concerning vector transport, under the following assumption.

Assumption 3 (Individual Retraction Lipschitzness)**.** *A function $f : \mathcal{M} \to \mathbb{R}$ is said to posses a Lipschitz continuous gradient with the uniform constant $L > 0$, or equivalently it belongs to the class $\mathcal{C}^{1,L}$, if we have the estimate*

$$\|\mathcal{T}_y^x \mathrm{grad} f(y) - \mathrm{grad} f(x)\| \leq L\|\eta\|, \quad \text{or} \quad \|\mathrm{grad} f(y) - \mathcal{T}^{-\sharp}_{\mathcal{R}_x(\eta)} \mathrm{grad} f(x)\| \leq L\|\eta\|$$

for all $x, y = \mathcal{R}_x(\eta) \in \mathcal{M}$.

The global convergence properties of Algorithm 1 are detailed in the following theorem, which addresses both the gradient sequences and the function values.

Theorem 1. *Suppose that Assumption 1, 2 and 3 holds. Let $\{x_k\}$ be generated by Algorithm 1 with stepsize and errors satisfying the conditions* (14)*. Then the following convergence properties hold:*

(i) $\mathrm{grad} f(x_k) \to 0$*, and thus every accumulation point of the iterative sequence $\{x_k\}$ is stationary for f.*

(ii) If $\overline{x}$ is an accumulation point of the sequence $\{x_k\}$, then $f(x_k) \to f(\overline{x})$.

Algorithm 2 IRGDr method

1: Choose some initial point $x_0 \in \mathcal{M}$, a relative error $\nu \geq 0$, and a sequence of stepsizes $\{t_k\} \subset [0, \infty)$. For $k = 1, 2, \ldots$, do the following
2: Set $x_{k+1} = \mathcal{R}_{x_k}(-t_k g_k)$ with $\|g_k - \mathrm{grad} f(x_k)\| \leq \nu \|\mathrm{grad} f(x_k)\|$, where $g_k \in T_{x_k}\mathcal{M}$

The following assumption on the desingularizing function, as defined in Definition 1, is utilized by [LMQ23].

Assumption 4. *There is some $C > 0$ such that*

$$C[\varphi'(x+y)]^{-1} \leq (\varphi'(x))^{-1} + (\varphi'(y))^{-1}$$

whenever $x, y \in (0, \delta)$ with $x + y < \delta$.

When f satisfies the Riemannian KL property in the set of all accumulation points, the following convergence result holds for the sequence of iterates.

Theorem 2. *Under the same condition as in the Theorem 1 and $\mathcal{S}$ denote the set of all accumulation points. Suppose that f satisfies the Riemannian KL property at every point in $\mathcal{S}$ with the desingularizing function φ satisfying Assumption 4. Assume in addition that*

$$\sum_{k=1}^{\infty} t_k \Big(\varphi'(\sum_{i=k}^{\infty} t_k \epsilon_k) \Big)^{-1} < \infty \tag{15}$$

and that $f(x_k) > f(\overline{x})$ for sufficiently large $k \in \mathbb{N}$. Then $x_k \to \overline{x}$ as $k \to \infty$. In particular, if $\overline{x}$ is a global minimizer of f, then either $f(x_k) = f(\overline{x})$ for some $k \in \mathbb{N}$, or $x_k \to \overline{x}$.

2.2.2 Normalized condition

In this subsection, we present a general framework for the standardized variants of our novel IRGD method with relative error and analyze the convergence properties of IRGDr under the inexact gradient normalized condition given by (8).

This algorithm differs from Algorithm 1 in that the convergence properties of IRGDr are established under stronger stepsize assumptions and provide additional conclusions on the convergence rate. This makes it particularly suitable for constructing the REG method discussed in Sect. 3. The following lemma verifies the descent property of the objective function.

Lemma 6. *Let $\{x_k\}$ be the sequence generated by Algorithm 2. It holds that*

$$(1 - \nu)\|\mathrm{grad} f(x_k)\| \leq \|g_k\| \leq (1 + \nu)\|\mathrm{grad} f(x_k)\|$$

Based on the assumption on the stepsize, we can derive the fundamental convergence properties of Algorithm 2. These properties include the stationarity of accumulation points, the convergence of the gradient sequence to the origin, and the convergence of the function values to the optimal value. These results are presented in the following theorem.

Theorem 3. *Suppose that Assumption 1, 2, 3 holds. Let $f : \mathcal{M} \to \mathbb{R}$ be a smooth function satisfying the descent condition for some constant $L > 0$, and let $\{x_k\}$ be the sequence generated by Algorithm 2 with the relative error $\nu \in [0, 1)$, and the stepsizes satisfying*

$$\sum_{k=1}^{\infty} t_k = \infty \quad \text{and} \quad t_k \in [0, \frac{2 - 2\nu - \delta}{L(1+\nu)^2}] \tag{16}$$

for sufficiently large $k \in \mathbb{N}$ and for some $\delta > 0$. Then either $f(x_k) \to -\infty$, or we have the assertions:

(i) Every accumulation point of $\{x_k\}$ is a stationary point of the cost function f.

(ii) If the sequence $\{x_k\}$ has any accumulation point $\overline{x}$, then $f(x_k) \downarrow f(\overline{x})$.

(iii) If $f \in \mathcal{C}^{1,L}$, then $\mathrm{grad} f(x_k) \to 0$.

The following theorem on the convergence of iterates to stationary points is ensured under the Riemannian KL property of the objective function, with convergence rates determined by the Riemannian KL exponent.

Theorem 4. *Under the same condition as in Theorem 3 and $\mathcal{S}$ denote the set of all accumulation points.*

(i) Assume the stepsizes are bounded away from 0, if f satisfies the Riemannian KL property at some accumulation point in $\mathcal{S}$, then $\{x_k\} \to \overline{x}$.

(ii) Assume the Riemannian KL property in (i) holds with the desingularizing function $\varphi(t) = \frac{C}{\theta} t^\theta$ with $C > 0$ and $\theta \in (0, 1)$. Then either $\{x_k\}$ stops finitely at a stationary point, or the following convergence rates are achieved:

- *If $\theta = 1$, then there exists k_1 such that $x_k = \overline{x}$ for all $k > k_1$.*
- *If $\theta \in [\frac{1}{2}, 1)$, then there exists $C_r > 0$ and $Q \in [0, 1)$ such that for all k*

$$\|x_k - \overline{x}\| < C_r Q^k$$

- *If $\theta \in (0, \frac{1}{2})$, then there exists a positive constant > 0 such that for all k*

$$\|x_k - \overline{x}\| < k^{-\frac{1-\theta}{2\theta-1}}$$

2.2.3 Proof of Theorem 3

Proof of Theorem 3. Using condition (16), we find $N \in \mathbb{N}$ so that $2 - 2\nu - Lt_k(1+\nu)^2 \geq \delta$ for all $k \geq N$. Select such a natural number k and use the Lipschitz continuity of $\mathrm{grad} f$ with constant L to deduce from Assumption 1, the relationship $x_{k+1} = \mathcal{R}_{x_k}(-t_k g_k)$, the estimates (46)–(48) and (50) that

$$\begin{aligned}
f(x_{k+1}) &\leq f(x_k) - t_k \langle \mathrm{grad} f(x_k), g_k \rangle + \frac{Lt_k^2}{2} \|g_k\|^2 \\
&\leq f(x_k) - t_k \langle \mathrm{grad} f(x_k), g_k \rangle + Lt_k^2 \langle \mathrm{grad} f(x_k), g_k \rangle - \frac{Lt_k^2(1-\nu^2)}{2} \|\mathrm{grad} f(x_k)\|^2 \\
&\leq f(x_k) - t_k(1-\nu)\|\mathrm{grad} f(x_k)\|^2 + Lt_k^2(1+\nu)\|\mathrm{grad} f(x_k)\|^2 - \frac{Lt_k^2(1-\nu^2)}{2} \|\mathrm{grad} f(x_k)\|^2 \\
&= f(x_k) - \frac{t_k}{2}\Big(2 - 2\nu - Lt_k(1+\nu)^2\Big)\|\mathrm{grad} f(x_k)\|^2 \\
&\leq f(x_k) - \frac{\delta t_k}{2} \|\mathrm{grad} f(x_k)\|^2 \text{ for all } k \geq N
\end{aligned} \tag{17}$$

It follows from the above that the sequence $\{f(x_k)\}_{k\geq N}$ is nonincreasing, and hence the condition $\inf_{k\in\mathbb{N}} f(x_k) > -\infty$ ensures the convergence of $\{f(x_k)\}$. This allows us to deduce from (17) that

$$\frac{\delta}{2}\sum_{k=N}^{\infty} t_k\|\text{grad}f(x_k)\|^2 \leq \sum_{K=N}^{\infty}(f(x_k)-f(x_{k+1})) \leq f(x_k) - \inf_{k\in\mathbb{N}} f(x_k) < \infty \tag{18}$$

Combining the latter with (50) and $x_{k+1} = \mathcal{R}_{x_k}(-t_k g_k)$ gives us

$$\sum_{k=1}^{\infty}\|x_{k+1}-x_k\|\cdot\|g_k\| \leq \sum_{k=1}^{\infty}\kappa t_k\|g_k\|^2 \leq \kappa(1+\nu)^2\sum_{k=1}^{\infty} t_k\|\text{grad}f(x_k)\|^2 < \infty \tag{19}$$

Now we are ready to verify all the assertions of the theorem. Let us start with (i) and show that 0 in an accumulation point of $\{g_k\}$. Indeed, supposing the contrary gives us $\epsilon > 0$ and $K \in \mathbb{N}$ such that $\|g_k\| \geq \epsilon$ for all $k \geq K$, and therefore

$$\infty > \sum_{k=K}^{\infty} t_k\|g_k\|^2 \geq \sum_{k=K}^{\infty} t_k = \infty$$

which is a contradiction justifying that 0 is an accumulation point of $\{g_k\}$. If $\overline{x}$ is an accumulation point of $\{x_k\}$, then by Lemma 4 and (19), we find an infinite set $J \subset N$ such that $x_k \xrightarrow{J} \overline{x}$ and $g_k \xrightarrow{J} 0$. The latter being combined with (50) gives us $\text{grad}f(x_k) \xrightarrow{J} 0$, which yields the stationary condition $\text{grad}f(\overline{x}) = 0$.

To verity (ii), let $\overline{x}$ be an accumulation point of $\{x_k\}$ and find an infinite set $J \subset \mathbb{N}$ such that $x_k \xrightarrow{J} \overline{x}$. Combining this with the continuity of f and the fact that $\{f(x_k)\}$ is convergent, we arrive at the equalities

$$f(\overline{x}) = \lim_{k\in J} f(x_k) = \lim_{k\in\mathbb{N}} f(x_k)$$

which therefore justify assertion (ii).

To proceed with the proof of (iii), assume that $\text{grad}f$ is Lipschitz continuous with constant $L > 0$ and employ Lemma 3 with $\alpha_k := \|\text{grad}f(x_k)\|$, $\beta_k := Lt_k(1+\nu)$, and $\gamma_k := 0$ for all $k \in \mathbb{N}$ to derive that $\text{grad}f(x_k) \to 0$. Observe first that condition (10) of this lemma is satisfied due to the the estimates

$$\begin{aligned}\alpha_{k+1}-\alpha_k &= \|\text{grad}f(x_{k+1})\| - \|\text{grad}f(x^k)\|\\ &\leq \|\mathcal{T}_{x_{k+1}}^{x_k}\text{grad}f(x_{k+1}) - \text{grad}f(x_k)\|\\ &= Lt_k\|g_k\| \leq Lt_k(1+\nu)\|\text{grad}f(x_k)\| = \beta_k\alpha_k\end{aligned}$$

The conditions in (11) of the lemma are satisfied since $\{t_k\}$ is bounded, $\sum_{k=1}^{\infty} t_k = \infty$ by (16), $\gamma_k = 0$, and

$$\sum_{k=1}^{\infty}\beta_k\alpha_k^2 = L(1+\nu)\sum_{k=1}^{\infty} t_k\|\text{grad}f(x_k)\|^2 < \infty$$

where the inequality follows from (18). Thus applying Lemma 3 gives us $\text{grad}f(x_k) \to 0$ as $k \to \infty$. □

2.2.4 Proof of Theorem 4

Proof of Theorem 4. To prove (i). Since 0 is an accumulation point of $\{g_k\}$, and the application of Lemma 2 and implies that

$$\|x_{k+1} - x_k\| = \|\mathcal{R}_{x_k}(t_k g_k) - x_k\| \leq \kappa t_k \|g_k\| \to 0 \tag{20}$$

Then by [BST14], we konw that $\mathcal{S}$ is a compect set. Morever, since $f(x_k)$ is nonincreasing, and f is continuous, thus f has the same value at all the points in $\mathcal{S}$. Therefore, by Lemma 1, there exists a single desingularizing function φ, for the Riemannian KL property of f to hold at all the points in $\mathcal{S}$. Since $f(x_k) \to f(\overline{x})$, $\inf_{\overline{x}\in\mathcal{S}} \|x_k - \overline{x}\| \to 0$, thus there exists an $l > 0$ such that

$$\varphi'(f(x_k) - f(\overline{x})) \geq \|\mathrm{grad} f(x_k)\|^{-1} \text{ for all } k \geq l \tag{21}$$

By Assumption 3, we have

$$\|\mathrm{grad} f(\mathcal{R}_{x_k}(-t_k g_k) - \mathcal{T}^{-\sharp}_{\mathcal{R}_{x_k}(-t_k g_k)} \mathrm{grad} f(x_k)\| \leq L t_k \|g_k\| \text{ for all } k \geq l \tag{22}$$

Since the vector transport is isometric, it holds that

$$\|\mathcal{T}^{-\sharp}_{x_{k+1}} \mathrm{grad} f(x_k)\| = \|\mathrm{grad} f(x_k)\| \tag{23}$$

Combining the triangle inequality with (22), and (23), we get

$$\begin{aligned}\|\mathrm{grad} f(x_{k+1})\| &\leq \|\mathcal{T}^{-\sharp}_{x_{k+1}} \mathrm{grad} f(x_k)\| + \|\mathrm{grad} f(x_{k+1}) - \mathcal{T}^{-\sharp}_{x_{k+1}} \mathrm{grad} f(x_k)\| \\ &\leq \|\mathrm{grad} f(x_k)\| + L t_k \|g_k\|\end{aligned}$$

Suppose in addition that the sequence $\{t_k\}$ is bounded away from 0 (i.e. there is some $\bar{t} > 0$ such that $t_k > \bar{t}$ for large $k \in \mathbb{N}$). Using the inequality (50), we get

$$\begin{aligned}\|\mathrm{grad} f(x_{k+1})\| &\leq \|\mathrm{grad} f(x_k)\| + L t_k \|g_k\| \leq \frac{1}{1-\nu}\|g_k\| + L t_k \|g_k\| \\ &\leq (\frac{1}{1-\nu}\frac{1}{t_k} + L) t_k \|g_k\| \leq (\frac{1}{1-\nu}\frac{1}{\bar{t}} + L) t_k \|g_k\| \\ &\leq \widetilde{L} t_k \|g_k\|\end{aligned}$$

where $\widetilde{L} = \frac{1}{(1-\nu)}\frac{1}{\bar{t}} + L$ is a constant. Thus, we have

$$\|\mathrm{grad} f(x_k)\| \leq \widetilde{L} t_{k-1} \|g_{k-1}\| \text{ for all } k \geq l \tag{24}$$

Inserting this into (21) gives

$$\varphi'(f(x_k) - f(\overline{x})) \geq \widetilde{L}^{-1} (t_{k-1}\|g_{k-1}\|)^{-1} \text{ for all } k \geq l \tag{25}$$

Moreover, it follows from (50) and (17) that

$$f(x_{k+1}) \leq f(x_k) - \frac{\delta t_k}{2(1+\nu)^2}\|g_k\|^2 \tag{26}$$

and the concavity of φ yields that

$$\begin{aligned}\varphi(f(x_k)-f(\overline{x}))-\varphi(f(x_{k+1})-f(\overline{x})) &\geq \varphi'(f(x_k)-f(\overline{x}))(f(x_k)-f(x_{k+1}))\\ &\geq \|\mathrm{grad} f(x_k)\|^{-1}\frac{\delta}{2(1+\nu)^2}t_k\|g_k\|^2\\ &\geq \widetilde{L}^{-1}\frac{\delta}{2(1+\nu)^2}\frac{t_k\|g_k\|^2}{t_{k-1}\|g_{k-1}\|} \end{aligned} \tag{27}$$

For convenience, we define for all $p,q\in\mathbb{N}$ and $\overline{x}$ the following quantities $\Delta_{p,q}:=\varphi(f(x_p)-f(\overline{x}))-\varphi(f(x_q)-f(\overline{x}))$ and $M:=\frac{2(1+\nu)^2\widetilde{L}}{\delta}\in(0,\infty)$. Consequently, (27) is equivalent to

$$\Delta_{k,k+1}\geq\frac{t_k\|g_k\|^2}{Mt_{k-1}\|g_{k-1}\|} \tag{28}$$

for all $k>l$ and hence

$$t_k\|g_k\|^2\leq M\Delta_{k,k+1}t_{k-1}\|g_{k-1}\|$$

Using the fact that the AM-GM inequality, we infer

$$2t_k\|g_k\|=2\sqrt{t_k}\sqrt{t_k\|g_k\|^2}\leq 2\sqrt{t_{k-1}\|g_{k-1}\|M\Delta_{k,k+1}}\leq t_{k-1}\|g_{k-1}\|+M\Delta_{k,k+1} \tag{29}$$

Let us now prove that for all $k>l$ the following inquality holds

$$\sum_{i=l+1}^{k}t_k\|g_k\|\leq t_l\|g_l\|+M\Delta_{l+1,k+1}$$

Summing up (29) for $i=l+1,\cdots,k$ yields

$$\begin{aligned}2\sum_{i=l+1}^{k}t_i\|g_i\| &\leq \sum_{i=l+1}^{k}t_{i-1}\|g_{i-1}\|+M\sum_{i=l+1}^{k}\Delta_{i,i+1}\\ &\leq \sum_{i=l+1}^{k}t_i\|g_i\|+t_l\|g_l\|+M\sum_{i=l+1}^{k}\Delta_{i,i+1}\\ &= \sum_{i=l+1}^{k}t_i\|g_i\|+t_l\|g_l\|+M\Delta_{l+1,k+1}\end{aligned}$$

where the last inequality follows from the fact that $\Delta_{p,q}+\Delta_{q,r}=\Delta_{p,r}$ for all $p,q,r\in\mathbb{N}$. Since $\varphi>0$, we thus have for all $k>l$ that

$$\sum_{i=l+1}^{k}t_i\|g_i\|\leq t_l\|g_l\|+M\varphi(f(x_{l+1})-f(\overline{x}))$$

Combining Assumption 2 and (20), this easily shows that the sequence $\{x_k\}_{k\in\mathbb{N}}$ has finite length, that is,

$$\sum_{k=1}^{\infty}\|x_{k+1}-x_k\|\leq\kappa\sum_{k=1}^{\infty}t_k\|g_k\|\leq\infty \tag{30}$$

The latter means that $\{x_k\}$ is a Cauchy sequence, and tell us that $\{x_k\}$ is converges to $\overline{x}$.

Let us now verify assertion (ii) of the theorem. Applying $\varphi(t) = \frac{C}{\theta}t^\theta$ to (27) yields

$$t_k\|g_k\|^2 \leq t_{k-1}\|g_{k-1}\|\frac{2(1+\nu)^2\widetilde{L}C}{\delta\theta}\Big((f(x_k)-f(\overline{x}))^\theta - (f(x_{k+1})-f(\overline{x}))^\theta\Big) \tag{31}$$

for all $k \geq l$. Taking square root to the both sides of (31) and using the AM-GM inequality, we have

$$2t_k\|g_k\| = t_{k-1}\|g_{k-1}\| + \frac{2(1+\nu)^2\widetilde{L}C}{\delta\theta}\Big((f(x_k)-f(\overline{x}))^\theta - (f(x_{k+1})-f(\overline{x}))^\theta\Big) \tag{32}$$

for all $k \geq l$. Summing the both sides from $p > l$ to ∞ yields

$$\sum_{k=p}^{\infty} t_k\|g_k\| \leq t_{p-1}\|g_{p-1}\| + \frac{2(1+\nu)^2\widetilde{L}C}{\delta\theta}(f(x_p)-f(\overline{x}))^\theta \tag{33}$$

By (21), we have $\frac{1}{C}(f(x_k)-f(\overline{x}))^{1-\theta} \leq \|\mathrm{grad} f(x_k)\|$. Combining this inequality with (24) yields

$$\frac{1}{C}(f(x_k)-f(\overline{x}))^{1-\theta} \leq \widetilde{L}t_{k-1}\|g_{k-1}\| \tag{34}$$

It follows from (33) and (34) that

$$\sum_{k=p}^{\infty} t_k\|g_k\| \leq t_{k-1}\|g_{k-1}\| + \frac{2(1+\nu)^2\widetilde{L}C}{\delta\theta}(CLt_{k-1}\|g_{k-1}\|)^{\frac{\theta}{1-\theta}}, \text{ for all } k \geq l \tag{35}$$

Define $\Delta_k = \sum_{i=k}^{\infty}\|g_i\|$. Therefore, inequality (35) becomes

$$\Delta_k \leq (\Delta_{k-1}-\Delta_k) + b_1(\Delta_{k-1}-\Delta_k)^{\frac{\theta}{1-\theta}}, \text{ for all } k \geq l \tag{36}$$

where $b_1 = \frac{2(1+\nu)^2\widetilde{L}C}{\delta\theta}(C\widetilde{L})^{\frac{\theta}{1-\theta}}$. Noting that (36) has the same form as [AB09], we have follow the same derivations in [AB09] and show that

- if $\theta = 1$, then Algorithm 2 terminates in finite steps,
- if $\theta \in [\frac{1}{2}, 1)$, then $\Delta_k < C_r Q^k$ for $C_r > 0$ and $Q \in [0, 1)$,
- if $\theta \in (0, \frac{1}{2})$, then $\Delta_k < k^{-\frac{1-\theta}{2\theta-1}}$ for > 0.

It only remains to show that $\mathrm{dist}(x_k, \overline{x}) < C_p\Delta_k$ for a positive constant C_p. This can be obtained by

$$\|x_k - \overline{x}\| \leq \sum_{i=k}^{\infty}\|x_{i+1}-x_i\| \leq \kappa\sum_{i=k}^{\infty} t_i\|g_i\| = \kappa\Delta_k$$

where the first inequality is by triangle inequality and the second inequality is from Assumption 4. This completes the proof. □

Algorithm 3 Riemannian Extragradient

1: Choose some initial point $x_1 \in \mathcal{M}$, a sequence of perturbation radii $\{\rho_k\} \subset [0, \infty)$, and a sequence of stepsizes $\{t_k\} \subset [0, \infty)$. For $k = 1, 2, \ldots,$ do the following

2: Set $x_{k+1} = \mathcal{R}_{x_k}(-t_k \mathcal{T}^{x_k}_{x_k^{adv}} \mathrm{grad} f(x_k^{adv}))$, where $x_k^{adv} = \mathcal{R}_{x_k}(-\rho_k \mathrm{grad} f(x_k))$.

3 Applications of IRGD to Riemannian Extragradient

In this section, we apply the inexact Riemannian gradient method for $\mathcal{C}^1$-smooth nonconvex optimization, as developed in Sect. 2.2, to design and justify a novel gradient-based inexact methods.

Riemannian extragradient (REG) was initially proposed as a novel first-order method designed for monotone Riemannian variational inequality problems, with both last-iterate and average-iterate convergence established under strong assumptions [HWW+24]. In contrast, in this subsection, we propose a novel REG method focused on analyzing nonconvex smooth problems from an inexact gradient perspective, offering convergence analysis under more general assumptions.

It is demonstrated that Algorithm 3 is a specific instance of IRGDr utilizing relative errors, as outlined in Algorithm 2. Consequently, it inherits all the convergence properties detailed in Theorem 3 and 4. The following theorem verifies that Algorithm 3 satisfies the inexact normalized gradient condition given in (8).

Theorem 5. *Suppose that Assumption 1, 2, 3 and 4 holds. Let $f : \mathcal{M} \to \mathbb{R}$ be a $\mathcal{C}^1$-smooth function satisfying the descent condition with some constant $L > 0$. Let $\{x_k\}$ be the sequence generated by Algorithm 3 with $\rho_k \le \frac{\nu}{L}$ for some $\nu \in [0, 1)$ and with the stepsize satisfying* (16)*. Then all the convergence properties in Theorem 3 and 4 hold.*

3.1 Proof of Theorem 5

Proof of Theorem 5. Defining $g_k := \mathcal{T}^{x_k}_{x_k^{adv}} \mathrm{grad} f(\mathcal{R}_{x_k}(-\rho_k \mathrm{grad} f(x_k)))$ and utilizing $\rho_k \le \frac{\nu}{L}$, we obtain

$$\begin{aligned} \|g_k - \mathrm{grad} f(x_k)\| &= \|\mathcal{T}^{x_k}_{x_k^{adv}} \mathrm{grad} f(\mathcal{R}_{x_k}(-\rho_k \mathrm{grad} f(x_k))) - \mathrm{grad} f(x_k)\| \\ &\le L\| - \rho_k \mathrm{grad} f(x_k)\| \le \nu \|\mathrm{grad} f(x_k)\| \end{aligned}$$

which verifies the inexact condition in Step 2 of Algorithm 2. Therefore, all the convergence properties in Theorem 3 and 4 hold for Algorithm 3. □

3.2 Deferred Auxiliary Proofs

3.2.1 Proof of Lemma 5

Proof of Lemma 5. First, fix $k \in \mathbb{N}$ and deduce from the Cauchy-Schwarz inequality that

$$\begin{aligned} \langle g_k, \mathrm{grad} f(x_k)\rangle &= \langle g_k - \mathrm{grad} f(x_k), \mathrm{grad} f(x_k)\rangle + \|\mathrm{grad} f(x_k)\|^2 \\ &\ge -\|g_k - \mathrm{grad} f(x_k)\| \cdot \|\mathrm{grad} f(x_k)\| + \|\mathrm{grad} f(x_k)\|^2 \\ &\ge -\epsilon_k \|\mathrm{grad} f(x_k)\| + \|\mathrm{grad} f(x_k)\|^2 \end{aligned} \tag{37}$$

By (14), we find some $c_1 > 0, c_2 \in (0, 1)$, and $K \in \mathbb{N}$ such that

$$\frac{1}{2}(2 - Lt_k - \epsilon_k + Lt_k\epsilon_k) \ge c_1, \quad \frac{1}{2}(1 - Lt_k) + \frac{Lt_k\epsilon_k}{2} \le c_2, \quad \text{and} \quad Lt_k < 1 \tag{38}$$

for all $k \geq K$. Since $\mathrm{grad} f$ is Lipschitz continuous with constant L, it follows from the descent condition in Assumption 1 and the estimate (37) that

$$
\begin{aligned}
f(\mathcal{R}_{x_k}(-t_k g_k)) &\leq f(x_k) - t_k \langle \mathrm{grad} f(x_k), g_k \rangle + \frac{L t_k^2}{2} \|g_k\|^2 \\
&= f(x_k) - t_k(1 - L t_k)\langle \mathrm{grad} f(x_k), g_k \rangle + \frac{L t_k^2}{2}(\|g_k - \mathrm{grad} f(x_k)\|^2 - \|\mathrm{grad} f(x_k)\|^2) \\
&\leq f(x_k) - t_k(1 - L t_k)\Big(- \epsilon_k \|\mathrm{grad} f(x_k) + \|\mathrm{grad} f(x_k)\|^2 \Big) + \frac{L t_k^2 \epsilon_k^2}{2} - \frac{L t_k^2}{2} \|\mathrm{grad} f(x_k)\|^2 \\
&= f(x_k) - \frac{t_k}{2}(2 - L t_k)\|\mathrm{grad} f(x_k)\|^2 + t_k(1 - L t_k)\epsilon_k \|\mathrm{grad} f(x_k)\| + \frac{L t_k^2 \epsilon_k^2}{2} \\
&\leq f(x_k) - \frac{t_k}{2}(2 - L t_k)\|\mathrm{grad} f(x_k)\|^2 + \frac{1}{2} t_k(1 - L t_k)\epsilon_k \Big(1 + \|\mathrm{grad} f(x_k)\|^2 \Big) + \frac{L t_k^2 \epsilon_k^2}{2} \\
&= f(x_k) - \frac{t_k}{2}(2 - L t_k - \epsilon_k + L t_k \epsilon_k)\|\mathrm{grad} f(x_k)\|^2 + \frac{1}{2} t_k \epsilon_k (1 - L t_k) + \frac{L t_k^2 \epsilon_k^2}{2} \\
&= f(x_k) - \frac{t_k}{2}(2 - L t_k - \epsilon_k + L t_k \epsilon_k)\|\mathrm{grad} f(x_k)\|^2 + t_k \epsilon_k \Big(\frac{1}{2}(1 - L t_k) + \frac{L t_k \epsilon_k}{2} \Big)
\end{aligned}
$$

Combining this with (38) gives us

$$
f(\mathcal{R}_{x_k}(-t_k g_k)) \leq f(x_k) - c_1 t_k \|\mathrm{grad} f(x_k)\|^2 + c_2 t_k \epsilon_k \ \text{ whenever } \ k \geq K \tag{39}
$$

□

Proof of Theorem 1. Defining $u_k := c_2 \sum_{i=k}^{\infty} t_i \epsilon_i$ for $k \in \mathbb{N}$, we get that $u_k \to 0$ as $k \to \infty$ and $u_k - u_{k+1} = c_2 t_k \epsilon_k$ for all $k \in \mathbb{N}$. Then (39) can be rewritten as

$$
f(\mathcal{R}_{x_k}(-t_k g_k)) + u_{k+1} \leq f(x_k) + u_k - c_1 t_k \|\mathrm{grad} f(x_k)\|^2, \quad k \geq K \tag{40}
$$

To proceed now with the proof of (i), we deduce from (40) combined with $\inf_{k \in \mathbb{N}} f(x_k) > -\infty$ and $u_k \to 0$ as $k \to \infty$ that

$$
\begin{aligned}
c_1 \sum_{k=K}^{\infty} t_k \|\mathrm{grad} f(x_k)\|^2 &\leq \sum_{k=K}^{\infty} (f(x_k) - f(\mathcal{R}_{x_k}(-t_k g_k)) + u_k - u_{k+1}) \\
&\leq f(x_K) - \inf_{k \in \mathbb{N}} f(x_k) + u_K < \infty
\end{aligned}
$$

Next we employ Lemma 3 with $\alpha_k := \|\mathrm{grad} f(x_k)\|$, $\beta_k := L t_k$, and $\gamma_k := L t_k \epsilon_k$ for all $k \in \mathbb{N}$ to derive $\mathrm{grad} f(x_k) \to 0$. Observe first that condition (10) is satisfied due to the the estimates

$$
\begin{aligned}
\alpha_{k+1} - \alpha_k &= \|\mathrm{grad} f(x_{k+1})\| - \|\mathrm{grad} f(x_k)\| \leq \|\mathcal{T}_{x_{k+1}}^{x_k} \mathrm{grad} f(x_{k+1}) - \mathrm{grad} f(x_k)\| \\
&= L t_k \|g_k\| \leq L t_k (\|\mathrm{grad} f(x_k)\| + \|g_k - \mathrm{grad} f(x_k)\|) \\
&\leq L t_k (\|\mathrm{grad} f(x_k)\| + \epsilon_k) = \beta_k \alpha_k + \gamma_k \ \text{ for all } \ k \in \mathbb{N}
\end{aligned}
$$

Further, the conditions in (11) hold by (14) and $\sum_{k=1}^{\infty} t_k \|\mathrm{grad} f(x_k)\|^2 < \infty$. As all the assumptions (10), (11) are satisfied, Lemma 3 tells us that $\|\mathrm{grad} f(x_k)\| = \alpha_k \to 0$ as $k \to \infty$.

To verify (ii), deduce from (40) that $\{f(x_k) + u_k\}$ is nonincreasing. Since $\inf_{k \in \mathbb{N}} f(x_k) > -\infty$ and $u_k \to 0$, it follows that $\{f(x_k) + u_k\}$ is bounded from below, and thus is convergent. Taking into account that $u_k \to 0$, it follows that $f(x_k)$ is convergent as well. Since $\overline{x}$ is an accumulation point of $\{x_k\}$, the continuity of f tells us that $f(\overline{x})$ is also an accumulation point of $\{f(x_k)\}$, which immediately yields $f(x_k) \to f(\overline{x})$ due to the convergence of $\{f(x_k)\}$. □

Proof of Theorem 2. First note that the global convergence result in (i) of Theorem 1 implies that every point in $\mathcal{S}$ is a stationary point. Since $\mathrm{grad}f(x_k) \to 0$, there exists a $\delta > 0$ such that $\|\mathrm{grad}f(x_k)\| \leq \delta$ for all k. Thus, the application of Lemma 2 implies that

$$\begin{aligned}\|x_{k+1} - x_k\| = \|\mathcal{R}_{x_k}(t_k g_k) - x_k\| &\leq \kappa t_k \|g_k\| \\ &\leq \kappa t_k(\|g_k - \mathrm{grad}f(x_k)\| + \|\mathrm{grad}f(x_k)\|) \\ &\leq \kappa t_k(\epsilon_k + \|\mathrm{grad}f(x_k)\|) \to 0 \end{aligned} \tag{41}$$

Then by [BST14], we konw that $\mathcal{S}$ is a compect set. Morever, since $f(x_k)$ is convergent, thus f has the same value at all the points in $\mathcal{S}$. Therefore, by Lemma 1, there exists a single desingularizing function satisfying Assumption 4, denote φ, for the Riemannian KL property of f to hold at all the points in $\mathcal{S}$.

In the case when $f(x_k) > f(\overline{x})$ for all k, since $f(x_k) \to f(\overline{x})$, $\inf_{\overline{x}\in\mathcal{S}} \|x_k - \overline{x}\| \to 0$, by the Riemannian KL property of f on $\mathcal{S}$, there exists an $l > 0$ such that

$$\varphi'(f(x_k) - f(\overline{x}))\|\mathrm{grad}f(x_k)\| \geq 1 \ \text{ for all } \ k \geq l \tag{42}$$

Define $\Delta_{p,q} := \varphi(f(x_p) - f(\overline{x}) + u_p) - \varphi(f(x_q) - f(\overline{x}) + u_q)$ for all $p, q \in \mathbb{N}$, and combining this with $u_k > 0$ and $f(x_k) - f(\overline{x}) > 0$ gives us

$$\Delta_{k,k+1} \geq \varphi'(f(x_k) - f(\overline{x}) + u_k)(f(x_k) + u_k - f(x_{k+1}) - u_{k+1}) \tag{43a}$$

$$\geq \frac{C}{(\varphi'(f(x_k) - f(\overline{x})))^{-1} + (\varphi'(u_k))^{-1}} c_1 t_k \|\mathrm{grad}f(x_k)\|^2 \tag{43b}$$

$$\geq \frac{C}{\|\mathrm{grad}f(x_k)\| + (\varphi'(u_k))^{-1}} c_1 t_k \|\mathrm{grad}f(x_k)\|^2 \tag{43c}$$

where (43a) follows from the concavity of φ, (43b) follows from (40) and Assumption 4, and (43c) follows from (42). Taking the square root of both sides in (43c) and employing the AM-GM inequality yield

$$t_k\|\mathrm{grad}f(x_k)\| = \sqrt{t_k} \cdot \sqrt{t_k\|\mathrm{grad}f(x_k)\|^2} \tag{44}$$

$$\leq \sqrt{\frac{1}{Cc_1}(\Delta_{k,k+1}) t_k(\|\mathrm{grad}f(x_k)\| + (\varphi'(u_k))^{-1})}$$

$$\leq \frac{1}{2Cc_1}(\Delta_{k,k+1}) + \frac{1}{2} t_k\Big((\varphi'(u_k))^{-1} + \|\mathrm{grad}f(x_k)\|\Big) \tag{45}$$

Using the nonincreasing property of φ' due to the concavity of φ and the choice of $c_2 \in (0, 1)$ ensures that

$$\Big(\varphi'(u_k)\Big)^{-1} = \Big(\varphi'(c_2 \sum_{i=k}^{\infty} t_i\epsilon_i)\Big)^{-1} \leq \Big(\varphi'(\sum_{i=k}^{\infty} t_i\epsilon_i)\Big)^{-1}$$

Rearranging terms and taking the sum over $i = l+1, \ldots, k$ of inequality (44) gives us

$$\begin{aligned}\sum_{i=l+1}^{k} t_i\|\mathrm{grad}f(x_i)\| &\leq \frac{1}{Cc_1}\sum_{i=l+1}^{k}(\Delta_{i,i+1}) + \sum_{i=l+1}^{k} t_i\Big(\varphi'(u_i)\Big)^{-1} \\ &= \frac{1}{Cc_1}(\Delta_{l+1,k+1}) + \sum_{i=l+1}^{k} t_i\Big(\varphi'(c_2\sum_{i=k}^{\infty} t_i\epsilon_i)\Big)^{-1} \\ &\leq \frac{1}{Cc_1}\Delta_{l+1} + \sum_{i=l+1}^{k} t_i\Big(\varphi'(\sum_{i=k}^{\infty} t_i\epsilon_i)\Big)^{-1}\end{aligned}$$

Taking the number C from Assumption 4, remembering that $\overline{x}$ is an accumulation point of $\{x_k\}$, and using $f(x_k) + u_k \downarrow f(\overline{x})$, $\Delta_k \downarrow 0$ as $k \to \infty$ together with condition (15), which yields $\sum_{k=1}^{\infty} t_k \|\mathrm{grad} f(x_k)\| < \infty$. Inserting this into (41) gives

$$\sum_{k=1}^{\infty} \|x_{k+1} - x_k\| \leq \kappa \sum_{k=1}^{\infty} t_k \epsilon_k + \kappa \sum_{k=1}^{\infty} t_k \|\mathrm{grad} f(x_k)\| < \infty$$

which justifies the convergence of $\{x_k\}$ and thus completes the proof of the theorem. □

3.2.2 Proof of Lemma 6

Proof of Lemma 6. Using $\|\mathrm{grad} f(x_k) - g_k\| \leq \nu \|\mathrm{grad} f(x_k)\|$ gives us the estimates

$$\begin{aligned} \|g_k\|^2 &= \|\mathrm{grad} f(x_k) - g_k\|^2 - \|\mathrm{grad} f(x_k)\|^2 + 2\langle \mathrm{grad} f(x_k), g_k \rangle \\ &\leq \nu^2 \|\mathrm{grad} f(x_k)\|^2 - \|\mathrm{grad} f(x_k)\|^2 + 2\langle \mathrm{grad} f(x_k), g_k \rangle \\ &= -(1-\nu^2)\|\mathrm{grad} f(x_k)\|^2 + 2\langle \mathrm{grad} f(x_k), g_k \rangle \end{aligned} \tag{46}$$

$$\begin{aligned} \langle \mathrm{grad} f(x_k), g_k \rangle &= \langle \mathrm{grad} f(x_k), g_k - \mathrm{grad} f(x_k) \rangle + \|\mathrm{grad} f(x_k)\|^2 \\ &\leq \|\mathrm{grad} f(x_k)\| \cdot \|g_k - \mathrm{grad} f(x_k)\| + \|\mathrm{grad} f(x_k)\|^2 \\ &\leq (1+\nu)\|\mathrm{grad} f(x_k)\|^2 \end{aligned} \tag{47}$$

$$\begin{aligned} -\langle \mathrm{grad} f(x_k), g_k \rangle &= -\langle \mathrm{grad} f(x_k), g_k - \mathrm{grad} f(x_k) \rangle - \|\mathrm{grad} f(x_k)\|^2 \\ &\leq \|\mathrm{grad} f(x_k)\| \cdot \|g_k - \mathrm{grad} f(x_k)\| - \|\mathrm{grad} f(x_k)\|^2 \\ &\leq -(1-\nu)\|\mathrm{grad} f(x_k)\|^2 \end{aligned} \tag{48}$$

$$\|\mathrm{grad} f(x_k)\| - \|g_k - \mathrm{grad} f(x_k)\| \leq \|g_k\| \leq \|\mathrm{grad} f(x_k)\| + \|g_k - \mathrm{grad} f(x_k)\| \tag{49}$$

which in turn implies that

$$(1-\nu)\|\mathrm{grad} f(x_k)\| \leq \|g_k\| \leq (1+\nu)\|\mathrm{grad} f(x_k)\| \text{ for all } k \in \mathbb{N} \tag{50}$$

□

4 Advancing Min-Max Optimization: First-Order Algorithms and Geometric Insights in Riemannian Manifolds

In this section, we switch gear to study the constrained optimization problems arise throughout machine learning, in classical settings such as dimension reduction [BA11], dictionary learning [SQW16a, SQW16b], and deep neural networks [HLL$^+$18], but also in emerging problems involving decision-making and multi-agent interactions. While simple convex constraints (such as norm constraints) can be easily incorporated in standard optimization formulations, notably (proximal) gradient descent [RM15, GVGM21b, GVGM21a, ABM20, VGFL$^+$20], in a range of other applications such as matrix recovery [FRW11, CWB08], low-rank matrix factorization [HMJG21]

and generative adversarial nets [GPAM+14], the constraints are fundamentally nonconvex and are often treated via special heuristics.

Thus, a general goal is to design algorithms that systematically take account of special geometric structure of the feasible set [MGD+21, Loj63, Pol63]. A long line of work in the machine learning (ML) community has focused on understanding the geometric properties of commonly used constraints and how they affect optimization; see, e.g., [GHJY15, AG16, SH16, JGN+17, GJZ17, DJL+17, RZS+18, CB19, JNG+21]. A prominent aspect of this agenda has been the re-expression of these constraints through the lens of Riemannian manifolds. This has given rise to new algorithms [SH15, HS15] with a wide range of ML applications, inclduing online principal component analysis (PCA), the computation of Mahalanobis distance from noisy measurements [Bon13], consensus distributed algorithms for aggregation in ad-hoc wireless networks [TAV12] and maximum likelihood estimation for certain non-Gaussian (heavy- or light-tailed) distributions [Wie12].

Going beyond simple minimization problems, the robustification of many ML tasks can be formulated as min-max optimization problems. Well-known examples in this domain include adversarial machine learning [KSF17, CKK+18], optimal transport [LFH+20], and online learning [MS18, BMSS19, ABM20]. Similar to their minimization counterparts, non-convex constraints have been widely applicable to the min-max optimization as well [HRU+17, DP18, BRM+18, MLZ+19, JNJ20]. Recently there has been significant effort in proving tighter results either under more structured assumptions [TJNO19, NSH+19, LTHC20, AMLJG20, Dia20, GPDO20, LJJ20a, LJJ20b, LRLY21, OLR21, KM21], and/or obtaining last-iterate convergence guarantees [DP18, DP19, MLZ+19, ADLH19, LS19, GHP+19, MRS20, LMR+20, MOP20, LJJ20a, HA21, ALW21, COZ22] for computing min-max solutions in convex-concave settings. Nonetheless, the analysis of the iteration complexity in the general *non-convex non-concave* setting is still in its infancy [VGFP19, VGFP21]. In response, the optimization community has recently studied how to extend standard min-max optimization algorithms such as gradient descent ascent (GDA) and extragradient (EG) to the Riemannian setting. In mathematical terms, given two Riemannian manifolds $\mathcal{M}, \mathcal{N}$ and a function $f : \mathcal{M} \times \mathcal{N} \to \mathbb{R}$, the Riemannian min-max optimization (RMMO) problem becomes

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

The change of geometry from Euclidean to Riemannian poses several difficulties. Indeed, a fundamental stumbling block has been that this problem may not even have theoretically meaningful solutions. In contrast with minimization where an optimal solution in a bounded domain is always guaranteed [FGHS21], existence of such saddle points necessitates typically the application of topological fixed point theorems [Bro11, Kak41], KKM Theory [KKM29]). For the case of convex-concave f with compact sets $\mathcal{X}$ and $\mathcal{Y}$, [Sio58] generalized the celebrated theorem [Neu28] and guaranteed that a solution $(x^\star, y^\star)$ with the following property exists

$$\min_{x \in \mathcal{X}} f(x, y^\star) = f(x^\star, y^\star) = \max_{y \in \mathcal{Y}} f(x^\star, y)$$

However, at the core of the proof of this result is an ingenuous application of Helly's lemma [Hel23] for the sublevel sets of f, and, until the work of [Iva14], it has been unclear how to formulate an analogous lemma for the Riemannian geometry. As a result, until recently have extensions of the min-max theorem been established, and only for restricted manifold families [Kom88, Kri14, Par19].

[ZZS22] was the first to establish a min-max theorem for a flurry of Riemannian manifolds equipped with unique geodesics. Notice that this family is not a mathematical artifact since it encompasses many practical applications of RMMO, including Hadamard and Stiefel ones used in

PCA [LKO+22]. Intuitively, the unique geodesic between two points of a manifold is the analogue of the a linear segment between two points in convex set: For any two points $x_1, x_2 \in \mathcal{X}$, their connecting geodesic is the unique shortest path contained in $\mathcal{X}$ that connects them.

Even when the RMMO is well defined, transferring the guarantees of traditional min-max optimization algorithms like Gradient Ascent Descent (GDA) and Extra-Gradient (EG) to the Riemannian case is non-trivial. Intuitively speaking, in the Euclidean realm the main leitmotif of the last-iterate analyses the aforementioned algorithms is a proof that $\delta_t = \|x_t - x^*\|^2$ is decreasing over time. To achieve this, typically the proof correlates δ_t and δ_{t-1} via a "square expansion," namely:

$$\underbrace{\|x_{t-1} - x^*\|^2}_{\alpha^2} = \underbrace{\|x_t - x^*\|^2}_{\beta^2} + \underbrace{\|x_{t-1} - x_t\|^2}_{\gamma^2} - \underbrace{2\langle x_t - x^*, x_{t-1} - x_t\rangle}_{2\beta\gamma\cos(\widehat{A})} \tag{51}$$

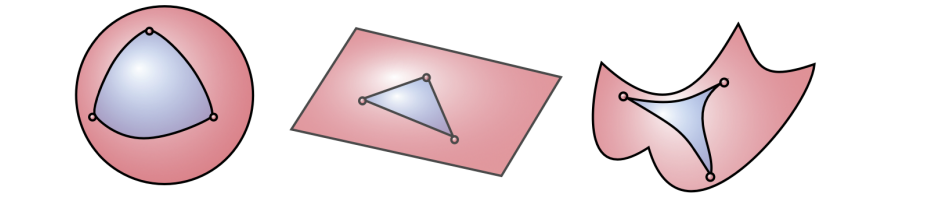
Notice, however that the above expression relies strongly on properties of Euclidean geometry (and the flatness of the corresponding line), namely that the the lines connecting the three points x_t, x_{t-1} and x^* form a triangle; indeed, it is the generalization of the Pythagorean theorem, known also as the law of cosines, for the induced triangle $(ABC) := \{(x_t, x_{t-1}, x^*)\}$. In a uniquely geodesic manifold such triangle may not belong to the manifold as discussed above. As a result, the difference of distances to the equilibrium using the geodesic paths $d^2_{\mathcal{M}}(x_t, x^*) - d^2_{\mathcal{M}}(x_{t-1}, x^*)$ generally cannot be given in a closed form. The manifold's curvature controls how close these paths are to forming a Euclidean triangle. In fact, the phenomenon of *distance distortion*, as it is typically called, was hypothesised by [ZZS22, Section 4.2] to be the cause of exponential slowdowns when applying EG to RMMO problems when compared to their Euclidean counterparts.

Multiple attempts have been made to bypass this hurdle. [HGH20] analyzed the Riemannian GDA (RGDA) for the non-convex non-concave setting. However, they do not present any last-iterate convergence results and, even in the average/best iterate setting, they only derive sub-optimal rates for the geodesic convex-concave setting due to the lack of the machinery that convex analysis and optimization offers they derive sub-optimal rates for the geodesic convex-concave case, which is the problem of our interest. The analysis of [HMJ+22] for Riemannian Hamiltonian Method (RHM), matches the rate of second-order methods in the Euclidean case. Although theoretically faster in terms of iterations, second-order methods are not preferred in practice since evaluating second order derivatives for optimization problems of thousands to millions of parameters quickly becomes prohibitive. Finally, [ZZS22] leveraged the standard averaging output trick in EG to derive a sublinear convergence rate of $O(1/\epsilon)$ for the general geodesically convex-concave Riemannian framework. In addition, they conjectured that the use of a different method could close the exponential gap for the geodesically strongly-convex-strongly-convex scenario and its Euclidean counterpart.

Given this background, a crucial question underlying the potential for successful application of first-order algorithms to Riemannian settings is the following: ***Is a performance gap necessary between Riemannian and Euclidean optimal convex-concave algorithms in terms of accuracy and the condition number?***

4.1 Preliminaries and Technical Background

We present the basic setup and optimality conditions for Riemannian min-max optimization. Indeed, we focus on some of key concepts that we need from Riemannian geometry, deferring a fuller presentation, including motivating examples and further discussion of related work, to Section B-4.3.



Positive Curvature $\kappa_{\mathcal{M}} > 0$ | Zero Curvature $\kappa_{\mathcal{M}} = 0$ | Negative Curvature $\kappa_{\mathcal{M}} < 0$

Riemannian geometry. An n-dimensional manifold $\mathcal{M}$ is a topological space where any point has a neighborhood that is homeomorphic to the n-dimensional Euclidean space. For each $x \in \mathcal{M}$, each tangent vector is tangent to all parametrized curves passing through x and the tangent space $T_x\mathcal{M}$ of a manifold $\mathcal{M}$ at this point is defined as the set of all tangent vectors. A Riemannian manifold $\mathcal{M}$ is a smooth manifold that is endowed with a smooth ("Riemannian") metric $\langle \cdot, \cdot \rangle_x$ on the tangent space $T_x\mathcal{M}$ for each point $x \in \mathcal{M}$. The inner metric induces a norm $\| \cdot \|_x$ on the tangent spaces.

A geodesic can be seen as the generalization of an Euclidean linear segment and is modeled as a smooth curve (map), $\gamma : [0, 1] \mapsto \mathcal{M}$, which is locally a distance minimizer. Additionally, because of the non-flatness of a manifold a different relation between the angles and the lengths of an arbitrary geodesic triangle is induced. This distortion can be quantified via the *sectional curvature* parameter $\kappa_{\mathcal{M}}$ thanks to Toponogov's theorem [CE75, BGP92]. A constructive consequence of this definition are the trigonometric comparison inequalities (TCIs) that will be essential in our proofs; see [AOBL20, Corollary 2.1] and [ZS16, Lemma 5] for detailed derivations. Assuming bounded sectional curvature, TCIs provide a tool for bounding Riemannian "inner products" that are more troublesome than classical Euclidean inner products.

The following proposition summarizes the TCIs that we will need; note that if $\kappa_{\min} = \kappa_{\max} = 0$ (i.e., Euclidean spaces), then the proposition reduces to the law of cosines.

Proposition 1. *Suppose that $\mathcal{M}$ is a Riemannian manifold and let Δ be a geodesic triangle in $\mathcal{M}$ with the side length a, b, c and let A be the angle between b and c. Then, we have*

(i) If $\kappa_{\mathcal{M}}$ that is upper bounded by $\kappa_{\max} > 0$ and the diameter of $\mathcal{M}$ is bounded by $\frac{\pi}{\sqrt{\kappa_{\max}}}$, then

$$a^2 \geq \underline{\xi}(\kappa_{\max}, c) \cdot b^2 + c^2 - 2bc\cos(A)$$

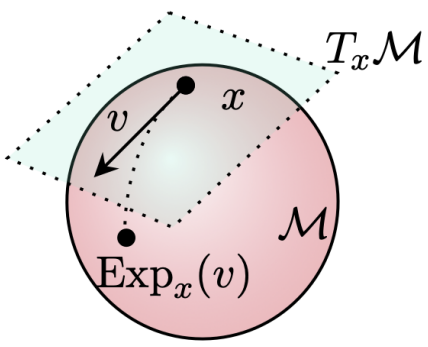
where $\underline{\xi}(\kappa, c) := 1$ for $\kappa \leq 0$ and $\underline{\xi}(\kappa, c) := c\sqrt{\kappa}\cot(c\sqrt{\kappa}) < 1$ for $\kappa > 0$.

(ii) If $\kappa_{\mathcal{M}}$ is lower bounded by $\kappa_{\min}$, then

$$a^2 \leq \overline{\xi}(\kappa_{\min}, c) \cdot b^2 + c^2 - 2bc\cos(A)$$

where $\overline{\xi}(\kappa, c) := c\sqrt{-\kappa}\coth(c\sqrt{-\kappa}) > 1$ if $\kappa < 0$ and $\overline{\xi}(\kappa, c) := 1$ if $\kappa \geq 0$.

Also, in contrast to the Euclidean case, x and $v = \mathrm{grad}_x f(x)$ do not lie in the same space, since $\mathcal{M}$ and $T_x\mathcal{M}$ respectively are distinct entities. The interplay between these dual spaces typically is carried out via the *exponential maps*. An exponential map at a point $x \in \mathcal{M}$ is a mapping from the tangent space $T_x\mathcal{M}$ to $\mathcal{M}$. In particular, $y := \mathbb{E}_x(v) \in \mathcal{M}$



is defined such that there exists a geodesic $\gamma : [0,1] \mapsto \mathcal{M}$ satisfying $\gamma(0) = x$, $\gamma(1) = y$ and $\gamma'(0) = v$. The inverse map exists since the manifold has a unique geodesic between any two points, which we denote as $\mathbb{E}_x^{-1} : \mathcal{M} \mapsto T_x\mathcal{M}$. Accordingly, we have $d_{\mathcal{M}}(x,y) = \|\mathbb{E}_x^{-1}(y)\|_x$ is the Riemannian distance induced by the exponential map.

Finally, in contrast again to Euclidean spaces, we cannot compare the tangent vectors at different points $x, y \in \mathcal{M}$ since these vectors lie in different tangent spaces. To resolve this issue, it suffices to define a transport mapping that moves a tangent vector along the geodesics and also preserves the length and Riemannian metric $\langle \cdot, \cdot \rangle_x$; indeed, we can define a parallel transport $\Gamma_x^y : T_x\mathcal{M} \mapsto T_y\mathcal{M}$ such that the inner product between any $u, v \in T_x\mathcal{M}$ is preserved; i.e., $\langle u, v \rangle_x = \langle \Gamma_x^y(u), \Gamma_x^y(v) \rangle_y$.

Riemannian min-max optimization and function classes. We let $\mathcal{M}$ and $\mathcal{N}$ be Riemannian manifolds with unique geodesic and bounded sectional curvature and assume that the function $f : \mathcal{M} \times \mathcal{N} \mapsto \mathbb{R}$ is defined on the product of these manifolds. The regularity conditions that we impose on the function f are as follows.

Definition 2. *A function $f : \mathcal{M} \times \mathcal{N} \mapsto \mathbb{R}$ is* geodesically L-Lipschitz *if for $\forall x, x' \in \mathcal{M}$ and $\forall y, y' \in \mathcal{N}$, the following statement holds true: $|f(x,y) - f(x',y')| \leq L(d_{\mathcal{M}}(x,x') + d_{\mathcal{N}}(y,y'))$. Additionally, if function f is also differentiable, it is called* geodesically ℓ-smooth *if for $\forall x, x' \in \mathcal{M}$ and $\forall y, y' \in \mathcal{N}$, the following statement holds true,*

$$\begin{aligned} \|\mathrm{grad}_x f(x,y) - \Gamma_{x'}^x \mathrm{grad}_x f(x',y')\| \leq &\quad \ell(d_{\mathcal{M}}(x,x') + d_{\mathcal{N}}(y,y')) \\ \|\mathrm{grad}_y f(x,y) - \Gamma_{y'}^y \mathrm{grad}_y f(x',y')\| \leq &\quad \ell(d_{\mathcal{M}}(x,x') + d_{\mathcal{N}}(y,y')) \end{aligned}$$

where $(\mathrm{grad}_x f(x',y'), \mathrm{grad}_y f(x',y')) \in T_{x'}\mathcal{M} \times T_{y'}\mathcal{N}$ is the Riemannian gradient of f at (x',y'), $\Gamma_{x'}^x$ is the parallel transport of $\mathcal{M}$ from x' to x, and $\Gamma_{y'}^y$ is the parallel transport of $\mathcal{N}$ from y' to y.

Definition 3. *A function $f : \mathcal{M} \times \mathcal{N} \to \mathbb{R}$ is* geodesically strongly-convex-strongly-concave *with the modulus $\mu > 0$ if the following statement holds true,*

$$\begin{aligned} f(x',y) \geq &\quad f(x,y) + \langle \mathrm{subgrad}_x f(x,y), \mathbb{E}_x^{-1}(x') \rangle_x + \frac{\mu}{2}(d_{\mathcal{M}}(x,x'))^2, \text{ for each } y \in \mathcal{N} \\ f(x,y') \leq &\quad f(x,y) + \langle \mathrm{subgrad}_y f(x,y), \mathbb{E}_y^{-1}(y') \rangle_y - \frac{\mu}{2}(d_{\mathcal{N}}(y,y'))^2, \text{ for each } x \in \mathcal{M} \end{aligned}$$

where $(\mathrm{subgrad}_x f(x',y'), \mathrm{subgrad}_y f(x',y')) \in T_{x'}\mathcal{M} \times T_{y'}\mathcal{N}$ is a Riemannian subgradient of f at a point (x',y'). A function f is geodesically convex-concave *if the above holds true with $\mu = 0$.*

Following standard conventions in Riemannian optimization [ZS16, AOBL20, ZZS22], we make the following assumptions on the manifolds and objective functions:[1]

Assumption 5. *The objective function $f : \mathcal{M} \times \mathcal{N} \mapsto \mathbb{R}$ and manifolds $\mathcal{M}$ and $\mathcal{N}$ satisfy*

(i) The diameter of the domain $\{(x,y) \in \mathcal{M} \times \mathcal{N} : -\infty < f(x,y) < +\infty\}$ is bounded by $D > 0$

(ii) $\mathcal{M}, \mathcal{N}$ admit unique geodesic paths for any $(x,y), (x',y') \in \mathcal{M} \times \mathcal{N}$

[1] In particular, our assumed upper and lower bounds $\kappa_{\min}, \kappa_{\max}$ guarantee that TCIs in Proposition 1 can be used in our analysis for proving finite-time convergence.

(iii) The sectional curvatures of $\mathcal{M}$ and $\mathcal{N}$ are both bounded in the range $[\kappa_{\min}, \kappa_{\max}]$ with $\kappa_{\min} \leq 0$. If $\kappa_{\max} > 0$, we assume that the diameter of manifolds is bounded by $\frac{\pi}{\sqrt{\kappa_{\max}}}$

Under these conditions, [ZZS22] proved an analog of Sion's minimax theorem [Sio58] in geodesic metric spaces. Formally, we have

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

which guarantees that there exists at least one global saddle point $(x^\star, y^\star) \in \mathcal{M} \times \mathcal{N}$ such that $\min_{x \in \mathcal{M}} f(x, y^\star) = f(x^\star, y^\star) = \max_{y \in \mathcal{Y}} f(x^\star, y)$. Note that the unicity of geodesics assumption is algorithm-independent and is imposed for guaranteeing that a saddle-point solution always exist. Even though this rules out many manifolds of interest, there are still many manifolds that satisfy such conditions. More specifically, the Hadamard manifold (manifolds with non-positive curvature, $\kappa_{\max} = 0$) has a unique geodesic between any two points. This also becomes a common regularity condition in Riemannian optimization [ZS16, AOBL20]. For any point $(\widehat{\mathbf{x}}, \widehat{\mathbf{y}}) \in \mathcal{M} \times \mathcal{N}$, the duality gap $f(\widehat{x}, y^\star) - f(x^\star, \widehat{y})$ thus gives an optimality criterion.

Definition 4. *A point $(\widehat{x}, \widehat{y}) \in \mathcal{M} \times \mathcal{N}$ is an* ϵ-saddle point *of a geodesically convex-concave function $f(\cdot, \cdot)$ if $f(\widehat{x}, y^\star) - f(x^\star, \widehat{y}) \leq \epsilon$ where $(x^\star, y^\star) \in \mathcal{M} \times \mathcal{N}$ is a global saddle point.*

In the setting where f is geodesically strongly-convex-strongly-concave with $\mu > 0$, it is not difficult to verify the uniqueness of a global saddle point $(x^\star, y^\star) \in \mathcal{M} \times \mathcal{N}$. Then, we can consider the distance gap $(d(\widehat{x}, x^\star))^2 + (d(\widehat{y}, y^\star))^2$ as an optimality criterion for any point $(\widehat{\mathbf{x}}, \widehat{\mathbf{y}}) \in \mathcal{M} \times \mathcal{N}$.

Definition 5. *A point $(\widehat{x}, \widehat{y}) \in \mathcal{M} \times \mathcal{N}$ is an ϵ-saddle point of a geodesically strongly-convex-strongly-concave function $f(\cdot, \cdot)$ if $(d(\widehat{x}, x^\star))^2 + (d(\widehat{y}, y^\star))^2 \leq \epsilon$, where $(x^\star, y^\star) \in \mathcal{M} \times \mathcal{N}$ is a global saddle point. If f is also geodesically ℓ-smooth, we denote $\kappa = \frac{\ell}{\mu}$ as the condition number.*

Given the above definitions, we can ask whether it is possible to find an ϵ-saddle point efficiently or not. In this context, [ZZS22] have answered this question in the affirmative for the setting where f is geodesically ℓ-smooth and geodesically convex-concave; indeed, they derive the convergence rate of Riemannian corrected extragradient (RCEG) method in terms of time-average iterates and also conjecture that *RCEG does not guarantee convergence at a linear rate in terms of last iterates when f is geodesically ℓ-smooth and geodesically strongly-convex-strongly-concave, due to the existence of distance distortion*; see [ZZS22, Section 4.2]. Surprisingly, we show in Section 4.2 that RCEG with constant stepsize can achieve last-iterate convergence at a linear rate. Moreover, we establish the optimal convergence rates of stochastic RCEG for certain choices of stepsize for both geodesically convex-concave and geodesically strongly-convex-strongly-concave settings.

4.2 Riemannian Corrected Extragradient Method

In this section, we revisit the scheme of Riemannian corrected extragradient (RCEG) method proposed by [ZZS22] and extend it to a stochastic algorithm that we refer to as *stochastic RCEG*. We present our main results on an optimal last-iterate convergence guarantee for the geodesically strongly-convex-strongly-concave setting (both deterministic and stochastic) and a time-average convergence guarantee for the geodesically convex-concave setting (stochastic). This complements the time-average convergence guarantee for geodesically convex-concave setting (deterministic) [ZZS22, Theorem 4.1] and resolves an open problem posted in [ZZS22, Section 4.2].

Algorithm 4 RCEG

Input: initial points (x_0, y_0) and stepsizes $\eta > 0$
for $t = 0, 1, 2, \ldots, T-1$ **do**
  Query $(g_x^t, g_y^t) \leftarrow (\mathrm{grad}_x f(x_t, y_t), \mathrm{grad}_y f(x_t, y_t))$, the Riemannian gradient of f at a point (x_t, y_t)
  $\widehat{x}_t \leftarrow \mathbb{E}_{x_t}(-\eta \cdot g_x^t)$
  $\widehat{y}_t \leftarrow \mathbb{E}_{y_t}(\eta \cdot g_y^t)$
  Query $(\widehat{g}_x^t, \widehat{g}_y^t) \leftarrow (\mathrm{grad}_x f(\widehat{x}_t, \widehat{y}_t), \mathrm{grad}_y f(\widehat{x}_t, \widehat{y}_t))$, the Riemannian gradient of f at a point $(\widehat{x}_t, \widehat{y}_t)$
  $x_{t+1} \leftarrow \mathbb{E}_{\widehat{x}_t}(-\eta \cdot \widehat{g}_x^t + \mathbb{E}_{\widehat{x}_t}^{-1}(x_t))$
  $y_{t+1} \leftarrow \mathbb{E}_{\widehat{y}_t}(\eta \cdot \widehat{g}_y^t + \mathbb{E}_{\widehat{y}_t}^{-1}(y_t))$
end for

Algorithm 5 SRCEG

Input: initial points (x_0, y_0) and stepsizes $\eta > 0$
for $t = 0, 1, 2, \ldots, T-1$ **do**
  Query (g_x^t, g_y^t) as a **noisy** estimator of Riemannian gradient of f at a point (x_t, y_t)
  $\widehat{x}_t \leftarrow \mathbb{E}_{x_t}(-\eta \cdot g_x^t)$
  $\widehat{y}_t \leftarrow \mathbb{E}_{y_t}(\eta \cdot g_y^t)$
  Query $(\widehat{g}_x^t, \widehat{g}_y^t)$ as a **noisy** estimator of Riemannian gradient of f at a point $(\widehat{x}_t, \widehat{y}_t)$
  $x_{t+1} \leftarrow \mathbb{E}_{\widehat{x}_t}(-\eta \cdot \widehat{g}_x^t + \mathbb{E}_{\widehat{x}_t}^{-1}(x_t))$
  $y_{t+1} \leftarrow \mathbb{E}_{\widehat{y}_t}(\eta \cdot \widehat{g}_y^t + \mathbb{E}_{\widehat{y}_t}^{-1}(y_t))$
end for

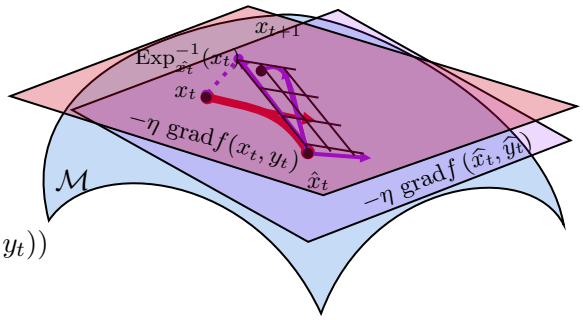
4.2.1 Algorithmic scheme

The recently proposed *Riemannian corrected extragradient* (RCEG) method [ZZS22] is a natural extension of the celebrated extragradient (EG) method to the Riemannian setting. Its scheme resembles that of EG in Euclidean spaces but employs a simple modification in the extrapolation step to accommodate the nonlinear geometry of Riemannian manifolds. Let us provide some intuition how such modifications work.

We start with a basic version of EG as follows, where $\mathcal{M}$ and $\mathcal{N}$ are classically restricted to be convex constraint sets in Euclidean spaces:

$$\begin{aligned} \widehat{x}_t &\leftarrow \Pi_{\mathcal{M}}(x_t - \eta \cdot \nabla_x f(x_t, y_t)), \quad & \widehat{y}_t &\leftarrow \Pi_{\mathcal{N}}(y_t + \eta \cdot \nabla_y f(x_t, y_t)) \\ x_{t+1} &\leftarrow \Pi_{\mathcal{M}}(x_t - \eta \cdot \nabla_x f(\widehat{x}_t, \widehat{y}_t)), \quad & y_{t+1} &\leftarrow \Pi_{\mathcal{N}}(y_t + \eta \cdot \nabla_y f(\widehat{x}_t, \widehat{y}_t)) \end{aligned} \tag{52}$$

Turning to the setting where $\mathcal{M}$ and $\mathcal{N}$ are Riemannian manifolds, the rather straightforward way to do the generalization is to replace the projection operator by the corresponding exponential map and the gradient by the corresponding Riemannian gradient. For the first line of Eq. (52), this approach works and leads to the following updates:



$$\widehat{x}_t \leftarrow \mathbb{E}_{x_t}(-\eta \cdot \mathrm{grad}_x f(x_t, y_t)), \quad \widehat{y}_t \leftarrow \mathbb{E}_{y_t}(\eta \cdot \mathrm{grad}_y f(x_t, y_t))$$

However, we encounter some issues for the second line of Eq. (52): The aforementioned approach leads to some problematic updates, $x_{t+1} \leftarrow \mathbb{E}_{x_t}(-\eta \cdot \mathrm{grad}_x f(\widehat{x}_t, \widehat{y}_t))$ and $y_{t+1} \leftarrow \mathbb{E}_{y_t}(\eta \cdot \mathrm{grad}_y f(\widehat{x}_t, \widehat{y}_t))$; indeed, the exponential maps $\mathbb{E}_{x_t}(\cdot)$ and $\mathbb{E}_{y_t}(\cdot)$ are defined from $T_{x_t}\mathcal{M}$ to $\mathcal{M}$ and from $T_{y_t}\mathcal{N}$ to $\mathcal{N}$ respectively. However, we have $-\mathrm{grad}_x f(\widehat{x}_t, \widehat{y}_t) \in T_{\widehat{x}_t}\mathcal{M}$ and $\mathrm{grad}_y f(\widehat{x}_t, \widehat{y}_t) \in T_{\widehat{y}_t}\mathcal{N}$. This motivates us to reformulate the second line of Eq. (52) as follows:

$$x_{t+1} \leftarrow \Pi_{\mathcal{M}}(\widehat{x}_t - \eta \cdot \nabla_x f(\widehat{x}_t, \widehat{y}_t) + (x_t - \widehat{x}_t)), \quad y_{t+1} \leftarrow \Pi_{\mathcal{N}}(\widehat{y}_t + \eta \cdot \nabla_y f(\widehat{x}_t, \widehat{y}_t) + (y_t - \widehat{y}_t))$$

In the general setting of Riemannian manifolds, the terms $x_t - \widehat{x}_t$ and $y_t - \widehat{y}_t$ become $\mathbb{E}_{\widehat{x}_t}^{-1}(x_t) \in T_{\widehat{x}_t}\mathcal{M}$ and $\mathbb{E}_{\widehat{y}_t}^{-1}(y_t) \in T_{\widehat{y}_t}\mathcal{N}$. This observation yields the following updates:

$$x_{t+1} \leftarrow \mathbb{E}_{\widehat{x}_t}(-\eta \cdot \mathrm{grad}_x f(\widehat{x}_t, \widehat{y}_t) + \mathbb{E}_{\widehat{x}_t}^{-1}(x_t)), \quad \widehat{y}_t \leftarrow \mathbb{E}_{\widehat{y}_t}(\eta \cdot \mathrm{grad}_y f(\widehat{x}_t, \widehat{y}_t) + \mathbb{E}_{\widehat{y}_t}^{-1}(y_t))$$

We summarize the resulting RCEG method in Algorithm 4 and present the stochastic extension with noisy estimators of Riemannian gradients of f in Algorithm 5.

4.2.2 Main results

We present our main results on global convergence for Algorithms 4 and 5. To simplify the presentation, we treat separately the following two cases:

Assumption 6. *The objective function f is geodesically ℓ-smooth and geodesically strongly-convex-strongly-concave with $\mu > 0$.*

Assumption 7. *The objective function f is geodesically ℓ-smooth and geodesically convex-concave.*

Letting $(x^\star, y^\star) \in \mathcal{M} \times \mathcal{N}$ be a global saddle point of f (which exists under either Assumption 6 or 7), we let $D_0 = (d_{\mathcal{M}}(x_0, x^\star))^2 + (d_{\mathcal{N}}(y_0, y^\star))^2 > 0$ and $\kappa = \ell/\mu$ for geodesically strongly-convex-strongly-concave setting. For simplicity of presentation, we also define a ratio $\tau(\cdot, \cdot)$ that measures how non-flatness changes in the spaces: $\tau([\kappa_{\min}, \kappa_{\max}], c) = \frac{\overline{\xi}(\kappa_{\min}, c)}{\underline{\xi}(\kappa_{\max}, c)} \geq 1$. We summarize our results for Algorithm 4 in the following theorem.

Theorem 6. *Given Assumptions 5 and 6, and letting $\eta = \min\{1/(2\ell\sqrt{\tau_0}), \underline{\xi}_0/(2\mu)\}$, there exists some $T > 0$ such that the output of Algorithm 4 satisfies that $(d(x_T, x^\star))^2 + (d(y_T, y^\star))^2 \leq \epsilon$ (i.e., an ϵ-saddle point of f in Definition 5) and the total number of Riemannian gradient evaluations is bounded by*

$$O\left(\left(\kappa\sqrt{\tau_0} + \frac{1}{\underline{\xi}_0}\right)\log\left(\frac{D_0}{\epsilon}\right)\right)$$

where $\tau_0 = \tau([\kappa_{\min}, \kappa_{\max}], D) \geq 1$ measures how non-flatness changes in $\mathcal{M}$ and $\mathcal{N}$ and $\underline{\xi}_0 = \underline{\xi}(\kappa_{\max}, D) \leq 1$ is properly defined in Proposition 1.

Remark. Theorem 6 illustrates the last-iterate convergence of Algorithm 4 for solving geodesically strongly-convex-strongly-concave problems, thereby resolving an open problem delineated by [ZZS22]. Further, the dependence on κ and $1/\epsilon$ cannot be improved since it matches the lower bound established for min-max optimization problems in Euclidean spaces [ZHZ21]. However, we believe that the dependence on τ_0 and $\underline{\xi}_0$ is not tight, and it is of interest to either improve the rate or establish a lower bound for general Riemannian min-max optimization.

Remark. The current theoretical analysis covers local geodesic strong-convex-strong-concave settings. The key ingredient is how to define the local region; indeed, if we say the set of $\{(x, y) : d_{\mathcal{M}}(x, x^\star) \leq \delta, d_{\mathcal{N}}(y_t, y^\star) \leq \delta\}$ is a local region where the function is geodesic strong-convex-strong-concave. Then, the set of $\{(x, y) : (d_{\mathcal{M}}(x, x^\star)^2 + d_{\mathcal{N}}(y_t, y^\star)^2) \leq \delta^2\}$ must be contained in the above local region and the objective function is also geodesic strong-convex-strong-concave. If $(x_0, y_0) \in \{(x, y) : (d_{\mathcal{M}}(x, x^\star)^2 + d_{\mathcal{N}}(y_t, y^\star)^2) \leq \delta^2\}$, our theoretical analysis guarantees the last-iterate linear convergence rate. Such argument and definition of local region were standard for min-max optimization in the Euclidean setting; see [LS19, Assumption 2.1]. For an important optimization problem that is globally geodesically strongly-convex-strongly-concave, we refer to Section A where *Robust matrix Karcher mean problem* is indeed the desired one.

In the scheme of SRECG, we highlight that (g_x^t, g_y^t) and $(\widehat{g}_x^t, \widehat{g}_y^t)$ are noisy estimators of Riemannian gradients of f at (x_t, y_t) and $(\widehat{x}_t, \widehat{y}_t)$. It is necessary to impose the conditions such that

these estimators are unbiased and has bounded variance. By abuse of notation, we assume that

$$\begin{aligned} g_x^t &= \mathrm{grad}_x f(x_t, y_t) + \xi_x^t \qquad g_y^t = \mathrm{grad}_y f(x_t, y_t) + \xi_y^t \\ \widehat{g}_x^t &= \mathrm{grad}_x f(\widehat{x}_t, \widehat{y}_t) + \widehat{\xi}_x^t \qquad \widehat{g}_y^t = \mathrm{grad}_y f(\widehat{x}_t, \widehat{y}_t) + \widehat{\xi}_y^t \end{aligned} \tag{53}$$

where the noises (ξ_x^t, ξ_y^t) and $(\widehat{\xi}_x^t, \widehat{\xi}_y^t)$ are independent and satisfy that

$$\begin{aligned} &\mathbb{E}[\xi_x^t] = 0, \quad \mathbb{E}[\xi_y^t] = 0, \quad \mathbb{E}[\|\xi_x^t\|^2 + \|\xi_y^t\|^2] \leq \sigma^2 \\ &\mathbb{E}[\widehat{\xi}_x^t] = 0, \quad \mathbb{E}[\widehat{\xi}_y^t] = 0, \quad \mathbb{E}[\|\widehat{\xi}_x^t\|^2 + \|\widehat{\xi}_y^t\|^2] \leq \sigma^2 \end{aligned} \tag{54}$$

We are ready to summarize our results for Algorithm 5 in the following theorems.

Theorem 7. *Given Assumptions 5 and 6, letting Eq.* (53) *and Eq.* (54) *hold with $\sigma > 0$ and letting $\eta > 0$ satisfy $\eta = \min\{\frac{1}{24\ell\sqrt{\tau_0}}, \frac{\underline{\xi}_0}{2\mu}, \frac{2(\log(T)+\log(\mu^2 D_0 \sigma^{-2}))}{\mu T}\}$, there exists some $T > 0$ such that the output of Algorithm 5 satisfies that $\mathbb{E}[(d(x_T, x^\star))^2 + (d(y_T, y^\star))^2] \leq \epsilon$ and the total number of noisy Riemannian gradient evaluations is bounded by*

$$O\left(\left(\kappa\sqrt{\tau_0} + \frac{1}{\underline{\xi}_0}\right)\log\left(\frac{D_0}{\epsilon}\right) + \frac{\sigma^2\overline{\xi}_0}{\mu^2\epsilon}\log\left(\frac{1}{\epsilon}\right)\right)$$

where $\tau_0 = \tau([\kappa_{\min}, \kappa_{\max}], D) \geq 1$ measures how non-flatness changes in $\mathcal{M}$ and $\mathcal{N}$ and $\underline{\xi}_0 = \underline{\xi}(\kappa_{\max}, D) \leq 1$ is properly defined in Proposition 1.

Theorem 8. *Given Assumptions 5 and 7 and assume that Eq.* (53) *and Eq.* (54) *hold with $\sigma > 0$ and let $\eta > 0$ satisfies that $\eta = \min\{\frac{1}{4\ell\sqrt{\tau_0}}, \frac{1}{\sigma}\sqrt{\frac{D_0}{\overline{\xi}_0 T}}\}$, there exists some $T > 0$ such that the output of Algorithm 5 satisfies that $\mathbb{E}[f(\bar{x}_T, y^\star) - f(x^\star, \bar{y}_T)] \leq \epsilon$ and the total number of noisy Riemannian gradient evaluations is bounded by*

$$O\left(\frac{\ell D_0\sqrt{\tau_0}}{\epsilon} + \frac{\sigma^2\overline{\xi}_0}{\epsilon^2}\right)$$

where $\tau_0 = \tau([\kappa_{\min}, \kappa_{\max}], D)$ measures how non-flatness changes in $\mathcal{M}$ and $\mathcal{N}$ and $\overline{\xi}_0 = \overline{\xi}(\kappa_{\min}, D) \geq 1$ is properly defined in Proposition 1. The time-average iterates $(\bar{x}_T, \bar{y}_T) \in \mathcal{M} \times \mathcal{N}$ can be computed by $(\bar{x}_0, \bar{y}_0) = (0, 0)$ and the inductive formula: $\bar{x}_{t+1} = \mathbb{E}_{\bar{x}_t}(\frac{1}{t+1} \cdot \mathbb{E}_{\bar{x}_t}^{-1}(\widehat{x}_t))$ and $\bar{y}_{t+1} = \mathbb{E}_{\bar{y}_t}(\frac{1}{t+1} \cdot \mathbb{E}_{\bar{y}_t}^{-1}(\widehat{y}_t))$ for all $t = 0, 1, \ldots, T-1$.

Remark. Theorem 7 presents the last-iterate convergence rate of Algorithm 5 for solving geodesically strongly-convex-strongly-concave problems while Theorem 8 gives the time-average convergence rate when the function f is only assumed to be geodesically convex-concave. Note that we carefully choose the stepsizes such that our upper bounds match the lower bounds established for stochastic min-max optimization problems in Euclidean spaces [JNT11, FOP20, KLL22], in terms of the dependence on κ, $1/\epsilon$ and σ^2, up to log factors.

Discussions: The last-iterate linear convergence rate in terms of Riemannian metrics is only limited to geodesically strongly convex-concave cases but other results, e.g., the average-iterate sublinear convergence rate, are derived under more mild conditions. This is consistent with classical results in the Euclidean setting where geodesic convexity reduces to convexity; indeed, the last-iterate linear convergence rate in terms of squared Euclidean norm is only obtained for strongly

convex-concave cases. As such, our setting is not restrictive. Moreover, [ZZS22] showed that the existence of a global saddle point is only guaranteed under the geodesically convex-concave assumption. For geodesically nonconvex-concave or geodesically nonconvex-nonconcave cases, a global saddle point might not exist and new optimality notions are required before algorithmic design. This question remains open in the Euclidean setting and is beyond the scope of this paper. However, we remark that an interesting class of robustification problems are nonconvex-nonconcave min-max problems in the Euclidean setting can be geodesically convex-concave in the Riemannian setting; see Section A.

4.3 Metric Geometry

To generalize the first-order methods in Euclidean setting, we introduce several basic concepts in metric geometry [BBB$^+$01], which are known to include both Euclidean spaces and Riemannian manifolds as special cases. Formally, we have

Definition 6 (Metric Space)**.** *A metric space (X, d) is a pair of a set X and a distance function $d(\cdot, \cdot)$ satisfying: (i) $d(x, x') \geq 0$ for any $x, x' \in X$; (ii) $d(x, x') = d(x', x)$ for any $x, x' \in X$; and (iii) $d(x, x'') \leq d(x, x') + d(x', x'')$ for any $x, x', x'' \in X$. In other words, the distance function $d(\cdot, \cdot)$ is non-negative, symmetrical and satisfies the triangle inequality.*

A *path* $\gamma : [0, 1] \mapsto X$ is a continuous mapping from the interval $[0, 1]$ to X and the *length* of γ is defined as $\text{length}(\gamma) := \lim_{n\to+\infty} \sup_{0=t_0<\ldots<t_n=1} \sum_{i=1}^n d(\gamma(t_{i-1}), \gamma(t_i))$. Note that the triangle inequality implies that $\sup_{0=t_0<\ldots<t_n=1} \sum_{i=1}^n d(\gamma(t_{i-1}), \gamma(t_i))$ is nondecreasing. Then, the length of a path γ is well defined since the limit is either $+\infty$ or a finite scalar. Moreover, for $\forall \epsilon > 0$, there exists $n \in \mathbb{N}$ and the partition $0 = t_0 < \ldots < t_n = 1$ of the interval $[0, 1]$ such that $\text{length}(\gamma) \leq \sum_{i=1}^n d(\gamma(t_{i-1}), \gamma(t_i)) + \epsilon$.

Definition 7 (Length Space)**.** *A metric space (X, d) is a length space if, for any $x, x' \in X$ and $\epsilon > 0$, there exists a path $\gamma : [0, 1] \mapsto X$ connecting x and x' such that* $\text{length}(\gamma) \leq d(x, x') + \epsilon$.

We can see from Definition 7 that a set of length spaces is strict subclass of metric spaces; indeed, for some $x, x' \in X$, there does not exist a path γ such that its length can be approximated by $d(x, x')$ for some tolerance $\epsilon > 0$. In metric geometry, a *geodesic* is a path which is locally a distance minimizer everywhere. More precisely, a path γ is a geodesic if there is a constant $\nu > 0$ such that for any $t \in [0, 1]$ there is a neighborhood I of $[0, 1]$ such that,

$$d(\gamma(t_1), \gamma(t_2)) = \nu|t_1 - t_2|, \quad \text{for any } t_1, t_2 \in I$$

Note that the above generalizes the notion of geodesic for Riemannian manifolds. Then, we are ready to introduce the geodesic space and uniquely geodesic space [Bac14].

Definition 8. *A metric space (X, d) is a geodesic space if, for any $x, x' \in X$, there exists a geodesic $\gamma : [0, 1] \mapsto X$ connecting x and x'. Furthermore, it is called uniquely geodesic if the geodesic connecting x and x' is unique for any $x, x' \in X$.*

Trigonometric geometry in nonlinear spaces is intrinsically different from Euclidean space. In particular, we remark that the law of cosines in Euclidean space (with $\|\cdot\|$ as ℓ_2-norm) is crucial for analyzing the convergence property of optimization algorithms, e.g.,

$$\|a\|^2 = \|b\|^2 + \|c\|^2 - 2bc\cos(A)$$

where a, b, c are sides of *a geodesic triangle* in Euclidean space and A is the angle between b and c. However, such nice property does not hold for nonlinear spaces due to the lack of flat geometry, further motivating us to extend the law of cosines under nonlinear trigonometric geometry. That is to say, given a geodesic triangle in X with sides a, b, c where A is the angle between b and c, we hope to establish the relationship between a^2, b^2, c^2 and $2bc\cos(A)$ in nonlinear spaces; see the main context for the comparing inequalities.

Finally, we specify the definition of *section curvature* of Riemannian manifolds and clarify how such quantity affects the trigonometric comparison inequalities. More specifically, the sectional curvature is defined as the Gauss curvature of a 2-dimensional sub-manifold that are obtained from the image of a two-dimensional subspace of a tangent space after exponential mapping. It is worth mentioning that the above 2-dimensional sub-manifold is locally isometric to a 2-dimensional sphere, a Euclidean plane, and a hyperbolic plane with the same Gauss curvature if its sectional curvature is positive, zero and negative respectively. Then we are ready to summarize the existing trigonometric comparison inequalities for Riemannian manifold with bounded sectional curvatures. Note that the following two propositions are the full version of Proposition 1 and will be used in our subsequent proofs.

Proposition 2. *Suppose that $\mathcal{M}$ is a Riemannian manifold with sectional curvature that is upper bounded by $\kappa_{\max}$ and let Δ be a geodesic triangle in $\mathcal{M}$ with the side length a, b, c and A which is the angle between b and c. If $\kappa_{\max} > 0$, we assume the diameter of $\mathcal{M}$ is bounded by $\frac{\pi}{\sqrt{\kappa_{\max}}}$. Then, we have*

$$a^2 \geq \underline{\xi}(\kappa_{\max}, c) \cdot b^2 + c^2 - 2bc\cos(A)$$

where $\underline{\xi}(\kappa, c) := 1$ for $\kappa \leq 0$ and $\underline{\xi}(\kappa, c) := c\sqrt{\kappa}\cot(c\sqrt{\kappa}) < 1$ for $\kappa > 0$.

Proposition 3. *Suppose that $\mathcal{M}$ is a Riemannian manifold with sectional curvature that is lower bounded by $\kappa_{\min}$ and let Δ be a geodesic triangle in $\mathcal{M}$ with the side length a, b, c and A which is the angle between b and c. Then, we have*

$$a^2 \leq \overline{\xi}(\kappa_{\min}, c) \cdot b^2 + c^2 - 2bc\cos(A)$$

where $\overline{\xi}(\kappa, c) := c\sqrt{-\kappa}\coth(c\sqrt{-\kappa}) > 1$ if $\kappa < 0$ and $\overline{\xi}(\kappa, c) := 1$ if $\kappa \geq 0$.

Remark. Proposition 2 and 3 are simply the restatement of [AOBL20, Corollary 2.1] and [ZS16, Lemma 5]. The former inequality is obtained when the sectional curvature is bounded from above while the latter inequality characterizes the relationship between the trigonometric lengths when the sectional curvature is bounded from below. If $\kappa_{\min} = \kappa_{\max} = 0$ (i.e., Euclidean spaces), we have $\overline{\xi}(\kappa_{\min}, c) = \underline{\xi}(\kappa_{\max}, c) = 1$. The proof is based on Toponogov's theorem and Riccati comparison estimate [Pet06, Proposition 25] and we refer the interested readers to [ZS16] and [AOBL20] for the details.

4.4 Missing Proofs for Riemannian Corrected Extragradient Method

In this section, we present some technical lemmas for analyzing the convergence property of Algorithm 4 and 5. We also give the proofs of Theorem 6, 7 and 8.

4.4.1 Technical lemmas

We provide two technical lemmas for analyzing Algorithm 4 and 5 respectively. Parts of the first lemma were presented in [ZZS22, Lemma C.1]. For the completeness, we provide the proof details.

Lemma 7. *Under Assumption 6 and let* $\{(x_t, y_t), (\widehat{x}_t, \widehat{y}_t)\}_{t=0}^{T-1}$ *be generated by Algorithm 4 with the stepsize* $\eta > 0$*. Then, we have*

$$\begin{aligned}0 \leq & \frac{1}{2}\left((d_{\mathcal{M}}(x_t, x^\star))^2 - (d_{\mathcal{M}}(x_{t+1}, x^\star))^2 + (d_{\mathcal{N}}(y_t, y^\star))^2 - (d_{\mathcal{N}}(y_{t+1}, y^\star))^2\right) \\ & + 2\overline{\xi}_0 \eta^2 \ell^2 ((d_{\mathcal{M}}(\widehat{x}_t, x_t))^2 + (d_{\mathcal{N}}(\widehat{y}_t, y_t))^2 - \frac{1}{2}\underline{\xi}_0 \left((d_{\mathcal{M}}(\widehat{x}_t, x_t))^2 + (d_{\mathcal{N}}(\widehat{y}_t, y_t))^2\right) \\ & - \frac{\mu\eta}{2}\left((d_{\mathcal{M}}(\widehat{x}_t, x^\star))^2 + (d_{\mathcal{N}}(\widehat{y}_t, y^\star))^2\right)\end{aligned}$$

where $(x^\star, y^\star) \in \mathcal{M} \times \mathcal{N}$ *is a global saddle point of* f*.*

Proof. Since f is geodesically ℓ-smooth, we have the Riemannian gradients of f, i.e., $(\mathrm{grad}_x f, \mathrm{grad}_y f)$, are well defined. Since f is geodesically strongly-concave-strongly-concave with the modulus $\mu \geq 0$ (here $\mu = 0$ means that f is geodesically concave-concave), we have

$$\begin{aligned}& f(\widehat{x}_t, y^\star) - f(x^\star, \widehat{y}_t) = f(\widehat{x}_t, \widehat{y}_t) - f(x^\star, \widehat{y}_t) - (f(\widehat{x}_t, \widehat{y}_t) - f(\widehat{x}_t, y^\star)) \\ & \overset{\text{Definition 3}}{\leq} -\langle \mathrm{grad}_x f(\widehat{x}_t, \widehat{y}_t), \mathbb{E}_{\widehat{x}_t}^{-1}(x^\star)\rangle + \langle \mathrm{grad}_y f(\widehat{x}_t, \widehat{y}_t), \mathbb{E}_{\widehat{y}_t}^{-1}(y^\star)\rangle - \frac{\mu}{2}(d_{\mathcal{M}}(\widehat{x}_t, x^\star))^2 - \frac{\mu}{2}(d_{\mathcal{N}}(\widehat{y}_t, y^\star))^2\end{aligned}$$

Since $(x^\star, y^\star) \in \mathcal{M} \times \mathcal{N}$ is a global saddle point of f, we have $f(\widehat{x}_t, y^\star) - f(x^\star, \widehat{y}_t) \geq 0$. Recalling also from the scheme of Algorithm 4 that we have

$$\begin{aligned}x_{t+1} &\leftarrow \mathbb{E}_{\widehat{x}_t}(-\eta \cdot \mathrm{grad}_x f(\widehat{x}_t, \widehat{y}_t) + \mathbb{E}_{\widehat{x}_t}^{-1}(x_t)) \\ y_{t+1} &\leftarrow \mathbb{E}_{\widehat{y}_t}(\eta \cdot \mathrm{grad}_y f(\widehat{x}_t, \widehat{y}_t) + \mathbb{E}_{\widehat{y}_t}^{-1}(y_t))\end{aligned}$$

By the definition of an exponential map, we have

$$\begin{aligned}\mathbb{E}_{\widehat{x}_t}^{-1}(x_{t+1}) &= -\eta \cdot \mathrm{grad}_x f(\widehat{x}_t, \widehat{y}_t) + \mathbb{E}_{\widehat{x}_t}^{-1}(x_t) \\ \mathbb{E}_{\widehat{y}_t}^{-1}(y_{t+1}) &= \eta \cdot \mathrm{grad}_y f(\widehat{x}_t, \widehat{y}_t) + \mathbb{E}_{\widehat{y}_t}^{-1}(y_t)\end{aligned} \tag{55}$$

This implies that

$$\begin{aligned}-\langle \mathrm{grad}_x f(\widehat{x}_t, \widehat{y}_t), \mathbb{E}_{\widehat{x}_t}^{-1}(x^\star)\rangle &= \frac{1}{\eta}(\langle \mathbb{E}_{\widehat{x}_t}^{-1}(x_{t+1}), \mathbb{E}_{\widehat{x}_t}^{-1}(x^\star)\rangle - \langle \mathbb{E}_{\widehat{x}_t}^{-1}(x_t), \mathbb{E}_{\widehat{x}_t}^{-1}(x^\star)\rangle) \\ \langle \mathrm{grad}_y f(\widehat{x}_t, \widehat{y}_t), \mathbb{E}_{\widehat{y}_t}^{-1}(y^\star)\rangle &= \frac{1}{\eta}(\langle \mathbb{E}_{\widehat{y}_t}^{-1}(y_{t+1}), \mathbb{E}_{\widehat{y}_t}^{-1}(y^\star)\rangle - \langle \mathbb{E}_{\widehat{y}_t}^{-1}(y_t), \mathbb{E}_{\widehat{y}_t}^{-1}(y^\star)\rangle)\end{aligned}$$

Putting these pieces together yields that

$$\begin{aligned}0 \leq & \frac{1}{\eta}(\langle \mathbb{E}_{\widehat{x}_t}^{-1}(x_{t+1}), \mathbb{E}_{\widehat{x}_t}^{-1}(x^\star)\rangle - \langle \mathbb{E}_{\widehat{x}_t}^{-1}(x_t), \mathbb{E}_{\widehat{x}_t}^{-1}(x^\star)\rangle) - \frac{\mu}{2}(d_{\mathcal{M}}(\widehat{x}_t, x^\star))^2 \\ & + \frac{1}{\eta}(\langle \mathbb{E}_{\widehat{y}_t}^{-1}(y_{t+1}), \mathbb{E}_{\widehat{y}_t}^{-1}(y^\star)\rangle - \langle \mathbb{E}_{\widehat{y}_t}^{-1}(y_t), \mathbb{E}_{\widehat{y}_t}^{-1}(y^\star)\rangle) - \frac{\mu}{2}(d_{\mathcal{N}}(\widehat{y}_t, y^\star))^2\end{aligned}$$

Equivalently, we have

$$\begin{aligned}0 \leq & \langle \mathbb{E}_{\widehat{x}_t}^{-1}(x_{t+1}), \mathbb{E}_{\widehat{x}_t}^{-1}(x^\star)\rangle - \langle \mathbb{E}_{\widehat{x}_t}^{-1}(x_t), \mathbb{E}_{\widehat{x}_t}^{-1}(x^\star)\rangle - \frac{\mu\eta}{2}(d_{\mathcal{M}}(\widehat{x}_t, x^\star))^2 \\ & + \langle \mathbb{E}_{\widehat{y}_t}^{-1}(y_{t+1}), \mathbb{E}_{\widehat{y}_t}^{-1}(y^\star)\rangle - \langle \mathbb{E}_{\widehat{y}_t}^{-1}(y_t), \mathbb{E}_{\widehat{y}_t}^{-1}(y^\star)\rangle - \frac{\mu\eta}{2}(d_{\mathcal{N}}(\widehat{y}_t, y^\star))^2\end{aligned} \tag{56}$$

It suffices to bound the terms in the right-hand side of Eq. (56) by leveraging the celebrated comparison inequalities on Riemannian manifold with bounded sectional curvature (see Proposition 2 and 3). More specifically, we define the constants using $\overline{\xi}(\cdot,\cdot)$ and $\underline{\xi}(\cdot,\cdot)$ from Proposition 2 and 3 as follows,

$$\overline{\xi}_0 = \overline{\xi}(\kappa_{\min}, D), \qquad \underline{\xi}_0 = \underline{\xi}(\kappa_{\max}, D)$$

By Proposition 2 and using that $\max\{d_{\mathcal{M}}(\widehat{x}_t, x^\star), d_{\mathcal{N}}(\widehat{y}_t, y^\star)\} \le D$, we have

$$\begin{aligned} -\langle \mathbb{E}_{\widehat{x}_t}^{-1}(x_t), \mathbb{E}_{\widehat{x}_t}^{-1}(x^\star)\rangle \le & \quad -\frac{1}{2}\left(\underline{\xi}_0 (d_{\mathcal{M}}(\widehat{x}_t, x_t))^2 + (d_{\mathcal{M}}(\widehat{x}_t, x^\star))^2 - (d_{\mathcal{M}}(x_t, x^\star))^2\right) \\ -\langle \mathbb{E}_{\widehat{y}_t}^{-1}(y_t), \mathbb{E}_{\widehat{y}_t}^{-1}(y^\star)\rangle \le & \quad -\frac{1}{2}\left(\underline{\xi}_0 (d_{\mathcal{N}}(\widehat{y}_t, y_t))^2 + (d_{\mathcal{N}}(\widehat{y}_t, y^\star))^2 - (d_{\mathcal{N}}(y_t, y^\star))^2\right) \end{aligned} \tag{57}$$

By Proposition 3 and using that $\max\{d_{\mathcal{M}}(\widehat{x}_t, x^\star), d_{\mathcal{N}}(\widehat{y}_t, y^\star)\} \le D$, we have

$$\langle \mathbb{E}_{\widehat{x}_t}^{-1}(x_{t+1}), \mathbb{E}_{\widehat{x}_t}^{-1}(x^\star)\rangle \le \frac{1}{2}\left(\overline{\xi}_0 (d_{\mathcal{M}}(\widehat{x}_t, x_{t+1}))^2 + (d_{\mathcal{M}}(\widehat{x}_t, x^\star))^2 - (d_{\mathcal{M}}(x_{t+1}, x^\star))^2\right)$$

and

$$\langle \mathbb{E}_{\widehat{y}_t}^{-1}(y_{t+1}), \mathbb{E}_{\widehat{y}_t}^{-1}(y^\star)\rangle \le \frac{1}{2}\left(\overline{\xi}_0 (d_{\mathcal{N}}(\widehat{y}_t, y_{t+1}))^2 + (d_{\mathcal{N}}(\widehat{y}_t, y^\star))^2 - (d_{\mathcal{N}}(y_{t+1}, y^\star))^2\right)$$

By the definition of an exponential map and Riemannian metric, we have

$$\begin{aligned} d_{\mathcal{M}}(\widehat{x}_t, x_{t+1}) = & \quad \|\mathbb{E}_{\widehat{x}_t}^{-1}(x_{t+1})\| \overset{\text{Eq. (55)}}{=} \|\eta \cdot \mathrm{grad}_x f(\widehat{x}_t, \widehat{y}_t) - \mathbb{E}_{\widehat{x}_t}^{-1}(x_t)\| \\ d_{\mathcal{N}}(\widehat{y}_t, y_{t+1}) = & \quad \|\mathbb{E}_{\widehat{y}_t}^{-1}(y_{t+1})\| \overset{\text{Eq. (55)}}{=} \|\eta \cdot \mathrm{grad}_y f(\widehat{x}_t, \widehat{y}_t) + \mathbb{E}_{\widehat{y}_t}^{-1}(y_t)\| \end{aligned} \tag{58}$$

Further, we see from the scheme of Algorithm 4 that we have

$$\begin{aligned} \widehat{x}_t & \leftarrow \mathbb{E}_{x_t}(-\eta \cdot \mathrm{grad}_x f(x_t, y_t)) \\ \widehat{y}_t & \leftarrow \mathbb{E}_{y_t}(\eta \cdot \mathrm{grad}_y f(x_t, y_t)) \end{aligned}$$

By the definition of an exponential map, we have

$$\mathbb{E}_{x_t}^{-1}(\widehat{x}_t) = -\eta \cdot \mathrm{grad}_x f(x_t, y_t), \qquad \mathbb{E}_{y_t}^{-1}(\widehat{y}_t) = \eta \cdot \mathrm{grad}_y f(x_t, y_t)$$

Using the definition of a parallel transport map and the above equations, we have

$$\mathbb{E}_{\widehat{x}_t}^{-1}(x_t) = \eta \cdot \Gamma_{x_t}^{\widehat{x}_t} \mathrm{grad}_x f(x_t, y_t), \qquad \mathbb{E}_{\widehat{y}_t}^{-1}(y_t) = -\eta \cdot \Gamma_{y_t}^{\widehat{y}_t} \mathrm{grad}_y f(x_t, y_t)$$

Since f is geodesically ℓ-smooth, we have

$$\begin{aligned} \|\mathrm{grad}_x f(\widehat{x}_t, \widehat{y}_t) - \Gamma_{x_t}^{\widehat{x}_t} \mathrm{grad}_x f(x_t, y_t)\| \le & \quad \ell(d_{\mathcal{M}}(\widehat{x}_t, x_t) + d_{\mathcal{N}}(\widehat{y}_t, y_t)) \\ \|\mathrm{grad}_y f(\widehat{x}_t, \widehat{y}_t) - \Gamma_{y_t}^{\widehat{y}_t} \mathrm{grad}_y f(x_t, y_t)\| \le & \quad \ell(d_{\mathcal{M}}(\widehat{x}_t, x_t) + d_{\mathcal{N}}(\widehat{y}_t, y_t)) \end{aligned}$$

Plugging the above inequalities into Eq. (58) yields that

$$\max\{d_{\mathcal{M}}(\widehat{x}_t, x_{t+1}), d_{\mathcal{N}}(\widehat{y}_t, y_{t+1})\} \le \eta\ell(d_{\mathcal{M}}(\widehat{x}_t, x_t) + d_{\mathcal{N}}(\widehat{y}_t, y_t))$$

Therefore, we have

$$\begin{aligned} \langle \mathbb{E}_{\widehat{x}_t}^{-1}(x_{t+1}), \mathbb{E}_{\widehat{x}_t}^{-1}(x^\star)\rangle \le & \quad \frac{1}{2}\left(2\overline{\xi}_0 \eta^2 \ell^2 ((d_{\mathcal{M}}(\widehat{x}_t, x_t))^2 + (d_{\mathcal{N}}(\widehat{y}_t, y_t))^2) + (d_{\mathcal{M}}(\widehat{x}_t, x^\star))^2 - (d_{\mathcal{M}}(x_{t+1}, x^\star))^2\right) \\ \langle \mathbb{E}_{\widehat{y}_t}^{-1}(y_{t+1}), \mathbb{E}_{\widehat{y}_t}^{-1}(y^\star)\rangle \le & \quad \frac{1}{2}\left(2\overline{\xi}_0 \eta^2 \ell^2 ((d_{\mathcal{M}}(\widehat{x}_t, x_t))^2 + (d_{\mathcal{N}}(\widehat{y}_t, y_t))^2) + (d_{\mathcal{N}}(\widehat{y}_t, y^\star))^2 - (d_{\mathcal{N}}(y_{t+1}, y^\star))^2\right) \end{aligned}$$

Plugging the above inequalities and Eq. (57) into Eq. (56) yields the desired inequality. □

The second lemma gives another key inequality that is satisfied by the iterates generated by Algorithm 5.

Lemma 8. *Under Assumption 6 (or Assumption 7) and the noisy model (cf. Eq.* (53) *and* (54)*) and let* $\{(x_t, y_t), (\widehat{x}_t, \widehat{y}_t)\}_{t=0}^{T-1}$ *be generated by Algorithm 5 with the stepsize* $\eta > 0$*. Then, we have*

$$\begin{aligned}\mathbb{E}[f(\widehat{x}_t, y^\star) - f(x^\star, \widehat{y}_t)] \leq \frac{1}{2\eta}\mathbb{E}\left[(d_{\mathcal{M}}(x_t, x^\star))^2 - (d_{\mathcal{M}}(x_{t+1}, x^\star))^2 + (d_{\mathcal{N}}(y_t, y^\star))^2 - (d_{\mathcal{N}}(y_{t+1}, y^\star))^2\right]\\ +6\overline{\xi}_0\eta\ell^2\mathbb{E}\left[(d_{\mathcal{M}}(\widehat{x}_t, x_t))^2 + (d_{\mathcal{N}}(\widehat{y}_t, y_t))^2\right] - \frac{1}{2\eta}\underline{\xi}_0\mathbb{E}\left[(d_{\mathcal{M}}(\widehat{x}_t, x_t))^2 + (d_{\mathcal{N}}(\widehat{y}_t, y_t))^2\right]\\ -\frac{\mu}{2}\mathbb{E}\left[(d_{\mathcal{M}}(\widehat{x}_t, x^\star))^2 + (d_{\mathcal{N}}(\widehat{y}_t, y^\star))^2\right] + 3\overline{\xi}_0\eta\sigma^2\end{aligned}$$

where $(x^\star, y^\star) \in \mathcal{M} \times \mathcal{N}$ *is a global saddle point of* f*.*

Proof. Using the same argument, we have ($\mu = 0$ refers to geodesically convex-concave case)

$$\begin{aligned}&f(\widehat{x}_t, y^\star) - f(x^\star, \widehat{y}_t) = f(\widehat{x}_t, \widehat{y}_t) - f(x^\star, \widehat{y}_t) - (f(\widehat{x}_t, \widehat{y}_t) - f(\widehat{x}_t, y^\star))\\ &\leq \ -\langle \mathrm{grad}_x f(\widehat{x}_t, \widehat{y}_t), \mathbb{E}_{\widehat{x}_t}^{-1}(x^\star)\rangle + \langle \mathrm{grad}_y f(\widehat{x}_t, \widehat{y}_t), \mathbb{E}_{\widehat{y}_t}^{-1}(y^\star)\rangle - \frac{\mu}{2}(d_{\mathcal{M}}(\widehat{x}_t, x^\star))^2 - \frac{\mu}{2}(d_{\mathcal{N}}(\widehat{y}_t, y^\star))^2\end{aligned}$$

Combining the arguments used in Lemma 7 and the scheme of Algorithm 5, we have

$$\begin{aligned}-\langle \widehat{g}_x^t, \mathbb{E}_{\widehat{x}_t}^{-1}(x^\star)\rangle &= \frac{1}{\eta}(\langle \mathbb{E}_{\widehat{x}_t}^{-1}(x_{t+1}), \mathbb{E}_{\widehat{x}_t}^{-1}(x^\star)\rangle - \langle \mathbb{E}_{\widehat{x}_t}^{-1}(x_t), \mathbb{E}_{\widehat{x}_t}^{-1}(x^\star)\rangle)\\ \langle \widehat{g}_y^t, \mathbb{E}_{\widehat{y}_t}^{-1}(y^\star)\rangle &= \frac{1}{\eta}(\langle \mathbb{E}_{\widehat{y}_t}^{-1}(y_{t+1}), \mathbb{E}_{\widehat{y}_t}^{-1}(y^\star)\rangle - \langle \mathbb{E}_{\widehat{y}_t}^{-1}(y_t), \mathbb{E}_{\widehat{y}_t}^{-1}(y^\star)\rangle)\end{aligned}$$

Putting these pieces together with Eq. (53) yields that

$$\begin{aligned}f(\widehat{x}_t, y^\star) - f(x^\star, \widehat{y}_t) \leq \frac{1}{\eta}(\langle \mathbb{E}_{\widehat{x}_t}^{-1}(x_{t+1}), \mathbb{E}_{\widehat{x}_t}^{-1}(x^\star)\rangle - \langle \mathbb{E}_{\widehat{x}_t}^{-1}(x_t), \mathbb{E}_{\widehat{x}_t}^{-1}(x^\star)\rangle)\\ +\frac{1}{\eta}(\langle \mathbb{E}_{\widehat{y}_t}^{-1}(y_{t+1}), \mathbb{E}_{\widehat{y}_t}^{-1}(y^\star)\rangle - \langle \mathbb{E}_{\widehat{y}_t}^{-1}(y_t), \mathbb{E}_{\widehat{y}_t}^{-1}(y^\star)\rangle) - \frac{\mu}{2}(d_{\mathcal{M}}(\widehat{x}_t, x^\star))^2 - \frac{\mu}{2}(d_{\mathcal{N}}(\widehat{y}_t, y^\star))^2\\ +\langle \widehat{\xi}_x^t, \mathbb{E}_{\widehat{x}_t}^{-1}(x^\star)\rangle - \langle \widehat{\xi}_y^t, \mathbb{E}_{\widehat{y}_t}^{-1}(y^\star)\rangle\end{aligned} \tag{59}$$

By the same argument as used in Lemma 7, we have

$$\begin{aligned}-\langle \mathbb{E}_{\widehat{x}_t}^{-1}(x_t), \mathbb{E}_{\widehat{x}_t}^{-1}(x^\star)\rangle \leq \ -\frac{1}{2}\left(\underline{\xi}_0(d_{\mathcal{M}}(\widehat{x}_t, x_t))^2 + (d_{\mathcal{M}}(\widehat{x}_t, x^\star))^2 - (d_{\mathcal{M}}(x_t, x^\star))^2\right)\\ -\langle \mathbb{E}_{\widehat{y}_t}^{-1}(y_t), \mathbb{E}_{\widehat{y}_t}^{-1}(y^\star)\rangle \leq \ -\frac{1}{2}\left(\underline{\xi}_0(d_{\mathcal{N}}(\widehat{y}_t, y_t))^2 + (d_{\mathcal{N}}(\widehat{y}_t, y^\star))^2 - (d_{\mathcal{N}}(y_t, y^\star))^2\right)\end{aligned} \tag{60}$$

and

$$\begin{aligned}\langle \mathbb{E}_{\widehat{x}_t}^{-1}(x_{t+1}), \mathbb{E}_{\widehat{x}_t}^{-1}(x^\star)\rangle \leq \ \frac{1}{2}\left(\overline{\xi}_0\eta^2\|\widehat{g}_x^t - \Gamma_{x_t}^{\widehat{x}_t} g_x^t\|^2 + (d_{\mathcal{M}}(\widehat{x}_t, x^\star))^2 - (d_{\mathcal{M}}(x_{t+1}, x^\star))^2\right)\\ \langle \mathbb{E}_{\widehat{y}_t}^{-1}(y_{t+1}), \mathbb{E}_{\widehat{y}_t}^{-1}(y^\star)\rangle \leq \ \frac{1}{2}\left(\overline{\xi}_0\eta^2\|\widehat{g}_y^t - \Gamma_{y_t}^{\widehat{y}_t} g_y^t\|^2 + (d_{\mathcal{N}}(\widehat{y}_t, y^\star))^2 - (d_{\mathcal{N}}(y_{t+1}, y^\star))^2\right)\end{aligned}$$

Since f is geodesically ℓ-smooth and Eq. (53) holds, we have

$$\begin{aligned}\|\widehat{g}_x^t - \Gamma_{x_t}^{\widehat{x}_t} g_x^t\|^2 &\leq 3\|\widehat{\xi}_x^t\|^2 + 3\|\xi_x^t\|^2 + 6\ell^2(d_{\mathcal{M}}(\widehat{x}_t, x_t))^2 + 6\ell^2(d_{\mathcal{N}}(\widehat{y}_t, y_t))^2\\ \|\widehat{g}_y^t - \Gamma_{y_t}^{\widehat{y}_t} g_y^t\|^2 &\leq 3\|\widehat{\xi}_y^t\|^2 + 3\|\xi_y^t\|^2 + 6\ell^2(d_{\mathcal{M}}(\widehat{x}_t, x_t))^2 + 6\ell^2(d_{\mathcal{N}}(\widehat{y}_t, y_t))^2\end{aligned}$$

Therefore, we have

$$\begin{aligned}
&\langle \mathbb{E}_{\widehat{x}_t}^{-1}(x_{t+1}), \mathbb{E}_{\widehat{x}_t}^{-1}(x^\star)\rangle + \langle \mathbb{E}_{\widehat{y}_t}^{-1}(y_{t+1}), \mathbb{E}_{\widehat{y}_t}^{-1}(y^\star)\rangle \\
&\leq \quad 6\overline{\xi}_0\eta^2\ell^2((d_{\mathcal{M}}(\widehat{x}_t, x_t))^2 + (d_{\mathcal{N}}(\widehat{y}_t, y_t))^2) + \frac{3}{2}\overline{\xi}_0\eta^2(\|\widehat{\xi}_x^t\|^2 + \|\xi_x^t\|^2 + \|\widehat{\xi}_y^t\|^2 + \|\xi_y^t\|^2) \\
&\quad + \frac{1}{2}\left((d_{\mathcal{M}}(\widehat{x}_t, x^\star))^2 - (d_{\mathcal{M}}(x_{t+1}, x^\star))^2 + (d_{\mathcal{N}}(\widehat{y}_t, y^\star))^2 - (d_{\mathcal{N}}(y_{t+1}, y^\star))^2\right)
\end{aligned}$$

Plugging the above inequalities and Eq. (60) into Eq. (59) yields that

$$\begin{aligned}
f(\widehat{x}_t, y^\star) - f(x^\star, \widehat{y}_t) \leq\ & \frac{1}{2\eta}\left((d_{\mathcal{M}}(x_t, x^\star))^2 - (d_{\mathcal{M}}(x_{t+1}, x^\star))^2 + (d_{\mathcal{N}}(y_t, y^\star))^2 - (d_{\mathcal{N}}(y_{t+1}, y^\star))^2\right) \\
&+ 6\overline{\xi}_0\eta\ell^2((d_{\mathcal{M}}(\widehat{x}_t, x_t))^2 + (d_{\mathcal{N}}(\widehat{y}_t, y_t))^2) + \frac{3}{2}\overline{\xi}_0\eta(\|\widehat{\xi}_x^t\|^2 + \|\xi_x^t\|^2 + \|\widehat{\xi}_y^t\|^2 + \|\xi_y^t\|^2) \\
&- \frac{1}{2\eta}\underline{\xi}_0\left((d_{\mathcal{M}}(\widehat{x}_t, x_t))^2 + (d_{\mathcal{N}}(\widehat{y}_t, y_t))^2\right) - \frac{\mu}{2}(d_{\mathcal{M}}(\widehat{x}_t, x^\star))^2 - \frac{\mu}{2}(d_{\mathcal{N}}(\widehat{y}_t, y^\star))^2 \\
&+ \langle \widehat{\xi}_x^t, \mathbb{E}_{\widehat{x}_t}^{-1}(x^\star)\rangle - \langle \widehat{\xi}_y^t, \mathbb{E}_{\widehat{y}_t}^{-1}(y^\star)\rangle
\end{aligned}$$

Taking the expectation of both sides and using Eq. (54) yields the desired inequality. □

4.4.2 Proof of Theorem 6

Since Riemannian metrics satisfy the triangle inequality, we have

$$(d_{\mathcal{M}}(\widehat{x}_t, x^\star))^2 + (d_{\mathcal{N}}(\widehat{y}_t, y^\star))^2 \geq \frac{1}{2}((d_{\mathcal{M}}(x_t, x^\star))^2 + (d_{\mathcal{N}}(y_t, y^\star))^2) - (d_{\mathcal{M}}(\widehat{x}_t, x_t))^2 + (d_{\mathcal{N}}(\widehat{y}_t, y_t))^2$$

Plugging the above inequality into the inequality from Lemma 7 yields that

$$\begin{aligned}
&(d_{\mathcal{M}}(x_{t+1}, x^\star))^2 + (d_{\mathcal{N}}(y_{t+1}, y^\star))^2 \\
&\leq \quad \left(1 - \frac{\mu\eta}{2}\right)\left((d_{\mathcal{M}}(x_t, x^\star))^2 + (d_{\mathcal{N}}(y_t, y^\star))^2\right) + (4\overline{\xi}_0\eta^2\ell^2 + \mu\eta - \underline{\xi}_0)((d_{\mathcal{M}}(\widehat{x}_t, x_t))^2 + (d_{\mathcal{N}}(\widehat{y}_t, y_t))^2
\end{aligned}$$

Since $\eta = \min\{\frac{1}{4\ell\sqrt{\tau_0}}, \frac{\underline{\xi}_0}{2\mu}\}$, we have $4\overline{\xi}_0\eta^2\ell^2 + \mu\eta - \underline{\xi}_0 \leq 0$. By the definition, we have $\tau_0 \geq 1$, $\kappa \geq 1$ and $\underline{\xi}_0 \leq 1$. This implies that

$$1 - \frac{\mu\eta}{2} = 1 - \min\left\{\frac{1}{8\kappa\sqrt{\tau_0}}, \frac{\underline{\xi}_0}{4}\right\} > 0$$

Putting these pieces together yields that

$$\begin{aligned}
(d_{\mathcal{M}}(x_T, x^\star))^2 + (d_{\mathcal{N}}(y_T, y^\star))^2 \quad &\leq \quad \left(1 - \min\left\{\frac{1}{8\kappa\sqrt{\tau_0}}, \frac{\underline{\xi}_0}{4}\right\}\right)^T (d_{\mathcal{M}}(x_0, x^\star))^2 + (d_{\mathcal{N}}(y_0, y^\star))^2 \\
&\leq \quad \left(1 - \min\left\{\frac{1}{8\kappa\sqrt{\tau_0}}, \frac{\underline{\xi}_0}{4}\right\}\right)^T D_0
\end{aligned}$$

This completes the proof.

4.4.3 Proof of Theorem 7

Since Riemannian metrics satisfy the triangle inequality, we have

$$(d_{\mathcal{M}}(\widehat{x}_t, x^\star))^2 + (d_{\mathcal{N}}(\widehat{y}_t, y^\star))^2 \geq \frac{1}{2}((d_{\mathcal{M}}(x_t, x^\star))^2 + (d_{\mathcal{N}}(y_t, y^\star))^2) - (d_{\mathcal{M}}(\widehat{x}_t, x_t))^2 + (d_{\mathcal{N}}(\widehat{y}_t, y_t))^2$$

Plugging the above inequality into the inequality from Lemma 8 yields that

$$\begin{aligned}\mathbb{E}[f(\widehat{x}_t, y^\star) - f(x^\star, \widehat{y}_t)] \leq{}& \frac{1}{2\eta}\mathbb{E}\left[(d_{\mathcal{M}}(x_t, x^\star))^2 - (d_{\mathcal{M}}(x_{t+1}, x^\star))^2 + (d_{\mathcal{N}}(y_t, y^\star))^2 - (d_{\mathcal{N}}(y_{t+1}, y^\star))^2\right] \\ &+(6\overline{\xi}_0\eta\ell^2 + \frac{\mu}{2} - \frac{1}{2\eta}\underline{\xi}_0)\mathbb{E}\left[(d_{\mathcal{M}}(\widehat{x}_t, x_t))^2 + (d_{\mathcal{N}}(\widehat{y}_t, y_t))^2\right] - \frac{\mu}{4}\mathbb{E}\left[(d_{\mathcal{M}}(\widehat{x}_t, x^\star))^2 + (d_{\mathcal{N}}(\widehat{y}_t, y^\star))^2\right] + 3\overline{\xi}_0\eta\sigma^2\end{aligned}$$

Since $(x^\star, y^\star) \in \mathcal{M} \times \mathcal{N}$ is a global saddle point of f, we have $\mathbb{E}[f(\widehat{x}_t, y^\star) - f(x^\star, \widehat{y}_t)] \geq 0$. Then, we have

$$\begin{aligned}&\mathbb{E}\left[(d_{\mathcal{M}}(x_{t+1}, x^\star))^2 + (d_{\mathcal{N}}(y_{t+1}, y^\star))^2\right] \\ &\leq \left(1 - \frac{\mu\eta}{2}\right)\mathbb{E}\left[(d_{\mathcal{M}}(x_t, x^\star))^2 + (d_{\mathcal{N}}(y_t, y^\star))^2\right] + (12\overline{\xi}_0\eta^2\ell^2 + \mu\eta - \underline{\xi}_0)\mathbb{E}\left[(d_{\mathcal{M}}(\widehat{x}_t, x_t))^2 + (d_{\mathcal{N}}(\widehat{y}_t, y_t))^2\right] \\ &\quad +6\overline{\xi}_0\eta^2\sigma^2\end{aligned}$$

Since $\eta \leq \min\{\frac{1}{24\ell\sqrt{\tau_0}}, \frac{\underline{\xi}_0}{2\mu}\}$, we have $12\overline{\xi}_0\eta^2\ell^2 + \mu\eta - \underline{\xi}_0 \leq 0$. This implies that

$$\mathbb{E}\left[(d_{\mathcal{M}}(x_{t+1}, x^\star))^2 + (d_{\mathcal{N}}(y_{t+1}, y^\star))^2\right] \leq \left(1 - \frac{\mu\eta}{2}\right)\mathbb{E}\left[(d_{\mathcal{M}}(x_t, x^\star))^2 + (d_{\mathcal{N}}(y_t, y^\star))^2\right] + 6\overline{\xi}_0\eta^2\sigma^2$$

By the definition, we have $\tau_0 \geq 1$, $\kappa \geq 1$ and $\underline{\xi}_0 \leq 1$. This implies that

$$1 - \frac{\mu\eta}{2} \geq 1 - \min\left\{\frac{1}{48\kappa\sqrt{\tau_0}}, \frac{\underline{\xi}_0}{4}\right\} > 0$$

By the inductive arguments, we have

$$\begin{aligned}&\mathbb{E}\left[(d_{\mathcal{M}}(x_T, x^\star))^2 + (d_{\mathcal{N}}(y_T, y^\star))^2\right] \\ \leq{}& \left(1 - \frac{\mu\eta}{2}\right)^T\left((d_{\mathcal{M}}(x_0, x^\star))^2 + (d_{\mathcal{N}}(y_0, y^\star))^2\right) + 6\overline{\xi}_0\eta^2\sigma^2\left(\sum_{t=0}^{T-1}\left(1 - \frac{\mu\eta}{2}\right)^t\right) \\ \leq{}& \left(1 - \frac{\mu\eta}{2}\right)^T D_0 + \frac{12\overline{\xi}_0\eta\sigma^2}{\mu}\end{aligned}$$

Since $\eta = \min\{\frac{1}{24\ell\sqrt{\tau_0}}, \frac{\underline{\xi}_0}{2\mu}, \frac{2(\log(T) + \log(\mu^2 D_0\sigma^{-2}))}{\mu T}\}$, we have

$$\begin{aligned}\left(1 - \frac{\mu\eta}{2}\right)^T D_0 \quad &\leq \quad \left(1 - \min\left\{\frac{1}{48\kappa\sqrt{\tau_0}}, \frac{\underline{\xi}_0}{4}\right\}\right)^T D_0 + \left(1 - \frac{\log(\mu^2 D_0\sigma^{-2}T)}{T}\right)^T D_0 \\ &\overset{1+x\leq e^x}{\leq} \quad \left(1 - \min\left\{\frac{1}{48\kappa\sqrt{\tau_0}}, \frac{\underline{\xi}_0}{4}\right\}\right)^T D_0 + \frac{\sigma^2}{\mu^2 T}\end{aligned}$$

and

$$\frac{12\overline{\xi}_0\eta\sigma^2}{\mu} \leq \frac{24\overline{\xi}_0\sigma^2}{\mu^2 T}\log\left(\frac{\mu^2 D_0 T}{\sigma^2}\right)$$

Putting these pieces together yields that

$$\mathbb{E}\left[(d_{\mathcal{M}}(x_T, x^\star))^2 + (d_{\mathcal{N}}(y_T, y^\star))^2\right] \leq \left(1 - \min\left\{\frac{1}{48\kappa\sqrt{\tau_0}}, \frac{\underline{\xi}_0}{4}\right\}\right)^T D_0 + \frac{\sigma^2}{\mu^2 T} + \frac{24\overline{\xi}_0\sigma^2}{\mu^2 T}\log\left(\frac{\mu^2 D_0 T}{\sigma^2}\right)$$

This completes the proof.

4.4.4 Proof of Theorem 8

By the inductive formulas of $\bar{x}_{t+1} = \mathbb{E}_{\bar{x}_t}(\frac{1}{t+1} \cdot \mathbb{E}^{-1}_{\bar{x}_t}(\widehat{x}_t))$ and $\bar{y}_{t+1} = \mathbb{E}_{\bar{y}_t}(\frac{1}{t+1} \cdot \mathbb{E}^{-1}_{\bar{y}_t}(\widehat{y}_t))$ and using [ZZS22, Lemma C.2], we have

$$f(\bar{x}_T, y^\star) - f(x^\star, \bar{y}_T) \leq \frac{1}{T}\left(\sum_{t=0}^{T-1} f(\widehat{x}_t, y^\star) - f(x^\star, \widehat{y}_t)\right)$$

Plugging the above inequality into the inequality from Lemma 8 yields that (recall that $\mu = 0$ in geodesically convex-concave setting here)

$$\begin{aligned}\mathbb{E}[f(\bar{x}_T, y^\star) - f(x^\star, \bar{y}_T)] &\leq \frac{1}{2\eta T}\left((d_{\mathcal{M}}(x_0, x^\star))^2 + (d_{\mathcal{N}}(y_0, y^\star))^2\right) \\ &+\frac{1}{T}\left(6\bar{\xi}_0\eta\ell^2 - \frac{1}{2\eta}\xi_0\right)\left(\sum_{t=0}^{T-1}\mathbb{E}\left[(d_{\mathcal{M}}(\widehat{x}_t, x_t))^2 + (d_{\mathcal{N}}(\widehat{y}_t, y_t))^2\right]\right) + 3\bar{\xi}_0\eta\sigma^2\end{aligned}$$

Since $\eta \leq \frac{1}{4\ell\sqrt{\tau_0}}$, we have $6\bar{\xi}_0\eta\ell^2 - \frac{1}{2\eta}\xi_0 \leq 0$. Then, this together with $(d_{\mathcal{M}}(x_0, x^\star))^2 + (d_{\mathcal{N}}(y_0, y^\star))^2 \leq D_0$ implies that

$$\mathbb{E}[f(\bar{x}_T, y^\star) - f(x^\star, \bar{y}_T)] \leq \frac{D_0}{2\eta T} + 3\bar{\xi}_0\eta\sigma^2$$

Since $\eta = \min\{\frac{1}{4\ell\sqrt{\tau_0}}, \frac{1}{\sigma}\sqrt{\frac{D_0}{\bar{\xi}_0 T}}\}$, we have

$$\frac{D_0}{2\eta T} \leq \frac{2\ell D_0\sqrt{\tau_0}}{T} + \frac{\sigma}{2}\sqrt{\frac{\bar{\xi}_0 D_0}{T}}$$

and

$$3\bar{\xi}_0\eta\sigma^2 \leq 3\sigma\sqrt{\frac{\bar{\xi}_0 D_0}{T}}$$

Putting these pieces together yields that

$$\mathbb{E}[f(\bar{x}_T, y^\star) - f(x^\star, \bar{y}_T)] \leq \frac{2\ell D_0\sqrt{\tau_0}}{T} + \frac{7\sigma}{2}\sqrt{\frac{\bar{\xi}_0 D_0}{T}}$$

This completes the proof.

5 Conclusion

In this paper, we introduced a comprehensive framework for *Inexact Riemannian Gradient Descent* (IRGD) and its variant, IRGDr, designed to solve nonconvex optimization problems on Riemannian manifolds. By generalizing inexact gradient conditions to the Riemannian setting, our framework efficiently handles gradient approximation errors while providing strong convergence guarantees. These include stationarity of accumulation points and convergence of gradient sequences under the Kurdyka-Łojasiewicz (KL) property.

Additionally, we extended the framework to include the *Riemannian extragradient* (REG) method, demonstrating its robust performance in min-max optimization scenarios. Our results

indicate that the proposed methods perform well under both deterministic and stochastic perturbations, offering theoretical guarantees and practical insights into first-order algorithms in geodesic metric spaces.

While this work establishes foundational results, future research could explore several directions, including extending the IRGD framework to stochastic settings, investigating stricter inexact gradient conditions, and applying these methods to a broader range of machine learning applications such as low-rank matrix completion and robust dimensionality reduction. Furthermore, tightening the dependence on manifold curvature and generalizing these results to Riemannian Monotone Variational Inequalities (RMVI) presents promising opportunities for advancing the field of Riemannian optimization.

References

[AB09] Hedy Attouch and Jérôme Bolte. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Mathematical Programming*, 116:5–16, 2009.

[ABM20] K. Antonakopoulos, E. V. Belmega, and P. Mertikopoulos. Online and stochastic optimization beyond Lipschitz continuity: A Riemannian approach. In *ICLR*, 2020.

[ADLH19] L. Adolphs, H. Daneshmand, A. Lucchi, and T. Hofmann. Local saddle point optimization: A curvature exploitation approach. In *AISTATS*, pages 486–495. PMLR, 2019.

[AG16] A. Anandkumar and R. Ge. Efficient approaches for escaping higher order saddle points in non-convex optimization. In *COLT*, pages 81–102. PMLR, 2016.

[AH19] P-A. Absil and S. Hosseini. A collection of nonsmooth Riemannian optimization problems. In *Nonsmooth Optimization and Its Applications*, pages 1–15. Springer, 2019.

[ALW21] J. Abernethy, K. A. Lai, and A. Wibisono. Last-iterate convergence rates for min-max optimization: Convergence of Hamiltonian gradient descent and consensus optimization. In *ALT*, pages 3–47. PMLR, 2021.

[AMLJG20] W. Azizian, I. Mitliagkas, S. Lacoste-Julien, and G. Gidel. A tight and unified analysis of gradient-based methods for a whole spectrum of differentiable games. In *AISTATS*, pages 2863–2873. PMLR, 2020.

[AMS08] P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, Princeton, NJ, 2008.

[AMS09] P-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2009.

[AOBL20] F. Alimisis, A. Orvieto, G. Bécigneul, and A. Lucchi. A continuous-time perspective for modeling acceleration in Riemannian optimization. In *AISTATS*, pages 1297–1307. PMLR, 2020.

[ATV13] Bijan Afsari, Roberto Tron, and René Vidal. On the convergence of gradient descent for finding the Riemannian center of mass. *SIAM Journal on Control and Optimization*, 51(3):2230–2260, 2013.

[BA11] N. Boumal and P-A. Absil. RTRMC: A Riemannian trust-region method for low-rank matrix completion. In *NIPS*, pages 406–414, 2011.

[Bac14] M. Bacak. *Convex Analysis and Optimization in Hadamard Spaces*, volume 22. Walter de Gruyter GmbH & Co KG, 2014.

[BAC19a] N. Boumal, P-A. Absil, and C. Cartis. Global rates of convergence for nonconvex optimization on manifolds. *IMA Journal of Numerical Analysis*, 39(1):1–33, 2019.

[BAC19b] Nicolas Boumal, Pierre-Antoine Absil, and Coralia Cartis. Global rates of convergence for nonconvex optimization on manifolds. *IMA Journal of Numerical Analysis*, 39(1):1–33, 2019.

[BBB+01] D. Burago, I. D. Burago, Y. Burago, S. Ivanov, S. V. Ivanov, and S. A. Ivanov. *A Course in Metric Geometry*, volume 33. American Mathematical Soc., 2001.

[BC11] H. H. Bauschke and P. L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, volume 408. Springer, 2011.

[BCCS22] Albert S Berahas, Liyuan Cao, Krzysztof Choromanski, and Katya Scheinberg. A theoretical and empirical comparison of gradient approximations in derivative-free optimization. *Foundations of Computational Mathematics*, 22(2):507–560, 2022.

[Ber97] Dimitri P Bertsekas. Nonlinear programming. *Journal of the Operational Research Society*, 48(3):334–334, 1997.

[BFM17] G. C. Bento, O. P. Ferreira, and J. G. Melo. Iteration-complexity of gradient, subgradient and proximal point methods on Riemannian manifolds. *Journal of Optimization Theory and Applications*, 173(2):548–562, 2017.

[BG19] G. Becigneul and O-E. Ganea. Riemannian adaptive optimization methods. In *ICLR*, 2019.

[BGP92] Y. Burago, M. Gromov, and G. Perel'man. A. D. Alexandrov spaces with curvature bounded below. *Russian Mathematical Surveys*, 47(2):1, 1992.

[BH19] R. Bergmann and R. Herzog. Intrinsic formulation of KKT conditions and constraint qualifications on smooth manifolds. *SIAM Journal on Optimization*, 29(4):2423–2444, 2019.

[BMSS19] I. M. Bomze, P. Mertikopoulos, W. Schachinger, and M. Staudigl. Hessian barrier algorithms for linearly constrained optimization problems. *SIAM Journal on Optimization*, 29(3):2100–2127, 2019.

[BNO11] GC Bento, JX Neto, and PR Oliveira. Convergence of inexact descent methods for nonconvex optimization on Riemannian manifolds. *arXiv preprint arXiv:1103.4828*, 2011.

[Bon13] S. Bonnabel. Stochastic gradient descent on Riemannian manifolds. *IEEE Transactions on Automatic Control*, 58(9):2217–2229, 2013.

[BRM+18] D. Balduzzi, S. Racaniere, J. Martens, J. Foerster, K. Tuyls, and T. Graepel. The mechanics of N-player differentiable games. In *ICML*, pages 354–363. PMLR, 2018.

[Bro11] L. E. J. Brouwer. Über abbildung von mannigfaltigkeiten. *Mathematische Annalen*, 71(1):97–115, 1911.

[BST14] Jérôme Bolte, Shoham Sabach, and Marc Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1):459–494, 2014.

[BTEGN09] A. Ben-Tal, L. EL Ghaoui, and A. Nemirovski. *Robust Optimization*, volume 28. Princeton University Press, 2009.

[Car91] Richard G Carter. On the global convergence of trust region algorithms using inexact gradient information. *SIAM Journal on Numerical Analysis*, 28(1):251–265, 1991.

[CB19] C. Criscitiello and N. Boumal. Efficiently escaping saddle points on manifolds. In *NeurIPS*, pages 5987–5997, 2019.

[CE75] J. Cheeger and D. G. Ebin. *Comparison Theorems in Riemannian Geometry*, volume 9. North-Holland Amsterdam, 1975.

[CKK+18] N. Chen, A. Klushyn, R. Kurle, X. Jiang, J. Bayer, and P. Smagt. Metrics for deep generative models. In *AISTATS*, pages 1540–1550. PMLR, 2018.

[CMSZ20] S. Chen, S. Ma, A. M-C. So, and T. Zhang. Proximal gradient method for nonsmooth optimization over the Stiefel manifold. *SIAM Journal on Optimization*, 30(1):210–239, 2020.

[COZ22] Y. Cai, A. Oikonomou, and W. Zheng. Tight last-iterate convergence of the extragradient method for constrained monotone variational inequalities. *ArXiv Preprint: 2204.09228*, 2022.

[CS18] Coralia Cartis and Katya Scheinberg. Global convergence rate analysis of unconstrained optimization methods based on probabilistic models. *Mathematical Programming*, 169:337–375, 2018.

[CWB08] E. J. Candes, M. B. Wakin, and S. P. Boyd. Enhancing sparsity by reweighted ℓ_1 minimization. *Journal of Fourier Analysis and Applications*, 14(5):877–905, 2008.

[Dia20] J. Diakonikolas. Halpern iteration for near-optimal and parameter-free monotone inclusion and strong solutions to variational inequalities. In *COLT*, pages 1428–1451. PMLR, 2020.

[DJL+17] S. S. Du, C. Jin, J. D. Lee, M. I. Jordan, B. Póczos, and A. Singh. Gradient descent can take exponential time to escape saddle points. In *NIPS*, pages 1067–1077, 2017.

[DP18] C. Daskalakis and I. Panageas. The limit points of (optimistic) gradient descent in min-max optimization. In *NIPS*, pages 9256–9266, 2018.

[DP19] C. Daskalakis and I. Panageas. Last-iterate convergence: Zero-sum games and constrained min-max optimization. In *ITCS*, 2019.

[FGHS21] J. Fearnley, P. W. Goldberg, A. Hollender, and R. Savani. The complexity of gradient descent: CLS = PPAD∩ PLS. In *STOC*, pages 46–59, 2021.

[FJ07] P. T. Fletcher and S. Joshi. Riemannian geometry for the statistical analysis of diffusion tensor data. *Signal Processing*, 87(2):250–262, 2007.

[FO98] O. P. Ferreira and P. R. Oliveira. Subgradient algorithm on Riemannian manifolds. *Journal of Optimization Theory and Applications*, 97(1):93–104, 1998.

[FO02] O. P. Ferreira and P. R. Oliveira. Proximal point algorithm on Riemannian manifolds. *Optimization*, 51(2):257–270, 2002.

[FOP20] A. Fallah, A. Ozdaglar, and S. Pattathil. An optimal multistage stochastic gradient method for minimax problems. In *CDC*, pages 3573–3579. IEEE, 2020.

[FP07] F. Facchinei and J-S. Pang. *Finite-Dimensional Variational Inequalities and Complementarity Problems*. Springer Science & Business Media, 2007.

[FPN05] O. P. Ferreira, L. R. Pérez, and S. Z. Németh. Singularities of monotone vector fields and an extragradient-type algorithm. *Journal of Global Optimization*, 31(1):133–151, 2005.

[FRW11] M. Fornasier, H. Rauhut, and R. Ward. Low-rank matrix recovery via iteratively reweighted least squares minimization. *SIAM Journal on Optimization*, 21(4):1614–1640, 2011.

[GHJY15] R. Ge, F. Huang, C. Jin, and Y. Yuan. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *COLT*, pages 797–842. PMLR, 2015.

[GHP+19] G. Gidel, R. A. Hemmat, M. Pezeshki, R. Le Priol, G. Huang, S. Lacoste-Julien, and I. Mitliagkas. Negative momentum for improved game dynamics. In *AISTATS*, pages 1802–1811. PMLR, 2019.

[GJZ17] R. Ge, C. Jin, and Y. Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *ICML*, pages 1233–1242. PMLR, 2017.

[GLCY18] B. Gao, X. Liu, X. Chen, and Y. Yuan. A new first-order algorithmic framework for optimization problems with orthogonality constraints. *SIAM Journal on Optimization*, 28(1):302–332, 2018.

[GPAM+14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. In *NIPS*, pages 2672–2680, 2014.

[GPDO20] N. Golowich, S. Pattathil, C. Daskalakis, and A. Ozdaglar. Last iterate is slower than averaged iterate in smooth convex-concave saddle point problems. In *COLT*, pages 1758–1784. PMLR, 2020.

[GVGM21a] A. Giannou, E. V. Vlatakis-Gkaragkounis, and P. Mertikopoulos. On the rate of convergence of regularized learning in games: From bandits and uncertainty to optimism and beyond. In *NeurIPS*, pages 22655–22666, 2021.

[GVGM21b] A. Giannou, E. V. Vlatakis-Gkaragkounis, and P. Mertikopoulos. Survival of the strictest: Stable and unstable equilibria under regularized learning with partial information. In *COLT*, pages 2147–2148. PMLR, 2021.

[HA21] E. Y. Hamedani and N. S. Aybat. A primal-dual algorithm with line search for general convex-concave saddle point problems. *SIAM Journal on Optimization*, 31(2):1299–1329, 2021.

[Hel23] E. D. Helly. Über mengen konvexer körper mit gemeinschaftlichen punkte. *Jahresbericht der Deutschen Mathematiker-Vereinigung*, 32:175–176, 1923.

[HG23] Feihu Huang and Shangqian Gao. Gradient descent ascent for minimax problems on Riemannian manifolds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):8466–8476, 2023.

[HGH20] F. Huang, S. Gao, and H. Huang. Gradient descent ascent for min-max problems on Riemannian manifolds. *ArXiv Preprint: 2010.06097*, 2020.

[HJL+19] J. Hu, B. Jiang, L. Lin, Z. Wen, and Y. Yuan. Structured quasi-Newton methods for optimization with orthogonality constraints. *SIAM Journal on Scientific Computing*, 41(4):A2239–A2269, 2019.

[HLL+18] L. Huang, X. Liu, B. Lang, A. Yu, Y. Wang, and B. Li. Orthogonal weight normalization: solution to optimization over multiple dependent stiefel manifolds in deep neural networks. In *AAAI*, pages 3271–3278, 2018.

[HLWY20] J. Hu, X. Liu, Z-W. Wen, and Y-X. Yuan. A brief introduction to manifold optimization. *Journal of the Operations Research Society of China*, 8(2):199–248, 2020.

[HLZ20] Thomas Y Hou, Zhenzhen Li, and Ziyun Zhang. Fast global convergence for low-rank matrix recovery via Riemannian gradient descent with random initialization. *arXiv preprint arXiv:2012.15467*, 2020.

[HMJ+22] A. Han, B. Mishra, P. Jawanpuria, P. Kumar, and J. Gao. Riemannian Hamiltonian methods for min-max optimization on manifolds. *ArXiv Preprint: 2204.11418*, 2022.

[HMJG21] A. Han, B. Mishra, P. K. Jawanpuria, and J. Gao. On Riemannian optimization over positive definite matrices with the Bures-Wasserstein geometry. In *NeurIPS*, pages 8940–8953, 2021.

[HMWY18] J. Hu, A. Milzarek, Z. Wen, and Y. Yuan. Adaptive quadratically regularized Newton method for Riemannian optimization. *SIAM Journal on Matrix Analysis and Applications*, 39(3):1181–1207, 2018.

[Hos15] S Hosseini. Convergence of nonsmooth descent methods via Kurdyka–Łojasiewicz inequality on riemannian manifolds. *Hausdorff Center for Mathematics and Institute for Numerical Simulation, University of Bonn (2015,(INS Preprint No. 1523))*, 2015.

[HRU+17] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *NIPS*, pages 6629–6640, 2017.

[HS15] R. Hosseini and S. Sra. Matrix manifold optimization for Gaussian mixtures. In *NIPS*, pages 910–918, 2015.

[HW22] Wen Huang and Ke Wei. Riemannian proximal gradient methods. *Mathematical Programming*, 194(1):371–413, 2022.

[HWW+24] Zihao Hu, Guanghui Wang, Xi Wang, Andre Wibisono, Jacob D Abernethy, and Molei Tao. Extragradient type methods for Riemannian variational inequality problems. In *International Conference on Artificial Intelligence and Statistics*, pages 2080–2088. PMLR, 2024.

[Iva14] S. Ivanov. On Helly's theorem in geodesic spaces. *Electronic Research Announcements*, 21:109, 2014.

[JGN+17] C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan. How to escape saddle points efficiently. In *ICML*, pages 1724–1732. PMLR, 2017.

[JM18] P. Jawanpuria and B. Mishra. A unified framework for structured low-rank matrix learning. In *ICML*, pages 2254–2263. PMLR, 2018.

[JNG+21] C. Jin, P. Netrapalli, R. Ge, S. M. Kakade, and M. I. Jordan. On nonconvex optimization for machine learning: Gradients, stochasticity, and saddle points. *Journal of the ACM (JACM)*, 68(2):1–29, 2021.

[JNJ20] C. Jin, P. Netrapalli, and M. I. Jordan. What is local optimality in nonconvex-nonconcave minimax optimization? In *ICML*, pages 4880–4889. PMLR, 2020.

[JNT11] A. Juditsky, A. Nemirovski, and C. Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011.

[Kak41] S. Kakutani. A generalization of Brouwer's fixed point theorem. *Duke Mathematical Journal*, 8(3):457–459, 1941.

[KJM19] H. Kasai, P. Jawanpuria, and B. Mishra. Riemannian adaptive stochastic gradient algorithms on matrix manifolds. In *ICML*, pages 3262–3271, 2019.

[KKM29] B. Knaster, C. Kuratowski, and S. Mazurkiewicz. Ein beweis des fixpunktsatzes für n-dimensionale simplexe. *Fundamenta Mathematicae*, 14(1):132–137, 1929.

[KLL22] G. Kotsalis, G. Lan, and T. Li. Simple and optimal methods for stochastic variational inequalities, I: operator extrapolation. *SIAM Journal on Optimization*, 32(3):2041–2073, 2022.

[KLMT24] Pham Duy Khanh, Hoang-Chau Luong, Boris S Mordukhovich, and Dat Ba Tran. Fundamental convergence analysis of sharpness-aware minimization. *arXiv preprint arXiv:2401.08060*, 2024.

[KM18] H. Kasai and B. Mishra. Inexact trust-region algorithms on Riemannian manifolds. In *NeurIPS*, pages 4249–4260, 2018.

[KM21] W. Kong and R. D. C. Monteiro. An accelerated inexact proximal point method for solving nonconvex-concave min-max problems. *SIAM Journal on Optimization*, 31(4):2558–2585, 2021.

[KMT22] Pham Duy Khanh, Boris S Mordukhovich, and Dat Ba Tran. Inexact reduced gradient methods in smooth nonconvex optimization. *arXiv preprint arXiv:2204.01806*, 2022.

[KMT23a] Pham Duy Khanh, Boris S Mordukhovich, and Dat Ba Tran. General derivative-free optimization methods under global and local lipschitz continuity of gradients. *arXiv preprint arXiv:2311.16850*, 2023.

[KMT23b] Pham Duy Khanh, Boris S Mordukhovich, and Dat Ba Tran. Inexact reduced gradient methods in nonconvex optimization. *Journal of Optimization Theory and Applications*, pages 1–41, 2023.

[KMT24a] Pham Duy Khanh, Boris S Mordukhovich, and Dat Ba Tran. Globally convergent derivative-free methods in nonconvex optimization with and without noise, 2024.

[KMT24b] Pham Duy Khanh, Boris S Mordukhovich, and Dat Ba Tran. A new inexact gradient descent method with applications to nonsmooth convex optimization. *Optimization Methods and Software*, pages 1–29, 2024.

[Kom88] H. Komiya. Elementary proof for Sion's minimax theorem. *Kodai Mathematical Journal*, 11(1):5–7, 1988.

[Kor76] G. M. Korpelevich. The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756, 1976.

[Kri14] A. Kristály. Nash-type equilibria on Riemannian manifolds: A variational approach. *Journal de Mathématiques Pures et Appliquées*, 101(5):660–688, 2014.

[KSF17] A. Kumar, P. Sattigeri, and P. T. Fletcher. Semi-supervised learning with GANs: Manifold invariance with improved inference. In *NIPS*, pages 5540–5550, 2017.

[LCD+21] X. Li, S. Chen, Z. Deng, Q. Qu, Z. Zhu, and A. Man-Cho So. Weakly convex optimization over Stiefel manifold using Riemannian subgradient-type methods. *SIAM Journal on Optimization*, 31(3):1605–1634, 2021.

[Lee12] J. Lee. *Introduction to Smooth Manifolds*, volume 218. Springer Science & Business Media, 2012.

[LFH+20] T. Lin, C. Fan, N. Ho, M. Cuturi, and M. I. Jordan. Projection robust Wasserstein distance and Riemannian optimization. In *NeurIPS*, pages 9383–9397, 2020.

[LH16] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

[LJJ20a] T. Lin, C. Jin, and M. I. Jordan. Near-optimal algorithms for minimax optimization. In *COLT*, pages 2738–2779. PMLR, 2020.

[LJJ20b] T. Lin, C. Jin, and M. I. Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *ICML*, pages 6083–6093. PMLR, 2020.

[LKO+22] J. Lee, G. Kim, M. Olfat, M. Hasegawa-Johnson, and C. D. Yoo. Fast and efficient MMD-based fair PCA via optimization over Stiefel manifold. In *AAAI*, pages 7363–7371, 2022.

[LLMM09] C. Li, G. López, and V. Martín-Márquez. Monotone vector fields and the proximal point algorithm on Hadamard manifolds. *Journal of the London Mathematical Society*, 79(3):663–683, 2009.

[LMQ23] Xiao Li, Andre Milzarek, and Junwen Qiu. Convergence of random reshuffling under the Kurdyka–Łojasiewicz inequality. *SIAM Journal on Optimization*, 33(2):1092–1120, 2023.

[LMR+20] M. Liu, Y. Mroueh, J. Ross, W. Zhang, X. Cui, P. Das, and T. Yang. Towards better understanding of adaptive gradient algorithms in generative adversarial nets. In *ICLR*, 2020.

[Loj63] S. Lojasiewicz. Une propriété topologique des sous-ensembles analytiques réels. *Les équations aux dérivées partielles*, 117:87–89, 1963.

[LRLY21] M. Liu, H. Rafique, Q. Lin, and T. Yang. First-order convergence theory for weakly-convex-weakly-concave min-max problems. *Journal of Machine Learning Research*, 22(169):1–34, 2021.

[LS19] T. Liang and J. Stokes. Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks. In *AISTATS*, pages 907–915. PMLR, 2019.

[LSW19] H. Liu, A. M-C. So, and W. Wu. Quadratic optimization with orthogonality constraint: explicit łojasiewicz exponent and linear convergence of retraction-based line-search and stochastic variance-reduced gradient methods. *Mathematical Programming*, 178(1-2):215–262, 2019.

[LTHC20] S. Lu, I. Tsaknakis, M. Hong, and Y. Chen. Hybrid block successive approximation for one-sided non-convex min-max problems: Algorithms and applications. *IEEE Transactions on Signal Processing*, 68:3676–3691, 2020.

[LZC+21] T. Lin, Z. Zheng, E. Chen, M. Cuturi, and M. I. Jordan. On projection robust optimal transport: Sample complexity and model misspecification. In *AISTATS*, pages 262–270. PMLR, 2021.

[Mar70] B. Martinet. Régularisation d'inéquations variationnelles par approximations successives. rev. française informat. *Recherche Opérationnelle*, 4:154–158, 1970.

[MGD+21] J. Mei, Y. Gao, B. Dai, C. Szepesvari, and D. Schuurmans. Leveraging non-uniformity in first-order non-convex optimization. In *ICML*, pages 7555–7564. PMLR, 2021.

[MLZ+19] P. Mertikopoulos, B. Lecouat, H. Zenati, C-S. Foo, V. Chandrasekhar, and G. Piliouras. Optimistic mirror descent in saddle-point problems: Going the extra(-gradient) mile. In *ICLR*, 2019.

[MOP20] A. Mokhtari, A. Ozdaglar, and S. Pattathil. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. In *AISTATS*, pages 1497–1507. PMLR, 2020.

[MRS20] E. Mazumdar, L. J. Ratliff, and S. S. Sastry. On gradient-based learning in continuous games. *SIAM Journal on Mathematics of Data Science*, 2(1):103–131, 2020.

[MS18] P. Mertikopoulos and W. H. Sandholm. Riemannian game dynamics. *Journal of Economic Theory*, 177:315–364, 2018.

[Nem04] A. 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.

[Neu28] J. V. Neumann. Zur theorie der gesellschaftsspiele. *Mathematische Annalen*, 100(1):295–320, 1928.

[NJLS09] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

[NSH+19] M. Nouiehed, M. Sanjabi, T. Huang, J. D. Lee, and M. Razaviyayn. Solving a class of non-convex min-max games using iterative first order methods. In *NeurIPS*, pages 14934–14942, 2019.

[NW99] Jorge Nocedal and Stephen J Wright. *Numerical Optimization*. Springer, New York, 1999.

[OLR21] D. M. Ostrovskii, A. Lowy, and M. Razaviyayn. Efficient search of first-order Nash equilibria in nonconvex-concave smooth min-max problems. *SIAM Journal on Optimization*, 31(4):2508–2538, 2021.

[Par19] S. Park. Riemannian manifolds are KKM spaces. *Advances in the Theory of Nonlinear Analysis and its Application*, 3(2):64–73, 2019.

[Ped04] Pablo Pedregal. *Introduction to Optimization*, volume 46. Springer, New York, 2004.

[Pet06] P. Petersen. *Riemannian Geometry*, volume 171. Springer, 2006.

[PFA06] X. Pennec, P. Fillard, and N. Ayache. A Riemannian framework for tensor computing. *International Journal of Computer Vision*, 66(1):41–66, 2006.

[PJ92] B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

[Pol63] B. T. Polyak. Gradient methods for minimizing functionals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 3(4):643–653, 1963.

[PS20] Courtney Paquette and Katya Scheinberg. A stochastic line search method with expected complexity analysis. *SIAM Journal on Optimization*, 30(1):349–376, 2020.

[RM15] G. Raskutti and S. Mukherjee. The information geometry of mirror descent. *IEEE Transactions on Information Theory*, 61(3):1451–1457, 2015.

[Roc76] R. T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14(5):877–898, 1976.

[Rup88] D. Ruppert. Efficient estimations from a slowly convergent Robbins-Monro process. Technical report, Cornell University Operations Research and Industrial Engineering, 1988.

[RZS+18] S. Reddi, M. Zaheer, S. Sra, B. Poczos, F. Bach, R. Salakhutdinov, and A. Smola. A generic approach for escaping saddle points. In *AISTATS*, pages 1233–1242. PMLR, 2018.

[SFF19] Y. Sun, N. Flammarion, and M. Fazel. Escaping from saddle points on Riemannian manifolds. In *NeurIPS*, pages 7274–7284, 2019.

[SH15] S. Sra and R. Hosseini. Conic geometric optimization on the manifold of positive definite matrices. *SIAM Journal on Optimization*, 25(1):713–739, 2015.

[SH16] S. Sra and R. Hosseini. Geometric optimization in machine learning. In *Algorithmic Advances in Riemannian Geometry and Applications*, pages 73–91. Springer, 2016.

[Sio58] M. Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171–176, 1958.

[Smi14] Steven Thomas Smith. Optimization techniques on Riemannian manifolds. *arXiv preprint arXiv:1407.5965*, 2014.

[SQW16a] J. Sun, Q. Qu, and J. Wright. Complete dictionary recovery over the sphere I: Overview and the geometric picture. *IEEE Transactions on Information Theory*, 63(2):853–884, 2016.

[SQW16b] J. Sun, Q. Qu, and J. Wright. Complete dictionary recovery over the sphere II: Recovery by Riemannian trust-region method. *IEEE Transactions on Information Theory*, 63(2):885–914, 2016.

[TAV12] R. Tron, B. Afsari, and R. Vidal. Riemannian consensus for manifolds with bounded curvature. *IEEE Transactions on Automatic Control*, 58(4):921–934, 2012.

[TFBJ18] N. Tripuraneni, N. Flammarion, F. Bach, and M. I. Jordan. Averaging stochastic gradient descent on Riemannian manifolds. In *COLT*, pages 650–687, 2018.

[TJNO19] K. K. Thekumprampil, P. Jain, P. Netrapalli, and S. Oh. Efficient algorithms for smooth minimax optimization. In *NeurIPS*, pages 12680–12691, 2019.

[Van13] B. Vandereycken. Low-rank matrix completion by Riemannian optimization. *SIAM Journal on Optimization*, 23(2):1214–1236, 2013.

[VGFL+20] E. V. Vlatakis-Gkaragkounis, L. Flokas, T. Lianeas, P. Mertikopoulos, and G. Piliouras. No-regret learning and mixed Nash equilibria: They do not mix. In *NeurIPS*, pages 1380–1391, 2020.

[VGFP19] E. V. Vlatakis-Gkaragkounis, L. Flokas, and G. Piliouras. Poincaré recurrence, cycles and spurious equilibria in gradient-descent-ascent for non-convex non-concave zero-sum games. In *NeurIPS*, pages 10450–10461, 2019.

[VGFP21] E. V. Vlatakis-Gkaragkounis, L. Flokas, and G. Piliouras. Solving min-max optimization with hidden structure via gradient descent ascent. In *NeurIPS*, pages 2373–2386, 2021.

[WCCL16] Ke Wei, Jian-Feng Cai, Tony F Chan, and Shingyu Leung. Guarantees of Riemannian optimization for low rank matrix recovery. *SIAM Journal on Matrix Analysis and Applications*, 37(3):1198–1222, 2016.

[Wie12] A. Wiesel. Geodesic convexity and covariance estimation. *IEEE Transactions on Signal Processing*, 60(12):6182–6189, 2012.

[WLMML10] J. H. Wang, G. López, V. Martín-Márquez, and C. Li. Monotone and accretive vector fields on Riemannian manifolds. *Journal of Optimization Theory and Applications*, 146(3):691–708, 2010.

[WY13] Z. Wen and W. Yin. A feasible method for optimization with orthogonality constraints. *Mathematical Programming*, 142(1-2):397–434, 2013.

[ZHZ21] J. Zhang, M. Hong, and S. Zhang. On lower iteration complexity bounds for the convex concave saddle point problems. *Mathematical Programming*, pages 1–35, 2021.

[ZMZ20] J. Zhang, S. Ma, and S. Zhang. Primal-dual optimization algorithms over Riemannian manifolds: An iteration complexity analysis. *Mathematical Programming*, 184(1):445–490, 2020.

[ZRS16] H. Zhang, S. J. Reddi, and S. Sra. Riemannian SVRG: Fast stochastic optimization on Riemannian manifolds. In *NeurIPS*, pages 4592–4600, 2016.

[ZS16] H. Zhang and S. Sra. First-order methods for geodesically convex optimization. In *COLT*, pages 1617–1638. PMLR, 2016.

[ZZS22] P. Zhang, J. Zhang, and S. Sra. Minimax in geodesic metric spaces: Sion's theorem and algorithms. *ArXiv Preprint: 2202.06950*, 2022.

A Motivating Examples on Riemannian Min-Max Optimization

We provide some examples of Riemannian min-max optimization to give a sense of their expressivity. Two of the examples are the generic models from the optimization literature [BTEGN09, AMS09, HLWY20] and the two others are the formulations of application problems arising from machine learning and data analytics [PFA06, FJ07, LFH+20].

Example 1 (Riemannian optimization with nonlinear constraints)**.** We can consider a rather straightforward generalization of constrained optimization problem from Euclidean spaces to Riemannian manifolds [BH19]. This formulation finds a wide range of real-world applications, e.g., non-negative principle component analysis, weighted max-cut and so on. Letting $\mathcal{M}$ be a finite-dimensional Riemannian manifold with unique geodesic, we focus on the following problem:

$$\min_{x\in\mathcal{M}} \ f(x), \quad \text{s.t. } g(x)\leq 0,\ h(x)=0$$

where $g:=(g_1,g_2,\ldots,g_m):\mathcal{M}\mapsto\mathbb{R}^m$ and $h:=(h_1,h_2,\ldots,h_n):\mathcal{M}\mapsto\mathbb{R}^n$ are two mappings. Then, we can introduce the dual variables λ and μ and reformulate the aforementioned constrained optimization problem as follows,

$$\min_{x\in\mathcal{M}}\max_{(\lambda,\mu)\in\mathbb{R}^m_+\times\mathbb{R}^n} \ f(x)+\langle\lambda,g(x)\rangle+\langle\mu,h(x)\rangle$$

Suppose that f and all of g_i and h_i are geodesically convex and smooth, the above problem is a geodesic-convex-Euclidean-concave min-max optimization problem.

Example 2 (Distributionally robust Riemannian optimization)**.** Distributionally robust optimization (DRO) is an effective method to deal with the noisy data, adversarial data, and imbalanced data. We consider the problem of DRO over Riemannian manifold; indeed, given a set of data samples $\{\xi_i\}_{i=1}^N$, the problem of DRO over Riemannian manifold $\mathcal{M}$ can be written in the form of

$$\min_{x\in\mathcal{M}}\max_{\mathbf{p}\in\mathcal{S}} \ \sum_{i=1}^N p_i\ell(x;\xi_i)-\left\|\mathbf{p}-\frac{1}{N}\mathbf{1}\right\|^2$$

where $\mathbf{p}=(p_1,p_2,\ldots,p_N)$ and $\mathcal{S}=\{\mathbf{p}\in\mathbb{R}^N:\sum_{i=1}^N p_i=1,p_i\geq 0\}$. In general, $\ell(x;\xi_i)$ denotes the loss function over Riemannian manifold $\mathcal{M}$. If ℓ is geodesically convex and smooth, the above problem is a geodesic-convex-Euclidean-concave min-max optimization problem.

Example 3 (Robust matrix Karcher mean problem)**.** We consider a robust version of classical matrix Karcher mean problem. More specifically, the Karcher mean of N symmetric positive definite matrices $\{A_i\}_{i=1}^N$ is defined as the matrix $X\in\mathcal{M}=\{X\in\mathbb{R}^{n\times n}:X\succ 0,\ X=X^\top\}$ that minimizes the sum of squared distance induced by the Riemannian metric:

$$d(X,Y)=\|\log(X^{-1/2}YX^{-1/2})\|_F$$

The loss function is thus defined by

$$f(X;\{A_i\}_{i=1}^N)=\sum_{i=1}^N(d(X,A_i))^2$$

which is known to be nonconvex in Euclidean spaces but geodesically strongly convex. Then, the robust version of classical matrix Karcher mean problem is aiming at solving the following problem:

$$\min_{X\in\mathcal{M}}\max_{Y_i\in\mathcal{M}} \ f(X;\{Y_i\}_{i=1}^N)-\gamma\left(\sum_{i=1}^N(d(Y_i,A_i))^2\right)$$

where $\gamma > 0$ stands for the trade-off between the computation of Karcher mean over a set of $\{Y_i\}_{i=1}^N$ and the difference between the observed samples $\{A_i\}_{i=1}^N$ and $\{Y_i\}_{i=1}^N$. It is clear that the above problem is a geodesically strongly-convex-strongly-concave min-max optimization problem.

Example 4 (Projection robust optimal transport problem)**.** We consider the projection robust optimal transport (OT) problem – a robust variant of the OT problem – that achieves superior sample complexity bound [LZC$^+$21]. Let $\{x_1, x_2, \ldots, x_n\} \subseteq \mathbb{R}^d$ and $\{y_1, y_2, \ldots, y_n\} \subseteq \mathbb{R}^d$ denote sets of n atoms, and let $(r_1, r_2, \ldots, r_n)$ and $(c_1, c_2, \ldots, c_n)$ denote weight vectors. We define discrete probability measures $\mu = \sum_{i=1}^n r_i \delta_{x_i}$ and $\nu = \sum_{j=1}^n c_j \delta_{y_j}$. In this setting, the computation of the k-dimensional projection robust OT distance between μ and ν resorts to solving the following problem:

$$\max_{U \in \mathrm{St}(d,k)} \min_{\pi \in \Pi(\mu,\nu)} \sum_{i=1}^n \sum_{j=1}^n \pi_{i,j} \|U^\top x_i - U^\top y_j\|^2$$

where $\mathrm{St}(d,k) = \{U \in \mathbb{R}^{d\times k} \mid U^\top U = I_k\}$ is a Stiefel manifold and $\Pi(r,c) = \{\pi \in \mathbb{R}_+^{n\times n} \mid \sum_{j=1}^n \pi_{ij} = r_i, \sum_{i=1}^n \pi_{ij} = c_j\}$ is a transportation polytope. It is worth mentioning that the above problem is a geodesically-nonconvex-Euclidean-concave min-max optimization problem with special structures, making the computation of stationary points tractable. While the global convergence guarantee for our algorithm does not apply, the above problem might be locally geodesically-convex-Euclidean-concave such that our algorithm with sufficiently good initialization works here.

In addition to these examples, it is worth mentioning that Riemannian min-max optimization problems contain all general min-max optimization problems in Euclidean spaces and all Riemannian minimization or maximization optimization problems. It is also an abstraction of many machine learning problems, e.g., principle component analysis [BA11], dictionary learning [SQW16a, SQW16b], deep neural networks (DNNs) [HLL$^+$18] and low-rank matrix learning [Van13, JM18]; indeed, the problem of principle component analysis resorts to optimization problems on Grassmann manifolds for example.

B More Related Works on Riemannian Min-Max Optimization

The literature for the geometric properties of Riemannian Manifolds is immense and hence we cannot hope to survey them here; for an appetizer, we refer the reader to [BBB$^+$01] and [Lee12] and references therein. On the other hand, as stated, it is not until recently that the long-run non-asymptotic behavior of optimization algorithms in Riemannian manifolds (even the smooth ones) has encountered a lot of interest. For concision, we have deferred here a detailed exposition of the rest of recent results to Section B of the section. Additionally, in Section A we also give a bunch of motivating examples which can be solved by Riemannian min-max optimization.

Minimization on Riemannian manifolds. Many application problems can be formulated as the minimization or maximization of a smooth function over Riemannian manifold and has triggered a line of research on the extension of the classical first-order and second-order methods to Riemannian setting with asymptotic convergence to first-order stationary points in general [AMS09]. Recent years have witnessed the renewed interests on nonasymptotic convergence analysis of solution methods. In particular, [BAC19a] proved the global sublinear convergence results for Riemannian gradient descent method and Riemannian trust region method, and further demonstrated that the Riemannian trust region method converges to a second-order stationary point in polynomial time;

see also similar results in some other works [KM18, HMWY18, HJL+19]. We are also aware of recent works on problem-specific methods [WY13, GLCY18, LSW19] and primal-dual methods [ZMZ20].

Compared to the smooth counterpart, Riemannian nonsmooth optimization is harder and relatively less explored [AH19]. A few existing works focus on optimizing geodesically convex functions over Riemannian manifold with subgradient methods [FO98, ZS16, BFM17]. In particular, [FO98] provided the first asymptotic convergence result while [ZS16] and [BFM17] proved an nonasymptotic global convergence rate of $\mathcal{O}(\epsilon^{-2})$ for Riemannian subgradient methods. Further, [FO02] assumed that the proximal mapping over Riemannian manifold is computationally tractable and proved the global sublinear convergence of Riemannian proximal point method. Focusing on optimization over Stiefel manifold, [CMSZ20] studied the composite objective function and proposed Riemannian proximal gradient method which only needs to compute the proximal mapping of nonsmooth component function over the tangent space of Stiefel manifold. [LCD+21] consider optimizing a weakly convex function over Stiefel manifold and proposed Riemannian subgradient methods that drive a near-optimal stationarity measure below ϵ within the number of iterations bounded by $O(\epsilon^{-4})$.

There are some results on stochastic optimization over Riemannian manifold. In particular, [Bon13] proved the first asymptotic convergence result for Riemannian stochastic gradient descent, which is extended by a line of subsequent works [ZRS16, TFBJ18, BG19, KJM19]. If the Riemannian Hessian is not positive definite, some recent works have suggested frameworks to escape saddle points [SFF19, CB19].

Min-Max optimization in Euclidean spaces. Focusing on solving specifically min-max problems, the algorithms under euclidean geometry have a very rich history in optimization that goes back at least to the original proximal point algorithms [Mar70, Roc76] for variational inequality (VI) problems; At a high level, if the objective function is Lipschitz and strictly convex-concave, the simple forward-backward schemes are known to converge – and if combined with a Polyak–Ruppert averaging scheme [Rup88, PJ92, NJLS09], they achieve an $O(1/\epsilon^2)$ complexity[2] without the caveat of strictness [BC11]. If, in addition, the objective admits Lipschitz continuous gradients, then the extragradient (EG) algorithm [Kor76] achieves trajectory convergence without strict monotonicity requirements, while the time-average iterate converges at $O(1/\epsilon)$ steps [Nem04]. Finally, if the problem is strongly convex-concave, forward-backward methods computes an ϵ-saddle point at $O(1/\epsilon)$ steps; and if the operator is also Lipschitz continuous, classical results in operator theory show that simple forward-backward methods suffice to achieve a linear convergence rate [FP07, BC11].

Min-Max optimization on Riemannian manifolds. In the case of nonlinear geometry, the literature has been devoted on two different orthogonal axes: *a)* the existence of saddle point for min-max objective bi-functions and *b)* the design of algorithms for the computation of such points. For the existence of saddle point, a long line of recent work tried to generalize the seminal minima theorem for quasi-convex-quasi-concave problems of [Sio58]. The crucial bottleneck of this generalization to Riemannian smooth manifolds had been the application of both Knaster–Kuratowski–Mazurkiewicz (KKM) theorem and Helly's theorem in non-flat spaces. Before [ZZS22], the existence of saddle points had been identified for the special case of Hadamard manifolds [Kom88, Kri14, BFM17, Par19].

[2]For the entire presentation, we adopt the convention of presenting the *fine-grained complexity* performance measure for computing an $O(\epsilon)$-close solution instead of the *convergence rate* of a method. Thus a rate of the form $\|\mathbf{x}_t - \mathbf{x}^*\| \leq O(1/t^{1/p})$ typically corresponds to $O(1/\epsilon^p)$ gradient computations and the geometric rate $\|\mathbf{x}_t - \mathbf{x}^*\| \leq O(\exp(-\mu t))$ matches usually up with the $O(\ln(1/\epsilon))$ computational complexity.

Similar with the existence results, initially the developed methods referred to the computation of singularities in monotone variational operators typically in hyperbolic Hadamard manifolds with negative curvature [LLMM09]. More recently, [HGH20] proposed a Riemannian gradient descent ascent method (RGDA), yet the analysis is restricted to $\mathcal{N}$ being a convex subset of the Euclidean space and $f(x, y)$ being strongly concave in y. It is worth mentioning that for the case Hadamard and generally hyperbolic manifolds, extra-gradient style algorithms have been proposed [WLMML10, FPN05] in the literature, establishing mainly their asymptotic convergence. However it was not until recent [ZZS22] that the riemannian correction trick has been analyzed for the case of the extra-gradient algorithm. Bearing in our mind the higher-order methods, [HMJ$^+$22] has recently proposed the Riemannian Hamiltonian Descent and versions of Newton's method for for geodesic convex geodesic concave functions. Since in this work, we focus only on first-order methods, we don't compare with the aforementioned Hamiltonian alternative since it incorporates always the extra computational burden of second-derivatives and hessian over a manifold.