Accelerated Methods for Riemannian Min-Max Optimization Ensuring Bounded Geometric Penalties

David Martínez-Rubio[*1,2,4] **Christophe Roux**[*1,2] **Christopher Criscitiello**[3] **Sebastian Pokutta**[1,2]
[1]Zuse Institute Berlin, Germany, [2]Technical University Berlin, Germany,
[3]Swiss Federal Technology Institute of Lausanne, Switzerland, [4]Carlos III University of Madrid, Spain

Abstract

In this work, we study optimization problems of the form $\min_x \max_y f(x, y)$, where $f(x, y)$ is defined on a product Riemannian manifold $\mathcal{M} \times \mathcal{N}$ and is μ_x-strongly geodesically convex (g-convex) in x and μ_y-strongly g-concave in y, for $\mu_x, \mu_y \geqslant 0$. We design accelerated methods when f is (L_x, L_y, L_{xy})-smooth and $\mathcal{M}$, $\mathcal{N}$ are Hadamard. To that aim we introduce new g-convex optimization results, of independent interest: we show global linear convergence for metric-projected Riemannian gradient descent and improve existing accelerated methods by reducing geometric constants. Additionally, we complete the analysis of two previous works applying to the Riemannian min-max case by removing an assumption about iterates staying in a pre-specified compact set.

1 Introduction

A wide array of recently developed machine learning methods can be phrased as min-max optimization problems. This has led to a renewed interest in optimization methods for min-max algorithms (Gidel et al., 2019; Mokhtari, Ozdaglar, and Pattathil, 2020b; Mokhtari, Ozdaglar, and Pattathil, 2020a; Lin, Jin, and Jordan, 2020; Wang and Li, 2020). Applications include generative adversarial networks (Goodfellow, Pouget-Abadie, et al., 2014), and adversarial as well as distributionally robust classifiers (Goodfellow, Shlens, and Szegedy, 2015; Duchi and Namkoong, 2018), among others. In this work, we study a class of min-max problems over Riemannian manifolds.

Riemannian optimization, the study of optimizing functions defined over Riemannian manifolds, is motivated by the following two reasons. First, some constrained optimization problems can be expressed as unconstrained optimization problems over Riemannian manifolds. And second, some non-convex Euclidean problems such as operator scaling (Allen-Zhu et al., 2018), can be rephrased as geodesically convex (g-convex) problems on Riemannian manifolds, which means global minima can be found efficiently despite nonconvexity. Geodesic convexity is a generalization of convexity to Riemannian manifolds. Some examples of machine learning tasks which can be phrased as g-convex, g-concave min-max problems are the robust matrix Karcher mean, distributionally robust linear quadratic control, constrained g-convex optimization on manifolds via the augmented Lagrangian, and more generally the distributionally robust version of any finite-sum, g-convex optimization problem, cf. (Zhang, Zhang, and Sra, 2022; Jordan, Lin, and Vlatakis-Gkaragkounis, 2022; Taskesen et al., 2023). Other related Riemannian min-max problems, which do not satisfy the g-convex, g-concave assumption include projection-robust optimal transport, decentralized PCA and geometry-aware robust PCA (Jiang and Liu, 2022; Gemp et al., 2021; Zhang, Zhang, and Sra, 2022).

Another motivation for studying the g-convex setting is its potential to shed light on the non-g-convex case. Indeed, in Euclidean optimization, a deep understanding of convex problems has led to optimal methods for approximating stationary points. In fact a variety of these algorithms run convex methods as subroutines (Jin et al., 2017; Carmon et al., 2017; Li and Lin, 2022).

Main results. Previous works on Riemannian min-max optimization (Zhang, Zhang, and Sra, 2022; Jordan, Lin, and Vlatakis-Gkaragkounis, 2022) proposed and analyzed a Riemannian version of the extragradient method. These important works have two major limitations:

Most notations in this work have a link to their definitions, using this code. For example, if you click or tap on any instance of $\mathrm{Exp}_x(\cdot)$, you will jump to the place where it is defined as the exponential map of a Riemannian manifold.

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

(a) They assume that the smoothness constants in x and y are similar, i.e., they consider the (L, L, L)-smooth and (μ, μ)-strongly convex case. However, in applications, often the smoothness constants in x and y differ, and this can be exploited to achieve faster algorithms.

(b) They rely on the assumption that the iterates of their algorithms stay in some compact set specified *a priori* but without a mechanism to enforce such constraints[1]. This property is key to bound geometric constants required to choose the algorithm parameters, as these grow with increasing distances between the iterates and a saddle point. Also, for some problems it is necessary to enforce in-manifold constraints, such as in Taskesen et al. (2023), which is not possible with previous algorithms.

We address both of these limitations, ensuring bounded geometric penalties. Our main contribution is a Riemannian Accelerated Min-Max Algorithm (RAMMA), which is able to exploit the fine-grained assumptions of (L_x, L_y, L_{xy})-smoothness and (μ_x, μ_y)-strong g-convexity, addressing limitation (a), and enforces the iterates to stay in a predefined compact set via projections, addressing limitation (b). RAMMA works by reducing the strongly g-convex, strongly g-concave (SCSC) min-max problem to a sequence of strongly g-convex minimization problems. Also, via reductions to the SCSC case, RAMMA can be used to solve g-convex, g-concave (CC) and strongly g-convex, g-concave (SCC) min-max problems, obtaining the accelerated rate on ε in the SCC case for the first time on Riemannian manifolds.

Further, we present two new results in g-convex minimization, which are crucial for implementing the subroutines of RAMMA, but are also of independent interest.

First, we prove linear convergence of Projected Riemannian Gradient Descent (PRGD) for smooth and constrained strongly g-convex functions defined on Hadamard manifolds. The best previous analysis for PRGD (Martínez-Rubio and Pokutta, 2023) was limited to the case where the diameter of the domain is sufficiently small and a point with 0 gradient is inside of this set: our algorithm works without any of these assumptions.

Second, we introduce RiemaconAbs, an accelerated inexact proximal point algorithms for g-convex Hadamard optimization. Our new method requires a less restrictive inexactness criterion than previous methods, which is crucial for its application to RAMMA. Further, it is directly applicable to strongly g-convex functions without a reduction and improves the rates from $\widetilde{O}(\zeta/\sqrt{\lambda\mu})$ to $\widetilde{O}(\sqrt{\zeta/(\lambda\mu)}+\zeta)$ for μ-strongly g-convex problems, where $\lambda > 0$ is a proximal parameter and $\zeta \geqslant 1$ is a geometric penalty, cf. Section 1.1.

To further address limitation (b), we show that, with the right choice of step sizes, the algorithms from Zhang, Zhang, and Sra (2022); Jordan, Lin, and Vlatakis-Gkaragkounis (2022) do stay in a ball around the global saddle point, whose radius is two times the initial distance to the saddle. Table 1 provides a detailed comparison between our results and prior work. Lastly, we provide a general version of Sion's theorem which holds on Riemannian manifolds, and ensures the existence of a saddle point $f(x^*, y^*) = \max_{y\in\mathcal{Y}} \min_{x\in\mathcal{X}} f(x, y) = \min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} f(x, y)$ under weak assumptions. In particular $\mathcal{X}$ and $\mathcal{Y}$ are not required to be compact in, for example, the strongly g-convex, strongly g-concave case.

1.1 Preliminaries: Definitions and Notations

The following definitions in Riemannian geometry cover the concepts used in this work, cf. (Petersen, 2006; Bacák, 2014). A Riemannian manifold $(\mathcal{M}, \mathfrak{g})$ is a real C^∞ manifold $\mathcal{M}$ equipped with a metric $\mathfrak{g}$, which is a smoothly varying inner product. For $x \in \mathcal{M}$, denote by $T_x\mathcal{M}$ the tangent space of $\mathcal{M}$ at x. For vectors $v, w \in T_x\mathcal{M}$, we use $\langle v, w\rangle_x$ for the inner product of the metric, $\|v\|_x \stackrel{\text{def}}{=} \sqrt{\langle v, v\rangle_x}$ for the corresponding norm, and we omit x when it is clear from context. A geodesic of length ℓ is a curve $\gamma : [0, \ell] \to \mathcal{M}$ of unit speed that is locally distance minimizing.

A set $\mathcal{X}$ is said to be g-convex if every two points are connected by a geodesic that remains in $\mathcal{X}$. The set $\mathcal{X}$ is said to be uniquely geodesic if every two points are connected by one and only one geodesic. The exponential map $\mathrm{Exp}_x : T_x\mathcal{M} \to \mathcal{M}$ takes a point $x \in \mathcal{M}$, and a vector $v \in T_x\mathcal{M}$ and returns the point y we obtain from following the geodesic from x in the direction v for length $\|v\|$, if this is possible. We denote its inverse by $\mathrm{Log}_x(\cdot)$, which is well defined for uniquely geodesic manifolds, so we have $\mathrm{Exp}_x(v) = y$ and $\mathrm{Log}_x(y) = v$. Note $d(x, y) = \|\mathrm{Log}_x(y)\|$. We exclusively work in uniquely geodesic manifolds, such as an open hemisphere. The manifold $\mathcal{M}$ comes with a natural parallel transport of vectors between tangent spaces, that formally is defined from a way of identifying nearby tangent spaces, known as the Levi-Civita connection ∇ (Levi-Civita, 1977). Throughout this work, we use $\Gamma_y^x(v) \in T_x\mathcal{M}$ to denote this parallel transport for a vector v in T_yM from $T_y\mathcal{M}$ to $T_x\mathcal{M}$ along the unique geodesic that connects y to x. We write $\Gamma^x(v)$ if y is clear from context. As for all of the related works in Section 1.3, we assume that we can compute the $\mathrm{Exp}_x(\cdot)$, $\mathrm{Log}_x(\cdot)$ and $\Gamma^x(\cdot)$. We use $\mathrm{inj}(x)$ to denote the injectivity radius at x, that is, the largest radius r of a ball $B(0, r) \subseteq T_x\mathcal{M}$ for which Exp_x is a diffeomorphism.

We use $\bar{B}(x, R)$ to denote a closed Riemannian ball with

[1] Note that this assumption is not the same as the iterates staying in *some* compact set a posteriori.

Table 1: Summary of our results and comparisons with previous work. See Section 1.1 for notations or *click on them*. We use κ_λ for $1/(\lambda\mu)$, where λ is a proximal parameter. In column **K**, H stands for Hadamard, R stands for Riemannian manifolds. Our contributions are in gray.

Method	**Complexity**	**Notes**	**K**
	G-CONVEX		
(MP22, PRGD)	$\widetilde{O}(\zeta_D L/\mu)$	small diam. D & $\nabla f(x^*)=0$	H
(MP22, Thm. 4)	$\widetilde{O}(\zeta\sqrt{\kappa_\lambda})$	accelerated, g-convex	H
PRGD	$\widetilde{O}(\zeta_R L/\mu)$	global, $R \stackrel{\text{def}}{=} L_p(f,\mathcal{X})/L$	H
Algorithm 3	$\widetilde{O}(\sqrt{\zeta\kappa_\lambda}+\zeta)$	accelerated, g-convex	H
	MIN-MAX		
(JLV22, RCEG-SCSC)	$\widetilde{O}(\sqrt{\zeta/\delta}\cdot L/\mu + 1/\delta)$	We remove strong assumptions (see Section 3)	R
(ZZS22, RCEG-CC)	$O\left(\sqrt{\zeta/\delta}\cdot LD^2/\varepsilon\right)$		R
RAMMA-SCSC	$\widetilde{O}(\zeta^{4.5}\sqrt{\frac{L_x}{\mu_x}+\frac{\zeta L L_{xy}}{\mu_x\mu_y}+\frac{L_y}{\mu_y}+\zeta^2})$	(L_x, L_y, L_{xy})-smooth case, where $L \stackrel{\text{def}}{=} \max\{L_x, L_y, L_{xy}\}$	H
RAMMA-SCC	$\widetilde{O}(\zeta^{4.5}\sqrt{\frac{L}{\mu_x}+\frac{\zeta^2 L_{xy} L D^2}{\mu_x\varepsilon}+\frac{L_y D^2}{\varepsilon}})$		H
RAMMA-CC	$\widetilde{O}(\zeta^{4.5}\sqrt{\frac{LD^2}{\varepsilon}}+\zeta^{5.5}\sqrt{L_{xy}L}\frac{D^2}{\varepsilon})$		H

center x and radius R, and denote by $d(x,y)$ the distance between x and y, A map $\mathcal{P}_\mathcal{X} : \mathcal{M} \to \mathcal{X}$ is a projection operator if it satisfies $d(x,\mathcal{P}_\mathcal{X}(x)) \leqslant d(x,z)$ for all $z \in \mathcal{X}$. For a uniquely geodesic g-convex set $\mathcal{Z}$ a point $z \in \mathcal{Z}$ and a closed Riemannian ball $\mathcal{X} \stackrel{\text{def}}{=} \bar{B}(x,D) \subset \mathcal{Z} \subset \mathcal{M}$ we have $\mathcal{P}_\mathcal{X}(z) = z$ if $z \in \mathcal{X}$ and $\mathcal{P}_\mathcal{X}(z) = \mathrm{Exp}_x(D\mathrm{Log}_x(z)/\|\mathrm{Log}_x(z)\|)$ is a relatively cheap projection operator, cf. (Martínez-Rubio and Pokutta, 2023). As in the Euclidean space, a differentiable function is L-smooth and μ-strongly g-convex in a uniquely geodesic g-convex set $\mathcal{X}$, if for any two points $x, y \in \mathcal{X}$ we have, respectively:

$$f(y) \leqslant f(x) + \langle \nabla f(x), \mathrm{Log}_x(y)\rangle + \frac{L}{2}d^2(x,y)$$

and

$$f(y) \geqslant f(x) + \langle \nabla f(x), \mathrm{Log}_x(y)\rangle + \frac{\mu}{2}d^2(x,y).$$

The latter property is equivalent to $f(\mathrm{Exp}_x(t\cdot\mathrm{Log}_x(x) + (1-t)\cdot\mathrm{Log}_x(y))) \leqslant tf(x)+(1-t)f(y)-\frac{t(1-t)\mu}{2}d^2(x,y)$, for $t \in [0,1]$ and also applies to non-differentiable functions. The function is said to be g-convex if $\mu = 0$. Further, it is μ-strongly g-concave if $-f$ is μ-strongly convex. The function f has $\bar{L}_x$-Lipschitz gradients in $\mathcal{X}$ if for all $x, y \in \mathcal{X}$ we have $\|\nabla f(x) - \Gamma^x \nabla f(y)\| \leqslant \bar{L}_x d(x,y)$. A function f is $L_p(f,\mathcal{X})$-Lipschitz in $\mathcal{X}$ if $|f(x)-f(y)| \leqslant L_p(f,\mathcal{X})d(x,y)$ for all $x, y \in \mathcal{X}$, where $\mathcal{X}$ is omitted if clear from context. A function $f(x,y)$ is (μ_x,μ_y)-SCSC in $\mathcal{X}\times\mathcal{Y}$ if it is μ_x-strongly g-convex in $\mathcal{X}$ and μ_y-strongly g-concave in $\mathcal{Y}$. If $\mu_y = 0$ we say the function is μ_x-SCC, if it is $\mu_x = \mu_y = 0$, then it is CC.

For a function $f : \mathcal{M}\times\mathcal{N} \to \mathbb{R}$ that is CC in $\mathcal{X}\times\mathcal{Y}$, a point $(\hat{x},\hat{y}) \in \mathcal{X}\times\mathcal{Y}$ is an ε-saddle point of f in $\mathcal{X}\times\mathcal{Y}$ if $\max_{y\in\mathcal{Y}} f(\hat{x},y) - \min_{x\in\mathcal{X}} f(x,\hat{y}) \leqslant \varepsilon$, assuming the max and min exist. We define it to be an ε-saddle point *in distance* if $d^2(\hat{x},x^*) + d^2(\hat{y},y^*) \leqslant \varepsilon$, where (x^*,y^*) is a 0-saddle point in $\mathcal{X}\times\mathcal{Y}$ that satisfies $f(x^*,y^*) = \min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}} f(x,y) = \max_{y\in\mathcal{Y}}\min_{x\in\mathcal{X}} f(x,y)$, whose existence we can guarantee under mild assumptions, cf. Theorem 16. We do not specify the saddle point is taken with respect to $\mathcal{X}\times\mathcal{Y}$ when it is clear from context.

The sectional curvature of a manifold $\mathcal{M}$ at a point $x \in \mathcal{M}$ for a 2-dimensional space $V \subset T_x\mathcal{M}$ is the Gauss curvature of $\mathrm{Exp}_x(V)$ at x. Hadamard manifolds are complete simply-connected Riemannian manifolds of non-positive sectional curvature, like the hyperbolic space or the space of $n\times n$ symmetric positive definite matrices with the metric $\langle X,Y\rangle_A \stackrel{\text{def}}{=} \mathrm{Tr}(A^{-1}XA^{-1}Y)$ where X, Y are in the tangent space of A. They are uniquely geodesic and $\mathrm{Exp}_x(\cdot)$ is well defined on them for every $v \in T_x\mathcal{M}$.

Given $R > 0$, and a manifold of sectional curvature bounded in $[\kappa_{\min},\kappa_{\max}]$, we define the geometric constants $\zeta_R \stackrel{\text{def}}{=} R\sqrt{|\kappa_{\min}|}\coth(R\sqrt{|\kappa_{\min}|}) \geqslant 1$ if $\kappa_{\min} < 0$ and $\zeta_R \stackrel{\text{def}}{=} 1$ otherwise, and $\delta_R \stackrel{\text{def}}{=} R\sqrt{\kappa_{\max}}\cot(R\sqrt{\kappa_{\max}}) \leqslant 1$ if $\kappa_{\max} > 0$ and $\delta_R \stackrel{\text{def}}{=} 1$ otherwise. For a set $\mathcal{X}$ of diameter D, we use $\zeta^\mathcal{X} \stackrel{\text{def}}{=} \zeta_D = \Theta(1+\sqrt{|\kappa_{\min}|D})$, or just ζ if $\mathcal{X}$ is clear from context. For the product $\mathcal{X}\times\mathcal{Y}$, we abuse the notation and use $\zeta \stackrel{\text{def}}{=} \max\{\zeta^\mathcal{X},\zeta^\mathcal{Y}\}$. Similarly for the notation $\delta^\mathcal{X}$ and δ. These constants appear in Riemannian optimization analy-

sis via the Riemannian cosine law Lemma 10 or other similar inequalities. The big-O notation $\widetilde{O}(\cdot)$ omits log factors.

1.2 Problem Setting

In this work, $\mathcal{M}$ and $\mathcal{N}$ always represent two uniquely geodesic finite-dimensional Riemannian manifolds of sectional curvature bounded by $[\kappa_{\min}, \kappa_{\max}]$, and $\mathcal{X} \subset \mathcal{M}$, $\mathcal{Y} \subset \mathcal{N}$ are compact g-convex subsets for which we have access to projection oracles. We consider the following optimization problem

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

where $f : \mathcal{M} \times \mathcal{N} \to \mathbb{R}$ denotes a function with a saddle point at $(x^*, y^*) \in \mathcal{X} \times \mathcal{Y}$, that satisfies $\nabla f(x^*, y^*) = 0$. Let $(x_0, y_0) \in \mathcal{X} \times \mathcal{Y}$ be an initial point. Define $D \stackrel{\text{def}}{=} \max\{\operatorname{diam}(\mathcal{X}), \operatorname{diam}(\mathcal{Y})\}$. Our aim is to compute an ε-saddle point of f over $\mathcal{X} \times \mathcal{Y}$, where f satisfies the following Assumption 1, with constants $\mu_x, \mu_y, L_x, L_y, L_{xy}$. We also assume without loss of generality that $L_x = L_y$, and $\mu_y \leqslant \mu_x$. Indeed, we can rescale the manifolds to obtain $L_x = L_y$ which keeps $L_{xy}, L_x/\mu_x, L_y/\mu_y, \mu_x\mu_y$ constant, as well as geometric penalties depending on ζ, see Appendix B. Also, if $\mu_y > \mu_x$, we can work with the function $h(x, y) = -f(y, x)$. We write $L \stackrel{\text{def}}{=} \max\{L_x, L_y, L_{xy}\}$, $\kappa_x \stackrel{\text{def}}{=} L_x/\mu_x$, and $\kappa_y \stackrel{\text{def}}{=} L_y/\mu_y$. We say a function is $(\bar{L}_x, \bar{L}_y, \bar{L}_{xy})$-smooth if it satisfies Statements 2 and 3 below.

Assumption 1 *Let $\mathcal{M}$, $\mathcal{N}$, $\mathcal{X}$, $\mathcal{Y}$ be as above, and let $g : \mathcal{M} \times \mathcal{N} \to \mathbb{R}$ be differentiable. For any $(x, y) \in \mathcal{X} \times \mathcal{Y}$, it holds:*

1. *g is $(\bar{\mu}_x, \bar{\mu}_y)$-SCSC in $\mathcal{X} \times \mathcal{Y}$.*
2. *$\nabla_x g(\cdot, y)$ is $\bar{L}_x$-Lipschitz in $\mathcal{X}$ and $\nabla_y g(x, \cdot)$ is $\bar{L}_y$-Lipschitz in $\mathcal{Y}$.*
3. *$\nabla_y g(\cdot, y)$ and $\nabla_x g(x, \cdot)$ are $\bar{L}_{xy}$-Lipschitz in $\mathcal{X}$ and $\mathcal{Y}$, respectively.*

We generalize Sion's theorem (Sion, 1958) to Riemannian manifolds in Appendix C under mild assumptions, which in particular are satisfied if f is SCSC, ensuring the existence of a saddle point in this case. For Riemannian manifolds, this theorem generalizes Zhang, Zhang, and Sra (2022) which required the sets $\mathcal{X}$, $\mathcal{Y}$ to be compact.

1.3 Related Work

Euclidean min-max optimization. The extragradient (EG) algorithm is an optimal algorithm for solving min-max problems achieving optimal rates of $O(L/\mu)$ for L-smooth and μ-SCSC functions. The convergence of EG was first introduced in Korpelevich (1976) under the assumption that the optimization was performed over compact sets. It was recently proven in Mokhtari, Ozdaglar, and Pattathil (2020b); Mokhtari, Ozdaglar, and Pattathil (2020a) that this compactness is not needed in either the SCSC case or the CC case. Recently, some works (Lin, Jin, and Jordan, 2020; Wang and Li, 2020) have achieved accelerated rates for more fine-grained smoothness and strong-convexity assumptions, where the constants with respect to each variable can differ, among other things, cf. Assumption 1. Lin, Jin, and Jordan (2020) introduced an algorithm with accelerated rates of $\widetilde{O}(\sqrt{\frac{L^2}{\mu_x\mu_y}})$ for (μ_x, μ_y)-SCSC and L-smooth functions. Later, Wang and Li (2020) proved accelerated rates of $\widetilde{O}(\sqrt{\frac{L_x}{\mu_x} + \frac{LL_{xy}}{\mu_x\mu_y} + \frac{L_y}{\mu_y}})$ for the more general case of (L_x, L_y, L_{xy})-smooth and (μ_x, μ_y)-SCSC functions. These accelerated methods reduce the solution of the SCSC min-max problem to a sequence of better conditioned strongly convex minimization problems carried out by variants of accelerated proximal point algorithms.

Riemannian min-max optimization. There are several works studying algorithms for solving monotone variational inequalities on compact subsets of Hadamard manifolds, which encompass CC min-max problems as a special case. The algorithms presented in these works are variations of the inexact proximal point algorithm (Bento, Ferreira, and Quiroz, 2021; Li, López, and Martín-Márquez, 2009) and the EG algorithm (Batista, Bento, and Ferreira, 2018; Ferreira, Pérez, and Németh, 2005) and come with asymptotic convergence guarantees.

Two recent works (Zhang, Zhang, and Sra, 2022; Jordan, Lin, and Vlatakis-Gkaragkounis, 2022) based on a variation of the Euclidean EG called Riemannian corrected extragradient (RCEG) have shown convergence rates for smooth, unconstrained min-max problems on Riemannian manifolds of bounded sectional curvature that match the Euclidean EG up to geometric constants for the CC (Zhang, Zhang, and Sra, 2022) and SCSC (Jordan, Lin, and Vlatakis-Gkaragkounis, 2022) case. The convergence guarantees of RCEG assume that the iterates stay in some pre-specified, compact set without any mechanism to enforce this. In contrast, the previously mentioned works (Bento, Ferreira, and Quiroz, 2021; Li, López, and Martín-Márquez, 2009; Batista, Bento, and Ferreira, 2018; Ferreira, Pérez, and Németh, 2005), treat Riemannian min-max problems over compact sets and explicitly enforce that the iterates stay inside these. Jordan, Lin, and Vlatakis-Gkaragkounis (2022) additionally analyze non-smooth min-max problems in the CC and SCSC case, as well as stochastic versions of both the smooth and non-smooth case.

Riemannian g-convex minimization. Optimization of g-convex functions with rates of convergence has been

studied more extensively than min-max problems. Zhang and Sra (2016) provided several first-order deterministic and stochastic methods applying to smooth or non-smooth problems. A long line of works have tackled the question of acceleration on Riemannian manifolds, see Martínez-Rubio and Pokutta (2023) for an overview. We improve over this work by reducing geometric penalties in the rates and allowing for a less restrictive inexactness criterion.

Furthermore, some works studied adaptive methods (Kasai, Jawanpuria, and Mishra, 2019), as well as projection-free (Weber and Sra, 2017; Weber and Sra, 2019), saddle-point-escaping (Criscitiello and Boumal, 2019; Sun, Flammarion, and Fazel, 2019; Zhou, Yuan, and Feng, 2019; Criscitiello and Boumal, 2022a), and variance-reduced methods (Sato, Kasai, and Mishra, 2019a; Sato, Kasai, and Mishra, 2019b; Zhang, Reddi, and Sra, 2016). Some lower bounds were provided by Hamilton and Moitra (2021); Criscitiello and Boumal (2022b). Analyses of methods for non-smooth problems that work with in-manifold constraints via projection oracles were provided in Zhang and Sra (2016) and further studied in other works, such as Wang, Tu, et al. (2021).

For the smooth case with in-manifold constraints, Martínez-Rubio and Pokutta (2023) discovered that the proof provided in Zhang and Sra (2016) is flawed and provided an analysis of PRGD under restrictive assumptions: the diameter D of the feasible set is required to be small enough so that $\zeta_D < 2$ and the global solution is required to be inside of the set. To the best of our knowledge, there is no other method or analysis for the smooth setting using a projection oracle. We remove both of these assumptions and show that PRGD enjoys linear convergence rates for strongly g-convex and smooth problems with a projection oracle to an arbitrary g-convex constraint.

2 Riemannian Accelerated Algorithms for Minimization and Min-Max

2.1 G-Convex Algorithms

We first present a global linear convergence analysis for Projected Riemannian Gradient Descent (PRGD) in Hadamard manifolds, an algorithm for constrained optimization of μ-strongly g-convex and L-smooth functions that makes use of a projection oracle and a gradient oracle. Several attempts have been made towards obtaining this result (Zhang and Sra, 2016; Martínez-Rubio and Pokutta, 2023), but the best analysis (Martínez-Rubio and Pokutta, 2023) only provides linear convergence of PRGD in a compact g-convex set with diameter D small enough so that $\zeta_D < 2$, for a function whose global minimizer is inside of this set. Our analysis provides linear global convergence without any of these assumptions. We believe this will unlock many new algorithms, as it was unknown how to deal with constraints in general smooth problems before.

Proposition 2 (PRGD) [↓] *Let $f : \mathcal{M} \to \mathbb{R}$ be a μ-strongly g-convex and L-smooth function in a g-convex compact subset $\mathcal{X} \subset \mathcal{M}$ of a Hadamard manifold $\mathcal{M}$. For an initial point $x_0 \in \mathcal{X}$ and $R \overset{\text{def}}{=} L_p(f, \mathcal{X})/L$, we have $f(x_T) - f(x^*) \leqslant \varepsilon$, after $T = \widetilde{O}(\zeta_R L/\mu)$ steps of Projected Riemannian Gradient Descent (PRGD) with update rule*

$$x_{t+1} \leftarrow \mathcal{P}_{\mathcal{X}}\left(\operatorname{Exp}_{x_t}\left(-\frac{1}{L}\nabla f(x_t)\right)\right),$$

we have $f(x_T) - f(x^*) \leqslant \varepsilon$.

We note that we show that at every iteration PRGD reduces the gap by a factor of $1 - \mu/(4L\zeta_{R_t})$, where $R_t \overset{\text{def}}{=} \|\nabla f(x_t)\|/L \leqslant R$, which is a quantity that does not require knowing $L_p(f, \mathcal{X})$.

We now present our results on g-convex accelerated optimization. For a μ-strongly g-convex function over a g-convex closed subset $\mathcal{X}$ of a finite-dimensional Hadamard manifold $\mathcal{M}$, we improve over the state of the art (Martínez-Rubio and Pokutta, 2023, Algorithm 1) in two important directions. First, we obtain the convergence rate from $\widetilde{O}(\sqrt{\zeta\kappa_\lambda} + \zeta)$ to $\widetilde{O}(\zeta\sqrt{\kappa_\lambda})$, where $\kappa_\lambda \overset{\text{def}}{=} 1/(\lambda\bar{\mu})$ and λ is the step size of the approximate implicit gradient descent step of the algorithm and we recall that $\zeta = \Theta(1 + \sqrt{|\kappa_{\min}|D})$. A similar improvement is obtained for the g-convex case via reductions, see Remark 22. The key idea for our improved algorithm is to work directly in the strongly g-convex case, as opposed to the g-convex case and make a careful choice of step sizes. Second, we require a less restrictive condition for the subroutine that Algorithm 3 (RiemaconAbs) uses, namely we only require to compute a minimizer of the subproblem in Line 12, $\min_{y\in\mathcal{X}}\{f(y) + \frac{1}{2\lambda}d^2(x_k, y)\}$, with *absolute* accuracy, as opposed to *relative* accuracy, i.e., accuracy proportional to the distance to the minimizer of the proximal problem, which is unknown in general. This second part is a requirement for the application to RAMMA, our min-max algorithm, and it is inspired by the analysis of the Euclidean APPA algorithm (Frostig et al., 2015).

Theorem 3 (RiemaconAbs) [↓] *Using the definitions and notation of Algorithm 3 to optimize a function f which is $\bar{\mu}$-strongly g-convex in $\mathcal{X}$, we have $f(y_T) - f(x^*) \leqslant \varepsilon$ after $T \geqslant 2\sqrt{\xi \max\{\frac{1}{\lambda\bar{\mu}}, 9\xi\}}\log_2\left(\frac{2\lambda^{-1}d^2(x_0,x^*)}{\varepsilon}\right) = \widetilde{O}(\sqrt{\zeta\kappa_\lambda} + \zeta)$ iterations. where $\kappa_\lambda \overset{\text{def}}{=} 1/(\lambda\bar{\mu})$ and λ is the step size of the approximate implicit gradient descent step of the algorithm.*

We emphasize that each iteration of Algorithm 3 (RiemaconAbs) in Theorem 3 requires a solution to a subproblem, which we can implement with PRGD, and RiemaconAbs *only* accesses f through these subproblem solutions: no other function values or gradients are asked of

f. This will prove important later on. If for an L-smooth $\bar{\mu}$-strongly g-convex function we apply RiemaconAbs with $\lambda = 1/L$ and PRGD as subroutine, we obtain the following Corollary 4.

Corollary 4 [↓] *If in addition to the assumptions from Theorem 3, f is also L-smooth in $\mathcal{X}$, Algorithm 3 with $\lambda = 1/L$ and PRGD as subroutine, yields an ε-minimizer after $\widetilde{O}(\zeta_R \zeta^{3/2}\sqrt{\kappa+\zeta})$ gradient and projection oracle calls, where $R \leqslant (L_p(f,\mathcal{X})/L + 2D_{\mathcal{X}})/\zeta$ and $D_{\mathcal{X}}$ is the diameter of $\mathcal{X}$.*

2.2 Min-Max Algorithms

Before turning to general Riemannian min-max problems, we first consider a specific class of functions $f(x,y)$ for which the interaction between x and y is weak, meaning that L_{xy} is small relative to other function parameters, i.e., the gradient of $f(x,y)$ with respect to x is only weakly dependent on y and vice versa. If the interaction between x and y is weak enough, alternating between minimizing $x \mapsto f(x, y_t)$ where y_t is kept fixed and maximizing $y \mapsto f(x_{t+1}, y)$ where x_{t+1} is kept fixed is sufficient to converge to the saddle point. The approach of computing the optimal value of x for a fixed y, or vice versa, can be seen as the *best response* of x given a fixed y, hence the name. In particular, for the case of $L_{xy} = 0$, x and y have no interaction and it suffices to independently compute the best response for x and y once to solve the min-max problem.

Our Riemannian Alternating Best Response (RABR) algorithm implements this approach by repeatedly applying approximate best responses using (Martínez-Rubio and Pokutta, 2023, Algorithm 1), cf. Appendix E.3. RABR applies only to a limited class of problems, but it will be used as a subroutine for RAMMA.

Theorem 5 (Convergence of RABR) [↓] *Let f satisfy Assumption 1 with $L_{xy} < \frac{1}{2}\sqrt{\mu_x \mu_y}$. Then Algorithm 4 requires $T = \widetilde{O}(\zeta_R \zeta^2 \sqrt{\kappa_x + \kappa_y})$ calls to the gradient and projection oracles to ensure $d^2(x_T, x^*) + d^2(y_T, y^*) \leqslant \varepsilon$, where $R = \max\{L_p(f(\cdot, y), \mathcal{X})/L_x, L_p(f(x,\cdot), \mathcal{Y})/L_y\}/\zeta + D/\zeta$.*

Now we turn to explain the intuition about our Algorithm 1 (RAMMA) for the general (L_x, L_y, L_{xy})-smooth and (μ_x, μ_y)-SCSC case. Defining $\phi(x) \overset{\text{def}}{=} \max_{y\in\mathcal{Y}} f(x,y)$, one can rephrase the problem $\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} f(x,y)$ as $\min_{x\in\mathcal{X}} \phi(x)$. By Lemma 31, if $f(\cdot, y)$ is μ_x-strongly g-convex, then $\phi(x)$ is as well. Hence, we can use RiemaconAbs to solve this minimization problem. This means that at each iteration, we need to solve the subroutine $\min_{x\in\mathcal{X}}\{\phi(x) + 1/(2\eta_x) d^2(\tilde{x}, x)\}$ (Line 12 of RiemaconAbs), which can be phrased as $\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}}\{f(x,y) + 1/(2\eta_x) d^2(\tilde{x}, x)\}$. By Theorem 16 we can exchange the min and the max, resulting in $\max_{y\in\mathcal{Y}} \min_{x\in\mathcal{X}}\{f(x,y) + \frac{1}{2\eta_x} d^2(\tilde{x}, x)\}$. Let $\psi(y) \overset{\text{def}}{=} -\max_{x\in\mathcal{X}}\{-f(x,y) - \frac{1}{2\eta_x} d^2(\tilde{x}, x)\}$, then by Lemma 31 we can interpret the latter min-max problem above as the μ_y-strongly g-convex problem $\min_{y\in\mathcal{Y}} \psi(y)$. We can solve this minimization problem via RiemaconAbs again, which involves solving a proximal step $\min_{y\in\mathcal{Y}}\{\psi(y) + 1/(2\eta_y) d^2(y, \tilde{y})\}$ at each iteration which can be phrased as the min-max problem $\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}}\{f(x,y) + 1/(2\eta_x) d^2(\tilde{x}, x) - 1/(2\eta_y) d^2(y, \tilde{y})\}$. By choosing η_x and η_y such that $L_{xy} \leqslant (4\eta_x\eta_y)^{-1/2}$, we ensure that x and y have weak interaction for this last regularized min-max problem. Hence, we reduce the original min-max problem to a series of min-max problems with weak interaction, which we can solve efficiently using RABR.

The convergence guarantee of RiemaconAbs holds for any proximal parameter $\lambda > 0$. However, when taking into account the computational cost of computing the proximal steps, the right choice of λ becomes crucial in order to ensure a good overall complexity. For RAMMA, we exploit the specific structure of the min-max problem by choosing the proximal parameters small enough such that the inner proximal problem is strongly decoupled while being large enough so that the overall computational complexity achieves the accelerated rate. Overall, the convergence rates of RAMMA are the following.

Theorem 6 (Convergence rates of RAMMA) [↓] *Consider a function $f : \mathcal{M} \times \mathcal{N} \to \mathbb{R}$ as defined in Section 1.2 with $\mu_x, \mu_y > 0$ and let $\mathcal{M}$ and $\mathcal{N}$ be Hadamard manifolds. Then, Algorithm 1 obtains an ε-saddle point after*

$$T = \widetilde{O}\left(\zeta^{4.5}\sqrt{\frac{L_x}{\mu_x} + \frac{L_y}{\mu_y} + \frac{\zeta L L_{xy}}{\mu_x \mu_y}} + \zeta^2\right).$$

calls to the gradient and projection oracles.

Now, by means of regularization, we reduce the CC and SCC cases to the SCSC case. Interestingly, we require regularization in both variables even if the function is strongly g-convex with respect to only one of them, because regularizing in both variables guarantees $d((x_0, y_0), (\hat{x}_\varepsilon^*, \hat{y}_\varepsilon^*)) \leqslant d((x_0, y_0), (x^*, y^*))$, where $(\hat{x}_\varepsilon^*, \hat{y}_\varepsilon^*)$ denotes the global saddle point of the regularized problem, and this is an important property in our analysis to reduce geometric penalties, see Remark 27.

Corollary 7 (SCC or CC to SCSC) [↓] *Let $f : \mathcal{M} \times \mathcal{N} \to \mathbb{R}$ be a function as defined in Section 1.2 and let $\mathcal{M}$ and $\mathcal{N}$ be Hadamard manifolds. Via regularization, Algorithm 1 obtains an ε-saddle point of f after*

$$T = \widetilde{O}\left(\zeta^{4.5}\sqrt{\frac{LD^2}{\varepsilon}} + \frac{\zeta^{5.5} D^2 \sqrt{L L_{xy}}}{\varepsilon}\right)$$

Algorithm 1 Riemannian Accelerated Min-Max Algorithm RAMMA$(f, (x_0, y_0), \varepsilon, \mathcal{X} \times \mathcal{Y})$

Input: Sets $\mathcal{X} \subset \mathcal{M}$, $\mathcal{Y} \subset \mathcal{N}$ that are g-convex in Hadamard manifolds $\mathcal{M}$ and $\mathcal{N}$, (μ_x, μ_y)-strongly g-convex (L_x, L_y, L_{xy})-smooth function $f : \mathcal{M} \times \mathcal{N} \to \mathbb{R}$, initial point $(x_0, y_0) \in \mathcal{X} \times \mathcal{Y}$, accuracy ε. Define $\xi \stackrel{\text{def}}{=} 4\max\{\zeta_{2D}^{\mathcal{X}}, \zeta_{2D}^{\mathcal{Y}}\} - 3 = O(\zeta)$. For T_i, $\hat{\varepsilon}_i$, see Table 2.

1: $\eta_x \leftarrow (9\xi\mu_x + \max\{L_{xy}, \mu_x\})^{-1}$, $\eta_y \leftarrow (9\xi\mu_y + \max\{L_{xy}, \mu_y\})^{-1}$, $\lambda_y \leftarrow (\max\{L_y, L_{xy}\} + 9\xi\mu_y)^{-1}$

2: $\hat{x} \leftarrow \text{RiemaconAbs}(\phi(x) \stackrel{\text{def}}{=} \max_{y\in\mathcal{Y}} f(x, y), x_0, T_1, \eta_x, \mathcal{X},$ Lines 5-8)

3: $\hat{y} \leftarrow \text{RiemaconAbs}(y \mapsto -f(\hat{x}, y), y_0, T_2, \lambda_y, \mathcal{Y}, \text{PRGD})$ ◇ One-Gap-to-Dist

4: **return** $\hat{x}, \hat{y}$

Subroutine for Line 2: With accuracy $\hat{\varepsilon}_1$, solve $\min_{x\in\mathcal{X}}\{\phi(x) + \frac{1}{2\eta_x}d^2(x_k, x)\}$ for some $x_k \in \mathcal{X}$.

5: $\eta_y \leftarrow (9\xi\mu_y + \max\{L_{xy}, \mu_y\})^{-1}$, $\widehat{\lambda} \leftarrow 1/(9\xi(\mu_x + \eta_x^{-1}) + L_x + \zeta\eta_x^{-1})$

6: $\tilde{y}_k \leftarrow \text{RiemaconAbs}(\psi(y) \stackrel{\text{def}}{=} \max_{x\in\mathcal{X}}\{-f(x, y) - \frac{1}{2\eta_x}d^2(x_k, x)\}, y_0, T_3, \eta_y, \mathcal{Y},$ Lines 9-10)

7: $\tilde{x}_k \leftarrow \text{RiemaconAbs}(x \mapsto f(x, \tilde{y}_k) + \frac{1}{2\eta_x}d^2(x_k, x), x_0, T_4, \mathcal{X}, \widehat{\lambda}, \text{PRGD})$ ◇ One-Gap-to-Dist

8: **return** $\tilde{x}_k$

Subroutine for Line 6: With accuracy $\hat{\varepsilon}_3$, solve $\min_{y\in\mathcal{Y}}\{\psi(y) + \frac{1}{2\eta_y}d^2(y_\ell, y)\}$ for some $y_\ell \in \mathcal{Y}$.

9: $\bar{x}_\ell, \bar{y}_\ell \leftarrow \text{RABR}(f(x, y) + \frac{1}{2\eta_x}d^2(x_k, x) - \frac{1}{2\eta_y}d^2(y_\ell, y), (x_0, y_0), T_5, \mathcal{X} \times \mathcal{Y})$

10: **return** $\bar{y}_\ell$

calls to the gradient and projection oracles in the CC case and

$$T = \widetilde{O}\left(\zeta^{5.5}\sqrt{\frac{L}{\mu_x} + \frac{\zeta^2 L_{xy} L D^2}{\mu_x \varepsilon} + \frac{L_y D^2}{\varepsilon}}\right)$$

calls in the μ_x-SCC case.

For the SCC case, we achieve a $\widetilde{O}(1/\sqrt{\varepsilon})$ rate in terms of ε compared to the previous state of the art rate method RCEG which achieves only a rate of $\widetilde{O}(\varepsilon^{-1})$ (Jordan, Lin, and Vlatakis-Gkaragkounis, 2022). To better understand the impact of L_{xy}, consider the case where x and y are independent, i.e., $L_{xy} = 0$. Then the min-max problem can be solved by independently minimizing and maximizing both variables. Omitting log factors and geometric terms for simplicity and setting $L_{xy} = 0$, the rates of RAMMA become $\widetilde{O}(\sqrt{L/\mu})$, $\widetilde{O}(\sqrt{L_x/\mu_x} + \sqrt{\frac{L_y D^2}{\varepsilon}})$ and $\widetilde{O}(\sqrt{\frac{LD^2}{\varepsilon}})$ for the SCSC, SCC and CC case, respectively. Recall that the accelerated rates for g-convex minimization are $\widetilde{O}(\sqrt{L/\mu})$ and $\widetilde{O}(\sqrt{\frac{LD^2}{\varepsilon}})$ in the strongly g-convex and the g-convex setting, respectively. Hence we see that the rates of RAMMA are equivalent (up to log factors and geometric terms) to applying accelerated g-convex methods to minimizing f in terms of x and maximizing f in terms of y independently. In contrast, the rates of RCEG correspond to applying unaccelerated methods for these problems.

Note that the proximal parameters $\eta_x, \eta_y, \lambda_x, \lambda_y$ as well as the precision parameters $\hat{\varepsilon}_i$ in RAMMA depend on the geometric constant ζ arising from the Riemannian cosine law Lemma 10. This geometric constant depends on the distance between the iterates and the saddle point. Thus, it is central to show that the iterates stay in a pre-specified compact set in order to bound the geometric constant ζ. We design an algorithm that enforces constraints, as opposed to guaranteeing that the iterates of the algorithm naturally stay in a set, as we did for RCEG in Section 3.

Recall that by definition of $\phi(x)$, the subproblem $\min_{x\in\mathcal{X}}\{\phi(x) + 1/(2\eta_x)d^2(x_k, x)\}$ can be phrased as $\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}}\{f(x, y) + 1/(2\eta_x)d^2(x, x_k)\}$. Let $(x_{k+1}^*, y^*(x_{k+1}^*))$ be the saddle point of this min-max problem (Theorem 16), and note that for x^*, the best response $y^*(x^*)$ for this regularized min-max problem is still y^*. Intuitively, if we do not constrain the min-max problem and the subproblems to $\mathcal{X} \times \mathcal{Y}$, RAMMA's parameters depend on the distance between its iterates and the saddle point, but this distance again depends on the parameters and so on. Denote by $\hat{x}^*$ and $\hat{y}^*$ the unconstrained optimizers. Then, if we considered an unconstrained version of our algorithm, the point $\tilde{y}_k$ computed in Line 6, would be close to the best response $\hat{y}^*(\hat{x}_{k+1}^*)$ and for this point we can only guarantee that its distance to $\hat{y}^*$ is bounded as $d(\hat{y}^*(\hat{x}_{k+1}^*), \hat{y}^*(\hat{x}^*)) \leqslant (L_{xy}/\mu_y)d(\hat{x}^*, \hat{x}_{k+1}^*)$, by using Statement 1 of Lemma 35. This would preclude acceleration because of the added extra polynomial dependency of L_{xy}/μ_y on the convergence rates via ζ, which grows with the distances between the iterates and the minimizer. For this reason, we constrain the algorithm. In order to implement a constrained algorithm, we require a linearly con-

vergent subroutine for strongly g-convex and *constrained* problems, which did not previously exist. To this end, we use our PRGD algorithm, cf. Section 2.1.

Another difficulty is that in order to apply the steps explained above, we require an accelerated algorithm for strongly g-convex optimization that bounds geometric penalties in our setting. Prior work relies on relative-accuracy proximal solutions which, for our setting, would rely on unknown quantities. Because of this reason, we designed $\mathrm{RiemaconAbs}$ that accesses f by solving proximal subproblems with absolute accuracy.

Lastly, there is a mismatch between the optimality criterion required for the proximal problems and the optimality criterion provided by the guarantees of the subroutines we used to solve them. For example, Line 5 requires computing an approximate minimizer $\tilde{x}_k$ of $\min_{x\in\mathcal{X}}\{\phi(x) + 1/(2\eta_x)d^2(x_k,x)\}$, but after Line 6 we only obtain an approximate minimizer $\tilde{y}_k$ of $\psi(y)$. In the next subsection, we present Lemma 8 which allows to solve this problem.

2.3 Converting Between Different Optimality Criteria

We now give a brief overview of the different optimality criteria in g-convex and min-max problems, before we formalize how to guarantee one criterion from another, in Lemma 8. In $\bar{L}$-smooth and $\bar{\mu}$-strongly g-convex optimization there are, among others, two well-known measures of optimality: The primal gap $\overline{\mathrm{gap}}(x) \overset{\mathrm{def}}{=} g(x) - g(x^*)$ and the squared distance to the solution $d^2(x,x^*)$, where $x^* \overset{\mathrm{def}}{=} \arg\min g(x)$ is the unique minimizer of g. By strong convexity, one can show that if a point x has a small gap, then it is also close in distance to the solution, up to some function parameters: $d^2(x,x^*) \leqslant \frac{2}{\bar{\mu}}(g(x) - g(x^*))$. In unconstrained optimization, a converse statement also holds, since $g(x) - g(x^*) \leqslant \frac{L}{2}d^2(x,x^*)$. Analogously to the Euclidean space, in optimization constrained to a g-convex closed set $\mathcal{X}$ a similar result can be obtained after a relatively cheap algorithmic computation. For the point $x' \overset{\mathrm{def}}{=} \mathcal{P}_{\mathcal{X}}(x - \frac{1}{L}\nabla f(x))$ defined as the result of one step of projected gradient descent, one can show that $\overline{\mathrm{gap}}(x') \leqslant \frac{\zeta_R L}{2}d^2(x,x^*)$, where now the gap is defined with respect to the constrained optimum $g(x) - \min_{x\in\mathcal{X}} g(x)$ and R is defined as in Proposition 2.

In the optimization of smooth and SCSC functions $g : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}$, where $\mathcal{X},\mathcal{Y}$ are g-convex closed sets, we have other notions of optimality. Define the functions $x^*(y) \overset{\mathrm{def}}{=} \arg\min_{x\in\mathcal{X}} g(x,y)$ and $y^*(x) \overset{\mathrm{def}}{=} \arg\max_{y\in\mathcal{X}}, g(x,y)$. Then we define the duality gap $\mathrm{gap}(\bar{x},\bar{y}) \overset{\mathrm{def}}{=} g(\bar{x},y^*(\bar{x})) - g(x^*(\bar{y}),\bar{y})$, the squared distance to the solution $d^2(\bar{x},x^*) + d^2(\bar{y},y^*)$, and the gap in each of the variables, i.e., $\mathrm{gap}(\bar{x}) \overset{\mathrm{def}}{=} g(\bar{x},y^*(\bar{x})) - g(x^*,y^*)$ and $\mathrm{gap}(\bar{y}) \overset{\mathrm{def}}{=} g(x^*,y^*) - g(y^*(\bar{y}),\bar{y})$. We note that $\mathrm{gap}(\bar{x},\bar{y}) = \mathrm{gap}(\bar{x}) + \mathrm{gap}(\bar{y})$ and by optimality of the points involved, all gaps are non-negative. The following lemma provides how one can relate these measures.

Lemma 8 [↓] *Let $\mathcal{X}\subset\mathcal{M}, \mathcal{Y}\subset\mathcal{N}$ be closed g-convex subsets of the Hadamard manifolds $\mathcal{M}$, $\mathcal{N}$, respectively. Let $g : \mathcal{M}\times\mathcal{N} \to \mathbb{R}$ satisfy Assumption 1 in $\mathcal{X}\times\mathcal{Y}$. The following holds:*

1. ***(Full Gap to One Gap)*** $\mathrm{gap}(\bar{x}) \leqslant \mathrm{gap}(\bar{x},\bar{y})$ *and* , $\mathrm{gap}(\bar{y}) \leqslant \mathrm{gap}(\bar{x},\bar{y})$.
2. ***(One Gap to One Dist)*** $d^2(\bar{x},x^*) \leqslant \frac{2}{\bar{\mu}_x}\mathrm{gap}(\bar{x})$, *and* $d^2(\bar{y},y^*) \leqslant \frac{2}{\bar{\mu}_y}\mathrm{gap}(\bar{y})$.
3. ***(One Gap to Dist - One Variable Optimization)*** *Suppose* $\mathrm{gap}(\bar{y}) \leqslant \varepsilon$. *If we compute an $\hat{\varepsilon}$-minimizer $\bar{x}'$ of the problem* $\min_{x\in\mathcal{X}} g(x,\bar{y})$, *then*
$$d^2(\bar{x}',x^*) + d^2(\bar{y},y^*) \leqslant \frac{4\hat{\varepsilon}}{\bar{\mu}_x} + \frac{2\varepsilon}{\bar{\mu}_y}\left(\frac{2\bar{L}_{xy}^2}{\bar{\mu}_x^2} + 1\right).$$
4. ***(Dist to Gap)*** *If in addition, $x \mapsto g(x,\bar{y})$ is $\bar{L}_p^x$-Lipschitz in $\mathcal{X}$ and $y \mapsto g(\bar{x},y)$ is $\bar{L}_p^y$-Lipschitz in $\mathcal{Y}$, we have that*
$$\begin{aligned} gap(\bar{x},\bar{y}) \leqslant\ & d(y^*,\bar{y})\left(\bar{L}_p^y + \bar{L}_p^x\frac{\bar{L}_{xy}}{\bar{\mu}_x}\right) \\ & + d(x^*,\bar{x})\left(\bar{L}_p^x + \bar{L}_p^y\frac{\bar{L}_{xy}}{\bar{\mu}_y}\right). \end{aligned}$$

Above, we showed that we can essentially guarantee any optimality criterion with another, by only increasing the accuracy by low polynomial factors depending on the problem parameters, which translates into logarithmic factors when applied to a method like our Algorithm 1. The most expensive reduction consists of going from having a low $\mathrm{gap}(x)$ or $\mathrm{gap}(y)$ to bounding the other measures Lemma 8.3, for which we require running an accelerated method on one variable. This is done in Lines 3 and 7 of Algorithm 1. The complexity of the main routine in Algorithm 1 still dominates these extra algorithmic steps.

3 Riemannian Corrected Extra-Gradient

Two previous works provide rates of convergence for Riemannian smooth min-max problems (Zhang, Zhang, and Sra, 2022; Jordan, Lin, and Vlatakis-Gkaragkounis, 2022) by using RCEG, a Riemannian adaptation of the extra-gradient method, see Algorithm 2. Specifically, Zhang, Zhang, and Sra (2022, Theorem 4.1) consider CC functions and Jordan, Lin, and Vlatakis-Gkaragkounis (2022, Theorem 3.1) consider SCSC functions. They work in the setting of Section 1.2 with $\mu = \min\{\mu_x,\mu_y\}$, and

$L = \max\{L_x, L_y, L_{xy}\}$, which is equivalent to regular gradient Lipschitzness with constant L. Possible different condition numbers or different Lipschitz assumptions are not exploited by RCEG, in contrast to our algorithm RAMMA.

The step size η of RCEG depends on the geometric constants ζ_D and $1/\delta_D$ arising from the Riemannian cosine law Lemma 10, which are larger the farther the iterates are from each other and from the global saddle. Thus, the current theory does not yield a complete algorithm unless we can guarantee that given a step size η depending on the diameter D of a set $\mathcal{X}$ specified a priori, the iterates of RCEG stay in $\mathcal{X}$. Zhang, Zhang, and Sra (2022) assumes the above occurs, and Jordan, Lin, and Vlatakis-Gkaragkounis (2022) also assumes this implicitly. Because step sizes depend on the diameter of this set, this assumption is not the same as the iterates staying in some compact set a posteriori.

The reliance on this assumption is a recurring issue in Riemannian algorithms, and multiple other works also make this assumption in other contexts, as detailed by Hosseini and Sra (2020). Prior to our work, no complete algorithm was given and the precise convergence rate was unknown, since the geometric constants were not shown to be bounded by problem parameters.

In the following Proposition 9, we show that for D depending on the initial distance to the saddle, for both the CC and SCSC cases, the iterates of RCEG stay in the closed ball $\bar{B}((x^*, y^*), D/2)$ around the saddle point, satisfying the assumption. Not relying on this assumption is one of the main difficulties in the design of our algorithm RAMMA.

Proposition 9 (RCEG) [↓] *Let $\mathcal{M}$ and $\mathcal{N}$ be uniquely geodesic Riemannian manifolds of sectional curvature bounded by $[\kappa_{\min}, \kappa_{\max}]$. Let $f : \mathcal{M} \times \mathcal{N} \to \mathbb{R}$ be a function with a saddle point at (x^*, y^*) and $\mathcal{D}^2 \stackrel{\text{def}}{=} d^2(x_0, x^*) + d^2(y_0, y^*)$. If f is (L, L, L)-smooth and (μ, μ)-SCSC in $\bar{B}((x^*, y^*), 2\mathcal{D})$, then the iterates of Algorithm 2 stay in $\bar{B}((x^*, y^*), 4\mathcal{D})$, and we obtain an ε-saddle point after $T = \widetilde{O}\left(\frac{L}{\mu}\sqrt{\frac{\zeta_D}{\delta_D}} + \frac{1}{\delta_D}\right)$ and $T = O\left(\frac{L\mathcal{D}^2}{\varepsilon}\sqrt{\frac{\zeta_D}{\delta_D}}\right)$ iterations in the μ-SCSC and the CC cases, respectively.*

Proposition 9 requires the manifolds to be uniquely geodesic, but does not impose any restrictions on their curvature bounds. Examples of such manifolds are Hadamard manifolds, or, when $\kappa_{\max} > 0$, open Riemannian balls $B(x, \hat{R})$, where $\hat{R} < \min\{\text{inj}(x)/2, \pi/(2\sqrt{\kappa_{\max}})\}$ (Chavel, 2006, Thm. IX.6.1).

4 Conclusion and Future Work

In this work, we obtained strong new algorithms for Riemannian optimization in multiple directions: The first global analysis of PRGD, an improved accelerated min method, and making use of these and other techniques we were able to obtain fast rates for min-max problems exploiting different values of smoothness and strong convexity (L_x, L_y, L_{xy}, μ_x, μ_y) and importantly, the first $\widetilde{O}(1/\sqrt{\varepsilon})$ accelerated convergence for SCC problems. None of our algorithmic results assume that the iterates will stay in some pre-specified bounded set, and consequently we can ensure we bound geometric penalties.

These new results have broad applicability to various problems, including the robust matrix Karcher mean and constrained g-convex optimization on manifolds using the augmented Lagrangian method. Notably, they enable the solution of problems with in-manifold constraints, such as distributionally robust linear quadratic control (Taskesen et al., 2023), which could not be solved using existing Riemannian min-max methods.

A promising future direction is whether our algorithms can avoid extra logarithmic factors or enjoy lower dependence on the geometric constants. Further, finding lower bounds for the min-max Riemannian case that feature additional hardness caused by the geometry remains an open challenge. An important open question, even for the Euclidean setting, is achieving rates for the general (L_x, L_y, L_{xy})-smooth and (μ_x, μ_y)-SCSC case matching lower bounds.

Acknowledgments

This research was partially funded by the Research Campus Modal funded by the German Federal Ministry of Education and Research (fund numbers 05M14ZAM,05M20ZBM) as well as the Deutsche Forschungsgemeinschaft (DFG) through the DFG Cluster of Excellence MATH$^+$ (EXC-2046/1, project ID 390685689). David Martínez-Rubio was partially funded by the project IDEA-CM (TEC-2024/COM-89) and by the DFG Cluster of Excellence MATH+ (EXC-2046/1, project id 390685689) funded by the Deutsche Forschungsgemeinschaft (DFG).

References

Allen-Zhu, Zeyuan, Ankit Garg, Yuanzhi Li, Rafael Mendes de Oliveira, and Avi Wigderson (2018). "Operator scaling via geodesically convex optimization, invariant theory and polynomial identity testing". In: *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2018, Los Angeles, CA, USA, June 25-29, 2018*. Ed. by Ilias Diakonikolas, David Kempe, and Monika Henzinger. ACM, pp. 172–181 (cit. on p. 1).

Bacák, Miroslav (2014). *Convex analysis and optimization in Hadamard spaces*. de Gruyter (cit. on p. 2).

Batista, E. E. A., G. C. Bento, and O. P. Ferreira (Apr. 2018). *An Extragradient-type Algorithm for Variational*

Inequality on Hadamard Manifolds. arXiv:1804.09292 [math] (cit. on p. 4).

Bento, GC, OP Ferreira, and EA Quiroz (2021). "An inexact proximal point method for variational inequality on Hadamard manifolds". In: *arXiv preprint arXiv:2103.02116* (cit. on p. 4).

Carmon, Yair, John C. Duchi, Oliver Hinder, and Aaron Sidford (2017). ""Convex Until Proven Guilty": Dimension-Free Acceleration of Gradient Descent on Non-Convex Functions". In: *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, pp. 654–663 (cit. on p. 1).

Chavel, Isaac (2006). *Riemannian geometry: a modern introduction*. 2nd ed. Vol. 98. Cambridge university press (cit. on p. 9).

Criscitiello, Chris and Nicolas Boumal (2019). "Efficiently escaping saddle points on manifolds". In: *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pp. 5985–5995 (cit. on p. 5).

Criscitiello, Christopher and Nicolas Boumal (2022a). "An accelerated first-order method for non-convex optimization on manifolds". In: *Foundations of Computational Mathematics*, pp. 1–77 (cit. on p. 5).

– (2022b). "Negative curvature obstructs acceleration for strongly geodesically convex optimization, even with exact first-order oracles". In: *Conference on Learning Theory, 2-5 July 2022, London, UK*. Ed. by Po-Ling Loh and Maxim Raginsky. Vol. 178. Proceedings of Machine Learning Research. PMLR, pp. 496–542 (cit. on pp. 5, 25).

Duchi, John C. and Hongseok Namkoong (2018). "Learning Models with Uniform Performance via Distributionally Robust Optimization". In: *CoRR* abs/1810.08750 (cit. on p. 1).

Ferreira, Orizon Pereira, L. R. Lucambio Pérez, and Sándor Zoltán Németh (2005). "Singularities of Monotone Vector Fields and an Extragradient-type Algorithm". In: *J. Glob. Optim.* 31.1, pp. 133–151 (cit. on p. 4).

Frostig, Roy, Rong Ge, Sham M. Kakade, and Aaron Sidford (2015). "Un-regularizing: approximate proximal point and faster stochastic algorithms for empirical risk minimization". In: *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*. Ed. by Francis R. Bach and David M. Blei. Vol. 37. JMLR Workshop and Conference Proceedings. JMLR.org, pp. 2540–2548 (cit. on p. 5).

Gemp, Ian, Brian McWilliams, Claire Vernade, and Thore Graepel (2021). "EigenGame: PCA as a Nash Equilibrium". In: *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net (cit. on p. 1).

Gidel, Gauthier, Hugo Berard, Gaëtan Vignoud, Pascal Vincent, and Simon Lacoste-Julien (2019). "A Variational Inequality Perspective on Generative Adversarial Networks". In: *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net (cit. on p. 1).

Goodfellow, Ian J., Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio (2014). "Generative Adversarial Nets". In: *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*. Ed. by Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger, pp. 2672–2680 (cit. on p. 1).

Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy (2015). "Explaining and Harnessing Adversarial Examples". In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun (cit. on p. 1).

Hamilton, Linus and Ankur Moitra (2021). "A No-go Theorem for Robust Acceleration in the Hyperbolic Plane". In: *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*. Ed. by Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, pp. 3914–3924 (cit. on p. 5).

Hartung, Joachim (1982). "An extension of Sion's minimax theorem with an application to a method for constrained games." In: *Pacific Journal of Mathematics* 103.2, pp. 401–408 (cit. on p. 15).

Hosseini, Reshad and Suvrit Sra (2020). "Recent advances in stochastic Riemannian optimization". In: *Handbook of Variational Methods for Nonlinear Geometric Data*, pp. 527–554 (cit. on p. 9).

Jiang, Bo and Ya-Feng Liu (2022). "A Riemannian exponential augmented Lagrangian method for computing the projection robust Wasserstein distance". In: *arXiv preprint arXiv:2211.13386* (cit. on p. 1).

Jin, Chi, Rong Ge, Praneeth Netrapalli, Sham M. Kakade, and Michael I. Jordan (2017). "How to Escape Saddle Points Efficiently". In: *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, pp. 1724–1732 (cit. on p. 1).

Jordan, Michael I., Tianyi Lin, and Emmanouil-Vasileios Vlatakis-Gkaragkounis (2022). "First-Order Algorithms for Min-Max Optimization in Geodesic Metric Spaces". In: *CoRR* abs/2206.02041 (cit. on pp. 1, 2, 4, 7–9, 18).

Kasai, Hiroyuki, Pratik Jawanpuria, and Bamdev Mishra (2019). "Riemannian adaptive stochastic gradient algorithms on matrix manifolds". In: *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pp. 3262–3271 (cit. on p. 5).

Kim, Jungbin and Insoon Yang (2022). "Accelerated Gradient Methods for Geodesically Convex Optimization: Tractable Algorithms and Convergence Analysis". In: *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*. Ed. by Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato. Vol. 162. Proceedings of Machine Learning Research. PMLR, pp. 11255–11282 (cit. on p. 13).

Komiya, Hidetoshi (1988). "Elementary proof for Sion's minimax theorem". In: *Kodai mathematical journal* 11.1, pp. 5–7 (cit. on p. 15).

Korpelevich, G. M. (1976). "The extragradient method for finding saddle points and other problems". en. In: *undefined* (cit. on p. 4).

Levi-Civita, Tullio (1977). *The absolute differential calculus (calculus of tensors)*. Courier Corporation. ISBN: 978-0-486-31625-3 (cit. on p. 2).

Li, Chong, Genaro López, and Victoria Martín-Márquez (2009). "Monotone vector fields and the proximal point algorithm on Hadamard manifolds". In: *Journal of the London Mathematical Society* 79.3, pp. 663–683 (cit. on p. 4).

Li, Huan and Zhouchen Lin (2022). "Restarted Nonconvex Accelerated Gradient Descent: No More Polylogarithmic Factor in the $O(\varepsilon^{-7/4})$ Complexity". In: *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*. Ed. by Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato. Vol. 162. Proceedings of Machine Learning Research. PMLR, pp. 12901–12916 (cit. on p. 1).

Lin, Tianyi, Chi Jin, and Michael I. Jordan (2020). "Near-Optimal Algorithms for Minimax Optimization". In: *Conference on Learning Theory, COLT 2020, 9-12 July 2020, Virtual Event [Graz, Austria]*. Ed. by Jacob D. Abernethy and Shivani Agarwal. Vol. 125. Proceedings of Machine Learning Research. PMLR, pp. 2738–2779 (cit. on pp. 1, 4).

Martínez-Rubio, David and Sebastian Pokutta (2023). "Accelerated Riemannian Optimization: Handling Constraints with a Prox to Bound Geometric Penalties". In: *COLT 2023, Conference on Learning Theory 12-15 July 2023, Bangalore, India* (cit. on pp. 2, 3, 5, 6, 13, 18, 24–26).

Mokhtari, Aryan, Asuman E. Ozdaglar, and Sarath Pattathil (2020a). "A Unified Analysis of Extra-gradient and Optimistic Gradient Methods for Saddle Point Problems: Proximal Point Approach". In: *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*. Ed. by Silvia Chiappa and Roberto Calandra. Vol. 108. Proceedings of Machine Learning Research. PMLR, pp. 1497–1507 (cit. on pp. 1, 4).

– (2020b). "Convergence Rate of (1/k) for Optimistic Gradient and Extragradient Methods in Smooth Convex-Concave Saddle Point Problems". In: *SIAM J. Optim.* 30.4, pp. 3230–3251 (cit. on pp. 1, 4).

Petersen, Peter (2006). *Riemannian geometry*. Ed. by S Axler and KA Ribet. Vol. 171. Springer. ISBN: 978-0-387-29403-2 (cit. on p. 2).

Sato, Hiroyuki, Hiroyuki Kasai, and Bamdev Mishra (2019a). "Riemannian Stochastic Variance Reduced Gradient Algorithm with Retraction and Vector Transport". In: *SIAM J. Optim.* 29.2, pp. 1444–1472 (cit. on p. 5).

– (2019b). "Riemannian Stochastic Variance Reduced Gradient Algorithm with Retraction and Vector Transport". In: *SIAM Journal on Optimization* 29.2, pp. 1444–1472 (cit. on p. 5).

Sion, Maurice (Mar. 1958). "On general minimax theorems." en. In: *Pacific Journal of Mathematics* 8.1, pp. 171–176. ISSN: 0030-8730, 0030-8730 (cit. on p. 4).

Sun, Yue, Nicolas Flammarion, and Maryam Fazel (2019). "Escaping from saddle points on Riemannian manifolds". In: *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pp. 7274–7284 (cit. on p. 5).

Taskesen, Bahar, Dan A. Iancu, Çagil Koçyigit, and Daniel Kuhn (2023). "Distributionally Robust Linear Quadratic Control". In: *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*. Ed. by Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (cit. on pp. 1, 2, 9).

Wang, Xi, Zhipeng Tu, Yiguang Hong, Yingyi Wu, and Guodong Shi (2021). "No-regret Online Learning over Riemannian Manifolds". In: *Advances in Neural Information Processing Systems* 34 (cit. on p. 5).

Wang, Yuanhao and Jian Li (2020). "Improved Algorithms for Convex-Concave Minimax Optimization". In: *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. Ed. by Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (cit. on pp. 1, 4).

Weber, Melanie and Suvrit Sra (2017). "Frank-Wolfe methods for geodesically convex optimization with application to the matrix geometric mean". In: *CoRR* abs/1710.10770 (cit. on p. 5).

Weber, Melanie and Suvrit Sra (2019). "Nonconvex stochastic optimization on manifolds via Riemannian Frank-Wolfe methods". In: *CoRR* abs/1910.04194 (cit. on p. 5).

Zhang, Hongyi, Sashank J. Reddi, and Suvrit Sra (2016). "Riemannian SVRG: Fast Stochastic Optimization on Riemannian Manifolds". In: *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 4592–4600 (cit. on p. 5).

Zhang, Hongyi and Suvrit Sra (2016). "First-order Methods for Geodesically Convex Optimization". In: *Proceedings of the 29th Conference on Learning Theory, COLT 2016, New York, USA, June 23-26, 2016*, pp. 1617–1638 (cit. on pp. 5, 13, 17).

Zhang, Peiyuan, Jingzhao Zhang, and Suvrit Sra (Feb. 2022). "Minimax in Geodesic Metric Spaces: Sion's Theorem and Algorithms". In: *arXiv:2202.06950 [cs, math, stat]*. arXiv: 2202.06950 (cit. on pp. 1, 2, 4, 8, 9, 15, 18).

Zhou, Pan, Xiao-Tong Yuan, and Jiashi Feng (2019). "Faster First-Order Methods for Stochastic Non-Convex Optimization on Riemannian Manifolds". In: *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, pp. 138–147 (cit. on p. 5).

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes, see Sections 1.1 and 1.2]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Not Applicable]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A Geometric Auxiliary Results

In this appendix, we use the following abuse of notation. Given points $x, y, z \in \mathcal{M}$, we write y to mean $\mathrm{Log}_x(y)$, if it is clear from context. For example, for $v \in T_x\mathcal{M}$ we have $\langle v, y - x\rangle = -\langle v, x - y\rangle = \langle v, \mathrm{Log}_x(y) - \mathrm{Log}_x(x)\rangle = \langle v, \mathrm{Log}_x(y)\rangle$; $\|v - y\| = \|v - \mathrm{Log}_x(y)\|$; $\|z - y\|_x = \|\mathrm{Log}_x(z) - \mathrm{Log}_x(y)\|$; and $\|y - x\|_x = \|\mathrm{Log}_x(y)\| = d(x, y)$.

In this section, we provide already established useful geometric results that will be used in our proofs in the sequel.

Lemma 10 (Riemannian Cosine-Law Inequalities) *For the vertices $x, y, p \in \mathcal{M}$ of a uniquely geodesic triangle of diameter D, we have*

$$\langle \mathrm{Log}_x(y), \mathrm{Log}_x(p)\rangle \geqslant \frac{\delta_D}{2} d^2(x, y) + \frac{1}{2} d^2(p, x) - \frac{1}{2} d^2(p, y).$$

and

$$\langle \mathrm{Log}_x(y), \mathrm{Log}_x(p)\rangle \leqslant \frac{\zeta_D}{2} d^2(x, y) + \frac{1}{2} d^2(p, x) - \frac{1}{2} d^2(p, y)$$

See Martínez-Rubio and Pokutta (2023) for a proof.

Remark 11 *Actually, in spaces with lower bounded sectional curvature, if we substitute the constants ζ_D in the previous Lemma 10 by the tighter constant and $\zeta_{d(p,x)}$, the result also holds. See Zhang and Sra (2016).*

The following lemmas allow us to bound functions defined in some tangent space by other functions defined in another tangent space. See Kim and Yang (2022) for a proof.

Lemma 12 *Let $x, y, p \in \mathcal{M}$ be the vertices of a uniquely geodesic triangle $\mathcal{T}$ of diameter D, and let $z^x \in T_x\mathcal{M}$, $z^y \stackrel{\text{def}}{=} \Gamma_x^y(z^x) + \mathrm{Log}_y(x)$, such that $y = \mathrm{Exp}_x(rz^x)$ for some $r \in [0, 1)$. If we take vectors $a^y \in T_y\mathcal{M}$, $a^x \stackrel{\text{def}}{=} \Gamma_y^x(a^y) \in T_x\mathcal{M}$, then we have the following, for all $\xi \geqslant \zeta_D$:*

$$\begin{aligned}&\|z^y + a^y - \mathrm{Log}_y(p)\|_y^2 + (\xi - 1)\|z^y + a^y\|_y^2 \\ &\leqslant \|z^x + a^x - \mathrm{Log}_x(p)\|_x^2 + (\xi - 1)\|z^x + a^x\|_x^2 + \frac{\xi - \delta_D}{2}\left(\frac{r}{1 - r}\right)\|a^x\|_x^2.\end{aligned}$$

Corollary 13 *Let $x, y, p \in \mathcal{M}$ be the vertices of a uniquely geodesic triangle of diameter D, and let $z^x \in T_x\mathcal{M}$, $z^y \stackrel{\text{def}}{=} \Gamma_x^y(z^x) + \mathrm{Log}_y(x)$, such that $y = \mathrm{Exp}_x(rz^x)$ for some $r \in [0, 1]$. Then, the following holds*

$$\|z^y - \mathrm{Log}_y(p)\|^2 + (\zeta_D - 1)\|z^y\|^2 \leqslant \|z^x - \mathrm{Log}_x(p)\|^2 + (\zeta_D - 1)\|z^x\|^2.$$

The case $r = 1$ above is obtained by taking the limit $r \to 1$.

Lemma 14 *Let $\mathcal{M}$ be a Riemannian manifold of sectional curvature bounded by $[\kappa_{\min}, \kappa_{\max}]$ that contains a uniquely g-convex set $\mathcal{X} \subset \mathcal{M}$ of diameter $D < \infty$. Then, given $x, y \in \mathcal{X}$ we have the following for the function $\Phi_x : \mathcal{M} \to \mathbb{R}$, $y \mapsto \frac{1}{2} d^2(x, y)$:*

$$\nabla \Phi_x(y) = -\mathrm{Log}_y(x) \qquad \text{and} \qquad \delta_D \|v\|^2 \leqslant \mathrm{Hess}\, \Phi_x(y)[v, v] \leqslant \zeta_D \|v\|^2.$$

These bounds are tight for spaces of constant sectional curvature. Consequently, Φ_x is δ_D-strongly g-convex and ζ_D-smooth in $\mathcal{X}$.

See Kim and Yang (2022) for a proof, for instance.

B Rescaling the metric to obtain $L_x = L_y$

If given $(\mathcal{M}, \mathfrak{g})$ we rescale the metric $\mathfrak{g}$ by a factor $c^2 \in \mathbb{R}_{>0}$, we obtain that any distance D is scaled by c. That is, if we consider $(\mathcal{M}, \tilde{\mathfrak{g}})$, where $\tilde{\mathfrak{g}} = c^2\mathfrak{g}$, then if we denote the distance induced by $\tilde{\mathfrak{g}}$ by $d_{\tilde{\mathfrak{g}}}(\cdot)$, we have $d_{\tilde{\mathfrak{g}}}(x, y) = cd(x, y)$ for all $x, y \in \mathcal{M}$. And similarly, if the bounds on the sectional curvature of $(\mathcal{M}, \mathfrak{g})$ are $[\kappa_{\min}, \kappa_{\max}]$, we obtain that the bounds on the sectional curvature for $(\mathcal{M}, \tilde{\mathfrak{g}})$ are $[\tilde{\kappa}_{\min}, \tilde{\kappa}_{\max}] = \frac{1}{c^2}[\kappa_{\min}, \kappa_{\max}]$. Given a geodesically convex set $\mathcal{X}$, the geometric constants $\zeta^{\mathcal{X}}$ and $\delta^{\mathcal{X}}$ remain invariant under this transformation. Indeed, let $\mathcal{X}$ be a set of diameter D measured with $d(\cdot)$ and $\tilde{D}$ measured with $d_{\tilde{\mathfrak{g}}}(\cdot)$. Then, if $\kappa_{\min} < 0$ we have $\zeta^{\mathcal{X}} = D\sqrt{|\kappa_{\min}|}\coth(D\sqrt{|\kappa_{\min}|}) =$

$\tilde{D}\sqrt{|\tilde{\kappa}_{\min}|}\coth(\tilde{D}\sqrt{|\tilde{\kappa}_{\min}|})$. For $\kappa_{\min} > 0$ it is $\zeta^{\mathcal{X}} = 1$ in both cases. Similarly, we obtain the result for $\delta^{\mathcal{X}}$. For a function $f : \mathcal{M} \to \mathbb{R}$, we have that L-smoothness and μ-strong convexity under $\mathfrak{g}$ transforms into $\tilde{L}$-smoothness and $\tilde{\mu}$-strong convexity under $\tilde{\mathfrak{g}}$, for $\tilde{L} = L/c^2$ and $\tilde{\mu} = \mu/c^2$ since by definition:

$$f(y) \leqslant f(x) + \langle \nabla f(x), \operatorname{Log}_x(y)\rangle + \frac{L}{2}d^2(x,y) = f(x) + \langle \nabla f(x), \operatorname{Log}_x(y)\rangle + \frac{L}{2c^2}d_{\tilde{\mathfrak{g}}}(x,y)^2,$$

and analogously for μ-strong convexity. In particular, the condition number $\tilde{L}/\tilde{\mu} = L/\mu$ remains constant, and for any two points $x, y \in \mathcal{M}$ we have $\tilde{L}d_{\tilde{\mathfrak{g}}}(x,y)^2 = Ld^2(x,y)$. Now, if we have a function $f : \mathcal{M} \times \mathcal{N} \to \mathbb{R}$ defined as in Section 1.2, and we rescale the metric of $\mathcal{M}$ by $c_1^2 \stackrel{\text{def}}{=} (L_x/L_y)^{1/2}$ and rescale the metric of $\mathcal{N}$ by $c_2^2 \stackrel{\text{def}}{=} (L_y/L_x)^{1/2}$, then we have $\tilde{L}_x = \sqrt{L_xL_y} = \tilde{L}_y$ and $\tilde{L}_x/\tilde{\mu}_x = L_x/\mu_x$ as well as $\tilde{L}_y/\tilde{\mu}_y = L_y/\mu_y$, and $\tilde{\mu}_x\tilde{\mu}_y = \mu_x(L_x/L_y)^{-1/2}\mu_y(L_y/L_x)^{-1/2} = \mu_x\mu_y$. Finally $\tilde{L}_{xy} = L_{xy}$, since

$$\begin{aligned}\|\nabla_x f(x,y) - \nabla_x f(x,y')\|_{x,\tilde{\mathfrak{g}}} &= \frac{1}{c_1}\|\nabla_x f(x,y) - \nabla_x f(x,y')\|_x \\ &\leqslant \frac{1}{c_1}L_{xy}d(y,y') = \frac{1}{c_1c_2}L_{xy}d_{\tilde{\mathfrak{g}}}(y,y') \\ &= L_{xy}d_{\tilde{\mathfrak{g}}}(y,y').\end{aligned}$$

So indeed one can assume without loss of generality that for one such function f, we have $L_x = L_y$.

C Generalized Riemannian Sion's Theorem

Definition 15 *A function $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is called inf-sup-compact at $(\tilde{x}, \tilde{y}) \in \mathcal{X} \times \mathcal{Y}$ if the sublevel sets of $f(\cdot, \tilde{y})$ and the superlevel sets of $f(\tilde{x}, \cdot)$ are compact. A function is quasi g-convex (resp. quasi g-concave) if their level sets are g-convex (resp. g-concave).*

Theorem 16 (Generalized Sion's Theorem) [↓] *Let $\mathcal{M}$ and $\mathcal{N}$ be finite-dimensional Riemannian manifolds. Further, let $\mathcal{X} \subset \mathcal{M}$ and $\mathcal{Y} \subset \mathcal{N}$ be g-convex and uniquely geodesic subsets. Let $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a function such that $f(\cdot, y)$ is lower semicontinuous and quasi g-convex for all $y \in \mathcal{Y}$, $f(x, \cdot)$ is upper semicontinuous and quasi g-concave for all $x \in \mathcal{X}$ and that is inf-sup compact for some $(x_1, y_1) \in \mathcal{X} \times \mathcal{Y}$. Then we have*

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

Proof of Theorem 16. Let $\mathcal{X}_{\sup} = \{x \in \mathcal{X} \mid \sup_{y\in\mathcal{Y}} f(x,y) \leqslant \alpha\}$ and $\mathcal{X}_{\cap} = \bigcap_{y\in\mathcal{Y}}\{x \in \mathcal{X} \mid f(x,y) \leqslant \alpha\}$. Note that $\mathcal{X}_{\cap} = \{x \in \mathcal{X} \mid f(x,y) \leqslant \alpha, \forall y \in \mathcal{Y}\}$ and hence $\mathcal{X}_{\cap} = \mathcal{X}_{\sup}$ by the definition of the supremum. We have that $\mathcal{X}_{\cap} \subset \{x \in \mathcal{X} \mid f(x,y_1) \leqslant \alpha\}$. For every $y \in \mathcal{Y}$, $\{x \in \mathcal{X} \mid f(x,y) \leqslant \alpha\}$ is closed because $f(\cdot, y)$ is lower semicontinuous for all $y \in \mathcal{Y}$ and by the inf-compactness of $f(\cdot, y_1)$, we have $\{x \in \mathcal{X} \mid f(x,y_1) \leqslant \alpha\}$ is compact. It follows that $\mathcal{X}_{\cap}$ and $\mathcal{X}_{\sup}$ are also compact. Note that $\mathcal{X}_{\sup}$ is a sublevel set of $\varphi(x) = \sup_{y\in\mathcal{Y}} f(x,y)$. Define $M_1 = \{x \in \mathcal{X} \mid \varphi(x) \leqslant \varphi(x_1)\}$, which is compact because the sublevel sets of φ are of the form of $\mathcal{X}_{\sup}$, and it is not empty because $x_1 \in M_1$. We can write $\inf_{x\in\mathcal{X}}\varphi(x) = \inf_{x\in M_1}\varphi(x)$ and since φ is lower semicontinuous and M_1 is compact and non-empty, we have $\min_{x\in\mathcal{X}}\varphi(x) = \min_{x\in\mathcal{X}}\sup_{y\in Y} f(x,y)$. One can analogously, show that $\max_{y\in\mathcal{Y}}\inf_{x\in\mathcal{X}} f(x,y)$ exists.

The inequality $\max_{y\in\mathcal{Y}}\inf_{x\in\mathcal{X}} f(x,y) \leqslant \min_{x\in\mathcal{X}}\sup_{y\in Y} f(x,y)$ holds. In the following, we show that the reverse inequality also holds. Let $\alpha < \min_{x\in\mathcal{X}}\sup_{y\in\mathcal{Y}} f(x,y)$ and $\phi_y(\alpha) \stackrel{\text{def}}{=} \{x \in \mathcal{X} \mid f(x,y) \leqslant \alpha\}$. By definition of ϕ_{y_1} and inf-compactness, $\phi_{y_1}(\alpha)$ is g-convex and compact. We have that the collection of complements

$$\phi^C \stackrel{\text{def}}{=} \left\{\phi_y^C(\alpha) = \{x \in \mathcal{X} \mid f(x,y) > \alpha\}\right\}_{y\in\mathcal{Y}}$$

is an open cover of $\mathcal{X}$. Assume for the sake of contraction that ϕ^C is not an open cover of $\mathcal{X}$. In this case there exists an $x_0 \in \mathcal{X}$ such that $f(x_0,y) \leqslant \alpha$ for all $y \in \mathcal{Y}$. For this x_0, it holds in particular that $\sup_{y\in\mathcal{Y}} f(x_0,y) \leqslant \alpha$. This contradicts the definition of α, since

$$\alpha < \min_x\sup_y f(x,y) \leqslant \sup_{y\in\mathcal{Y}} f(x_0,y) \leqslant \alpha$$

cannot hold.

Since ϕ^C covers $\mathcal{X}$, it also covers $\phi_{y_1}(\alpha) \subset \mathcal{X}$. The set $\phi_{y_1}(\alpha)$ is compact, so it has a finite cover $\left\{\phi^C_{y_i}(\alpha)\right\}_{i=2,\dots,m}$, and thus $\left\{\phi^C_{y_i}(\alpha)\right\}_{i=1,\dots,m}$ is a finite cover for $\mathcal{X}$. We have found a set $\mathcal{Y}_1 = \{y_1, \dots, y_m\}$ such that for all $x \in \mathcal{X}$, there exists $\tilde{y} \in \mathcal{Y}_1$, with $x \in \phi^C_{\tilde{y}}(\alpha)$. This implies $\alpha < \max_{\tilde{y}\in\mathcal{Y}_1} f(x, \tilde{y})$ for all $x \in \mathcal{X}$. Let $\varphi_1(x) = \max_{\tilde{y}\in\mathcal{Y}_1} f(x, \tilde{y})$; then $\widetilde{M}_1 = \{x \in \mathcal{X} \mid \varphi_1(x) \leqslant \varphi_1(x_1)\}$ is compact and non-empty. It follows that $\inf_{x\in\mathcal{X}} \varphi_1(x) = \inf_{x\in\widetilde{M}_1} \varphi_1(x)$ and hence $\min_{x\in\mathcal{X}} \varphi_1(x) = \min_{x\in\mathcal{X}} \max_{\tilde{y}\in\mathcal{Y}_1} f(x, y)$ exists. Since $\alpha < \max_{\tilde{y}\in\mathcal{Y}_1} f(x, \tilde{y})$ for all $x \in \mathcal{X}$, it holds in particular for $\alpha < \min_{x\in\mathcal{X}} \max_{\tilde{y}\in\mathcal{Y}_1} f(x, \tilde{y})$. By Lemma 17, there exists a $y_0 \in \mathcal{Y}$ with $\alpha < \min_{x\in\mathcal{X}} f(x, y_0) \leqslant \sup_{y\in\mathcal{Y}} \min_{x\in\mathcal{X}} f(x, y)$. Consider a monotonic increasing sequence $\alpha_k \to \min_{x\in\mathcal{X}} \sup_{y\in\mathcal{Y}} f(x, y)$, then since $\alpha_k < \sup_{y\in\mathcal{Y}} \min_{x\in\mathcal{X}} f(x, y)$, we have shown the reverse inequality

$$\min_{x\in\mathcal{X}} \sup_{y\in\mathcal{Y}} f(x, y) \leqslant \max_{y\in\mathcal{Y}} \inf_{x\in\mathcal{X}} f(x, y).$$

We have shown that $\min_{x\in\mathcal{X}} \sup_{y\in Y} f(x, y)$ and $\max_{y\in\mathcal{Y}} \inf_{x\in\mathcal{X}} f(x, y)$ exist, i.e., there exists $x_0 = \arg\min_{x\in\mathcal{X}} \sup_{y\in\mathcal{Y}} f(x, y)$ and $y_0 = \arg\max_{y\in\mathcal{Y}} \inf_{x\in\mathcal{X}} f(x, y)$ and we can write

$$f(x_0, \tilde{y}) \leqslant \sup_{y\in\mathcal{Y}} f(x_0, y) = \inf_{x\in\mathcal{X}} f(x, y_0) \leqslant f(\tilde{x}, y_0), \quad \forall (\tilde{x}, \tilde{y}) \in \mathcal{X} \times \mathcal{Y}.$$

Setting $\tilde{x} \leftarrow x_0$ and $\tilde{y} \leftarrow y_0$ we have that

$$f(x_0, y_0) = \min_{x\in\mathcal{X}} f(x, y_0) = \max_{y\in\mathcal{Y}} f(x_0, y).$$

It follows that

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

which concludes the proof. ■

The previous proof was inspired by ideas from three different generalizations of Sion's theorem, namely from Hartung (1982); Komiya (1988); Zhang, Zhang, and Sra (2022). The following lemma, that we used in the proof of Theorem 16 above, appeared in Zhang, Zhang, and Sra (2022). We add the lemma with a proof for completeness.

Lemma 17 *Let $(\mathcal{M}, d_{\mathcal{M}})$ and $(\mathcal{N}, d_{\mathcal{N}})$ be finite-dimensional, unique geodesic metric spaces. Suppose $\mathcal{X} \subseteq \mathcal{M}$, $\mathcal{Y} \subseteq \mathcal{N}$ are geodesically convex sets. Let $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a function such that $f(\cdot, y)$ is geodesically-quasi-convex and lower semi-continuous and $f(x, \cdot)$ is geodesically-quasi-concave and upper semi-continuous. Then for any finite k points $y_1, \dots, y_k \in \mathcal{Y}$ and any real number $\alpha < \min_{x\in X} \max_{i\in[k]} f(x, y_i)$, there exists $y_0 \in \mathcal{Y}$ s.t. $\alpha < \min_{x\in X} f(x, y_0)$.*

Proof We prove the lemma for two points, and then the general lemma holds by induction. Suppose it does not hold, so assume that for such an α, we have $\min_{x\in\mathcal{X}} f(x, y) \leqslant \alpha$ for any $y \in \mathcal{Y}$. As a consequence, there is at least a constant β such that

$$\min_{x\in\mathcal{X}} f(x, y) \leqslant \alpha < \beta < \min_{x\in\mathcal{X}} \max\{f(x, y_1), f(x, y_2)\}. \tag{1}$$

Consider the geodesic $\gamma : [0, d(y_1, y_2)] \to \mathcal{Y}$ connecting y_1 and y_2. For any $t \in [0, d(y_1, y_2)]$ and corresponding $z = \gamma(t)$ on the geodesic, the level sets $\phi_z(\alpha), \phi_z(\beta)$ are nonempty due to (1) and closed due to lower semi-continuity of f regarding the first variable. Since f is geodesic quasi-concave in the second variable, we obtain

$$f(x, z) \geqslant \min\{f(x, y_1), f(x, y_2)\}, \quad \forall x \in \mathcal{X}.$$

This is equivalent to say $\phi_z(\alpha) \subseteq \phi_z(\beta) \subseteq \phi_{y_1}(\beta) \cup \phi_{y_2}(\beta)$. We then argue the intersection $\phi_{y_1}(\beta) \cap \phi_{y_2}(\beta)$ should be empty. Otherwise, there exists $x \in \mathcal{X}$ such that $\max\{f(x, y_1), f(x, y_2)\} \leqslant \beta$, contradicting (1). Next, by quasi-convexity, since level set $\phi_z(\beta)$ is geodesic convex for any z, it is also connected. Consider the three facts:

- $\phi_z(\alpha) \subseteq \phi_{y_1}(\beta) \cup \phi_{y_2}(\beta)$
- $\phi_{y_1}(\beta) \cap \phi_{y_2}(\beta)$ is empty
- $\phi_z(\alpha), \phi_{y_1}(\beta)$ and $\phi_{y_2}(\beta)$ are closed (due to lower semi-continuity), connected and convex

We claim that either $\phi_z(\alpha) \subseteq \phi_{y_1}(\alpha)$ or $\phi_z(\alpha) \subseteq \phi_{y_2}(\alpha)$ holds for any point z on the geodesic γ. Suppose not, then we can always find two points $w_1 \in \phi_{y_1}(\beta), w_2 \in \phi_{y_2}(\beta)$ such that $w_1, w_2 \in \phi_z(\alpha)$. Since $\phi_z(\alpha)$ is convex, then there is a geodesic $\gamma : [0,1] \to \mathcal{X}$ in $\phi_z(\alpha)$ connecting w_1, w_2. Therefore γ also lies in $\phi_z(\alpha) \subseteq \phi_{y_1}(\beta) \cup \phi_{y_2}(\beta)$. Because $\phi_{y_1}(\beta) \cap \phi_{y_2}(\beta)$ is empty, γ^{-1} induces a partition on $[0,1]$ as $J_1 \cap J_2 = \varnothing$ and $J_1 \cup J_2 = [0,1]$ where $\gamma(J_1) \subseteq \phi_{y_1}(\beta), \gamma(J_2) \subseteq \phi_{y_2}(\beta)$. Therefore at least one of J_1, J_2 is not closed. Since γ is a continuous map, at least one of $\phi_{y_2}(\beta)$ or $\phi_{y_2}(\beta)$ is also not closed, contradicting known conditions. Since either $\phi_z(\alpha) \subseteq \phi_{y_1}(\alpha)$ or $\phi_z(\alpha) \subseteq \phi_{y_2}(\alpha)$, the two sets below

$$I_1 \stackrel{\text{def}}{=} \left\{ t \in [0,1] \mid \phi_{\gamma(t)}(\alpha) \subseteq \phi_{y_1}(\beta) \right\},$$
$$I_2 \stackrel{\text{def}}{=} \left\{ t \in [0,1] \mid \phi_{\gamma(t)}(\alpha) \subseteq \phi_{y_2}(\beta) \right\}$$

form a partition of the interval $[0,1]$. We prove I_1 is closed and nonempty. The latter is obvious since at least $\gamma^{-1}(y_1) \in I_1$. Now we turn to prove closedness. Let t_k be an infinite sequence in I_1 with a limit point of t. We consider any $x \in \phi_{\gamma(t)}(\alpha)$. The upper semi-continuity of $f(x,\cdot)$ implies

$$\limsup_{k \to \infty} f(x, \gamma(t_k)) \leqslant f(x, \gamma(t)) \leqslant \alpha < \beta.$$

Therefore, there exists a large enough integer l such that $f(x, \gamma(t_l)) < \beta$. This implies $x \in \phi_{\gamma(t_l)}(\beta) \subseteq \phi_{y_1}(\beta)$. Therefore for any $x \in \phi_{\gamma(t)}(\alpha)$, $x \in \phi_{y_1}(\beta)$ also holds. This is equivalent to $\phi_{\gamma(t)}(\alpha) \subseteq \phi_{y_1}(\beta)$. Hence by the definition of level set, we know the $t \in I_1$ and I_1 is then closed. By a similar argument, I_2 is also closed and nonempty. This contradicts the definition of partition and hence proves the lemma. ■

Corollary 18 *For $\mu_x, \mu_y > 0$, a (μ_x, μ_y)-SCSC function $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is inf-sup compact for any point in $\mathcal{X} \times \mathcal{Y}$. If we have an $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ that is CC, then if $\mathcal{X}$, $\mathcal{Y}$ are compact then $f(x,y) + I_{\mathcal{X}}(x) - I_{\mathcal{Y}}(y)$ is inf-sup compact for any point, where $I_{\mathcal{C}}$ denotes the indicator function of a set $\mathcal{C}$, which is 0 if $x \in \mathcal{C}$ and it is $+\infty$ otherwise. Similarly, if the function is $(\mu_x, 0)$-SCC and $\mathcal{Y}$ is compact then $f(x,y) - I_{\mathcal{Y}}(y)$ is inf-sup compact for any point.*

D Proofs of Riemannian Corrected Extra-Gradient

Proof of Proposition 9. We show that if $d(x_t, x^*)^2 + d(y_t, y^*)^2 \leqslant \mathcal{D}^2$ then $d(w_t, x^*)^2 + d(z_t, y^*)^2 \leqslant 4\mathcal{D}^2$ and $d(x_{t+1}, x^*)^2 + d(y_{t+1}, y^*)^2 \leqslant \mathcal{D}^2$. Then, these two latter properties are satisfied for all $t \geqslant 0$, since $d(x_0, x^*)^2 + d(y_0, y^*)^2 \leqslant \mathcal{D}^2$. Recall our notation $\zeta \stackrel{\text{def}}{=} \zeta_D$ and $\delta \stackrel{\text{def}}{=} \delta_D$.

We start by showing that in both the CC and the SCSC cases, the secondary iterates (w_t, z_t) are not far from the saddle point:

$$\begin{aligned} d^2(w_t, x^*) + d^2(z_t, y^*) &\leqslant 2[d^2(w_t, x_t) + d^2(x_t, x^*) + d^2(z_t, y_t) + d^2(y_t, y^*)] \\ &\overset{①}{\leqslant} (2 + 4\eta^2 L^2)(d^2(x_t, x^*) + d^2(y_t, y^*)) \\ &\overset{②}{\leqslant} (2 + \frac{\delta}{\zeta})(d^2(x_t, x^*) + d^2(y_t, y^*)) \\ &\overset{③}{\leqslant} 4\mathcal{D}^2. \end{aligned}$$

Here, ① holds since by definition of w_t we have $d^2(w_t, x_t) = \|\eta \nabla_x f(x_t, y_t)\|^2 \leqslant 2\eta^2 L^2 (d^2(x_t, x^*) + d^2(y_t, y^*))$ and similarly $d^2(z_t, y_t) \leqslant 2\eta^2 L^2 (d^2(x_t, x^*) + d^2(y_t, y^*))$. Further ② holds because $\eta \leqslant \sqrt{\frac{1}{4L^2\zeta}}$. And ③ holds since we have $\frac{\delta}{\zeta} \leqslant 1$ and by our hypothesis on x_t, y_t. We bounded $3\mathcal{D} \leqslant 4\mathcal{D}$ for convenience. Now, since f is μ-SCSC, we have

that

$$\begin{aligned} f(w_t, y^*) - f(x^*, z_t) &= f(w_t, y^*) - f(w_t, z_t) + f(w_t, z_t) - f(x^*, z_t) \\ &\overset{①}{\leqslant} +\langle \nabla_y f(w_t, z_t), \operatorname{Log}_{z_t}(y^*)\rangle - \frac{\mu}{2} d^2(w_t, x^*) \\ &\quad - \langle \nabla_x f(w_t, z_t), \operatorname{Log}_{w_t}(x^*)\rangle - \frac{\mu}{2} d^2(z_t, y^*) \\ &= \frac{1}{\eta}\langle -\eta \nabla_x f(w_t, z_t) \pm \operatorname{Log}_{w_t}(x_t), \operatorname{Log}_{w_t}(x^*)\rangle - \frac{\mu}{2} d^2(w_t, x^*) \\ &\quad + \frac{1}{\eta}\langle \eta \nabla_y f(w_t, z_t) \pm \operatorname{Log}_{z_t}(y_t), \operatorname{Log}_{z_t}(y^*)\rangle - \frac{\mu}{2} d^2(z_t, y^*) \\ &\overset{②}{\leqslant} \frac{1}{\eta}\langle \operatorname{Log}_{w_t}(x_{t+1}), \operatorname{Log}_{w_t}(x^*)\rangle - \frac{1}{\eta}\langle \operatorname{Log}_{w_t}(x_t), \operatorname{Log}_{w_t}(x^*)\rangle - \frac{\mu}{2} d^2(w_t, x^*) \\ &\quad + \frac{1}{\eta}\langle \operatorname{Log}_{z_t}(y_{t+1}), \operatorname{Log}_{z_t}(y^*)\rangle - \frac{1}{\eta}\langle \operatorname{Log}_{z_t}(y_t), \operatorname{Log}_{z_t}(y^*)\rangle - \frac{\mu}{2} d^2(z_t, y^*), \end{aligned} \tag{2}$$

where we used the μ-SCSC property in ① and the definition of the iterates x_{t+1}, y_{t+1} in ②. We now use the Riemannian cosine inequalities to obtain the inequalities below, cf. Lemma 10, (Zhang and Sra, 2016, Lemma 1). The first two inequalities use the fact that the diameters of the geodesic triangles with vertices w_t, x_t, x^* and z_t, y_t, y^*, respectively, are upper bounded by $4\mathcal{D} \leqslant D$, because all of those points are in $\mathcal{B} \overset{\text{def}}{=} \bar{B}((x^*, y^*), 2\mathcal{D})$ and this ball is geodesically convex and uniquely convex and hence it contains the triangles. If $\kappa_{\min} \geqslant 0$, we have that $\zeta_c = 1$ for any $c > 0$, so we can just use Lemma 10 to obtain the last two inequalities. If $\kappa_{\min} < 0$, we use the more fine-grained inequality (Zhang and Sra, 2016, Lemma 1). This lemma establishes the cosine inequalities with constants $\zeta_{d(w_t, x^*)}$ and $\zeta_{d(z_t, y^*)}$, respectively. The inequalities also hold for greater values of these constants, so we use ζ because we already established that $d(w_t, x^*), d(z_t, y^*) \leqslant 2\mathcal{D} \leqslant 4\mathcal{D}$.

$$\begin{aligned} -2\langle \operatorname{Log}_{w_t}(x_t), \operatorname{Log}_{w_t}(x^*)\rangle &\leqslant -\delta d^2(w_t, x_t) - d^2(w_t, x^*) + d^2(x_t, x^*) \\ -2\langle \operatorname{Log}_{z_t}(y_t), \operatorname{Log}_{z_t}(y^*)\rangle &\leqslant -\delta d^2(z_t, y_t) - d^2(z_t, y^*) + d^2(y_t, y^*) \\ 2\langle \operatorname{Log}_{w_t}(x_{t+1}), \operatorname{Log}_{w_t}(x^*)\rangle &\leqslant \zeta d^2(w_t, x_{t+1}) + d^2(w_t, x^*) - d^2(x_{t+1}, x^*) \\ 2\langle \operatorname{Log}_{z_t}(y_{t+1}), \operatorname{Log}_{z_t}(y^*)\rangle &\leqslant \zeta d^2(z_t, y_{t+1}) + d^2(z_t, y^*) - d^2(y_{t+1}, y^*). \end{aligned} \tag{3}$$

We can further bound the following term using the update rules and gradient Lipschitzness,

$$\begin{aligned} d^2(w_t, x_{t+1}) &= \|\operatorname{Log}_{w_t}(x_{t+1})\|^2 = \|\operatorname{Log}_{w_t}(x_t) - \eta \nabla_x f(w_t, z_t)\|^2 \\ &= \|\eta \Gamma_{x_t}^{w_t} \nabla_x f(x_t, y_t) - \eta \nabla_x f(w_t, z_t)\|^2 \\ &= 2\eta^2 \left(\|\Gamma_{x_t}^{w_t} \nabla_x f(x_t, y_t) - \nabla_x f(w_t, y_t)\|^2 + \|\nabla_x f(w_t, y_t) - \nabla_x f(w_t, z_t)\|^2 \right) \\ &\leqslant 2\eta^2 L^2 (d^2(w_t, x_t) + d^2(z_t, y_t)). \end{aligned} \tag{4}$$

Analogously, we obtain

$$d^2(z_t, y_{t+1}) \leqslant 2\eta^2 L^2 (d^2(w_t, x_t) + d^2(z_t, y_t)). \tag{5}$$

Using the triangle inequality and $(a+b)^2 \leqslant 2a^2 + 2b^2$, we have

$$\begin{aligned} -\frac{\mu}{2} d^2(w_t, x^*) &\leqslant \frac{\mu}{2} d^2(w_t, x_t) - \frac{\mu}{4} d^2(x_t, x^*), \\ -\frac{\mu}{2} d^2(z_t, y^*) &\leqslant \frac{\mu}{2} d^2(z_t, y_t) - \frac{\mu}{4} d^2(y_t, y^*). \end{aligned} \tag{6}$$

So finally, we now bound (2) using (3) in combination with (4) to (6). We will use the following inequality to study the CC and SCSC cases separately.

$$\begin{aligned} 0 &\leqslant f(w_t, y^*) - f(x^*, z_t) \\ &\leqslant \frac{1}{2\eta}[(4\zeta\eta^2 L^2 - \delta + \mu\eta) d^2(w_t, x_t) + (1 - \frac{\mu\eta}{2}) d^2(x_t, x^*) - d^2(x_{t+1}, x^*)] \\ &\quad + \frac{1}{2\eta}[(4\zeta\eta^2 L^2 - \delta + \mu\eta) d^2(z_t, y_t) + (1 - \frac{\mu\eta}{2}) d^2(y_t, y^*) - d^2(y_{t+1}, y^*)]. \end{aligned} \tag{7}$$

Case CC We have $\mu = 0$. It follows that for $\eta \leqslant \sqrt{\frac{\delta}{4\zeta L^2}}$, we have by (7) that

$$d^2(x_{t+1}, x^*) + d^2(y_{t+1}, y^*) \leqslant d^2(y_t, y^*) + d^2(x_t, x^*) \leqslant \mathcal{D}^2.$$

Case SCSC We have $\mu > 0$. It follows that for $\eta \leqslant \min\left\{\sqrt{\frac{\delta}{8L^2\zeta}}, \frac{\delta}{2\mu}\right\}$, we have by (7) that

$$d^2(x_{t+1}, x^*) + d^2(y_{t+1}, y^*) \leqslant \left(1 - \frac{\mu\eta}{2}\right)\left(d^2(y_t, y^*) + d^2(x_t, x^*)\right) \leqslant d^2(y_t, y^*) + d^2(x_t, x^*) \leqslant \mathcal{D}^2.$$

Now to conclude the first statement, recall $\mathcal{D} \stackrel{\text{def}}{=} d((x_0, y_0), (x^*, y^*))$ and $D \geqslant 4\mathcal{D}$ by definition. Since we just showed that the iterates do not go farther than $2\mathcal{D} \leqslant D/2$ to (x^*, y^*), then they stay in the closed ball $\mathcal{B} \stackrel{\text{def}}{=} \bar{B}((x^*, y^*), D/2)$, whose diameter is D. Therefore, our choice of η in the pseudocode in Algorithm 2 is a valid one. Note that the knowledge of this set $\mathcal{B}$ is not needed for the algorithm.

For the second statement, we note that the proofs of the convergence rates stated in the proposition were provided by Jordan, Lin, and Vlatakis-Gkaragkounis (2022) for the SCSC case and by Zhang, Zhang, and Sra (2022) for the CC case under the following additional assumption: when η is chosen with respect to some geometric constants δ_D and ζ_D, then the iterates do not leave a set of diameter D that contains (x^*, y^*). We showed that the iterates satisfy this assumption for our choice of step size η and therefore the convergence follows. Note that Jordan, Lin, and Vlatakis-Gkaragkounis (2022) proves for the SCSC case that (x_T, y_T) is an ε'-saddle point in distance. By Statement 4 of Lemma 8, we have that $\operatorname{gap}(x_T, y_T) \leqslant [d(x_T, x^*) + d(y_T, y^*)]LD\,(1 + L/\mu)$. This holds, as by assumption $\nabla f(x^*, y^*) = 0$ and hence $L_p(f(x, y)) \leqslant LD$. Hence, we have that $\operatorname{gap}(x_T, y_T) \leqslant 2LD\left(1 + \frac{L}{\mu}\right)\sqrt{\varepsilon'}$. Thus after

$$T = \widetilde{O}\left(\frac{L}{\mu}\sqrt{\frac{\zeta_D}{\delta_D} + \frac{1}{\delta_D}}\right).$$

iterations of RCEG, we obtain a ε-saddle point for the SCSC case.

■

Algorithm 2 Riemannian Corrected Extragradient (RCEG)

Input: Initialization (x_0, y_0), $f : \mathcal{M} \times \mathcal{N} \to \mathbb{R}$, manifolds $\mathcal{M}$ and $\mathcal{N}$, g-strong convexity constant μ (for SCSC), smoothness L, bound $D \geqslant 4\mathcal{D} = 4(d^2(x_0, x^*) + d^2(y_0, y^*))^{1/2}$.

For SCSC, choose $\eta \leftarrow \min\left\{\sqrt{\frac{\delta_D}{8L^2\zeta_D}}, \frac{\delta_D}{2\mu}\right\}$, for CC, choose $\eta \leftarrow \sqrt{\frac{\delta_D}{4L^2\zeta_D}}$
$(w_0, z_0) \leftarrow \left(\exp_{x_0}(-\eta\nabla_x f(x_0, y_0)), \exp_{y_0}(\eta\nabla_y f(x_0, y_0))\right)$
$\bar{z}_0 \leftarrow z_0$, $\bar{w}_0 \leftarrow w_0$
$(x_1, y_1) \leftarrow \left(\exp_{w_0}(-\eta\nabla_x f(w_0, z_0) + \operatorname{Log}_{w_0}(x_0)), \exp_{z_0}(\eta\nabla_y f(w_0, z_0) + \operatorname{Log}_{z_0}(y_0))\right)$
for $t = 1$ **to** $T - 1$ **do**
  $(w_t, z_t) \leftarrow \left(\exp_{x_t}(-\eta\nabla_x f(x_t, y_t)), \exp_{y_t}(\eta\nabla_y f(x_t, y_t))\right)$
  $(x_{t+1}, y_{t+1}) \leftarrow \left(\exp_{w_t}(-\eta\nabla_x f(w_t, z_t) + \operatorname{Log}_{w_t}(x_t)), \exp_{z_t}(\eta\nabla_y f(w_t, z_t) + \operatorname{Log}_{z_t}(y_t))\right)$
  if $\mu = 0$ **then** ◇ Geodesic averaging for CC
    $(\bar{w}_t, \bar{z}_t) \leftarrow \left(\exp_{\bar{w}_{t-1}}\left(t^{-1}\log_{\bar{w}_{t-1}}(w_t)\right), \exp_{\bar{z}_{t-1}}\left(t^{-1}\log_{\bar{z}_{t-1}}(z_t)\right)\right)$
  end if
end for
Output: (x_T, y_T) **if** $\mu > 0$, **else** $(\bar{w}_{T-1}, \bar{z}_{T-1})$

E Proofs of Convergence Rates for our Accelerated Algorithms

E.1 Acceleration for G-Convex Functions

In the following, we discuss our results on g-convex accelerated optimization. For a μ-strongly g-convex function over a g-convex closed subset $\mathcal{X}$ of a finite-dimensional Hadamard manifold $\mathcal{M}$, we improve over the state of the art (Martínez-Rubio and Pokutta, 2023, Algorithm 1) in two important directions. First, we obtain better convergence rates: $\widetilde{O}(\sqrt{\zeta\kappa_\lambda} + \zeta)$

as opposed to $\widetilde{O}(\zeta\sqrt{\kappa_\lambda})$, where $\kappa_\lambda \stackrel{\text{def}}{=} 1/(\lambda\bar{\mu})$ and λ is the step size of the approximate implicit gradient descent step of the algorithm. Recall that $\zeta = \Theta(1+\sqrt{|\kappa_{\min}|D})$. A similar improvement is obtained for the g-convex case via reductions, see Remark 22. The key idea for our improved algorithm is to work directly in the strongly g-convex case, as opposed to the g-convex case and make a careful choice of step sizes. Second, we require a less restrictive condition for the subroutine that Algorithm 3 (RiemaconAbs) uses, namely we only require to compute a minimizer of the subproblem in Line 12, $\min_{y\in\mathcal{X}}\{f(y)+\frac{1}{2\lambda}d^2(x_k,y)\}$, with *absolute* accuracy, as opposed to *relative* accuracy, i.e., accuracy proportional to the distance to the minimizer of the proximal problem, which is unknown in general.

Algorithm 3 RiemaconAbs$(f, x_0, T$ or $\varepsilon, \lambda, \mathcal{X}, \texttt{subroutine})$. Absolute accuracy criterion.

Input: Finite-dimensional Hadamard manifold $\mathcal{M}$ of bounded sectional curvature, feasible g-convex compact set $\mathcal{X}$ of diameter $D_{\mathcal{X}}$. Initial point $x_0 \in \mathcal{X} \subset \mathcal{M}$. Function $f : \mathcal{M} \to \mathbb{R}$ that is $\bar{\mu}$-strongly g-convex in $\mathcal{X}$. Parameter $\lambda > 0$. Final iteration T or accuracy ε. If ε is provided, compute the corresponding T and vice versa, cf. Theorem 3, `subroutine` to solve the proximal problems of Lines 5 and 12.

1: $\xi \leftarrow 4\zeta_{2D_{\mathcal{X}}} - 3$ ◇ ξ is $O(\zeta_{D_{\mathcal{X}}})$
2: $\mu \leftarrow \min\{\bar{\mu}, 1/(9\xi\lambda)\}$
3: $\kappa \leftarrow 1/(\lambda\mu)$
4: $\hat{\varepsilon} \leftarrow \varepsilon \cdot (8\sqrt{\xi}\kappa^{3/2})^{-1}$
5: $y_0 \leftarrow \hat{\varepsilon}$-minimizer of the proximal problem $\min_{y\in\mathcal{X}}\{f(y)+\frac{1}{2\lambda}d^2(x_0,y)\}$
6: $\bar{z}_0^{y_0} \leftarrow z_0^{y_0} \leftarrow 0 \in T_{y_0}\mathcal{M}$
7: $A_0 \leftarrow 1$

8: **for** $k = 1$ **to** T **do**
9: $A_k \leftarrow (1 + 1/(2\sqrt{\xi\kappa}))^k$
10: $a_k \leftarrow \xi(A_k - A_{k-1})$
11: $x_k \leftarrow \mathrm{Exp}_{y_{k-1}}(\frac{a_k}{A_{k-1}+a_k}\bar{z}_{k-1}^{y_{k-1}} + \frac{A_{k-1}}{A_{k-1}+a_k}y_{k-1}) = \mathrm{Exp}_{y_{k-1}}(\frac{a_k}{A_{k-1}+a_k}\bar{z}_{k-1}^{y_{k-1}})$ ◇ Coupling
12: $y_k \leftarrow \hat{\varepsilon}$-minimizer of problem $\min_{y\in\mathcal{X}}\{f(y)+\frac{1}{2\lambda}d^2(x_k,y)\}$ ◇ Approximate implicit RGD
13: $v_k^x \leftarrow -\mathrm{Log}_{x_k}(y_k)/\lambda$ ◇ Approximate subgradient
14: $z_{k-1}^{x_k} \leftarrow \mathrm{Log}_{x_k}(\mathrm{Exp}_{y_{k-1}}(\bar{z}_{k-1}^{y_{k-1}}))$
15: $z_k^{x_k} \leftarrow A_k^{-1}(A_{k-1}z_{k-1}^{x_k} + \frac{a_k}{\xi}(-\lambda - \frac{2}{\mu})v_k^x)$ ◇ Mirror Descent step
16: $z_k^{y_k} \leftarrow \Gamma_{x_k}^{y_k}(z_k^{x_k}) + \mathrm{Log}_{y_k}(x_k)$ ◇ Moving the dual point to $T_{y_k}\mathcal{M}$
17: $\bar{z}_k^{y_k} \leftarrow \Pi_{\bar{B}(0,D_{\mathcal{X}})}(z_k^{y_k}) \in T_{y_k}\mathcal{M}$ ◇ Easy projection done so the dual point is not very far
18: **end for**
19: **return** y_T.

We use RiemaconAbs to refer to Algorithm 3 for μ-strongly g-convex minimization. We write RiemaconAbs$(f, x_0, T, \mathcal{X}, \texttt{subroutine})$ to specify the output of the algorithm initialized at x_0 for optimizing the function f constrained to $\mathcal{X}$, run for T steps and making use of the subroutine `subroutine`. The subroutine solves a proximal problem approximately, effectively implementing an approximate implicit Riemannian Gradient Descent (RGD) step. In the pseudocode, we use the notation $\Pi_C(p)$ to refer to the Euclidean projection of a point p onto a closed convex set C.

Proof of Theorem 3. We note that it is $\xi \stackrel{\text{def}}{=} 4\zeta_{2D_{\mathcal{X}}} - 3 \leqslant 8\zeta_{D_{\mathcal{X}}} - 3 = O(\zeta_{D_{\mathcal{X}}})$. Let $\kappa \stackrel{\text{def}}{=} \frac{1}{\lambda\mu}$ and let $c \stackrel{\text{def}}{=} \frac{1}{2\sqrt{\xi\kappa}}$, so that $A_k = (1+c)^k$ for all $k \geqslant 1$, and $a_k = \xi((1+c)^k - (1+c)^{k-1}) = \xi c(1+c)^{k-1}$ for all $k \geqslant 1$. We also note that in order to satisfy $\sqrt{\xi/\kappa} \leqslant 1/3$, to be used later, we only use μ-strong g-convexity of f, instead of $\bar{\mu}$-strong convexity, where $\mu \stackrel{\text{def}}{=} \min\{\bar{\mu}, 1/(9\xi\lambda)\}$. We want to show the following is almost a Lyapunov function for our problem

$$\Psi_k \stackrel{\text{def}}{=} A_k\left(f(y_k) - f(x^*) + \frac{\mu}{4}\|z_k^{y_k} - x^*\|_{y_k}^2 + \frac{\mu(\xi-1)}{4}\|z_k^{y_k}\|_{y_k}^2\right),$$

in the sense that we can show

$$\Psi_k \leqslant \Psi_{k-1} + 2(\kappa+1)A_k\hat{\varepsilon}. \tag{8}$$

If we show (8), then we can conclude the theorem since for $T \geqslant 2\sqrt{\xi\kappa}\log_2(\frac{2\lambda^{-1}d^2(x_0,x^*)}{\varepsilon})$, we would have

$$f(y_T) - f(x^*) \leqslant \frac{\Psi_T}{A_T} \leqslant \frac{\Psi_0}{A_T} + 2\hat{\varepsilon}(\kappa+1)\frac{\sum_{i=1}^T A_i}{A_T} \leqslant \frac{\Psi_0}{A_T} + 2\hat{\varepsilon}(\kappa+1)\frac{A_{T+1}}{cA_T}$$
$$\overset{①}{\leqslant} \frac{\Psi_0}{2^{Tc}} + 4\hat{\varepsilon}\kappa^{3/2}\sqrt{\xi} \overset{②}{\leqslant} \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

where in ① we used $c = 1/(2\sqrt{\xi\kappa}) \leqslant 1/6$ (by $\sqrt{\xi/\kappa} \leqslant 1/3$ and $\xi \geqslant 1$) and so $(1+c)^{1/c} > 2$. We also bounded $1+c \leqslant 2$. For ②, we used $\hat{\varepsilon} \overset{\text{def}}{=} \varepsilon/(8\sqrt{\kappa^3\xi})$ and $\Psi_0 \leqslant f(y_0) - f(x^*) \leqslant \lambda^{-1}d^2(x_0,x^*)$ due to Lemma 24 and the definition of y_0. In short, we find an ε-minimizer for $T = O(\sqrt{\zeta\kappa}\log(\frac{\lambda^{-1}d^2(x_0,x^*)}{\varepsilon}))$.

We now focus on proving (8). We can assume without loss of generality that $x_k = 0$. We work in the tangent space of x_k all of the time except when applying Lemma 20 that moves lower bounds, so according to our notation, points x^*, y_{k-1}, and y_k should be interpreted as $\operatorname{Log}_{x_k}(x^*)$, $\operatorname{Log}_{x_k}(y_{k-1})$ and $\operatorname{Log}_{x_k}(y_k)$, respectively. We note that our choice of dual point $z_k^{x_k}$ comes from optimizing the regularized lower bound that we have at iteration k:

$$\begin{aligned} z_k^{x_k} &= \operatorname*{arg\,min}_{x\in T_{x_k}\mathcal{M}} \left\{ \left(\frac{A_{k-1}\mu}{4}\right)\|z_{k-1}^{x_k} - x\|_{x_k}^2 + \frac{a_k}{\xi}\frac{\mu}{4}\left\|\left(1+\frac{2}{\mu\lambda}\right)y_k - \frac{2}{\mu\lambda}x_k - x\right\|_{x_k}^2 \right\} \\ &\overset{①}{=} \frac{A_{k-1}z_{k-1}^{x_k} + \frac{a_k}{\xi}\left[\left(1+\frac{2}{\mu\lambda}\right)y_k - \frac{2}{\lambda\mu}x_k\right]}{A_{k-1}+a_k/\xi} \overset{②}{=} \frac{A_{k-1}z_{k-1}^{x_k} - \frac{a_k}{\xi}\left(\lambda+\frac{2}{\mu}\right)v_k^x}{A_{k-1}+a_k/\xi}. \end{aligned} \tag{9}$$

The equality ① can be obtained by just taking a derivative and checking when we have global optimality. Equality ② uses $x_k = 0$ and $\lambda v_k^x = x_k - y_k = -y_k$. The definition of $z_k^{x_k}$ as the $\arg\min$ above is derived from minimizing a convex combination of the previously computed regularized lower bound, a quadratic with minimizer at $z_{k-1}^{x_k}$, plus the new bound which we obtain by Lemma 19. Indeed, one can check that the second summand is a quadratic that has the same minimizer as the right hand side in Lemma 19. In order to show (8), by Lemma 20 it is enough to show

$$\begin{aligned} &A_k\left(f(y_k) - f(x^*) + \frac{\mu}{4}\|z_k^{x_k} - x^*\|_{x_k}^2 + \frac{\mu(\xi-1)}{4}\|z_k^{x_k}\|_{x_k}^2 - 2(\kappa+1)\hat{\varepsilon}\right) \\ &\quad\leqslant A_{k-1}\left(f(y_{k-1}) - f(x^*) + \frac{\mu}{4}\|z_{k-1}^{x_k} - x^*\|_{x_k}^2 + \frac{\mu(\xi-1)}{4}\|z_{k-1}^{x_k}\|_{x_k}^2\right). \end{aligned} \tag{10}$$

The following identities involving our parameters will be useful in the sequel

$$A_k = A_{k-1} + a_k/\xi \tag{11}$$
$$A_{k-1}(x_k - y_{k-1}) = -a_k(x_k - z_{k-1}^{x_k}) \text{ (equiv. to) } y_{k-1} = -\frac{a_k}{A_{k-1}}z_{k-1}^{x_k} \tag{12}$$
$$y_k = -\lambda v_k^x \tag{13}$$

We regroup the terms in (10) with evaluations of f to the left hand side to yield $A_{k-1}(f(y_k) - f(y_{k-1})) + (a_k/\xi)\cdot(f(y_k) - f(x^*))$ and then we apply Lemma 19 twice to show that it is enough to prove:

$$\begin{aligned} &A_{k-1}\langle v_k^x, y_k - y_{k-1}\rangle - A_{k-1}\frac{\mu}{4}\|y_{k-1} - y_k\|^2 + \frac{a_k}{\xi}\langle v_k^x, y_k - x^*\rangle - \frac{a_k}{\xi}\frac{\mu}{4}\|x^* - y_k\|^2 \\ &\leqslant \frac{\mu}{4}\left[A_{k-1}(\|z_{k-1}^{x_k} - x^*\|^2 + (\xi-1)\|z_{k-1}^{x_k}\|^2) - A_k\left(\|z_k^{x_k} - x^*\|^2 + (\xi-1)\|z_k^{x_k}\|^2\right)\right] \end{aligned}$$

Note that the errors with respect to $\hat{\varepsilon}$ cancel each other. Now, we will just check that the terms involving $\langle x^*, \cdot\rangle$ and $\|x^*\|^2$ cancel each other, given our choice of $z_k^{x_k}$. Indeed, for $\|x^*\|^2$ we have the following weights on each side:

$$-\frac{a_k}{\xi}\frac{\mu}{4} = \frac{\mu}{4}(A_{k-1} - A_k),$$

which holds by (11). Then, on each side of the inequality we have the following that holds by our choice of $z_k^{x_k}$, and (13)

$$\langle x^*, -\frac{a_k}{\xi}v_k^x + \frac{a_k}{\xi}\frac{\mu}{2}\cdot(-\lambda)v_k^x\rangle = \langle x^*, \frac{\mu}{4}A_{k-1}\cdot(-2z_{k-1}^{x_k}) - \frac{\mu}{4}A_k\cdot(-2z_k^{x_k})\rangle.$$

We remove those terms involving x^*, and we use the properties (12) and (13) and the definition of $z_k^{x_k}$ so the only variables left are a_k, A_{k-1}, A_k, v_k^x and $z_{k-1}^{x_k}$:

$$
\begin{aligned}
&-\lambda A_{k-1}\|v_k^x\|^2 + a_k\langle v_k^x, z_{k-1}^{x_k}\rangle - \frac{\mu}{4}A_{k-1}\lambda^2\|v_k^x\|^2 - \frac{\mu}{4}\frac{a_k^2}{A_{k-1}}\|z_{k-1}^{x_k}\|^2 + 2\langle v_k^x, z_{k-1}^{x_k}\rangle\frac{\lambda\mu a_k}{4}\\
&-\frac{a_k}{\xi}\lambda\|v_k^x\|^2 - \frac{a_k}{\xi}\frac{\mu}{4}\lambda^2\|v_k^x\|^2\\
&\leqslant \frac{\mu}{4}A_{k-1}\xi\|z_{k-1}^{x_k}\|^2 - \frac{\mu}{4}\frac{\xi}{A_k}\left(A_{k-1}^2\|z_{k-1}^{x_k}\|^2 + \frac{a_k^2}{\xi^2}(\lambda+\frac{2}{\mu})^2\|v_k^x\|^2 - 2\langle v_k^x, z_{k-1}^{x_k}\rangle A_{k-1}\frac{a_k}{\xi}(\lambda+\frac{2}{\mu})\right)
\end{aligned}
$$

The strategy now is to complete squares to make appear a factor proportional to $-\|av_k^x + bz_{k-1}^{x_k}\|^2$ on the left hand side, and show that we can prove the inequality without that term, where $a, b \in \mathbb{R}$. We pick a so that $-a^2\|v_k^x\|^2$ is precisely the term involving $\|v_k^x\|^2$ that we have above, if we move all of those to the left hand side. In other words, after completing squares we will just need to prove that the resulting factor multiplying $\|z_{k-1}^{x_k}\|^2$ is non-positive. Let's first regroup all the coefficients with respect to $\|v_k^x\|^2$, $\|z_{k-1}^{x_k}\|^2$ and $\langle v_k^x, z_{k-1}^{x_k}\rangle$ and place them on the left hand side:

$$
\begin{aligned}
&\langle v_k^x, z_{k-1}^{x_k}\rangle \cdot a_k\left(1+\frac{\lambda\mu}{2}\right)\left(1-\frac{A_{k-1}}{A_k}\right) + \|v_k^x\|^2\cdot\left(-\lambda A_k - \frac{\mu}{4}\lambda^2 A_k + \frac{\mu}{4}\frac{a_k^2/\xi}{A_k}\left(\lambda+\frac{2}{\mu}\right)^2\right)\\
&+\|z_{k-1}^{x_k}\|^2\left(-\frac{\mu}{4}\frac{a_k^2}{A_{k-1}} - \frac{\mu\xi}{4}A_{k-1} + \frac{\mu\xi}{4}\frac{A_{k-1}^2}{A_k}\right)\leqslant 0.
\end{aligned}
\tag{14}
$$

So now we pick $-a^2$ as the resulting factor multiplying $\|v_k^x\|^2$, i.e., $-a^2 \overset{\text{def}}{=} -\lambda A_k(1+\frac{\mu\lambda}{4}) + \frac{1}{4\mu}\frac{a_k^2/\xi}{A_k}(\lambda\mu+2)^2$ and therefore we have

$$
b^2 = \frac{(2ab)^2}{4a^2} = a_k^2\left(1+\frac{\lambda\mu}{2}\right)^2\left(1-\frac{A_{k-1}}{A_k}\right)^2\frac{1}{4a^2}.
$$

For this computation to be valid, we need to show our choice for $-a^2$ is non-positive. We recall it holds that $A_k = (1+c)^k$, $A_{k-1} = (1+c)^{k-1}$, $a_k = \xi c(1+c)^{k-1}$ and use $\kappa = \frac{1}{\mu\lambda}$, and $c = \frac{1}{2\sqrt{\xi\kappa}}$. So $-a^2 \leqslant 0$ if and only if ① below holds:

$$
\xi\left(\frac{1+4\kappa+4\kappa^2}{1+4\kappa}\right) = \frac{\xi}{4\lambda\mu}\left(\frac{(\lambda\mu+2)^2}{(1+\frac{\lambda\mu}{4})}\right) \overset{①}{\leqslant} \frac{A_k^2\xi^2}{a_k^2} = \frac{(1+c)^2}{c^2} = 1+\frac{2}{c}+\frac{1}{c^2} = 1+4\sqrt{\xi\kappa}+4\xi\kappa.
$$

And this inequality is clearly satisfied, since we assumed $\kappa \geqslant 1$. Indeed, drop the two first summands on the right hand side and multiply by $1+4\kappa$.

So now we can just add $2ab\langle v_k^x, z_{k-1}^{x_k}\rangle + a^2\|v_k^x\|^2 + b^2\|z_{k-1}^{x_k}\|^2 = \|v_k^x - z_{k-1}^{x_k}\|^2 \geqslant 0$ to the left hand side of (14) and show the resulting inequality, in order to prove the result. So it is enough to prove the resulting factor multiplying $\|z_{k-1}^{x_k}\|^2$ is non-positive:

$$
\begin{aligned}
&a_k^2\left(1+\frac{\lambda\mu}{2}\right)^2\left(1-\frac{A_{k-1}}{A_k}\right)^2\left(\lambda A_k(4+\mu\lambda) - \frac{1}{\mu}\frac{a_k^2/\xi}{A_k}(\lambda\mu+2)^2\right)^{-1}\\
&-\frac{\mu a_k^2}{4A_{k-1}} - \frac{\mu\xi A_{k-1}}{4} + \frac{\mu\xi A_{k-1}^2}{4A_k}\leqslant 0.
\end{aligned}
$$

Now, this inequality holds by substituting the values of A_k, A_{k-1}, a_k, using $\kappa = (\lambda\mu)^{-1}$ and $c = \frac{1}{2\sqrt{\xi\kappa}}$ and doing some simple computations. Indeed, by substituting, combining the last two summands, dividing by $\mu(1+c)^{k-2}/4$, reducing the $4+\mu\lambda$ to $2+\mu\lambda$ and simplifying terms, we obtain it is enough to prove

$$
\frac{c^4(1+c)^{k-2}\xi^2(2+\lambda\mu)}{\lambda\mu(1+c)^k - (1+c)^{k-2}c^2\xi(\lambda\mu+2)} - c^2(1+c)\xi^2 - c\xi\leqslant 0.
$$

From here, the inequality follows by operating out and comparing terms. If we just substitute the value of $c^2 = 1/(2\xi\kappa)$ in the instances where there is c^4 or c^2, then κ disappears from the equation and one can compare terms to reach the result, by using $c \in (0,1)$, $\xi > 1$. ■

We now prove the auxiliary lemmas that were used to prove Theorem 3.

Lemma 19 (Approx. st. g-convexity by approx. subgradient) *Let y_k be an $\hat{\varepsilon}$ minimizer of $h_k(x) \stackrel{\text{def}}{=} \min_{x\in\mathcal{X}}\{f(x) + \frac{1}{2\lambda}d^2(x_k, x)\}$, and let $v_k^x \stackrel{\text{def}}{=} -\lambda^{-1}\mathrm{Log}_{x_k}(y_k)$. Then, for all $x \in \mathcal{X}$, we have*

$$f(x) \geqslant f(y_k) + \langle v_k^x, x - y_k\rangle_{x_k} + \frac{\mu}{4}\|x - y_k\|_{x_k}^2 - \left(\frac{2}{\lambda\mu} + 2\right)\hat{\varepsilon}.$$

Proof Let $y_k^* \stackrel{\text{def}}{=} \arg\min_{x\in\mathcal{X}} h_k(x)$. The function h_k is $(\frac{1}{\lambda} + \mu)$-strongly g-convex because by Lemma 14 the function $\frac{1}{2}d^2(x_k, x)$ is 1-strongly g-convex in a Hadamard manifold. This strong convexity and optimality of the point y_k^* yield ① below. Besides, we have

$$\begin{aligned}
f(x) &\overset{①}{\geqslant} \left(f(y_k^*) + \frac{1}{2\lambda}d^2(x_k, y_k^*)\right) - \frac{1}{2\lambda}d^2(x_k, x) + \left(\frac{1}{2\lambda} + \frac{\mu}{2}\right)d^2(y_k^*, x)\\
&\overset{②}{\geqslant} \left(f(y_k) + \frac{1}{2\lambda}\|x_k - y_k\|_{x_k}^2 - \hat{\varepsilon}\right) - \frac{1}{2\lambda}\|x_k - x\|_{x_k}^2 + \left(\frac{1}{2\lambda} + \frac{\mu}{2}\right)\|y_k^* - x\|_{x_k}^2\\
&= f(y_k) + \langle v_k^x, x - y_k\rangle_{x_k} + \frac{\mu}{2}\|x - y_k\|_{x_k}^2 + \left(\frac{1}{\lambda} + \mu\right)(\langle x - y_k, y_k - y_k^*\rangle_{x_k} + \frac{1}{2}\|y_k^* - y_k\|_{x_k}^2) - \hat{\varepsilon} \qquad (15)\\
&\overset{③}{\geqslant} f(y_k) + \langle v_k^x, x - y_k\rangle_{x_k} + \frac{\mu}{4}\|x - y_k\|_{x_k}^2 - \left(\frac{1}{\lambda\mu} + \frac{1}{2}\right)\left(\frac{1}{\lambda} + \mu\right)\|y_k^* - y_k\|_{x_k}^2 - \hat{\varepsilon}\\
&\overset{④}{\geqslant} f(y_k) + \langle v_k^x, x - y_k\rangle_{x_k} + \frac{\mu}{4}\|x - y_k\|_{x_k}^2 - \left(\frac{2}{\lambda\mu} + 2\right)\hat{\varepsilon}.
\end{aligned}$$

where in ② we used the $\hat{\varepsilon}$-optimality of y_k for $h_k(\cdot)$ and we used for the last summand that in a Hadamard manifold we have $d(x, y) \geqslant \|x - y\|_z$ for any three points x, y, z.

In ③, we used Young's inequality:

$$\langle x - y_k, \left(\frac{1}{\lambda} + \mu\right)(y_k - y_k^*)\rangle_{x_k} \geqslant -\frac{\mu}{4}\|x - y_k\|_{x_k}^2 - \frac{1}{\mu}\left(\frac{1}{\lambda} + \mu\right)^2\|y_k - y_k^*\|^2,$$

and grouped some terms. Finally, in ④ we used $-\|y_k^* - y_k\|_{x_k}^2 \geqslant -d^2(y_k^*, y_k)$ and then $(\frac{1}{\lambda} + \mu)$-strong convexity of h_k along with $\hat{\varepsilon}$-optimality of y_k: $(\frac{1}{\lambda} + \mu)\|y_k^* - y_k\|_{x_k}^2 \leqslant h_k(y_k) - h_k(y_k^*) \leqslant \varepsilon$. ■

Lemma 20 (Translating potentials with no geometric penalty) *Using the notation in Algorithm 3, we have*

$$\begin{aligned}
&A_{k-1}(\|z_{k-1}^{x_k} - x^*\|_{x_k}^2 + (\xi - 1)\|z_{k-1}^{x_k}\|_{x_k}^2) - A_k(\|z_k^{x_k} - x^*\|_{x_k}^2 + (\xi - 1)\|z_k^{x_k}\|_{x_k}^2)\\
&\quad\leqslant A_{k-1}(\|z_{k-1}^{y_{k-1}} - x^*\|_{y_{k-1}}^2 + (\xi - 1)\|z_{k-1}^{y_{k-1}}\|_{y_{k-1}})\\
&\quad\quad - A_k(\|z_k^{y_k} - x^*\|_{y_k}^2 + (\xi - 1)\|z_{k-1}^{y_{k-1}}\|_{y_{k-1}}^2 - \|y_k - z_k^{y_k}\|_{y_k}^2).
\end{aligned}$$

Proof Firstly, by the projection step in Line 17, we have

$$\|z_{k-1}^{y_{k-1}} - x^*\|_{y_k}^2 \geqslant \|\bar{z}_{k-1}^{y_{k-1}} - x^*\|_{y_k}^2 \qquad \text{and} \qquad (\xi - 1)\|z_{k-1}^{y_{k-1}}\|_{y_k}^2 \geqslant (\xi - 1)\|\bar{z}_{k-1}^{y_{k-1}}\|_{y_k}^2 \qquad (16)$$

since the operation is a simple Euclidean projection onto the closed ball $\bar{B}(0, D)$ in $T_{y_k}\mathcal{M}$. Now, the following holds

$$\begin{aligned}
\|\bar{z}_{k-1}^{y_{k-1}} - x^*\|_{y_{k-1}}^2 + (\xi - 1)\|\bar{z}_{k-1}^{y_{k-1}}\|_{y_{k-1}}^2 &\overset{①}{\geqslant} \|z_{k-1}^{x_k} - x^*\|_{x_k}^2 + (\zeta_{2D} - 1)\|z_{k-1}^{x_k}\|_{x_k}^2 + (\xi - \zeta_{2D})\|\bar{z}_{k-1}^{y_{k-1}}\|_{y_{k-1}}^2\\
&\overset{②}{\geqslant} \|z_{k-1}^{x_k} - x^*\|_{x_k}^2 + (\xi - 1)\|z_{k-1}^{x_k}\|_{x_k}^2 + (\xi - \zeta_{2D})\left(\left(\frac{A_{k-1} + a_k}{A_{k-1}}\right)^2 - 1\right)\|z_{k-1}^{x_k}\|_{x_k}^2 \qquad (17)\\
&\overset{③}{\geqslant} \|z_{k-1}^{x_k} - x^*\|_{x_k}^2 + (\xi - 1)\|z_{k-1}^{x_k}\|_{x_k}^2 + \frac{3(\xi - 1)}{2}\left(\left(\frac{A_{k-1} + a_k}{A_{k-1}}\right)^2 - 1\right)\|z_{k-1}^{x_k}\|_{x_k}^2,
\end{aligned}$$

where ① is due to Corollary 13, with $y \leftarrow x_k$ and $x \leftarrow y_{k-1}$ and to $d(x_k,p) \leqslant d(x_k,y_{k-1}) + d(y_{k-1},p) \leqslant \|z_{k-1}^{y_{k-1}}\|_{y_{k-1}} + D \leqslant 2D$ for any $p \in \mathcal{M}$. Inequality ② uses the definition of x_k. In ③, we used the definition of $\xi = 4\zeta_{2D} - 3$ that implies $\xi - \zeta_{2D} \geqslant \frac{3}{4}(\xi - 1)$. Now, we use Lemma 12 with $y \leftarrow y_k$, $x \leftarrow x_k$ $z^x = -v_k^x \cdot \frac{a_k}{\xi}(\lambda + \frac{4}{\mu})/A_k$, $a^x \leftarrow z_{k-1}^{x_k}(A_{k-1}/A_k)$, so that $z^x + a^x = z_k^{x_k}$ and $z^y + a^y = z_k^{y_k}$ and

$$r = \frac{\|\mathrm{Log}_{x_k}(y_k)\|}{\|z^x\|} = \frac{\lambda\|v_k^x\|}{\|v_k^x\| \cdot \frac{a_k}{\xi}(\lambda + \frac{4}{\mu})A_k^{-1}} = \frac{\xi A_k}{a_k(1 + \frac{4}{\lambda\mu})} \overset{①}{\leqslant} \sqrt{\frac{\xi}{\kappa}} \overset{②}{\leqslant} \frac{1}{3} < 1.$$

We will now explain why ① holds. But first, note that by the previous inequality, by the choice of parameters and the fact that $r < 1$, the assumptions in Lemma 12 are satisfied. Also, note that ② holds by the assumption on λ. We have ① above if and only if the following holds

$$\xi A_{k-1} \leqslant \sqrt{\frac{\xi}{\kappa}} a_k \left(-\sqrt{\frac{\kappa}{\xi}} + 1 + 4\kappa \right), \tag{18}$$

and this is implied by $\xi \leqslant \sqrt{\frac{\xi}{\kappa}} c\xi \cdot 3\kappa$, by substituting the values of A_{k-1} and a_k and using $3\kappa \leqslant 4\kappa - \sqrt{\kappa/\xi} \leqslant 4\kappa - \sqrt{\kappa/\xi} + 1$, which comes from our assumption $\kappa \geqslant 9\xi \geqslant \xi$. Consequently, a sufficient condition is $c \geqslant 1/(3\sqrt{\xi\kappa})$, which is satisfied by our choice $c = 1/(2\sqrt{\xi\kappa})$.

The result in Lemma 12, applied as above results in

$$\|z_k^{x_k} - x^*\|_{x_k}^2 + (\xi - 1)\|z_k^{x_k}\|_{x_k}^2 + \frac{\xi - 1}{2}\left(\frac{r}{1-r}\right)\frac{A_{k-1}^2}{A_k^2}\|z_{k-1}^{x_k}\|^2 \geqslant \|z_k^{y_k} - x^*\|_{y_k}^2 + (\xi - 1)\|z_k^{y_k}\|_{y_k}^2. \tag{19}$$

Combining (17) multiplied by A_{k-1} with (19) multiplied by A_k, we obtain that in order to conclude, it suffices to show

$$A_k \frac{\xi - 1}{2}\frac{r}{1-r}\frac{A_{k-1}^2}{A_k^2} - A_{k-1}\frac{3(\xi-1)}{2}\left(\left(\frac{a_k + A_{k-1}}{A_{k-1}}\right)^2 - 1\right) \leqslant 0.$$

We substitute the value of r and after simplifying we obtain that we want to show

$$\xi A_{k-1}^3 - 3(a_k^2 + 2a_k A_{k-1})\left(a_k\left(1 + \frac{4}{\lambda\mu}\right) - \xi A_k\right) \leqslant 0$$

After substituting the value of A_k, A_{k-1}, a_k, operating out and dividing by $\xi(1+c)^{3k-3}$, we obtain that the previous inequality is equivalent to

$$1 + 3\xi^2(1+c)c^2 + 6\xi c + 6\xi c^2 \leqslant 3\xi^2(1+4\kappa)c^3 + 6\xi c^2(1+4\kappa) = \frac{3\xi c}{2\kappa} + 6\xi c + \frac{3}{\kappa} + 12$$

where in the last equality we used $c^2 = 1/(4\xi\kappa)$. This inequality holds. After simplifying the terms $6\xi c$ we get that the left hand side is $\leqslant 1 + \frac{3(1+c)}{2\kappa} + \frac{3}{\kappa} \leqslant 7$ where we used the value of $c = 1/(2\sqrt{\xi\kappa}) \leqslant 1$ and $\kappa \geqslant 1$.

■

Proof of Corollary 4. By Theorem 3, it suffices to run Algorithm 3 for

$$T'' \geqslant 2\sqrt{\frac{\xi}{\lambda\bar{\mu}} + 9\xi^2}\log_2\left(\frac{2\lambda^{-1}d^2(x_0,x^*)}{\varepsilon}\right)$$

iterations in order to obtain an ε-minimizer. We use $\lambda = 1/L$, and recall $\xi = \widetilde{O}(\zeta)$. Step k of Algorithm 3 requires computing a $\hat{\varepsilon}$-minimizer of $\min_{x\in\mathcal{X}} h_k(x)$, where $h_k(x) \overset{\text{def}}{=} f(x) + \frac{1}{2\lambda}d^2(x_k,x)$ and

$$\hat{\varepsilon} = \frac{\varepsilon}{8\sqrt{\xi\left(\frac{L}{\min\{\mu, L/(9\xi)\}}\right)^3}}.$$

We implement the subroutine with PRGD with step size $\frac{1}{L'}$, where $L' \stackrel{\text{def}}{=} L(1+\zeta)$ is a bound on the smoothness of h, cf. Lemma 14. We require the following number of steps

$$T' \geqslant 1 + 2\frac{L'}{\mu'}\zeta_R \log\left(\frac{L'\zeta_R D_{\mathcal{X}}^2}{2\hat{\varepsilon}}\right),$$

where $R = L_p(h, \mathcal{X})/L'$. We can bound

$$R \leqslant \frac{\max_{x\in\mathcal{X}} \|\nabla f(x)\| + Ld(x, x_k)}{L(1+\zeta)} \overset{①}{\leqslant} \frac{\max_{x\in\mathcal{X}}\{\|\nabla f(x)\|/L\} + 2D_{\mathcal{X}}}{\zeta},$$

where in ① we used that for all $x \in \mathcal{X}$, it is

$$d(x, x_k) \leqslant d(y_{k-1}, x_k) + d(x, y_{k-1}) \overset{②}{<} d(y_{k-1}, \bar{z}_{k-1}^{y_{k-1}}) + D_{\mathcal{X}} \overset{③}{\leqslant} 2D_{\mathcal{X}}.$$

where ② holds by definition of x_k and the fact $y_{k-1}, x \in \mathcal{X}$, while ③ is due to the projection defining $\bar{z}_{k-1}^{y_{k-1}}$.

The condition number of h is bounded by

$$\frac{L'}{\mu'} \leqslant \lambda L + \zeta \leqslant 1 + \zeta.$$

So we have

$$T' \geqslant 1 + 2\zeta_R(1+\zeta)\log\left(4\frac{L(1+\zeta)\zeta_R D_{\mathcal{X}}^2\sqrt{(\frac{L}{\mu}+9\xi)^3\xi}}{\varepsilon}\right),$$

The complete complexity of RiemaconAbs is

$$T = T'T'' = \widetilde{O}(\zeta_R\zeta^{\frac{3}{2}}\sqrt{\kappa+\zeta}).$$

■

Corollary 21 *Under the assumptions from Corollary 4, if a global minimizer $x^* \in \arg\min_{\mathcal{M}} f(x)$ is in $\mathcal{X}$, so that $\nabla f(x^*) = 0$, then Algorithm 3 with $\lambda = 1/L$ and PRGD as subroutine, as in Corollary 4 yields an ε-minimizer after $\widetilde{O}(\zeta^{3/2}\sqrt{\kappa+\zeta})$ gradient and projection oracle calls.*

Proof By the assumption on x^* and the smoothness assumption, we have that $L_p(f, \mathcal{X}) \leqslant LD_{\mathcal{X}}$ and since $\zeta \geqslant D_{\mathcal{X}}\sqrt{|\kappa_{\min}|}$, we obtain $R \leqslant (L_p(f, \mathcal{X})/L + 2D_{\mathcal{X}})/\zeta = O(\frac{1}{\sqrt{|\kappa_{\min}|}})$ and thus $\zeta_R = O(1)$. We obtain the result by applying Corollary 4. Note that we can also use PRGD with RiemaconRel and Corollary 26 and similarly it is $\zeta_R = O(1)$. ■

Note that Martínez-Rubio and Pokutta (2023, Theorem 6) had to assume a mild condition on the Hadamard manifold and obtained overall query complexity $\widetilde{O}(\zeta^2\sqrt{\kappa})$ whereas we obtain lower complexity in the general Hadamard case with bounded sectional curvature. Note that to compare to Martínez-Rubio and Pokutta (2023, Theorem 6) we would run our algorithm for $\mathcal{X}$ a ball of center x_0 and radius an upper bound on $d(x_0, x^*)$ and in such a case the implementation of the projection oracle consist of computing a direct and an inverse exponential and simple operations, cf. Section 1.1, so it does not increase the order of our computational complexity. We note that Martínez-Rubio and Pokutta (2023) provided another instantiation of their algorithm for g-convex minimization but under the assumption of having access to a projection oracle that is not a metric-projected oracle.

Remark 22 *We can obtain the accelerated result for the g-convex case with reduced geometric penalties via a reduction to the μ-strongly g-convex case. Assume the existence of a global minimizer x^* and let $D/2 \geqslant d(x_0, x^*)$. Given an $\varepsilon > 0$, we optimize the regularized function $f_\varepsilon(x) \stackrel{\text{def}}{=} f + \frac{\varepsilon}{D^2}d(x_0, x)^2$. Denote by x_ε^* to the minimizer of f_ε. We have $d(x_0, x_\varepsilon^*) \leqslant d(x_0, x^*) \leqslant D/2$ (see Martínez-Rubio and Pokutta (2023, Lemma 10)). We run Algorithm 3 on f_ε on a ball $\bar{B}(x_0, D/2)$. We have that f_ε is strongly g-convex with constant $\frac{2\varepsilon}{D^2}$, cf. Lemma 14. Hence, the algorithm finds an $\varepsilon/2$ minimizer $x_{T'}$ of f_ε after $T' = \widetilde{O}(\zeta + \sqrt{\zeta D^2/(\lambda\varepsilon)})$ iterations. By definition, it is $d(x_0, x^*) \leqslant D/2$ so the regularization at x^* is $\frac{\varepsilon}{D^2}d(x_0, x^*)^2 \leqslant \frac{\varepsilon}{4}$ and thus $x_{T'}$ is an ε-minimizer of f:*

$$f(x_{T'}) \leqslant f_\varepsilon(x_{T'}) \leqslant f_\varepsilon(x^*) + \frac{\varepsilon}{4} \leqslant f(x^*) + \varepsilon.$$

E.2 Convergence of Projected Gradient Descent

Proof of Proposition 2. Below, we prove that for any point $x_t \in \mathcal{X}$, PRGD yields

$$f(x_{t+1}) - f(x^*) \leqslant (f(x_t) - f(x^*)) \left(1 - \frac{\mu}{4L\zeta_{R_t}}\right), \tag{20}$$

where $R_t \stackrel{\text{def}}{=} \|\nabla f(x_t)\|/L$. Recall our notation $L_p(f, \mathcal{X})$ for denoting the Lipschitz constant of f in $\mathcal{X}$. Given (20) above, and defining $R \stackrel{\text{def}}{=} L_p(f, \mathcal{X})/L$, we have by applying (20) T times from x_0, that the following holds

$$f(x_T) - f(x^*) \leqslant \min\left\{(f(x_0) - f(x^*))\left(1 - \frac{\mu}{4L\zeta_R}\right)^T, \frac{L\zeta_R}{2} d^2(x_0, x^*)\left(1 - \frac{\mu}{4L\zeta_R}\right)^{T-1}\right\},$$

since by Lemma 24 we have $f(x_1) - f(x^*) \leqslant \frac{L\zeta_R}{2} d^2(x_0, x^*)$. The result follows by bounding the right hand side of the expression above by ε and reorganizing.

We now prove (20). The following holds:

$$\begin{aligned}
f(x_{t+1}) &\overset{①}{\leqslant} \min_{x\in\mathcal{X}}\left\{f(x) + \frac{L\zeta_{R_t}}{2} d^2(x, x_t)\right\} \\
&\overset{②}{\leqslant} \min_{\alpha\in[0,1]}\left\{\alpha f(x^*) + (1-\alpha) f(x_t) + \frac{L\zeta_{R_t}\alpha^2}{2} d^2(x^*, x_t)\right\} \\
&\overset{③}{\leqslant} \min_{\alpha\in[0,1]}\left\{f(x_t) - \alpha\left(1 - \alpha\frac{L\zeta_{R_t}}{\mu}\right)(f(x_t) - f(x^*))\right\} \\
&\overset{④}{=} f(x_t) - \frac{\mu}{4L\zeta_{R_t}}(f(x_t) - f(x^*)).
\end{aligned}$$

Above, we used Lemma 24 to conclude ①, and ② results from restricting the minimum to the geodesic segment between x^* and x_t so that $x = \operatorname{Exp}_{x_t}(\alpha x^* + (1-\alpha)x_t)$. We also use g-convexity of f. In ③, we used strong convexity of f to bound $\frac{\mu}{2} d^2(x^*, x_t) \leqslant f(x_t) - f(x^*)$. Finally, in ④ we substituted α by the value that minimizes the expression, which is $\mu/(2L\zeta_{R_t})$. The result in (20) follows by subtracting $f(x^*)$ to the inequality above. The final statement is a direct consequence of (20) and the definition of R, along with $f(x_1) - f(x^*) \leqslant \frac{L\zeta_R}{2} d^2(x_0, x^*)$ which holds due to Lemma 24.

■

Remark 23 *Let $\mathcal{M}$ be a Hadamard manifold with curvatures in the interval $[\kappa_{\min}, \kappa_{\max}]$ where $\kappa_{\max} < 0$. Criscitiello and Boumal (2022b, Theorem 2) provide the query complexity lower bound $\tilde{\Omega}(\sqrt{\frac{\kappa_{\max}}{\kappa_{\min}}}\zeta)$ for minimizing a smooth strongly g-convex function in a ball. On the other hand, Martínez-Rubio and Pokutta (2023, Prop 17) provide an $\widetilde{O}(\kappa)$ gradient query complexity upper bound in this setting. From this, we conclude that $\kappa \geqslant \tilde{\Omega}(\sqrt{\frac{\kappa_{\max}}{\kappa_{\min}}}\zeta)$. This fact is also shown by Criscitiello and Boumal (2022b, Prop 28), albeit in a very different way.*

Lemma 24 (Dist to Gap and Warm Start) *Let $\mathcal{M}$ be a Hadamard manifold $\mathcal{X} \subseteq \mathcal{M}$ be a uniquely g-convex set of diameter D, $\bar{x} \in \mathcal{X}$, and $g : \mathcal{M} \to \mathbb{R}$ a g-convex and L-smooth function in $\mathcal{X}$ with a minimizer at $x^* \in \arg\min_{x\in\mathcal{X}} g(x)$. Assume access to a projection operator $\mathcal{P}_{\mathcal{X}}$ on $\mathcal{X}$ and let $x' \stackrel{\text{def}}{=} \mathcal{P}_{\mathcal{X}}(\operatorname{Exp}_{\bar{x}}(-\frac{1}{L}\nabla g(\bar{x})))$, and $R \stackrel{\text{def}}{=} d(x', \bar{x}) = \|\nabla g(\bar{x})\|/L$. The following holds for all $p \in \mathcal{X}$:*

$$g(x') - g(p) \leqslant \frac{\zeta_R L}{2} d^2(\bar{x}, p).$$

In particular, we have

$$g(x') - g(x^*) \leqslant \frac{\zeta_R L}{2} d^2(\bar{x}, x^*).$$

See Martínez-Rubio and Pokutta (2023, Lemma 18 (Warm start)) for a proof.

Algorithm 4 Riemannian Alternating Best Response RABR($f, (x_0, y_0), T$ or $\varepsilon, \mathcal{X} \times \mathcal{Y}$)

Input: G-convex subsets $\mathcal{X} \subset \mathcal{M}$, $\mathcal{Y} \subset \mathcal{N}$ of Hadamard manifolds $\mathcal{M}$ and $\mathcal{N}$, initialization $(x_0, y_0) \in \mathcal{X} \times \mathcal{Y}$, function $f : \mathcal{M} \times \mathcal{N} \to \mathbb{R}$ that is (μ_x, μ_y)-SCSC and (L_x, L_y, L_{xy})-smooth, T (if ε is given, compute T, see Theorem 5). Define $\xi \stackrel{\text{def}}{=} 4\max\{\zeta_{2D}^{\mathcal{X}}, \zeta_{2D}^{\mathcal{Y}}\} - 3 = O(\zeta)$

$T_x \leftarrow 90\xi\sqrt{\kappa_x}\log(512)$, $T_y \leftarrow 90\xi\sqrt{\kappa_y}\log(512)$

for $t = 0$ **to** $T - 1$ **do**

  $x_{t+1} \leftarrow \text{RiemaconRel}(f(\cdot, y_t), x_t, T_x, \mathcal{X}, \text{PRGD})$

  $y_{t+1} \leftarrow \text{RiemaconRel}(-f(x_{t+1}, \cdot), y_t, T_y, \mathcal{Y}, \text{PRGD})$

end for

Output: (x_T, y_T)

E.3 Convergence of Riemannian Alternating Best Response

We use RiemaconRel to refer to the accelerated algorithm for μ-strongly convex functions presented in Martínez-Rubio and Pokutta (2023, Theorem 4), and $\text{RiemaconRel}(f, x_0, T, \mathcal{X}, \texttt{subroutine})$ to specify the output of the algorithm initialized at x_0 for optimizing the function f constrained to $\mathcal{X}$, run for T steps and making use of the subroutine `subroutine`.

Fact 25 (Convergence of RiemaconRel**)** *Let $\mathcal{M}$ be a finite-dimensional Hadamard manifold of bounded sectional curvature, and consider $f : \mathcal{X} \subset \mathcal{M} \to \mathbb{R}$, a g-convex function in a compact g-convex set $\mathcal{X}$ of diameter $D_{\mathcal{X}}$, $\lambda > 0$, and $x^* \in \arg\min_{x \in \mathcal{X}} f(x)$. Define $\xi \stackrel{\text{def}}{=} 4\zeta_{2D} - 3$. If f is μ-strongly g-convex then, running* RiemaconRel *as defined in Martínez-Rubio and Pokutta (2023, Theorem 4) for $T = (90\xi/\sqrt{\mu\lambda})\log(\mu d^2(x_0, x^*)/\varepsilon)$ iterations, returns a point y_T that satisfies $f(y_T) - f(x^*) \leqslant \varepsilon$.*

See Martínez-Rubio and Pokutta (2023, Theorem 4) for a proof.

Corollary 26 [↓] *If f as defined in Fact 25 is in addition L-smooth, then* RiemaconRel *with $\lambda = 1/L$ and PRGD as subroutine, yields an ε-minimizer after $\widetilde{O}(\zeta_R\zeta^2\sqrt{\kappa})$ gradient and projection oracle calls, where $R \leqslant (L_p(f, \mathcal{X})/L + 2D_{\mathcal{X}})/\zeta$.*

Proof of Corollary 26. By Fact 25, it suffices to run RiemaconRel for

$$T'' = \frac{90\xi}{\sqrt{\mu\lambda}}\log\left(\frac{\mu d^2(x_0, x^*)}{\varepsilon}\right) \tag{21}$$

iterations in order to obtain an ε-minimizer. Note $\xi = \widetilde{O}(\zeta)$. We use $\lambda = 1/L$. Note that RiemaconRel uses a series of restarts, so in the following, k refers to the k-th iteration in one of the calls to (Martínez-Rubio and Pokutta, 2023, Algorithm 1). This detail can be ignored in the following, as we bound k uniformly by T'' in (22). Step k of RiemaconRel requires computing a σ_k-minimizer of $\min_{x \in \mathcal{X}} h_k(x)$, where $h_k(x) \stackrel{\text{def}}{=} f(x) + \frac{1}{2\lambda}d^2(x_k, x)$ and $\sigma_k \stackrel{\text{def}}{=} d^2(x_k, x_k^*)/(78\lambda(k+1)^2)$. We solve the prox problem using PRGD with step size $\frac{1}{L'}$, where $L' \stackrel{\text{def}}{=} L(1+\zeta)$ is a bound on the smoothness of h, cf. Lemma 14. By Proposition 2, this requires

$$T' \geqslant 1 + 2\frac{L'}{\mu'}\zeta_R\log\left(\frac{L'\zeta_R D_{\mathcal{X}}^2}{2\sigma_k}\right)$$

iterations, where L' and μ' are smoothness and strong g-convexity constants of h respectively and $R = \|\nabla h(x_0)\|/L'$. We can bound

$$R \leqslant \frac{\max_{x \in \mathcal{X}}\|\nabla f(x)\| + Ld(x, x_k)}{L(1+\zeta)} \leqslant \frac{\max_{x \in \mathcal{X}}\|\nabla f(x)\|/L + 2D_{\mathcal{X}}}{\zeta},$$

where the last inequality uses $d(\mathcal{X}, x_k) \leqslant 2D_{\mathcal{X}}$, cf. (Martínez-Rubio and Pokutta, 2023). We have that κ', the condition number of h, is bounded by $\kappa' \leqslant \frac{L'}{\mu'} \leqslant \lambda L + \zeta \leqslant 1 + \zeta$. By the definition of κ' and bounding $k \leqslant T''$, we obtain

$$T' \geqslant 1 + 4\zeta_R\zeta\log(78\zeta\zeta_R(T''+1)^2) \geqslant 1 + 4\zeta_R\zeta\log(78\zeta\zeta_R(k+1)^2). \tag{22}$$

All in all RiemaconRel requires

$$T = T'T'' = \widetilde{O}(\zeta_R\zeta^2\sqrt{\kappa})$$

calls to the gradient and metric projection oracle respectively. ■

Proof of Theorem 5. We begin by connecting the inexactness of the iterates (x_t, y_t) from Algorithm 4 to the number of RiemaconRel iterations,

$$\begin{aligned} d^2(x_{t+1}, x^*(y_t)) &\overset{①}{\leqslant} \frac{2}{\mu_x}[f(x_{t+1}, y_t) - f(x^*(y_t), y_t)] \\ &\overset{②}{\leqslant} 2d^2(x_t, x^*(y_t)) \exp\left(\frac{-T_x}{90\xi\sqrt{\kappa_x}}\right). \end{aligned} \tag{23}$$

We used strong g-convexity of $x \mapsto f(x, y_t)$ in ①, and ② follows from running RiemaconRel on $f(\cdot, y_t)$ for T_x iterations starting from x_t and ending up with x_{t+1}. Noting that since $-f(x_{t+1}, \cdot)$ is strongly g-convex, we can repeat the arguments for y,

$$\begin{aligned} d^2(y_{t+1}, y^*(x_{t+1})) &\leqslant \frac{2}{\mu_y}[f(x_{t+1}, y^*(x_{t+1})) - f(x_{t+1}, y_{t+1})] \\ &\leqslant 2d^2(y_t, y^*(x_{t+1})) \exp\left(\frac{-T_y}{90\xi\sqrt{\kappa_y}}\right). \end{aligned} \tag{24}$$

Note that we use RiemaconRel instead of our RiemaconAbs in Algorithm 4, since the absolute error criterion of RiemaconAbs creates an unwanted dependence between the required precision of the proximal problems and the precision of the original problem. Choosing $T_x = 90\xi\sqrt{\kappa_x}\log(512)$ and $T_y = 90\xi\sqrt{\kappa_y}\log(512)$, it follows from (23) and (24) that,

$$d^2(x_{t+1}, x^*(y_t)) \leqslant \frac{1}{256} d^2(x_t, x^*(y_t)) \tag{25}$$

$$d^2(y_{t+1}, y^*(x_{t+1})) \leqslant \frac{1}{256} d^2(y_t, y^*(x_{t+1})) \tag{26}$$

Further,

$$\begin{aligned} d(x_{t+1}, x^*) &\overset{①}{\leqslant} d(x_{t+1}, x^*(y_t)) + d(x^*(y_t), x^*) \\ &\overset{②}{\leqslant} \frac{1}{16} d(x_t, x^*) + \frac{17}{16} d(x^*(y_t), x^*) \\ &\overset{③}{\leqslant} \frac{1}{16} d(x_t, x^*) + \frac{17 L_{xy}}{16\mu_x} d(y_t, y^*). \end{aligned} \tag{27}$$

We used the triangle inequality in ①, and ② follows from (25) and the triangular inequality again. Finally, ③ uses Lemma 35, noting that $x^* = x^*(y^*)$. For y, we follow the same argument and then use (27) in ① below

$$\begin{aligned} d(y_{t+1}, y^*) &\leqslant d(y_{t+1}, y^*(x_{t+1})) + d(y^*(x_{t+1}), y^*) \\ &\leqslant \frac{1}{16} d(y_t, y^*) + \frac{17}{16} d(y^*(x_{t+1}), y^*) \\ &\overset{①}{\leqslant} \frac{1}{16} d(y_t, y^*) + \frac{17 L_{xy}}{16\mu_y} \left(\frac{1}{16} d(x_t, x^*) + \frac{17 L_{xy}}{16\mu_x} d(y_t, y^*)\right) \\ &\leqslant \left(\frac{1}{16} + \frac{17^2 {L_{xy}}^2}{16^2 \mu_x \mu_y}\right) d(y_t, y^*) + \frac{17 L_{xy}}{16^2 \mu_y} d(x_t, x^*). \end{aligned} \tag{28}$$

Now we define $C \stackrel{\text{def}}{=} \mu_y/\mu_x$ and obtain

$$\begin{aligned} d^2(x_{t+1}, x^*) + Cd^2(y_{t+1}, y^*) &\overset{①}{\leqslant} 2\left(\frac{1}{16^2} + C\left(\frac{17L_{xy}}{16^2\mu_y}\right)^2\right) d^2(x_t, x^*) \\ &\quad + 2\left(\frac{C}{C}\left(\frac{17L_{xy}}{16\mu_x}\right)^2 + C\left(\frac{1}{16^2} + \left(\frac{17^2 {L_{xy}}^2}{16^2\mu_x\mu_y}\right)^2\right)\right) d^2(y_t, y^*) \\ &\overset{②}{\leqslant} 2d^2(x_t, x^*)\left(\frac{1}{16^2} + \frac{17^2}{16^2}\frac{1}{4}\right) + 2Cd^2(y_t, y^*)\left(\frac{17^2}{16^2}\frac{1}{4} + \left(\frac{1}{16^2} + \frac{1}{4^2}\frac{17^2}{16^2}\right)\right) \\ &\leqslant \frac{3}{5}(d^2(x_t, x^*) + Cd^2(y_t, y^*)). \end{aligned}$$

where ① follows from $(a+b)^2 \leqslant 2a^2 + 2b^2$, (27) and (28). Inequality ② is obtained by using $L_{xy} < \frac{1}{2}\sqrt{\mu_x\mu_y}$, which holds by assumption, and by the definition of C. By expanding the previous inequality, it follows

$$d^2(x_T, x^*) + Cd^2(y_T, y^*) \leqslant \left(\frac{3}{5}\right)^T \left(d^2(x_0, x^*) + Cd^2(y_0, y^*)\right).$$

Now we study two cases. If $C \geqslant 1$, we have ① below

$$d^2(x_T, x^*) + d^2(y_T, y^*) \overset{①}{\leqslant} \left(\frac{3}{5}\right)^T \frac{\mu_y}{\mu_x}\left(d^2(x_0, x^*) + d^2(y_0, y^*)\right) \overset{②}{\leqslant} \left(\frac{3}{5}\right)^T \frac{L_x}{\mu_x}\left(d^2(x_0, x^*) + d^2(y_0, y^*)\right),$$

where ② is due to $\mu_y \leqslant L_y$ and $L_x = L_y$. Recall that we assumed the latter without loss of generality. Similarly, if $C \in (0, 1)$ we have, for $C \leqslant 1$,

$$d^2(x_T, x^*) + d^2(y_T, y^*) \leqslant \left(\frac{3}{5}\right)^T \frac{\mu_x}{\mu_y}\left(d^2(x_0, x^*) + d^2(y_0, y^*)\right) \leqslant \left(\frac{3}{5}\right)^T \frac{L_y}{\mu_y}\left(d^2(x_0, x^*) + d^2(y_0, y^*)\right).$$

Thus, for all $C > 0$, we obtain

$$d^2(x_T, x^*) + d^2(y_T, y^*) \leqslant \left(\frac{3}{5}\right)^T \max\{\kappa_x, \kappa_y\}\left(d^2(x_0, x^*) + d^2(y_0, y^*)\right). \tag{29}$$

Hence, we require

$$T' = \mathcal{O}\left(\log\left(\frac{(d^2(x_0, x^*) + d^2(y_0, y^*))(\kappa_x + \kappa_y)}{\varepsilon}\right)\right),$$

iterations of Algorithm 4 to ensure $d^2(x_T, x^*) + d^2(y_T, y^*) \leqslant \varepsilon$. By Corollary 26, each RiemaconRel call requires $T'_x = \widetilde{O}(\zeta_R\zeta\sqrt{\kappa_x})$ and $T'_y = \widetilde{O}(\zeta_R\zeta\sqrt{\kappa_y})$ calls to the gradient and metric projection oracle respectively. In total, RABR requires

$$T'(T_xT'_x + T_yT'_y) = \widetilde{O}(\zeta_R{\zeta}^2\sqrt{\kappa_x + \kappa_y})$$

calls to the gradient and metric projection oracle. ■

E.4 Convergence analysis of RAMMA

We first prove this important proposition, that describes how to go from one measure of convergence to another, possibly after performing some optimization steps.

Proof of Lemma 8. Statement 1 follows from the non-negativity of the gaps and $\text{gap}(\bar{x}) + \text{gap}(\bar{y}) = \text{gap}(\bar{x}, \bar{y})$. Statement 2 follows from the strong convexity of ϕ_x and ϕ_y.

Table 2: Overview of precision and iteration parameters for RAMMA

Number of iterations	Required precisions
$T_1 = \widetilde{O}(\sqrt{\frac{\zeta}{\mu_x \eta_x}})$	$\varepsilon_1 = \frac{\varepsilon \mu_x}{4C}\left(\frac{2{L_{xy}}^2}{{\mu_y}^2}+1\right)^{-1}, \quad \hat{\varepsilon}_1 = \frac{\varepsilon_1 (\eta_x \mu_x)^{\frac{3}{2}}}{4\sqrt{\xi}}$
$T_2 = \widetilde{O}(\zeta^{\frac{5}{2}}\sqrt{\zeta + \frac{L_{xy}+L_y}{\mu_y}})$	$\varepsilon_2 = \frac{\mu_y \varepsilon}{8C}$
$T_3 = \widetilde{O}\left(\sqrt{\frac{\zeta}{\mu_y \eta_y}}\right)$	$\varepsilon_3 = \frac{\mu_y {\varepsilon_1}^2 (\mu_x \eta_x)^3}{64 C_k^2 \xi \left(\frac{2{L_{xy}}^2}{\mu_x + \eta_{-x}^{-1}}+1\right)}, \quad \hat{\varepsilon}_3 = \varepsilon_3 (\eta_y \mu_y)^{-3/2}/(8\sqrt{\xi})$
$T_4 = \widetilde{O}(\zeta^{\frac{5}{2}}\sqrt{\kappa_x + \zeta})$	$\varepsilon_4 = \frac{(\mu_x + \eta_x^{-1})(\mu_x \eta_x)^3 {\varepsilon_1}^2}{128 C_k^2 \xi}$
$T_5 = \widetilde{O}(\zeta^3\sqrt{L_x \eta_x + L_y \eta_y + \zeta})$	$\varepsilon_5 = \frac{{\varepsilon_3}^2 (\mu_y \eta_y)}{32 \xi C_\ell^2}$

We now prove Statement 3. By strong concavity of ϕ_y we have $d^2(\bar{y}, y^*) \leqslant \frac{2}{\bar{\mu}_y}\mathrm{gap}(y) \leqslant \frac{2\varepsilon}{\bar{\mu}_y}$. The optimizer of $g(\cdot, \bar{y})$ is $x^*(\bar{y})$, so by $\bar{\mu}_x$-strong g-convexity of this function and optimality of $\bar{x}'$, we have $d^2(\bar{x}', x^*(\bar{y})) \leqslant \frac{2\hat{\varepsilon}}{\bar{\mu}_x}$. Thus, we have

$$\begin{aligned} d^2(\bar{x}', x^*) + d^2(\bar{y}, y^*) &\overset{①}{\leqslant} 2d(\bar{x}', x^*(\bar{y}))^2 + 2d(x^*(\bar{y}), x^*)^2 + d^2(\bar{y}, y^*) \\ &\overset{②}{\leqslant} \frac{4\hat{\varepsilon}}{\bar{\mu}_x^2} + \left(\frac{2\bar{L}_{xy}^2}{\bar{\mu}_x} + 1\right) d^2(\bar{y}, y^*) \leqslant \frac{4\hat{\varepsilon}}{\bar{\mu}_x} + \frac{2\varepsilon}{\bar{\mu}_y}\left(\frac{2\bar{L}_{xy}^2}{\bar{\mu}_x^2} + 1\right). \end{aligned} \tag{30}$$

We used the triangular inequality and Young's in ① and for the second summand of ② we used the $(\bar{L}_{xy}/\bar{\mu}_x)$-Lipschitzness of $x^*(\cdot)$, due to Lemma 35.

Under the assumption of Statement 4, we have that

$$\begin{aligned} \mathrm{gap}(\bar{x}, \bar{y}) &= g(\bar{x}, y^*(\bar{x})) - g(\bar{x}, \bar{y}) + g(\bar{x}, \bar{y}) - g(x^*(\bar{y}), \bar{y}) \\ &\leqslant \bar{L}_p^y d(y^*(\bar{x}), \bar{y}) + \bar{L}_p^x d(x^*(\bar{y}), \bar{x}) \\ &\leqslant \bar{L}_p^y (d(y^*(\bar{x}), y^*) + d(y^*, \bar{y})) + \bar{L}_p^x (d(x^*(\bar{y}), x^*) + d(x^*, \bar{x})) \\ &\overset{①}{\leqslant} \bar{L}_p^y \left(\frac{\bar{L}_{xy}}{\bar{\mu}_y} d(x^*, \bar{x}) + d(y^*, \bar{y})\right) + \bar{L}_p^x \left(\frac{\bar{L}_{xy}}{\bar{\mu}_x} d(\bar{y}, y^*) + d(x^*, \bar{x})\right) \\ &= d(y^*, \bar{y}) \left(\bar{L}_p^y + \bar{L}_p^x \frac{\bar{L}_{xy}}{\bar{\mu}_x}\right) + d(x^*, \bar{x}) \left(\bar{L}_p^x + \bar{L}_p^y \frac{\bar{L}_{xy}}{\bar{\mu}_y}\right). \end{aligned}$$

We used Lemma 35 in ① above. ■

Before we go on to prove Theorem 6, we briefly discuss a technical detail.

Remark 27 (Saddle point assumption) *In Section 1.2 we assume for the sake of clarity that f admits a saddle point $(x^*, y^*) \in \mathcal{X} \times \mathcal{Y}$ satisfying $\nabla f(x^*, y^*) = 0$. However, it is not necessary to assume that the saddle point has zero gradient and a slightly weaker assumption suffices to show our convergence result. From now on, let $(x^*, y^*) \in \mathcal{X} \times \mathcal{Y}$ be a saddle point in $\mathcal{X} \times \mathcal{Y}$ which does not necessarily have zero gradient, and let $(\hat{x}^*, \hat{y}^*) \in \mathcal{M} \times \mathcal{N}$, be a global saddle point such that $\nabla f(\hat{x}^*, \hat{y}^*) = 0$. Then, it suffices to assume that $d^2(x_0, \hat{x}^*) + d^2(y_0, \hat{y}^*) \leqslant D^2$. This allows global saddle points to lie outside $\mathcal{X} \times \mathcal{Y}$. We also note that in fact, our algorithms can work without any assumption on $d^2(x_0, \hat{x}^*) + d^2(y_0, \hat{y}^*)$ by using an upper bound of this distance in our algorithmic parameters.*

Proof of Theorem 6. The total number of gradient and projection oracle calls of Algorithm 1 can be calculated as follows

$$T_1(T_3 T_5 + T_4) + T_2,$$

where T_1 to T_5 refer to the complexity of the different routines, which are provided in Lemmas 28 to 30. We provide an overview of the required ε_i, $\hat{\varepsilon}_i$, T_i in Table 2. We have that

$$T_1 T_3 T_5 = \widetilde{O}\left(\zeta^4 \sqrt{\frac{L_x}{\mu_x \mu_y \eta_y} + \frac{L_y}{\mu_x \mu_y \eta_x} + \frac{\zeta}{\mu_x \mu_y \eta_x \eta_y}}\right).$$

Recall that we assumed without loss of generality that $\mu_y \leqslant \mu_x$. We analyze some cases now. If $\mu_x \leqslant L_{xy}$, we have that $\eta_x^{-1} = L_{xy} + 9\xi\mu_x$, $\eta_y^{-1} = L_{xy} + 9\xi\mu_y$ and

$$O\left(\frac{L_x}{\mu_x\mu_y\eta_y} + \frac{L_y}{\mu_x\mu_y\eta_x} + \frac{\zeta}{\mu_x\mu_y\eta_x\eta_y}\right) = O\left(\frac{\zeta^2 L L_{xy}}{\mu_x\mu_y} + \zeta^3\right).$$

If $L_{xy} \leqslant \mu_x, \mu_y$, we have that $\eta_x^{-1} = (1+9\xi)\mu_x$, $\eta_y^{-1} = (1+9\xi)\mu_y$ and

$$O\left(\frac{L_x}{\mu_x\mu_y\eta_y} + \frac{L_y}{\mu_x\mu_y\eta_x} + \frac{\zeta}{\mu_x\mu_y\eta_x\eta_y}\right) = O\left(\zeta(\kappa_x+\kappa_y) + \zeta^3\right).$$

If $\mu_y \leqslant L_{xy} \leqslant \mu_x$, we have $\eta_x^{-1} = (1+9\xi)\mu_x$,$\eta_y^{-1} = L_{xy} + 9\xi\mu_y$ and it is

$$O\left(\frac{L_x}{\mu_x\mu_y\eta_y} + \frac{L_y}{\mu_x\mu_y\eta_x} + \frac{\zeta}{\mu_x\mu_y\eta_x\eta_y}\right) = O\left(\zeta(\kappa_x+\kappa_y) + \frac{L L_{xy}}{\mu_x\mu_y} + \frac{\zeta^2 L_{xy}}{\min\{\mu_x,\mu_y\}} + \zeta^3\right).$$

All in all the worst case complexity is

$$O\left(\frac{L_x}{\mu_x\mu_y\eta_y} + \frac{L_y}{\mu_x\mu_y\eta_x} + \frac{\zeta}{\mu_x\mu_y\eta_x\eta_y}\right) = O\left(\frac{\zeta^2 L L_{xy}}{\mu_x\mu_y} + \zeta(\kappa_x+\kappa_y) + \zeta^3\right).$$

Hence

$$T_1T_3T_5 = \widetilde{O}(\zeta^{9/2}\sqrt{\frac{\zeta L L_{xy}}{\mu_x\mu_y} + \kappa_x + \kappa_y + \zeta^2}). \tag{31}$$

Further, we have

$$T_1T_4 = \widetilde{O}\left(\zeta^3\sqrt{\frac{\kappa_x+\zeta}{\mu_x\eta_x}}\right).$$

It holds that

$$O\left(\frac{\kappa_x+\zeta}{\mu_x\eta_x}\right) = O\left(\zeta^2 + \kappa_x\zeta + \frac{L_x L_{xy}}{{\mu_x}^2} + \frac{\zeta L_{xy}}{\mu_x}\right),$$

and hence

$$T_1T_4 = \widetilde{O}(\zeta^3\sqrt{\zeta^2 + \kappa_x\zeta + \frac{L_x L_{xy}}{{\mu_x}^2} + \frac{\zeta L_{xy}}{\mu_x}}). \tag{32}$$

Finally

$$T_2 = \widetilde{O}(\zeta^{\frac{5}{2}}\sqrt{\zeta + \frac{L_{xy}+L_y}{\mu_y}}). \tag{33}$$

Using (31) to (33), we conclude that

$$T = \widetilde{O}(\zeta^{9/2}\sqrt{\frac{\zeta L L_{xy}}{\mu_x\mu_y} + \kappa_x + \kappa_y + \zeta^2}),$$

where we used $\mu_y \leqslant \mu_x$, which we recall that was assumed to hold without loss of generality. We note that the dependence on ε is $\log^3(\varepsilon^{-1})$ and the log contains a polylog expression on D and the smoothness and strong convexity constants of f. ■

Lemma 28 (Guarantees of Lines 1-4) *Running Lines 1-4 of Algorithm 1 with* $T_1 = \widetilde{O}(\sqrt{\frac{\zeta}{\mu_x\eta_x}})$ *and* $T_2 = \widetilde{O}(\zeta^{\frac{5}{2}}\sqrt{\zeta + \frac{L_{xy}+L_y}{\mu_y}})$ *ensures that gap*$(\hat{x},\hat{y}) \leqslant \varepsilon$.

Proof We show the lemma by first finding sufficient error criteria ε_1 and ε_2 for obtaining $\text{gap}(\hat{x},\hat{y}) \leqslant \varepsilon$ and then we compute the number of iterations T_1 and T_2 required to achieve these error criteria.

Error criterion Let $f_x \overset{\text{def}}{=} f(\cdot, y)$ and $f_y = f(x, \cdot)$, then using Statement 4 of Lemma 8, we have that

$$\begin{aligned}\text{gap}(\hat{x}, \hat{y}) &\leqslant d(\hat{x}, x^*)(L_p(f_x) + \frac{L_{xy}}{\mu_y} L_p(f_y)) + d(\hat{y}, y^*)(L_p(f_y) + \frac{L_{xy}}{\mu_x} L_p(f_x)) \\ &\leqslant C(d(\hat{x}, x^*) + d(\hat{y}, y^*))\end{aligned} \tag{34}$$

where $L_p(f_x)$ and $L_p(f_y)$ denote the Lipschitz constant of f_x and f_y respectively and

$$C = \max\{L_p(f_x) + \frac{L_{xy}}{\mu_y} L_p(f_y), L_p(f_y) + \frac{L_{xy}}{\mu_x} L_p(f_x)\}.$$

Recall that by assumption, we have that $\nabla_x f(\hat{x}^*, \hat{y}^*) = \nabla_y f(\hat{x}^*, \hat{y}^*) = 0$. We leverage this fact in order to bound the Lipschitz constants. We have for some $x \in \mathcal{X}$

$$\begin{aligned}L_p(f_y) = \max_{y \in \mathcal{Y}} \|\nabla_y f_y(y)\| &= \max_{y \in \mathcal{Y}} \|\nabla_y f(x, y) \pm \nabla_y f(\hat{x}^*, y) - \Gamma_{\hat{y}^*}^y \nabla_y f(\hat{x}^*, \hat{y}^*)\| \\ &\leqslant \max_{y \in \mathcal{Y}} L_y d(\hat{y}^*, y) + L_{xy} d(x, \hat{x}^*) \leqslant D(L_y + L_{xy})\end{aligned} \tag{35}$$

and similarly

$$L_p(f_y) = \max_{x \in \mathcal{X}} \|\nabla_x f_x(x)\| \leqslant D(L_x + L_{xy}). \tag{36}$$

Since $\hat{x}$ is an ε_1-minimizer of the problem $\min_{x \in \mathcal{X}} \phi(x)$ and $\hat{y}$ is an ε_2-minimizer of the problem $\min_{y \in \mathcal{Y}} -f(\hat{x}, y)$, Statement 3 of Lemma 8 implies that

$$d^2(\hat{x}, x^*) + d^2(\hat{y}, y^*) \leqslant \frac{4\varepsilon_2}{\mu_y} + \frac{2\varepsilon_1}{\mu_x} \left(\frac{2{L_{xy}}^2}{{\mu_y}^2} + 1 \right). \tag{37}$$

Using (37), we obtain

$$\text{gap}(\hat{x}, \hat{y}) \leqslant C \left(\frac{4\varepsilon_2}{\mu_y} + \frac{2\varepsilon_1}{\mu_x} \left(\frac{2{L_{xy}}^2}{{\mu_y}^2} + 1 \right) \right).$$

It suffices to choose

$$\varepsilon_1 \leqslant \frac{\varepsilon \mu_x}{4C} \left(\frac{2{L_{xy}}^2}{{\mu_y}^2} + 1 \right)^{-1}, \quad \varepsilon_2 \leqslant \frac{\mu_y \varepsilon}{8C}, \tag{38}$$

in order to ensure that $\text{gap}(\hat{x}, \hat{y}) \leqslant \varepsilon$.

Complexity By Theorem 3, computing $\hat{x}$ with RiemaconAbs takes, by the choice of the corresponding ε_1, computed in (38):

$$T_1 = \widetilde{O}\left(\sqrt{\frac{\zeta}{\mu_x \eta_x}} \right).$$

Using Theorem 3 and by definition of λ_y, we have that running RiemaconAbs in Line 3 requires

$$T_2{}'' = \widetilde{O}\left(\sqrt{\zeta \frac{L_{xy} + L_y}{\mu_y} + \zeta^2} \right)$$

iterations. Further, computing a σ-minimizer of $\min_{y \in \mathcal{Y}} \hat{F}_y(y)$, where $\hat{F}_y(y) \overset{\text{def}}{=} -f(\hat{x}, y) + \frac{1}{2\lambda_y} d^2(y, \bar{y})$ using PRGD costs $T_2{}' = \widetilde{O}(\tilde{\kappa} \zeta_R)$, where

$$\tilde{\kappa} \overset{\text{def}}{=} \frac{L_y}{\mu_y + \lambda_y^{-1}} + \frac{\zeta/\lambda_y}{\mu_y + \lambda_y^{-1}} \leqslant 1 + \zeta$$

is the condition number of $\hat{F}_y$. Note we used $\lambda_y^{-1} = (\max\{L_{xy}, L_y\} + 9\xi\mu_y) \geqslant L_y$. We now show that $R \leqslant 2D$ in order to bound $\zeta_R \leqslant \zeta_{2D} = O(\zeta)$. We bound $R \leqslant L_p(\hat{F}_y)/(L_y + \lambda_y^{-1})$, where $L_p(\hat{F}_y)$ is the Lipschitz constant of $\hat{F}_y(y)$ for $y \in \mathcal{Y}$ and $\bar{y} \in \mathcal{Y}$. Note that $\nabla_y f(\hat{x}^*, \hat{y}^*) = 0$ by assumption. Hence, for all $x \in \mathcal{X}$

$$\begin{aligned}L_p(\hat{F}_y) &\leqslant \max_{y \in \mathcal{Y}} \|\nabla_y \hat{F}(y)\| \\ &= \max_{y \in \mathcal{Y}} \| - \nabla_y f(\hat{x}, y) - \lambda_y^{-1} \log_y(\bar{y}) + \Gamma_{\hat{y}^*}^y \nabla_y f(\hat{x}^*, \hat{y}^*)\| \\ &\leqslant \max_{y \in \mathcal{Y}} \| - \nabla_y f(\hat{x}, y) \pm \nabla_y f(\hat{x}^*, y) + \Gamma_{\hat{y}^*}^y \nabla_y f(\hat{x}^*, \hat{y}^*)\| + \max_{y \in \mathcal{Y}} \lambda_y^{-1} d(y, \bar{y}) \\ &\leqslant (L_y + L_{xy} + \lambda_y^{-1}) D.\end{aligned}$$

And thus it holds that $R \leqslant 2D$. Therefore, the total complexity of computing $\hat{y}$ is

$$T_2 = T_2' T_2'' = \widetilde{O}\left(\zeta^{\frac{5}{2}} \sqrt{\zeta + \frac{L_{xy} + L_y}{\mu_y}}\right).$$

■

Lemma 29 (Guarantees of Lines 6-7) *Running Lines 6-7 of Algorithm 1 with* $T_3 = \widetilde{O}\left(\sqrt{\frac{\zeta}{\mu_y \eta_y}}\right)$ *and* $T_4 = \widetilde{O}(\zeta^{\frac{5}{2}} \sqrt{\kappa_x + \zeta})$ *ensures that* $gap_k(\tilde{x}_k) \leqslant \hat{\varepsilon}_1$.

Proof We show the lemma by first finding sufficient error criteria ε_3 and ε_4 for obtaining $\text{gap}_k(\tilde{x}_k) \leqslant \hat{\varepsilon}_1$ and then computing the number of iterations T_3 an T_4 required to achieve these error criteria.

Error criterion Let $G_x(x) \stackrel{\text{def}}{=} f(x,y) + \frac{1}{2\eta_x} d^2(x_k, x)$ and $G_y(y) \stackrel{\text{def}}{=} f(x,y) + \frac{1}{2\eta_x} d^2(x_k, x)$, then we have

$$\begin{aligned}
\text{gap}_k(\tilde{x}_k) &\overset{①}{\leqslant} \text{gap}_k(\tilde{x}_k, \tilde{y}_k) \\
&\overset{②}{\leqslant} d(\hat{x}, x_k^*)\left(L_p(G_x) + \frac{L_{xy}}{\mu_x} L_p(G_y)\right) + d(\hat{y}, y_k^*)\left(L_p(G_x)\frac{L_{xy}}{\mu_y} + L_p(G_y)\right) \\
&\overset{③}{\leqslant} C_k(d(\hat{x}, x_k^*) + d(\hat{y}, y_k^*)) \\
&\overset{④}{\leqslant} C_k \sqrt{2(d^2(\hat{x}, x_k^*) + d^2(\hat{y}, y_k^*))}.
\end{aligned} \tag{39}$$

Here ① and ② hold by Statements 1 and 4 in Lemma 8, respectively, and ③ holds with

$$C_k = \max\left\{L_p(G_x) + \frac{L_{xy}}{\mu_x} L_p(G_y), L_p(G_x)\frac{L_{xy}}{\mu_y} + L_p(G_y), \right\}.$$

Finally, ④ follows from $a + b \leqslant \sqrt{2(a^2 + b^2)}$. We bound the Lipschitz constant of G_x by bounding the following, for all $x \in \mathcal{X}$

$$\|\nabla_x F(x,y)\| = \|\nabla_x f(x,y) \pm \nabla_x f(x, \hat{y}^*) - \frac{1}{\eta_x} \log_x(x_k) - \Gamma_{\hat{x}^*}^x \nabla_x f(\hat{x}^*, \hat{y}^*)\| \leqslant D(L_x + L_{xy} + \eta_x^{-1}),$$

and thus $L_p(G_x) \leqslant D(L_x + L_{xy} + \eta_x^{-1})$. Similarly, we obtain $L_p(G_y) \leqslant D(L_y + L_{xy})$. Since $\tilde{y}_k$ is an ε_3-minimizer of the problem $\min_{y \in \mathcal{Y}} \psi(y)$ and $\tilde{x}_k$ is an ε_4-minimizer of the problem $\min_{x \in \mathcal{X}}\{f(x, \tilde{y}_k) + \frac{1}{2\eta_x} d^2(x_k, x)\}$, Statement 3 in Lemma 8 implies that

$$d^2(\tilde{x}_k, x_k^*) + d^2(\tilde{y}_k, y_k^*) \leqslant \frac{4\varepsilon_4}{\mu_x + \eta_x^{-1}} + \frac{2\varepsilon_3}{\mu_y}\left(\frac{2{L_{xy}}^2}{(\mu_x + \eta_x^{-1})^2} + 1\right). \tag{40}$$

Then, using (40) and (39), we have that

$$\text{gap}_k(\tilde{x}_k) \leqslant \sqrt{2} C_k \sqrt{\frac{4\varepsilon_4}{\mu_x + \eta_x^{-1}} + \frac{2\varepsilon_3}{\mu_y}\left(\frac{2{L_{xy}}^2}{(\mu_x + \eta_x^{-1})^2} + 1\right)}.$$

We have that $\hat{\varepsilon}_1 = \frac{\varepsilon_1 (\eta_x \mu_x)^{\frac{3}{2}}}{4\sqrt{\xi}}$ Hence, choosing

$$\varepsilon_3 = \frac{\mu_y {\varepsilon_1}^2 (\mu_x \eta_x)^3}{64 C_k^2 \xi \left(\frac{2{L_{xy}}^2}{\mu_x + \eta_x^{-1}} + 1\right)}, \quad \varepsilon_4 = \frac{(\mu_x + \eta_x^{-1})(\mu_x \eta_x)^3 {\varepsilon_1}^2}{128 C_k^2 \xi}. \tag{41}$$

suffices to satisfy $\text{gap}_k(\tilde{x}_k) \leqslant \hat{\varepsilon}_1$.

Complexity By Theorem 3, and using ε_3 computed in (41), we have that computing $\tilde{y}_k$ takes

$$T_3 = \widetilde{O}\left(\sqrt{\frac{\zeta}{\mu_y \eta_y}}\right)$$

iterations. Further, by Corollary 4, we have

$$T_4 = \widetilde{O}(\zeta_R \zeta^{\frac{3}{2}} \sqrt{\tilde{\kappa}_x + \zeta}),$$

where $\tilde{\kappa}_x$ is the condition number of $G_x(x)$. Further, let $\hat{F}_x(x) \stackrel{\text{def}}{=} f(x, \hat{y}) + \frac{1}{2\eta_x} d^2(x_k, x) + \frac{1}{2\lambda_x} d^2(x, \bar{x})$, where $\lambda_x = (L_x + \zeta\eta_x^{-1} + 9\xi\mu_x)$. We bound the Lipschitz constant $L_p(\hat{F}_x)$ for all $x \in \mathcal{X}$ by bounding

$$\begin{aligned} \|\nabla \hat{F}_x(x)\| &\leqslant \|\nabla_x f(x, \hat{y}) - \eta_x^{-1} \log_x(x_k) - \lambda_x^{-1} \log_x(\bar{x}) - \Gamma_{\hat{x}*}^x \nabla_x f(\hat{x}^*, \hat{y}^*)\| \\ &\leqslant D(L_x + L_{xy} + \eta_x^{-1} + \lambda_x^{-1}). \end{aligned} \tag{42}$$

Thus $L_p(\hat{F}_x) \leqslant D(L_x + L_{xy} + \eta_x^{-1}\lambda_x^{-1})$. Hence,

$$R = \frac{L_p(\hat{F}_x)}{L_x + \zeta\eta_x^{-1} + \zeta\lambda_x^{-1}} \leqslant \frac{d(L_x + L_{xy} + \eta_x^{-1} + \lambda_x^{-1})}{L_x + \zeta\eta_x^{-1} + \zeta\lambda_x^{-1}} \leqslant 2D.$$

And so $\zeta_R \leqslant \zeta_{2D} = O(\zeta)$. Further, the condition number of $G_x(x)$ can be bounded by

$$\tilde{\kappa}_x = \frac{L_x}{\mu_x + \eta_x^{-1} + \lambda_x^{-1}} + \frac{\zeta(\eta_x^{-1} + \lambda_x^{-1})}{\mu_x + \eta_x^{-1} + \lambda_x^{-1}} \leqslant 1 + \zeta.$$

Finally, we obtain

$$T_4 = \widetilde{O}(\zeta^{5/2} \sqrt{\kappa_x + \zeta}).$$

■

Lemma 30 (Guarantees of Lines 9-10) *Let* gap_ℓ *refers to the gap of the problem* $\min\max\{f(x,y) + \frac{1}{2\eta_x} d^2(x_k, x) - \frac{1}{2\eta_y} d^2(y_\ell, y)\}$. *Running Lines 9-10 of Algorithm 1 with* $T_5 = \widetilde{O}(\zeta^3 \sqrt{L_x\eta_x + L_y\eta_y + \zeta})$ *ensures that* $\text{gap}_\ell(\bar{y}_\ell) \leqslant \hat{\varepsilon}_3$.

Proof We show the lemma by first finding a sufficient error criterion ε_5 for obtaining $\text{gap}_\ell(\bar{x}_\ell, \bar{y}_\ell) \leqslant \hat{\varepsilon}_3$, and then we computing the number of iterations T_5 required to achieve it. This bound implies the result since by Statement 1 of Lemma 8, it is $\text{gap}_\ell(\bar{y}_\ell) \leqslant \text{gap}_\ell(\bar{x}_\ell, \bar{y}_\ell)$.

Error criterion Write $h(x, y) \stackrel{\text{def}}{=} f(x, y) + \frac{1}{2\eta_x} d^2(x, x_k) - \frac{1}{2\eta_y} d^2(y, y_\ell)$. Then, we can bound the Lipschitz constant of $h(\cdot, \bar{y}_\ell)$ as

$$\begin{aligned} L_p(h(\cdot, \bar{y}_\ell)) &= \max_{x \in \mathcal{X}} \|\nabla_x h(x, \bar{y}_\ell)\| \\ &= \max_{x \in \mathcal{X}} \|\nabla_x f(x, \bar{y}_\ell) - \eta_x^{-1} \log_x(x_k) \pm \nabla_x f(x, \hat{y}^*) - \Gamma_{\hat{x}*}^x \nabla_x f(\hat{x}^*, \hat{y}^*)\| \\ &\leqslant D(\eta_x^{-1} + L_x + L_{xy}). \end{aligned} \tag{43}$$

and similarly, for any $x \in \mathcal{X}$, we bound the Lipschitz constant of $h(\bar{x}_\ell, \cdot)$ as follows

$$L_p(h(\bar{x}_\ell, \cdot)) \leqslant \max_{y \in \mathcal{Y}} \|\nabla_y h(\bar{x}_\ell, y)\| \leqslant D(\eta_y^{-1} + L_y + L_{xy}). \tag{44}$$

we have that for

$$C_\ell = D \max\left\{ \frac{L_{xy}(\eta_x^{-1} + L_x + L_{xy})}{\mu_x} + \eta_y^{-1} + L_y + L_{xy}, \frac{L_{xy}(\eta_y^{-1} + L_y + L_{xy})}{\mu_y} + \eta_x^{-1} + L_x + L_{xy} \right\},$$

and defining

$$\varepsilon_5 \stackrel{\text{def}}{=} \frac{{\varepsilon_3}^2(\mu_y \eta_y)}{32\xi C_\ell^2},$$

the following holds, as desired:

$$\begin{aligned}
&\mathrm{gap}_\ell(\bar{x}_\ell, \bar{y}_\ell) \\
&\overset{①}{\leqslant} d(\bar{x}_\ell, x_\ell^*)\left(\frac{L_{xy}}{\mu_x} L_p(h(\cdot, \bar{y}_\ell)) + L_p(h(\bar{x}_\ell, \cdot))\right) + d(\bar{y}_\ell, y_\ell^*)\left(\frac{L_{xy}}{\mu_y} L_p(h(\cdot, \bar{x}_\ell)) + L_p(h(\bar{y}_\ell, \cdot))\right) \\
&\overset{②}{\leqslant} C_\ell[d(\bar{x}_\ell, x_\ell^*) + d(\bar{y}_\ell, y_\ell^*)] \overset{③}{\leqslant} C_\ell\sqrt{2\varepsilon_5} \overset{④}{\leqslant} \varepsilon_3(\eta_y\mu_y)^{-3/2}/(8\sqrt{\xi}) \overset{⑤}{=} \hat{\varepsilon}_3.
\end{aligned} \tag{45}$$

We used Statement 4 in Lemma 8 for ① and the definition of C_ℓ in ②. Theorem 5 implies $d^2(\bar{x}_\ell, x_\ell^*) + d^2(\bar{y}_\ell, y_\ell^*) \leqslant \varepsilon_5$ which was used for ③ along with $(a+b)^2 \leqslant 2a^b + 2b^2$. We defined ε_5 in order to satisfy ④. The definition of the accuracy $\hat{\varepsilon}_3$ that we require in RiemaconAbs was used in ⑤.

Complexity By Theorem 5, and the definition of ε_5, computing $\bar{y}_\ell$ takes

$$T_5 = \widetilde{O}(\zeta_R \zeta^2 \sqrt{\tilde{\kappa}_x + \tilde{\kappa}_y + \zeta}),$$

iterations of RABR, where $\tilde{\kappa}_x$ is the condition number of $h(\cdot, \bar{y}_\ell)$ and $\tilde{\kappa}_y$ is the condition number of h_y. Using (43) and the definition of η_x, we have

$$R \leqslant \frac{\max_{y\in\mathcal{Y}} L_p(h(\cdot, y))}{L_x + \frac{\zeta}{\eta_x}} \leqslant \frac{D(\eta_x^{-1} + L_x + L_{xy})}{L_x + \frac{\zeta}{\eta_x}} \leqslant 2D$$

and similarly, by (44) and the definition of η_y it holds

$$R \leqslant \frac{\max_{x\in\mathcal{X}} L_p(h(x, \cdot))}{L_y + \frac{\zeta}{\eta_y}} \leqslant 2D.$$

Hence, we have $\zeta_R \leqslant \zeta_{2D} = \widetilde{O}(\zeta)$. Given that $\tilde{\kappa}_x \leqslant \eta_x L_x + \zeta$ and $\tilde{\kappa}_y \leqslant L_y\eta_y + \zeta$, we conclude that

$$T_5 = \widetilde{O}(\zeta^3 \sqrt{L_x\eta_x + L_y\eta_y + \zeta}).$$

■

Proof of Corollary 7. Let $(\bar{x}, \bar{y}) \in \mathcal{X} \times \mathcal{Y}$ be the initial point of our algorithm, and define the following regularized function

$$f_\varepsilon(x, y) \overset{\text{def}}{=} f(x, y) + \frac{\varepsilon}{4D^2} d^2(\bar{x}, x) - \frac{\varepsilon}{4D^2} d^2(\bar{y}, y).$$

Interestingly, we require the use of this function for both the CC and the SCC case. That is, we require regularizing both variables even when the function is strongly g-convex with respect to one. This is done in order to show that the global saddle point of the regularized problem $(\hat{x}_\varepsilon^*, \hat{y}_\varepsilon^*)$ is not further away from the initial point $(\bar{x}, \bar{y})$ than the global saddle point $(\hat{x}^*, \hat{y}^*)$ of the unregularized problem , i.e. $d^2(\bar{x}, \hat{x}_\varepsilon^*) + d^2(\bar{y}, \hat{y}_\varepsilon^*) \leqslant d^2(\bar{x}, \hat{x}^*) + d^2(\bar{y}, \hat{y}^*) = D^2$ which is required in order bound the geometric penalties. Now let $(\hat{x}, \hat{y})$ be an $\varepsilon/2$ saddle point of f_ε, i.e.

$$\max_{y\in\mathcal{Y}} f_\varepsilon(\hat{x}, y) - \min_{x\in\mathcal{X}} f(x, \hat{y}) \leqslant \frac{\varepsilon}{2}.$$

Let $y^*(\hat{x}) = \arg\max_{y\in\mathcal{Y}} f(\hat{x}, y)$, then

$$\max_{y\in\mathcal{Y}} f_\varepsilon(\hat{x}, y) \geqslant f_\varepsilon(\hat{x}, y^*(\hat{x})) = f(\hat{x}, y^*(\hat{x})) + \frac{\varepsilon}{4D^2} d^2(\bar{x}, \hat{x}) - \frac{\varepsilon}{4D^2} d^2(\bar{y}, \hat{y}^*) \geqslant f(\hat{x}, y^*(\hat{x})) - \frac{\varepsilon}{4}.$$

Similarly, for $x^*(\hat{y}) = \arg\min_{x\in\mathcal{X}} f(x, \hat{y})$, we have $\min_{x\in\mathcal{X}} f_\varepsilon(x, \hat{y}) \leqslant f(x^*(\hat{y}), \hat{y}) + \frac{\varepsilon}{4}$. Combining these inequalities, we conclude

$$\max_{y\in\mathcal{Y}} f(\hat{x}, y) - \min_{x\in\mathcal{X}} f(x, \hat{y}) = f(\hat{x}, y^*(\hat{x})) - f(x^*(\hat{y}), \hat{y}) \leqslant \max_{y\in\mathcal{Y}} f_\varepsilon(\hat{x}, y) - \min_{x\in\mathcal{X}} f_\varepsilon(x, \hat{y}) + \frac{\varepsilon}{4} + \frac{\varepsilon}{4} \leqslant \varepsilon.$$

Hence if $(\hat{x}, \hat{y})$ is an $\varepsilon/2$-saddle point of f_ε it is an ε-saddle point of f. By Lemma 32 and by the definition of $\mathcal{X}$ and $\mathcal{Y}$, we have that the saddle point of f_ε satisfies $d^2(\bar{x}, \hat{x}_\varepsilon^*) + d^2(\bar{y}, \hat{y}_\varepsilon^*) \leqslant D^2$ which is required to use Algorithm 1 on f_ε. Recall that the complexity of the algorithm is

$$\widetilde{O}\left(\zeta^{9/2}\sqrt{\frac{\zeta\tilde{L}L_{xy}}{\tilde{\mu}_x\tilde{\mu}_y} + \tilde{\kappa}_x + \tilde{\kappa}_y + \zeta^2}\right),$$

where the variables noted with a tilde are the constants of f_ε.

We first analyze the SCC case. We have $\tilde{\mu}_x = \mu_x + \varepsilon/(2D^2)$, $\tilde{\mu}_y = \varepsilon/(2D^2)$, $\tilde{L}_x \leqslant L_x + \zeta\varepsilon/(2D^2)$ and therefore the condition numbers are

$$\tilde{\kappa}_x = \frac{L_x + \zeta\frac{\varepsilon}{2D^2}}{\mu_x + \frac{\varepsilon}{2D^2}} \leqslant \frac{L_x}{\mu_x} + \zeta \qquad \text{and} \qquad \tilde{\kappa}_y = \frac{L_y + \zeta\frac{\varepsilon}{2D^2}}{\frac{\varepsilon}{2D^2}} \leqslant \frac{2L_yD^2}{\varepsilon} + \zeta. \tag{46}$$

Note that L_{xy} is not influenced by regularization. We now bound the terms inside the square root of (46). First, assume that $L_{xy} \geqslant \tilde{L}_x$, then $\tilde{L} = L_{xy}$ (recall that $L_x = L_y$ without loss of generality)

$$\frac{\zeta\tilde{L}L_{xy}}{\tilde{\mu}_x\tilde{\mu}_y} \leqslant \frac{2\zeta {L_{xy}}^2D^2}{\mu_x\varepsilon} = \frac{2\zeta LL_{xy}D^2}{\mu_x\varepsilon}.$$

Now assume that $L_{xy} \leqslant \tilde{L}_x$, then $\tilde{L} = L_x + \zeta\frac{\varepsilon}{2D^2}$ and

$$\frac{\zeta\tilde{L}L_{xy}}{\tilde{\mu}_x\tilde{\mu}_y} = \frac{\zeta L_{xy}(L_x + \zeta\frac{\varepsilon}{2D^2})}{\tilde{\mu}_x\tilde{\mu}_y} \leqslant \frac{2\zeta L_{xy}LD^2}{\mu_x\varepsilon} + \frac{\zeta^2L_{xy}}{\tilde{\mu}_x} \leqslant \frac{2\zeta L_{xy}LD^2}{\mu_x\varepsilon} + \frac{\zeta^2L}{\mu_x}.$$

Using this bound in (46), we have

$$\widetilde{O}\left(\frac{\zeta\tilde{L}L_{xy}}{\tilde{\mu}_x\tilde{\mu}_y} + \tilde{\kappa}_x + \tilde{\kappa}_y + \zeta^2\right) = \widetilde{O}\left(\frac{\zeta^2L}{\mu_x} + \frac{\zeta L_{xy}LD^2}{\mu_x\varepsilon} + \frac{L_yD^2}{\varepsilon}\right).$$

The resulting complexity is

$$\widetilde{O}\left(\zeta^{9/2}\sqrt{\frac{\zeta^2L}{\mu_x} + \frac{\zeta L_{xy}LD^2}{\mu_x\varepsilon} + \frac{L_yD^2}{\varepsilon}}\right)$$

Now we proceed to analyze the CC case. We have $\tilde{\mu}_x = \mu_x + \varepsilon/(2D^2)$, $\tilde{\mu}_y = \varepsilon/(2D^2)$, $\tilde{L}_x \leqslant L_x + \zeta\varepsilon/(2D^2)$ and therefore the condition numbers are

$$\tilde{\kappa}_x = \frac{L_x + \zeta\frac{\varepsilon}{2D^2}}{\mu_x + \frac{\varepsilon}{2D^2}} \leqslant \frac{2L_xD^2}{\varepsilon} + \zeta \qquad \text{and} \qquad \tilde{\kappa}_y = \frac{L_y + \zeta\frac{\varepsilon}{2D^2}}{\mu_y + \frac{\varepsilon}{2D^2}} \leqslant \frac{2L_yD^2}{\varepsilon} + \zeta.$$

First, assume that $L_{xy} \geqslant \tilde{L}_x$, then $\tilde{L} = L_{xy}$ (recall that $L_x = L_y$ without loss of generality)

$$\frac{\zeta\tilde{L}L_{xy}}{\tilde{\mu}_x\tilde{\mu}_y} \leqslant \frac{4\zeta LL_{xy}D^4}{\varepsilon^2}$$

Now assume that $L_{xy} \leqslant \tilde{L}_x$, then $\tilde{L} = L_x + \zeta\frac{\varepsilon}{2D^2}$ and

$$\frac{\zeta\tilde{L}L_{xy}}{\tilde{\mu}_x\tilde{\mu}_y} \leqslant \frac{4\zeta L_{xy}L_xD^4}{\varepsilon^2} + \frac{2\zeta^2L_{xy}D^2}{\varepsilon}.$$

Together, we have that

$$\widetilde{O}\left(\frac{\zeta\tilde{L}L_{xy}}{\tilde{\mu}_x\tilde{\mu}_y} + \tilde{\kappa}_x + \tilde{\kappa}_y + \zeta^2\right) = \widetilde{O}\left(\frac{\zeta LL_{xy}D^4}{\varepsilon^2} + \frac{\zeta^2L_{xy}D^2}{\varepsilon} + \frac{D^2(L_x + L_y)}{\varepsilon}\right).$$

Thus, the complexity is bounded by

$$\widetilde{O}\left(\zeta^{9/2}\left(\sqrt{\frac{L_xD^2}{\varepsilon} + \frac{L_yD^2}{\varepsilon} + \frac{\zeta L_{xy}D^2}{\varepsilon}\left(\frac{LD^2}{\varepsilon} + \zeta\right)}\right)\right) = \widetilde{O}\left(\zeta^{9/2}\sqrt{\frac{LD^2}{\varepsilon}} + \zeta^{11/2}\frac{D^2\sqrt{LL_{xy}}}{\varepsilon}\right).$$

■

E.5 Technical Results

Lemma 31 *Let $\mathcal{M}, \mathcal{N}$ be Riemannian manifolds and let $\mathcal{X} \subset \mathcal{M}$, $\mathcal{Y} \subset \mathcal{N}$ be g-convex subsets that are uniquely geodesic. Let $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be such that $f(\cdot, y)$ is lower semicontinuous, $f(x, \cdot)$ is upper semicontinuous, and $f(x, y)$ is $(\mu_x, 0)$-SCSC in $\mathcal{X} \times \mathcal{Y}$. Then $\phi(x) \stackrel{\text{def}}{=} \sup_{y \in \mathcal{Y}} f(x, y)$ is μ_x-strongly g-convex in its domain. Also, if f is sup-compact, $\phi(x)$ is well defined for all $x \in \mathcal{X}$ and it holds $\phi(x) = \max_{y \in \mathcal{Y}} f(x, y)$.*

Proof Let x_1, x_2 be two points in the domain of ϕ and let γ be the geodesic joining $\gamma(0) = x_1$ and $\gamma(t) = x_2$ with $t \in [0, 1]$. Then, we have for all $y \in \mathcal{Y}$ that

$$f(\gamma(t), y) \overset{①}{\leqslant} tf(x_1, y) + (1-t)f(x_2, y) - \frac{t(1-t)\mu_x}{2} d^2(x_1, x_2)$$
$$\overset{②}{\leqslant} t\phi(x_1) + (1-t)\phi(x_2) - \frac{t(1-t)\mu_x}{2} d^2(x_1, x_2).$$

Here, ① holds by μ_x-strong g-convexity of $f(\cdot, y)$ and ② uses the definition of ϕ. Since the inequality holds for all y, it also holds for the supremum, proving that ϕ is μ_x strongly g-convex.

Now we show that if f is sup-compact for some $\tilde{x} \in \mathcal{X}$, then $\phi(x) = \max_{y \in \mathcal{Y}} f(x, y)$ for all $x \in \mathcal{X}$. To that aim, we show that the superlevel sets of $f(x, \cdot)$ are compact for all $x \in \mathcal{X}$. We have that $\{y \in \mathcal{Y} | f(\tilde{x}, y) \geqslant \alpha\}$ is compact because $f(\tilde{x}, \cdot)$ is sup-compact. We have that $\mathcal{Y}_\cap = \{y \in \mathcal{Y} | f(x, y) \geqslant \alpha, \forall x \in \mathcal{X}\} = \bigcap_{x \in \mathcal{X}} \{y \in \mathcal{Y} | f(x, y) \geqslant \alpha\}$ is closed because $f(x, \cdot)$ is upper semicontinuous. Further, $\mathcal{Y}_\cap \subset \{y \in \mathcal{Y} | f(\tilde{x}, y) \geqslant \alpha\}$, hence $\mathcal{Y}_\cap$ is compact since it is the intersection of a closed and a compact set. By the extreme value theorem, an upper semicontinuous function reaches its maximum over a compact set, hence $\phi(x) = \max_{y \in \mathcal{Y}} f(x, y)$. ■

Lemma 32 [↓] *Consider a function $f : \mathcal{M} \times \mathcal{N} \to \mathbb{R}$ as described in Section 1.2. Further, let $h(x, y) = f(x, y) + \frac{1}{2\eta} d^2(\tilde{x}, x) - \frac{1}{2\eta} d^2(\tilde{y}, y)$ with*

$$(\tilde{x}^*, \tilde{y}^*) \stackrel{\text{def}}{=} \arg\min_{x \in \mathcal{M}} \arg\max_{y \in \mathcal{N}} h(x, y).$$

Then, $d^2(\tilde{x}, \tilde{x}^*) + d^2(\tilde{y}, \tilde{y}^*) \leqslant d^2(\tilde{y}, \hat{y}^*) + d^2(\tilde{x}, \hat{x}^*)$, where $(\hat{x}^*, \hat{y}^*)$ is the unconstrained saddle point of f.

Proof of Lemma 32. Note that, by Theorem 16, h admits an unconstrained saddle point $(\tilde{x}^*, \tilde{y}^*)$. We have that,

$$\begin{aligned}
&\frac{1}{2\eta} d^2(\tilde{x}, \tilde{x}^*) - \frac{1}{2\eta} d^2(\tilde{y}, \hat{y}^*) + \frac{1}{2\eta} d^2(\tilde{y}, \tilde{y}^*) - \frac{1}{2\eta} d^2(\tilde{x}, \hat{x}^*) \\
\leqslant &f(\tilde{x}^*, \hat{y}^*) + \frac{1}{2\eta} d^2(\tilde{x}, \tilde{x}^*) - \frac{1}{2\eta} d^2(\tilde{y}, \hat{y}^*) - f(\hat{x}^*, \tilde{y}^*) + \frac{1}{2\eta} d^2(\tilde{y}, \tilde{y}^*) - \frac{1}{2\eta} d^2(\tilde{x}, \hat{x}^*) \\
= &h(\tilde{x}^*, \hat{y}^*) - h(\hat{x}^*, \tilde{y}^*) \leqslant 0.
\end{aligned}$$

It follows that

$$d^2(\tilde{x}, \tilde{x}^*) + d^2(\tilde{y}, \tilde{y}^*) \leqslant d^2(\tilde{y}, \hat{y}^*) + d^2(\tilde{x}, \hat{x}^*).$$

■

Proposition 33 [↓] *Consider a g-convex function $f : \mathcal{X} \to \mathbb{R}$, where $\mathcal{X} \subset \mathcal{M}$ is a compact convex subset of a Riemannian Manifold $\mathcal{M}$. Then for $x^* \in \arg\min_{x \in \mathcal{X}} f(x)$, it holds that*

$$\langle \nabla f(x^*), \operatorname{Log}_{x^*}(y) \rangle \geqslant 0, \forall y \in \mathcal{X}. \tag{47}$$

Note that this also directly implies that if we have a g-concave function f and $x^* \in \arg\max_{x \in \mathcal{X}} f(x)$, then $\langle \nabla f(x^*), \operatorname{Log}_{x^*}(y) \rangle \leqslant 0, \forall y \in \mathcal{X}$.

Proof of Proposition 33. Let f be g-convex and $x^* \in \arg\min_{x \in \mathcal{X}} f(x)$. Let $F(t) \stackrel{\text{def}}{=} f(\gamma(t))$, where $\gamma(t)$ is a geodesic such that $\gamma(0) = x^*$ and $\gamma(1) = x$. Then F reaches its minimum at $t = 0$ and we have that $0 \leqslant F'(0) = \langle \nabla f(x^*), \operatorname{Log}_{x^*}(x) \rangle$. ■

Lemma 34 *For a μ-strongly g-convex function $f : \mathcal{X} \to \mathbb{R}$ where $\mathcal{X} \subset \mathcal{M}$ is a compact g-convex subset we have*

$$\begin{aligned} \frac{\mu}{2} d^2(x^*, y) &\leqslant f(y) - f(x^*) \\ \mu d^2(x, y) &\leqslant \langle \mathrm{Log}_x(y), \Gamma_y^x \nabla f(y) - \nabla f(x) \rangle \\ &= \langle \log_y(x), \Gamma_x^y \nabla f(x) - \nabla f(y) \rangle \end{aligned} \tag{48}$$

where $x^* = \arg\min_{x \in \mathcal{X}} f(x)$. Equivalently if f is μ-strongly g-concave then

$$\begin{aligned} \frac{\mu}{2} d^2(x^*, y) &\leqslant f(x^*) - f(y) \\ \mu d^2(x, y) &\leqslant \langle -\mathrm{Log}_x(y), \Gamma_y^x \nabla f(y) - \nabla f(x) \rangle \\ &= \langle -\log_y(x), \nabla f(x) - \Gamma_y^x \nabla f(y) \rangle \end{aligned} \tag{49}$$

where $x^* = \arg\max_{x \in \mathcal{X}} f(x)$.

Proof By strong convexity, we have

$$f(x^*) \leqslant f(y) + \langle \nabla f(x^*), -\mathrm{Log}_{x^*}(y) \rangle - \frac{\mu}{2} d^2(x^*, y).$$

Using (47), we conclude that

$$\frac{\mu}{2} d^2(x^*, y) \leqslant f(y) - f(x^*).$$

Further, by strong convexity, we can write

$$\begin{aligned} f(x) &\leqslant f(y) + \langle \nabla f(x), -\mathrm{Log}_x(y) \rangle - \frac{\mu}{2} d^2(x, y) \\ f(y) &\leqslant f(x) + \langle \nabla f(y), -\mathrm{Log}_y(x) \rangle - \frac{\mu}{2} d^2(x, y). \end{aligned}$$

Adding both inequalities, we get

$$\begin{aligned} \mu d^2(x, y) &\leqslant \langle \nabla f(y), -\mathrm{Log}_y(x) \rangle + \langle \nabla f(x), -\mathrm{Log}_x(y) \rangle \\ &= \langle \Gamma_y^x \nabla f(y) - \nabla f(x), \log_x(y) \rangle \\ &= \langle \Gamma_x^y \nabla f(x) - \nabla f(y), \log_y(x) \rangle \end{aligned}$$

where the last two equalities follow from transporting the scalar products to $\mathcal{T}_x$ and $\mathcal{T}_y$ respectively. The proof for the strongly g-concave follows by the same argument. ■

Lemma 35 *Assume f satisfies Assumption 1, define $y^*(x) \stackrel{\text{def}}{=} \arg\max_{y \in \mathcal{Y}} f(x, y)$, $x^*(y) \stackrel{\text{def}}{=} \arg\max_{x \in \mathcal{X}} f(x, y)$, $\phi(x) \stackrel{\text{def}}{=} \max_{y \in \mathcal{Y}} f(x, y)$ and $\Psi(y) \stackrel{\text{def}}{=} \min_{x \in \mathcal{X}} f(x, y)$ then*

1. *$y^*(\cdot)$ is $\frac{L_{xy}}{\mu_y}$-Lipschitz.*
2. *$x^*(\cdot)$ is $\frac{L_{xy}}{\mu_x}$-Lipschitz.*
3. *$\phi(x)$ is μ_x-strongly g-convex.*
4. *$\Psi(y)$ is μ_y-strongly g-concave.*

Proof By Proposition 33, we have

$$\langle \mathrm{Log}_{y^*(x)}(y), \nabla_y f(x, y^*(x)) \rangle \leqslant 0, \quad \forall y \in \mathcal{Y} \tag{50}$$

$$\langle \mathrm{Log}_{y^*(z)}(y), \nabla_y f(z, y^*(z)) \rangle \leqslant 0, \quad \forall y \in \mathcal{Y} \tag{51}$$

Sum up (50) with $y = y^*(z)$ transporting the scalar product to $\mathcal{T}_{y^*(z)}$ and (51) with $y = y^*(x)$,

$$\langle \mathrm{Log}_{y^*(z)}(y^*(x)), \nabla_y f(z, y^*(z)) - \Gamma^{y^*(z)} \nabla_y f(x, y^*(x)) \rangle \leqslant 0.$$

Then, since $f(x, \cdot)$ is μ_y-strongly g-concave, we have by Lemma 34

$$\mu_y d^2(y^*(x), y^*(z)) + \langle \mathrm{Log}_{y^*(z)}(y^*(x)), \Gamma^{y^*(z)} \nabla_y f(x, y^*(x)) - \nabla_y f(x, y^*(z) \rangle \leqslant 0.$$

Summing these two equations we get

$$\mu_y d^2(y^*(x), y^*(z)) + \langle \mathrm{Log}_{y^*(z)}(y^*(x)), \nabla_y f(z, y^*(z)) - \nabla_y f(x, y^*(z) \rangle \leqslant 0.$$

Further,

$$\begin{aligned} d(y^*(x), y^*(z)) &\leqslant d^{-1}(y^*(x), y^*(z)) {\mu_y}^{-1} \langle \mathrm{Log}_{y^*(z)}(y^*(x)), \nabla_y f(x, y^*(z) - \nabla_y f(z, y^*(z)) \rangle \\ &\leqslant {\mu_y}^{-1} \| \nabla_y f(x, y^*(z) - \nabla_y f(z, y^*(z)) \| \\ &\leqslant \frac{L_{xy}}{\mu_y} d(x, z), \end{aligned}$$

where we used gradient Lipschitzness. This concludes the proof for Statement 1. The proof of Statement 2 works in the same way using strong g-convexity instead of strong g-concavity. The proofs of Statements 3 and 4 follow directly by Lemma 31. ■