Revisiting Riemannian Gradient Descent and Riemannian Proximal Point

David Martínez-Rubio
Zuse Institute and Technische Universität Berlin
Berlin, Germany

Christophe Roux
Zuse Institute Berlin
Berlin, Germany

Sebastian Pokutta
Zuse Institute and Technische Universität Berlin
Berlin, Germany

Abstract

Convergence rates of Riemannian optimization algorithms often depend on geometric quantities determined by the sectional curvature and the distance between iterates and an optimizer. Numerous previous works bound the latter only by assumption, leading to an incomplete analysis and unquantified rates. In this work, we remove this limitation for two of the most fundamental algorithms in geodesically convex optimization: Riemannian gradient descent and inexact Riemannian proximal point. As a consequence, we are able to quantify their rates of convergence. The iterates stay naturally in a ball around an optimizer, of radius depending on the initial distance and, in some cases, on the curvature. Convergence of the proximal point was unknown when positive curvature was present. We also prove new properties of the Riemannian proximal operator, of independent interest: it does not move points away from any optimizer, and we quantify the smoothness of its induced Moreau envelope. We explore beyond our theory with empirical tests.

1 Introduction

Riemannian optimization is the study of optimizing functions defined over Riemannian manifolds. This paradigm is used in cases that naturally present Riemannian constraints, which allows for exploiting the geometric structure of our problem, and for transforming it into an unconstrained one by working in the manifold. In addition, there are non-convex Euclidean problems, such as operator scaling (Allen-Zhu et al., 2018) that, when phrased over a Riemannian manifold with the right metric, become convex when restricted to every geodesic, that is, they are geodesically convex (g-convex) (Cruz Neto et al., 2006; Carvalho Bento and Melo, 2012; Bento et al., 2015).

Some other applications in machine learning are Gaussian mixture models (Hosseini and Sra, 2015), Karcher mean (Zhang et al., 2016), dictionary learning (Cherian and Sra, 2017; Sun et al., 2017), low-rank matrix completion (Vandereycken, 2013; Mishra and Sepulchre, 2014; Tan et al., 2014; Cambier and Absil, 2016; Heidel and Schulz, 2018), and optimization under orthogonality constraints (Edelman et al., 1998; Lezcano-Casado and Martínez-Rubio, 2019). Riemannian optimization is a wide, active area of research, and numerous generalizations of Euclidean methods, such as the following first-order methods have been designed: projection-free (Weber and Sra, 2017, 2019), accelerated (Martínez-Rubio, 2020; Kim and Yang, 2022; Martínez-Rubio and Pokutta, 2023), min-max (Zhang et al., 2022; Jordan et al., 2022; Martínez-Rubio et al., 2023; Cai et al., 2023), stochastic (Tripuraneni et al., 2018; Khuzani and Li, 2017; Hosseini and Sra, 2017), and variance-reduced methods (Zhang et al., 2016; Sato et al., 2017, 2019), among many others.

A recurrent problem many works in Riemannian optimization incur is that geometric deformations appear in their analyses and while they scale with the distance between iterates and between those and an optimizer, these distances are often bounded and quantified only by assumption (Zhang and Sra, 2016; Zhang et al., 2016; Zhang and Sra, 2018; Ahn and Sra, 2020; Kim and Yang, 2022; Zhang et al., 2022; Jordan et al., 2022), among others.

On the other hand, some works obtain convergence rates which are seemingly independent of the curvature, but they make use of conditions like smoothness or strong convexity without specifying where these have to hold (Smith, 1994; Udriste, 1994; Cai et al., 2023), among others. This is a problem, since unlike in the Euclidean space, where we can have globally smooth and strongly g-convex functions with constant condition number, in general Riemannian manifolds the condition number is lower bounded by a value

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

that depends on the curvature and the diameter of the optimization domain (Martínez-Rubio, 2020; Criscitiello and Boumal, 2021). For this reason, in order to quantify convergence rates, one has to assume problem parameters such as smoothness or strong g-convexity hold in a specific region where the iterates lie.

Most works in the Riemannian optimization literature do not ensure their algorithms stay in some particular set $\mathcal{X}$ while they assume the smoothness or strong convexity inequalities hold between their iterates and between them and the optimizer. This often leads to an unfinished argument due to a circularity in which the step sizes depend on the problem parameters, that depend on the set $\mathcal{X}$ where the iterates lie, that depends on the step sizes. See also Hosseini and Sra (2020, Section 19.6) for a description of these issues and other references that suffer from it.

One way of tackling these two problems is showing that our algorithms naturally stay in a bounded region that we can quantify. Martínez-Rubio (2020) presents an algorithm with this property, that reduces an unconstrained g-convex problem to a series of problems in Riemannian balls of constant diameter. In the context of g-convex g-concave optimization, Martínez-Rubio et al. (2023) proved this property holds for an extragradient algorithm and then Wang et al. (2023); Hu et al. (2023) showed it for other related algorithms. The latter work applies to the more general variational inequalities setting. An alternative approach is to add in-manifold constraints to the problem and design methods that can enforce those constraints. Projection-free algorithms like those of Weber and Sra (2017, 2019), the Projected RGD algorithms surveyed in Section 3, and the accelerated constrained first-order methods in Martínez-Rubio (2020); Martínez-Rubio and Pokutta (2023); Martínez-Rubio et al. (2023) are designed to work with constraints and therefore they do not present the aforementioned problems.

Bounding the iterates of Riemannian algorithms has often been overlooked in the literature. Because of this reason, two of the most fundamental classes of first-order methods, namely Riemannian gradient descent (RGD) and Riemannian inexact proximal point algorithms (RIPPA), are not fully understood, and their convergence rates have not been fully quantified. Moreover, the Riemannian proximal point algorithm (RPPA), inexact versions of it, and implementations of such inexact methods in the smooth g-convex case, were also not well understood, especially in the non-Hadamard setting. In this work, we provide several convergence results in this direction and prove various fundamental properties of these methods. The following are the main results that we provide in this work:

- **RGD**: Among other RGD results, we show that for g-convex L-smooth Riemannian functions with a minimizer x^*, RGD with step size $\eta = 1/L$ stays in a closed ball $\bar{B}(x^*, O(\zeta_R R))$, where $R \stackrel{\text{def}}{=} d(x_0, x^*)$ and ζ_R is a geometric constant. If we use a step size that is $\eta = 1/(\zeta_{O(R)} L)$, the iterates stay in $\bar{B}(x^*, O(R))$. We quantify the rates of RGD in different settings as a result.

- **RPPA**: A general analysis of RPPA. It was only known in Hadamard manifolds before. An inexact RPPA and an implementation of it with first-order methods for smooth g-convex functions with quantified geometric penalties. A composite RGD that implies the inexact prox can be computed with $\widetilde{O}(1/\delta_R)$ gradient queries, where δ_R is a geometric constant.

- **Prox properties**: For any optimizer x^*, we have $d(\operatorname{prox}_{\eta f}(x), x^*) \leq d(x, x^*)$ and the Moreau envelope $M(x) \stackrel{\text{def}}{=} \min_{y \in \mathcal{X}}\{f(y) + \frac{1}{2\eta} d(x,y)^2\}$ is $(\zeta_{\operatorname{diam}(\mathcal{X})}/\eta)$-smooth in $\mathcal{X}$.

- **Experiments**: Numerical tests exploring beyond our theory. In our experiments, we observe that RGD presents a monotonic decrease in distance to an optimizer.

Outline We begin by introducing relevant definitions and notation in Section 2. Then we provide a detailed review of prior works on RGD and RPPA in Section 3. We present our new results regarding RGD and Riemannian proximal methods in Section 4. Then we present some empirical results in Section 5 and a conclusion in Section 6.

2 Preliminaries and notation

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 space is uniquely geodesic if every two points in that space are connected by one and only one geodesic.

The exponential map $\operatorname{Exp}_x : T_x\mathcal{M} \to \mathcal{M}$ 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 $\operatorname{Log}_x(\cdot)$, which is well defined for uniquely geodesic manifolds, so we have $\operatorname{Exp}_x(v) = y$ and $\operatorname{Log}_x(y) = v$. We denote the distance between two points by $d(x, y)$. 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 ∇. In that case, we use $\Gamma_x^y(v) \in T_y\mathcal{M}$ to denote the parallel transport of a vector v in $T_x\mathcal{M}$ to $T_y\mathcal{M}$ along the unique geodesic that connects x to y.

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. The Gauss curvature at a point x can be defined as the product of the maximum and minimum curvatures of the curves resulting from intersecting the surface with planes that are normal to the surface at x. We denote by $\mathcal{R}_{\mathrm{LB}}$ the set of uniquely geodesic Riemannian manifolds of sectional curvature lower bounded by $\kappa_{\min}$ and by $\mathcal{R}_{\mathrm{LUB}}$ the set of uniquely geodesic Riemannian manifolds of sectional curvature that is lower and upper bounded in $[\kappa_{\min}, \kappa_{\max}]$.

A set $\mathcal{X}$ is said to be g-convex if every two points are connected by a geodesic that remains in $\mathcal{X}$. We note that if a manifold $\mathcal{M} \in \mathcal{R}_{\mathrm{LUB}}$ has some positive sectional curvature, it may not be allowed to have arbitrarily large diameter. For example, since we work with uniquely geodesic manifolds, if $\kappa_{\min} > 0$, it is necessary that the diameter of the manifold is $< \pi/\sqrt{\kappa_{\min}}$. Thus, take into account that when we assume a ball of a certain radius is in $\mathcal{M}$, for some manifolds it is implicit that the radius is small enough. This is not a restriction for instance in Hadamard manifolds. A Hadamard manifold is a complete simply-connected Riemannian manifold of non-positive sectional curvature, which in particular is diffeomorphic to $\mathbb{R}^n$ and is uniquely geodesic.

Let $\mathcal{X}$ be a uniquely geodesic g-convex set. A differentiable function is μ-strongly g-convex (resp., L-smooth) in $\mathcal{X}$, if we have ① (resp. ②) for any two points $x, y \in \mathcal{X}$:

$$\frac{\mu d(x,y)^2}{2} \overset{①}{\leq} f(y) - f(x) - \langle \nabla f(x), \mathrm{Log}_x(y) \rangle \overset{②}{\leq} \frac{L d(x,y)^2}{2}$$

The function is said to be g-convex if $\mu = 0$. If we parametrize a geodesic joining x and y as the constant speed curve $\gamma : [0,1] \to \mathcal{M}$ such that $\gamma(0) = x$ and $\gamma(1) = y$, we have that convexity can be written as $f(\gamma(t)) \leq tf(x) + (1-t)f(y)$ and this also applies to non-differentiable functions. A function f is L_{p}-Lipschitz in $\mathcal{X}$ if $|f(x) - f(y)| \leq L_{\mathrm{p}} d(x,y)$ for all $x, y \in \mathcal{X}$.

Given a g-convex set $\mathcal{X} \subseteq \mathcal{M}$ for $\mathcal{M} \in \mathcal{R}_{\mathrm{LB}}$, we denote by $\mathcal{F}(\mathcal{X})$ the class of functions $f : \mathcal{M} \to \mathbb{R} \cup \{+\infty\}$ which are proper, lower semicontinuous and g-convex in $\mathcal{X}$. We denote by $\mathcal{F}_L(\mathcal{X}) \subset \mathcal{F}(\mathcal{X})$ the subclass of functions which are also differentiable in an open subset $\mathcal{N} \subset \mathcal{M}$ containing $\mathcal{X}$, and are L-smooth and g-convex in $\mathcal{X}$. We denote by $\mathcal{F}_{\mu,L}(\mathcal{X}) \subset \mathcal{F}_L(\mathcal{X})$ the subset of those functions that are μ-strongly g-convex in $\mathcal{X}$. Note the dependence on $\mathcal{X}$ is important, since the condition number of functions in $\mathcal{F}_{\mu,L}(\mathcal{X})$ depends on $\mathcal{X}$. We denote by $\mathcal{F}_{L_{\mathrm{p}}}(\mathcal{X}) \subset \mathcal{F}(\mathcal{X})$ the subclass of functions which are also L_{p}-Lipschitz and g-convex in $\mathcal{X}$.

Given $r > 0$, and a manifold $\mathcal{M} \in \mathcal{R}_{\mathrm{LUB}}$, we define the geometric constants $\zeta_r \stackrel{\mathrm{def}}{=} r\sqrt{|\kappa_{\min}|}\coth(r\sqrt{|\kappa_{\min}|}) = \Theta(1 + r\sqrt{|\kappa_{\min}|})$ if $\kappa_{\min} < 0$ and $\zeta_r \stackrel{\mathrm{def}}{=} 1$ otherwise, and $\delta_r \stackrel{\mathrm{def}}{=} r\sqrt{\kappa_{\max}}\cot(r\sqrt{\kappa_{\max}}) \leq 1$ if $\kappa_{\max} > 0$ and $\delta_r \stackrel{\mathrm{def}}{=} 1$ otherwise. It is $\delta_r \leq 1 \leq \zeta_r$. For a g-convex set $\mathcal{X} \subseteq \mathcal{M}$ of diameter bounded by D and containing $x \in \mathcal{M}$, the function $\Phi_x(y) \stackrel{\mathrm{def}}{=} \frac{1}{2}d(x,y)^2$ is δ_D-strongly g-convex and ζ_D-smooth in $\mathcal{X}$, cf. Lemma 21.

We define the indicator $I_{\mathcal{X}}(x)$ as 0 if $x \in \mathcal{X}$ and $+\infty$ if $x \notin \mathcal{X}$. A metric-projection operator $\mathcal{P}_{\mathcal{X}} : \mathcal{M} \to \mathcal{X}$ onto a closed g-convex set $\mathcal{X}$ is a map satisfying $d(\mathcal{P}_{\mathcal{X}}(x), y) \leq d(x,y)$ for all $y \in \mathcal{X}$. $\bar{B}(x,r)$ is a closed Riemannian ball of center x and radius r. The big-O notation $\widetilde{O}(\cdot)$ omits log factors.

In this paper $x_0 \in \mathcal{M}$ always represents an initial point of the algorithm we consider in that context. We assume the functions $f : \mathcal{M} \to \mathbb{R}$ we optimize contain at least one minimizer denoted by x^*, and we denote the initial distance to it by $R \stackrel{\mathrm{def}}{=} d(x_0, x^*)$. A point x is an ε-minimizer of f if $f(x) - f(x^*) \leq \varepsilon$. For an algorithm which runs for T iterations, we define $R_{\max} \stackrel{\mathrm{def}}{=} \max_{i \in T} d(x_i, x^*)$. We define the update rule of the algorithms we analyze in this paper. The RGD update rule is:

$$x_{t+1} \leftarrow \mathrm{Exp}_{x_t}(-\eta \nabla f(x_t)). \tag{1}$$

Uniform geodesic averaging of the iterates $\{x_1, \ldots, x_T\}$ is defined recursively as

$$\bar{x}_{t+1} \leftarrow \mathrm{Exp}_{\bar{x}_t}\left(\frac{1}{t+1}\mathrm{Log}_{\bar{x}_t}(x_{t+1})\right) \tag{2}$$

for $t \in \{1, \ldots, T-1\}$ where $\bar{x}_1 \leftarrow x_1$. The metric-projected RGD (PRGD) update rule is:

$$x_{t+1} \leftarrow \mathcal{P}_{\mathcal{X}}(\mathrm{Exp}_{x_t}(-\eta \nabla f(x_t))). \tag{3}$$

And given an $\eta > 0$, the RPPA update rule is:

$$x_{t+1} \leftarrow \mathrm{prox}_{\eta f}(x_t), \tag{4}$$

where $\mathrm{prox}_{\eta f}(x) \stackrel{\mathrm{def}}{=} \arg\min_{z \in \mathcal{M}} \{f(z) + \frac{1}{2\eta} d(z,x)^2\}$ if it exists, which is always the case in our setting.

3 Related work

3.1 Riemannian Gradient Descent

We limit this section to non-asymptotic analyses of RGD. Unless we specify otherwise, RGD refers to (1) and for a g-convex compact $\mathcal{X}$ of diameter D, it assumes $x_t \in \mathcal{X}$ for $t = 0, \ldots, T$, while properties like L-smoothness are assumed to hold in $\mathcal{X}$. Previous works either take this property as an assumption, rely on projections to enforce a bound or incur slow converge rates. In contrast, we ensure this property holds for $\mathcal{X}$ being a ball around a minimizer without using projections to enforce it.

For $\mathcal{F}_{\mu,L}(\mathcal{X})$, Gabay (1982, Thm. 4.4) showed an analysis of RGD with per-iteration descent factor of $1-\frac{\mu^2}{L^2}$, but only in the limit. Smith (1994) presents an analysis of RGD with rates $\widetilde{O}(\frac{L^2}{\mu^2})$ and Udriste (1994, Theorem 4.2) obtained the better rate $\widetilde{O}(\frac{L}{\mu})$. Zhang and Sra (2016) present several results on stochastic or deterministic RGD, under a variety of assumptions on the function. The rates depend on $\zeta_{R_{\max}}$ but $R_{\max} = \max_{t\in[T]} d(x_t, x^*)$ is not quantified. They analyze PRGD for non-smooth optimization, where they can use $D \geq R_{\max}$. They claim a PRGD analysis for smooth functions but the proof was found to be flawed (Martínez-Rubio and Pokutta, 2023). Bento et al. (2016b) obtained a curvature-independent rate of RGD for $\mathcal{F}_L(\mathcal{X})$ when the manifold is of non-negative sectional curvature. In this case, $\zeta_r = 1$ for every $r > 0$ so this result is an instance of the one in Zhang and Sra (2016). Ferreira et al. (2019) analyzed RGD for $\mathcal{F}_L(\mathcal{X})$ but with some exponential constants depending on the sectional curvature and the initial gap. Martínez-Rubio and Pokutta (2023) achieve linear rates of PRGD for $\mathcal{F}_{\mu,L}(\mathcal{X})$ assuming $\nabla f(x^*) = 0$ and $\zeta_D < 2$. For $\mathcal{F}_L(\mathcal{X})$ they obtain the curvature-independent rates $O(\frac{LR_{\max}^2}{\varepsilon})$ for RGD, but $R_{\max}$ is not quantified. They also provide an analysis of PRGD for $\mathcal{F}_{\mu,L}(\mathcal{X})$ with a projection oracle that is not a metric projection, obtaining $\widetilde{O}(\frac{L}{\mu})$ rates. Martínez-Rubio et al. (2023) present a general convergence analysis of PRGD for $\mathcal{F}_{\mu,L}(\mathcal{X})$ for Hadamard manifolds with rates depending on the Lipschitz constant of f in $\mathcal{X}$, namely $\widetilde{O}(\frac{L}{\mu}\zeta_C\zeta_D)$, for $D \stackrel{\text{def}}{=} \operatorname{diam}(\mathcal{X})$ and $C \stackrel{\text{def}}{=} (L_{\text{p}}/L + 2D)/\zeta_D$. If $\nabla f(x^*) = 0$, it is $\zeta_C = O(1)$.

In this work we show, for different fixed step sizes, that the iterates of RGD for several function classes stay naturally bounded in a set $\mathcal{X}$ being a ball around x^* whose radius we quantify. This allows us to bound $R_{\max}$ in a principled way instead resorting to assuming such a bound.

3.2 Riemannian Proximal Methods

To the best of our knowledge, the first work on the Riemannian proximal point algorithm is due to Ferreira and Oliveira (2002), with an asymptotic convergence in Hadamard manifolds with an exact proximal operator. They also established some properties of the algorithm in these manifolds.

There are numerous works on asymptotic convergence of exact or inexact RPPA for g-convex optimization or more in general for variational inequalities, but we focus on discussing works with convergence rates. Bačák (2013) obtained rates for RPPA in Hadamard manifolds (and more in general CAT(0) metric spaces), analogous to the classical Euclidean rates. Under a growth condition, Tang and Huang (2014) present linear rates for an inexact RPPA for a monotone operator F in Hadamard manifolds. Bento et al. (2016b) rediscover the results of Bačák (2013) regarding the convergence rates for RPPA in Hadamard. Bento et al. (2016a) obtains asymptotic convergence of RPPA under Kurdyka–Lojasiewicz inequality, without assuming the manifold is Hadamard. Espinola and Nicolae (2016); Kimura and Kohsaka (2017) also work in the general Riemannian case and obtain non-asymptotic convergence of an RPPA, but with a convolving function that is not the distance squared. In this work, we provide non-asymptotic rates for inexact RPPA in the general Riemannian case, which are the first of its kind when allowing positive sectional curvature, and we show how this framework can be implemented with first-order methods in the g-convex smooth case.

4 Convergence Results and Bounded Iterates

We summarize the convergence results presented in this section in Table 1. We include the proofs in the supplementary material. Note that for the same function, initialized at the same point, the smoothness or strong convexity constants in each row are not necessarily the same, since they refer to these parameters holding in balls of different diameters D.

Table 1: Summary of the convergence results in this work for g-convex functions in a ball $\mathcal{X}$ of diameter D centered at x^*. All iterates stay in $\mathcal{X}$. Note that μ, L and the Lipschitz constant L_{p} depend on the respective different sets $\mathcal{X}$. The value $\eta > 0$ is a proximal parameter.

Method	g-convex	μ-str. g-cvx	D
	L-SMOOTH		
RGD$_{L^{-1}}$	$O(\zeta_R^2\frac{LR^2}{\varepsilon})$	$\widetilde{O}(\frac{L}{\mu})$	$O(R\zeta_R)$
*Red. RGD$_{L^{-1}}$	$\widetilde{O}(\zeta_R^2+\frac{LR^2}{\varepsilon})$	–	$O(R\zeta_R)$
RGD$_{L^{-1}\zeta_{O(R)}^{-1}}$	$O(\zeta_R\frac{LR^2}{\varepsilon})$	$\widetilde{O}(\zeta_R\frac{L}{\mu})$	$O(R)$
RCEGmin	$O(\frac{\sqrt{\zeta_R}LR^2}{\sqrt{\delta_{4R}\varepsilon}})$	$\widetilde{O}(\frac{1}{\delta_{4R}}+\frac{\sqrt{\zeta_R}L}{\sqrt{\delta_{4R}\mu}})$	$O(R)$
RIPPA-CRGD	$O(\frac{LR^2}{\delta_{4R}\varepsilon})$	$\widetilde{O}(\frac{L}{\delta_{4R}\mu})$	$O(R)$
†RIPPA-PRGD	$O(\zeta_R\frac{LR^2}{\varepsilon})$	$\widetilde{O}(\zeta_R\frac{L}{\mu})$	$O(R)$
	NON-SMOOTH		
RGD NSm	$O(\sqrt{\zeta_R\frac{L_{\text{p}}R}{\varepsilon^2}})$	–	$O(R)$
RIPPA	$O(\frac{R^2}{\eta\varepsilon})$	$\widetilde{O}(1+\frac{1}{\mu\eta})$	$O(R)$

Consider as an example the n-dimensional hyperbolic space $\mathbb{H}^d \in \mathcal{R}_{\mathrm{LUB}}$, i.e., a complete simply connected, n-dimensional Riemannian manifold of constant sectional curvature equal to -1. For a point $x \in \mathbb{H}^d$, for any $r > 0$, and for the ball $\bar{B}(x, r)$, we have that $\Phi_x(y) \stackrel{\mathrm{def}}{=} \frac{1}{2}d(x, y)^2$ is exactly ζ_r-smooth and 1-strongly convex, cf. Lemma 21. In fact this is the function with minimum possible condition number L/μ in $\bar{B}(x, r)$, cf. Martínez-Rubio (2020). Using this fact and $\zeta_{O(R\zeta_R)} = O(\zeta_R^2)$, we have that for $\mathcal{M} = \mathbb{H}^d$, the expressions for the rates in Table 1 for strongly g-convex smooth functions of RGD with both $\eta = L^{-1}$ and $\eta = (L\zeta_{O(R)})^{-1}$ are both $\widetilde{O}(\zeta_R^2)$, despite of the seemingly better rate of the former.

4.1 Riemannian Gradient Descent

We start by showing that for g-convex L-smooth functions, the iterates of RGD with the standard $\eta = 1/L$ step size naturally stay in a Riemannian ball around the optimizer. In the proof, we perform a careful analysis of the different terms playing a role in the convergence in order to bound the distances. Recall that $R \stackrel{\mathrm{def}}{=} d(x_0, x^*)$, and let $\varphi \stackrel{\mathrm{def}}{=} (1+\sqrt{5})/2$.

Theorem 1. [↓] *Consider a manifold $\mathcal{M}_{\mathrm{LB}} \in \mathcal{R}_{\mathrm{LB}}$, and $f \in \mathcal{F}_L(\mathcal{X})$ for $\mathcal{X} \stackrel{\mathrm{def}}{=} \bar{B}(x^*, \varphi R\zeta_R) \subset \mathcal{M}_{\mathrm{LB}}$. The iterates of RGD with $\eta = 1/L$ satisfy $x_t \in \mathcal{X}$. In addition, if $\mathcal{M}_{\mathrm{LB}}$ is a hyperbolic space, and $\mathcal{X}_{\mathcal{H}} \stackrel{\mathrm{def}}{=} \bar{B}(x^*, \varphi R)$, $f \in \mathcal{F}_L(\mathcal{X})$, then $x_t \in \mathcal{X}_{\mathcal{H}}$.*

This result allows us to fully quantify the convergence rate of RGD by completing previous proofs of convergence.

Corollary 2. [↓] *Under the assumptions of Theorem 1, we obtain an ε-minimizer in $O(\zeta_R^2 \frac{LR^2}{\varepsilon})$ iterations. If f is also μ-strongly g-convex in $\mathcal{X}$, it takes $O(\frac{L}{\mu}\log(\frac{LR^2}{\varepsilon}))$.*

We also note that RGD with any step size $< 2/L$ never increases the function value, and so in fact, we only need to assume smoothness and g-convexity in the intersection of $\mathcal{X}$ and the level set of f with respect to x_0. We discuss the size of the level set and convergence results which assume these properties hold in the level set in Remark 12.

Interestingly, the rate for $f \in \mathcal{F}_L(\mathcal{X})$ that we obtain in Corollary 2 by using our iterate bounds coincides with both the rate obtained from the curvature-dependent rate $O(\zeta_{R_{\max}} \frac{LR^2}{\varepsilon})$ in Zhang and Sra (2016) and the seemingly curvature-independent rate $O(\frac{LR_{\max}^2}{\varepsilon})$ in Martínez-Rubio and Pokutta (2023). This example highlights the importance of providing iterate bounds to fully quantify convergence rates. Note that the result from Zhang and Sra (2016) holds for the geodesic average of the iterates, while the result from Martínez-Rubio and Pokutta (2023) holds for the last iterate. This difference is due to the fact that the proof of the former is based on mirror descent while the latter directly uses gradient descent techniques.

We note that among all the algorithms in Table 1 applying to smooth functions, RGD with $\eta = 1/L$ is the only one that does not require knowing the initial distance to a minimizer or a bound of it. If we know R, we can reduce the minimization of a function $f \in \mathcal{F}_L(\mathcal{X})$ to minimizing the strongly g-convex function $F(x) \stackrel{\mathrm{def}}{=} f(x) + \frac{\varepsilon}{2R^2}d(x_0, x)^2$. An upper bound on R can also be used instead. Indeed, applying RGD with $\eta = 1/L$ on $F(x)$, we obtain rates $\widetilde{O}(\zeta_R^2 + \frac{\hat{L}R^2}{\varepsilon})$ to find an ε-minimizer of f defined in a Hadamard manifold, where $\hat{L}$ is the smoothness constant of f in $\bar{B}(x_0, (1+\sqrt{5})R\zeta_R/2 + R)$. We can also quantify the rate in the general case, see Remark 13. Without some iterate boundedness like the one in Theorem 1 we do not know what rates this reduction would yield, or what step size we should use, even though we have curvature independent rates for strongly g-convex smooth problems. This occurs because the smoothness and condition number of the regularized function depend on the sets where the iterates lie and both increase with the diameter of this set.

Alternatively, we can use RGD with a step size $\eta = 1/(L\zeta_{O(R)})$. As for the reduction described above, an upper bound on R can be used instead of its value. Using RGD with this step size, we show that the iterates stay away from the minimizer an amount of the same order as the initial distance. Note that this step size is not in general smaller than the one in Theorem 1, since the smoothness constants in both step sizes are taken with respect to sets of different sizes and are not necessarily identical. In fact, for the problem we implement in Section 5, the step size given by Theorem 3 is slightly larger.

Theorem 3. [↓] *Consider a manifold $\mathcal{M}_{\mathrm{LB}} \in \mathcal{R}_{\mathrm{LB}}$. Let $f \in \mathcal{F}_L(\mathcal{X})$, for $\mathcal{X} \stackrel{\mathrm{def}}{=} \bar{B}(x^*, R\sqrt{3/2}) \subset \mathcal{M}_{\mathrm{LB}}$. The iterates of RGD with step size $\eta \stackrel{\mathrm{def}}{=} 1/(\zeta_{\sqrt{3/2}R}L)$ satisfy $x_t \in \mathcal{X}$. The convergence rate is $O(\zeta_R LR^2/\varepsilon)$. If f is also μ-strongly g-convex in $\mathcal{X}$, then it takes $O((\zeta_R L/\mu)\log(LR^2/\varepsilon))$ iterations.*

We can obtain similar results using Riemannian Corrected Extra-Gradient (RCEG) for minimization. Note however that RCEG requires computing two gradients per iteration.

Remark 4 (RCEG for minimization)**.** *Consider a manifold $\mathcal{M} \in \mathcal{R}_{\mathrm{LUB}}$ and $f \in \mathcal{F}_L(\mathcal{X})$ for $\mathcal{X} \stackrel{\mathrm{def}}{=} \bar{B}(x^*, 2R) \subset \mathcal{M}$. Using the g-convex g-concave algorithm RCEG in (Martínez-Rubio et al., 2023, Prop. 5, Alg. 3) with $F : \mathcal{M} \times \mathbb{R} \to \mathbb{R}$, $F(\cdot, y) = f(\cdot)$ for all y, we obtain an ε-minimizer in $O(\sqrt{\zeta_R/\delta_{4R}}LR^2/\varepsilon)$ iterations. If f is also μ-strongly g-convex in $\mathcal{X}$, then it takes $O((1/\delta_{4R} + L/\mu\sqrt{\zeta_R/\delta_{4R}})\log(LR^2/\varepsilon))$ iterations. The iterates stay in $\mathcal{X}$.*

---

[*] This is the rate for Hadamard manifolds only, for the general case see Remark 13.

[†] This result only applies to Hadamard manifolds.

It is worth noting that the rates of this algorithm involve the constant δ_{4R}, whereas this does not appear in the other analyses presented above. Studying whether a Riemannian extra-gradient algorithm converges without any dependence on $\delta_{O(R)}$ in the rates is an interesting direction of future research.

We present a composite RGD (CRGD) algorithm, of independent interest. Note that while the update rule of CRGD requires just one call to the gradient oracle, its implementation could consist of a hard computational problem. The interest of this result is that it yields better information theoretical upper bounds on the gradient oracle complexity than other approaches in some contexts. For instance, for the implementation of proximal subroutines for the optimization of functions in $\mathcal{F}_L(\mathcal{X})$, see Proposition 9.

Proposition 5 (Composite RGD)**.** [↓] *Let $\mathcal{M} \in \mathcal{R}_{\mathrm{LUB}}$ and let $\mathcal{X} \subset \mathcal{M}$ be closed and g-convex. Given $f \in \mathcal{F}_L(\mathcal{X})$, and $g \in \mathcal{F}(\mathcal{X})$, such that $F \stackrel{\mathrm{def}}{=} f+g$ is μ-strongly g-convex in $\mathcal{X}$, and $x^* \stackrel{\mathrm{def}}{=} \arg\min_{x\in\mathcal{X}} F(x)$. We get, iterating the rule*

$$x_{t+1} = \operatorname*{arg\,min}_{y\in\mathcal{X}} \langle \nabla f(x_t), \mathrm{Log}_{x_t}(y)\rangle + \frac{L}{2} d(x_t, y)^2 + g(y),$$

an ε-minimizer of F in $O(\frac{L}{\mu}\log(\frac{F(x_0)-F(x^*)}{\varepsilon}))$ iterations.

We note that in the proof of Proposition 5 we showed that the method above is well defined, in particular, that the argmin in the problem defining x_{t+1} above exists.

Finally, we show that the iterates of Riemannian subgradient descent, that uses the same update rule as RGD but making use of subgradients, move away from an optimizer by at most a $\sqrt{2}$ factor farther than the initial distance for Lipschitz functions.

Theorem 6 (Non-smooth RGD)**.** [↓] *Consider a manifold $\mathcal{M}_{\mathrm{LB}} \in \mathcal{R}_{\mathrm{LUB}}$ and $f : \mathcal{M}_{\mathrm{LB}} \to R$ that is L_{p}-Lipschitz in $\mathcal{X}$, for $\mathcal{X} \stackrel{\mathrm{def}}{=} \bar{B}(x^*, \sqrt{2}R) \subset \mathcal{M}_{\mathrm{LB}}$. The iterates of Riemannian subgradient descent with $\eta \stackrel{\mathrm{def}}{=} R/(L_{\mathrm{p}}\sqrt{\zeta_{\sqrt{2}R}T})$ lie in $\mathcal{X}$ and the geodesic average of the iterates, cf. (2), is an ε-minimizer of f after $O(\sqrt{\zeta_R}L_{\mathrm{p}}R/\varepsilon^2)$ iterations.*

4.2 Riemannian Proximal Methods

We start by showing that the iterates of the exact RPPA never move farther from an optimizer than the initial distance. This fact was proven in the more restricted Hadamard case by Ferreira and Oliveira (2002).

Proposition 7 (RPPA)**.** [↓] *Consider $\mathcal{M} \in \mathcal{R}_{\mathrm{LUB}}$ and $f \in \mathcal{F}(\mathcal{M})$ with $\mathcal{X} \stackrel{\mathrm{def}}{=} \bar{B}(x^*, R) \subset \mathcal{M}$. For any $\eta > 0$ and all $t \geq 0$, the iterates of the exact RPPA, cf. (4), satisfy $d(x_{t+1}, x^*) \leq d(x_t, x^*)$. In particular, it is $x_t \in \mathcal{X}$.*

Further, we show that if the iterates are computed inexactly as described in Algorithm 1 the iterates only move away from an optimizer by a small constant factor from the initial distance, and we quantify the convergence rates. We note that we can make the iterates stay in $\bar{B}(x^*, r)$ for $r > R$ as close as we want to R, by making the criterion in Line 2 more strict. The convergence rates of RPPA can be derived from the one of RIPPA when setting the error to 0, and in general they are the same up to constant factors.

Theorem 8 (RIPPA)**.** [↓] *Let $\mathcal{M} \in \mathcal{R}_{\mathrm{LUB}}$ and let $f \in \mathcal{F}(\mathcal{M})$ and assume $\bar{B}(x^*, 3R) \subset \mathcal{M}$. Using the notation in Algorithm 1, it holds that $x_t \in \bar{B}(x^*, \sqrt{2}R)$ for every $t \geq 0$, and the output of Algorithm 1 after $T = O(\frac{R^2}{\eta\varepsilon})$ iterations is an ε-minimizer of f. If f is μ-strongly convex in $\bar{B}(x^*, \sqrt{2}R)$, then $d(x_{t+1}, x^*)^2 \leq \frac{1}{1+\eta\mu/2} d(x_t, x^*)^2$ and in particular $d(x_T, x^*)^2 \leq \varepsilon_d$ after $T = O((1 + \frac{1}{\mu\eta})\log(\frac{R^2}{\varepsilon_d}))$ iterations.*

Algorithm 1 Riemannian Inexact Proximal Point Algorithm (RIPPA)

Input: Manifold $\mathcal{M} \in \mathcal{R}_{\mathrm{LUB}}$, initial point $x_0 \in \mathcal{M}$, μ-strongly g-convex function $f : \mathcal{M} \to \mathbb{R}$, for $\mu \geq 0$, and proximal parameter $\eta > 0$.

Definitions:

- Exact prox: $x^*_{t+1} \stackrel{\mathrm{def}}{=} \mathrm{prox}_{\eta f}(x_t)$.
- Subgradient: $v_{t+1} \in \partial f(x_{t+1})$.
- Error $r_{t+1} \stackrel{\mathrm{def}}{=} \eta v_{t+1} - \mathrm{Log}_{x_{t+1}}(x_t)$.
- For $\mu = 0$: $\Delta_t \stackrel{\mathrm{def}}{=} (t+1)^{-2}$. For $\mu > 0$: $\Delta_t \stackrel{\mathrm{def}}{=} \eta\mu/2$.

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

2: $$x_{t+1} \leftarrow \text{approx.} \operatorname*{arg\,min}_{z\in\mathcal{M}} \{f(z) + \frac{1}{2\eta} d(x_t, z)^2\}$$
$$\text{s.t.} \quad d(x_{t+1}, x^*_{t+1})^2 \leq \frac{1}{4} d(x_t, x^*_{t+1})^2,$$
$$\|r_{t+1}\|^2 \leq \Delta_{t+1}\delta_{5R} d(x_{t+1}, x_t)^2.$$

3: **end for**

Output: x_T **if** $\mu > 0$, **else** uniform geodesic averaging of $x_1, \ldots x_T$, cf. Corollary 26

While Theorem 8 does not require smoothness, this condition ensures that the proximal subproblems can be efficiently implemented via first-order methods, as we show in the following.

Proposition 9. [↓] *In the setting of Theorem 8, suppose that in addition $\eta = 1/L$, $\bar{B}(x^*, 4R) \subset \mathcal{M}$, and f is g-convex and L-smooth in $\bar{B}(x^*, 4R)$. The composite Riemannian Gradient Descent of Proposition 5 in $\mathcal{X} \stackrel{\mathrm{def}}{=} \bar{B}(x_t, 2R)$ implements the criterion in Line 2 of Algorithm 1 at iteration t using $\widetilde{O}(1/\delta_{4R})$ gradient oracle queries. If $\mathcal{M}$ is Hadamard, PRGD in $\mathcal{X}$ implements the* *criterion after $\widetilde{O}(\zeta_R)$ iterations.*

Recall that for a Hadamard manifold, it is $\delta_r = 1$, for all $r \geq 0$, so in this case CRGD uses $\widetilde{O}(1)$ gradient queries. Although the rates of CRGD are better than the ones of PRGD, implementing each step of the former could be a hard computational problem. On the other hand, each step of PRGD can be implemented easily, since projecting onto a Riemannian ball like $\mathcal{X}$ is as simple as joining the point with the center of the ball and computing the intersection with its border, which only requires computing and exponential map and an inverse exponential map, cf. Martínez-Rubio and Pokutta (2023).

We conclude this section studying the smoothness of the Moreau envelope, tightly related to the proximal point algorithm. In particular, this algorithm is equivalent to performing RGD on the Moreau envelope. This envelope is a useful tool in the design of optimization algorithms (Parikh and Boyd, 2014; Davis and Drusvyatskiy, 2019). First, we provide an expression for the gradient of the Moreau envelope.

Lemma 10 (Gradient of Moreau envelope)**.** [↓] *Let $\mathcal{M}$ be a uniquely geodesically Riemannian manifold, let $\mathcal{X} \subset \mathcal{M}$ be a g-convex closed set. For $f \in \mathcal{F}(\mathcal{X})$, we define the Moreau envelope of $g \stackrel{\text{def}}{=} f + I_{\mathcal{X}}$ with $\eta > 0$ as $M(x) \stackrel{\text{def}}{=} \min_{z\in\mathcal{M}}\{f(z)+I_{\mathcal{X}}(z)+\frac{1}{2\eta}d(x,z)^2\}$. We have $\nabla M(x) = -\frac{1}{\eta}\mathrm{Log}_x(\mathrm{prox}_{\eta g}(x))$.*

Now, we can show the smoothness of the Moreau envelope.

Theorem 11 (Moreau envelope smoothness)**.** [↓] *Consider a manifold $\mathcal{M}_{\mathrm{LB}} \in \mathcal{R}_{\mathrm{LB}}$ and let $\mathcal{X} \subset \mathcal{M}_{\mathrm{LB}}$ be a g-convex closed set. For a function $f \in \mathcal{F}(\mathcal{M})$, we have that the Moreau envelope of $g \stackrel{\text{def}}{=} f + I_{\mathcal{X}}$ with parameter $\eta > 0$, defined for all $x \in \mathcal{M}_{\mathrm{LB}}$ as $M(x) \stackrel{\text{def}}{=} \min_{z\in\mathcal{M}_{\mathrm{LB}}}\{f(z) + I_{\mathcal{X}}(z) + \frac{1}{2\eta}d(x,z)^2\}$, satisfies for all $x, y \in \mathcal{M}$:*

$$M(y) \leq M(x) + \langle \nabla M(x), \mathrm{Log}_x(y)\rangle + \frac{\zeta_{d(x,\mathrm{prox}_{\eta g}(x))}}{2\eta} d(x,y)^2.$$

In particular, if $\mathcal{X}$ is compact and its diameter is D, the Moreau envelope $M(x)$ is (ζ_D/η)-smooth in $\mathcal{X}$.

We note that if $\kappa_{\min} \geq 0$, the Moreau envelope is $(1/\eta)$-smooth. That is, the smoothness is not degraded by the curvature with respect to the Euclidean case. However, in this case, the Moreau envelope may not be g-convex.

The intuition about the proof of Theorem 11 is the following. The epigraph of the Moreau envelope can be seen as the union of the epigraphs $\{(x, f(y) + I_{\mathcal{X}}(y) + \frac{1}{2\eta}d(x,y)^2) \mid x \in \mathcal{M}_{\mathrm{LB}}\}$ for all $y \in \mathcal{M}_{\mathrm{LB}}$. Consequently, given the $\mathrm{prox}_{\eta f}(x)$, we have that the quadratic $U(y) \stackrel{\text{def}}{=} f(x) + \frac{1}{2\eta}d(x,y)^2$ satisfies $U(y) \geq M(y)$, for all $y \in \mathcal{X}$. And in fact, in light of the definition of $M(\cdot)$ and of Lemma 10 that shows $\nabla M(x) = -\frac{1}{\eta}\mathrm{Log}_x(\mathrm{prox}_{\eta g}(x))$, we have $U(x) = M(x)$ and $\nabla U(x) = \nabla M(x)$. The quadratic $U(\cdot)$ is itself smooth by the cosine inequality, cf. Remark 20, so it has a quadratic in $T_x\mathcal{M}_{\mathrm{LB}}$ whose induced function in $\mathcal{M}_{\mathrm{LB}}$ upper bounds $M(\cdot)$. In the supplementary material we present a proof based on this intuition and then we present another analysis, that although suboptimal, shows a different point of view on the problem.

5 Experiments

In this section, we verify our theoretical results empirically by computing the Karcher mean (Zhang et al., 2016; Zhang and Sra, 2016; Lou et al., 2020) in the d-dimensional hyperbolic space $\mathbb{H}^d$, which is used in training Hyperbolic Graph Convolutional Networks (Lou et al., 2020) and in hyperbolic embeddings (Sala et al., 2018), among others. We focus on validating our results bounding the distance of the iterates of RGD with $\eta = 1/L$ and $\eta = 1/(L\zeta_{O(R)})$. Theorems 1 and 3 guarantee that iterates naturally stay in a ball around the optimizers with a radius that is larger than the initial distance $R \stackrel{\text{def}}{=} d(x^*, x_0)$. Based on these results, one might expect to see an initial increase in distance to the optimizer. However, in our experiments we did not observe this increase in the distance of the iterates to the optimizer. In fact, we ran the problem for different settings, different initializations, and we performed a grid search on the step sizes. In all instances except those in which the step size was too large and the algorithm diverged, the distances where monotonically decreasing. This indicates that our bounds on the distance of the iterates to optimizers could potentially be improved. Alternatively, it might be that the distance between the optimizers and the iterates only increases for some pathological functions which do not arise in practice. Finding out whether a bound on the distance of the iterates to the optimizers which is larger than the initial distance R is unavoidable or if it can be improved is an interesting direction of future research.

We explain now the setting of our experiments. For $\mathcal{M} \in \mathcal{R}_{\mathrm{LUB}}$, the Karcher mean is defined as

$$\min_{x\in\mathcal{M}} \left\{ F(x) \stackrel{\text{def}}{=} \frac{1}{2n}\sum_{i=1}^{n} d^2(x, y_i) \right\}. \tag{5}$$

For a g-convex set $\mathcal{X} \subset \mathcal{M}$ containing all points y_i, the function F is ζ_D-smooth and δ_D-strongly g-convex in $\mathcal{X}$, where $D \stackrel{\text{def}}{=} \mathrm{diam}(\mathcal{X})$. This is a consequence of Lemma 21. By the g-strong convexity of F, the optimizer is unique. Further the optimizer x^* lies in $\mathcal{X}$, as we show in Proposition 28.

We generate the points y_i as follows: First, we define an anchor point $\bar{x} \in \mathcal{M}$, then we sample a d-dimensional vector $v_i \in \mathbb{R}^d$ uniformly in $[0, 1]$ in each dimension independently and compute $y_i \leftarrow \mathrm{Exp}_{\bar{x}}(rv_i/\sqrt{d})$, where $r \in \mathbb{R}^+$.

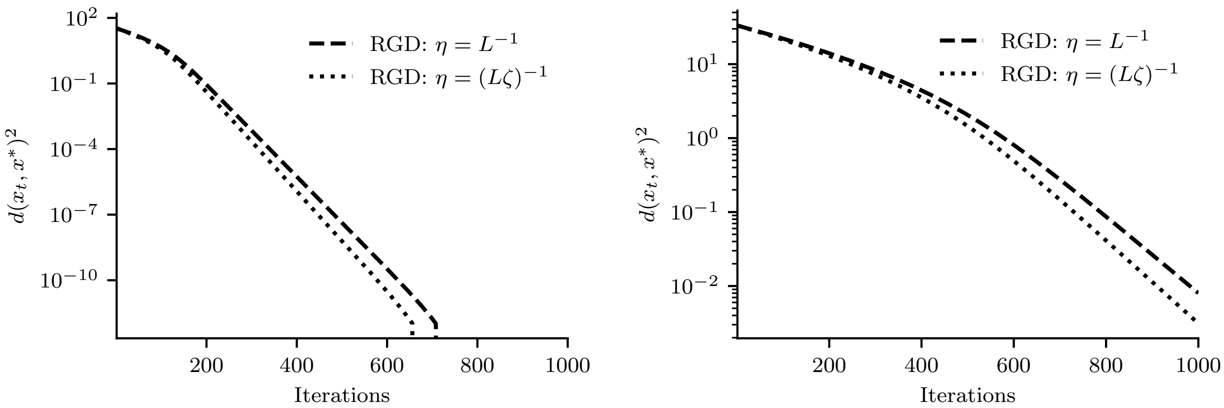

Figure 1: Comparison of RGD with $\eta = L^{-1}$ and $\eta = (L\zeta_{O(R)})^{-1}$ for solving (5) in the hyperbolic space in terms of squared distance to the optimizer x^* for two of our Karcher mean instances. Left: Our regular Karcher mean problem instance. Right: Our harder instance. We observe monotonous decrease in all of our experiments.

This ensures that $y_i \in \bar{\mathcal{X}} \stackrel{\text{def}}{=} \bar{B}(\bar{x}, r)$ for all y_i and hence $x^* \in \bar{\mathcal{X}}$. It follows that if we initialize our algorithm with $x_0 \in \bar{\mathcal{X}}$, we have $R = d(x_0, x^*) \le 2r$. This allows us to upper bound R a priori by $R \le 2\max_{i\in\{1,\dots,n\}} d(y_i, \bar{x})$.

We also generate a Karcher mean problem (5) in a different way, by dividing all but one of the randomly generated v_i by a factor of 10. This has an impact on the condition number and makes the problem harder.

Note that our bounds for both step sizes and the convergence results of RGD with $\eta = 1/(L\zeta_{O(R)})$ and $\eta = 1/L$ coincide up to constant factors for the Karcher mean on any $\mathcal{M} \in \mathcal{R}_{\text{LUB}}$. When using RGD with $\eta = 1/(L\zeta_{O(R)})$ to solve (5), we can ensure that the iterates stay in $\bar{B}(x^*, \sqrt{6}r)$. Hence F is $\zeta_{\sqrt{6}r}$-smooth, which means that the step size is $\eta = O(\frac{1}{\zeta_r^2})$ and the convergence rate simplifies to $\widetilde{O}\left(\zeta_R^2\right)$. If we use RGD $\eta = 1/L$, the iterates stay in $\bar{B}(x^*, (1+\sqrt{5})r\zeta_{2r})$. Hence, F is ζ_D smooth in that set with $D = 2(1+\sqrt{5})r\zeta_{2r}$ and since $\zeta_D = O(\zeta_R^2)$, we have $\eta = O(\frac{1}{\zeta_r^2})$ and the convergence rate simplifies to $\widetilde{O}\left(\zeta_R^2\right)$. This exemplifies the issue with stating smoothness and strong g-convexity constants of a function without specifying the size of the set in which they hold, since we obtained the same rate up to constants, but this was not obvious a priori. Note that since the sectional curvature of $\mathbb{H}^d$ is -1 everywhere, we have that $\mathbb{H}^d \in \mathcal{R}_{\text{LUB}}$. We do not consider manifolds of non-negative sectional curvature like the sphere, since we are interested in analyzing the influence of the sectional curvature on RGD and for manifolds with positive curvature, we have that $\zeta_r = 1$ for all $r \ge 0$, thus the rates of RGD are curvature-independent.

In Figure 1 we show the performance of RGD with both step sizes with $d = 10^3$, $n = 10^4$ and $r = 10$ for the hard and the normal Karcher mean, see Appendix F for results with different parameters. We implement this problem using the geoopt library (Kochurov et al., 2020), published under the Apache License, Version 2.0. We run the experiments on a computer with an 8-core i7-9700K @ 3.60GHz CPU with 64 GB RAM and a GeForce RTX 2080 Super GPU. The results for both step sizes are similar, however $\eta = 1/L$ performs slightly worse than $\eta = 1/(\zeta_R L)$. This is due to the fact that our computed step size is larger for the latter by a factor of about 1.1. While the convergence for the harder instances is slower, the distance to the optimizer still decreases monotonically.

6 Conclusion

In spite of recent advances in Riemannian optimization, the interplay between the curvature of the manifolds and the behaviour of optimization algorithms is still not fully understood. In this article, we advance the understanding on this connection for RGD algorithms by showing that the iterates are naturally bounded around a minimizer, where the bound depends on the step size. Further, we present the first analysis of the inexact Riemannian proximal point algorithm which holds in general manifolds. We provide non-asymptotic convergence guarantees and explicitly show that its iterates also stay in a set around an optimizer. One presented RGD algorithm guarantees that the iterates move away from an optimizer at most a small constant factor farther than the initial iterate.

An interesting future direction of research is whether this constant factor is unavoidable or if there is a variant of RGD which never moves away from an optimizer, as it is the case in the Euclidean space. Or more generally, whether the RGD rule is a non-expansive operator for some choice of the step size. This would have implications to algorithmic stability and differential privacy. Furthermore, for smooth functions, it is of interest to explore whether one can implement a Riemannian inexact proximal point algorithm by using a constant number of iterations in the subroutine, and therefore avoiding an extra logarithmic factor in the convergence results.

References

Ahn, K. and Sra, S. (2020). From Nesterov's estimate sequence to Riemannian acceleration. *arXiv preprint arXiv:2001.08876*.

Allen-Zhu, Z., Garg, A., Li, Y., de Oliveira, R. M., and Wigderson, A. (2018). Operator scaling via geodesically convex optimization, invariant theory and polynomial identity testing. In Diakonikolas, I., Kempe, D., and Henzinger, M., editors, *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2018, Los Angeles, CA, USA, June 25-29, 2018*, pages 172–181. ACM.

Bačák, M. (2013). The proximal point algorithm in metric spaces. *Israel journal of mathematics*, 194:689–701.

Bacák, M. (2014). *Convex analysis and optimization in Hadamard spaces*, volume 22. Walter de Gruyter GmbH & Co KG.

Bento, G., da Cruz Neto, J. X., and Oliveira, P. R. (2016a). A new approach to the proximal point method: Convergence on general riemannian manifolds. *J. Optim. Theory Appl.*, 168(3):743–755.

Bento, G., Ferreira, O., and Oliveira, P. (2015). Proximal point method for a special class of nonconvex functions on Hadamard manifolds. *Optimization*, 64(2):289–319.

Bento, G. C., Ferreira, O. P., and Melo, J. G. (2016b). Iteration-complexity of gradient, subgradient and proximal point methods on Riemannian manifolds. *arXiv:1609.04869 [math]*. arXiv: 1609.04869.

Bertsekas, D., Nedic, A., and Ozdaglar, A. (2003). *Convex analysis and optimization*, volume 1. Athena Scientific.

Boumal, N. (2023). *An introduction to optimization on smooth manifolds*. Cambridge University Press.

Cai, Y., Jordan, M. I., Lin, T., Oikonomou, A., and Vlatakis-Gkaragkounis, E.-V. (2023). Curvature-independent last-iterate convergence for games on riemannian manifolds. *arXiv preprint arXiv:2306.16617*.

Cambier, L. and Absil, P. (2016). Robust low-rank matrix completion by Riemannian optimization. *SIAM J. Scientific Computing*, 38(5).

Carvalho Bento, G. d. and Melo, J. G. (2012). Subgradient method for convex feasibility on Riemannian manifolds. *J. Optim. Theory Appl.*, 152(3):773–785.

Cherian, A. and Sra, S. (2017). Riemannian dictionary learning and sparse coding for positive definite matrices. *IEEE Trans. Neural Networks Learn. Syst.*, 28(12):2859–2871.

Criscitiello, C. and Boumal, N. (2021). Negative curvature obstructs acceleration for geodesically convex optimization, even with exact first-order oracles. *CoRR*, abs/2111.13263.

Criscitiello, C. and Boumal, N. (2023). Curvature and complexity: Better lower bounds for geodesically convex optimization. In Neu, G. and Rosasco, L., editors, *The Thirty Sixth Annual Conference on Learning Theory, COLT 2023, 12-15 July 2023, Bangalore, India*, volume 195 of *Proceedings of Machine Learning Research*, pages 2969–3013. PMLR.

Cruz Neto, J. X. d., Ferreira, O. P., Pérez, L. R. L., and Németh, S. Z. (2006). Convex- and monotone-transformable mathematical programming problems and a proximal-like point method. *J. Glob. Optim.*, 35(1):53–69.

Davis, D. and Drusvyatskiy, D. (2019). Stochastic model-based minimization of weakly convex functions. *SIAM J. Optim.*, 29(1):207–239.

Edelman, A., Arias, T. A., and Smith, S. T. (1998). The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Analysis Applications*, 20(2):303–353.

Espinola, R. and Nicolae, A. (2016). Proximal minimization in cat (k) spaces. *arXiv preprint arXiv:1607.03660*.

Ferreira, O. and Oliveira, P. (2002). Proximal point algorithm on riemannian manifolds. *Optimization*, 51(2):257–270.

Ferreira, O. P., Louzeiro, M. S., and da Fonseca Prudente, L. (2019). Gradient method for optimization on riemannian manifolds with lower bounded curvature. *SIAM J. Optim.*, 29(4):2517–2541.

Gabay, D. (1982). Minimizing a differentiable function over a differential manifold. *Journal of Optimization Theory and Applications*, 37:177–219.

Heidel, G. and Schulz, V. (2018). A Riemannian trust-region method for low-rank tensor completion. *Numerical Lin. Alg. with Applic.*, 25(6).

Hosseini, R. and Sra, S. (2015). Matrix manifold optimization for Gaussian mixtures. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 910–918.

Hosseini, R. and Sra, S. (2017). An alternative to EM for gaussian mixture models: Batch and stochastic Riemannian optimization. *CoRR*, abs/1706.03267.

Hosseini, R. and Sra, S. (2020). Recent advances in stochastic Riemannian optimization. *Handbook of Variational Methods for Nonlinear Geometric Data*, pages 527–554.

Hu, Z., Wang, G., Wang, X., Wibisono, A., Abernethy, J., and Tao, M. (2023). Extragradient type methods for riemannian variational inequality problems. *arXiv preprint arXiv:2309.14155v1*.

Jordan, M. I., Lin, T., and Vlatakis-Gkaragkounis, E. (2022). First-order algorithms for min-max optimization in geodesic metric spaces. *CoRR*, abs/2206.02041.

Jost, J. (1995). Convex functionals and generalized harmonic maps into spaces of non positive curvature. *Commentarii mathematici helvetici*, 70:659–673.

Khuzani, M. B. and Li, N. (2017). Stochastic primal-dual method on Riemannian manifolds of bounded sectional curvature. In *16th IEEE International Conference on Machine Learning and Applications, ICMLA 2017, Cancun, Mexico, December 18-21, 2017*, pages 133–140.

Kim, J. and Yang, I. (2022). Accelerated gradient methods for geodesically convex optimization: Tractable algorithms and convergence analysis. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvári, C., Niu, G., and Sabato, S., editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 11255–11282. PMLR.

Kimura, Y. and Kohsaka, F. (2017). The proximal point algorithm in geodesic spaces with curvature bounded above. *arXiv preprint arXiv:1704.05721*.

Kochurov, M., Karimov, R., and Kozlukov, S. (2020). Geoopt: Riemannian optimization in pytorch.

Lezcano-Casado, M. (2020). Curvature-dependant global convergence rates for optimization on manifolds of bounded geometry. *arXiv preprint arXiv:2008.02517*.

Lezcano-Casado, M. and Martínez-Rubio, D. (2019). Cheap orthogonal constraints in neural networks: A simple parametrization of the orthogonal and unitary group. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pages 3794–3803.

Lou, A., Katsman, I., Jiang, Q., Belongie, S., Lim, S.-N., and Sa, C. D. (2020). Differentiating through the Fréchet Mean. In *Proceedings of the 37th International Conference on Machine Learning*, pages 6393–6403. PMLR. ISSN: 2640-3498.

Martínez-Rubio, D. (2020). Global Riemannian acceleration in hyperbolic and spherical spaces. *arXiv preprint arXiv:2012.03618*.

Martínez-Rubio, D. and Pokutta, S. (2023). Accelerated riemannian optimization: Handling constraints with a prox to bound geometric penalties. In Neu, G. and Rosasco, L., editors, *The Thirty Sixth Annual Conference on Learning Theory, COLT 2023, 12-15 July 2023, Bangalore, India*, volume 195 of *Proceedings of Machine Learning Research*, pages 359–393. PMLR.

Martínez-Rubio, D., Roux, C., Criscitiello, C., and Pokutta, S. (2023). Accelerated methods for riemannian min-max optimization ensuring bounded geometric penalties. *CoRR*, abs/2305.16186.

Mayer, U. F. (1998). Gradient flows on nonpositively curved metric spaces and harmonic maps. *Communications in Analysis and Geometry*, 6(2):199–253.

Mishra, B. and Sepulchre, R. (2014). R3MC: A Riemannian three-factor algorithm for low-rank matrix completion. In *53rd IEEE Conference on Decision and Control, CDC 2014, Los Angeles, CA, USA, December 15-17, 2014*, pages 1137–1142.

Parikh, N. and Boyd, S. P. (2014). Proximal algorithms. *Found. Trends Optim.*, 1(3):127–239.

Petersen, P. (2006). *Riemannian geometry*, volume 171. Springer.

Sala, F., Sa, C. D., Gu, A., and Ré, C. (2018). Representation tradeoffs for hyperbolic embeddings. In Dy, J. G. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 4457–4466. PMLR.

Sato, H., Kasai, H., and Mishra, B. (2017). Riemannian stochastic variance reduced gradient. *CoRR*, abs/1702.05594.

Sato, H., Kasai, H., and Mishra, B. (2019). Riemannian stochastic variance reduced gradient algorithm with retraction and vector transport. *SIAM Journal on Optimization*, 29(2):1444–1472.

Smith, S. T. (1994). Optimization techniques on riemannian manifolds. *Fields Institute Communications*, 3.

Sun, J., Qu, Q., and Wright, J. (2017). Complete dictionary recovery over the sphere II: recovery by Riemannian trust-region method. *IEEE Trans. Inf. Theory*, 63(2):885–914.

Tan, M., Tsang, I. W., Wang, L., Vandereycken, B., and Pan, S. J. (2014). Riemannian pursuit for big matrix recovery. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pages 1539–1547.

Tang, G. and Huang, N. (2014). Rate of convergence for proximal point algorithms on hadamard manifolds. *Oper. Res. Lett.*, 42(6-7):383–387.

Tripuraneni, N., Flammarion, N., Bach, F., and Jordan, M. I. (2018). Averaging stochastic gradient descent on Riemannian manifolds. *CoRR*, abs/1802.09128.

Udriste, C. (1994). *Convex functions and optimization methods on Riemannian manifolds*, volume 297. Springer Science & Business Media.

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

Wang, X., Yuan, D., Hong, Y., Hu, Z., Wang, L., and Shi, G. (2023). Riemannian optimistic algorithms. *arXiv preprint arXiv:2308.16004v1*.

Weber, M. and Sra, S. (2017). Frank-Wolfe methods for geodesically convex optimization with application to the matrix geometric mean. *CoRR*, abs/1710.10770.

Weber, M. and Sra, S. (2019). Nonconvex stochastic optimization on manifolds via Riemannian Frank-Wolfe methods. *CoRR*, abs/1910.04194.

Zhang, H., Reddi, S. J., and Sra, S. (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*, pages 4592–4600.

Zhang, H. and Sra, S. (2016). First-order Methods for Geodesically Convex Optimization. *arXiv:1602.06053 [cs, math, stat]*. arXiv: 1602.06053.

Zhang, H. and Sra, S. (2018). An estimate sequence for geodesically convex optimization. In Bubeck, S., Perchet, V., and Rigollet, P., editors, *Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018*, volume 75 of *Proceedings of Machine Learning Research*, pages 1703–1723. PMLR.

Zhang, P., Zhang, J., and Sra, S. (2022). Minimax in geodesic metric spaces: Sion's theorem and algorithms. *CoRR*, abs/2202.06950.

A RGD proofs

Theorem 1. [↓] *Consider a manifold $\mathcal{M}_{\mathrm{LB}} \in \mathcal{R}_{\mathrm{LB}}$, and $f \in \mathcal{F}_L(\mathcal{X})$ for $\mathcal{X} \stackrel{\mathrm{def}}{=} \bar{B}(x^*, \varphi R\zeta_R) \subset \mathcal{M}_{\mathrm{LB}}$. The iterates of RGD with $\eta = 1/L$ satisfy $x_t \in \mathcal{X}$. In addition, if $\mathcal{M}_{\mathrm{LB}}$ is a hyperbolic space, and $\mathcal{X}_{\mathcal{H}} \stackrel{\mathrm{def}}{=} \bar{B}(x^*, \varphi R)$, $f \in \mathcal{F}_L(\mathcal{X})$, then $x_t \in \mathcal{X}_{\mathcal{H}}$.*

Proof. (Theorem 1) Define $\varphi \stackrel{\mathrm{def}}{=} (1+\sqrt{5})/2$ and $\zeta \stackrel{\mathrm{def}}{=} \zeta_{\varphi R\zeta_R}$. We first check that $\zeta = O(\zeta_R^2)$. Since it is $\zeta_r \in [r\sqrt{|\kappa_{\min}|}, r\sqrt{|\kappa_{\min}|}+1]$ and $\zeta_r \geq 1$, for all $r \geq 0$, we have $\zeta_{\varphi R\zeta_R} \leq \varphi R\sqrt{|\kappa_{\min}|}\zeta_R + 1 \leq \varphi\zeta_R^2 + 1 = O(\zeta_R^2)$. Now denote $\Delta_i \stackrel{\mathrm{def}}{=} f(x_i) - f(x^*)$. We show by induction that $d(x_t, x^*) \leq \varphi R\zeta_R$, for all $t \geq 0$. This holds for $t = 0$ by definition so suppose the property holds for all $i \leq t$ and let's prove it for $t+1$. Let $\eta_* \stackrel{\mathrm{def}}{=} \max_{\eta > 0}\{\eta \mid \mathrm{Exp}_{x_t}(-\eta_*\nabla f(x_t)) \in \bar{B}(x^*, \varphi R\zeta_R)\}$. Note we can write a maximum because the ball is compact. It is enough to show $\eta_* \geq \frac{1}{L}$. Suppose $\eta_* < \frac{1}{L}$ and let $x'_{t+1} \stackrel{\mathrm{def}}{=} \mathrm{Log}_{x_t}(-\eta_*\nabla f(x_t))$. By definition, it must be $d(x'_{t+1}, x^*) = \varphi R\zeta_R$. We will arrive to a contradiction. We have for all $i < t$:

$$\Delta_{i+1} - \Delta_i \leq \langle \nabla f(x_i), \mathrm{Log}_{x_i}(x_{i+1})\rangle + \frac{L}{2} d\left(x_{i+1}, x_i\right)^2 = -\frac{1}{2L}\|\nabla f(x_i)\|^2, \tag{6}$$

where we used L-smoothness of f in $\bar{B}(x^*, \varphi R\zeta_R)$ the definition of x_{i+1} for $i < t$, and the induction hypothesis that allows us to use the L-smoothness property. Similarly, by the definition of x'_{t+1} and defining $\Delta'_{t+1} \stackrel{\mathrm{def}}{=} f(x'_{t+1}) - f(x^*)$, we have:

$$\Delta'_{t+1} - \Delta_t \leq \langle \nabla f(x_t), \mathrm{Log}_{x_t}(x'_{t+1})\rangle + \frac{L}{2} d\left(x'_{t+1}, x_t\right)^2 = \left(-\eta_* + \frac{L\eta_*^2}{2}\right)\|\nabla f(x_t)\|^2 \stackrel{①}{<} -\frac{\eta_*}{2}\|\nabla f(x_t)\|^2, \tag{7}$$

where ① is equivalent to $\eta_* \in (0, 1/L)$. We also have the following bound, for all $i < t$:

$$\begin{aligned}\Delta_i &\stackrel{①}{\leq} \langle -\nabla f(x_i), \mathrm{Log}_{x_i}(x^*)\rangle \stackrel{②}{\leq} \frac{L}{2}\left[d(x_i, x^*)^2 - d(x_{i+1}, x^*)^2 + \zeta d(x_i, x_{i+1})^2\right] \\ &= \frac{L}{2}\left[d(x_i, x^*)^2 - d(x_{i+1}, x^*)^2\right] + \frac{\zeta}{2L}\|\nabla f(x_i)\|^2,\end{aligned} \tag{8}$$

where ① uses g-convexity of f, ② uses the cosine inequality Remark 20 and the bound $\zeta_{d(x_i, x^*)} \leq \zeta$ which holds by induction hypothesis and monotonicity of $r \mapsto \zeta_r$. Likewise, we have

$$\Delta_t \leq \frac{1}{2\eta_*}\left[d(x_t, x^*)^2 - d(x'_{t+1}, x^*)^2\right] + \frac{\zeta\eta_*}{2}\|\nabla f(x_t)\|^2 \tag{9}$$

Multiplying (6) by ζ and adding it to (8), and similarly with (7) and (9) we obtain:

$$\begin{aligned}\zeta\Delta_{i+1} - (\zeta - 1)\Delta_i &\leq \frac{L}{2}\left(d(x_i, x^*)^2 - d(x_{i+1}, x^*)^2\right) \text{ for } i < t \\ \eta_* L(\zeta\Delta'_{t+1} - (\zeta-1)\Delta_t) &< \frac{L}{2}\left(d(x_t, x^*)^2 - d(x'_{t+1}, x^*)^2\right)\end{aligned} \tag{10}$$

Adding up from $i = 0$ to t, we obtain

$$\begin{aligned}\eta_* L\zeta\Delta'_{t+1} + \zeta\Delta_t(1 - \eta_* L) + \eta_* L\Delta_t + \sum_{i=1}^{t-1}\Delta_i + \frac{L}{2}d(x'_{t+1}, x^*)^2 &< (\zeta - 1)\Delta_0 + \frac{Ld(x_0, x^*)^2}{2} \\ &\stackrel{①}{\leq} \frac{\zeta L R^2}{2}.\end{aligned} \tag{11}$$

where ① uses that by smoothness $\Delta_0 \leq \frac{Ld(x_0, x^*)^2}{2} \leq \frac{LR^2}{2}$. Using $\eta_* L \in (0,1)$ dropping all the terms with $\Delta_i, \Delta'_{t+1} \geq 0$, and simplifying, we obtain ② below, while the rest of the following holds by the definition of ζ, and the fact that for all $r \geq 0$, we have $\zeta_r \geq 1$ and $\zeta_r \in [r\sqrt{|\kappa_{\min}|}, r\sqrt{|\kappa_{\min}|}+1]$:

$$d(x'_{t+1}, x^*)^2 \stackrel{②}{<} \zeta R^2 \leq (\varphi\zeta_R R\sqrt{|\kappa_{\min}|} + 1)R^2 \leq (\varphi\zeta_R^2 + 1)R^2 \leq (\varphi + 1)\zeta_R^2 R^2 = \varphi^2\zeta_R^2 R^2.$$

But this contradicts the definition of x'_{t+1} for which $d(x'_{t+1}, x^*)^2 = \varphi^2\zeta_R^2 R^2$. Thus $\eta_* \geq 1/L$, and the inductive statement holds.

For the statement about the hyperbolic space, it is enough to show it for $\kappa_{\min} = \kappa_{\max} = -1$, since the other cases can be reduced to this one by rescaling, see (Martínez-Rubio, 2020, Remark 24). We prove $d(x_t, x^*) \leq \varphi R$ by induction in a similar way. It holds for $t = 0$ by definition and notice that the proof above also works for this case until (11) if we now consider $\eta_* \stackrel{\text{def}}{=} \max_{\eta>0}\{\eta \mid \operatorname{Exp}_{x_t}(-\eta_*\nabla f(x_t)) \in \bar{B}(x^*, \varphi R)\}$, which yields $d(x'_{t+1}, x^*) = \phi R$, we make use of $\zeta \stackrel{\text{def}}{=} \varphi\zeta_R$, and use L-smoothness in $\bar{B}(x^*, \varphi R)$. However, we substitute the right hand side of 1 in (11) by $\varphi^2 LR^2/2$ which holds since by (Criscitiello and Boumal, 2023, Proposition 13) we have $\Delta_0 \leq \frac{\varphi LR^2}{2\zeta_R}$ for the hyperbolic space, and using $\zeta \leq \varphi R + 1$ and $\zeta_R \geq R$, we bound $(\zeta - 1)/\zeta_R \leq \varphi$, and use $\varphi + 1 = \varphi^2$. We note that (Criscitiello and Boumal, 2023, Proposition 13) states global g-convexity as an assumption, but the proof only uses $f \in \mathcal{F}_L(\mathcal{X})$. Now we conclude as before. Using $\eta_* L \in (0, 1)$ dropping all the terms with $\Delta_i, \Delta'_{t+1} \geq 0$, and simplifying, we obtain $d(x'_{t+1}, x^*)^2 < \varphi^2 R^2$. This contradicts the definition of x'_{t+1} for which $d(x'_{t+1}, x^*)^2 = \varphi^2 R^2$. Consequently, $\eta_* \geq 1/L$, and we showed the inductive statement.

□

We note that if we were to assume smoothness in the closed ball $\bar{B}(x^*, 2\varphi R\zeta_R)$, we could just write (6) for all $i = 0$ to t, by arguing that x_{t+1} is in such ball since $d(x_{t+1}, x^*) \leq d(x_t, x^*) + d(x_{t+1}, x_t) \leq \varphi R\zeta_R + \varphi R\zeta_R$, where the bound on the first term is due to the induction hypothesis and the one of the second term is due to the definition of x_{t+1} and that L-smoothness implies $\frac{1}{L}\|\nabla f(x_t)\| \leq d(x_t, x^*)$. In this case, the proof proceeds in a similar but simpler way, without having to argue by contradiction or having to talk about the last iterate using a different learning rate. But the proof we presented requires smoothness only in a smaller region.

Corollary 2. [↓] *Under the assumptions of Theorem 1, we obtain an ε-minimizer in $O(\zeta_R^2\frac{LR^2}{\varepsilon})$ iterations. If f is also μ-strongly g-convex in $\mathcal{X}$, it takes $O(\frac{L}{\mu}\log(\frac{LR^2}{\varepsilon}))$.*

Proof. (Corollary 2) By Theorem 1 our iterates stay in $\mathcal{X}$, that is, $R_{\max} \leq \varphi R\zeta_R$. Note $\zeta_{\varphi R\zeta_R} = O(\zeta_R^2)$. For the g-convex case, the corollary is an immediate consequence of this iterate bound and of the convergence result in (Zhang and Sra, 2016, Theorems 13). This theorem proves rates $O(\zeta_{R_{\max}}\frac{LR^2}{\varepsilon})$, which by using the bound on $R_{\max}$ yields $O(\zeta_R^2\frac{LR^2}{\varepsilon})$. Similarly, if we apply the RGD result in Martínez-Rubio and Pokutta (2023) that has rates $O(\frac{LR_{\max}^2}{\varepsilon})$ for a function in $\mathcal{F}_L(\mathcal{X})$, we obtain the convergence rate $O(\zeta_R^2\frac{LR^2}{\varepsilon})$. Note that the two approaches give the same rates.

For the μ-strongly g-convex case, the corollary is an immediate consequence of our iterate bound and both the result by Udriste (1994) and the one by (Martínez-Rubio and Pokutta, 2023, Proposition 17) with $\mathcal{X}$ being the ball Theorem 1, so that the algorithm becomes RGD. Both have rates of $O(\frac{L}{\mu}\log(\frac{f(x_0)-f(x^*)}{\varepsilon}))$. The result is derived by the bound $f(x_0) - f(x^*) \leq LR^2/2$ due to smoothness. Note that due to our iterate bounds, we just need to assume L-smoothness and the μ-strong g-convex in $\mathcal{X} = \bar{B}(x^*, R\zeta_R(1+\sqrt{5})/2)$ and without these bounds, this rate is not obtained for the previous results for RGD since otherwise we cannot even specify where the L-smoothness and the μ-strong g-convex properties hold and we cannot necessarily take the value of some global properties since for many manifolds there is no globally smooth strongly g-convex function with finite condition number, namely all Hadamard manifolds for which $\kappa_{\max} < 0$ Criscitiello and Boumal (2021).

We note that for the strongly g-convex case, we could also use the result in (Zhang and Sra, 2016, Theorems 15) yielding the rate $O((\zeta_{R_{\max}} + \frac{L}{\mu})\log(\frac{LR^2}{\varepsilon})) = O((\zeta_R^2 + \frac{L}{\mu})\log(\frac{LR^2}{\varepsilon}))$.

□

Remark 12. *We note that RGD with step size $\eta < 2/L$ does not increase the function value. Indeed, assume smoothness holds between x_t and $x_{t+1} \stackrel{\text{def}}{=} \operatorname{Exp}_{x_t}(-\eta\nabla f(x_t))$. We have:*

$$f(x_{t+1}) \leq f(x_t) + \langle \nabla f(x_t), \operatorname{Log}_{x_t}(x_{t+1})\rangle + \frac{L}{2}d(x_t, x_{t+1})^2 = f(x_t) + \left(-\eta + L\eta^2\right)\|\nabla f(x_t)\|^2 \stackrel{1}{<} f(x_t),$$

where 1 is equivalent to $\eta < 2/L$. This means that if $\mathcal{Y} \stackrel{\text{def}}{=} \{y \in \mathcal{M} \mid f(y) \leq f(x_0)\}$ is the level set of f with respect to x_0, the iterates of RGD in the setting of Theorem 1 stay in $\mathcal{X} \cap \mathcal{Y}$ and we only need to assume g-convexity and L-smoothness *in in that set. Note that if f satisfies these properties in $\mathcal{Y}$, one can also obtain a convergence result by using* $\operatorname{diam}(\mathcal{Y})$ *as a bound for* $R_{\max}$*. And if f is μ-strongly g-convex in $\mathcal{Y}$ we obtain the rate $\widetilde{O}(L/\mu)$. But this result is strictly worse than that one of Theorem 1. Besides, we can compute a bound that suggests that in general, the level set could have points that are $R\sqrt{L/\mu}$ away from the minimizer. Indeed, let $y \in \mathcal{Y}$, we obtain*

$$\frac{\mu}{2}d(y,x)^2 \overset{①}{\leq} f(y) - f(x^*) \overset{②}{\leq} f(x_0) - f(x^*) \overset{③}{\leq} \frac{L}{2}d(x_0,x^*)^2,$$

where above ① uses μ-strong g-convexity, ② uses $y \in \mathcal{Y}$ and ③ uses L-smoothness. This bound can be much larger than the $O(R\zeta_R)$ bound in Theorem 1.

Remark 13. *We note that if we know a valid upper bound R, we can instead minimize $F(x) \stackrel{\text{def}}{=} f(x) + r(x)$ where $r(x) \stackrel{\text{def}}{=} \frac{\varepsilon}{2R^2}d(x_0,x)^2$. In that case, a point $\hat{x}$ that is an $(\varepsilon/2)$-minimizer of r will be an ε minimizer of f, since*

$$f(\hat{x}) \leq f(\hat{x}) + r(\hat{x}) \leq f(x^*) + r(x^*) + \frac{\varepsilon}{2} \leq f(x^*) + \varepsilon.$$

Denote $\hat{x}^* \stackrel{\text{def}}{=} \arg\min_x r(x)$. We have by Lemma 27 that $d(\hat{x}^*, x_0) \leq d(x^*, x_0) \leq R$. The smoothness constant of $F(\cdot)$ in $\mathcal{X} \stackrel{\text{def}}{=} \bar{B}(\hat{x}^*, \varphi R\zeta_R) \subseteq \bar{B}(x_0, \varphi R\zeta_R + R)$ is $\hat{L} + \zeta_{\varphi R\zeta_R + R}\frac{\varepsilon}{R^2} = O(\hat{L} + \zeta_R^2\frac{\varepsilon}{R^2})$, where $\hat{L}$ is the smoothness constant of f in $\mathcal{X} \subseteq \bar{B}(x_0, \varphi R\zeta_R + R)$. The strong convexity constant of $F(\cdot)$ in $\mathcal{X}$ is at least $\frac{\varepsilon}{R^2}\delta_{\varphi R\zeta_R + R}$. The computation of these smoothness and strong convexity constants comes from Lemma 21. We know by Theorem 1 that the iterates of RGD on $F(\cdot)$ with step size $\eta \stackrel{\text{def}}{=} (\hat{L} + \zeta_{\varphi R\zeta_R + R})^{-1}$ stay in $\mathcal{X}$ and thus by Corollary 2 we obtain the convergence rate $\widetilde{O}(\frac{\zeta_R^2}{\delta_{\varphi R\zeta_R + R}} + \frac{\hat{L}R^2}{\varepsilon\delta_{\varphi R\zeta_R + R}})$. If we consider for instance a Hadamard manifold, that satisfies $\delta_r = 1$ for any r, we obtain the convergence rate $\widetilde{O}(\zeta_R^2 + \frac{\hat{L}R^2}{\varepsilon})$. We note that we can obtain essentially the same result but without logarithmic factors if we use the reduction in Martínez-Rubio (2020).

Now we prove our results for RGD with a different step size. The proof of iterate boundedness is similar to our proof of Theorem 1.

Theorem 3. [↓] *Consider a manifold $\mathcal{M}_{\text{LB}} \in \mathcal{R}_{\text{LB}}$. Let $f \in \mathcal{F}_L(\mathcal{X})$, for $\mathcal{X} \stackrel{\text{def}}{=} \bar{B}(x^*, R\sqrt{3/2}) \subset \mathcal{M}_{\text{LB}}$. The iterates of RGD with step size $\eta \stackrel{\text{def}}{=} 1/(\zeta_{\sqrt{3/2}R}L)$ satisfy $x_t \in \mathcal{X}$. The convergence rate is $O(\zeta_R LR^2/\varepsilon)$. If f is also μ-strongly g-convex in $\mathcal{X}$, then it takes $O((\zeta_R L/\mu)\log(LR^2/\varepsilon))$ iterations.*

Proof. (Theorem 3) Define $\zeta \stackrel{\text{def}}{=} \zeta_{R\sqrt{3/2}}$. We begin by showing that the choice of η suffices to ensure that $d(x_t, x^*) \leq \sqrt{3/2}R$ for all $t \geq 0$ via induction. By assumption, the statement holds for $t = 0$. Suppose the property holds for t, then we show that it also holds for $t+1$. By the induction statement, we have that $x_t \in \bar{B}(x^*, \sqrt{3/2}R)$. Let $\eta_* \stackrel{\text{def}}{=} \max_{\eta>0}\{\eta \mid \operatorname{Exp}_{x_t}(-\eta_*\nabla f(x_t)) \in \bar{B}(x^*, \sqrt{3/2}R)\}$. Note we can write a maximum because the ball is compact. It is enough to show $\eta_* \geq \frac{1}{\zeta L}$. Suppose for the sake of contradiction that $\eta_* < \frac{1}{\zeta L}$ and let $x'_{t+1} \stackrel{\text{def}}{=} \operatorname{Log}_{x_t}(-\eta_*\nabla f(x_t))$. By definition, it must be $d(x'_{t+1}, x^*) = \sqrt{3/2}R$. We use the shorthands $\Delta_i \stackrel{\text{def}}{=} f(x_i) - f(x^*)$ and $g_i \stackrel{\text{def}}{=} \nabla f(x_i)$. By L-smoothness of f in $\mathcal{X}$, we have for all $i < t$:

$$\Delta_{i+1} - \Delta_i \leq \langle g_i, \operatorname{Log}_{x_i}(x_{i+1})\rangle + \frac{L}{2}d(x_{i+1}, x_i)^2 = \left(-\frac{1}{\zeta L} + \frac{1}{2\zeta^2 L}\right)\|g_i\|^2 = -\frac{2\zeta - 1}{2\zeta^2 L}\|g_i\|^2. \tag{12}$$

Similarly, we have

$$\Delta'_{t+1} - \Delta_t \leq \langle g_i, \operatorname{Log}_{x_t}(x'_{t+1})\rangle + \frac{L}{2}d(x_{t+1}, x_t)^2 = \left(-\eta_* + \frac{\eta_*^2 L}{2}\right)\|g_t\|^2 \overset{①}{<} -\eta_*\frac{2\zeta - 1}{2\zeta}\|g_t\|^2. \tag{13}$$

where ① is equivalent to $\eta_* \in (0, 1/(L\zeta))$. By the g-convexity of f, the cosine inequality Lemma 19 and $\zeta_{d(x_i,x^*)} \leq \zeta$ which holds by the induction hypothesis and the monotonicity of $r \mapsto \zeta_r$ we obtain ① below for $i < t$

$$\begin{aligned}\Delta_i \leq \langle -g_i, \operatorname{Log}_{x_i}(x^*)\rangle &\overset{①}{\leq} \frac{1}{2\eta}\left[d(x_i,x^*)^2 - d(x_{i+1},x^*)^2 + \zeta d(x_i,x_{i+1})^2\right] \\ &\overset{②}{=} \frac{L\zeta}{2}\left[d(x_i,x^*)^2 - d(x_{i+1},x^*)^2\right] + \frac{1}{2L}\|g_i\|^2,\end{aligned} \tag{14}$$

where ② follows by the definition of x_{i+1} and η. Similarly, we have

$$\Delta_t \leq \frac{1}{2\eta_*}\left(d(x_t, x^*)^2 - d(x'_{t+1}, x^*)^2\right) + \frac{\zeta\eta_*}{2}\|g_t\|^2. \tag{15}$$

Multiplying (12) by $C \stackrel{\text{def}}{=} \zeta^2/(2\zeta-1)$ and adding it to (14), we have for $i < t$

$$C\Delta_{i+1} - (C-1)\Delta_i \leq \frac{L\zeta}{2}\left(d(x_i, x^*)^2 - d(x_{i+1}, x^*)^2\right). \tag{16}$$

And analogously multiplying (13) by C and adding it to (15) and multiplying by $\eta_* L\zeta < 1$:

$$\eta_* L\zeta(C\Delta'_{t+1} - (C-1)\Delta_t) < \frac{L\zeta}{2}\left(d(x_t, x^*)^2 - d(x'_{t+1}, x^*)^2\right). \tag{17}$$

Now summing (16) from $i = 0$ to $t-1$ and adding (17), we obtain ② below:

$$\begin{aligned}
0 &\overset{①}{\leq} \eta_* L\zeta C\Delta'_{t+1} + (1-\eta_* L\zeta)(C-1)\Delta_t + \sum_{i=1}^{t}\Delta_i \\
&\overset{②}{<} (C-1)\Delta_0 + \frac{L\zeta d(x_0, x^*)^2}{2} - \frac{L\zeta d(x'_{t+1}, x^*)^2}{2} \\
&\overset{③}{\leq} (C-1+\zeta)\frac{LD^2}{2} - \frac{L\zeta d(x'_{t+1}, x^*)^2}{2} \\
&\overset{④}{\leq} \frac{3\zeta}{2}\cdot\frac{LD^2}{2} - \frac{L\zeta d(x'_{t+1}, x^*)^2}{2}
\end{aligned} \tag{18}$$

where ① holds since $\Delta_i, \Delta'_{t+1} \geq 0$, $\eta_* L\zeta < 1$, and ③ follows from $\Delta_0 \leq \frac{LD^2}{2}$, which is implied by the L-smoothness of f. Finally ④ can be shown since $C - 1 + \zeta$ is increasing for $\zeta \in [1, \infty)$ and the limit at $+\infty$ is $3/2$. In particular, it follows that $d(x'_{t+1}, x^*)^2 < (3/2)R^2$, which contradicts the fact that $d(x'_{t+1}, x^*)^2 = (3/2)R^2$. We note that for $\zeta = 1$, we would have $C - 1 + \zeta = 1$ and we could make the radius of the ball equal to RD. Hence, $\eta_* \geq \frac{1}{L\zeta}$ and the inductive statement is proven.

Now, using $\eta = 1/(\zeta L)$ to define x_{t+1}, we have that (18) holds for $i = t$ as well. And adding (18) up from $i = 0$ to t, we conclude

$$0 \leq (t+1)\Delta_{t+1} \overset{①}{\leq} C\Delta_{t+1} + \sum_{i=1}^{t}\Delta_i \leq \frac{3\zeta LR^2}{4} - \frac{L\zeta}{2}d(x_{t+1}, x^*)^2, \tag{19}$$

where ① holds since $C \geq 1$, and also $\Delta_{t+1} \leq \Delta_i$ for $i \leq t$ which follows from (12). Now, (19) with $t \leftarrow (T-1)$ implies that $\Delta_T \leq \frac{3\zeta LR^2}{4T}$. Thus, we have that $\Delta_T \leq \varepsilon$ for $T \geq \frac{3\zeta LD^2}{4\varepsilon}$. We now prove the result for the μ-strongly g-convex case. The algorithm is the same, and thus the iterates stay in $\bar{B}(x^*, \sqrt{3/2}R)$. The guarantee we just showed for the g-convex case implies ② below:

$$\frac{\mu}{2}d(x_T, x^*)^2 \overset{①}{\leq} f(x_T) - f(x^*) \overset{②}{\leq} \frac{3\zeta Ld(x_0, x^*)^2}{4T} \overset{③}{\leq} \frac{\mu}{4}d(x_0, x^*)^2$$

where ① holds by μ-strong g-convexity and ③ holds if $T = \lceil 3\zeta\frac{L}{\mu}\rceil$. Consequently, after $O(\zeta\frac{L}{\mu})$ iterations we reduce the distance squared to the minimizer by a factor of 2. Applying the same argument again sequentially for $r \stackrel{\text{def}}{=} \lceil \log_2(\frac{LD^2}{\varepsilon})\rceil$ stages of length T, we obtain that after $\hat{T} \stackrel{\text{def}}{=} rT = O(\zeta\frac{L}{\mu}\log(\frac{LD^2}{\varepsilon}))$, iterations we have

$$f(x_{\hat{T}}) - f(x^*) \leq Ld(x_{\hat{T}}, x^*)^2 \leq \frac{Ld(x_0, x^*)^2}{2^r} \leq \varepsilon.$$

□

Theorem 6 (Non-smooth RGD). [↓] *Consider a manifold $\mathcal{M}_{\mathrm{LB}} \in \mathcal{R}_{\mathrm{LUB}}$ and $f : \mathcal{M}_{\mathrm{LB}} \to R$ that is L_{p}-Lipschitz in $\mathcal{X}$, for $\mathcal{X} \stackrel{\text{def}}{=} \bar{B}(x^*, \sqrt{2}R) \subset \mathcal{M}_{\mathrm{LB}}$. The iterates of Riemannian subgradient descent with $\eta \stackrel{\text{def}}{=} R/(L_{\mathrm{p}}\sqrt{\zeta_{\sqrt{2}R}T})$ lie in $\mathcal{X}$ and the geodesic average of the iterates, cf. (2), is an ε-minimizer of f after $O(\sqrt{\zeta_R}L_{\mathrm{p}}R/\varepsilon^2)$ iterations.*

Proof. (Theorem 6) Denote by $v_i \in \partial f(x_i)$ the subgradients obtained and used by the algorithm. By the Lipschitzness assumption, it is $\|v_i\| \leq L_{\mathrm{p}}$. We show by induction that $d(x_t, x^*) \leq R$ for all $t \geq 0$. For $t = 0$ the statement holds by definition. Assume that the statement holds for $t \leq T-1$, then we show that is also holds for $t+1$. By g-convexity of f, we have for $i \leq t$

$$\begin{aligned} f(x_i) - f(x^*) \leq \langle v_i, -\mathrm{Log}_{x_i}(x^*)\rangle &= \frac{1}{\eta}\langle -\mathrm{Log}_{x_i}(x_{i+1}), -\mathrm{Log}_{x_i}(x^*)\rangle \\ &\overset{①}{\leq} \frac{1}{2\eta}\left[\zeta_R d(x_i, x_{i+1})^2 + d(x_i, x^*)^2 - d(x_{i+1}, x^*)^2\right] \\ &\overset{②}{\leq} \frac{1}{2\eta}\left[d(x_i, x^*)^2 - d(x_{i+1}, x^*)^2\right] + \frac{\zeta_R L_{\mathrm{p}}^2 \eta}{2}, \end{aligned}$$

where ① holds by the cosine inequality Remark 20 and the monotonicity of $R \mapsto \zeta_R$. Further, ② uses the definition of x_{i+1} and Lipschitzness of f in $\mathcal{X}$. Summing up the previous equation from $i = 0$ to t, using $d(x_0, x^*) = R \leq \sqrt{2}R$, $t \leq T-1$, and $\eta = \frac{R}{L_{\mathrm{p}}\sqrt{\zeta_R T}}$ yields

$$0 \leq 2\eta \sum_{i=0}^{t}[f(x_i) - f(x^*)] \leq R^2 - d(x_{t+1}, x^*)^2 + \zeta_R (t+1) {L_{\mathrm{p}}}^2 \eta^2 \leq -d(x_{t+1}, x^*)^2 + 2R^2. \tag{20}$$

This proves the induction statement, since $d(x_{t+1}, x^*) \leq \sqrt{2}R$. From (20) with $t \leftarrow T-1$ and dropping the negative distance term, we obtain

$$\frac{1}{T}\sum_{t=0}^{T-1}[f(x_t) - f(x^*)] \leq \frac{R^2}{\eta T} = \frac{\sqrt{\zeta_R} L_{\mathrm{p}} R}{\sqrt{T}}.$$

Lastly, note that geodesic average of $\{x_0, \ldots, x_{T-1}\}$ denoted by $\bar{x}_{T-1}$ as defined by (2) satisfies $f(\bar{x}_{T-1}) \leq \frac{1}{T}\sum_{t=0}^{T-1} f(x_t)$ by Corollary 26. $\square$

B RPPA proofs

Proposition 7 (RPPA). [↓] *Consider $\mathcal{M} \in \mathcal{R}_{\mathrm{LUB}}$ and $f \in \mathcal{F}(\mathcal{M})$ with $\mathcal{X} \overset{\mathrm{def}}{=} \bar{B}(x^*, R) \subset \mathcal{M}$. For any $\eta > 0$ and all $t \geq 0$, the iterates of the exact RPPA, cf. (4), satisfy $d(x_{t+1}, x^*) \leq d(x_t, x^*)$. In particular, it is $x_t \in \mathcal{X}$.*

Proof. (Proposition 7) For $t = 0$, $x_0 \in \bar{B}(x^*, R)$ by definition. Fix $t \geq 0$ and assume $d(x_t, x^*) \leq R$, then we are done if we show $d(x_{t+1}, x^*) \leq d(x_t, x^*) \leq R$. By Lemma 27, it is $d(x_t, x_{t+1}) \leq d(x_t, x^*) \leq R$. By the triangular inequality, we have that the diameter of the geodesic triangle $\triangle x_{t+1}x_t x^*$ is $2R$. This fact along with the monotonicity $r \mapsto \delta_r$ and the cosine inequality Lemma 19 implies ② below

$$\begin{aligned} 0 \leq f(x_{t+1}) - f(x^*) &\overset{①}{\leq} \frac{1}{\eta}\langle -\mathrm{Log}_{x_{t+1}}(x_t), \mathrm{Log}_{x_{t+1}}(x^*)\rangle \\ &\overset{②}{\leq} -\frac{\delta_{2R}}{2} d(x_{t+1}, x_t)^2 - \frac{1}{2} d(x_{t+1}, x^*)^2 + \frac{1}{2} d(x_t, x^*)^2 \\ &\overset{③}{\leq} -\frac{1}{2} d(x_{t+1}, x^*)^2 + \frac{1}{2} d(x_t, x^*)^2, \end{aligned} \tag{21}$$

where ① holds because by the first-order optimality condition in the definition of x_{t+1}, we have that $-\frac{1}{\eta}\mathrm{Log}_{x_{t+1}}(x_t) \in \partial f(x_{t+1})$. In ② we use Lemma 19 and in ③, we drop one negative term. The conclusion from this inequality is what we desired to prove. $\square$

Theorem 8 (RIPPA). [↓] *Let $\mathcal{M} \in \mathcal{R}_{\mathrm{LUB}}$ and let $f \in \mathcal{F}(\mathcal{M})$ and assume $\bar{B}(x^*, 3R) \subset \mathcal{M}$. Using the notation in Algorithm 1, it holds that $x_t \in \bar{B}(x^*, \sqrt{2}R)$ for every $t \geq 0$, and the output of Algorithm 1 after $T = O(\frac{R^2}{\eta\varepsilon})$ iterations is an ε-minimizer of f. If f is μ-strongly convex in $\bar{B}(x^*, \sqrt{2}R)$, then $d(x_{t+1}, x^*)^2 \leq \frac{1}{1+\eta\mu/2} d(x_t, x^*)^2$ and in particular $d(x_T, x^*)^2 \leq \varepsilon_d$ after $T = O((1 + \frac{1}{\mu\eta})\log(\frac{R^2}{\varepsilon_d}))$ iterations.*

Proof. (Theorem 8) We show by induction that $x_t \in B(x^*, \sqrt{2}R)$ for all $t \geq 0$. For $t = 0$ the property holds by definition. Now assume it holds for t, we will prove it for $t+1$. We first show that $d(x_{t+1}, x^*) \leq 3R$. We have that

$$d(x_{t+1}, x^*) \leq d(x_{t+1}, x^*_{t+1}) + d(x^*, x^*_{t+1}) \overset{①}{\leq} \frac{1}{2} d(x_t, x^*_{t+1}) + d(x_t, x^*) \overset{②}{\leq} \frac{3}{2} d(x_t, x^*) \overset{③}{\leq} 3R. \tag{22}$$

where in ① we used the criterion in Line 2 and that by Proposition 7 it is $d(x^*_{t+1}, x^*) \leq d(x_t, x^*)$. In ② we used Lemma 27, and we use the induction hypothesis in ③. We conclude that $\operatorname{diam}(\triangle x_{t+1} x_t x^*) \leq d(x_{t+1}, x^*) + d(x_t, x^*) \leq 5R$. By μ-strong g-convexity of f, with possibly $\mu = 0$, $v_{t+1} \in \partial f(x_{t+1})$ and the definition of r_{t+1}, we have

$$\begin{aligned}
0 &\leq f(x_{t+1}) - f(x^*) \\
&\leq -\langle v_{t+1}, \operatorname{Log}_{x_{t+1}}(x^*)\rangle - \frac{\mu}{2} d(x_{t+1}, x^*)^2 \\
&= \frac{1}{\eta}\langle -\operatorname{Log}_{x_{t+1}}(x_t) - r_{t+1}, \operatorname{Log}_{x_{t+1}}(x^*)\rangle - \frac{\mu}{2} d(x_{t+1}, x^*)^2 \\
&\overset{①}{\leq} \frac{1}{\eta}\left(-\frac{\delta_{5R}}{2} d(x_{t+1}, x_t)^2 - \frac{1}{2} d(x_{t+1}, x^*)^2 + \frac{1}{2} d(x_t, x^*)^2\right) + \frac{1}{\eta}\left(\frac{\|r_{t+1}\|^2}{2\Delta_{t+1}} + \frac{\Delta_{t+1}}{2} d(x_{t+1}, x^*)^2\right) - \frac{\mu}{2} d(x_{t+1}, x^*)^2 \\
&\overset{②}{\leq} -\frac{1}{2\eta} d(x_{t+1}, x^*)^2 + \frac{1}{2\eta} d(x_t, x^*)^2 + \frac{\Delta_{t+1}}{2\eta} d(x_{t+1}, x^*)^2 - \frac{\mu}{2} d(x_{t+1}, x^*)^2 \\
&= \frac{1}{2\eta}\left((\Delta_{t+1} - 1 - \eta\mu) d(x_{t+1}, x^*)^2 + d(x_t, x^*)^2\right).
\end{aligned} \tag{23}$$

where in ①, we used the cosine inequality Lemma 19 for the first term in the inner product. We also bounded the second term in the inner product by Young's inequality. In ② we use the criterion in Line 2 to bound $\|r_{t+1}\|^2$ and cancel the result with the first summand. We now separate two cases:

G-convex setting. From (23) with $\mu = 0$ and $\Delta_{t+1} = (t+2)^{-2}$, we have that

$$d(x_{t+1}, x^*)^2 \leq (1 - \Delta_{t+1})^{-1} d(x_t, x^*)^2 \leq \prod_{i=0}^{t} \frac{1}{1 - \Delta_{i+1}} d(x_0, x^*)^2 \overset{①}{\leq} 2R^2, \tag{24}$$

where ① holds since $\prod_{i=0}^{t} \frac{1}{1-(t+c)^{-2}} \leq \frac{c}{c-1}$, by Proposition 18. This proves the induction statement. Note that changing the value of Δ_{t+1}, we could have reduced the constant $2R^2$ above to something as close to R^2 as we want. For the convergence, we now sum (23) from $t = 0$ to $T-1$, divide by T and use $d(x_0, x^*) \leq R$:

$$\frac{1}{T}\sum_{t=0}^{T-1} f(x_{t+1}) - f(x^*) \leq \frac{1}{2\eta T}\left(R^2 - d(x_T, x^*)^2 + \sum_{t=0}^{T-1} \Delta_{t+1} d^2(x_{t+1}, x^*)\right) \overset{①}{\leq} \frac{1}{2\eta T}\left(R^2 + 2R^2 \sum_{t=0}^{T-1} \Delta_{t+1}\right) \overset{②}{\leq} \frac{3R^2}{\eta T}$$

In ①, we dropped a negative term and used (24). Then, ② follows from $\sum_{t=0}^{T-1} \Delta_{t+1} \leq \sum_{t=1}^{\infty} \frac{1}{t^2} \leq \frac{\pi^2}{6} \leq 2$. By using the uniform averaging scheme in Corollary 26, we obtain $f(\bar{x}_T) \leq \frac{1}{T}\sum_{t=0}^{T-1} f(x_{t+1})$, which concludes the proof for the g-convex case.

Strongly g-convex setting By the definition of η and Δ_{t+1} in the strongly g-convex case, (23) implies

$$d(x_{t+1}, x^*)^2 \leq \frac{d(x_t, x^*)^2}{1 + \eta\mu/2} \leq \ldots \leq (1 + \eta\mu/2)^{-(t+1)} R^2, \tag{25}$$

Since $\eta > 0$, we have that $0 < \frac{1}{1+\eta\mu/2} < 1$ and hence $d(x_{t+1}, x^*) \leq R$, which proves the induction statement. Now, if we set $T \geq (1 + 2/(\mu\eta)) \ln(\frac{R^2}{\varepsilon_d}) = \widetilde{O}(1 + \frac{1}{\mu\eta})$, we have that (25) with $t \leftarrow T-1$ implies

$$d(x_T, x^*)^2 \leq \left(1 - \frac{1}{1 + 2/(\mu\eta)}\right)^T R^2 \leq R^2 \exp\left(\frac{-T}{1 + 2/(\mu\eta)}\right) \leq \varepsilon_d,$$

which concludes the proof. $\square$

B.1 RIPPA implementation via composite RGD or PRGD

We start by showing that a particular implementation of composite RGD enjoys linear convergence, and as a corollary, we obtain that the subroutine in Line 2 of Algorithm 1 can be implemented by using only $\widetilde{O}(1)$ gradient oracle calls when applied to smooth g-convex optimization defined in Hadamard manifolds, and using $O(\zeta_R)$ iterations of PRGD. We note that if g is an indicator function, the composite RGD algorithm below is not the same as PRGD in general. That is, the resulting algorithm is a projected RGD that does not use a metric-projection.

Proposition 5 (Composite RGD). [↓] *Let $\mathcal{M} \in \mathcal{R}_{\mathrm{LUB}}$ and let $\mathcal{X} \subset \mathcal{M}$ be closed and g-convex. Given $f \in \mathcal{F}_L(\mathcal{X})$, and $g \in \mathcal{F}(\mathcal{X})$, such that $F \stackrel{\text{def}}{=} f + g$ is μ-strongly g-convex in $\mathcal{X}$, and $x^* \stackrel{\text{def}}{=} \arg\min_{x\in\mathcal{X}} F(x)$. We get, iterating the rule*

$$x_{t+1} = \arg\min_{y\in\mathcal{X}} \langle \nabla f(x_t), \mathrm{Log}_{x_t}(y)\rangle + \frac{L}{2} d(x_t, y)^2 + g(y),$$

an ε-minimizer of F in $O(\frac{L}{\mu}\log(\frac{F(x_0)-F(x^*)}{\varepsilon}))$ iterations.

Proof. (Proposition 5) We first note that the $\arg\min$ in the update rule exists. Since g is proper, lower semicontinuous and g-convex in $\mathcal{X}$, we have that $\mathcal{Y} \stackrel{\text{def}}{=} \mathcal{X} \cap \mathrm{dom}(g)$ is non-empty, closed and if $x \in \mathcal{Y}$ and $v \in \partial g(x)$, we have that $\{y \in \mathcal{Y} \mid \frac{L}{4} d(x_t, y)^2 + \langle v, \mathrm{Log}_x(y)\rangle \leq \frac{L}{4} d(x_t, x)^2\}$ is compact by strong convexity of $x \mapsto d(x_t, x)^2$. We also have that $\{y \in Y \mid \frac{L}{4} d(x_t, y)^2 + \langle \nabla f(x), \mathrm{Log}_{x_t}(y)\rangle \leq \frac{L}{4} d(x_t, x)^2 + \langle \nabla f(x), \mathrm{Log}_{x_t}(x)\rangle\}$ is compact. The union of these two compact sets is compact and if we consider z not in this union, we have ② below

$$\begin{aligned}\langle \nabla f(x_t), \mathrm{Log}_{x_t}(z)\rangle + \frac{L}{2} d(x_t, z)^2 + g(z) &\stackrel{①}{\geq} \langle \nabla f(x_t), \mathrm{Log}_{x_t}(z)\rangle + \frac{L}{2} d(x_t, z)^2 + g(x) + \langle v, \mathrm{Log}_x(z)\rangle \\ &\stackrel{②}{>} \langle \nabla f(x_t), \mathrm{Log}_{x_t}(x)\rangle + \frac{L}{2} d(x_t, x)^2 + g(x),\end{aligned}$$

where ① uses $v \in \partial g(x)$. This means that the minimization problem can be constrained to this union only and since it is compact the $\arg\min$ exists.

Now we prove the convergence result. We have

$$\begin{aligned}F(x_{t+1}) &\stackrel{①}{\leq} \min_{x\in\mathcal{X}} \left\{ f(x_t) + \langle \nabla f(x_t), x - x_t\rangle_{x_t} + \frac{L}{2} d(x, x_t)^2 + g(x) \right\} \\ &\stackrel{②}{\leq} \min_{x\in\mathcal{X}} \left\{ F(x) + \frac{L}{2} d(x, x_t)^2 \right\} \\ &\stackrel{③}{\leq} \min_{\alpha\in[0,1]} \left\{ \alpha F(x^*) + (1-\alpha) F(x_t) + \frac{L\alpha^2}{2} d(x^*, x_t)^2 \right\} \\ &\stackrel{④}{\leq} \min_{\alpha\in[0,1]} \left\{ F(x_t) - \alpha \left(1 - \alpha \frac{L}{\mu}\right) (F(x_t) - F(x^*)) \right\} \\ &\stackrel{⑤}{=} F(x_t) - \frac{\mu}{4L} (F(x_t) - F(x^*)).\end{aligned}$$

Above, ① holds by smoothness and the update rule of the composite Riemannian gradient descent algorithm. The g-convexity of f implies ②. Inequality ③ results from restricting the min to the geodesic segment between x^* and x_t so that $x = \mathrm{Exp}_{x_t}(\alpha \mathrm{Log}_{x_t}(x^*) + (1-\alpha)\mathrm{Log}_{x_t}(x_t))$. We also use the g-convexity of F. In ④, we used strong convexity of F to bound $\frac{\mu}{2} d(x^*, x_t)^2 \leq F(x_t) - F(x^*)$. Finally, in ⑤ we substituted α by the value that minimizes the expression, which is $\mu/2L$. The result follows by subtracting $F(x^*)$ to the inequality above and recursively applying the resulting inequality from $t = 1$ to $T \geq \frac{L}{4\mu}\log(\frac{F(x_0)-F(x^*)}{\varepsilon})$. □

We now show that for smooth functions, we can simplify the inexactness criterion of RIPPA.

Lemma 14. *Under the assumptions of Theorem 8, suppose that in addition $\bar{B}(x^*, 3R) \subset \mathcal{M}$, f is L-smooth in $\bar{B}(x^*, 2R)$ and let*

$$C_t \stackrel{\text{def}}{=} \min\left\{ 1/4, \frac{\Delta_{t+1}\delta_{3R}}{2(\eta L + \zeta_{3R})^2 + 2\Delta_{t+1}\delta_{3R}} \right\}. \tag{26}$$

It is enough that we guarantee $d(x_{t+1}^*, x_{t+1})^2 \leq C_t d(x_{t+1}^*, x_t)^2$ in order to satisfy the inexactness criterion in Line 2 of Algorithm 1 at iteration t.

Proof. Due to the definition of C_t, we just need to show the first part of the criterion in Line 2 of Algorithm 1. Fix $t \geq 0$. Firstly, we have

$$d(x_t, x_{t+1}) \leq d(x_t, x_{t+1}^*) + d(x_{t+1}^*, x_{t+1}) \overset{①}{\leq} \frac{3}{2} d(x_t, x^*),$$

where in ① we used $C_t \leq 1/4$ and the fact that by Lemma 27, it is $d(x_t, x_{t+1}^*) \leq d(x_t, x^*)$. So the diameter of $\triangle x_{t+1} x_t x_{t+1}^*$, which bounded by $\frac{1}{2}(d(x_t, x_{t+1}) + d(x_{t+1}^*, x_{t+1}) + d(x_t, x_{t+1}^*))$, is thus most $\frac{1}{2}(\frac{3}{2} + 1 + 1) d(x_t, x^*) \leq 2d(x_t, x^*)$. If the statement of Lemma 14 holds from iteration 0 to $t-1$, then by Theorem 8 we have $d(x_t, x^*) \leq 2R$. Now let $h_t(z) \stackrel{\text{def}}{=} f(z) + \frac{1}{2\eta} d(z, x_t)^2$ be the proximal function at step t which is thus smooth in $\triangle x_{t+1} x_t x_{t+1}^*$ with constant $\bar{L} \stackrel{\text{def}}{=} L + \zeta_{3R}/\eta$, where the $3R$ comes from $\max\{d(x_t, x_{t+1}), d(x_t, x^*)\}$ and the bound above. Note that by definition $r_{t+1} = \eta \nabla h_t(x_{t+1})$. Hence, we have

$$\|r_{t+1}\|^2 = \eta^2 \|\nabla h_t(x_{t+1})\|^2 \overset{①}{\leq} \bar{L}^2 \eta^2 d(x_{t+1}, x_{t+1}^*)^2 \overset{②}{\leq} \frac{2C_t}{1 - 2C_t} (\eta L + \zeta_{3R})^2 d(x_t, x_{t+1})^2 \tag{27}$$

Where ① is due to the $\bar{L}$-smoothness of h_t we just showed, and the fact $x_{t+1}^* \in \arg\min_{z \in \mathcal{M}} h_t(z)$. Further, ② follows by the inexactness criterion, i.e.,

$$\begin{aligned} & d(x_{t+1}, x_{t+1}^*)^2 \leq C_t d(x_t, x_{t+1}^*)^2 \leq 2C_t (d(x_t, x_{t+1})^2 + d(x_{t+1}, x_{t+1}^*)^2) \\ \Leftrightarrow\ & d(x_{t+1}, x_{t+1}^*)^2 \leq \frac{2C_t}{1 - 2C_t} d(x_t, x_{t+1})^2. \end{aligned}$$

Note that we want to prove $\|r_{t+1}\|^2 \leq \Delta_{t+1} \delta_{5R/2} d(x_t, x_{t+1})^2$, so by (27) it is enough that

$$C_t \leq \frac{\Delta_{t+1} \delta_{3R}}{2(\eta L + \bar{\zeta})^2 + 2\Delta_{t+1} \delta_{3R}}, \tag{28}$$

as specified in (26). Note that for simplicity, we used δ_{3R} which is less than $\delta_{5R/2}$. □

Finally, we can show that we can implement RIPPA for smooth functions.

Proposition 9. [↓] *In the setting of Theorem 8, suppose that in addition $\eta = 1/L$, $\bar{B}(x^*, 4R) \subset \mathcal{M}$, and f is g-convex and L-smooth in $\bar{B}(x^*, 4R)$. The composite Riemannian Gradient Descent of Proposition 5 in $\mathcal{X} \stackrel{\text{def}}{=} \bar{B}(x_t, 2R)$ implements the criterion in Line 2 of Algorithm 1 at iteration t using $\widetilde{O}(1/\delta_{4R})$ gradient oracle queries. If $\mathcal{M}$ is Hadamard, PRGD in $\mathcal{X}$ implements the criterion after $\widetilde{O}(\zeta_R)$ iterations.*

Proof. (Proposition 9) We show that we can implement the subroutine at iteration t, assuming that it was successfully implemented in previous iterations and thus according to Theorem 8, we have $x_t \in \bar{B}(x^*, 2R)$. Note that in such a case $\mathcal{X} \stackrel{\text{def}}{=} \bar{B}(x_t, 2R) \subset \bar{B}(x^*, 4R)$ and so by assumption f is L-smooth in $\mathcal{X}$. Also, note that $x^* \in \mathcal{X}$.

We now use the composite Riemannian Gradient Descent in Proposition 5 with the function f of the statement and with $g(x) = \frac{1}{2\eta} d(x_t, x)^2 = \frac{L}{2} d(x_t, x)^2$, which is strongly g-convex in $\mathcal{X}$ with parameter $\mu \stackrel{\text{def}}{=} L/\delta_{4R}$, since $\operatorname{diam}(\mathcal{X}) = 4R$. Similarly, note that F is $L(1 + \zeta_{4R})$-smooth in $\mathcal{X}$. If we use $T \geq \frac{L}{4\mu} \log(\frac{2L(1+\zeta_{4R})}{\mu C}) = \widetilde{O}(\frac{1}{\delta_{4R}})$ iterations, we obtain

$$\begin{aligned} \frac{\mu}{2} d(z_T, x^*)^2 &\overset{①}{\leq} F(z_T) - F(x^*) \overset{②}{\leq} \exp(-T\mu/4L)(F(x_0) - F(x^*)) \\ &\overset{③}{\leq} \exp(-T\mu/4L) L (1 + \zeta_{4R}) d(x_0, x^*)^2 \overset{④}{\leq} \frac{\mu C d(x_0, x^*)^2}{2}. \end{aligned}$$

Above, ① holds by μ-strong g-convexity, ② holds by the convergence guarantees in Proposition 5, ③ holds by smoothness of F and the fact that $\nabla F(x^*) = 0$ and finally ④ holds by the definition of T. This inequality is the criterion in Lemma 14, so the statement for composite RGD is proven.

For Hadamard manifolds, Martínez-Rubio et al. (2023) showed convergence of PRGD for functions in $\mathcal{F}_{\mu,L}(\mathcal{X})$, and in particular if the global minimizer is in the feasible set $\mathcal{X}$, the rates become $O(\zeta_{\operatorname{diam}(\mathcal{X})}\frac{L}{\mu})$. Thus, taking into account that in Hadamard $\delta_r = 1$ for all $r \geq 0$ and using PRGD and the exact same argument as above except for ③ and ④ in which we use these other convergence rates and $T = \widetilde{O}(\zeta_{4R}\frac{L}{\mu}) = \widetilde{O}(\zeta_R)$, we also arrive to the criterion in Lemma 14. Note that the constant C in the Lemma 14 is polynomial in problems parameters, such as ζ_R, $\frac{1}{\delta_{4R}}$, so the logarithm is benign. $\square$

C Prox properties

We start by showing that the smoothness of the Moreau envelope. After the proof, we provide some other alternative proofs that obtain a less general result. We include them because their techniques are very different and could be of independent interest.

C.1 Smoothness of Moreau envelope

Theorem 11 (Moreau envelope smoothness). [↓] *Consider a manifold $\mathcal{M}_{\mathrm{LB}} \in \mathcal{R}_{\mathrm{LB}}$ and let $\mathcal{X} \subset \mathcal{M}_{\mathrm{LB}}$ be a g-convex closed set. For a function $f \in \mathcal{F}(\mathcal{M})$, we have that the Moreau envelope of $g \stackrel{\mathrm{def}}{=} f + I_{\mathcal{X}}$ with parameter $\eta > 0$, defined for all $x \in \mathcal{M}_{\mathrm{LB}}$ as $M(x) \stackrel{\mathrm{def}}{=} \min_{z\in\mathcal{M}_{\mathrm{LB}}}\{f(z) + I_{\mathcal{X}}(z) + \frac{1}{2\eta}d(x,z)^2\}$, satisfies for all $x, y \in \mathcal{M}$:*

$$
\begin{aligned}
M(y) &\leq M(x) + \langle \nabla M(x), \operatorname{Log}_x(y)\rangle \\
&\quad + \frac{\zeta_{d(x,\operatorname{prox}_{\eta g}(x))}}{2\eta} d(x,y)^2.
\end{aligned}
$$

In particular, if $\mathcal{X}$ is compact and its diameter is D, the Moreau envelope $M(x)$ is (ζ_D/η)-smooth in $\mathcal{X}$.

Proof. (Theorem 11) Recall that we define $\operatorname{prox}_{\eta f}(x) \stackrel{\mathrm{def}}{=} \arg\min_{z\in\mathcal{M}_{\mathrm{LB}}}\{f(z) + I_{\mathcal{X}}(z) + \frac{1}{2\eta}d(x,z)^2\} \in \mathcal{X}$. The result is derived from the following.

$$
\begin{aligned}
M(y) &= \min_{z\in\mathcal{M}_{\mathrm{LB}}}\{f(z) + I_{\mathcal{X}}(z) + \frac{1}{2\eta}d(y,z)^2\} \\
&\overset{①}{\leq} f(\operatorname{prox}_{\eta f}(x)) + \frac{1}{2\eta}d(y,\operatorname{prox}_{\eta f}(x))^2 \\
&\overset{②}{=} M(x) - \frac{1}{2\eta}d(x,\operatorname{prox}_{\eta f}(x))^2 + \frac{1}{2\eta}d(y,\operatorname{prox}_{\eta f}(x))^2 \\
&\overset{③}{\leq} M(x) - \langle \operatorname{Log}_x(\operatorname{prox}_{\eta f}(x)), \operatorname{Log}_x(y)\rangle + \frac{\zeta_{d(x,\operatorname{prox}_{\eta f}(x))}}{2\eta}d(y,x)^2 \\
&\overset{④}{=} M(x) + \langle \nabla M(x), \operatorname{Log}_x(y)\rangle + \frac{\zeta_{d(x,\operatorname{prox}_{\eta f}(x))}}{2\eta}d(x,y)^2.
\end{aligned}
$$

Above, we just substituted in ① the variable in the $\min$ by $\operatorname{prox}_{\eta f}(x)$ yielding a possibly greater value. In ②, we used the definition of $M(x)$ and of $\operatorname{prox}_{\eta f}(x)$ and we used the cosine inequality Remark 20 in ③. In ④, we used Lemma 10. Since $\operatorname{prox}_{\eta f}(x) \in \mathcal{X}$, then if $x \in \mathcal{X}$ we have $d(x,\operatorname{prox}_{\eta f}(x)) \leq D$ and so given the inequality above, we have that $M(x)$ is indeed (ζ_D/η)-smooth in $\mathcal{X}$. $\square$

We now include alternative less general proofs of the fact proven in Theorem 11. They are strictly worse, but the techniques used can be of independent interest.

C.2 Alternative proofs

We start by showing the non-expansivity of the proximal operator. After we finished our result, we discovered that this fact was already proven in Jost (1995); Mayer (1998). We still include our proof since it is very different. A proof can also be found in the book (Bacák, 2014, Theorem 2.2.22).

Lemma 15 (Non-expansivity of the prox)**.** *Consider a function $f \in \mathcal{F}(\mathcal{H})$, where $\mathcal{H}$ is a Hadamard manifold, and let $x, y \in \mathcal{H}$, $x^+ \stackrel{\mathrm{def}}{=} \operatorname{prox}_f(x)$, $y^+ \stackrel{\mathrm{def}}{=} \operatorname{prox}_f(y)$. Then*

$$d(x^+, y^+) \le d(x, y).$$

Proof. Let $h_p(x) \stackrel{\mathrm{def}}{=} f(x) + \frac{1}{2\eta} d(x,p)^2$. Note that h_p is $(1/\eta)$-strongly g-convex, since f is g-convex. Define $x^+ = \arg\min_{z\in\mathcal{H}} h_x(z)$ and $y^+ = \arg\min_{z\in\mathcal{H}} h_y(z)$. We note that $\partial f(\cdot) + \frac{1}{\eta}\operatorname{Log}_y(\cdot) = \partial h_y(\cdot)$ and similarly for h_x. We choose a subgradient $g_{y^+}^f \in \partial f(y^+)$ and define subgradients $g_{x^+}^{h_x} \in \partial h_x(x^+)$, $g_{y^+}^{h_y} \stackrel{\mathrm{def}}{=} g_{y^+}^f + \frac{1}{\eta}\operatorname{Log}_{y^+}(y) \in \partial h_y(y^+)$ and $g_{y^+}^{h_x} \stackrel{\mathrm{def}}{=} g_{y^+}^f + \frac{1}{\eta}\operatorname{Log}_{y^+}(x) \in \partial h_x(y^+)$ so that

$$g_{y^+}^{h_x} - g_{y^+}^{h_y} = \frac{1}{\eta}\operatorname{Log}_{y^+}(x) - \frac{1}{\eta}\operatorname{Log}_{y^+}(y). \tag{29}$$

By Lemma 23, we have:

$$0 \le \langle g_{x^+}^{h_x}, \operatorname{Log}_{x^+}(z)\rangle, \forall z \in \mathcal{H} \qquad \text{and} \qquad 0 \le \langle g_{y^+}^{h_y}, \operatorname{Log}_{y^+}(z')\rangle, \forall z' \in \mathcal{H}.$$

Choosing $z = y^+$, $z' = x^+$, adding up and using Gauss lemma to transport to y^+, we obtain

$$0 \le \langle g_{y^+}^{h_y} - \Gamma_{x^+}^{y^+} g_{x^+}^{h_x}, \operatorname{Log}_{y^+}(x^+)\rangle. \tag{30}$$

Furthermore, by the $(1/\eta)$-strong g-convexity of h_x, we have

$$\begin{aligned} \frac{1}{2\eta} d(x^+, y^+)^2 + \frac{1}{2\eta} d(x^+, y^+)^2 &\le \Big(h_x(x^+) - h_x(y^+) - \langle g_{y^+}^{h_x}, \operatorname{Log}_{y^+}(x^+)\rangle_{y^+}\Big) \\ &\quad + \Big(h_x(y^+) - h_x(x^+) - \langle g_{x^+}^{h_x}, \operatorname{Log}_{x^+}(y^+)\rangle_{x^+}\Big) \\ &= \langle \Gamma_{x^+}^{y^+} g_{x^+}^{h_x} - g_{y^+}^{h_x}, \operatorname{Log}_{y^+}(x^+)\rangle. \end{aligned} \tag{31}$$

Summing up (30) and (31), we get ① below

$$\begin{aligned} \frac{1}{\eta} d(x^+, y^+)^2 &\stackrel{①}{\le} \langle g_{y^+}^{h_y} - g_{y^+}^{h_x}, \operatorname{Log}_{y^+}(x^+)\rangle \le \|g_{y^+}^{h_y} - g_{y^+}^{h_x}\| d(x^+, y^+) \\ &\stackrel{②}{=} \frac{1}{\eta}\|\operatorname{Log}_{y^+}(x) - \operatorname{Log}_{y^+}(y)\| d(x^+, y^+) \stackrel{③}{\le} \frac{1}{\eta} d(x,y) d(x^+, y^+). \end{aligned}$$

Therefore $d(x^+, y^+) \le d(x, y)$. We used (29) in ②. In ③ we use Lemma 22 which holds for Hadamard manifolds. ☐

Now we can give a slightly worse proof of the smoothness of the Moreau envelope by using Lemma 15.

Proposition 16. *Consider a $\mathcal{X} \subset \mathcal{H}$ closed and g-convex set where $\mathcal{H}$ is Hadamard Manifold. Let $f : \mathcal{H} \to \mathbb{R}$ be a g-convex function in $\mathcal{X}$. Then the gradient of the Moreau envelope $M(x) \stackrel{\mathrm{def}}{=} \min_{z\in\mathcal{H}}\{f(z) + I_{\mathcal{X}}(z) + \frac{1}{2\eta} d(x,z)^2\}$ is Lipschitz with constant $L \stackrel{\mathrm{def}}{=} \frac{1+\zeta}{\eta} = O(\zeta/\eta)$, i.e.,*

$$\|\nabla M(x) - \Gamma_y^x \nabla M(y)\|_x \le L d(x, y).$$

Proof. Let $x^+ \stackrel{\mathrm{def}}{=} \operatorname{prox}(x) = \arg\min_{z\in\mathcal{M}}\{f(z) + I_X(z) + \frac{1}{2\eta} d(x,z)^2\}$. The following holds:

$$\begin{aligned} \|\nabla M(x) - \Gamma_y^x \nabla M(y)\|_x &\stackrel{①}{=} \frac{1}{\eta}\|\operatorname{Log}_x(x^+) - \Gamma_y^x \operatorname{Log}_y(y^+)\|_x \\ &\stackrel{②}{\le} \frac{1}{\eta}\|\operatorname{Log}_x(x^+) - \operatorname{Log}_x(y^+)\|_x + \frac{1}{\eta}\|\operatorname{Log}_x(y^+) - \Gamma_y^x \operatorname{Log}_y(y^+)\|_x \\ &\stackrel{③}{\le} \frac{1}{\eta} d(x^+, y^+) + \frac{\zeta}{\eta} d(x, y) \stackrel{④}{\le} \frac{1+\zeta}{\eta} d(x, y). \end{aligned}$$

where ① holds by Lemma 10 and ② is the triangular inequality. The bound of the first summand in ③ is Lemma 22, which holds in Hadamard manifolds, and the bound of the second summand holds by Lemma 21. Finally ④ uses the non-expansivity of the prox, cf. Lemma 15. □

We also have the following proof of the smoothness of the Moreau envelope for twice differentiable functions by making use of partial differential equations.

Proposition 17. *Let $f : \mathcal{M} \to \mathbb{R}$ for a manifold $\mathcal{M}_{\mathrm{LB}} \in \mathcal{R}_{\mathrm{LB}}$ be a twice-differentiable g-convex function in some level set $\mathcal{X} \stackrel{\text{def}}{=} \{x \mid f(x) \leq f(p)\}$. Let $\zeta_D \stackrel{\text{def}}{=} \zeta_{\operatorname{diam}(\mathcal{X})}$. The Riemannian Moreau envelope $M(y) \stackrel{\text{def}}{=} \min_{x\in\mathcal{M}}\{f(x) + \frac{1}{2\eta}d(x,y)^2\}$ is (ζ_D/η)-smooth in $\mathcal{X}$.*

Proof. Let $y \in \mathcal{X}$ and define y^* as the $\arg\min$ in the problem that defines M, that is $y^* = \operatorname{prox}_{\eta f}(y)$. It is $y^* \in \mathcal{X}$ because any $z \notin \mathcal{X}$ yields $f(z) + \frac{1}{2\eta}d(y,z)^2 > f(y) + \frac{1}{2\eta}d(y,y)^2 \geq f(y^*) + \frac{1}{2\eta}d(y,y^*)^2$. Consider the first-order optimality condition of the problem in the definition of $M(y)$ and note y^* is a function of y. We have

$$\eta\nabla f(y^*) - \operatorname{Log}_{y^*}(y) = 0.$$

Differentiating with respect to y we obtain

$$\eta\nabla^2 f(y^*)\cdot\frac{\mathrm{d}y^*}{\mathrm{d}y} - \frac{\partial \operatorname{Log}_{y^*}(y)}{\partial y^*}\cdot\frac{\mathrm{d}y^*}{\mathrm{d}y} - \frac{\partial \operatorname{Log}_{y^*}(y)}{\partial y} = 0,$$

from which we deduce

$$\frac{\mathrm{d}y^*}{\mathrm{d}y} = \left(\eta\nabla^2 f(y^*) - \frac{\partial \operatorname{Log}_{y^*}(y)}{\partial y^*}\right)^{-1}\frac{\partial \operatorname{Log}_{y^*}(y)}{\partial y}. \tag{32}$$

Note that we can invert the matrix above since it is the Hessian of the strongly g-convex function $\eta f + \Phi_y$ at y^*. By Lemma 10, we obtain $\nabla M(y) = -\eta^{-1}\operatorname{Log}_y(y^*)$. We differentiate again with respect to y and use (32) to deduce:

$$\nabla^2 M(y) = -\frac{1}{\eta}\frac{\partial \operatorname{Log}_y(y^*)}{\partial y^*}\left(\eta\nabla^2 f(y^*) - \frac{\partial \operatorname{Log}_{y^*}(y)}{\partial y^*}\right)^{-1}\frac{\partial \operatorname{Log}_{y^*}(y)}{\partial y} + \frac{1}{\eta}\frac{-\partial \operatorname{Log}_y(y^*)}{\partial y}.$$

Now, the second of the two matrices being added above is $\eta^{-1}\nabla^2\Phi_{y^*}(y) \preccurlyeq (\zeta_{d(y,y^*)}/\eta)I$. The last inequality is due to Lemma 21. Note that $\nabla^2 M(y)$ is symmetric. The first matrix is symmetric and positive semidefinite. Indeed, we note that by (Lezcano-Casado, 2020, Theorem 3.12) one can conclude

$$\frac{\sqrt{|\kappa_{\min}|}d(y,y^*)}{\sinh(\sqrt{|\kappa_{\min}|}d(y,y^*))}I \preccurlyeq \frac{\partial \operatorname{Log}_y(y^*)}{\partial y^*},$$

and the symmetric statement with respect to y and y^*. For symmetric positive semidefinite matrices A, B it is $A - B \preccurlyeq A$ and so $\nabla^2 M(y) \preccurlyeq \eta^{-1}\nabla^2\Phi_{y^*}(y) \preccurlyeq (\zeta_{d(y,y^*)}/\eta)I$, so M is (ζ_D/η)-smooth. □

D Auxiliary results

Proposition 18. *For $c > 1$, and $T \in \mathbb{N}_0$ we have that*

$$\prod_{t=0}^{T}\frac{1}{1-(t+c)^{-2}} = \frac{c(c+T)}{(c-1)(c+T+1)} \leq \frac{c}{c-1}.$$

Proof. (Proposition 18) We show $\prod_{t=0}^{T}\frac{1}{1-(t+c)^{-2}} = \frac{c(c+T)}{(c-1)(c+T+1)}$ by induction. The statement holds for $T = 0$. Now assume that the statement holds for $T-1$. Then the statement also holds for T, which can be shown by noting that ① below holds by the induction hypothesis and rearranging

$$\prod_{t=0}^{T}\frac{1}{1-(t+c)^{-2}} \stackrel{①}{=} \frac{c(c+T-1)}{(c-1)(c+T)}\frac{1}{1-(T+c)^{-2}} = \frac{c(c+T)}{(c-1)(c+T+1)} \leq \frac{c}{c-1}.$$

□

E Geometric Auxiliary Results

In this section, we provide already established useful geometric results that we use in our proofs.

Lemma 19 (Riemannian Cosine-Law Inequalities)**.** *For the vertices $x, y, p \in \mathcal{M}$ of a uniquely geodesic triangle of diameter D, we have*

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

and

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

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

Remark 20. *In spaces with lower bounded sectional curvature, if we substitute the constants ζ_D in the previous Lemma 19 by the tighter constant and $\zeta_{d(p,x)}$, the result also holds. See Zhang and Sra (2016).*

We note that if $\kappa_{\min} < 0$, it is $\zeta_D = \Theta(1 + D\sqrt{|\kappa_{\min}|})$ and therefore if c is a constant, we have $\zeta_{cD} = O(\zeta_D)$. If $\kappa_{\min} \geq 0$ it is $\zeta_r = 1$, for all $r \geq 0$, so it also holds $\zeta_{cD} = O(\zeta_D)$.

Lemma 21. *Consider a manifold $\mathcal{M} \in \mathcal{R}_{\mathrm{LUB}}$ 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 $\Phi_x : \mathcal{M} \to \mathbb{R}$, $y \mapsto \frac{1}{2} d(x,y)^2$:*

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

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

See Kim and Yang (2022) for a proof, for instance. Note that the expression of $\nabla \Phi_x(y)$ along with Lemma 19 yields the smoothness and strong convexity inequalities.

Lemma 22. *Let $\mathcal{H}$ be a Hadamard manifold of sectional curvature bounded below by $\kappa_{\min}$. For any $x, y, z \in \mathcal{M}$, we have*

$$\|\mathrm{Log}_z(x) - \mathrm{Log}_z(y)\|_z \leq d(x,y).$$

Proof. Note that Hadamard manifolds are uniquely geodesic. Let D be the diameter of the geodesic triangle with vertices x, y, and z. Using the Euclidean cosine theorem in $T_x\mathcal{M}$ and Lemma 19 with $\delta_D = 1$, respectively, we have

$$\begin{aligned}
&2\langle \mathrm{Log}_z(x), \mathrm{Log}_z(y)\rangle = \|\mathrm{Log}_z(x)\|^2 + \|\mathrm{Log}_z(y)\|^2 - \|\mathrm{Log}_z(x) - \mathrm{Log}_z(y)\|^2, \\
&2\langle \mathrm{Log}_z(x), \mathrm{Log}_z(y)\rangle \geq d(z,x)^2 + d(z,y)^2 - d(x,y)^2.
\end{aligned}$$

Subtracting the first equation from the inequality below it, we obtain

$$0 \geq \|\mathrm{Log}_z(x) - \mathrm{Log}_z(y)\|^2 - d(x,y)^2.$$

□

Lemma 23. *Let $\mathcal{X} \subseteq \mathcal{M}$ be a closed uniquely geodesically convex set and let $f : \mathcal{M} \to \mathbb{R}$ be a differentiable g-convex function in $\mathcal{X}$. Let $x^* \in \arg\min_{x \in \mathcal{X}} f(x)$. We have*

$$\langle \nabla f(x^*), \mathrm{Log}_{x^*}(x)\rangle \geq 0, \textit{ for all } x \in \mathcal{X}.$$

Proof. Let f be g-convex and $x^* \in \arg\min_{x \in \mathcal{X}} f(x)$. Let $F(t) \stackrel{\mathrm{def}}{=} f(\gamma(t))$, where γ is a geodesic such that $\gamma(0) = x^*$ and $\gamma(d(x,x^*)) = x$. Then F reaches its minimum at $t = 0$ and we have that $0 \leq F'(0) = \langle \nabla f(x^*), \mathrm{Log}_{x^*}(x)\rangle$. □

Corollary 24 (Projection onto Geodesically Convex Sets)**.** *Consider a closed geodesically convex set $\mathcal{X} \subset \mathcal{M}$ in a manifold $\mathcal{M} \in \mathcal{R}_{\mathrm{LUB}}$, and let $\tilde{x} \in \mathcal{M}$. If $\kappa_{\max} > 0$, assume $\max_{x \in \mathcal{X}}\{d(x,\tilde{x})\} < \min\{\frac{\pi}{2\sqrt{\kappa_{\max}}}, \mathrm{inj}(\tilde{x})\}$ where $\mathrm{inj}(x)$ is the injectivity radius. We have $P_{\mathcal{X}}(\tilde{x})$ is unique and equal to $x^* = \arg\min_{x \in \mathcal{X}} \frac{1}{2} d(x,\tilde{x})^2$. Further, we have*

$$\langle \mathrm{Log}_{x^*}(\tilde{x}), \mathrm{Log}_{x^*}(z)\rangle_{x^*} \leq 0, \quad \forall z \in \mathcal{X}.$$

Proof. Apply Lemma 23 to the function $\Phi_{\tilde{x}} : \mathcal{X} \to \mathbb{R}, y \mapsto \frac{1}{2}d(\tilde{x}, y)^2$ whose gradient at the optimizer $x^* \in \mathcal{X}$ is $-\mathrm{Log}_{x^*}(\tilde{x})$. Finally, since the assumption implies $\Phi_{\tilde{x}}$ is strictly convex in $\mathcal{X}$, we have $d(x^*, z) < d(\tilde{x}, z)$ for all $z \in \mathcal{X} \setminus \{x^*\}$, so indeed $\mathcal{P}_{\mathcal{X}}(\tilde{x})$ is unique and is x^*. □

Lemma 25 (Geodesic averaging)**.** *Let $\mathcal{M} \in \mathcal{R}_{\mathrm{LUB}}$ and $f : \mathcal{M} \to \mathbb{R}$ be a g-convex function in a g-convex set $\mathcal{X} \subset \mathcal{M}$ and let $\{x_1, \ldots, x_T\}$ be points in $\mathcal{X}$. The geodesic average $\bar{x}_T$ defined recursively by*

$$\bar{x}_1 \leftarrow x_1, \quad t \in \{1, \ldots, T-1\} : \; \bar{x}_{t+1} \leftarrow \mathrm{Exp}_{\bar{x}_t}\left(\frac{w_{t+1}}{\sum_{j=1}^{t+1} w_j} \mathrm{Log}_{\bar{x}_t}(x_{t+1}) \right) \tag{33}$$

with $w_t > 0$ for all t satisfies $f(\bar{x}_T) \le \frac{1}{\sum_{t=1}^T w_t} \sum_{t=1}^T w_t f(x_t)$.

Proof. We prove the statement by induction. The statement holds for $T = 1$ by definition. Now assume that the statement holds for $T-1$, i.e., $f(\bar{x}_{T-1}) \le \frac{1}{\sum_{t=1}^{T-1} w_t} \sum_{t=1}^{T-1} w_t f(x_t)$. We show that the statement holds for T as well. By definition, $\bar{x}_T$ lies on the geodesic joining $\bar{x}_{T-1}$ and x_T. In particular, if we parametrize a geodesic segment joining $\bar{x}_{T-1}$ and $\gamma(1) = x_T$ as $\gamma : [0,1] \to \mathcal{M}$ with $\gamma(0) = \bar{x}_{T-1}$ and $\gamma(1) = x_T$, then $\gamma\left(\frac{w_T}{\sum_{t=1}^T w_t}\right) = \bar{x}_T$. Hence by g-convexity of f we have that,

$$\begin{aligned} f(\bar{x}_T) &\le \left(1 - \frac{w_T}{\sum_{t=1}^T w_t}\right) f(\bar{x}_{T-1}) + \frac{w_T}{\sum_{t=1}^T w_t} f(x_T) \\ &\overset{①}{\le} \frac{1}{\sum_{t=1}^T w_t} \sum_{t=1}^{T-1} w_t f(x_t) + \frac{w_T}{\sum_{t=1}^T w_t} f(x_T) = \frac{1}{\sum_{t=1}^T w_t} \sum_{t=1}^{T} w_t f(x_t) \end{aligned}$$

where ① holds by the induction hypothesis and the fact that $1 - \frac{w_T}{\sum_{t=1}^T w_t} = \frac{\sum_{t=1}^{T-1} w_t}{\sum_{t=1}^T w_t}$. □

Corollary 26. *Let $w_t = 1$ for all t, then the update rule simplifies to*

$$\bar{x}_1 \leftarrow x_1, \quad t \in \{1, \ldots, T-1\} : \; \bar{x}_{t+1} \leftarrow \mathrm{Exp}_{\bar{x}_t}\left(\frac{1}{t+1} \mathrm{Log}_{\bar{x}_t}(x_{t+1}) \right) \tag{34}$$

and we have $f(\bar{x}_T) \le \frac{1}{T} \sum_{t=1}^T f(x_t)$. We call this procedure uniform geodesic averaging.

The following lemma was proven in Martínez-Rubio and Pokutta (2023), but it was only stated in the context of Hadamard manifolds. We note that it works in the general Riemannian case.

Lemma 27. *Let $\mathcal{M} \in \mathcal{R}_{\mathrm{LUB}}$ and let $f : \mathcal{M} \to \mathbb{R}$ be g-convex, proper and lower semicontinuous in a set $\mathcal{X} \subset \mathcal{M}$. Let $x^* \in \arg\min_{x \in \mathcal{M}} f(x)$ and let $x^+ \overset{\mathrm{def}}{=} \mathrm{prox}_{\eta f}(x)$ for some $x \in \mathcal{M}$. Then $d(x, x^+)^2 \le d(x, x^*)^2$.*

Proof. Let $h(y) \overset{\mathrm{def}}{=} f(y) + \frac{1}{2\eta} d(y, x)^2$. By definition, we have that $f(x^*) \le f(x^+)$ and $h(x^+) \le h(x^*)$, hence

$$\frac{1}{2\eta} d(x, x^+)^2 - \frac{1}{2\eta} d(x, x^*)^2 \le f(x^+) - f(x^*) \frac{1}{2\eta} d(x, x^+)^2 - \frac{1}{2\eta} d(x, x^*)^2 = h(x^+) - h(x^*) \le 0.$$

It follows that $d(x, x^+) \le d(x, x^*)$. □

Proposition 28. *The optimizer x^* of (5) lies in $\mathcal{X}$.*

Proof. For the sake of contradiction, assume that $x^* \notin \mathcal{X}$. Denote by $\bar{x}^* \overset{\mathrm{def}}{=} \mathcal{P}_{\mathcal{X}}(x^*)$ the projection of x^* onto $\mathcal{X}$. By Corollary 24, we have $d(z, \bar{x}^*) \le d(z, x^*)$ for all $z \in \mathcal{X}$. By definition, $y_i \in \mathcal{X}$ for all i, hence

$$F(\bar{x}^*) \le \frac{1}{2} \sum_{i=1}^n d(\bar{x}^*, y_i)^2 \le \frac{1}{2} \sum_{i=1}^n d(x^*, y_i)^2 = F(x^*),$$

which contradicts the assumption. Hence $x^* \in \mathcal{X}$ which concludes the proof. □

E.1 Riemannian Generalized Danskin's theorem

We note that the Generalized Danskin's theorem (Bertsekas et al., 2003, Proposition 4.5.1) works in Riemannian manifolds. The reason is essentially that Danskin's theorem does not require convexity of the functions involved and we can talk about the functions retracted to the tangent space of $(x, y^*(x))$, apply the Euclidean Danskin's theorem and then use that the first-order information of the Riemannian function and the retracted function at $(x, y^*(x))$ is the same since retracting with the exponential map is a local isometry. Alternatively, one can see that the proof works without a problem in the Riemannian case.

Proposition 29 (Riemannian Generalized Danskin's Theorem)**.** *Let $\mathcal{M}, \mathcal{N}$ be uniquely geodesic Riemannian manifolds and let $Y \subset \mathcal{N}$ be g-convex and compact. Let $f : \mathcal{M} \times \mathcal{Y} \to \mathbb{R}$ be a continuous function. Then, the function $\phi(x) \stackrel{\text{def}}{=} \max_{y\in\mathcal{Y}} f(x, y)$ has directional derivative*

$$\phi'(x; v) = \max_{y\in\mathcal{Y}(x)} f'(x, y; v)$$

where $f'(x, y; v)$ is the directional derivative of $f(\cdot, y)$ at x with direction v, and $\mathcal{Y}(x)$ is the set of maximizing points in the definition of ϕ, that is $\mathcal{Y}(x) \stackrel{\text{def}}{=} \arg\max_{y\in\mathcal{Y}} f(x, y)$. If $\mathcal{Y}(x)$ is a singleton y^* and $f(\cdot, y^*(x))$ is differentiable at x, then ψ is differentiable at x and $\nabla\psi(x) = \nabla_x f(x, y^*(x))$.

Using the result above, we can provide the proof of Lemma 10 about the gradient of the Moreau envelope.

Lemma 10 (Gradient of Moreau envelope)**.** [↓] *Let $\mathcal{M}$ be a uniquely geodesically Riemannian manifold, let $\mathcal{X} \subset \mathcal{M}$ be a g-convex closed set. For $f \in \mathcal{F}(\mathcal{X})$, we define the Moreau envelope of $g \stackrel{\text{def}}{=} f + I_{\mathcal{X}}$ with $\eta > 0$ as $M(x) \stackrel{\text{def}}{=} \min_{z\in\mathcal{M}}\{f(z) + I_{\mathcal{X}}(z) + \frac{1}{2\eta} d(x, z)^2\}$. We have $\nabla M(x) = -\frac{1}{\eta}\mathrm{Log}_x(\mathrm{prox}_{\eta g}(x))$.*

Proof. (Lemma 10) In order to compute $\nabla M(\hat{x})$ it is enough to consider the function $F(x, y) \stackrel{\text{def}}{=} f(y)+I_{\mathcal{X}}(y)+\frac{1}{2\eta} d(x, y)^2$ for $x \in \hat{\mathcal{X}} \stackrel{\text{def}}{=} \bar{B}(\hat{x}, \delta)$ for any $\delta > 0$. In such a case, it is easy to see that we can restrict to y being in a compact $\mathcal{Y}$ in order to define $M(x) \stackrel{\text{def}}{=} \min_{y\in\mathcal{X}}\{f(y) + \frac{1}{2\eta} d(x, y)^2\}$ for all $x \in \hat{\mathcal{X}}$, that is, $M(x) \stackrel{\text{def}}{=} \min_{y\in\mathcal{Y}}\{f(y) + \frac{1}{2\eta} d(x, y)^2\}$. Indeed, consider $\mathcal{Y} = \{y \in \mathcal{X} : \langle v, \mathrm{Log}_{\hat{x}}(y)\rangle + \frac{1}{2\eta} d(\hat{x}, y)^2 \leq f(\hat{x})\}$ for a $v \in \partial f(\hat{x})$, then $\mathrm{Log}_{\hat{x}}(\mathcal{Y}) \subseteq T_{\hat{x}}\mathcal{M}$ is the level set of a quadratic plus I_X, which is compact and so $\mathcal{Y}$ is compact as well. Note that by definition, if $y \notin \mathcal{Y}$ then for all $x \in \hat{\mathcal{X}}$ we have $F(\hat{x}, y) \geq \langle v, \mathrm{Log}_{\hat{x}}(y)\rangle + \frac{1}{2\eta} d(\hat{x}, y)^2 > f(\hat{x}) = F(\hat{x}, \hat{x})$ so $y \notin \arg\min_{y\in\mathcal{X}} F(x, y)$. Thus, we can apply Proposition 29 with $\phi(x) = -M(x) = \max_{y\in\mathcal{Y}} -F(x, y)$ for F defined in the compact $\hat{\mathcal{X}} \times \mathcal{Y}$. The optimizer of $\max_{y\in\mathcal{Y}} -F(\hat{x}, y)$ is unique by strong convexity of $y \mapsto \frac{1}{2\eta} d(\hat{x}, y)^2$ and this point is $\mathrm{prox}_{\eta g}(\hat{x})$. Thus, $M(\cdot)$ is differentiable at $\hat{x}$ and $\nabla M(\hat{x}) = \nabla_x F(\hat{x}, \mathrm{prox}_{\eta g}(\hat{x})) = -\frac{1}{\eta}\mathrm{Log}_{\hat{x}}(\mathrm{prox}_{\eta g}(\hat{x}))$, as desired. □

F Numerical Results

We present numerical results for the normal and the hard Karcher mean for different values of n and d.

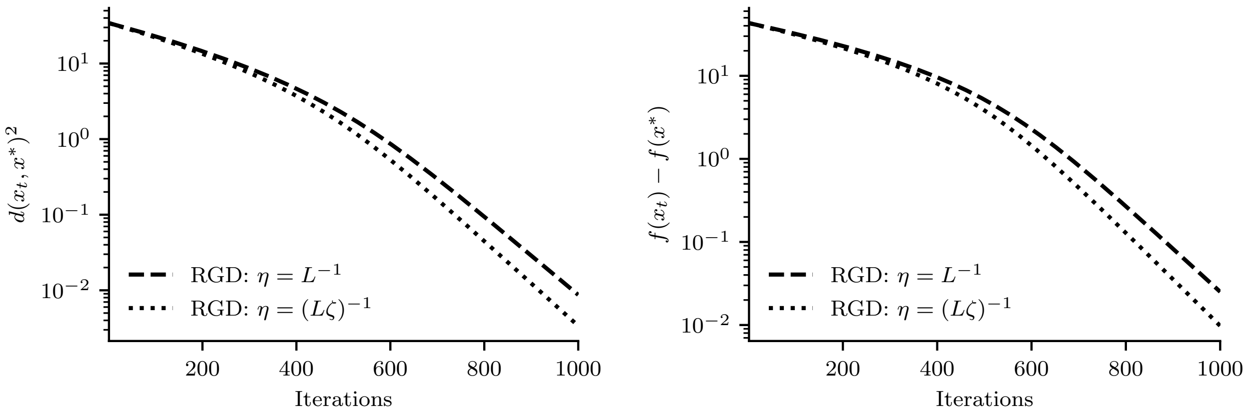

Figure 2: $n = 10^3$, $d = 10^3$, hard Karcher mean

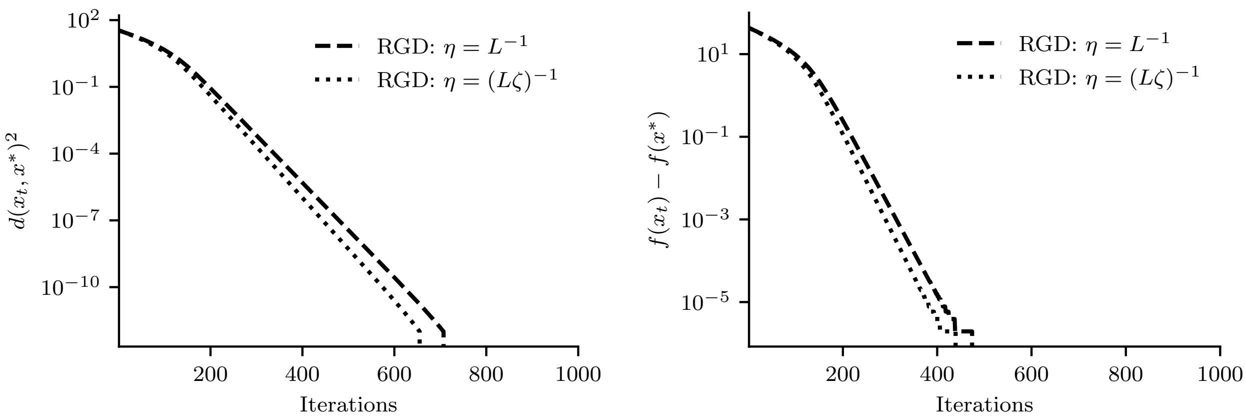

Figure 3: $n = 10^3$, $d = 10^3$, normal Karcher mean

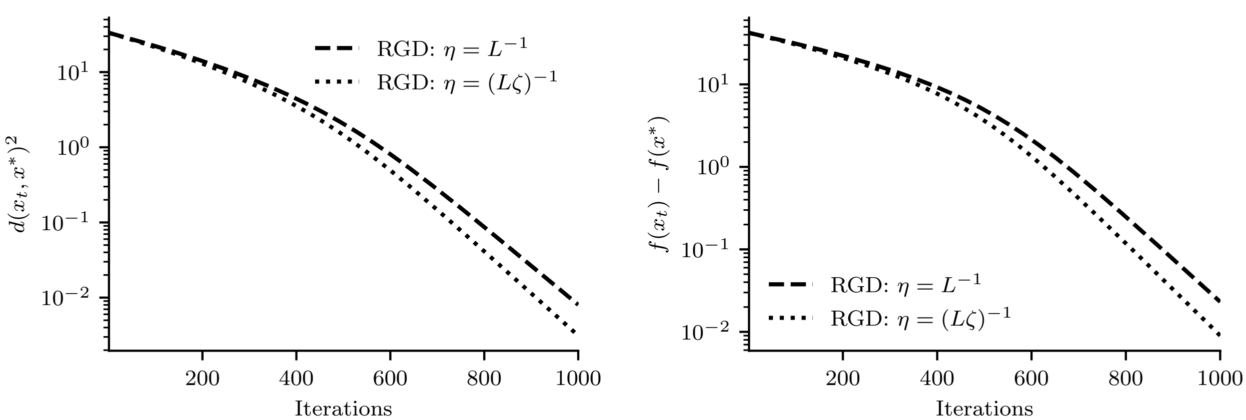

Figure 4: $n = 10^3$, $d = 10^4$, hard Karcher mean

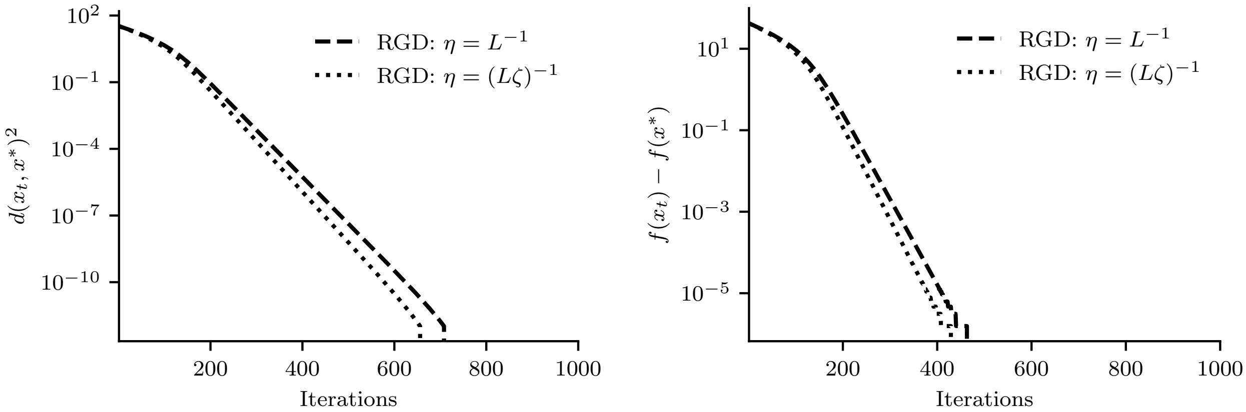

Figure 5: $n = 10^3$, $d = 10^4$, normal Karcher mean

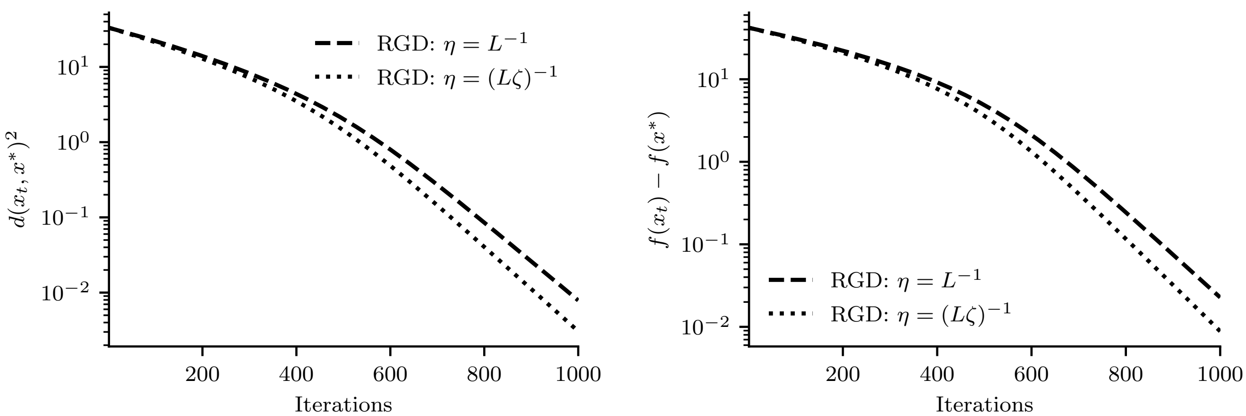

Figure 6: $n = 10^4$, $d = 10^4$, hard Karcher mean

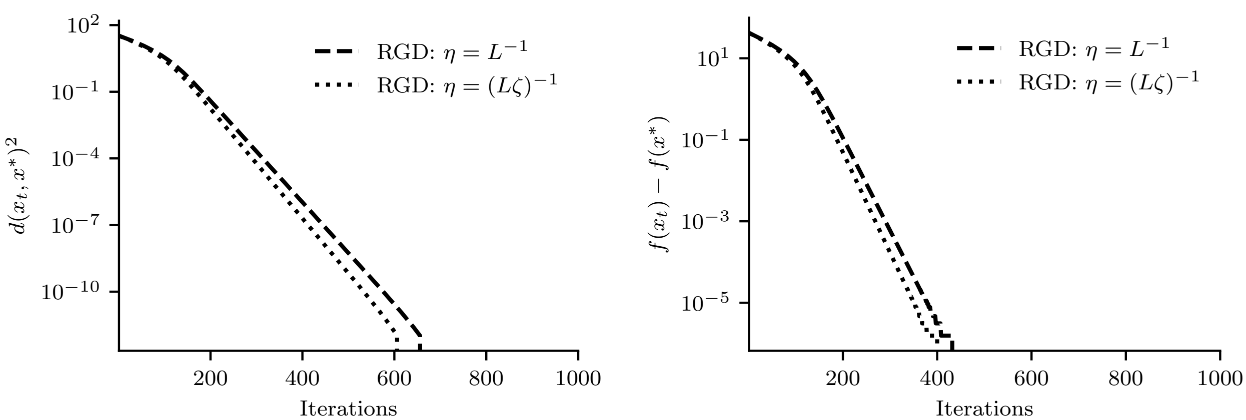

Figure 7: $n = 10^4$, $d = 10^4$, normal Karcher mean