Convergence and Trade-Offs in Riemannian Gradient Descent and Riemannian Proximal Point

David Martínez-Rubio [*1 2] **Christophe Roux** [*1 2] **Sebastian Pokutta** [1 2]

Abstract

In this work, we analyze two of the most fundamental algorithms in geodesically convex optimization: Riemannian gradient descent and (possibly inexact) Riemannian proximal point. We quantify their rates of convergence and produce different variants with several trade-offs. Crucially, we show the iterates naturally stay in a ball around an optimizer, of radius depending on the initial distance and, in some cases, on the curvature. Previous works simply assumed bounded iterates, resulting in rates that were not fully quantified. We also provide an implementable inexact proximal point algorithm and prove several new useful properties of Riemannian proximal methods: they work when positive curvature is present, the proximal operator does not move points away from any optimizer, and we quantify the smoothness of its induced Moreau envelope. Further, we explore beyond our theory with empirical tests.

1. Introduction

Riemannian optimization consists of the study of function optimization 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 & Melo, 2012; Bento et al., 2015).

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

A recurrent problem in Riemannian optimization algorithms is that geometric deformations appearing in their analyses scale with the distance between the iterates and between those and an optimizer. These distances are often bounded and quantified only by assumption (Zhang & Sra, 2016; Zhang et al., 2016; Zhang & Sra, 2018; Ahn & Sra, 2020; Kim & Yang, 2022; Zhang et al., 2022; Jordan et al., 2022). This assumption is the following: *there is compact g-convex set $\mathcal{X}$, that the algorithm has a priori access to, in which the iterates stay, i.e., $x_t \in \mathcal{X}$.*

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, see (Smith, 1994; Udriste, 1994; Cai et al., 2023) among others. This can be a problem, since unlike in the Euclidean space, where we can have globally smooth and strongly g-convex functions with constant condition number, in many Riemannian manifolds the condition number is lower bounded by a value that depends on the curvature and the diameter of the op-

[*]Equal contribution [1]Zuse Institute Berlin, Germany [2]Technische Universität Berlin, Germany. Correspondence to: Christophe Roux <roux@zib.de>, David Martínez-Rubio <martinez-rubio.zib.de>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

Most of the notations in this work have a link to their definitions, using this code, such as $\mathrm{Exp}_x(\cdot)$, which links to where it is defined as the exponential map of a Riemannian manifold.

timization domain (Martínez-Rubio, 2020; Criscitiello & 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.

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 unconstrained g-convex problems to a sequence 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 (Wang et al., 2023b; 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 & Sra, 2017; 2019), the Projected RGD algorithms surveyed in Section 3, and the accelerated constrained first-order methods in (Martínez-Rubio, 2020; Martínez-Rubio & 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, the convergence of two of the most fundamental classes of first-order methods is not fully understood. One of them is Riemannian gradient descent (RGD). Our first contribution is removing this limitation and quantifying the convergence rate and its dependence on geometric constants like the curvature. Secondly, we study the Riemannian proximal point algorithm (RPPA). We provide an inexact version (RIPPA) of it and convergence rates in general manifolds. Thirdly, for smooth functions we show how to implement the criterion of RIPPA in different ways and provide some variants of RGD. The iterate boundedness is the starting point of our work and importantly, under this framework, we obtain several algorithms with different convergence rates, where we trade off some dependence on the curvature for another or for optimizing in a larger set. The latter entails, for instance, greater lower bounds on the minimum possible condition number L/μ for μ-strongly g-convex L-smooth functions in the set. Lastly, we prove several new useful properties of Riemannian proximal methods. More precisely, our contributions are summarized in the following and in Table 1.

- **RGD**: Among other 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 instead we use a step size $\eta = 1/(\zeta_R L)$, the RGD update rule is quasi-nonexpansive. We quantify the rates of RGD in different settings as a result. A composite RGD, which implies reduced gradient complexity of solving the inexact prox in RIPPA.
- **RPPA**: A general analysis of RPPA. It was only known in Hadamard manifolds before. An inexact RPPA (RIPPA) and an implementation of it with first-order methods for smooth g-convex functions with quantified dependence on the curvature.
- **Prox properties**: The prox is quasi-nonexpansive, 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 $(\zeta_{\text{diam}(\mathcal{X})}/\eta)$-smooth in $\mathcal{X}$.
- **Experiments**: Numerical tests exploring beyond our theory. We observe that RGD presents a monotonic decrease in distance to an optimizer and show that RIPPA is competitive.

A future direction of research is studying whether one efficient algorithm can have the best rates and iterate bounds of all our algorithms at the same time, without assuming knowledge of the initial distance. Elucidating whether one such algorithm exists is a fundamental open problem. For the hyperbolic space, we do obtain such an algorithm. In Hadamard manifolds our method RIPPA-CRGD already obtains the best rate in terms of gradient complexity and iterate bounds, at the expense of solving subproblems that could be hard, and knowing the initial distance. The other methods can be implemented efficiently but have worse gradient complexity or iterate bounds.

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 $\langle v, w\rangle_x$ and $\|v\|_x \stackrel{\text{def}}{=} \sqrt{\langle v, v\rangle_x}$ for the metric's inner product and norm, and 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 $\mathrm{Exp}_x : T_x\mathcal{M} \to \mathcal{M}$ takes a point $x \in \mathcal{M}$, and a vector $v \in T_x\mathcal{M}$ and returns the point y we obtain from following the geodesic from x in the direction v for length $\|v\|$, if this is possible. We denote its inverse by $\mathrm{Log}_x(\cdot)$. It is well defined for uniquely geodesic manifolds, so we have $\mathrm{Exp}_x(v) = y$ and $\mathrm{Log}_x(y) = v$. 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 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. 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 less than $\pi/\sqrt{\kappa_{\min}}$. Along these lines, take into account that when we assume a ball of a certain radius is in $\mathcal{M}$, if $\kappa_{\max} > 0$ the radius may be restricted. This is not a limitation for instance in Hadamard manifolds, which are complete simply-connected Riemannian manifold of non-positive sectional curvature, and in particular are diffeomorphic to $\mathbb{R}^n$ and 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 g-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 possible condition numbers for functions in $\mathcal{F}_{\mu,L}(\mathcal{X})$ depends on $\mathcal{X}$.

Given $r > 0$, and a manifold $\mathcal{M} \in \mathcal{R}_{\mathrm{LUB}}$, we define the geometric constants $\zeta_r \overset{\mathrm{def}}{=} r\sqrt{|\kappa_{\min}|}\coth(r\sqrt{|\kappa_{\min}|}) = \Theta(1 + r\sqrt{|\kappa_{\min}|})$ if $\kappa_{\min} < 0$ and $\zeta_r \overset{\mathrm{def}}{=} 1$ otherwise, and $\delta_r \overset{\mathrm{def}}{=} r\sqrt{\kappa_{\max}}\cot(r\sqrt{\kappa_{\max}}) \leq 1$ if $\kappa_{\max} > 0$ and $\delta_r \overset{\mathrm{def}}{=} 1$ 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 $\Phi_x(y) \overset{\mathrm{def}}{=} \frac{1}{2}d(x, y)^2$ 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. We call a mapping $T : \mathcal{M} \to \mathcal{M}$ quasi-nonexpansive, if $d(T(x), x^*) \leq d(x, x^*)$, where x^* is a fixed-point of T.

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 $R \overset{\mathrm{def}}{=} d(x_0, x^*)$. 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 $R_{\max} \overset{\mathrm{def}}{=} \max_{i \in T} d(x_i, x^*)$. RGD is defined by the recursive update:

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

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

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

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

$$x_{t+1} \leftarrow \mathrm{prox}_{\eta f}(x_t), \tag{3}$$

where $\mathrm{prox}_{\eta f}(x) \overset{\mathrm{def}}{=} \arg\min_{z \in \mathcal{M}}\{f(z) + \frac{1}{2\eta}d(z, x)^2\}$ 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 $t = 0, \dots, T$, 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 $c = 1 - \frac{\mu^2}{L^2}$, i.e., satisfying $f(x_{t+1}) - f(x^*) \leq c(f(x_t) - f(x^*))$, 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 & 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 Lipschitz g-convex 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 & 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 & 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 & 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 convergence rates of different variants of unconstrained RGD, providing different trade-offs. Showing that the iterates stay naturally bounded in a set whose diameter we quantify allowed 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 & 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 generally for CAT(0) metric spaces, analogous to the classical Euclidean rates. Under a growth condition, (Tang & 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 the Kurdyka–Lojasiewicz inequality, without assuming the manifold is Hadamard. Espinola & Nicolae (2016); Kimura & Kohsaka (2017) also work in the general Riemannian case and obtain non-asymptotic convergence of an RPPA, but with a proximal operator not using the squared distance. In this work, we provide non-asymptotic rates for inexact RPPA in the general Riemannian case, which are the first of their 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. We present empirical results in Section 5.

4. Convergence Results and Bounded Iterates

We summarize the convergence results for general manifolds presented in this section in Table 1. Note that we obtain better results for the hyperbolic space. The proofs can be found in the appendix. Consider as an example a general Hadamard manifold $\mathcal{H}$. For a point $x \in \mathcal{H}$, for any $r > 0$, and for the ball $\bar{B}(x, r)$, we have that $\Phi_x(y) \stackrel{\text{def}}{=} \frac{1}{2}d(x, y)^2$ is $O(\zeta_r)$-smooth and 1-strongly convex, cf. Lemma 21. Using this fact and $\zeta_{O(R\zeta_R)} = O(\zeta_R^2)$, we have that for $\mathcal{H}$, 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. We note that for RGD in the hyperbolic space, we obtained better convergence rates than for the general case, namely $\widetilde{O}(\frac{LR^2}{\varepsilon})$, $\widetilde{O}(\frac{L}{\mu})$, and $D = O(R)$, respectively, matching the Euclidean rates, up to log factors.

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 $R \stackrel{\text{def}}{=} d(x_0, x^*)$, and let $\varphi \stackrel{\text{def}}{=}$

*This is the rate for Hadamard manifolds only, for the general case see Remark 12.

†This result only applies to Hadamard manifolds.

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^*. The column R? has a tick if the knowledge of the initial distance to an optimizer R is *not* required. All iterates stay in $\mathcal{X}$. Note that μ, L and the Lipschitz constant L_{p} depend on the respective different sets $\mathcal{X}$, so L in two rows need not mean the same. The value $\eta > 0$ is a proximal parameter.

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

$(1+\sqrt{5})/2$.

Theorem 1. [↓] *Consider a manifold $\mathcal{M}_{\mathrm{LB}} \in \mathcal{R}_{\mathrm{LB}}$, and $f \in \mathcal{F}_L(\mathcal{X})$ for $\mathcal{X} \overset{\mathrm{def}}{=} \bar{B}(x^*, \varphi R\zeta_R) \subset \mathcal{M}_{\mathrm{LB}}$. 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 $\mathcal{X}_{\mathcal{H}} \overset{\mathrm{def}}{=} \bar{B}(x^*, \varphi R)$, $f \in \mathcal{F}_L(\mathcal{X})$, then $x_t \in \mathcal{X}_{\mathcal{H}}$.*

This result allows us to fully quantify the convergence rate of RGD, without resorting to assumptions about the distances of the iterates to the optimizers, as shown in the following.

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

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 11.

Interestingly, the general rate for $f \in \mathcal{F}_L(\mathcal{X})$ that we obtain in Proposition 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 & Sra, 2016) and the seemingly curvature-independent rate $O(\frac{LR_{\max}^2}{\varepsilon})$ in (Martínez-Rubio & Pokutta, 2023). This fact highlights the importance of providing iterate bounds to fully quantify convergence rates.

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 or an upper bound thereof, we can reduce the minimization of a function $f \in \mathcal{F}_L(\mathcal{X})$ to minimizing the strongly g-convex function $F(x) \overset{\mathrm{def}}{=} f(x) + \frac{\varepsilon}{2R^2} d(x_0, x)^2$. 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 $\bar{B}(x_0, O(R\zeta_R))$. For the hyperbolic space we obtain $\widetilde{O}(\frac{\hat{L}R^2}{\varepsilon})$ with the smoothness property required only in $B(x_0, O(R))$. We can also quantify the rate in the general case, see Remark 12.

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 they increase with the diameter of this set.

Alternatively, we can use RGD with a step size $\eta = 1/(L\zeta_R)$. As for the reduction described above, an upper bound on R can be used instead of its value. Using this step size, we show that the iterates do not move away from the minimizer more than 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 L and ζ are not necessarily identical as they are taken with respect to sets of different sizes. In fact, for the problem we implement in Section 5, both step sizes coincide up to a small constant.

Theorem 3. [↓] *Consider a manifold $\mathcal{M}_{\mathrm{LB}} \in \mathcal{R}_{\mathrm{LB}}$. Let $f : \mathcal{M} \to \mathbb{R}$ be g-convex in $\mathcal{X} \overset{\mathrm{def}}{=} \bar{B}(x^*, R) \subset \mathcal{M}_{\mathrm{LB}}$ and L-smooth in $\bar{B}(x^*, 2R) \subset \mathcal{M}_{\mathrm{LB}}$. The iterates of RGD with step size $\eta = 1/(\zeta_R L)$ are quasi-nonexpansive. The convergence rate is $O(\zeta_R LR^2/\varepsilon)$ and if f is also μ-strongly g-convex in $\mathcal{X}$, then it takes $O((\zeta_R L/\mu)\log(LR^2/\varepsilon))$ iterations.*

Note that even though RGD with $\eta = 1/(\zeta_R L)$ is quasi-nonexpansive, we assume smoothness with respect to $B(x^*, 2R)$ for technical reasons. We can also extend our techniques to show that for g-convex and Lipschitz functions, the iterates of Riemannian subgradient descent move away from an optimizer by at most a $\sqrt{2}$ factor farther than the initial distance.

Theorem 4 (Non-smooth RGD)**.** [↓] *Consider a manifold $\mathcal{M}_{\mathrm{LB}} \in \mathcal{R}_{\mathrm{LUB}}$ and $f : \mathcal{M}_{\mathrm{LB}} \to \mathbb{R}$ that is L_{p}-*

Lipschitz and g-convex in $\mathcal{X} \overset{\text{def}}{=} \bar{B}(x^*, \sqrt{2}R) \subset \mathcal{M}_{\text{LB}}$. *The iterates of Riemannian subgradient descent with* $\eta \overset{\text{def}}{=} R/(L_{\text{p}}\sqrt{\zeta_{\sqrt{2}R}T})$ *lie in* $\mathcal{X}$ *and the geodesic average of the iterates, cf. (2), is an* ε*-minimizer of* f *after* $O(\zeta_R L_{\text{p}}^2 R^2/\varepsilon^2)$ *iterations.*

Lastly, we present an analysis of a composite RGD (CRGD) algorithm, of independent interest. This algorithm exploits the ability to solve a structured problem for improved convergence. Note that while the update rule of CRGD requires just one call to the gradient oracle, its implementation could be a hard computational problem. The interest of this result is that it can yield better information-theoretical upper bounds on the gradient oracle complexity than other approaches. For instance, for the implementation of proximal subroutines for the optimization of functions in $\mathcal{F}_L(\mathcal{X})$, see Proposition 8.

Proposition 5 (Composite RGD)**.** [↓] *Let* $\mathcal{M} \in \mathcal{R}_{\text{LUB}}$ *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* $F \overset{\text{def}}{=} f + g$ *is* μ*-strongly g-convex in* $\mathcal{X}$, *and* $x^* \overset{\text{def}}{=} \arg\min_{x\in\mathcal{X}} F(x)$. *Iterating the rule*

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

we get 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.

4.2. Riemannian Proximal Methods

We present our results on proximal methods. For Hadamard manifolds, it is known that the prox is a non-expansive operator, cf. Appendix C.2. However, that is not the case in general manifolds (Wang et al., 2023a, Section 6.1). Still, we are able to show that the iterates of RPPA never move farther from an optimizer than the initial distance in general manifolds, which allows us to provide fully quantified rates of convergence of it and its inexact version.

Proposition 6 (RPPA)**.** [↓] *Consider a manifold* $\mathcal{M} \in \mathcal{R}_{\text{LUB}}$ *and a function* $f \in \mathcal{F}(\mathcal{M})$ *with* $\mathcal{X} \overset{\text{def}}{=} \bar{B}(x^*, R) \subset \mathcal{M}$. *For any* $\eta > 0$ *and all* $t \geq 0$, *the iterates of the exact RPPA, cf. (3), 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, they 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. This convergence result is surprising, since RPPA is equivalent to RGD on the Moreau envelope, cf. Lemma 9, which can be non g-convex when positive curvature is present.

Theorem 7 (RIPPA)**.** [↓] *Consider a manifold* $\mathcal{M} \in \mathcal{R}_{\text{LUB}}$ *and a function* $f \in \mathcal{F}(\mathcal{M})$, *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 $\mathcal{M} \in \mathcal{R}_{\text{LUB}}$, 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: $x_{t+1}^* \overset{\text{def}}{=} \text{prox}_{\eta f}(x_t)$.
- Subgradient: $v_{t+1} \in \partial f(x_{t+1})$.
- Error $r_{t+1} \overset{\text{def}}{=} \eta v_{t+1} - \text{Log}_{x_{t+1}}(x_t)$.
- For $\mu = 0$: $\Delta_t \overset{\text{def}}{=} (t+1)^{-2}$. For $\mu > 0$: $\Delta_t \overset{\text{def}}{=} \eta\mu/2$.

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

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

3: **end for**

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

While Theorem 7 does not require smoothness, we can exploit this condition to give an efficient implementation via first-order methods. Note that we can guarantee the condition in Algorithm 1 without knowing x_{t+1}^*, which is what we are trying to approximate.

Proposition 8. [↓] *In the setting of Theorem 7, 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* $\mathcal{X} \overset{\text{def}}{=} \bar{B}(x_t, 2R)$ *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 $\widetilde{O}(\zeta_R^2)$ 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, while implementing each step of CRGD could be expensive. On the other hand PRGD has a worse gradient complexity but can be implemented easily, since projection onto a ball can be done in closed form (Martínez-Rubio & Pokutta, 2023).

We note that the first iteration of CRGD with the squared-distance regularization in Proposition 8 is equivalent to a step of PRGD with a different step-size and thus it can be easily implemented, see Appendix B.1.

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 this envelope. This is a very useful tool in the design of optimization algorithms (Parikh & Boyd, 2014; Davis & Drusvyatskiy, 2019). First, we provide an expression for the gradient of the Moreau envelope.

Lemma 9 (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 $f \in \mathcal{F}(\mathcal{X})$, and the Moreau envelope of $g \stackrel{\text{def}}{=} f + I_{\mathcal{X}}$ with $\eta > 0$, define 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 $\nabla M(x) = -\frac{1}{\eta} \operatorname{Log}_x(\operatorname{prox}_{\eta g}(x))$.

Now, we can provide a bound for the value of the Moreau envelope smoothness.

Theorem 10 (Moreau envelope smoothness)**.** [↓] *Consider $\mathcal{M}_{\text{LB}} \in \mathcal{R}_{\text{LB}}$, and let $\mathcal{X} \subset \mathcal{M}_{\text{LB}}$ be a g-convex closed set. For $f \in \mathcal{F}(\mathcal{M})$, we have that the Moreau envelope of $g \stackrel{\text{def}}{=} f + I_{\mathcal{X}}$ with parameter $\eta > 0$, defined for all $x \in \mathcal{M}_{\text{LB}}$ as $M(x) \stackrel{\text{def}}{=} \min_{z \in \mathcal{M}_{\text{LB}}} \{f(z) + I_{\mathcal{X}}(z) + \frac{1}{2\eta} d(x,z)^2\}$, satisfies for all $x, y \in \mathcal{M}$:*

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

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, in this case the smoothness is not degraded by the curvature with respect to the Euclidean case, while the g-convexity can be lost.

The intuition about the proof of Theorem 10 is the following. The epigraph of the Moreau envelope can be seen as the union of the epigraphs $\{(x, f(y) + I_{\mathcal{X}}(y) + \frac{1}{2\eta} d(x,y)^2) \mid x \in \mathcal{M}_{\text{LB}}\}$ for all $y \in \mathcal{M}_{\text{LB}}$. Consequently, given the $\operatorname{prox}_{\eta f}(x)$, 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 9 that shows $\nabla M(x) = -\frac{1}{\eta} \operatorname{Log}_x(\operatorname{prox}_{\eta g}(x))$, 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 $T_x \mathcal{M}_{\text{LB}}$ whose induced function in $\mathcal{M}_{\text{LB}}$ 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.

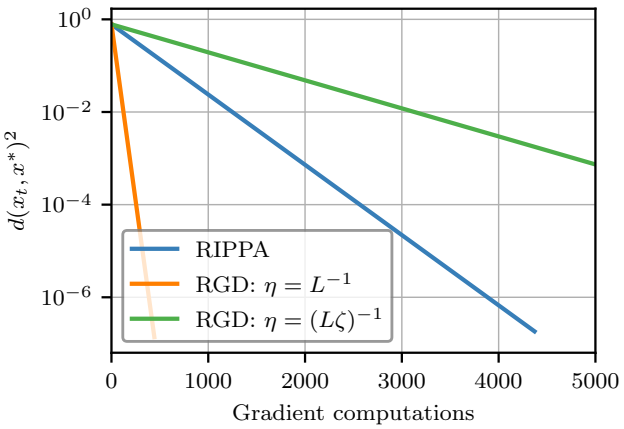

Figure 1: Comparison of RIPPA and of RGD with $\eta = L^{-1}$ and $\eta = (L\zeta_R)$ for solving (4) in terms of squared distance to the optimizer x^* for the hyperbolic space $\mathbb{H}^d$ with $n = 1000$ centers and dimension $d = 1000$. We observe monotonous decrease in distance in all of our experiments.

5. Experiments

We present experimental results for computing the Karcher mean in the d-dimensional hyperbolic space $\mathbb{H}^d$, and the $d(d+1)/2$-dimensional manifold of symmetric positive definite matrices $\mathcal{S}_+^d \stackrel{\text{def}}{=} \{M \in \mathbb{R}^{d \times d} : M = M^T, M \succ 0\}$ with the affine-invariant metric (Hosseini & Sra, 2015). $\mathbb{H}^d$ has constant negative sectional curvature everywhere and we scale it such that the curvature is -1. The sectional curvature of $\mathcal{S}_+^d$ equipped with the affine-invariant metric lies in $[-0.5, 0]$ (Criscitiello & Boumal, 2020, Prop I.1). We implement RGD with step sizes $\eta = 1/L$ and $\eta = 1/(L\zeta_R)$ as well as RIPPA performing a constant number of iterations of PRGD to approximately solve the proximal problems. The first step can be taken as CRGD, as we explained in Proposition 8.

We implement this problem using the Pymanopt library

(Townsend et al., 2016), published under the BSD-3-Clause license. We run until a fixed precision is reached in function value, and because of this, different algorithms stop at points at different distances from x^*. We provide the results in function value and more experiments in Appendix G, where we observe similar behavior. The experiments show that (A) RIPPA is a competitive algorithm for solving g-convex and smooth optimization problems and that (B) the distance of the iterates of RGD and RIPPA to the optimizer x^* monotonically decrease in practice, which goes beyond what our theoretical results predict.

For $\mathcal{M} \in \mathcal{R}_{\text{LUB}}$, the Karcher mean is defined as

$$\min_{x \in \mathcal{M}} \left\{ F(x) \stackrel{\text{def}}{=} \frac{1}{2n} \sum_{i=1}^{n} d^2(x, y_i) \right\}. \tag{4}$$

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 $D \stackrel{\text{def}}{=} \operatorname{diam}(\mathcal{X})$. In this problem, we can analytically compute an upper bound on R, which allows us to choose the step sizes in a principled manner, see Appendix F.

While RGD with $\eta = (\zeta_R L)^{-1}$ is quasi-nonexpansive by Theorem 3, Theorem 1 and Proposition 8 only guarantee that RGD with $\eta = 1/L$ and RIPPA 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 some increase in distance to the optimizer at some point for the latter two algorithms. However, in Figures 1 and 2 and in the results for different parameters in Appendix G, the distance of the iterates to the optimizer is monotonically decreasing for *all* algorithms. In fact, we ran the algorithms 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 in Theorem 1 and Proposition 8 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.

In Figure 2, RIPPA outperforms both RGD variants. The results for RGD with both step sizes are similar, which is be due to the fact that for the Karcher mean the step sizes and the convergence rates coincide up to constant factors on any $\mathcal{M} \in \mathcal{R}_{\text{LUB}}$, see Appendix F. 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. Due to Theorem 1, we have that the iterates of RGD with $\eta = 1/L$ stay in $\bar{B}(x^*, \varphi R)$ in $\mathbb{H}^d$, which allows for larger step sizes than in general manifolds. This means that one would not expect $\eta = 1/(L\zeta_R)$ to provide an advantage in this setting, which is what we observe. We do not have such a refined result for RIPPA in hyperbolic space. This may be why in this special case, we see in Figure 1 that RGD with $\eta = 1/L$ converges significantly faster than RIPPA while RGD with $\eta = 1/(L\zeta_R)$ is the slowest method.

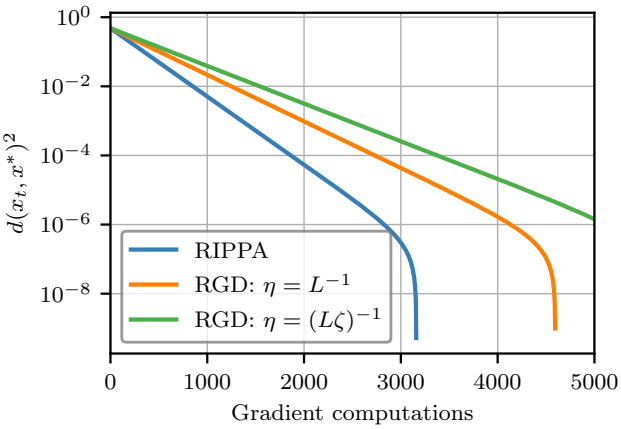

Figure 2: Comparison of RIPPA and of RGD with $\eta = L^{-1}$ and $\eta = (L\zeta_R)^{-1}$ for solving (4) in terms of squared distance to the optimizer x^* for $\mathcal{S}_+^{100}$ with $n = 1000$ centers, and dimension $d(d+1)/2 = 5050$. We observe monotonous decrease in distance in all of our experiments.

6. Conclusion and Discussion

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 and RPPA algorithms by providing full convergence rates without artificial assumptions, and different variants that enjoy different convergence rates, trading off its dependence on geometric constants with some guarantee on iterate boundedness or on efficiency of implementation. Further, we presented 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, and provided new properties of the Riemannian proximal operator.

One presented 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 studying whether the RGD rule is a nonexpansive operator for some choice of the step size. This would have implications to algorithmic stability and differential privacy.

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. Additionally, studying whether we can get an efficiently implementable algorithm that obtains the best of all of our rates and iterate bounds is a very interesting open question.

Acknowledgements

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

Impact statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Ahn, K. and Sra, S. From Nesterov's estimate sequence to Riemannian acceleration. *arXiv preprint arXiv:2001.08876*, 2020. URL https://arxiv.org/abs/2001.08876.

Allen-Zhu, Z., Garg, A., Li, Y., de Oliveira, R. M., and Wigderson, A. Operator scaling via geodesically convex optimization, invariant theory and polynomial identity testing. In Diakonikolas, I., Kempe, D., and Henzinger, M. (eds.), *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2018, Los Angeles, CA, USA, June 25-29, 2018*, pp. 172–181. ACM, 2018. doi: 10.1145/3188745.3188942. URL https://doi.org/10.1145/3188745.3188942.

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

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

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

Bento, G., da Cruz Neto, J. X., and Oliveira, P. R. A new approach to the proximal point method: Convergence on general riemannian manifolds. *J. Optim. Theory Appl.*, 168(3):743–755, 2016a. doi: 10.1007/s10957-015-0861-2. URL https://doi.org/10.1007/s10957-015-0861-2.

Bento, G. C., Ferreira, O. P., and Melo, J. G. Iteration-complexity of gradient, subgradient and proximal point methods on Riemannian manifolds. *arXiv:1609.04869 [math]*, September 2016b. URL http://arxiv.org/abs/1609.04869. arXiv: 1609.04869.

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

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

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

Cambier, L. and Absil, P. Robust low-rank matrix completion by Riemannian optimization. *SIAM J. Scientific Computing*, 38(5), 2016. doi: 10.1137/15M1025153. URL https://doi.org/10.1137/15M1025153.

Carvalho Bento, G. d. and Melo, J. G. Subgradient method for convex feasibility on Riemannian manifolds. *J. Optim. Theory Appl.*, 152(3):773–785, 2012. doi: 10.1007/s10957-011-9921-4. URL https://doi.org/10.1007/s10957-011-9921-4.

Cherian, A. and Sra, S. Riemannian dictionary learning and sparse coding for positive definite matrices. *IEEE Trans. Neural Networks Learn. Syst.*, 28(12):2859–2871, 2017. doi: 10.1109/TNNLS.2016.2601307. URL https://doi.org/10.1109/TNNLS.2016.2601307.

Criscitiello, C. and Boumal, N. An accelerated first-order method for non-convex optimization on manifolds. *CoRR*, abs/2008.02252, 2020. URL https://arxiv.org/abs/2008.02252.

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

Criscitiello, C. and Boumal, N. Curvature and complexity: Better lower bounds for geodesically convex optimization. In Neu, G. and Rosasco, L. (eds.), *The Thirty Sixth Annual Conference on Learning Theory, COLT 2023, 12-15 July 2023, Bangalore, India*, volume 195 of *Proceedings of Machine Learning Research*, pp. 2969–3013. PMLR, 2023. URL https://proceedings.mlr.press/v195/criscitiello23a.html.

Cruz Neto, J. X. d., Ferreira, O. P., Pérez, L. R. L., and Németh, S. Z. Convex- and monotone-transformable mathematical programming problems and a proximal-like point method. *J. Glob. Optim.*, 35(1):53–69, 2006. doi: 10.1007/s10898-005-6741-9. URL https://doi.org/10.1007/s10898-005-6741-9.

Davis, D. and Drusvyatskiy, D. Stochastic model-based minimization of weakly convex functions. *SIAM J. Optim.*, 29(1):207–239, 2019. doi: 10.1137/18M1178244. URL https://doi.org/10.1137/18M1178244.

Edelman, A., Arias, T. A., and Smith, S. T. The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Analysis Applications*, 20(2):303–353, 1998. doi: 10.1137/S0895479895290954. URL https://doi.org/10.1137/S0895479895290954.

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

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

Ferreira, O. P., Louzeiro, M. S., and da Fonseca Prudente, L. Gradient method for optimization on riemannian manifolds with lower bounded curvature. *SIAM J. Optim.*, 29 (4):2517–2541, 2019. doi: 10.1137/18M1180633. URL https://doi.org/10.1137/18M1180633.

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

Heidel, G. and Schulz, V. A Riemannian trust-region method for low-rank tensor completion. *Numerical Lin. Alg. with Applic.*, 25(6), 2018. doi: 10.1002/nla.2175. URL https://doi.org/10.1002/nla.2175.

Hosseini, R. and Sra, S. 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*, pp. 910–918, 2015. URL http://papers.nips.cc/paper/5812-matrix-manifold-optimization-for-gaussian-mixtures.

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

Hu, Z., Wang, G., Wang, X., Wibisono, A., Abernethy, J., and Tao, M. Extragradient type methods for riemannian variational inequality problems. *arXiv preprint arXiv:2309.14155v1*, 2023. URL https://arxiv.org/abs/2309.14155v1.

Jordan, M. I., Lin, T., and Vlatakis-Gkaragkounis, E. First-order algorithms for min-max optimization in geodesic metric spaces. *CoRR*, abs/2206.02041, 2022. doi: 10.48550/arXiv.2206.02041. URL https://doi.org/10.48550/arXiv.2206.02041.

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

Khuzani, M. B. and Li, N. 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*, pp. 133–140, 2017. doi: 10.1109/ICMLA.2017.0-167. URL https://doi.org/10.1109/ICMLA.2017.0-167.

Kim, J. and Yang, I. 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. (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 11255–11282. PMLR, 2022. URL https://proceedings.mlr.press/v162/kim22k.html.

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

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

Lezcano-Casado, M. and Martínez-Rubio, D. 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*, pp. 3794–3803, 2019. URL http://proceedings.mlr.press/v97/lezcano-casado19a.html.

Martínez-Rubio, D. Global Riemannian acceleration in hyperbolic and spherical spaces. *arXiv preprint arXiv:2012.03618*, 2020. URL https://arxiv.org/abs/2012.03618v5.

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

URL https://proceedings.mlr.press/v195/martinez-rubio23a.html.

Martínez-Rubio, D., Roux, C., Criscitiello, C., and Pokutta, S. Accelerated methods for riemannian min-max optimization ensuring bounded geometric penalties. *CoRR*, abs/2305.16186, 2023. doi: 10.48550/arXiv.2305.16186. URL https://doi.org/10.48550/arXiv.2305.16186.

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

Mishra, B. and Sepulchre, R. 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*, pp. 1137–1142, 2014. doi: 10.1109/CDC.2014.7039534. URL https://doi.org/10.1109/CDC.2014.7039534.

Parikh, N. and Boyd, S. P. Proximal algorithms. *Found. Trends Optim.*, 1(3):127–239, 2014. doi: 10.1561/2400000003. URL https://doi.org/10.1561/2400000003.

Petersen, P. *Riemannian geometry*, volume 171. Springer, 2006. ISBN 978-0-387-29403-2.

Sato, H., Kasai, H., and Mishra, B. Riemannian stochastic variance reduced gradient. *CoRR*, abs/1702.05594, 2017. URL http://arxiv.org/abs/1702.05594.

Sato, H., Kasai, H., and Mishra, B. Riemannian stochastic variance reduced gradient algorithm with retraction and vector transport. *SIAM Journal on Optimization*, 29(2):1444–1472, 2019. doi: 10.1137/17M1116787. URL https://doi.org/10.1137/17M1116787.

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

Sun, J., Qu, Q., and Wright, J. Complete dictionary recovery over the sphere II: recovery by Riemannian trust-region method. *IEEE Trans. Inf. Theory*, 63(2):885–914, 2017. doi: 10.1109/TIT.2016.2632149. URL https://doi.org/10.1109/TIT.2016.2632149.

Tan, M., Tsang, I. W., Wang, L., Vandereycken, B., and Pan, S. J. Riemannian pursuit for big matrix recovery. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pp. 1539–1547, 2014. URL http://proceedings.mlr.press/v32/tan14.html.

Tang, G. and Huang, N. Rate of convergence for proximal point algorithms on hadamard manifolds. *Oper. Res. Lett.*, 42(6-7):383–387, 2014. doi: 10.1016/j.orl.2014.06.009. URL https://doi.org/10.1016/j.orl.2014.06.009.

Townsend, J., Koep, N., and Weichwald, S. Pymanopt: A python toolbox for optimization on manifolds using automatic differentiation. *Journal of Machine Learning Research*, 17(137):1–5, 2016. URL http://jmlr.org/papers/v17/16-177.html.

Tripuraneni, N., Flammarion, N., Bach, F., and Jordan, M. I. Averaging stochastic gradient descent on Riemannian manifolds. *CoRR*, abs/1802.09128, 2018. URL http://arxiv.org/abs/1802.09128.

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

Vandereycken, B. Low-rank matrix completion by Riemannian optimization. *SIAM Journal on Optimization*, 23 (2):1214–1236, 2013. doi: 10.1137/110845768. URL https://doi.org/10.1137/110845768.

Wang, X., Tu, Z., Hong, Y., Wu, Y., and Shi, G. Online optimization over riemannian manifolds. *J. Mach. Learn. Res.*, 24:84:1–84:67, 2023a. URL http://jmlr.org/papers/v24/21-1308.html.

Wang, X., Yuan, D., Hong, Y., Hu, Z., Wang, L., and Shi, G. Riemannian optimistic algorithms. *arXiv preprint arXiv:2308.16004v1*, 2023b. URL https://arxiv.org/abs/2308.16004v1.

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

Weber, M. and Sra, S. Nonconvex stochastic optimization on manifolds via Riemannian Frank-Wolfe methods. *CoRR*, abs/1910.04194, 2019. URL http://arxiv.org/abs/1910.04194.

Zhang, H. and Sra, S. First-order Methods for Geodesically Convex Optimization. *arXiv:1602.06053 [cs, math, stat]*, February 2016. URL http://arxiv.org/abs/1602.06053. arXiv: 1602.06053.

Zhang, H. and Sra, S. An estimate sequence for geodesically convex optimization. In Bubeck, S., Perchet, V., and Rigollet, P. (eds.), *Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018*, volume 75 of *Proceedings of Machine Learning Research*, pp. 1703–1723. PMLR, 2018. URL http://proceedings.mlr.press/v75/zhang18a.html.

Zhang, H., Reddi, S. J., and Sra, S. Riemannian SVRG: fast stochastic optimization on Riemannian manifolds. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 4592–4600, 2016. URL http://papers.nips.cc/paper/6515-riemannian-svrg-fast-stochastic-optimization-on-riemannian-manifolds.

Zhang, P., Zhang, J., and Sra, S. Minimax in geodesic metric spaces: Sion's theorem and algorithms. *CoRR*, abs/2202.06950, 2022. URL https://arxiv.org/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 $\mathcal{X} \overset{\text{def}}{=} \bar{B}(x^*, \varphi R\zeta_R) \subset \mathcal{M}_{\mathrm{LB}}$. 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 $\mathcal{X}_{\mathcal{H}} \overset{\text{def}}{=} \bar{B}(x^*, \varphi R)$, $f \in \mathcal{F}_L(\mathcal{X})$, then $x_t \in \mathcal{X}_{\mathcal{H}}$.*

Proof. (Theorem 1) Define $\varphi \overset{\text{def}}{=} (1+\sqrt{5})/2$ and $\zeta \overset{\text{def}}{=} \zeta_{\varphi R\zeta_R}$. 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 $\Delta_i \overset{\text{def}}{=} f(x_i) - f(x^*)$. 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 $\eta_* \overset{\text{def}}{=} \max_{\eta>0}\{\eta \mid \mathrm{Exp}_{x_t}(-\eta\nabla f(x_t)) \in \bar{B}(x^*, \varphi R\zeta_R)\}$. 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 $x'_{t+1} \overset{\text{def}}{=} \mathrm{Log}_{x_t}(-\eta_*\nabla f(x_t))$. By definition, it must be $d(x'_{t+1}, x^*) = \varphi R\zeta_R$. We will arrive to a contradiction. We have for all $i < t$:

$$\Delta_{i+1} - \Delta_i \leq \langle \nabla f(x_i), \mathrm{Log}_{x_i}(x_{i+1})\rangle + \frac{L}{2} d\left(x_{i+1}, x_i\right)^2 = -\frac{1}{2L}\|\nabla f(x_i)\|^2, \tag{5}$$

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 $\Delta'_{t+1} \overset{\text{def}}{=} f(x'_{t+1}) - f(x^*)$, we have:

$$\Delta'_{t+1} - \Delta_t \leq \langle \nabla f(x_t), \mathrm{Log}_{x_t}(x'_{t+1})\rangle + \frac{L}{2} d\left(x'_{t+1}, x_t\right)^2 = \left(-\eta_* + \frac{L\eta_*^2}{2}\right)\|\nabla f(x_t)\|^2 \overset{①}{\leq} -\frac{\eta_*}{2}\|\nabla f(x_t)\|^2, \tag{6}$$

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

$$\begin{aligned}\Delta_i &\overset{①}{\leq} \langle -\nabla f(x_i), \mathrm{Log}_{x_i}(x^*)\rangle \overset{②}{\leq} \frac{L}{2}\left[d(x_i, x^*)^2 - d(x_{i+1}, x^*)^2 + \zeta d(x_i, x_{i+1})^2\right] \\ &= \frac{L}{2}\left[d(x_i, x^*)^2 - d(x_{i+1}, x^*)^2\right] + \frac{\zeta}{2L}\|\nabla f(x_i)\|^2,\end{aligned} \tag{7}$$

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

$$\Delta_t \leq \frac{1}{2\eta_*}\left[d(x_t, x^*)^2 - d(x'_{t+1}, x^*)^2\right] + \frac{\zeta\eta_*}{2}\|\nabla f(x_t)\|^2 \tag{8}$$

Multiplying (5) by ζ and adding it to (7), and similarly with (6) and (8) we obtain:

$$\begin{aligned}\zeta\Delta_{i+1} - (\zeta-1)\Delta_i &\leq \frac{L}{2}\left(d(x_i, x^*)^2 - d(x_{i+1}, x^*)^2\right) \text{ for } i < t \\ \eta_* L(\zeta\Delta'_{t+1} - (\zeta-1)\Delta_t) &< \frac{L}{2}\left(d(x_t, x^*)^2 - d(x'_{t+1}, x^*)^2\right)\end{aligned} \tag{9}$$

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

$$\begin{aligned}\eta_* L\zeta\Delta'_{t+1} + \zeta\Delta_t(1-\eta_* L) + \eta_* L\Delta_t + \sum_{i=1}^{t-1}\Delta_i + \frac{L}{2}d(x'_{t+1}, x^*)^2 &< (\zeta-1)\Delta_0 + \frac{Ld(x_0, x^*)^2}{2} \\ &\overset{①}{\leq} \frac{\zeta LR^2}{2}.\end{aligned} \tag{10}$$

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 $\zeta_r \in [r\sqrt{|\kappa_{\min}|}, r\sqrt{|\kappa_{\min}|}+1]$:

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

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 (10) if we now consider $\eta_* \overset{\text{def}}{=} \max_{\eta>0}\{\eta \mid \operatorname{Exp}_{x_t}(-\eta_* \nabla f(x_t)) \in \bar{B}(x^*, \varphi R)\}$, which yields $d(x'_{t+1}, x^*) = \phi R$, we make use of $\zeta \overset{\text{def}}{=} \varphi\zeta_R$, and use L-smoothness in $\bar{B}(x^*, \varphi R)$. However, we substitute the right hand side of ① in (10) by $\varphi^2 L R^2/2$ which holds since by (Criscitiello & Boumal, 2023, Proposition 13) we have $\Delta_0 \leq \frac{\varphi L R^2}{2\zeta_R}$ 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 & 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 (5) 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.

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

Proof. (Proposition 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 & 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 & 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 & 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 & Boumal, 2021).

We note that for the strongly g-convex case, we could also use the result in (Zhang & 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 11. *We note that RGD with step size $\eta < 2/L$ does not increase the function value. Indeed, assume smoothness holds between x_t and $x_{t+1} \overset{\text{def}}{=} \operatorname{Exp}_{x_t}(-\eta\nabla f(x_t))$. We have:*

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

where ① is equivalent to $\eta < 2/L$. This means that if $\mathcal{Y} \overset{\text{def}}{=} \{y \in \mathcal{M} \mid f(y) \leq f(x_0)\}$ 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 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 $\widetilde{O}(L/\mu)$. 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 12. *We note that if we know a valid upper bound R on the initial distance, we can instead minimize $F(x) \overset{\text{def}}{=} f(x) + r(x)$ where $r(x) \overset{\text{def}}{=} \frac{\varepsilon}{2R^2} d(x_0, x)^2$. 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 $\hat{x}^* \overset{\text{def}}{=} \arg\min_x r(x)$. 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 $\mathcal{X} \overset{\text{def}}{=} \bar{B}(\hat{x}^*, \varphi R \zeta_R) \subseteq \bar{B}(x_0, \varphi R \zeta_R + R)$ 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 $\eta \overset{\text{def}}{=} (\hat{L} + \zeta_{\varphi R\zeta_R + R})^{-1}$ stay in $\mathcal{X}$ and thus by Proposition 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 reduce the logarithmic factor if we use the reduction in (Martínez-Rubio, 2020).

In the hyperbolic space $\mathbb{H}$ of curvature -1, we make use of (Criscitiello & Boumal, 2023, Proposition 13), which says that for a differentiable g-convex L-smooth function $f : \mathbb{H} \to \mathbb{R}$ in $\bar{B}(x_0, R)$ with a global minimizer x^* in that ball, it is $f(x_0) - f(x^*) \leq 4Ld(x_0, x^*)^2/\zeta_{d(x_0,x^*)}$. With this proposition, we conclude that without loss of generality, we work with $\varepsilon \leq LR^2/\zeta_R$. On the other hand, using Theorem 1 and the reasoning above, we conclude that with $\hat{L}$ being the smoothness constant of f in $\mathcal{X} \overset{\text{def}}{=} \bar{B}(\hat{x}^*, \varphi R)$, we obtain a convergence rate of $\widetilde{O}(\zeta_R + \frac{\hat{L}R^2}{\varepsilon}) = \widetilde{O}(\frac{\hat{L}R^2}{\varepsilon})$. The last expression holds by our remark on the value of ε.

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 $f : \mathcal{M} \to \mathbb{R}$ be g-convex in $\mathcal{X} \overset{\text{def}}{=} \bar{B}(x^*, R) \subset \mathcal{M}_{\text{LB}}$ and L-smooth in $\bar{B}(x^*, 2R) \subset \mathcal{M}_{\text{LB}}$. The iterates of RGD with step size $\eta = 1/(\zeta_R L)$ are quasi-nonexpansive. The convergence rate is $O(\zeta_R LR^2/\varepsilon)$ and 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) We show $d(x_{t+1}, x^*) \leq R_t \overset{\text{def}}{=} d(x_t, x^*)$, for any minimizer x^* and by induction, for any $R_t \leq R$. By the cosine inequality Lemma 19, and the monotonicity of $r \mapsto \zeta_r$, we have

$$\begin{aligned} d(x_{t+1}, x^*)^2 &\leq d(x_t, x^*)^2 + \zeta_R d(x_t, x_{t+1})^2 + 2\langle \eta \nabla f(x_t), \log_{x_t}(x^*) \rangle \\ &\overset{①}{\leq} d(x_t, x^*)^2 + \zeta_R \eta^2 \|\nabla f(x_t)\|^2 - 2\Delta_t \eta \\ &\overset{②}{\leq} d(x_t, x^*)^2 + 2\eta\Delta_t(L\zeta_R \eta - 1) \\ &\overset{③}{\leq} d(x_t, x^*)^2 \end{aligned} \tag{11}$$

where $\Delta_t = f(x_t) - f(x^*)$. ① holds by g-convexity, ② follows by $\|\nabla f(x_t)\|^2 \leq 2L\Delta_t$, which holds by the following

argument: Let $\hat{x}_{t+1} \leftarrow \exp_{x_t}(-\frac{1}{L}\nabla f(x_t))$, then

$$\begin{aligned}\Delta_t = f(x_t) - f(x^*) &\geq f(x_t) - f(\hat{x}_{t+1}) \\ &\geq \langle -\nabla f(x_t), \log_{x_t}(\hat{x}_{t+1})\rangle - \frac{L}{2} d(x_t, \hat{x}_{t+1})^2 \\ &\geq \frac{1}{2L}\|\nabla f(x_t)\|^2.\end{aligned}$$

Note that we chose a different step size than the one used in the algorithm in this argument. We used smoothness between x_t and $\hat{x}_{t+1}$. By our assumption that f is L-smooth in $B(x^*, 2R)$, it suffices to show that $\hat{x}_{t+1} \in B(x^*, 2R)$. We have that $d(x_t, \hat{x}_{t+1}) = \frac{1}{L}\|\nabla f(x_t) - \nabla f(x^*)\| \leq d(x_t, x^*)$. Hence $d(\hat{x}_{t+1}, x^*) \leq d(x_t, x^*) + d(x_t, \hat{x}_{t+1}) \leq 2R$.

Further, ③ holds by choosing $\eta = 1/(\zeta_R L)$. This shows our desideratum $d(x_{t+1}, x^*) \leq d(x_t, x^*)$, which means that RGD with $\eta = \frac{1}{\zeta_R L}$ is quasi-nonexpansive. The proof convergence follows in analogy to Proposition 2. We provide the proof in Corollary 13 for the sake of completeness. □

Corollary 13. *Under the assumption of Theorem 3, the convergence rate of RGD is $O(\zeta_R L R^2/\varepsilon)$ and 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. In Theorem 3 we have shown that RGD with $\eta = 1/(\zeta_R L)$ is quasi-nonexpansive. In the following, we show the resulting convergence rates. By L-smoothness of f in $\mathcal{X}$, we have

$$\Delta_{t+1} - \Delta_t \leq \langle \nabla f(x_t), \mathrm{Log}_{x_t}(x_{t+1})\rangle + \frac{L}{2} d\left(x_{t+1}, x_t\right)^2 = \left(-\frac{1}{\zeta_R L} + \frac{1}{2\zeta_R^2 L}\right)\|\nabla f(x_t)\|^2 = -\frac{2\zeta_R - 1}{2\zeta_R^2 L}\|\nabla f(x_t)\|^2. \tag{12}$$

By the g-convexity of f, the cosine inequality Lemma 19 and $\zeta_{d(x_t, x^*)} \leq \zeta_R$ which holds by the non-expansivity of RGD and the monotonicity of $r \mapsto \zeta_r$ we obtain ① below

$$\begin{aligned}\Delta_t \leq \left\langle -\nabla f(x_t), \mathrm{Log}_{x_t}(x^*)\right\rangle &\overset{①}{\leq} \frac{1}{2\eta}\left[d(x_t, x^*)^2 - d(x_{t+1}, x^*)^2 + \zeta_R d(x_t, x_{t+1})^2\right] \\ &\overset{②}{=} \frac{L\zeta_R}{2}\left[d(x_t, x^*)^2 - d(x_{t+1}, x^*)^2\right] + \frac{1}{2L}\|\nabla f(x_t)\|^2,\end{aligned} \tag{13}$$

where ② follows by the definition of x_{t+1} and η. Multiplying (12) by $C \overset{\text{def}}{=} \zeta_R^2/(2\zeta_R - 1)$ and adding it to (13), we have

$$C\Delta_{t+1} - (C-1)\Delta_t \leq \frac{L\zeta_R}{2}\left(d(x_t, x^*)^2 - d(x_{t+1}, x^*)^2\right). \tag{14}$$

Now summing (14) from $i = 0$ to $T-1$, we obtain ② below:

$$\begin{aligned}T\Delta_T &\overset{①}{\leq} C\Delta_T + \sum_{t=1}^{T-1}\Delta_t \\ &\overset{②}{\leq} (C-1)\Delta_0 + \frac{L\zeta_R d(x_0, x^*)^2}{2} - \frac{L\zeta_R d(x_T, x^*)^2}{2} \\ &\overset{③}{\leq} (C - 1 + \zeta_R)\frac{LR^2}{2} - \frac{L\zeta_R d(x_T, x^*)^2}{2} \\ &\overset{④}{\leq} \frac{3\zeta_R}{2}\cdot\frac{LR^2}{2} - \frac{L\zeta_R d(x_T, x^*)^2}{2}\end{aligned} \tag{15}$$

where ① holds since $\Delta_t \geq 0$, $C \geq 1$ 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 $C - 1 + \zeta_R$ is increasing for $\zeta_R \in [1, \infty)$ and the limit at $+\infty$ is $3\zeta_R/2$.

Now, dividing (15) by T and dropping the negative terms, we have $\Delta_T \leq \frac{3\zeta_R L R^2}{4T}$. Thus, we have that $\Delta_T \leq \varepsilon$ for $T \geq \frac{3\zeta_R L D^2}{4\varepsilon}$. We now prove the result for the μ-strongly g-convex case. The algorithm is the same, and thus the iterates stay in $\bar{B}(x^*, R)$. The guarantee we just showed for the g-convex case implies ② below:

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

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

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

□

Theorem 4 (Non-smooth RGD). [↓] *Consider a manifold $\mathcal{M}_{\text{LB}} \in \mathcal{R}_{\text{LUB}}$ and $f : \mathcal{M}_{\text{LB}} \to \mathbb{R}$ that is L_{p}-Lipschitz and g-convex in $\mathcal{X} \overset{\text{def}}{=} \bar{B}(x^*, \sqrt{2}R) \subset \mathcal{M}_{\text{LB}}$. The iterates of Riemannian subgradient descent with $\eta \overset{\text{def}}{=} R/(L_{\text{p}}\sqrt{\zeta_{\sqrt{2}R} T})$ lie in $\mathcal{X}$ and the geodesic average of the iterates, cf. (2), is an ε-minimizer of f after $O(\zeta_R L_{\text{p}}^2 R^2/\varepsilon^2)$ iterations.*

Proof. (Theorem 4) Denote by $v_i \in \partial f(x_i)$ the subgradients obtained and used by the algorithm. By the Lipschitzness assumption, it is $\|v_i\| \leq L_{\text{p}}$. 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 $i \leq t$

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

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 $\eta = \frac{R}{L_{\text{p}}\sqrt{\zeta_R T}}$ yields

$$0 \leq 2\eta \sum_{i=0}^{t} [f(x_i) - f(x^*)] \leq R^2 - d(x_{t+1}, x^*)^2 + \zeta_R (t+1) L_{\text{p}}^2 \eta^2 \leq -d(x_{t+1}, x^*)^2 + 2R^2. \tag{16}$$

This proves the induction statement, since $d(x_{t+1}, x^*) \leq \sqrt{2}R$. From (16) with $t \leftarrow T - 1$ and dropping the negative distance term, we obtain

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

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. □

B. RPPA proofs

Proposition 6 (RPPA). [↓] *Consider a manifold $\mathcal{M} \in \mathcal{R}_{\text{LUB}}$ and a function $f \in \mathcal{F}(\mathcal{M})$ with $\mathcal{X} \overset{\text{def}}{=} \bar{B}(x^*, R) \subset \mathcal{M}$. For any $\eta > 0$ and all $t \geq 0$, the iterates of the exact RPPA, cf. (3), satisfy $d(x_{t+1}, x^*) \leq d(x_t, x^*)$. In particular, it is $x_t \in \mathcal{X}$.*

Proof. (Proposition 6) 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

$$\begin{aligned} 0 \leq f(x_{t+1}) - f(x^*) &\overset{①}{\leq} \frac{1}{\eta}\langle -\mathrm{Log}_{x_{t+1}}(x_t), \mathrm{Log}_{x_{t+1}}(x^*)\rangle \\ &\overset{②}{\leq} -\frac{\delta_{2R}}{2} d(x_{t+1}, x_t)^2 - \frac{1}{2} d(x_{t+1}, x^*)^2 + \frac{1}{2} d(x_t, x^*)^2 \\ &\overset{③}{\leq} -\frac{1}{2} d(x_{t+1}, x^*)^2 + \frac{1}{2} d(x_t, x^*)^2, \end{aligned} \tag{17}$$

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. □

Theorem 7 (RIPPA). [↓] *Consider a manifold $\mathcal{M} \in \mathcal{R}_{\mathrm{LUB}}$ and a function $f \in \mathcal{F}(\mathcal{M})$, 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 7) 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

$$d(x_{t+1}, x^*) \leq d(x_{t+1}, x^*_{t+1}) + d(x^*, x^*_{t+1}) \overset{①}{\leq} \frac{1}{2} d(x_t, x^*_{t+1}) + d(x_t, x^*) \overset{②}{\leq} \frac{3}{2} d(x_t, x^*) \overset{③}{\leq} 3R. \tag{18}$$

where in ① we used the criterion in Line 2 and that by Proposition 6 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 $\mathrm{diam}(\triangle x_{t+1} x_t x^*) \leq d(x_{t+1}, x^*) + d(x_t, x^*) \leq 5R$. 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

$$\begin{aligned} &0 \leq f(x_{t+1}) - f(x^*) \\ &\leq -\langle v_{t+1}, \mathrm{Log}_{x_{t+1}}(x^*)\rangle - \frac{\mu}{2} d(x_{t+1}, x^*)^2 \\ &= \frac{1}{\eta}\langle -\mathrm{Log}_{x_{t+1}}(x_t) - r_{t+1}, \mathrm{Log}_{x_{t+1}}(x^*)\rangle - \frac{\mu}{2} d(x_{t+1}, x^*)^2 \\ &\overset{①}{\leq} \frac{1}{\eta}\left(-\frac{\delta_{5R}}{2} d(x_{t+1}, x_t)^2 - \frac{1}{2} d(x_{t+1}, x^*)^2 + \frac{1}{2} d(x_t, x^*)^2\right) + \frac{1}{\eta}\left(\frac{\|r_{t+1}\|^2}{2\Delta_{t+1}} + \frac{\Delta_{t+1}}{2} d(x_{t+1}, x^*)^2\right) - \frac{\mu}{2} d(x_{t+1}, x^*)^2 \\ &\overset{②}{\leq} -\frac{1}{2\eta} d(x_{t+1}, x^*)^2 + \frac{1}{2\eta} d(x_t, x^*)^2 + \frac{\Delta_{t+1}}{2\eta} d(x_{t+1}, x^*)^2 - \frac{\mu}{2} d(x_{t+1}, x^*)^2 \\ &= \frac{1}{2\eta}\left((\Delta_{t+1} - 1 - \eta\mu) d(x_{t+1}, x^*)^2 + d(x_t, x^*)^2\right). \end{aligned} \tag{19}$$

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 (19) with $\mu = 0$ and $\Delta_{t+1} = (t + 2)^{-2}$, we have that

$$d(x_{t+1}, x^*)^2 \leq (1 - \Delta_{t+1})^{-1} d(x_t, x^*)^2 \leq \prod_{i=0}^{t} \frac{1}{1 - \Delta_{i+1}} d(x_0, x^*)^2 \overset{①}{\leq} 2R^2, \tag{20}$$

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 (19) 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 (20). 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, (19) implies

$$d(x_{t+1}, x^*)^2 \leq \frac{d(x_t, x^*)^2}{1 + \eta\mu/2} \leq \ldots \leq (1 + \eta\mu/2)^{-(t+1)} R^2, \tag{21}$$

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 (21) 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. □

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 $F \overset{\text{def}}{=} f + g$ is μ-strongly g-convex in $\mathcal{X}$, and $x^* \overset{\text{def}}{=} \arg\min_{x\in\mathcal{X}} F(x)$. Iterating the rule*

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

we get 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 $\mathcal{Y} \overset{\text{def}}{=} \mathcal{X} \cap \mathrm{dom}(g)$ 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

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

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

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

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 7, suppose that in addition $\bar{B}(x^*,3R) \subset \mathcal{M}$, f is L-smooth in $\bar{B}(x^*,2R)$ and let*

$$C_t \overset{\text{def}}{=} \min\left\{1/4, \frac{\Delta_{t+1}\delta_{3R}}{2(\eta L + \zeta_{3R})^2 + 2\Delta_{t+1}\delta_{3R}}\right\}. \tag{22}$$

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 7 we have $d(x_t,x^*) \leq 2R$. Now let $h_t(z) \overset{\text{def}}{=} f(z) + \frac{1}{2\eta}d(z,x_t)^2$ be the proximal function at step t which is thus smooth in $\triangle x_{t+1}x_t x^*_{t+1}$ with constant $\bar{L} \overset{\text{def}}{=} L + \zeta_{3R}/\eta$, 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

$$\|r_{t+1}\|^2 = \eta^2\|\nabla h_t(x_{t+1})\|^2 \overset{①}{\leq} \bar{L}^2\eta^2 d(x_{t+1}, x^*_{t+1})^2 \overset{②}{\leq} \frac{2C_t}{1-2C_t}(\eta L + \zeta_{3R})^2 d(x_t, x_{t+1})^2 \tag{23}$$

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 (23) it is enough that

$$C_t \leq \frac{\Delta_{t+1}\delta_{3R}}{2(\eta L + \bar{\zeta})^2 + 2\Delta_{t+1}\delta_{3R}}, \tag{24}$$

as specified in (22). 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 8. [↓] *In the setting of Theorem 7, 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 $\widetilde{O}(\zeta_R^2)$ iterations.*

Proof. (Proposition 8) We show that we can implement the subroutine at iteration t, starting from x_t, assuming that it was successfully implemented in previous iterations and thus according to Theorem 7, we have $x_t \in \bar{B}(x^*, 2R)$. The exact optimizer of the prox x_{t+1}^* satisfies, by Lemma 27, that $d(x_t, x_t^*) \leq d(x_t, x^*) \leq 2R$. Thus $x_{t+1}^* \in \mathcal{X} \stackrel{\text{def}}{=} \bar{B}(x_t, 2R) \subset \bar{B}(x^*, 4R)$ and by assumption f is L-smooth in $\mathcal{X}$.

We now use the composite Riemannian Gradient Descent in Proposition 5 with the L-smooth function f of the statement the set $\mathcal{X}$ defined above 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 $\mu \stackrel{\text{def}}{=} L/\delta_{2R}$, cf. Lemma 21. If we use $T \geq \frac{L}{4\mu}\log(\frac{L(1+\zeta_{2R})}{\mu C}) = \widetilde{O}(\frac{1}{\delta_{2R}})$ iterations on $F = f + g$, we obtain

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

Above, ① holds by μ-strong g-convexity, ② holds by the convergence guarantees in Proposition 5. The initial gap ③ holds by the fact that $\nabla F(x_{t+1}^*) = 0$ and by $L(1+\zeta_{2R})$-smoothness of F, since the smoothness of g is $L\zeta_{2R}$ by Lemma 21. 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 $O(\zeta_{\text{diam}(\mathcal{X})}\frac{L}{\mu})$. Thus, taking into account that in Hadamard $\delta_r = 1$ for all $r \geq 0$ and using PRGD on F which is smooth with constant $\hat{L} = \zeta_R L$, and the exact same argument as above except for the new joint smoothness constant and except for ③ and ④ in which we use these other convergence rates and $T = \widetilde{O}(\zeta_{2R}\frac{L}{\mu}) = \widetilde{O}(\zeta_R^2)$, 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, $\frac{1}{\delta_{2R}}$, so the logarithm is benign.

□

Note that the for the optimization required in the first iterate of CRGD for Proposition 8 is the optimization of a quadratic in $T_{x_t}\mathcal{M}$ with a ball constraint and therefore it can be easily implemented. It is in fact, equivalent to the first step of PRGD with possibly a different step-size.

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

Proof. (Theorem 10) Recall that we define $\mathrm{prox}_{\eta f}(x) \overset{\text{def}}{=} \arg\min_{z \in \mathcal{M}_{\mathrm{LB}}} \{f(z) + I_{\mathcal{X}}(z) + \frac{1}{2\eta} d(x,z)^2\} \in \mathcal{X}$. The result is derived from the following.

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

Above, we just substituted in ① the variable in the min by $\mathrm{prox}_{\eta f}(x)$ yielding a possibly greater value. In ②, we used the definition of $M(x)$ and of $\mathrm{prox}_{\eta f}(x)$ and we used the cosine inequality Remark 20 in ③. In ④, we used Lemma 9. Since $\mathrm{prox}_{\eta f}(x) \in \mathcal{X}$, then if $x \in \mathcal{X}$ we have $d(x, \mathrm{prox}_{\eta f}(x)) \leq D$ and so given the inequality above, we have that $M(x)$ is indeed (ζ_D/η)-smooth in $\mathcal{X}$. □

We now include alternative less general proofs of the fact proven in Theorem 10. 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). One can find a review of this proofs in the book (Bacák, 2014, Theorem 2.2.22). We still include our proof since it is very different and arguably simpler.

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 $x, y \in \mathcal{H}$, $x^+ \overset{\text{def}}{=} \mathrm{prox}_f(x)$, $y^+ \overset{\text{def}}{=} \mathrm{prox}_f(y)$. Then*

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

Proof. Let $h_p(x) \overset{\text{def}}{=} f(x) + \frac{1}{2\eta} d(x,p)^2$. 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 $\partial f(\cdot) + \frac{1}{\eta}\mathrm{Log}_y(\cdot) = \partial h_y(\cdot)$ and similarly for h_x. We choose a subgradient $g^f_{y^+} \in \partial f(y^+)$ and define subgradients $g^{h_x}_{x^+} \in \partial h_x(x^+)$, $g^{h_y}_{y^+} \overset{\text{def}}{=} g^f_{y^+} + \frac{1}{\eta}\mathrm{Log}_{y^+}(y) \in \partial h_y(y^+)$ and $g^{h_x}_{y^+} \overset{\text{def}}{=} g^f_{y^+} + \frac{1}{\eta}\mathrm{Log}_{y^+}(x) \in \partial h_x(y^+)$ so that

$$g^{h_x}_{y^+} - g^{h_y}_{y^+} = \frac{1}{\eta}\mathrm{Log}_{y^+}(x) - \frac{1}{\eta}\mathrm{Log}_{y^+}(y). \tag{25}$$

By Lemma 23, we have:

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

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

$$0 \leq \langle g_{y^+}^{h_y} - \Gamma_{x^+}^{y^+} g_{x^+}^{h_x}, \mathrm{Log}_{y^+}(x^+)\rangle. \tag{26}$$

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

$$\begin{aligned}\frac{1}{2\eta}d(x^+,y^+)^2 + \frac{1}{2\eta}d(x^+,y^+)^2 &\leq \left(h_x(x^+) - h_x(y^+) - \langle g_{y^+}^{h_x}, \mathrm{Log}_{y^+}(x^+)\rangle_{y^+}\right) \\ &\quad + \left(h_x(y^+) - h_x(x^+) - \langle g_{x^+}^{h_x}, \mathrm{Log}_{x^+}(y^+)\rangle_{x^+}\right) \\ &= \langle \Gamma_{x^+}^{y^+} g_{x^+}^{h_x} - g_{y^+}^{h_x}, \mathrm{Log}_{y^+}(x^+)\rangle.\end{aligned} \tag{27}$$

Summing up (26) and (27), we get ① below

$$\frac{1}{\eta}d(x^+,y^+)^2 \overset{①}{\leq} \langle g_{y^+}^{h_y} - g_{y^+}^{h_x}, \mathrm{Log}_{y^+}(x^+)\rangle \leq \|g_{y^+}^{h_y} - g_{y^+}^{h_x}\| d(x^+,y^+)$$
$$\overset{②}{=} \frac{1}{\eta}\|\mathrm{Log}_{y^+}(x) - \mathrm{Log}_{y^+}(y)\| d(x^+,y^+) \overset{③}{\leq} \frac{1}{\eta}d(x,y)d(x^+,y^+).$$

Therefore $d(x^+,y^+) \leq d(x,y)$. We used (25) in ②. In ③ we use Lemma 22 which holds for Hadamard manifolds. □

Now we can prove a slightly worse result for 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 $M(x) \overset{\text{def}}{=} \min_{z\in\mathcal{H}}\{f(z) + I_{\mathcal{X}}(z) + \frac{1}{2\eta}d(x,z)^2\}$ is Lipschitz with constant $L \overset{\text{def}}{=} \frac{1+\zeta}{\eta} = O(\zeta/\eta)$, i.e.,*

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

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

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

where ① holds by Lemma 9 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 $\mathcal{X} \overset{\text{def}}{=} \{x \mid f(x) \leq f(p)\}$. Let $\zeta_D \overset{\text{def}}{=} \zeta_{\mathrm{diam}(\mathcal{X})}$. The Riemannian Moreau envelope $M(y) \overset{\text{def}}{=} \min_{x\in\mathcal{M}}\{f(x) + \frac{1}{2\eta}d(x,y)^2\}$ 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

$$\frac{\mathrm{d}y^*}{\mathrm{d}y} = \left(\eta \nabla^2 f(y^*) - \frac{\partial \operatorname{Log}_{y^*}(y)}{\partial y^*} \right)^{-1} \frac{\partial \operatorname{Log}_{y^*}(y)}{\partial y}. \tag{28}$$

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 9, we obtain $\nabla M(y) = -\eta^{-1} \operatorname{Log}_y(y^*)$. We differentiate again with respect to y and use (28) 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 semi-definite. 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 semi-definite 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

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

□

E. Geometric Auxiliary Results

In this section, we provide already established useful geometric results that we use in our proofs. Note that as we mentioned in the preliminaries, we may need to restrict the size of our set if positive curvature is present. Recall for instance that a g-convex function $f : \mathcal{M} \to R$ defined over a compact manifold $\mathcal{M}$ must be constant (Boumal, 2023, Corollary 11.10).

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 & 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 & 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 & 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

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

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 $F(t) \stackrel{\text{def}}{=} f(\gamma(t))$, 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*

$$\bar{x}_1 \leftarrow x_1, \quad t \in \{1, \ldots, T-1\} : \ \bar{x}_{t+1} \leftarrow \mathrm{Exp}_{\bar{x}_t}\left(\frac{w_{t+1}}{\sum_{j=1}^{t+1} w_j}\mathrm{Log}_{\bar{x}_t}(x_{t+1})\right) \tag{29}$$

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

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., $f(\bar{x}_{T-1}) \leq \frac{1}{\sum_{t=1}^{T-1} w_t} \sum_{t=1}^{T-1} w_t f(x_t)$. 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,

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

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*

$$\bar{x}_1 \leftarrow x_1, \quad t \in \{1, \ldots, T-1\} : \ \bar{x}_{t+1} \leftarrow \mathrm{Exp}_{\bar{x}_t}\left(\frac{1}{t+1}\mathrm{Log}_{\bar{x}_t}(x_{t+1})\right) \tag{30}$$

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

The following lemma was proven in (Martínez-Rubio & 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 closed g-convex set $\mathcal{X} \subset \mathcal{M}$. Let $x^* \in \arg\min_{x \in \mathcal{X}} f(x)$ and let $x^+ \overset{\mathrm{def}}{=} \mathrm{prox}_{\eta f + I_{\mathcal{X}}}(x)$ for some $x \in \mathcal{M}$. Then $d(x, x^+)^2 \leq d(x, x^*)^2$.*

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

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

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

Proposition 28. *The optimizer x^* of (4) 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 $d(z, \bar{x}^*) \leq d(z, x^*)$ for all $z \in \mathcal{X}$. By definition, $y_i \in \mathcal{X}$ for all i, hence

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

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 9 about the gradient of the Moreau envelope.

Lemma 9 (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 $f \in \mathcal{F}(\mathcal{X})$, and the Moreau envelope of $g \stackrel{\text{def}}{=} f + I_{\mathcal{X}}$ with $\eta > 0$, define 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 $\nabla M(x) = -\frac{1}{\eta}\text{Log}_x(\text{prox}_{\eta g}(x))$.

Proof. (Lemma 9) 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 $\mathcal{Y} = \{y \in \mathcal{X} : \langle v, \text{Log}_{\hat{x}}(y)\rangle + \frac{1}{2\eta} d(\hat{x}, y)^2 \le f(\hat{x})\}$ for a $v \in \partial f(\hat{x})$, then $\text{Log}_{\hat{x}}(\mathcal{Y}) \subseteq T_{\hat{x}}\mathcal{M}$ 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 $F(\hat{x}, y) \ge \langle v, \text{Log}_{\hat{x}}(y)\rangle + \frac{1}{2\eta} d(\hat{x}, y)^2 > f(\hat{x}) = F(\hat{x}, \hat{x})$ 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 $\text{prox}_{\eta g}(\hat{x})$. Thus, $M(\cdot)$ is differentiable at $\hat{x}$ and $\nabla M(\hat{x}) = \nabla_x F(\hat{x}, \text{prox}_{\eta g}(\hat{x})) = -\frac{1}{\eta}\text{Log}_{\hat{x}}(\text{prox}_{\eta g}(\hat{x}))$, as desired. □

F. Experiment Details

Computing the step sizes. 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. Recall that $\mathcal{X}$ was defined as a g-convex set containing all of the Karcher mean centers y_i. We generate these centers y_i as follows: First, we define an anchor point $\bar{x} \in \mathcal{M}$, then we sample a d-dimensional vector $v_i \in \bar{B}(0, r) \subset T_{\bar{x}}$ uniformly, for some fixed radius r. Then, we divide all but one of the randomly generated v_i by a factor of 10 and compute $y_i \leftarrow \text{Exp}_{\bar{x}}(v_i)$. This has an impact on the condition number and makes the problem harder. This procedure 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 $R \le 2\max_{i\in\{1,\dots,n\}} d(y_i, \bar{x})$.

In order to compute an upper bound on the smoothness of F, we further need a lower bound on $\kappa_{\min}$. Both manifolds have non-positive sectional curvature. In particular $\mathbb{H}^d$ has sectional curvature -1 everywhere. Further, using the so-called affine-invariant metric,

$$\langle X, Y\rangle_P = \text{tr}(P^{-1}XP^{-1}Y) \text{ for } P \in \mathcal{S}_+^d \text{ and } X, Y \in T_P\mathcal{S}_+^d,$$

the sectional curvature of $\mathcal{S}_+^d$ lies in $[-0.5, 0]$ (Criscitiello & Boumal, 2020, Prop I.1).

Step sizes for the Karcher Mean When using RGD with $\eta = 1/(L\zeta_R)$ to solve (4), we can ensure that the iterates stay in $\bar{B}(x^*, r)$. Hence F is ζ_{2r}-smooth (note that we need to assume smoothness in $B(x^*, 2R)$ for technical reasons in Theorem 3), 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 $\widetilde{O}\left(\zeta_R^2\right)$. For RIPPA, in the first step of the subroutine we minimize the quadratic upper bound given by smoothness plus the regularizer in the tangent space $T_{x_t}\mathcal{M}$. For the rest of the steps, these are in different tangent spaces so we use a regular gradient step whose step size is given by the smoothness that is estimated as above. We performed 3 iterations in each subroutine.

G. Numerical Results

We present numerical results for the Karcher mean on $\mathbb{H}^d$ and $\mathcal{S}_+^d$ for different values of n and d.

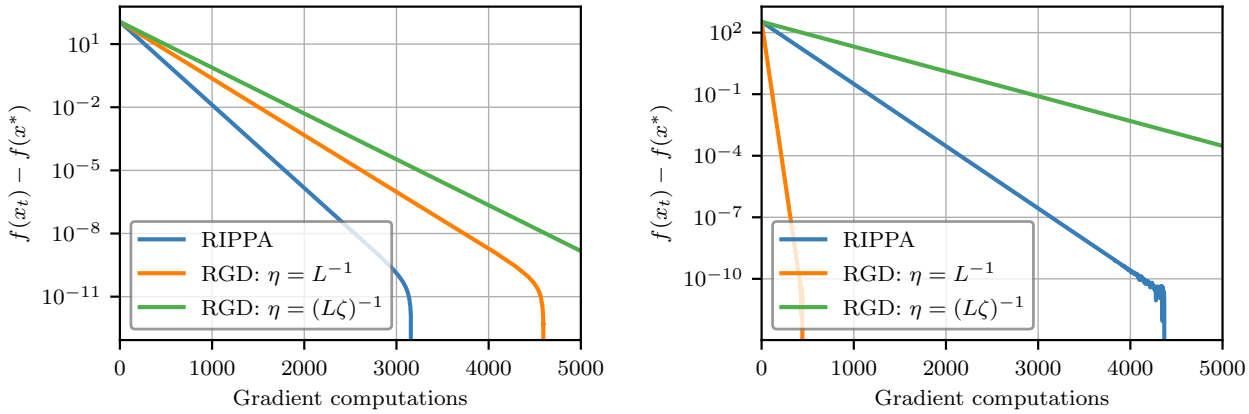

Figure 3: Corresponding plots for Figures 1 and 2 comparing RIPPA and of RGD with $\eta = L^{-1}$ and $\eta = (L\zeta_R)$ for solving (4) in terms of the primal gap. Left: $\mathbb{H}^d$ with $n = 1000$ centers and dimension $d = 1000$. Right: $\mathcal{S}_+^{100}$ with $n = 1000$ centers, and dimension $d(d+1)/2 = 5050$

.

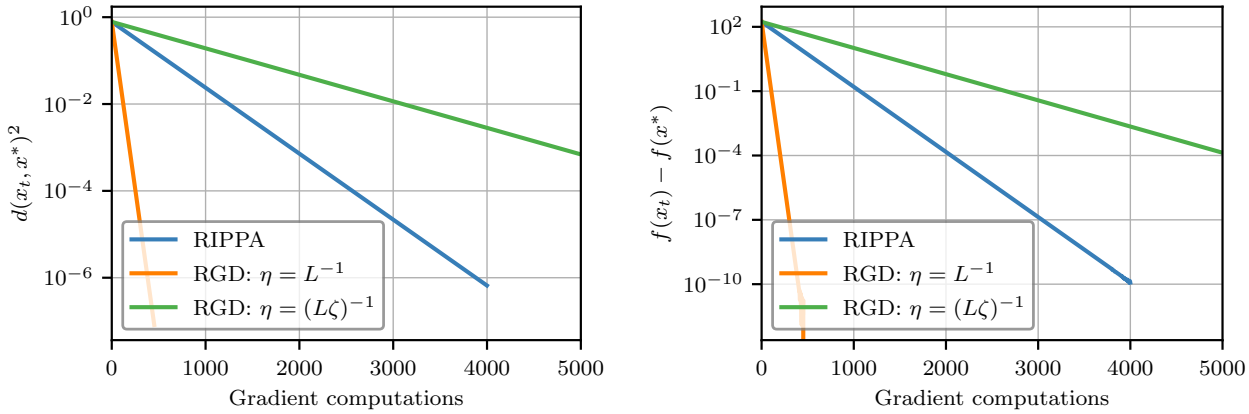

Figure 4: Karcher Mean on $\mathbb{H}^d$: $d = 500$, $n = 1000$, error in squared distance to the optimizer (left) and primal gap (right).

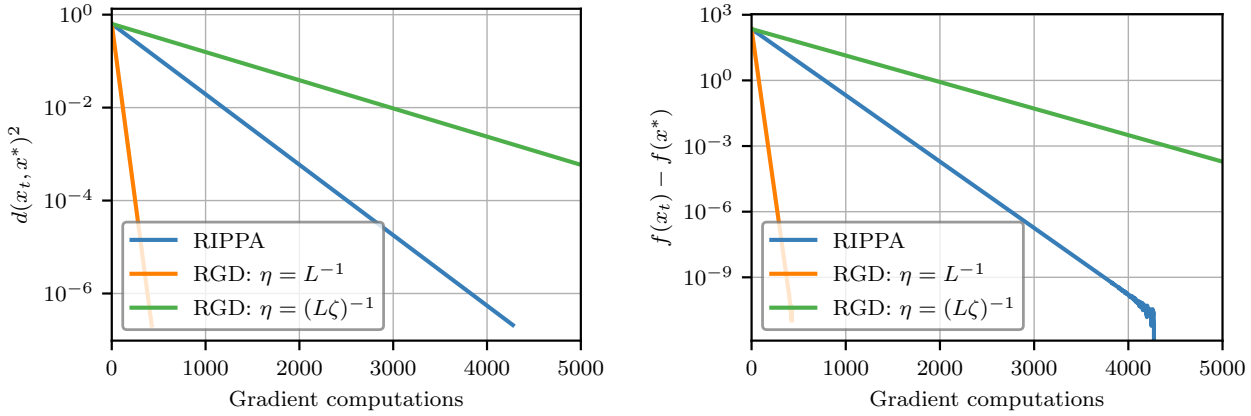

Figure 5: Karcher Mean on $\mathbb{H}^d$: $d = 1000$, $n = 500$, error in squared distance to the optimizer (left) and primal gap (right).

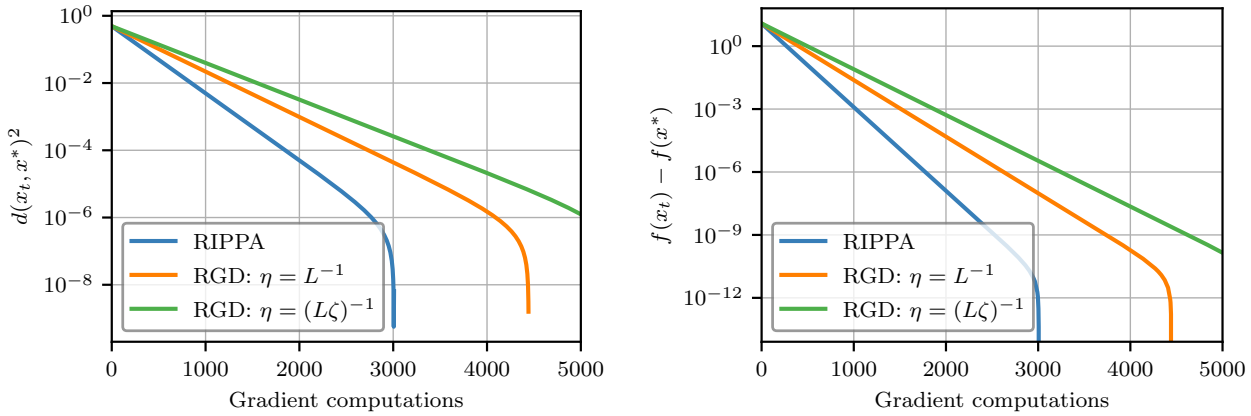

Figure 6: Karcher Mean on $\mathcal{S}_+^d$: $d = 100$, $n = 100$, error in squared distance to the optimizer (left) and primal gap (right).

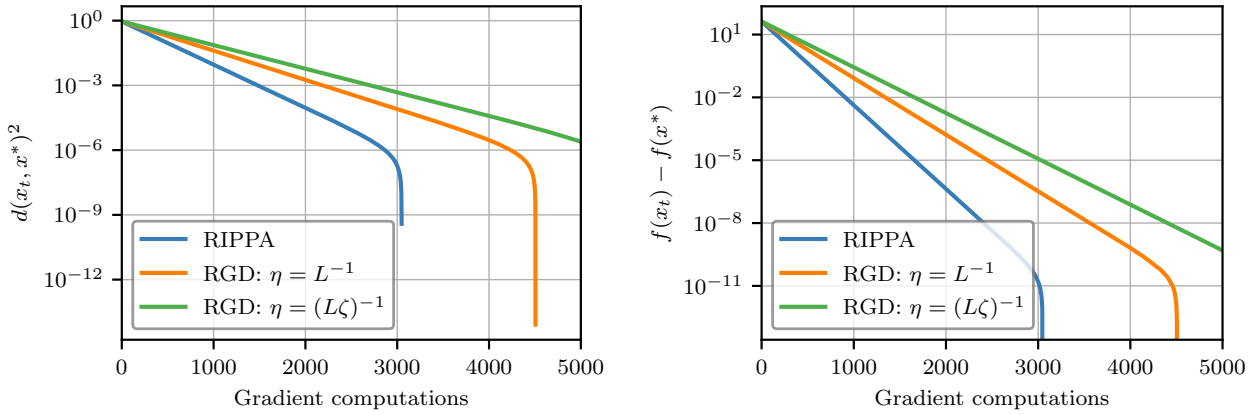

Figure 7: Karcher Mean on $\mathcal{S}_+^d$: $d = 50$, $n = 100$, error in squared distance to the optimizer (left) and primal gap (right).