Global Riemannian Acceleration in Hyperbolic and Spherical Spaces

David Martínez-Rubio   DAVID.MARTINEZ@CS.OX.AC.UK
Department of Computer Science, University of Oxford, Oxford, United Kingdom

Abstract

We further research on the accelerated optimization phenomenon on Riemannian manifolds by introducing accelerated global first-order methods for the optimization of L-smooth and geodesically convex (g-convex) or μ-strongly g-convex functions defined on the hyperbolic space or a subset of the sphere. For a manifold other than the Euclidean space, these are the first methods to *globally* achieve the same rates as accelerated gradient descent in the Euclidean space with respect to L and ε (and μ if it applies), up to log factors. Previous results with these accelerated rates only worked, given strong g-convexity, in a small neighborhood (initial distance R to a minimizer being $R = O((\mu/L)^{3/4})$). Our rates have a polynomial factor on $1/\cos(R)$ (spherical case) or $\cosh(R)$ (hyperbolic case). Thus, we completely match the Euclidean case for a constant initial distance, and for larger R we incur greater constants due to the geometry.

As a proxy for our solution, we solve a constrained non-convex Euclidean problem, under a condition between convexity and *quasar-convexity*, of independent interest. Additionally, for any Riemannian manifold of bounded sectional curvature, we provide reductions from optimization methods for smooth and g-convex functions to methods for smooth and strongly g-convex functions and vice versa.

Keywords: Riemannian optimization, geodesic convexity, accelerated optimization.

1. Introduction

Acceleration in convex optimization is a phenomenon that has drawn plenty of attention and has yielded many important results, since the renowned Accelerated Gradient Descent (AGD) method of Nesterov (1983). Having been proved successful for deep learning (Sutskever et al., 2013), among other fields, there have been recent efforts to better understand this phenomenon (Allen Zhu and Orecchia, 2017; Diakonikolas and Orecchia, 2019; Su et al., 2016; Wibisono et al., 2016). These have yielded numerous new results going beyond convexity or the standard oracle model, in a wide variety of settings (Allen-Zhu, 2017, 2018a,b; Allen Zhu and Orecchia, 2015; Allen Zhu et al., 2016; Allen-Zhu et al., 2017; Carmon et al., 2017; Cohen et al., 2018; Cutkosky and Sarlós, 2019; Diakonikolas and Jordan, 2019; Diakonikolas and Orecchia, 2018; Gasnikov et al., 2019; Wang et al., 2016). This surge of research that applies tools of convex optimization to models going beyond convexity has been fruitful. One of these models is the setting of geodesically convex Riemannian optimization. In this setting, the function to optimize is geodesically convex (g-convex), i.e. convex restricted to any geodesic (cf. Definition 1).

Riemannian optimization, g-convex and non-g-convex alike, is an extensive area of research. In recent years there have been numerous efforts towards obtaining Riemannian optimization algorithms that share analogous properties to the more broadly studied Euclidean first-order methods:

---

0. Most of the notations in this work have a link to their definitions. For example, if you click or tap on any instance of L, you will jump to the place where it is defined as the smoothness constant of the function we consider in this work.

deterministic (de Carvalho Bento et al., 2017; Wei et al., 2016; Zhang and Sra, 2016), stochastic (Hosseini and Sra, 2017; Khuzani and Li, 2017; Tripuraneni et al., 2018), variance-reduced (Sato et al., 2017, 2019; Zhang et al., 2016), adaptive (Kasai et al., 2019), saddle-point-escaping (Criscitiello and Boumal, 2019; Sun et al., 2019; Zhang et al., 2018; Zhou et al., 2019; Criscitiello and Boumal, 2020), and projection-free methods (Weber and Sra, 2017, 2019), among others. Unsurprisingly, Riemannian optimization has found many applications in machine learning, including low-rank matrix completion (Cambier and Absil, 2016; Heidel and Schulz, 2018; Mishra and Sepulchre, 2014; Tan et al., 2014; Vandereycken, 2013), dictionary learning (Cherian and Sra, 2017; Sun et al., 2017), optimization under orthogonality constraints (Edelman et al., 1998), with applications to Recurrent Neural Networks (Lezcano-Casado, 2019; Lezcano-Casado and Martínez-Rubio, 2019), robust covariance estimation in Gaussian distributions (Wiesel, 2012), Gaussian mixture models (Hosseini and Sra, 2015), operator scaling (Allen-Zhu et al., 2018), and sparse principal component analysis (Genicot et al., 2015; Huang and Wei, 2019b; Jolliffe et al., 2003).

However, the acceleration phenomenon, largely celebrated in the Euclidean space, is still not understood in Riemannian manifolds, although there has been some progress on this topic recently (cf. Related work). This poses the following question, which is the central subject of this paper:

Can a Riemannian first-order method enjoy the same rates as AGD does in the Euclidean space?

In this work, we provide an answer in the affirmative for functions defined on hyperbolic and spherical spaces, up to constants depending on the sectional curvature K and the initial distance to a minimizer R, and up to log factors. We summarize our main results in the following:

- *Full acceleration*. We design algorithms that provably achieve the same rates of convergence as AGD in the Euclidean space, up to constants and log factors. More precisely, we obtain the rates $\widetilde{O}(\sqrt{L/\varepsilon})$ and $O^*(\sqrt{L/\mu}\log(\mu/\varepsilon))$ when obtaining ε-minimizers of L-smooth functions that are, respectively, g-convex and μ-strongly g-convex, defined on the hyperbolic space or a subset of the sphere. The notation $\widetilde{O}(\cdot)$ and $O^*(\cdot)$ omits $\log(L/\varepsilon)$ and $\log(L/\mu)$ factors, respectively, and constants depending on R and K. We discuss these constants in Appendix D, noting a difference with respect to the Euclidean case and discussing lower bounds. Previous accelerated approaches only showed local results (Zhang and Sra, 2018) or obtained results with rates in between the ones obtainable by Riemannian Gradient Descent (RGD) and AGD (Ahn and Sra, 2020). Moreover, these previous works only apply to functions that are smooth and strongly g-convex and not to smooth functions that are only g-convex. As a proxy, we design an accelerated algorithm under a condition between convexity and *quasar-convexity* in the constrained setting, of independent interest.

- *Reductions*. We present two reductions for any Riemannian manifold of bounded sectional curvature. Given an optimization method for smooth and g-convex functions they provide a method for optimizing smooth and strongly g-convex functions, and vice versa.

It is often the case that methods and key geometric inequalities that apply to manifolds with bounded sectional curvatures are obtained from the ones existing for the spaces of constant extremal sectional curvature (Grove et al., 1997; Zhang and Sra, 2016, 2018). Consequently, our contribution is relevant not only because we establish an algorithm achieving global acceleration on functions defined on a manifold other than the Euclidean space, but also because understanding the constant sectional curvature case is an important step towards understanding the more general case

of obtaining algorithms that optimize g-convex functions, strongly or not, defined on manifolds of bounded sectional curvature.

Structure of the paper. We provide some definitions, notations, and related work in the rest of this section. We introduce our algorithms and their ideas and a proof sketch in Section 2 and we present our reductions in Section 3. Appendix A contains the convergence proofs of the accelerated algorithms. Appendix B contains the proofs of the reductions and the corollaries showing how to apply them to our algorithms. In Appendix C, we prove our geometric lemmas that show how to reduce our Riemannian problem to the Euclidean non-convex problem that we solve in an accelerated way. In Appendix D we comment on the constants of our algorithms, on rates of related work and on hardness results. We also show a lower bound on the condition number of any strongly g-convex function defined on a bounded domain.

Basic Geometric Definitions. We recall basic definitions of Riemannian geometry that we use in this work. For a thorough introduction we refer to (Petersen et al., 2006). A Riemannian manifold $(\mathcal{M}, \mathfrak{g})$ is a real smooth manifold $\mathcal{M}$ equipped with a metric $\mathfrak{g}$, which is a smoothly varying inner product. For $x \in \mathcal{M}$ and any two vectors $v, w \in T_x\mathcal{M}$ in the tangent space of $\mathcal{M}$, the inner product $\langle v, w \rangle_x$ is $\mathfrak{g}(v, w)$. For $v \in T_x\mathcal{M}$, the norm is defined as usual $\|v\|_x \stackrel{\text{def}}{=} \sqrt{\langle v, v \rangle_x}$. Typically, x is known given v or w, so we will just write $\langle v, w \rangle$ or $\|v\|$ if x 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 uniquely geodesic space is a space such that for every two points there is one and only one geodesic that joins them. In such a case the exponential map $\text{Exp}_x : T_x\mathcal{M} \to \mathcal{M}$ and inverse exponential map $\text{Exp}_x^{-1} : \mathcal{M} \to T_x\mathcal{M}$ are well defined for every pair of points, and are as follows. Given $x, y \in \mathcal{M}$, $v \in T_x\mathcal{M}$, and a geodesic γ of length $\|v\|$ such that $\gamma(0) = x$, $\gamma(\|v\|) = y$, $\gamma'(0) = v/\|v\|$, we have that $\text{Exp}_x(v) = y$ and $\text{Exp}_x^{-1}(y) = v$. Note, however, that $\text{Exp}_x(\cdot)$ might not be defined for each $v \in T_x\mathcal{M}$. We denote by $d(x, y)$ the distance between x and y. Its value is the same as $\|\text{Exp}_x^{-1}(y)\|$. Given a 2-dimensional subspace $V \subseteq T_x\mathcal{M}$, the sectional curvature at x with respect to V is defined as the classical notion of Gauss curvature, for the surface $\text{Exp}_x(V)$ at x. The Gauss curvature at a point x can be defined as the product of the maximum and minimum curvatures of the curves resulting from intersecting the surface with planes normal to the surface at x.

Notation. Let $\mathcal{M}$ be a n-dimensional Riemannian manifold. Given two points $x, y \in \mathcal{M}$ and a vector $v \in T_x\mathcal{M}$ in the tangent space of x, we use the formal notation $\langle v, y - x \rangle \stackrel{\text{def}}{=} -\langle v, x - y \rangle \stackrel{\text{def}}{=} \langle v, \text{Exp}_x^{-1}(y) \rangle$. We call $F : \mathcal{M} \to \mathbb{R}$ a function we want to optimize and that has at least one global minimum. We denote by x_0 an initial point, in $\mathcal{M}$, of an optimization algorithm. We denote $R \geq d(x_0, x^*)$ a bound on the initial distance to a minimizer x^*. We use the notation $\text{Exp}_{x_0}(\bar{B}(0, R)) \subset \mathcal{M}$ to mean that $\mathcal{M}$ is such that Exp_{x_0} is defined on the closed ball $\bar{B}(0, R) \subset T_{x_0}\mathcal{M}$. We use $\mathcal{M}_K$ to denote any manifold that is a subset of an n-dimensional complete and simply connected manifold of constant sectional curvature K, namely a subset of the hyperbolic space or sphere (Petersen et al., 2006), with the inherited metric, and such that $\text{Exp}_{x_0}(\bar{B}(0, R)) \subset \mathcal{M}_K$. In such a case, we use $\mathcal{B}_R$ for $\text{Exp}_{x_0}(\bar{B}(0, R))$. Note that we are not making explicit the dependence on n, x_0, $\mathcal{M}_K$, and K. We want to work with the standard choice of uniquely geodesic manifolds (Ahn and Sra, 2020; Liu et al., 2017; Zhang and Sra, 2016, 2018). Therefore, if $K > 0$ we restrict ourselves to $R < \pi/2\sqrt{K}$, so $\mathcal{B}_R$ is uniquely geodesic (it is contained in an open hemisphere). Note that by definition $x^* \in \mathcal{B}_R$. For $M \subseteq \mathbb{R}^n$, we denote by $h : \mathcal{M} \to M$ a geodesic map (Kreyszig, 1991), which is a diffeomorphism such that the image and the inverse image of a geodesic is a geodesic. Unless specified otherwise, we will have $h(x_0) = 0$. Given a point $x \in \mathcal{M}$ we use the notation $\tilde{x} \stackrel{\text{def}}{=} h(x)$ and vice versa; any point in M will use a tilde. Given a vector $v \in T_x\mathcal{M}$, we call $\tilde{v} \in \mathbb{R}^n$

the vector of the same norm such that $\{\tilde{x} + \tilde{\lambda}\tilde{v}|\tilde{\lambda} \in \mathbb{R}^+, \tilde{x} + \tilde{\lambda}\tilde{v} \in M\} = \{h(\mathrm{Exp}_x(\lambda v))|\lambda \in I \subseteq \mathbb{R}^+\}$, for some interval I. Likewise, given x and a vector $\tilde{v} \in \mathbb{R}^n$, we define $v \in T_x\mathcal{M}$. In the case of $\mathcal{M}_K$, we call $\mathcal{X} = h(\mathcal{B}_R)$. The big O notations $\widetilde{O}(\cdot)$ and $O^*(\cdot)$ omit $\log(L/\varepsilon)$ and $\log(L/\mu)$ factors, respectively, and constant factors depending on R and K.

We define now the main properties that will be assumed on the function F to be minimized.

Definition 1 (Geodesic Convexity and Smoothness) *Let $F : \mathcal{M} \to \mathbb{R}$ be a differentiable function defined on a Riemannian manifold $(\mathcal{M}, \mathfrak{g})$. Given $L \geq \mu > 0$, we say that F is L-smooth, and respectively μ-strongly g-convex, if for any two points $x, y \in \mathcal{M}$, F satisfies*

$$F(y) \leq F(x) + \langle \nabla F(x), y - x \rangle + \frac{L}{2}d(x,y)^2, \; resp. \; F(y) \geq F(x) + \langle \nabla F(x), y - x \rangle + \frac{\mu}{2}d(x,y)^2.$$

We say F is g-convex if the second inequality above, i.e. μ-strong g-convexity, is satisfied with $\mu = 0$. We have used the formal notation above for the subtraction of points in the inner product.

Our main technique consists of mapping the function domain to a subset M of the Euclidean space via a geodesic map h. Given the gradient of a point $x \in \mathcal{M}$, convexity defines a lower bound on the function that is affine over the tangent space of x, namely $\ell(y) = F(x) + \langle \nabla F(x), y - x \rangle \leq F(y)$ and it implies a minimizer must be in the halfspace $H = \{y|\langle \nabla F(x), y - x \rangle \leq 0\}$, since $\ell(\cdot)$ is greater than $F(x)$ outside of H. This lower bound induces, via the geodesic map, a function on M. And H is mapped to a halfspace H' in the Euclidean space, because $\{h(y)|\langle \nabla F(x), y-x \rangle = 0\}$ is mapped to a hyperplane by the definition of geodesic map. We find a lower bound of $\ell \circ h^{-1}$ that is affine over H' and such that it is equal to $F(x)$ at $h(x)$, despite the geodesic map being non-conformal, deforming distances, and breaking convexity, cf. [Lemma 3](). This allows to aggregate the lower bounds easily in the Euclidean space by taking an average, in the same spirit as mirror descent algorithms do. We believe that effective lower bound aggregation is key to achieving Riemannian acceleration and optimality and it has been the main hurdle of previous algorithms. Using this strategy, we are able to define a continuous method that we discretize using an approximate implementation of the implicit Euler method, achieving the same rates as the Euclidean AGD, up to constants and log factors, for the optimization of g-convex smooth functions. Our reductions take into account the deformations produced by the geometry to generalize existing optimal Euclidean reductions ([Allen Zhu and Hazan, 2016](); [Allen Zhu and Orecchia, 2017]()). Applying them, we obtain an analogous algorithm for strongly g-convex and smooth functions. Applying them again to the latter they yield an algorithm for g-convex smooth functions with the rates of the same order as the first one.

Comparison with Related Work. There are a number of works that study the problem of first-order acceleration in Riemannian manifolds of bounded sectional curvature. The first study is ([Liu et al., 2017]()). In this work, the authors develop an accelerated method with the same rates as AGD for both g-convex and strongly g-convex functions, provided that at each step a given non-linear equation can be solved. No algorithm for solving this equation has been found and, in principle, it could be intractable or infeasible. In ([Alimisis et al., 2019]()) a continuous method analogous to the continuous approach to accelerated methods is presented, but it is not known if there exists an accelerated discretization of it. In ([Alimisis et al., 2020]()), an algorithm presented is claimed to enjoy an accelerated rate of convergence, but fails to provide convergence when the function value gets below a potentially large constant that depends on the manifold and smoothness constant. The work ([Lin et al., 2020]()) is inspired by accelerated algorithms and focuses on adapting to the strong g-convex

4

parameter but does not obtain accelerated algorithms. In (Huang and Wei, 2019a) an accelerated algorithm is presented but relying on strong geometric inequalities that are not proved to be satisfied. Zhang and Sra (2018) obtain a *local* algorithm that optimizes L-smooth and μ-strongly g-convex functions achieving the same rates as AGD in the Euclidean space, up to constants. That is, the initial point needs to start close to the optimum, $O((\mu/L)^{3/4})$ close, to be precise. Their approach consists of adapting Nesterov's estimate sequence technique by keeping a quadratic on $T_{x_i}\mathcal{M}$ that induces on $\mathcal{M}$ a regularized lower bound on $F(x^*)$ via $\mathrm{Exp}_{x_i}(\cdot)$. They build another lower bound by aggregating the information yielded by the gradient $\nabla F(x_i)$ to it, and use a geometric lemma to find a quadratic in $T_{x_{i+1}}\mathcal{M}$ whose induced function lower bounds the previous one. Ahn and Sra (2020) generalize the previous algorithm and, by using similar ideas for the lower bound, they adapt it to work globally, obtaining strictly better rates than RGD, recovering the local acceleration of the previous paper, but not achieving global rates comparable to the ones of AGD. In fact, they prove that their algorithm eventually decreases the function value at a rate close to AGD but this can take as many iterations as the ones needed by RGD to reach the neighborhood of the previous local algorithm, cf. Remark 30. In our work, we take a step back and focus on the constant sectional curvature case to provide a global algorithm that achieves the same rates as AGD, up to constants on R, K, and log factors. It is common to characterize the properties of spaces of bounded sectional curvature by using the ones of the spaces of constant extremal sectional curvature (Grove et al., 1997; Zhang and Sra, 2016, 2018), which makes the study of the constant sectional curvature case critical to the development of full accelerated algorithms in the general bounded sectional curvature case. Our work also studies g-convexity besides strong g-convexity. No previous accelerated algorithms applied to this case. Some of the results of Ahn and Sra (2020), among others, require to assume that the iterates of the algorithm remain in a fixed ball around the minimizer, while our algorithms and analysis can work with this constraint directly. Because of the hardness of the geometry, our convergence rates incur greater constants depending on R and K with respect to the Euclidean case. They are a small polynomial on $1/\cos(R\sqrt{|K|})$ in spherical spaces and $\cosh(R\sqrt{|K|})$ in hyperbolic spaces, the latter being an exponential dependence. It is not clear if these constants can be avoided in fully accelerated algorithms and we provide a discussion about this in Appendix D. We also show a lower bound on the condition number of any strongly g-convex function defined on $\mathcal{B}_R$. Due to the geometry, there is a lower bound (Hamilton and Moitra, 2021; Criscitiello and Boumal, 2021) of $\widetilde{\Omega}(R)$ on several negatively curved Riemannian manifolds, but this does not preclude to globally accelerate, unless this lower bound preponderates over the condition number, cf. Remark 29.

On Euclidean optimization, a related work is the *approximate duality gap technique* (Diakonikolas and Orecchia, 2019), which presents a unified view of the analysis of first-order methods. It defines a continuous duality gap and by enforcing a natural invariant, it obtains accelerated continuous dynamics and their discretizations for most classical first-order methods. A derived work (Diakonikolas and Orecchia, 2018) obtains Euclidean acceleration in a fundamentally different way from previous acceleration approaches, namely using an approximate implicit Euler method for the discretization of the acceleration dynamics. Our convergence analysis of Theorem 5 draws ideas from these two works. We will see in the sequel that, for our manifolds of interest, g-convexity is related to a model known as quasar-convexity or weak-quasi-convexity (Guminov and Gasnikov, 2017; Nesterov et al., 2018; Hinder et al., 2019).

Table 1: Worse-case rates of related works for smooth problems. AGD is a Euclidean algorithm. Our Algorithm 1 works in manifolds of constant sectional curvature $K \neq 0$. The rest of the algorithms work in manifolds of sectional curvature that is bounded above and below.

Method	g-convex	μ-st. g-convex
AGD (Nesterov, 1983)	$O(\sqrt{L/\varepsilon})$	$O(\sqrt{L/\mu}\log(\mu/\varepsilon))$
(Zhang and Sra, 2018) (it only works locally)	-	$O(\sqrt{L/\mu}\log(\mu/\varepsilon))$
(Ahn and Sra, 2020)	-	$O^*(L/\mu + \sqrt{L/\mu}\log(\mu/\varepsilon))$
Remark 30 (**RGD**+(**Zhang and Sra, 2018**))	-	$O^*(L/\mu + \sqrt{L/\mu}\log(\mu/\varepsilon))$
Theorem 6 and **Corollary 8** resp.	$\widetilde{O}(\sqrt{L/\varepsilon})$	$O^*(\sqrt{L/\mu}\log(\mu/\varepsilon))$

2. Algorithm

We study the minimization problem $\min_{x \in \mathcal{M}_K} F(x)$ with a gradient oracle, for a twice differentiable smooth function $F : \mathcal{M}_K \to \mathbb{R}$ that is g-convex or strongly g-convex, for an initial point x_0, a minimizer x^* of F that is assumed to exist, and a constant $R > d(x_0, x^*)$. We recall $\mathcal{M}_K$ refers to any manifold of constant non-zero sectional curvature such that $\mathcal{B}_R = \mathrm{Exp}_{x_0}(\bar{B}(0, R)) \subset \mathcal{M}_K$. We work in this setting in this entire section. We perform constrained optimization over $\mathcal{B}_R$ in order to control the deformations caused by the geometry. We defer the proofs of the lemmas and theorems in this and following sections to the appendix. We assume without loss of generality that the sectional curvature of $\mathcal{M}_K$ is $K \in \{1, -1\}$, since for any other value of K and any function $F : \mathcal{M}_K \to \mathbb{R}$ defined on such a manifold, we can reparametrize F by a rescaling, so it is defined over a manifold of constant sectional curvature $K \in \{1, -1\}$. The parameters L, μ and R are rescaled accordingly as a function of K, cf. Remark 22. We denote the special cosine by $\mathrm{C}_K(\cdot)$, which is $\cos(\cdot)$ if $K = 1$ and $\cosh(\cdot)$ if $K = -1$. For a geodesic map $h : \mathcal{M} \to M$, we define $\mathcal{X} \stackrel{\mathrm{def}}{=} h(\mathcal{B}_R) \subseteq M \subseteq \mathbb{R}^n$. We use classical geodesic maps for the manifolds that we consider: the Gnomonic projection for $K = 1$ and the Beltrami-Klein projection for $K = -1$ (Greenberg, 1993). They map an open hemisphere and the hyperbolic space of curvature $K \in \{1, -1\}$ to $\mathbb{R}^n$ and $B(0, 1) \subseteq \mathbb{R}^n$, respectively. We will derive our results from the following characterization of h (Greenberg, 1993). Let $\tilde{x}, \tilde{y} \in \mathcal{X}$ be two points. Recall that we denote $x = h^{-1}(\tilde{x}), y = h^{-1}(\tilde{y}) \in \mathcal{B}_R$. Then we have that $d(x, y)$, the distance between x and y with the metric of $\mathcal{M}_K$, satisfies

$$\mathrm{C}_K(d(x, y)) = \frac{1 + K\langle \tilde{x}, \tilde{y} \rangle}{\sqrt{1 + K\|\tilde{x}\|^2} \cdot \sqrt{1 + K\|\tilde{y}\|^2}}. \tag{1}$$

Observe that the expression is symmetric with respect to rotations. In particular, $\mathcal{X}$ is a closed ball of some radius $\tilde{R}$. Using $\tilde{x} = 0$ and $\tilde{y}$ such that $d(x_0, y) = R$, we have $\mathrm{C}_K(R) = (1 + K\tilde{R}^2)^{-1/2}$.

Consider a point $x \in \mathcal{B}_R$ and the lower bound provided by the g-convexity assumption when computing $\nabla F(x)$. Dropping the μ term in case of strong g-convexity, this bound is affine over $T_x\mathcal{B}_R$. In order to define a duality gap, as we show in Section 2.1, we would like our algorithm to aggregate effectively the lower bounds it computes during the course of the optimization. The deformations of the geometry make the aggregation a difficult task, despite the fact that we have a simple description of each individual lower bound: each of them is affine over $T_{x_i}\mathcal{B}_R$ but these

simple functions are defined on different tangent spaces. We deal with this problem by obtaining a lower bound that is looser by a constant depending on R, and that is affine over $\mathcal{X} \subset \mathbb{R}^n$. In this way the aggregation becomes easier: all of them are simple and are in the same space. Then, we are able to combine this lower bound with decreasing upper bounds in the fashion some other accelerated methods work in the Euclidean space (Allen Zhu and Orecchia, 2017; Diakonikolas and Orecchia, 2018, 2019; Nesterov, 1983). Alternatively, we can see the approach in this work as the constrained optimization problem of minimizing the non-convex function $f : \mathcal{X} \to \mathbb{R}, \tilde{x} \mapsto F(h^{-1}(\tilde{x}))$

$$\text{minimize } f(\tilde{x}), \quad \text{for } \tilde{x} \in \mathcal{X}.$$

In the rest of the section, we will focus on the g-convex case. For simplicity, instead of solving the strongly g-convex case directly in an analogous way by finding a lower bound that is quadratic over $\mathcal{X}$, we rely on the reductions of Section 3 to obtain the accelerated algorithm in this case.

The following two lemmas show that finding the affine lower bound is possible, and is defined as a function of $\nabla f(\tilde{x})$. We first gauge the deformations caused by the geodesic map h. Distances are deformed, the map h is not conformal, and the image of the geodesic $\mathrm{Exp}_x(\lambda \nabla F(x))$ is not mapped into the image of the geodesic $\tilde{x} + \tilde{\lambda} \nabla f(\tilde{x})$, i.e. the direction of the gradient changes. We are able to find the affine lower bound after bounding these deformations.

Lemma 2 [↓] *Let $K \in \{1, -1\}$. Let $x, y \in \mathcal{B}_R$ be two different points, and in part b) different from x_0. Let $\tilde{\alpha}$ be the angle $\angle \tilde{x}_0 \tilde{x} \tilde{y}$, formed by the vectors $\tilde{x}_0 - \tilde{x}$ and $\tilde{y} - \tilde{x}$. Let α be the corresponding angle, the one between the vectors $\mathrm{Exp}_x^{-1}(x_0)$ and $\mathrm{Exp}_x^{-1}(y)$. Assume without loss of generality that $\tilde{x} \in \mathrm{span}\{\tilde{e}_1\}$ and $\nabla f(\tilde{x}) \in \mathrm{span}\{\tilde{e}_1, \tilde{e}_2\}$ for the canonical orthonormal basis $\{\tilde{e}_i\}_{i=1}^n$. Let $e_i \in T_x \mathcal{M}_K$ be the unit vector such that h maps the image of the geodesic $\mathrm{Exp}_x(\lambda e_i)$ to the image of the geodesic $\tilde{x} + \tilde{\lambda} e_i$, for $i = 1, \ldots, n$, and $\lambda, \tilde{\lambda} \geq 0$. Then, the following holds.*

a) *Distance deformation:*

$$K \, \mathrm{C}_K^2(R) \leq K \frac{d(x, y)}{\|\tilde{x} - \tilde{y}\|} \leq K.$$

b) *Angle deformation:*

$$\sin(\alpha) = \sin(\tilde{\alpha}) \sqrt{\frac{1 + K\|\tilde{x}\|^2}{1 + K\|\tilde{x}\|^2 \sin^2(\tilde{\alpha})}}, \qquad \cos(\alpha) = \cos(\tilde{\alpha}) \sqrt{\frac{1}{1 + K\|\tilde{x}\|^2 \sin^2(\tilde{\alpha})}}.$$

c) *Gradient deformation:*

$$\nabla F(x) = (1 + K\|\tilde{x}\|^2) \nabla f(\tilde{x})_1 e_1 + \sqrt{1 + K\|\tilde{x}\|^2} \nabla f(\tilde{x})_2 e_2 \quad \text{and} \quad e_i \perp e_j \text{ for } i \neq j.$$

And if $v \in T_x \mathcal{M}_K$ is a vector that is normal to $\nabla F(x)$, then $\tilde{v}$ is normal to $\nabla f(x)$.

The previous lemma shows that $\nabla f(\tilde{x})$ can be easily computed from $\nabla F(x)$. The following lemma uses the deformations described in Lemma 2 to obtain the affine lower bound on the function, given a gradient at a point $\tilde{x}$. Note that Lemma 2.c implies that we have $\langle \nabla f(\tilde{x}), \tilde{y} - \tilde{x} \rangle = 0$ if and only if $\langle \nabla F(x), y - x \rangle = 0$. In the proof we lower bound, generally, affine functions defined on $T_x \mathcal{M}_K$ by affine functions in the Euclidean space $\mathcal{X}$. This generality allows to obtain a result with constants that only depend on R. See Remark 28 for a discussion on these constants.

Lemma 3 [↓] *Let $F : \mathcal{M}_K \to \mathbb{R}$ be a differentiable function and let $f = F \circ h^{-1}$. Then, there are constants $\gamma_{\mathrm{n}}, \gamma_{\mathrm{p}} \in (0, 1]$ depending on R such that for all $x, y \in \mathcal{B}_R$ satisfying $\langle \nabla f(\tilde{x}), \tilde{y} - \tilde{x} \rangle \neq 0$ we have:*

$$\gamma_{\mathrm{p}} \leq \frac{\langle \nabla F(x), y - x \rangle}{\langle \nabla f(\tilde{x}), \tilde{y} - \tilde{x} \rangle} \leq \frac{1}{\gamma_{\mathrm{n}}}. \tag{2}$$

In particular, if F is g-convex we have the following condition, that we call tilted-convexity:

$$
\begin{aligned}
f(\tilde{x}) + \frac{1}{\gamma_{\mathrm{n}}} \langle \nabla f(\tilde{x}), \tilde{y} - \tilde{x} \rangle \leq f(\tilde{y}) \quad &\text{if } \langle \nabla f(\tilde{x}), \tilde{y} - \tilde{x} \rangle \leq 0, \\
f(\tilde{x}) + \gamma_{\mathrm{p}} \langle \nabla f(\tilde{x}), \tilde{y} - \tilde{x} \rangle \leq f(\tilde{y}) \quad &\text{if } \langle \nabla f(\tilde{x}), \tilde{y} - \tilde{x} \rangle \geq 0.
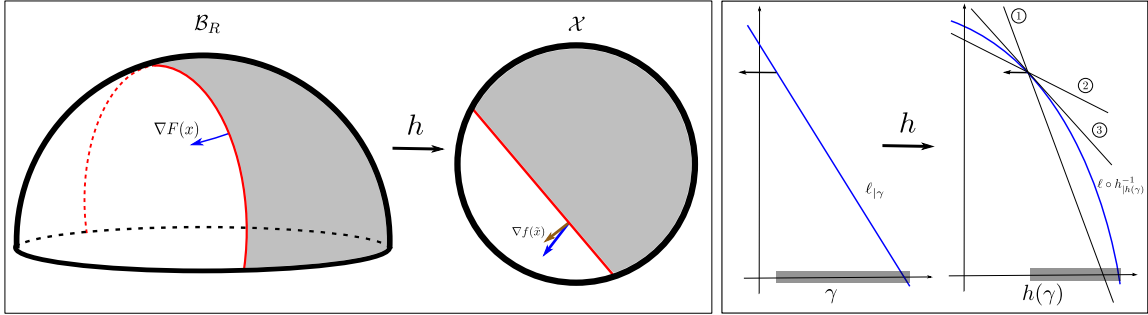\end{aligned}
\tag{3}
$$

Figure 1: Deformations of the map h.

We provide intuition for the previous lemma through Figure 1. The geodesic that $\nabla F(x)$ induces on $\mathcal{B}_R$ corresponds to the geodesic that the blue vector would induce on $\mathcal{X}$, which is in a different direction than the one induced by $\nabla f(\tilde{x})$. The angle and gradient deformations of Lemma 2 allow to show that, for any direction, inducing a geodesic γ, the slope of the affine function induced by $\nabla f(\tilde{x})$ on $\mathcal{X}$ is within a constant factor c_1 of the one of the lower bound ℓ defined by $\nabla F(x)$ in $\mathcal{B}_R$. Our main aim is to bound $F(x^*)$ and the shaded area of each image represents where x^* can be. On the right, we exemplify the deformation on a geodesic γ passing through x. Initially, we have the affine lower bound ℓ, but the map h deforms the domain. To lower bound the function on the shaded region, we can use an affine lower bound ①. Its slope is within a constant factor of the one of the tangent line ③ by the distance deformation of Lemma 2 and the factor c_1—the latter bounds the change of the directional derivatives, in black. This gives the first line of (3). The other one is analogous, using another affine function ②.

The first inequality in tilted-convexity shows the affine lower bound, which can be used to bound $f(\tilde{x}^*) = F(x^*)$. This first inequality, only applied to $\tilde{y} = \tilde{x}^*$ for a function $f : \mathbb{R}^n \to \mathbb{R}$, defines a model known in the literature as quasar-convexity or weak-quasi-convexity (Guminov and Gasnikov, 2017; Nesterov et al., 2018; Hinder et al., 2019), for which accelerated algorithms exist in the *unconstrained case*, provided smoothness is also satisfied. However, to the best of our knowledge, there is no known algorithm for solving the constrained case in an accelerated way. The condition in (3) is a relaxation of convexity that is stronger than quasar-convexity. We will make use of (3) in order to obtain acceleration in the constrained setting. This is of independent interest.

Recall that we need the constraint to guarantee bounded deformation due to the geometry. We also require smoothness of f. We prove in the following lemma that f is as smooth as F up to a constant depending on R.

Lemma 4 [↓] *The function f is $O(L)$-smooth in $\mathcal{X} = h(\mathcal{B}_R)$ if $F : \mathcal{M}_K \to \mathbb{R}$ is L-smooth.*

Inspired by the *approximate duality gap technique* (Diakonikolas and Orecchia, 2019) we obtain accelerated continuous dynamics, for the optimization of the function f. Then we achieve acceleration by obtaining an implicit Euler discretization of the dynamics. Diakonikolas and Orecchia (2018) obtained an accelerated method for convex functions by also making use of an implicit Euler discretization. Their algorithm is fundamentally different from AGD and techniques as Linear Coupling (Allen Zhu and Orecchia, 2017) or Nesterov's estimate sequence (Nesterov, 1983). The latter techniques use a balancing gradient step at each iteration to compensate the regret of an implicit or explicit dual algorithm, like mirror descent. Our use of a looser lower bound makes this regret greater by a constant factor and it complicates guaranteeing finding a gradient step within the constraints to compensate this greater regret. We state here the accelerated theorem and provide a sketch of the proof in Section 2.1.

Theorem 5 [↓] *Let $Q \subseteq \mathbb{R}^n$ be a closed convex set. Let $f : Q \to \mathbb{R}$ be an $\tilde{L}$-smooth, tilted-convex function with constants $\gamma_{\mathrm{n}}, \gamma_{\mathrm{p}} \in (0, 1]$. Assume there is a point $\tilde{x}^* \in Q$ such that $\nabla f(\tilde{x}^*) = 0$. We can obtain an ε-minimizer of f using $\widetilde{O}([\tilde{L}/(\gamma_{\mathrm{n}}^2 \gamma_{\mathrm{p}} \varepsilon)]^{1/2})$ queries to the gradient oracle of f.*

Finally, we have Riemannian acceleration as a consequence of Lemma 3, Lemma 4 and Theorem 5.

Theorem 6 (g-Convex Acceleration) [↓] *Let $F : \mathcal{M}_K \to \mathbb{R}$ be an L-smooth and g-convex function with a point $x^* \in \mathcal{B}_R$ satisfying $\nabla F(x^*) = 0$. Algorithm 1 computes a point $x_T \in \mathcal{B}_R$ satisfying $F(x_T) - F(x^*) \leq \varepsilon$ using $\widetilde{O}(\sqrt{L/\varepsilon})$ queries to the gradient oracle.*

We provide a sketch of the main optimization theorem in the section below. The full proof can be found in Appendix A. Our use of geodesic maps was a choice we used to be able to aggregate lower bounds. Our method showcases that an effective lower bound aggregation makes possible to achieve global full acceleration. It suggests that acceleration could also be achieved for functions defined on other manifolds by using our accelerated techniques if we can effectively aggregate the lower bounds yielded by the gradient at each iteration to build a lower bound on $F(x^*)$, similarly as in (4) below. We observe that if there is a geodesic map mapping a manifold into a convex subset of the Euclidean space then the manifold must necessarily have constant sectional curvature, cf. Beltrami's Theorem (Busemann and Phadke, 1984; Kreyszig, 1991). This means that lower bound aggregation in other manifolds would need to use a different kind of transformations. The field of comparison geometry allows to obtain properties of spaces of bounded sectional curvature by using the properties of the spaces that have constant curvature equal to the bounds of the former (Grove et al., 1997). Other Riemannian optimization algorithms have used comparison theorems that allow to obtain convergence bounds after computing the maximum possible deformations in spaces of extremal constant sectional curvature and relating them to the spaces of bounded sectional curvature (Zhang and Sra, 2016, 2018). The generalization to algorithms to optimize functions defined on manifolds of bounded sectional curvature is a future direction of research.

2.1. Sketch of the proof of Theorem 5

Inspired by the *approximate duality gap technique* (Diakonikolas and Orecchia, 2019), we let α_t be an increasing function of time t, and denote $A_t = \int_{t_0}^t d\alpha_\tau = \int_{t_0}^t \dot\alpha_\tau d\tau$. We define a continuous method that keeps a solution $\tilde{x}_t$, along with a differentiable upper bound U_t on $f(\tilde{x}_t)$ and a lower bound L_t on $f(\tilde{x}^*)$. In our case f is differentiable so we can just take $U_t = f(\tilde{x}_t)$. The lower bound comes from

$$f(\tilde{x}^*) \geq \frac{\int_{t_0}^t f(\tilde{x}_\tau)d\alpha_\tau}{A_t} + \frac{\int_{t_0}^t \frac{1}{\gamma_{\mathrm{n}}}\langle \nabla f(\tilde{x}_\tau), \tilde{x}^* - \tilde{x}_\tau\rangle d\alpha_\tau}{A_t}, \tag{4}$$

after adding and subtracting a regularizer ψ, which is a 1-strongly convex function, and after removing the unknown $\tilde{x}^*$ by taking a minimum over $\mathcal{X}$. Note (4) comes from averaging (3) for $\tilde{y} = \tilde{x}^*$. Then, if we define the gap $G_t = U_t - L_t$ and design a method that forces $\alpha_t G_t$ to be non-increasing, we can deduce $f(x_t) - f(\tilde{x}^*) \leq G_t \leq \alpha_{t_0} G_{t_0}/\alpha_t$. By forcing $\frac{d}{dt}(\alpha_t G_t) = 0$, we naturally obtain the following continuous dynamics, where z_t is a mirror point and ψ^* is the Fenchel dual of ψ.

$$\dot{z}_t = -\frac{1}{\gamma_{\mathrm{n}}}\dot\alpha_t \nabla f(\tilde{x}_t); \quad \dot{\tilde{x}}_t = \frac{1}{\gamma_{\mathrm{n}}}\dot\alpha_t \frac{\nabla\psi^*(\tilde{z}_t) - \tilde{x}_t}{\alpha_t}; \quad z_{t_0} = \nabla\psi^*(\tilde{x}_{t_0}), \tilde{x}_{t_0} \in \mathcal{X}. \tag{5}$$

We note that except for the constant γ_{n}, these dynamics match the accelerated dynamics used in the optimization of convex functions (Krichene et al., 2015; Diakonikolas and Orecchia, 2018, 2019). The AXGD algorithm (Diakonikolas and Orecchia, 2018), designed for the accelerated optimization of convex functions, discretizes the dynamics coming from the optimization of convex functions by using an approximate implementation of implicit Euler discretization. This has the advantage of not needing a gradient step per iteration to compensate for some positive discretization error. In our case, the extra error we incur by using a looser lower bound does not seem to be able to be compensated by a gradient step in the constrained case, as other acceleration techniques like Linear Coupling (Allen Zhu and Orecchia, 2017) or Nesterov's estimate sequence (Nesterov, 1983) do, so obtaining an approximate implicit Euler discretization proves to be a better approach. However, our dynamics are different and in our case we must use tilted-convexity (3) instead of convexity. We are able to obtain the following discretization coming from an approximate implicit Euler discretization:

$$\begin{cases} \tilde{\chi}_i = \frac{\hat\gamma_i A_i}{A_i\hat\gamma_i + a_{i+1}/\gamma_{\mathrm{n}}}\tilde{x}_i + \frac{a_{i+1}/\gamma_{\mathrm{n}}}{A_i\hat\gamma_i + a_{i+1}/\gamma_{\mathrm{n}}}\nabla\psi^*(\tilde{z}_i) \\ \tilde{\zeta}_i = \tilde{z}_i - \frac{a_{i+1}}{\gamma_{\mathrm{n}}}\nabla f(\tilde{\chi}_i) \\ \tilde{x}_{i+1} = \frac{\hat\gamma_i A_i}{A_i\hat\gamma_i + a_{i+1}/\gamma_{\mathrm{n}}}\tilde{x}_i + \frac{a_{i+1}/\gamma_{\mathrm{n}}}{A_i\hat\gamma_i + a_{i+1}/\gamma_{\mathrm{n}}}\nabla\psi^*(\tilde{\zeta}_i) \\ \tilde{z}_{i+1} = \tilde{z}_i - \frac{a_{i+1}}{\gamma_{\mathrm{n}}}\nabla f(\tilde{x}_{i+1}) \end{cases} \tag{6}$$

where $\hat\gamma_i \in [\gamma_{\mathrm{p}}, 1/\gamma_{\mathrm{n}}]$ is a parameter, $\tilde{x}_0 \in \mathcal{X}$ is an arbitrary point, $\tilde{z}_0 = \nabla\psi(\tilde{x}_0)$ and now α_t is a discrete measure and $\dot\alpha_t$ is a weighted sum of Dirac delta functions $\dot\alpha_t = \sum_{i=1}^\infty a_i\delta(t - (t_0+i-1))$. However, not having convexity, in order to have per-iteration discretization error less than $\hat\varepsilon/A_T$, we require $\hat\gamma_i$ to be such that $\tilde{x}_{i+1}$ satisfies

$$f(\tilde{x}_{i+1}) - f(\tilde{x}_i) \leq \hat\gamma_i\langle\nabla f(\tilde{x}_{i+1}), \tilde{x}_{i+1} - \tilde{x}_i\rangle + \hat\varepsilon, \tag{7}$$

where $\hat\varepsilon$ is chosen so that the accumulated discretization error is $< \varepsilon/2$, after having performed the steps necessary to obtain an $\varepsilon/2$ minimizer. We would like to use (3) to find such a $\hat\gamma_i$ but we need to take into account that we only know $\tilde{x}_{i+1}$ a posteriori. Indeed, using (3) we conclude that

setting $\hat{\gamma}_i$ to $1/\gamma_n$ or γ_p then we either satisfy (7) or there is a point $\hat{\gamma}_i \in (\gamma_p, 1/\gamma_n)$ for which $\langle \nabla f(\tilde{x}_{i+1}), \tilde{x}_{i+1} - \tilde{x}_i \rangle = 0$, which satisfies the inequality for $\hat{\varepsilon} = 0$. Then, using smoothness of f, the fact that $\nabla f(\tilde{x}^*) = 0$, and boundedness of $\mathcal{X}$ we can guarantee that a binary search finds a point satisfying (7) in $O(\log(\tilde{L}i/\gamma_n \hat{\varepsilon}))$ iterations. Each iteration of the binary search requires to run (6), that is, one step of the discretization. Computing the final discretization error, we obtain acceleration after choosing appropriate learning rates a_i. Algorithm 1 contains the pseudocode of this algorithm along with the reduction of the problem from minimizing F to minimizing f. We chose $\psi(\tilde{x}) \stackrel{\text{def}}{=} \frac{1}{2}\|\tilde{x}\|^2$ as our strongly convex regularizer.

Algorithm 1 Global Fully Accelerated g-Convex Minimization

Input: Initial point $x_0 \in \mathcal{M}_K$. Constants $\tilde{L}, \gamma_p, \gamma_n$. Geodesic map h satisfying (1) and $h(x_0) = 0$.
    Smooth and g-convex function $F : \mathcal{M}_K \to \mathbb{R}$ with a minimizer $x^* \in \mathcal{B}_R$.
    Bound on the distance to a minimum $R \geq d(x_0, x^*)$. Accuracy ε and number of iterations T.

1: $\mathcal{X} \stackrel{\text{def}}{=} h(\mathcal{B}_R) \subseteq M$; $f \stackrel{\text{def}}{=} F \circ h^{-1}$ and $\psi(\tilde{x}) \stackrel{\text{def}}{=} \frac{1}{2}\|\tilde{x}\|^2$
2: $\tilde{z}_0 \leftarrow \nabla\psi(\tilde{x}_0)$; $A_0 \leftarrow 0$
3: **for** $i = 0$ **to** $T - 1$ **do**
4:     $a_{i+1} \leftarrow (i+1)\gamma_n^2\gamma_p/2\tilde{L}$
5:     $A_{i+1} \leftarrow A_i + a_{i+1}$
6:     $\lambda \leftarrow \text{BinaryLineSearch}(\tilde{x}_i, \tilde{z}_i, f, \mathcal{X}, a_{i+1}, A_i, \varepsilon, \tilde{L}, \gamma_n, \gamma_p)$ (cf. Algorithm 2 in Appendix A)
7:     $\tilde{\chi}_i \leftarrow (1 - \lambda)\tilde{x}_i + \lambda\nabla\psi^*(\tilde{z}_i)$
8:     $\tilde{\zeta}_i \leftarrow \tilde{z}_i - (a_{i+1}/\gamma_n)\nabla f(\tilde{\chi}_i)$
9:     $\tilde{x}_{i+1} \leftarrow (1 - \lambda)\tilde{x}_i + \lambda\nabla\psi^*(\tilde{\zeta}_i)$ $\left[\nabla\psi^*(\tilde{p}) = \arg\min_{\tilde{z}\in\mathcal{X}}\{\|\tilde{z} - \tilde{p}\|\} = \Pi_{\mathcal{X}}(\tilde{p})\right]$
10:    $\tilde{z}_{i+1} \leftarrow \tilde{z}_i - (a_{i+1}/\gamma_n)\nabla f(\tilde{x}_{i+1})$
11: **end for**
12: **return** x_T.

---

3. Reductions

The construction of reductions proves to be very useful in order to facilitate the design of algorithms in different settings. Moreover, reductions are a helpful tool to infer new lower bounds without extra ad hoc analyses. We present two reductions. We will see in Corollary 8 and Example 1 that one can obtain full accelerated methods to minimize smooth and strongly g-convex functions from our accelerated methods for smooth and g-convex functions and vice versa. These are generalizations of some reductions designed to work in the Euclidean space (Allen Zhu and Hazan, 2016; Allen Zhu and Orecchia, 2017). The reduction to strongly g-convex functions takes into account the effect of the deformation of the space on the strong convexity of the function $F_y(x) = d(x, y)^2/2$, for $x, y \in \mathcal{M}$. It does not entail an extra $\log(1/\varepsilon)$ factor. The reduction to g-convexity requires the rate of the algorithm that applies to g-convex functions to be proportional to the squared distance between the initial point and the optimum $d(x_0, x^*)^2$ or to a bound R^2. The proofs of the statements in this section can be found in the appendix. We will use $\text{Time}_{\text{ns}}(\cdot)$ and $\text{Time}(\cdot)$ to denote the time algorithms $\mathcal{A}_{\text{ns}}$ and $\mathcal{A}$ below require, respectively, to perform the tasks we define below.

Theorem 7 [↓] *Let $\mathcal{M}$ be a Riemannian manifold, let $F : \mathcal{M} \to \mathbb{R}$ be an L-smooth and μ-strongly g-convex function, with a minimizer x^*. Suppose we have an algorithm $\mathcal{A}_{\text{ns}}$ to minimize*

F, such that for any starting point x_0 such that $d(x_0, x^*) \leq R$, it produces a point $\hat{x}_T$ in time $T = \mathrm{Time}_{\mathrm{ns}}(L, \mu, R)$ satisfying $F(\hat{x}_T) - F(x^*) \leq \mu R^2/4$. Then we can compute an ε-minimizer of F in time $O(\mathrm{Time}_{\mathrm{ns}}(L, \mu, R) \log(\mu/\varepsilon))$.

[Theorem 7](#) implies that if we forget about the strong g-convexity of a function and we treat it as if it is just g-convex then we can run in stages an algorithm designed for optimizing g-convex functions and still achieve acceleration. The fact that the function is strongly g-convex is only used between stages. We exemplify the power of the reduction by applying it to [Algorithm 1](#).

Corollary 8 [↓] *We can compute an ε-minimizer of an L-smooth and μ-strongly g-convex function $F : \mathcal{M}_K \to \mathbb{R}$ in $O^*(\sqrt{L/\mu} \log(\mu/\varepsilon))$ queries to the gradient oracle.*

We note that in the strongly convex case, by decreasing the function value by a factor we can guarantee we decrease the distance to x^* by another factor, so we can periodically recenter the geodesic map to reduce the constants produced by the deformations of the geometry, see the proof of [Corollary 8](#). Finally, we show the reverse reduction.

Theorem 9 (simplified, cf. [Theorem 19](#)) *Let $F : \mathcal{M} \to \mathbb{R}$ be L-smooth and g-convex, and let $\mathcal{M}$ be of bounded sectional curvature. Let Δ satisfy $F(x_0) - F(x^*) \leq \Delta$. Let $\mathcal{A}$ be an algorithm that in time $\hat{T} = \mathrm{Time}(L, \mu, \mathcal{M}, R)$ produces $\hat{x} \in \mathrm{Exp}_{x_0}(\bar{B}(0, R))$ that reduces the gap of an L-smooth and μ-strongly g-convex function $\hat{F} : \mathcal{M} \to \mathbb{R}$, with minimizer in $\mathrm{Exp}_{x_0}(\bar{B}(0, R))$, by a factor of $1/4$, i.e., $\hat{F}(\hat{x}) - \min_{x \in \mathcal{M}} \hat{F}(x) \leq (\hat{F}(\hat{x}_0) - \min_{x \in \mathcal{M}} \hat{F}(x))/4$. Let $T = \lceil \log_2(\Delta/\varepsilon) \rceil + 1$. Then, we can compute an ε-minimizer in time $\sum_{t=0}^{T-1} \mathrm{Time}(L + O(2^{-t}\Delta), O(2^{-t}\Delta), \mathcal{M}, R)$.*

Example 1 [↓] Applying [Theorem 9](#) to the algorithm in [Corollary 8](#) we can optimize L-smooth and g-convex functions defined on $\mathcal{M}_K$ with a gradient oracle complexity of $\widetilde{O}(L/\sqrt{\varepsilon})$.

Note that this reduction cannot be applied to the locally accelerated algorithm in ([Zhang and Sra](#), [2018](#)), that we discussed in the related work section. The reduction runs in stages by regularizing each time with a strongly g-convex regularizer whose parameter decreases exponentially until we use a regularizer with $O(\varepsilon)$ maximum function value. The local assumption required by the algorithm in ([Zhang and Sra](#), [2018](#)) on the closeness to the minimum cannot be guaranteed. In ([Ahn and Sra](#), [2020](#)), the authors give an unconstrained global algorithm whose rates are strictly better than RGD. The reduction could be applied to a constrained version of this algorithm to obtain a method for smooth and g-convex functions defined on manifolds of bounded sectional curvature and whose rates are strictly better than RGD.

4. Conclusion

In this work, we proposed an algorithm with the same rates as AGD, for the optimization of smooth and strongly g-convex functions, up to constants and log factors, while previous approaches essentially only reached this for a ball around the minimizer of radius $O((\mu/L)^{3/4})$. Our algorithm also applies to g-convex functions while previous accelerated algorithms did not apply. We focused on hyperbolic and spherical spaces, that have constant sectional curvature. The study of geometric properties for this is often employed to conclude that a space of bounded sectional curvature satisfies a property that is in between the ones for the cases of constant extremal sectional curvature.

Several previous algorithms have been developed for the general case by utilizing this philosophy, for instance (Ahn and Sra, 2020; Ferreira et al., 2019; Wang et al., 2015; Zhang and Sra, 2016, 2018). In future work, we will attempt to use the techniques and insights developed in this work to give an algorithm with the same rates as AGD for manifolds of bounded sectional curvature.

The key technique of our algorithm is the effective lower bound aggregation. Indeed, lower bound aggregation is the main hurdle to obtain accelerated first-order methods defined on Riemannian manifolds. Whereas the process of obtaining decreasing upper bounds on the function works similarly as in the Euclidean space—the same approach of locally minimizing the upper bound given by the smoothness assumption is used—obtaining adequate lower bounds proves to be a difficult task. We usually want a simple lower bound such that it, or a regularized version of it, can be easily optimized globally. We also want that the lower bound combines the knowledge that the g-convexity or strong g-convexity provides for all the queried points, commonly an average. These Riemannian convexity assumptions provide simple lower bounds, namely linear or quadratic, but each with respect to each of the tangent spaces of the queried points only. The deformations of the space complicate the aggregation of the lower bounds. Our work deals with this problem by finding appropriate lower bounds via the use of a geodesic map and takes into account the deformations incurred to derive a fully accelerated algorithm. We also used other tools for designing the accelerated algorithm. We worked with a relaxation of convexity that allowed to perform a binary search to reduce the discretization error. We had to use an implicit discretization of some accelerated continuous dynamics, since at least the vanilla application of usual approaches like Linear Coupling (Allen Zhu and Orecchia, 2017) or Nesterov's estimate sequence (Nesterov, 1983), that can be seen as a forward Euler discretization of the accelerated dynamics combined with a balancing gradient step (Diakonikolas and Orecchia, 2019), did not work in our constrained case. We interpret that the difficulty arises from trying to keep the gradient step inside the constraints while being able to compensate for a lower bound that is looser by a constant factor.

Acknowledgments

References

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

Foivos Alimisis, Antonio Orvieto, Gary Bécigneul, and Aurelien Lucchi. A continuous-time perspective for modeling acceleration in Riemannian optimization. *arXiv preprint arXiv:1910.10782*, 2019. URL *https://arxiv.org/abs/1910.10782*.

Foivos Alimisis, Antonio Orvieto, Gary Bécigneul, and Aurelien Lucchi. Practical accelerated optimization on Riemannian manifolds. *arXiv preprint arXiv:2002.04144*, 2020. URL *https://arxiv.org/abs/2002.04144*.

Zeyuan Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. *J. Mach. Learn. Res.*, 18:221:1–221:51, 2017. URL *http://jmlr.org/papers/v18/16-410.html*.

Zeyuan Allen-Zhu. Natasha 2: Faster non-convex optimization than SGD. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, pages 2680–2691, 2018a. URL *http://papers.nips.cc/paper/7533-natasha-2-faster-non-convex-optimization-than-sgd*.

Zeyuan Allen-Zhu. Katyusha X: practical momentum method for stochastic sum-of-nonconvex optimization. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 179–185, 2018b. URL *http://proceedings.mlr.press/v80/allen-zhu18a.html*.

Zeyuan Allen Zhu and Elad Hazan. Optimal black-box reductions between optimization objectives. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 1606–1614, 2016. URL *http://papers.nips.cc/paper/6364-optimal-black-box-reductions-between-optimization-objectives*.

Zeyuan Allen Zhu and Lorenzo Orecchia. Nearly-linear time positive LP solver with faster convergence rate. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC 2015, Portland, OR, USA, June 14-17, 2015*, pages 229–236, 2015. doi: 10.1145/2746539.2746573. URL *https://doi.org/10.1145/2746539.2746573*.

Zeyuan Allen Zhu and Lorenzo Orecchia. Linear coupling: An ultimate unification of gradient and mirror descent. In *8th Innovations in Theoretical Computer Science Conference, ITCS 2017, January 9-11, 2017, Berkeley, CA, USA*, pages 3:1–3:22, 2017. doi: 10.4230/LIPIcs.ITCS.2017.3. URL *https://doi.org/10.4230/LIPIcs.ITCS.2017.3*.

Zeyuan Allen Zhu, Zheng Qu, Peter Richtárik, and Yang Yuan. Even faster accelerated coordinate descent using non-uniform sampling. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, pages 1110–1119, 2016. URL *http://proceedings.mlr.press/v48/allen-zhuc16.html*.

Zeyuan Allen-Zhu, Yuanzhi Li, Rafael Mendes de Oliveira, and Avi Wigderson. Much faster algorithms for matrix scaling. In *58th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2017, Berkeley, CA, USA, October 15-17, 2017*, pages 890–901, 2017. doi: 10.1109/FOCS.2017.87. URL *https://doi.org/10.1109/FOCS.2017.87*.

Zeyuan Allen-Zhu, Ankit Garg, Yuanzhi Li, Rafael Oliveira, and Avi Wigderson. Operator scaling via geodesically convex optimization, invariant theory and polynomial identity testing. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 172–181, 2018. URL *https://arxiv.org/abs/1804.01076*.

Dimitri Bertsekas and Angelia Nedic. Convex analysis and optimization. 2003.

Herbert Busemann and Bhalchandra Phadke. A general version of Beltrami's theorem in the large. *Pacific Journal of Mathematics*, 115(2):299–315, 1984.

Léopold Cambier and Pierre-Antoine Absil. 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*.

Yair Carmon, John C. Duchi, Oliver Hinder, and Aaron Sidford. "Convex until proven guilty": Dimension-free acceleration of gradient descent on non-convex functions. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 654–663, 2017. URL *http://proceedings.mlr.press/v70/carmon17a.html*.

Anoop Cherian and Suvrit Sra. 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*.

Michael Cohen, Jelena Diakonikolas, and Lorenzo Orecchia. On acceleration with noise-corrupted gradients. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 1018–1027, 2018. URL *http://proceedings.mlr.press/v80/cohen18a.html*.

Chris Criscitiello and Nicolas Boumal. Efficiently escaping saddle points on manifolds. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 5985–5995, 2019. URL *http://papers.nips.cc/paper/8832-efficiently-escaping-saddle-points-on-manifolds*.

Chris Criscitiello and Nicolas Boumal. An accelerated first-order method for non-convex optimization on manifolds. *arXiv preprint arXiv:2008.02252*, 2020. URL *https://arxiv.org/abs/2008.02252*.

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

Ashok Cutkosky and Tamás Sarlós. Matrix-free preconditioning in online learning. In *Proceed-ings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pages 1455–1464, 2019. URL *http://proceedings.mlr.press/v97/cutkosky19b.html*.

Glaydston de Carvalho Bento, Orizon P. Ferreira, and Jefferson G. Melo. Iteration-complexity of gradient, subgradient and proximal point methods on Riemannian manifolds. *J. Optim. Theory Appl.*, 173(2):548–562, 2017. doi: 10.1007/s10957-017-1093-4. URL *https://doi.org/10.1007/s10957-017-1093-4*.

Jelena Diakonikolas and Michael I. Jordan. Generalized momentum-based methods: A hamiltonian perspective. *CoRR*, abs/1906.00436, 2019. URL *http://arxiv.org/abs/1906.00436*.

Jelena Diakonikolas and Lorenzo Orecchia. Accelerated extra-gradient descent: A novel accelerated first-order method. In *9th Innovations in Theoretical Computer Science Conference, ITCS 2018, January 11-14, 2018, Cambridge, MA, USA*, pages 23:1–23:19, 2018. doi: 10.4230/LIPIcs.ITCS. 2018.23. URL *https://doi.org/10.4230/LIPIcs.ITCS.2018.23*.

Jelena Diakonikolas and Lorenzo Orecchia. The approximate duality gap technique: A unified theory of first-order methods. *SIAM Journal on Optimization*, 29(1):660–689, 2019. doi: 10. 1137/18M1172314. URL *https://doi.org/10.1137/18M1172314*.

Alan Edelman, Tomás A. Arias, and Steven Thomas Smith. 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*.

OP Ferreira, MS Louzeiro, and LF Prudente. Gradient method for optimization on Riemannian man-ifolds with lower bounded curvature. *SIAM Journal on Optimization*, 29(4):2517–2541, 2019.

Alexander Gasnikov, Pavel E. Dvurechensky, Eduard A. Gorbunov, Evgeniya A. Vorontsova, Daniil Selikhanovych, César A. Uribe, Bo Jiang, Haoyue Wang, Shuzhong Zhang, Sébastien Bubeck, Qijia Jiang, Yin Tat Lee, Yuanzhi Li, and Aaron Sidford. Near optimal methods for minimizing convex functions with lipschitz p-th derivatives. In *Conference on Learning Theory, COLT 2019, 25-28 June 2019, Phoenix, AZ, USA*, pages 1392–1393, 2019. URL *http://proceedings.mlr.press/v99/gasnikov19b.html*.

Matthieu Genicot, Wen Huang, and Nickolay T. Trendafilov. Weakly correlated sparse com-ponents with nearly orthonormal loadings. In *Geometric Science of Information - Second International Conference, GSI 2015, Palaiseau, France, October 28-30, 2015, Proceedings*, pages 484–490, 2015. doi: 10.1007/978-3-319-25040-3_52. URL *https://doi.org/10.1007/978-3-319-25040-3_52*.

Marvin J Greenberg. *Euclidean and non-Euclidean geometries: Development and history*. Macmil-lan, 1993.

Karsten Grove, Peter Petersen, and Silvio Levy. *Comparison geometry*, volume 30. Cambridge University Press, 1997.

Sergey Guminov and Alexander Gasnikov. Accelerated methods for alpha-weakly-quasi-convex problems. *arXiv preprint arXiv:1710.00797*, 2017. URL *https://arxiv.org/abs/1710.00797*.

Linus Hamilton and Ankur Moitra. A no-go theorem for acceleration in the hyperbolic plane. *arXiv preprint arXiv:2101.05657*, 2021. URL *https://arxiv.org/abs/2101.05657*.

Gennadij Heidel and Volker Schulz. 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*.

Oliver Hinder, Aaron Sidford, and Nimit Sharad Sohoni. Near-optimal methods for minimizing star-convex functions and beyond. *CoRR*, abs/1906.11985, 2019. URL *http://arxiv.org/abs/1906.11985*.

Reshad Hosseini and Suvrit Sra. Matrix manifold optimization for gaussian mixtures. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 910–918, 2015. URL *http://papers.nips.cc/paper/5812-matrix-manifold-optimization-for-gaussian-mixtures*.

Reshad Hosseini and Suvrit Sra. 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*.

Wen Huang and Ke Wei. Extending FISTA to Riemannian optimization for sparse PCA. *arXiv preprint arXiv:1909.05485*, 2019a. URL *https://arxiv.org/abs/1909.05485*.

Wen Huang and Ke Wei. Riemannian proximal gradient methods. *arXiv preprint arXiv:1909.06065*, 2019b. URL *https://arxiv.org/abs/1909.06065*.

Ian T Jolliffe, Nickolay T Trendafilov, and Mudassir Uddin. A modified principal component technique based on the lasso. *Journal of computational and Graphical Statistics*, 12(3):531–547, 2003. URL *https://doi.org/10.1198/1061860032148*.

Hiroyuki Kasai, Pratik Jawanpuria, and Bamdev Mishra. Riemannian adaptive stochastic gradient algorithms on matrix manifolds. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pages 3262–3271, 2019. URL *http://proceedings.mlr.press/v97/kasai19a.html*.

Masoud Badiei Khuzani and Na Li. Stochastic primal-dual method on Riemannian manifolds of bounded sectional curvature. In *16th IEEE International Conference on Machine Learning and Applications, ICMLA 2017, Cancun, Mexico, December 18-21, 2017*, pages 133–140, 2017. doi: 10.1109/ICMLA.2017.0-167. URL *https://doi.org/10.1109/ICMLA.2017.0-167*.

E. Kreyszig. *Differential Geometry*. Differential Geometry. Dover Publications, 1991. ISBN 9780486667218. URL *https://books.google.es/books?id=P73DrhE9F0QC*.

Walid Krichene, Alexandre M. Bayen, and Peter L. Bartlett. Accelerated mirror descent in continuous and discrete time. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 2845–2853, 2015. URL *http://papers.nips.cc/paper/5843-accelerated-mirror-descent-in-continuous-and-discrete-time*.

Mario Lezcano-Casado. Trivializations for gradient-based optimization on manifolds. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 9154–9164, 2019. URL *http://papers.nips.cc/paper/9115-trivializations-for-gradient-based-optimization-on-manifolds*.

Mario Lezcano-Casado. Curvature-dependant global convergence rates for optimization on manifolds of bounded geometry. *arXiv preprint arXiv:2008.02517*, 2020. URL *https://arxiv.org/abs/2008.02517*.

Mario Lezcano-Casado and David Martínez-Rubio. Cheap orthogonal constraints in neural networks: A simple parametrization of the orthogonal and unitary group. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pages 3794–3803, 2019. URL *http://proceedings.mlr.press/v97/lezcano-casado19a.html*.

Lizhen Lin, Bayan Saparbayeva, Michael Minyi Zhang, and David B. Dunson. Accelerated algorithms for convex and non-convex optimization on manifolds. *CoRR*, abs/2010.08908, 2020. URL *https://arxiv.org/abs/2010.08908*.

Yuanyuan Liu, Fanhua Shang, James Cheng, Hong Cheng, and Licheng Jiao. Accelerated first-order methods for geodesically convex optimization on Riemannian manifolds. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 4868–4877, 2017. URL *https://papers.nips.cc/paper/2017/hash/6ef80bb237adf4b6f77d0700e1255907-Abstract.html*.

Bamdev Mishra and Rodolphe Sepulchre. R3MC: A Riemannian three-factor algorithm for low-rank matrix completion. In *53rd IEEE Conference on Decision and Control, CDC 2014, Los Angeles, CA, USA, December 15-17, 2014*, pages 1137–1142, 2014. doi: 10.1109/CDC.2014.7039534. URL *https://doi.org/10.1109/CDC.2014.7039534*.

Vaishnavh Nagarajan and J Zico Kolter. Generalization in deep networks: The role of distance from initialization. *arXiv preprint arXiv:1901.01672*, 2019. URL *https://arxiv.org/abs/1901.01672*.

Yurii Nesterov. A method of solving a convex programming problem with convergence rate o(1/k2). In *Soviet Mathematics Doklady*, volume 27, pages 372–376, 1983.

Yurii Nesterov, Alexander Gasnikov, Sergey Guminov, and Pavel Dvurechensky. Primal-dual accelerated gradient descent with line search for convex and nonconvex optimization problems. *arXiv preprint arXiv:1809.05895*, 2018. URL *https://arxiv.org/abs/1809.05895*.

Peter Petersen, S Axler, and KA Ribet. *Riemannian geometry*, volume 171. Springer, 2006.

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

Hiroyuki Sato, Hiroyuki Kasai, and Bamdev Mishra. 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*.

Weijie Su, Stephen P. Boyd, and Emmanuel J. Candès. A differential equation for modeling nesterov's accelerated gradient method: Theory and insights. *J. Mach. Learn. Res.*, 17:153:1–153:43, 2016. URL *http://jmlr.org/papers/v17/15-084.html*.

Ju Sun, Qing Qu, and John Wright. 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*.

Yue Sun, Nicolas Flammarion, and Maryam Fazel. Escaping from saddle points on Riemannian manifolds. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 7274–7284, 2019. URL *http://papers.nips.cc/paper/8948-escaping-from-saddle-points-on-riemannian-manifolds*.

Ilya Sutskever, James Martens, George E. Dahl, and Geoffrey E. Hinton. On the importance of initialization and momentum in deep learning. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, pages 1139–1147, 2013. URL *http://proceedings.mlr.press/v28/sutskever13.html*.

Mingkui Tan, Ivor W. Tsang, Li Wang, Bart Vandereycken, and Sinno Jialin Pan. Riemannian pursuit for big matrix recovery. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pages 1539–1547, 2014. URL *http://proceedings.mlr.press/v32/tan14.html*.

Nilesh Tripuraneni, Nicolas Flammarion, Francis Bach, and Michael I. Jordan. Averaging stochastic gradient descent on Riemannian manifolds. *CoRR*, abs/1802.09128, 2018. URL *http://arxiv.org/abs/1802.09128*.

Bart Vandereycken. 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*.

Di Wang, Satish Rao, and Michael W. Mahoney. Unified acceleration method for packing and covering problems via diameter reduction. In *43rd International Colloquium on Automata, Languages, and Programming, ICALP 2016, July 11-15, 2016, Rome, Italy*, pages 50:1–50:13, 2016. doi: 10.4230/LIPIcs.ICALP.2016.50. URL *https://doi.org/10.4230/LIPIcs.ICALP.2016.50*.

X. M. Wang, C. Li, and J. C. Yao. Subgradient projection algorithms for convex feasibility on Riemannian manifolds with lower bounded curvatures. *J. Optim. Theory Appl.*, 164(1):202–217, 2015. doi: 10.1007/s10957-014-0568-9. URL *https://doi.org/10.1007/s10957-014-0568-9*.

Melanie Weber and Suvrit Sra. 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*.

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

Ke Wei, Jian-Feng Cai, Tony F Chan, and Shingyu Leung. Guarantees of Riemannian optimization for low rank matrix completion. *arXiv preprint arXiv:1603.06610*, 2016. URL *https://arxiv.org/abs/1603.06610*.

Andre Wibisono, Ashia C. Wilson, and Michael I. Jordan. A variational perspective on accelerated methods in optimization. *CoRR*, abs/1603.04245, 2016. URL *http://arxiv.org/abs/1603.04245*.

Ami Wiesel. Geodesic convexity and covariance estimation. *IEEE Trans. Signal Process.*, 60(12): 6182–6189, 2012. doi: 10.1109/TSP.2012.2218241. URL *https://doi.org/10.1109/TSP.2012.2218241*.

Hongyi Zhang and Suvrit Sra. First-order methods for geodesically convex optimization. In *Proceedings of the 29th Conference on Learning Theory, COLT 2016, New York, USA, June 23-26, 2016*, pages 1617–1638, 2016. URL *http://proceedings.mlr.press/v49/zhang16b.html*.

Hongyi Zhang and Suvrit Sra. An estimate sequence for geodesically convex optimization. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018*, volume 75 of *Proceedings of Machine Learning Research*, pages 1703–1723. PMLR, 2018. URL *http://proceedings.mlr.press/v75/zhang18a.html*.

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

Jingzhao Zhang, Hongyi Zhang, and Suvrit Sra. R-SPIDER: A fast Riemannian stochastic optimization algorithm with curvature independent rate. *CoRR*, abs/1811.04194, 2018. URL *http://arxiv.org/abs/1811.04194*.

Pan Zhou, Xiao-Tong Yuan, and Jiashi Feng. Faster first-order methods for stochastic non-convex optimization on Riemannian manifolds. In *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, pages 138–147, 2019. URL *http://proceedings.mlr.press/v89/zhou19a.html*.

Xingyu Zhou. On the fenchel duality between strong convexity and lipschitz continuous gradient. *arXiv preprint arXiv:1803.06573*, 2018. URL *https://arxiv.org/abs/1803.06573*.

We divide the appendix in four sections. Appendix A contains the proofs related to the accelerated algorithm, i.e. the proofs of Theorems 5 and 6. In Appendix B we prove the results related to the reductions in Section 3. In Appendix C, we prove the geometric lemmas that take into account the geodesic map h to obtain relationships between F and f, namely Lemmas 2, 3 and 4. Finally, Appendix D contains a discussion on the constants of our algorithms, on rates of related work and on hardness results. We also show a lower bound on the condition number for any strongly g-convex function defined on $\mathcal{B}_R$.

Appendix A. Acceleration. Proofs of Theorem 5 and Theorem 6

Diakonikolas and Orecchia (2019) developed the *approximate duality gap technique* which is a technique that provides a structure to design and prove first order methods and their guarantees for the optimization of convex problems. We take inspiration from these ideas to apply them to the non-convex problem we have at hand Theorem 5, as it was sketched in Section 2.1. We start with two basic definitions.

Definition 10 *Given two points $\tilde{x}, \tilde{y}$, we define the Bregman divergence with respect to $\psi(\cdot)$ as*

$$D_\psi(\tilde{x}, \tilde{y}) \stackrel{\text{def}}{=} \psi(\tilde{x}) - \psi(\tilde{y}) - \langle \nabla \psi(\tilde{y}), \tilde{x} - \tilde{y} \rangle.$$

Definition 11 *Given a closed convex set Q and a function $\psi : Q \to \mathbb{R}$, we define the convex conjugate of ψ, also known as its Fenchel dual, as the function*

$$\psi^*(\tilde{z}) = \max_{\tilde{x} \in Q} \{\langle \tilde{z}, \tilde{x} \rangle - \psi(\tilde{x})\}.$$

For simplicity, we will use $\psi(\tilde{x}) = \frac{1}{2}\|\tilde{x}\|^2 + i_Q(\tilde{x})$ in Algorithm 1, but any strongly convex map works. Here $i_Q(x) = 0$ if $x \in Q$ and $i_Q(x) = +\infty$ otherwise. The gradient of the Fenchel dual of $\psi(\cdot)$ is $\nabla \psi^*(\tilde{z}) = \arg\min_{\tilde{z}' \in Q}\{\|\tilde{z}' - \tilde{z}\|\}$, that is, the Euclidean projection $\Pi_Q(\tilde{z})$ of the point $\tilde{z}$ onto Q. Note that when we apply Theorem 5 to Theorem 6 our constraint Q will be $\mathcal{X}$, that is, a ball centered at 0 of radius $\tilde{R}$, so the projection of a point $\tilde{z}$ outside of $\mathcal{X}$ will be the vector normalization $\tilde{R}\tilde{z}/\|\tilde{z}\|$. Any continuously differentiable strongly convex ψ would work, provided that $\nabla \psi^*(z)$ is easily computable, preferably in closed form. Note that by the Fenchel-Moreau theorem we have for any such map that $\psi^{**} = \psi$.

We recall we assume that f satisfies tilted-convexity (3):

$$f(\tilde{x}) + \frac{1}{\gamma_{\mathrm{n}}}\langle \nabla f(\tilde{x}), \tilde{y} - \tilde{x} \rangle \le f(\tilde{y}) \quad \text{if } \langle \nabla f(\tilde{x}), \tilde{y} - \tilde{x} \rangle \le 0,$$

$$f(\tilde{x}) + \gamma_{\mathrm{p}}\langle \nabla f(\tilde{x}), \tilde{y} - \tilde{x} \rangle \le f(\tilde{y}) \quad \text{if } \langle \nabla f(\tilde{x}), \tilde{y} - \tilde{x} \rangle \ge 0.$$

Let α_t be an increasing function of time t. We use Lebesgue-Stieltjes integration and its notation, so that $\int_0^t f(x_\tau)\dot{\alpha}_\tau d\tau = \int_0^t f(x_\tau)d\alpha_\tau$. We want to work with continuous and discrete approaches in a unified way. Thus, when α_t is a discrete measure, we have that $\dot{\alpha}_t = \sum_{i=1}^{\infty} a_i\delta(t - (t_0 + i - 1))$ is a weighted sum of Dirac delta functions. We define $A_t \stackrel{\text{def}}{=} \int_{t_0}^t d\alpha_\tau = \int_{t_0}^t \dot{\alpha}_\tau d\tau$. In discrete time, it is $A_t = \sum_{i=1}^{\lfloor t-t_0+1 \rfloor} a_i = \alpha_t$. In the continuous case note that we have $\alpha_t - A_t = \alpha_{t_0}$.

We start defining a continuous method that we discretize with an approximate implementation of the implicit Euler method. Let $\tilde{x}_t$ be the solution obtained by the algorithm at time t. We define

the duality gap $G_t \stackrel{\text{def}}{=} U_t - L_t$ as the difference between a differentiable upper bound U_t on the function at the current point and a lower bound on $f(\tilde{x}^*)$. Since in our case f is differentiable we use $U_t \stackrel{\text{def}}{=} f(\tilde{x}_t)$. The idea is to enforce the invariant $\frac{d}{dt}(\alpha_t G_t) = 0$, so we have at any time $f(\tilde{x}_t) - f(\tilde{x}^*) \leq G_t = G_{t_0}\alpha_{t_0}/\alpha_t$.

Note that for a global minimum $\tilde{x}^*$ of f and any other point $\tilde{x} \in Q$, we have $\langle \nabla f(\tilde{x}), \tilde{x}^* - \tilde{x} \rangle \leq 0$. Otherwise, we would obtain a contradiction since by tilted-convexity (3) we would have

$$f(\tilde{x}) < f(\tilde{x}) + \gamma_{\mathrm{p}}\langle \nabla f(\tilde{x}), \tilde{x}^* - \tilde{x} \rangle \leq f(\tilde{x}^*).$$

Therefore, in order to define an appropriate lower bound, we will make use of the inequality $f(\tilde{x}^*) \geq f(\tilde{x}) + \frac{1}{\gamma_{\mathrm{n}}}\langle \nabla f(\tilde{x}), \tilde{x}^* - \tilde{x} \rangle$, for any $\tilde{x} \in Q$, which holds true by tilted-convexity (3), for $\tilde{y} = \tilde{x}^*$. Combining this inequality for all the points visited by the continuous method we have

$$f(\tilde{x}^*) \geq \frac{\int_{t_0}^t f(\tilde{x}_\tau)d\alpha_\tau}{A_t} + \frac{\int_{t_0}^t \frac{1}{\gamma_{\mathrm{n}}}\langle \nabla f(\tilde{x}_\tau), \tilde{x}^* - \tilde{x}_\tau \rangle d\alpha_\tau}{A_t}.$$

We cannot compute this lower bound, since the right hand side depends on the unknown point $\tilde{x}^*$. We could compute a looser lower bound by taking the minimum over $\tilde{u} \in Q$ of this expression, substituting $\tilde{x}^*$ by $\tilde{u}$. However, this would make the lower bound be non-differentiable and we could have problems at t_0. In order to solve the first problem, we first add a regularizer and then take the minimum over $\tilde{u} \in Q$.

$$f(\tilde{x}^*) + \frac{D_\psi(\tilde{x}^*, \tilde{x}_{t_0})}{A_t}$$
$$\geq \frac{\int_{t_0}^t f(\tilde{x}_\tau)d\alpha_\tau}{A_t} + \frac{\min_{\tilde{u} \in Q}\left\{\int_{t_0}^t \frac{1}{\gamma_{\mathrm{n}}}\langle \nabla f(\tilde{x}_\tau), \tilde{u} - \tilde{x}_\tau \rangle d\alpha_\tau + D_\psi(\tilde{u}, \tilde{x}_{t_0})\right\}}{A_t}$$

In order to solve the second problem, we mix this lower bound with the optimal lower bound $f(\tilde{x}^*)$ with weight $\alpha_t - A_t$ (this is only necessary in continuous time, in discrete time this term is 0). Not knowing $f(\tilde{x}^*)$ or $D_\psi(\tilde{x}^*, \tilde{x}_{t_0})$ will not be problematic. Indeed, we only need to guarantee $\frac{d}{dt}(\alpha_t G_t) = 0$. After taking the derivative, these terms will vanish. After rescaling the normalization factor, we finally obtain the lower bound

$$f(\tilde{x}^*) \geq L_t \stackrel{\text{def}}{=} \frac{\int_{t_0}^t f(\tilde{x}_\tau)d\alpha_\tau}{\alpha_t} + \frac{\min_{\tilde{u} \in Q}\left\{\int_{t_0}^t \langle \frac{1}{\gamma_{\mathrm{n}}}\nabla f(\tilde{x}_\tau), \tilde{u} - \tilde{x}_\tau \rangle d\alpha_\tau + D_\psi(\tilde{u}, \tilde{x}_{t_0})\right\}}{\alpha_t}$$
$$+ \frac{(\alpha_t - A_t)f(\tilde{x}^*) - D_\psi(\tilde{x}^*, \tilde{x}_{t_0})}{\alpha_t}. \tag{8}$$

Let $\tilde{z}_t = \nabla \psi(\tilde{x}_{t_0}) - \int_{t_0}^t \frac{1}{\gamma_{\mathrm{n}}}\nabla f(\tilde{x}_\tau)d\alpha_\tau$. Then, by Fact 16, we can compute the optimum $\tilde{u} \in Q$ above as

$$\nabla \psi^*(\tilde{z}_t) = \arg\min_{\tilde{u} \in Q}\left\{\int_{t_0}^t \langle \frac{1}{\gamma_{\mathrm{n}}}\nabla f(\tilde{x}_\tau), \tilde{u} - \tilde{x}_\tau \rangle d\alpha_\tau + D_\psi(\tilde{u}, \tilde{x}_{t_0})\right\}. \tag{9}$$

Recalling $U_t = f(\tilde{x}_t)$ and using (8) and (9) we obtain:

$$\frac{d}{dt}(\alpha_t G_t) = \frac{d}{dt}(\alpha_t f(\tilde{x}_t)) - \dot{\alpha}_t f(\tilde{x}_t) - \dot{\alpha}_t \frac{1}{\gamma_{\mathrm{n}}}\langle \nabla f(\tilde{x}_t), \nabla \psi^*(\tilde{z}_t) - \tilde{x}_t \rangle$$
$$= \frac{1}{\gamma_{\mathrm{n}}}\langle \nabla f(\tilde{x}_t), \gamma_{\mathrm{n}}\alpha_t \dot{\tilde{x}}_t - \dot{\alpha}_t(\nabla \psi^*(\tilde{z}_t) - \tilde{x}_t)\rangle.$$

Thus, to satisfy the invariant $\frac{d}{dt}(\alpha_t G_t) = 0$, it is enough to set $\gamma_{\mathrm{n}} \alpha_t \dot{\tilde{x}}_t = \dot{\alpha}_t(\nabla \psi^*(\tilde{z}_t) - \tilde{x}_t)$, yielding the following continuous accelerated dynamics

$$
\begin{aligned}
\dot{\tilde{z}}_t &= -\frac{1}{\gamma_{\mathrm{n}}} \dot{\alpha}_t \nabla f(\tilde{x}_t), \\
\dot{\tilde{x}}_t &= \frac{1}{\gamma_{\mathrm{n}}} \dot{\alpha}_t \frac{\nabla \psi^*(\tilde{z}_t) - \tilde{x}_t}{\alpha_t}, \\
\tilde{z}_{t_0} &= \nabla \psi(\tilde{x}_{t_0}), \\
\tilde{x}_{t_0} &\in Q \text{ is an arbitrary initial point.}
\end{aligned}
\tag{10}
$$

Now we proceed to discretize the dynamics, so from now on we will use a discrete measure α_t, as we described above. We set t_0 to 1. Let $E_{i+1} \overset{\text{def}}{=} A_{i+1} G_{i+1} - A_i G_i$ be the discretization error. Then we have

$$
G_t = \frac{A_1}{A_t} G_1 + \frac{\sum_{i=1}^{t-1} E_{i+1}}{A_t}.
$$

Lemma 12 *If we have*

$$
f(\tilde{x}_{i+1}) - f(\tilde{x}_i) \leq \hat{\gamma}_i \langle \nabla f(\tilde{x}_{i+1}), \tilde{x}_{i+1} - \tilde{x}_i \rangle + \hat{\varepsilon}_i,
\tag{11}
$$

for some $\hat{\gamma}_i, \hat{\varepsilon}_i \geq 0$, then the discretization error satisfies

$$
E_{i+1} \leq \langle \nabla f(\tilde{x}_{i+1}), (A_i \hat{\gamma}_i + \frac{a_{i+1}}{\gamma_{\mathrm{n}}}) \tilde{x}_{i+1} - \hat{\gamma}_i A_i \tilde{x}_i - \frac{a_{i+1}}{\gamma_{\mathrm{n}}} \nabla \psi^*(\tilde{z}_{i+1})) \rangle - D_{\psi^*}(\tilde{z}_i, \tilde{z}_{i+1}) + A_i \hat{\varepsilon}_i.
$$

Proof In a similar way to (Diakonikolas and Orecchia, 2018), we could compute the discretization error as the difference between the gap and the gap computed allowing continuous integration rules in the integrals that it contains. However, we will directly bound E_{i+1} as $A_{i+1} G_{i+1} - A_i G_i$ instead. Recall that in discrete time we have $\alpha_i = A_i$ so the definition of the lower bound in discrete time becomes the following, by combining (8) and (9):

$$
L_i = \sum_{j=1}^{i} a_j f(\tilde{x}_j) + \sum_{j=1}^{i} \langle \frac{a_j}{\gamma_{\mathrm{n}}} \nabla f(\tilde{x}_j), \nabla \psi^*(\tilde{z}_i) - \tilde{x}_j \rangle + D_\psi(\nabla \psi^*(\tilde{z}_i), \tilde{x}_{t_0}) - D_\psi(\tilde{x}^*, \tilde{x}_{t_0}).
$$

Hence, using the definition of G_i, U_i, L_i we have

$$A_{i+1}G_{i+1} - A_iG_i$$
$$= (A_{i+1}f(\tilde{x}_{i+1}) - A_if(\tilde{x}_i)) - A_{i+1}L_{i+1} + A_iL_i$$
$$\overset{①}{=} (A_if(\tilde{x}_{i+1}) - A_if(\tilde{x}_i) + a_{i+1}f(\tilde{x}_{i+1}))$$
$$- \sum_{j=1}^{i+1} a_j f(\tilde{x}_j) - \sum_{j=1}^{i+1} \frac{a_j}{\gamma_{\mathrm{n}}} \langle \nabla f(\tilde{x}_j), \nabla\psi^*(\tilde{z}_{i+1}) - \tilde{x}_j \rangle - D_\psi(\nabla\psi^*(\tilde{z}_{i+1}), \tilde{x}_{t_0})$$
$$+ \sum_{j=1}^{i} a_j f(\tilde{x}_j) + \sum_{j=1}^{i} \frac{a_j}{\gamma_{\mathrm{n}}} \langle \nabla f(\tilde{x}_j), \nabla\psi^*(\tilde{z}_i) - \tilde{x}_j \rangle + D_\psi(\nabla\psi^*(\tilde{z}_i), \tilde{x}_{t_0})$$
$$\overset{②}{=} A_i(f(\tilde{x}_{i+1}) - f(\tilde{x}_i)) - \langle \frac{a_{i+1}}{\gamma_{\mathrm{n}}} \nabla f(\tilde{x}_{i+1}), \nabla\psi^*(\tilde{z}_{i+1}) - \tilde{x}_{i+1} \rangle$$
$$+ \sum_{j=1}^{i} \langle \frac{a_j}{\gamma_{\mathrm{n}}} \nabla f(\tilde{x}_j), \nabla\psi^*(\tilde{z}_i) - \nabla\psi^*(\tilde{z}_{i+1}) \rangle$$
$$[-\langle \nabla\psi(\tilde{x}_{t_0}), \nabla\psi^*(\tilde{z}_i) - \nabla\psi^*(\tilde{z}_{i+1}) \rangle + \psi(\nabla\psi^*(\tilde{z}_i)) - \psi(\nabla\psi^*(\tilde{z}_{i+1}))]$$
$$\overset{③}{=} A_i(f(\tilde{x}_{i+1}) - f(\tilde{x}_i)) - \langle \frac{a_{i+1}}{\gamma_{\mathrm{n}}} \nabla f(\tilde{x}_{i+1}), \nabla\psi^*(\tilde{z}_{i+1}) - \tilde{x}_{i+1} \rangle - D_{\psi^*}(\tilde{z}_i, \tilde{z}_{i+1})$$
$$\overset{④}{\leq} \langle \nabla f(\tilde{x}_{i+1}), (A_i\hat{\gamma}_i + \frac{a_{i+1}}{\gamma_{\mathrm{n}}})\tilde{x}_{i+1} - \hat{\gamma}_i A_i \tilde{x}_i - \frac{a_{i+1}}{\gamma_{\mathrm{n}}} \nabla\psi^*(\tilde{z}_{i+1}) \rangle - D_{\psi^*}(\tilde{z}_i, \tilde{z}_{i+1}) + A_i\hat{\varepsilon}_i.$$

In ① we write down the definitions of L_{i+1} and L_i and split the first summand so it is clear that in ② we cancel all the $a_j f(\tilde{x}_j)$. In ② we also cancel some terms involved in the inner products, we write the definitions of the Bregman divergences and cancel some of their terms. For equality ③, we recall $\tilde{z}_i = \nabla\psi(\tilde{x}_{t_0}) - \sum_{j=1}^{i} \frac{a_j}{\gamma_{\mathrm{n}}} \nabla f(\tilde{x}_j)$ so we use this fact and $\psi^*(\tilde{z}) = \langle \nabla\psi^*(\tilde{z}), \tilde{z} \rangle - \psi(\nabla\psi^*(\tilde{z}))$ (which holds by Fact 16) for $\tilde{z} = \tilde{z}_i$ and $\tilde{z} = \tilde{z}_{i+1}$ to conclude that the last two lines equal $-D_{\psi^*}(\tilde{z}_i, \tilde{z}_{i+1})$. Inequality ④ uses (11). ∎

We show now how to cancel out the discretization error by an approximate implementation of implicit Euler discretization of (10). Note that we need to take into account the tilted-convexity assumption (3) instead of the usual convexity assumption. According to the previous lemma, we can set $\tilde{x}_{i+1}$ so that the right hand side of the inner product in the bound of E_{i+1} is 0. Assume for the moment, that the point $\tilde{x}_{i+1}$ we are going to compute satisfies the assumption of the previous lemma for some $\hat{\gamma}_i \in [\gamma_{\mathrm{p}}, 1/\gamma_{\mathrm{n}}]$. Thus, the implicit equation that defines the ideal method we would like to have is

$$\tilde{x}_{i+1} = \frac{\hat{\gamma}_i A_i}{A_i\hat{\gamma}_i + a_{i+1}/\gamma_{\mathrm{n}}} \tilde{x}_i + \frac{a_{i+1}/\gamma_{\mathrm{n}}}{A_i\hat{\gamma}_i + a_{i+1}/\gamma_{\mathrm{n}}} \nabla\psi^*(\tilde{z}_i - \frac{a_{i+1}}{\gamma_{\mathrm{n}}} \nabla f(\tilde{x}_{i+1})).$$

Note that $\tilde{x}_{i+1}$ is a convex combination of the other two points so it stays in Q. Indeed, the initial point is in Q and by (9) we have that $\nabla\psi^*(\tilde{z}_j) \in Q$ for all $j \geq 0$. However this method is implicit and possibly computationally expensive to implement. Nonetheless, two steps of a fixed point iteration procedure of this equation will be enough to have discretization error that is bounded

by the term $A_i \hat{\varepsilon}_i$: the last term of our bound. The error in the bound of E_{i+1} that the inner product incurs is compensated by the Bregman divergence term. In such a case, the equations of this method become, for $i \geq 0$:

$$\begin{cases} \tilde{\chi}_i = \frac{\hat{\gamma}_i A_i}{A_i \hat{\gamma}_i + a_{i+1}/\gamma_{\rm n}} \tilde{x}_i + \frac{a_{i+1}/\gamma_{\rm n}}{A_i \hat{\gamma}_i + a_{i+1}/\gamma_{\rm n}} \nabla \psi^*(\tilde{z}_i) \\ \tilde{\zeta}_i = \tilde{z}_i - \frac{a_{i+1}}{\gamma_{\rm n}} \nabla f(\tilde{\chi}_i) \\ \tilde{x}_{i+1} = \frac{\hat{\gamma}_i A_i}{A_i \hat{\gamma}_i + a_{i+1}/\gamma_{\rm n}} \tilde{x}_i + \frac{a_{i+1}/\gamma_{\rm n}}{A_i \hat{\gamma}_i + a_{i+1}/\gamma_{\rm n}} \nabla \psi^*(\tilde{\zeta}_i) \\ \tilde{z}_{i+1} = \tilde{z}_i - \frac{a_{i+1}}{\gamma_{\rm n}} \nabla f(\tilde{x}_{i+1}) \end{cases} \quad (12)$$

We prove now that this indeed leads to an accelerated algorithm. After this, we will show that we can perform a binary search at each iteration, to ensure that even if we do not know $\tilde{x}_{i+1}$ a priori, we can compute a $\hat{\gamma}_i \in [\gamma_{\rm p}, 1/\gamma_{\rm n}]$ satisfying assumption (11). This will only add a log factor to the overall complexity.

Lemma 13 *Consider the method given in* (12), *starting from an arbitrary point* $\tilde{x}_0 \in Q$ *with* $\tilde{z}_0 = \nabla \psi(\tilde{x}_0)$ *and* $A_0 = 0$. *Assume we can compute* $\hat{\gamma}_i$ *such that* $\tilde{x}_{i+1}$ *satisfies* (11). *Then, the error from Lemma 12 is bounded by*

$$E_{i+1} \leq \frac{a_{i+1}}{\gamma_{\rm n}} \langle \nabla f(\tilde{x}_{i+1}) - \nabla f(\tilde{\chi}_i), \nabla \psi^*(\tilde{\zeta}_i) - \nabla \psi^*(\tilde{z}_{i+1}) \rangle - D_{\psi^*}(\tilde{\zeta}_i, \tilde{z}_{i+1}) - D_{\psi^*}(\tilde{z}_i, \tilde{\zeta}_i) + A_i \hat{\varepsilon}_i.$$

Proof Using Lemma 12 and the third line of (12) we have

$$E_{i+1} - A_i \hat{\varepsilon}_i \leq \frac{a_{i+1}}{\gamma_{\rm n}} \langle \nabla f(\tilde{x}_{i+1}), \nabla \psi^*(\tilde{\zeta}_i) - \nabla \psi^*(\tilde{z}_{i+1}) \rangle - D_{\psi^*}(\tilde{z}_i, \tilde{z}_{i+1})$$

$$\leq \frac{a_{i+1}}{\gamma_{\rm n}} \langle \nabla f(\tilde{x}_{i+1}) - \nabla f(\tilde{\chi}_i) + \nabla f(\tilde{\chi}_i), \nabla \psi^*(\tilde{\zeta}_i) - \nabla \psi^*(\tilde{z}_{i+1}) \rangle - D_{\psi^*}(\tilde{z}_i, \tilde{z}_{i+1})$$

By the definition of $\tilde{\zeta}_i$ we have $(a_{i+1}/\gamma_{\rm n})\nabla f(\tilde{\chi}_i) = \tilde{z}_i - \tilde{\zeta}_i$. Using this fact and the triangle equality of Bregman divergences Lemma 17, we obtain

$$\frac{a_{i+1}}{\gamma_{\rm n}} \langle \nabla f(\tilde{\chi}_i), \nabla \psi^*(\tilde{\zeta}_i) - \nabla \psi^*(\tilde{z}_{i+1}) \rangle = \langle \tilde{z}_i - \tilde{\zeta}_i, \nabla \psi^*(\tilde{\zeta}_i) - \nabla \psi^*(\tilde{z}_{i+1}) \rangle$$

$$= D_{\psi^*}(\tilde{z}_i, \tilde{z}_{i+1}) - D_{\psi^*}(\tilde{\zeta}_i, \tilde{z}_{i+1}) - D_{\psi^*}(\tilde{z}_i, \tilde{\zeta}_i).$$

The lemma follows after combining these two equations. ∎

Theorem 14 *Let* Q *be a closed convex set of diameter* D. *Let* $f : Q \to \mathbb{R}$ *be an* $\tilde{L}$-*smooth tilted-convex function with constants* $\gamma_{\rm n}, \gamma_{\rm p}$. *Assume there is a point* $\tilde{x}^* \in Q$ *such that* $\nabla f(\tilde{x}^*) = 0$. *Let* $\psi : Q \to \mathbb{R}$ *be a* σ-*strongly convex map. Let* $\tilde{x}_i, \tilde{z}_i, \tilde{\chi}_i, \tilde{\zeta}_i$ *be updated according to* (12), *for* $i \geq 0$ *starting from an arbitrary initial point* $\tilde{x}_0 \in Q$ *with* $\tilde{z}_0 = \nabla \psi(\tilde{x}_0)$ *and* $A_0 = 0$, *assuming we can find* $\hat{\gamma}_i$ *at each iteration satisfying* (11). *If* $\tilde{L} a_{i+1}^2/\gamma_{\rm n}\sigma \leq a_{i+1} + A_i \gamma_{\rm n} \gamma_{\rm p}$, *then for all* $T \geq 1$ *we have*

$$f(\tilde{x}_T) - f(\tilde{x}^*) \leq \frac{D_\psi(\tilde{x}^*, \tilde{\chi}_0)}{A_T} + \sum_{i=1}^{T-1} \frac{A_i \hat{\varepsilon}_i}{A_T}.$$

In particular, if $a_i = \frac{i}{2}\frac{\sigma}{\tilde{L}}\gamma_{\mathrm{n}}^2\gamma_{\mathrm{p}}$, $\psi(\tilde{x}) = \frac{\sigma}{2}\|\tilde{x}\|^2$, $\hat{\varepsilon}_i = \frac{A_T\varepsilon}{2(T-1)A_i}$ *and* $T = \left\lceil \sqrt{4\tilde{L}\|\tilde{x}_0 - \tilde{x}^*\|^2/(\gamma_{\mathrm{n}}^2\gamma_{\mathrm{p}}\varepsilon)} \right\rceil = O(\sqrt{\tilde{L}/(\gamma_{\mathrm{n}}^2\gamma_{\mathrm{p}}\varepsilon)})$ *then*

$$f(\tilde{x}_T) - f(\tilde{x}^*) < \varepsilon.$$

Proof We bound the right hand side of the discretization error given by Lemma 13. Define $a = \|\nabla\psi^*(\tilde{\zeta}_i) - \nabla\psi^*(\tilde{z}_{i+1})\|$ and $b = \|\nabla\psi^*(\tilde{\zeta}_i) - \nabla\psi^*(\tilde{z}_i)\|$. We have

$$E_{i+1} - A_i\hat{\varepsilon}_i \overset{①}{\leq} \frac{a_{i+1}}{\gamma_{\mathrm{n}}}\langle\nabla f(\tilde{x}_{i+1}) - \nabla f(\tilde{\chi}_i), \nabla\psi^*(\tilde{\zeta}_i) - \nabla\psi^*(\tilde{z}_{i+1})\rangle - D_{\psi^*}(\tilde{\zeta}_i, \tilde{z}_{i+1}) - D_{\psi^*}(\tilde{z}_i, \tilde{\zeta}_i)$$

$$\overset{②}{\leq} \frac{a_{i+1}}{\gamma_{\mathrm{n}}}\tilde{L}\|\tilde{x}_{i+1} - \tilde{\chi}_i\| \cdot a - D_{\psi^*}(\tilde{\zeta}_i, \tilde{z}_{i+1}) - D_{\psi^*}(\tilde{z}_i, \tilde{\zeta}_i)$$

$$\overset{③}{\leq} \frac{a_{i+1}}{\gamma_{\mathrm{n}}}\tilde{L}\|\tilde{x}_{i+1} - \tilde{\chi}_i\| \cdot a - \frac{\sigma}{2}(a^2 + b^2)$$

$$\overset{④}{\leq} \frac{a_{i+1}^2/\gamma_{\mathrm{n}}^2}{A_i\hat{\gamma}_i + a_{i+1}/\gamma_{\mathrm{n}}}\tilde{L} \cdot ab - \frac{\sigma}{2}(a^2 + b^2)$$

$$\overset{⑤}{\leq} ab\left(\frac{a_{i+1}^2/\gamma_{\mathrm{n}}^2}{A_i\hat{\gamma}_i + a_{i+1}/\gamma_{\mathrm{n}}}\tilde{L} - \sigma\right).$$

Here ① follows from Lemma 13, ② uses the Cauchy-Schwartz inequality and gradient Lipschitzness, which is equivalent to smoothness for differentiable tilted convex functions, as we point out in the proof of Lemma 4. In ③, we used Lemma 18, and ④ uses the fact that by the definition of the method (12) we have $\tilde{x}_{i+1} - \tilde{\chi}_i = \frac{a_{i+1}/\gamma_{\mathrm{n}}}{A_i\hat{\gamma}_i + a_{i+1}/\gamma_{\mathrm{n}}}(\nabla\psi^*(\tilde{\zeta}_i) - \nabla\psi^*(\tilde{z}_i))$. Finally ⑤ uses $-(a^2 + b^2) \leq -2ab$, which comes from $(a - b)^2 \geq 0$. By the previous inequality, if we want $E_{i+1} \leq A_i\hat{\varepsilon}_i$, it is enough to guarantee the right hand side of the last expression is ≤ 0 which is implied by

$$\frac{\tilde{L}}{\sigma\gamma_{\mathrm{n}}}a_{i+1}^2 \leq a_{i+1} + A_i\gamma_{\mathrm{n}}\gamma_{\mathrm{p}}, \tag{13}$$

since $\gamma_{\mathrm{p}} \leq \hat{\gamma}_i$. And this is the assumption we made in the theorem. By inspection, if we use the value in the second part of the statement of the theorem $a_i = \frac{i}{2} \cdot \frac{\sigma}{\tilde{L}} \cdot \gamma_{\mathrm{n}}^2\gamma_{\mathrm{p}}$ into the previous inequality and noting that $A_i = \frac{i(i+1)}{4} \cdot \frac{\sigma}{\tilde{L}} \cdot \gamma_{\mathrm{n}}^2\gamma_{\mathrm{p}}$ we prove that the previous inequality is satisfied:

$$\frac{\tilde{L}}{\sigma\gamma_{\mathrm{n}}}a_{i+1}^2 = \frac{(i+1)^2}{4} \cdot \frac{\sigma}{\tilde{L}} \cdot \gamma_{\mathrm{n}}^3\gamma_{\mathrm{p}}^2$$

$$\leq \left(\frac{i+1}{2} + \frac{i(i+1)}{4}\right)\frac{\sigma}{\tilde{L}} \cdot \gamma_{\mathrm{n}}^3\gamma_{\mathrm{p}}^2$$

$$\leq \frac{i+1}{2}\frac{\sigma}{\tilde{L}} \cdot \gamma_{\mathrm{n}}^2\gamma_{\mathrm{p}} + \frac{i(i+1)}{4}\frac{\sigma}{\tilde{L}} \cdot \gamma_{\mathrm{n}}^3\gamma_{\mathrm{p}}^2$$

$$= a_{i+1} + A_i\gamma_{\mathrm{n}}\gamma_{\mathrm{p}}.$$

So this choice, and in particular any choice that satisfies (13), guarantees discretization error $E_{i+1} \leq A_i \hat{\varepsilon}_i$. By the definition of G_i and E_i we have

$$f(\tilde{x}_T) - f(\tilde{x}^*) \leq \frac{A_1 G_1}{A_T} + \sum_{i=1}^{T-1} \frac{A_i \hat{\varepsilon}_i}{A_T}$$

So it only remains to bound the initial gap G_1. In order to do this, we note that the initial conditions and the method imply the following computation of the first points, from $\tilde{x}_0 \in Q$, which is an arbitrary initial point:

$$\begin{cases} \tilde{z}_0 = \nabla \psi(\tilde{x}_0) \\ \tilde{\chi}_0 = \frac{\hat{\gamma}_0 A_0}{A_0 \hat{\gamma}_0 + a_1/\gamma_{\mathrm{n}}} \tilde{x}_0 + \frac{a_1/\gamma_{\mathrm{n}}}{A_0 \hat{\gamma}_0 + a_1/\gamma_{\mathrm{n}}} \nabla \psi^*(\tilde{z}_0) = \nabla \psi^*(\nabla \psi(\tilde{x}_0)) = \tilde{x}_0 \\ \tilde{\zeta}_0 = \tilde{z}_0 - \frac{a_1}{\gamma_{\mathrm{n}}} \nabla f(\tilde{\chi}_0) = \tilde{z}_0 - \frac{a_1}{\gamma_{\mathrm{n}}} \nabla f(\tilde{x}_0) \\ \tilde{x}_1 = \frac{\hat{\gamma}_0 A_0}{A_0 \hat{\gamma}_0 + a_1/\gamma_{\mathrm{n}}} \tilde{x}_0 + \frac{a_1/\gamma_{\mathrm{n}}}{A_0 \hat{\gamma}_0 + a_1/\gamma_{\mathrm{n}}} \nabla \psi^*(\tilde{\zeta}_0) = \nabla \psi^*(\tilde{\zeta}_0) \end{cases} \quad (14)$$

We have used $A_0 = 0$. Note this first iteration does not depend on $\hat{\gamma}_0$. Also, by using this discretization we start at $\tilde{x}_0$ so we modify the definition of the lower bound (8) so the regularizer added measures the distance from $\tilde{x}_0$. This change of $\tilde{x}_{t_0}$ to $\tilde{x}_0 = \tilde{\chi}_0$ only changes the initial gap. Thus, the first lower bound computed is

$$L_1 = f(\tilde{x}_1) + \frac{1}{\gamma_{\mathrm{n}}} \langle \nabla f(\tilde{x}_1), \nabla \psi^*(\tilde{z}_1) - \tilde{x}_1 \rangle + \frac{1}{A_1} D_\psi(\nabla \psi^*(\tilde{z}_1), \tilde{\chi}_0) - \frac{1}{A_1} D_\psi(\tilde{x}^*, \tilde{\chi}_0).$$

Using $a_1 = A_1$, $\tilde{x}_1 = \nabla \psi^*(\tilde{\zeta}_0)$, $(a_1/\gamma_{\mathrm{n}}) \nabla f(\tilde{\chi}_0) = \tilde{z}_0 - \tilde{\zeta}_0$, and the triangle equality for Bregman divergences Lemma 17 we obtain

$$\begin{aligned} \frac{1}{\gamma_{\mathrm{n}}} \langle \nabla f(\tilde{\chi}_0), \nabla \psi^*(\tilde{z}_1) - \tilde{x}_1 \rangle &= \frac{1}{A_1} \langle \tilde{z}_0 - \tilde{\zeta}_0, \nabla \psi^*(\tilde{z}_1) - \nabla \psi^*(\tilde{\zeta}_0) \rangle \\ &= \frac{1}{A_1} \left(D_{\psi^*}(\tilde{z}_0, \tilde{\zeta}_0) - D_{\psi^*}(\tilde{z}_0, \tilde{z}_1) + D_{\psi^*}(\tilde{\zeta}_0, \tilde{z}_1) \right). \end{aligned} \quad (15)$$

On the other hand, by smoothness of f and the initial condition we have

$$\frac{1}{\gamma_{\mathrm{n}}} \langle \nabla f(\tilde{x}_1) - \nabla f(\tilde{\chi}_0), \nabla \psi^*(\tilde{z}_1) - \tilde{x}_1 \rangle \geq -\frac{\tilde{L}}{\gamma_{\mathrm{n}}} \| \nabla \psi^*(\tilde{\zeta}_0) - \tilde{\chi}_0 \| \| \nabla \psi^*(\tilde{z}_1) - \tilde{x}_1 \|. \quad (16)$$

27

We can now finally bound G_1:

$$
\begin{aligned}
G_1 &\overset{\text{①}}{\leq} \frac{\tilde{L}}{\gamma_{\mathrm{n}}} \|\nabla \psi^*(\tilde{\zeta}_0) - \tilde{\chi}_0\| \cdot \|\nabla \psi^*(\tilde{z}_1) - \tilde{x}_1\| \\
&\quad - \frac{1}{A_1} \left(D_{\psi^*}(\tilde{z}_0, \tilde{\zeta}_0) + D_{\psi^*}(\tilde{\zeta}_0, \tilde{z}_1) \right) + \frac{1}{A_1} D_\psi(\tilde{x}^*, \tilde{\chi}_0) \\
&\overset{\text{②}}{\leq} \frac{\tilde{L}}{\gamma_{\mathrm{n}}} \|\nabla \psi^*(\tilde{\zeta}_0) - \tilde{\chi}_0\| \cdot \|\nabla \psi^*(\tilde{z}_1) - \tilde{x}_1\| \\
&\quad - \frac{\sigma}{2A_1} \left(\|\nabla \psi^*(\tilde{\zeta}_0) - \tilde{\chi}_0\|^2 + \|\nabla \psi^*(\tilde{z}_1) - \tilde{x}_1\|^2 \right) + \frac{1}{A_1} D_\psi(\tilde{x}^*, \tilde{\chi}_0) \\
&\overset{\text{③}}{\leq} \|\nabla \psi^*(\tilde{\zeta}_0) - \tilde{\chi}_0\| \cdot \|\nabla \psi^*(\tilde{z}_1) - \tilde{x}_1\| \left(\frac{\tilde{L}}{\gamma_{\mathrm{n}}} - \frac{\sigma}{A_1} \right) + \frac{1}{A_1} D_\psi(\tilde{x}^*, \tilde{\chi}_0) \\
&\overset{\text{④}}{\leq} \frac{1}{A_1} D_\psi(\tilde{x}^*, \tilde{\chi}_0).
\end{aligned}
$$

We used in ① the definition of $G_1 = U_1 - L_1 = f(\tilde{x}_1) - L_1$ and we bound the inner product in L_1 using $-((15) + (16))$. Also, since $\tilde{z}_0 = \nabla \psi(\tilde{\chi}_0)$ we have $D_{\psi^*}(\tilde{z}_0, \tilde{z}_1) = D_{\psi^*}(\nabla \psi(\tilde{\chi}_0), \tilde{z}_1) = D_\psi(\nabla \psi^*(\tilde{z}_1), \tilde{\chi}_0)$, so we can cancel two of the Bregman divergences. In ②, we used Lemma 18, $\nabla \psi^*(\tilde{z}_0) = \tilde{x}_0 = \tilde{\chi}_0$, and $\nabla \psi^*(\tilde{\zeta}_0) = \tilde{x}_1$. In ③ we used again the inequality $-(a^2 + b^2) \leq -2ab$. Finally ④ is deduced from $A_1 = a_1 \leq \sigma \gamma_{\mathrm{n}} / \tilde{L}$ which comes from the assumption $\tilde{L} a_{i+1}^2 / \gamma_{\mathrm{n}} \sigma \leq a_{i+1} + A_i \gamma_{\mathrm{n}} \gamma_{\mathrm{p}}$ for $i = 0$.

The first part of the theorem follows. The second one is a straightforward application of the first one as we see below. Indeed, taking into account $A_T = \frac{T(T+1)\sigma \gamma_{\mathrm{n}}^2 \gamma_{\mathrm{p}}}{4\tilde{L}}$, and the choice of $T = \left\lceil \sqrt{4\tilde{L} \|\tilde{x}_0 - \tilde{x}^*\|^2 / (\gamma_{\mathrm{n}}^2 \gamma_{\mathrm{p}} \varepsilon)} \right\rceil$, $\psi(\tilde{x}) = \frac{\sigma}{2} \|\tilde{x}\|^2$, and $\hat{\varepsilon}_i = \frac{A_T \varepsilon}{2(T-1)A_i}$ we derive the second statement.

$$
f(\tilde{x}_T) - f(\tilde{x}^*) \leq \frac{A_1 G_1}{A_T} + \sum_{i=1}^{T-1} \frac{A_i \hat{\varepsilon}_i}{A_T} \leq \frac{\frac{\sigma}{2} \|\tilde{x}_0 - \tilde{x}^*\|^2}{A_T} + \frac{\varepsilon}{2} < \frac{2\tilde{L} \|\tilde{x}_0 - \tilde{x}^*\|^2}{\gamma_{\mathrm{n}}^2 \gamma_{\mathrm{p}} T^2} + \frac{\varepsilon}{2} \leq \varepsilon.
$$

∎

We present now the final lemma, that proves that $\hat{\gamma}_i$ can be found efficiently. As we advanced in the sketch of the main paper, we use a binary search. The idea behind it is that due to tilted-convexity (3) we satisfy the equation for $\hat{\gamma}_i = \frac{1}{\gamma_{\mathrm{n}}}$ or $\hat{\gamma}_i = \gamma_{\mathrm{p}}$, or there is $\hat{\gamma}_i \in (\gamma_{\mathrm{p}}, 1/\gamma_{\mathrm{n}})$ such that $\langle \nabla f(\tilde{x}_{i+1}), \tilde{x}_{i+1} - \tilde{x}_i \rangle = 0$. The existence of $\tilde{x}^*$ that satisfies $\nabla f(\tilde{x}^*) = 0$ along with the bounded-ness of Q and smoothness, imply the Lipschitzness of f. Both Lipschitzness and smoothness allow to prove that a binary search finds efficiently a suitable point.

Lemma 15 *Let $Q \subseteq \mathbb{R}^n$ be a convex set of diameter $2\tilde{R}$. Let $f : Q \to \mathbb{R}$ be a function that satisfies tilted-convexity (3), is $\tilde{L}$ smooth and such that there is $\tilde{x}^* \in Q$ such that $\nabla f(\tilde{x}^*) = 0$. Let the strongly convex parameter of $\psi(\cdot)$ be $\sigma = O(1)$. Let $i \geq 1$ be an index. Given two points*

$\tilde{x}_i, \tilde{z}_i \in Q$ *and the method in* (6) *using the learning rates* $a_i = \frac{i}{2} \cdot \frac{\sigma}{L} \cdot \gamma_{\mathrm{n}}^2 \gamma_{\mathrm{p}}$ *prescribed in Theorem 14, we can compute* $\hat{\gamma}_i$ *satisfying* (11)*, i.e.,*

$$f(\tilde{x}_{i+1}) - f(\tilde{x}_i) \leq \hat{\gamma}_i \langle \nabla f(\tilde{x}_{i+1}), \tilde{x}_{i+1} - \tilde{x}_i \rangle + \hat{\varepsilon}_i. \tag{17}$$

And the computation of $\hat{\gamma}_i$ *requires no more than*

$$O\left(\log \left(\frac{\tilde{L}\tilde{R}}{\gamma_{\mathrm{n}} \hat{\varepsilon}_i} \cdot i \right) \right)$$

queries to the gradient oracle.

Proof Let $\hat{\Gamma}_i(\lambda) : [\frac{a_{i+1}}{A_{i+1}}, \frac{a_{i+1}/\gamma_{\mathrm{n}}}{A_i \gamma_{\mathrm{p}} + a_{i+1}/\gamma_{\mathrm{n}}}] \to \mathbb{R}$ be defined as

$$\hat{\Gamma}_i \left(\frac{a_{i+1}/\gamma_{\mathrm{n}}}{A_i \tilde{\mathrm{x}} + a_{i+1}/\gamma_{\mathrm{n}}} \right) = \tilde{\mathrm{x}}, \text{ for } \tilde{\mathrm{x}} \in [\gamma_{\mathrm{p}}, \frac{1}{\gamma_{\mathrm{n}}}]. \tag{18}$$

By monotonicity, it is well defined. Let $\tilde{x}_{i+1}^{\lambda}$ be the point computed by one iteration of (6) using the parameter $\hat{\gamma}_i = \hat{\Gamma}_i(\lambda)$. Likewise, we define the rest of the points in iteration (6) depending on λ. We first try $\hat{\gamma}_i = 1/\gamma_{\mathrm{n}}$ and $\hat{\gamma}_i = \gamma_{\mathrm{p}}$ and use any of them if they satisfy the conditions. If neither of them do, it means that for the first choice we had $\langle \nabla f(\tilde{x}_{i+1}^{\lambda_1}), \tilde{x}_{i+1}^{\lambda_1} - \tilde{x}_i \rangle < 0$ and for the second one, it is $\langle \nabla f(\tilde{x}_{i+1}^{\lambda_2}), \tilde{x}_{i+1}^{\lambda_2} - \tilde{x}_i \rangle > 0$, for $\lambda_1 = \hat{\Gamma}_i^{-1}(1/\gamma_{\mathrm{n}})$ and $\lambda_2 = \hat{\Gamma}_i^{-1}(\gamma_{\mathrm{p}})$. Therefore, by continuity, there is $\lambda^* \in [\lambda_1, \lambda_2]$ such that $\langle \nabla f(\tilde{x}_{i+1}^{\lambda^*}), \tilde{x}_{i+1}^{\lambda^*} - \tilde{x}_i \rangle = 0$. The continuity condition is easy to prove. We omit it because it is derived from the Lipschitzness condition that we will prove below. Such a point satisfies (11) for $\hat{\varepsilon}_i = 0$. We will prove that the function $\mathcal{G}_i : [\frac{a_{i+1}}{A_{i+1}}, \frac{a_{i+1}/\gamma_{\mathrm{n}}}{A_i \gamma_{\mathrm{p}} + a_{i+1}/\gamma_{\mathrm{n}}}] \to \mathbb{R}$, defined as

$$\mathcal{G}_i(\lambda) \stackrel{\mathrm{def}}{=} -\hat{\Gamma}_i(\lambda) \langle \nabla f(\tilde{x}_{i+1}^{\lambda}), \tilde{x}_{i+1}^{\lambda} - \tilde{x}_i \rangle + (f(\tilde{x}_{i+1}^{\lambda}) - f(\tilde{x}_i)), \tag{19}$$

is Lipschitz so we can guarantee that (11) holds for a large enough interval around λ^*. Finally, we will be able to perform a binary search to efficiently find a point in such interval or another interval around another point that satisfies that the inner product is 0.

So

$$\begin{aligned}
|\mathcal{G}_i(\lambda) - \mathcal{G}_i(\lambda')| \leq{}& |f(\tilde{x}_{i+1}^{\lambda}) - f(\tilde{x}_{i+1}^{\lambda'})| \\
&+ |\hat{\Gamma}_i(\lambda')| \cdot |\langle \nabla f(\tilde{x}_{i+1}^{\lambda'}), \tilde{x}_{i+1}^{\lambda'} - \tilde{x}_i \rangle - \langle \nabla f(\tilde{x}_{i+1}^{\lambda}), \tilde{x}_{i+1}^{\lambda} - \tilde{x}_i \rangle| \\
&+ |\langle \nabla f(\tilde{x}_{i+1}^{\lambda}), \tilde{x}_{i+1}^{\lambda} - \tilde{x}_i \rangle| \cdot |\hat{\Gamma}_i(\lambda') - \hat{\Gamma}_i(\lambda)|
\end{aligned} \tag{20}$$

We have used the triangular inequality and the inequality

$$|\alpha_1 \beta_1 - \alpha_2 \beta_2| \leq |\alpha_1||\beta_1 - \beta_2| + |\beta_2||\alpha_1 - \alpha_2|, \tag{21}$$

which is a direct consequence of the triangular inequality, after adding and subtracting $\alpha_1 \beta_2$ in the $|\cdot|$ on the left hand side. We bound each of the three summands of the previous inequality separately,

but first we bound the following which will be useful for our other bounds,

$$
\begin{aligned}
\|\tilde{x}_{i+1}^{\lambda'} - \tilde{x}_{i+1}^{\lambda}\| &\overset{①}{=} \|(\lambda'\nabla\psi^*(\tilde{\zeta}_i^{\lambda'}) + (1-\lambda')\tilde{x}_i) - (\lambda\nabla\psi^*(\tilde{\zeta}_i^{\lambda}) + (1-\lambda)\tilde{x}_i)\| \\
&\overset{②}{\leq} \|\nabla\psi^*(\tilde{\zeta}_i^{\lambda}) - \tilde{x}_i\||\lambda'-\lambda| + \|\lambda'\nabla\psi^*(\tilde{\zeta}_i^{\lambda'}) - \lambda'\nabla\psi^*(\tilde{\zeta}_i^{\lambda})\| \\
&\overset{③}{\leq} 2\tilde{R}|\lambda-\lambda'| + \|\nabla\psi^*(\tilde{\zeta}_i^{\lambda'}) - \nabla\psi^*(\tilde{\zeta}_i^{\lambda})\| \overset{④}{\leq} 2\tilde{R}|\lambda-\lambda'| + \frac{a_{i+1}}{\gamma_n\sigma}\|\nabla f(\tilde{\chi}_i^{\lambda}) - \nabla f(\tilde{\chi}_i^{\lambda'})\| \\
&\overset{⑤}{\leq} 2\tilde{R}|\lambda-\lambda'| + \frac{a_{i+1}\tilde{L}}{\gamma_n\sigma}\|\tilde{\chi}_i^{\lambda} - \tilde{\chi}_i^{\lambda'}\| \overset{⑥}{\leq} \left(2\tilde{R} + \frac{2a_{i+1}\tilde{L}\tilde{R}}{\gamma_n\sigma}\right)|\lambda-\lambda'|
\end{aligned}
\tag{22}
$$

Here, ① uses the definition of $\tilde{x}_{i+1}^{\lambda}$ as a convex combination of $\tilde{x}_i$ and $\nabla\psi^*(\tilde{\zeta}_i^{\lambda})$. ② adds and substracts $\lambda'\nabla\psi^*(\tilde{\zeta}_i^{\lambda})$, groups terms and uses the triangular inequality. In ③ we use the fact that the diameter of Q is $2\tilde{R}$ and bound $\lambda' \leq 1$, and $|\lambda| \leq 1$. ④ uses the $\frac{1}{\sigma}$ Lipschitzness of $\nabla\psi^*(\cdot)$, which is a consequence of the σ-strong convexity of $\psi(\cdot)$. ⑤ uses the smoothness of f. In ⑥, from the definition of $\tilde{\chi}_i^{\lambda}$ we have that $\|\tilde{\chi}_i^{\lambda} - \tilde{\chi}_i^{\lambda'}\| \leq \|\tilde{x}_i - \tilde{z}_i\||\lambda-\lambda'|$. We bounded this further using the diameter of Q.

Note that f is Lipschitz over Q. By the existence of $\tilde{x}^*$, $\tilde{L}$-smoothness, and the diameter of Q we have $\|\nabla f(\tilde{x})\| = \|\nabla f(\tilde{x}) - \nabla f(\tilde{x}^*)\| \leq \tilde{L}\|\tilde{x} - \tilde{x}^*\| \leq 2R\tilde{L}$. So the Lipschitz constant L_p of f is $L_p \leq 2R\tilde{L}$. Now we can proceed and bound the three summands of (20). The first one reduces to the inequality above after using Lipschitzness of $f(\cdot)$:

$$
|f(\tilde{x}_{i+1}^{\lambda}) - f(\tilde{x}_{i+1}^{\lambda'})| \leq L_p\|\tilde{x}_{i+1}^{\lambda'} - \tilde{x}_{i+1}^{\lambda}\|.
\tag{23}
$$

We prove Lipschitzness of $\hat{\Gamma}_i$. Note that

$$
|(\hat{\Gamma}_i^{-1})'(\tilde{\mathbf{x}})| = \left|\frac{A_i a_{i+1}/\gamma_n}{(A_i\tilde{\mathbf{x}} + a_{i+1}/\gamma_n)^2}\right| \geq \frac{\gamma_n A_i a_{i+1}}{A_{i+1}^2},
\tag{24}
$$

so $\hat{\Gamma}_i'(\lambda)$ is bounded by $A_{i+1}^2/(\gamma_n A_i a_{i+1})$ for any λ. In order to bound the second summand, we use $\tilde{\mathbf{x}} \in [\gamma_p, 1/\gamma_n]$ and obtain $|\hat{\Gamma}_i(\lambda)| \leq \frac{1}{\gamma_n}$. For the second factor, we add and subtract $\langle\nabla f(\tilde{x}_{i+1}^{\lambda}), \tilde{x}_{i+1}^{\lambda'} - \tilde{x}_i\rangle$ and use the triangular inequality and then Cauchy-Schwartz. Thus, we obtain

$$
\begin{aligned}
&|\langle\nabla f(\tilde{x}_{i+1}^{\lambda'}), \tilde{x}_{i+1}^{\lambda'} - \tilde{x}_i\rangle - \langle\nabla f(\tilde{x}_{i+1}^{\lambda}), \tilde{x}_{i+1}^{\lambda} - \tilde{x}_i\rangle| \\
&\leq \|\nabla f(\tilde{x}_{i+1}^{\lambda})\| \cdot \|\tilde{x}_{i+1}^{\lambda'} - \tilde{x}_{i+1}^{\lambda}\| + \|\nabla f(\tilde{x}_{i+1}^{\lambda'}) - \nabla f(\tilde{x}_{i+1}^{\lambda})\| \cdot \|\tilde{x}_{i+1}^{\lambda'} - \tilde{x}_i\| \\
&\overset{①}{\leq} (L_p + 2\tilde{L}\tilde{R})\|\tilde{x}_{i+1}^{\lambda'} - \tilde{x}_{i+1}^{\lambda}\|.
\end{aligned}
\tag{25}
$$

In ①, we used Lipschitzness to bound the first factor. We also used the diameter of Q to bound the last factor and the smoothness of $f(\cdot)$ to bound the first factor of the second summand.

For the third summand, we will bound the first factor using Cauchy-Schwartz, Lipschitzness of $f(\cdot)$ and the diameter of Q. We just proved in (24) that $\hat{\Gamma}_i$ is Lipschitz, so use this property for the

second factor. The result is the following

$$|\langle \nabla f(\tilde{x}_{i+1}^{\lambda}), \tilde{x}_{i+1}^{\lambda} - \tilde{x}_i \rangle| \cdot |\hat{\Gamma}_i(\lambda') - \hat{\Gamma}_i(\lambda)| \leq 2L_{\mathrm{p}}\tilde{R}\frac{A_{i+1}^2}{\gamma_{\mathrm{n}}A_i a_{i+1}}|\lambda' - \lambda|. \tag{26}$$

Applying the bounds of the three summands (23), (24), (25), (26) into (20) we obtain the inequality $|\mathcal{G}_i(\lambda') - \mathcal{G}_i(\lambda)| \leq \hat{L}|\lambda' - \lambda|$ for

$$\hat{L} = \left(2\tilde{R} + \frac{2a_{i+1}\tilde{L}\tilde{R}}{\gamma_{\mathrm{n}}\sigma}\right)\left(L_{\mathrm{p}} + (L_{\mathrm{p}} + 2\tilde{L}\tilde{R})\frac{1}{\gamma_{\mathrm{n}}}\right) + 2L_{\mathrm{p}}\tilde{R}\frac{A_{i+1}^2}{\gamma_{\mathrm{n}}A_i a_{i+1}}.$$

We will use the following to bound $\hat{L}$. If we use the learning rates prescribed in Theorem 14, namely $a_i = \frac{i\sigma\gamma_{\mathrm{n}}^2\gamma_{\mathrm{p}}}{2L}$ and thus $A_i = \frac{i(i+1)\sigma\gamma_{\mathrm{n}}^2\gamma_{\mathrm{p}}}{4L}$ we can bound $A_{i+1}^2/(A_i a_{i+1}) \leq 4(i+2)$, using that $i \geq 1$. We recall we computed $L_{\mathrm{p}} \leq 2\tilde{R}\tilde{L}$ and that we assumed $\sigma = O(1)$. In Algorithm 1 we use $\sigma = 1$.

On the other hand the initial length of the search interval, which is the domain of definition of $\mathcal{G}_i$ is at most 1 since the interval is in $(0, 1)$. Recall we are denoting by λ^* a value such that $\langle \nabla f(\tilde{x}_{i+1}^{\lambda^*}), \tilde{x}_{i+1}^{\lambda^*} - \tilde{x}_i \rangle = 0$ so $\mathcal{G}_i(\lambda^*) \leq 0$. Lipschitzness of G implies that if $\mathcal{G}_i(\lambda^*) \leq 0$ then $\mathcal{G}_i(\lambda) \leq \hat{\varepsilon}_i$ for

$$\lambda \in [\lambda^* - \frac{\hat{\varepsilon}_i}{\hat{L}}, \lambda^* + \frac{\hat{\varepsilon}_i}{\hat{L}}] \cap [\hat{\Gamma}_i^{-1}(1/\gamma_{\mathrm{n}}), \hat{\Gamma}_i^{-1}(\gamma_{\mathrm{p}})].$$

If the extremal points, $\hat{\Gamma}_i^{-1}(1/\gamma_{\mathrm{n}}), \hat{\Gamma}_i^{-1}(\gamma_{\mathrm{p}})$ did not satisfy (17), then this interval is of length $\frac{2\hat{\varepsilon}_i}{\hat{L}}$ and a point in such interval or another interval that is around another point $\bar{\lambda}^*$ that satisfies $\langle \nabla f(\tilde{x}_{i+1}^{\bar{\lambda}^*}), \tilde{x}_{i+1}^{\bar{\lambda}^*} - \tilde{x}_i \rangle = 0$ can be found with a binary search in at most

$$O\left(\log\left(\frac{\hat{L}}{\hat{\varepsilon}_i}\right)\right) \overset{\text{①}}{=} O\left(\log\left(\frac{\tilde{L}\tilde{R}}{\gamma_{\mathrm{n}}\hat{\varepsilon}_i} \cdot i\right)\right)$$

iterations, provided that at each step we can ensure we halve the size of the search interval. The bounds of the previous paragraph are applied in ①.The binary search can be done easily: we start with $[\hat{\Gamma}_i^{-1}(1/\gamma_{\mathrm{n}}), \hat{\Gamma}_i^{-1}(\gamma_{\mathrm{p}})]$ and assume the extremes do not satisfy (17), so the sign of $\langle \nabla f(\tilde{x}_{i+1}^{\lambda}), \tilde{x}_{i+1}^{\lambda} - \tilde{x}_i \rangle$ is different for each extreme. Each iteration of the binary search queries the midpoint of the current working interval and if (17) is not satisfied, we keep the half of the interval such that the extremes keep having the sign of $\langle \nabla f(\tilde{x}_{i+1}^{\lambda}), \tilde{x}_{i+1}^{\lambda} - \tilde{x}_i \rangle$ different from each other, ensuring that there is a point in which this expression evaluates to 0 and thus keeping the invariant. We include the pseudocode of this binary search in Algorithm 2. ∎

We proceed to prove Theorem 5, which is an immediate consequence of the previous results.

Proof of Theorem 5. The proof follows from Theorem 14, provided that we can find $\hat{\gamma}_i$ satisfying (11). Lemma 15 shows that this is possible after performing a logarithmic number of queries to the gradient oracle. Note that given our choice of $\hat{\varepsilon}_i$, T and a_i, the number of queries to the gradient oracle Lemma 15 requires is no more than $O(\log(\tilde{L}R/\gamma_{\mathrm{n}}\varepsilon))$ for any $i \leq T$. So we find an ε-minimizer of f after $\widetilde{O}(\sqrt{\tilde{L}/(\gamma_{\mathrm{n}}^2\gamma_{\mathrm{p}}\varepsilon)})$ queries to the gradient oracle. ∎

Proof of Theorem 6. Given the function to optimize $F : \mathcal{M}_K \to \mathbb{R}$ and the geodesic map h, we define $f = F \circ h^{-1}$. Using Lemma 4 we know that f is $\tilde{L}$-smooth, with $\tilde{L} = O(L)$. Lemma 3

Algorithm 2 BinaryLineSearch($\tilde{x}_i, \tilde{z}_i, f, \mathcal{X}, a_{i+1}, A_i, \varepsilon, \tilde{L}, \gamma_{\mathrm{n}}, \gamma_{\mathrm{p}}$)

Input: Points $\tilde{x}_i, \tilde{z}_i$, function f, domain $\mathcal{X}$, learning rate a_{i+1}, accumulated learning rate A_i, final target accuracy ε, final number of iterations T, smoothness constant $\tilde{L}$, constants $\gamma_{\mathrm{n}}, \gamma_{\mathrm{p}}$. Define $\hat{\varepsilon}_i \leftarrow (A_T \varepsilon)/(2(T-1)A_i)$ as in Theorem 14, i.e. with $A_T = T(T+1)\gamma_{\mathrm{n}}^2 \gamma_{\mathrm{p}}/4\tilde{L}$. $\hat{\Gamma}_i$ defined as in (18) and $\mathcal{G}_i$ defined as in (19) i.e.

$$\mathcal{G}_i(\lambda) \stackrel{\text{def}}{=} -\hat{\Gamma}_i(\lambda)\langle \nabla f(\tilde{x}_{i+1}^{\lambda}), \tilde{x}_{i+1}^{\lambda} - \tilde{x}_i \rangle + (f(\tilde{x}_{i+1}^{\lambda}) - f(\tilde{x}_i)),$$

for x_{i+1}^{λ} being the result of method (12) when $\hat{\gamma}_i = \hat{\Gamma}_i(\lambda)$.

Output: $\lambda = \frac{a_{i+1}/\gamma_{\mathrm{n}}}{A_i \hat{\gamma}_i + a_{i+1}/\gamma_{\mathrm{n}}}$ for $\hat{\gamma}_i$ such that $\mathcal{G}_i(\hat{\Gamma}_i^{-1}(\hat{\gamma}_i)) \leq \hat{\varepsilon}_i$.

1: **if** $\mathcal{G}_i(\hat{\Gamma}_i^{-1}(1/\gamma_{\mathrm{n}})) \leq \hat{\varepsilon}_i$ **then** $\lambda = \hat{\Gamma}_i^{-1}(1/\gamma_{\mathrm{n}})$
2: **else if** $\mathcal{G}_i(\hat{\Gamma}_i^{-1}(\gamma_{\mathrm{p}})) \leq \hat{\varepsilon}_i$ **then** $\lambda = \hat{\Gamma}_i^{-1}(\gamma_{\mathrm{p}})$
3: **else**
4:      left $\leftarrow \hat{\Gamma}_i^{-1}(1/\gamma_{\mathrm{n}})$
5:      right $\leftarrow \hat{\Gamma}_i^{-1}(\gamma_{\mathrm{p}})$
6:      $\lambda \leftarrow (\text{left} + \text{right})/2$
7:      **while** $\mathcal{G}_i(\lambda) > \hat{\varepsilon}_i$ **do**
8:          **if** $\langle \nabla f(\tilde{x}_{i+1}^{\lambda}), \tilde{x}_{i+1}^{\lambda} - \tilde{x}_i \rangle < 0$ **then** right $\leftarrow \lambda$
9:          **else** left $\leftarrow \lambda$
10:         **end if**
11:         $\lambda \leftarrow (\text{left} + \text{right})/2$
12:      **end while**
13: **end if**
14: **return** λ

---

proves that f satisfies tilted-convexity (3) for constants γ_{n} and γ_{p} depending on R. So Theorem 5 applies and the total number of queries to the oracle needed to obtain an ε-minimizer of f is $\widetilde{O}(\sqrt{\tilde{L}/\gamma_{\mathrm{n}}^2 \gamma_{\mathrm{p}} \varepsilon}) = \widetilde{O}(\sqrt{L/\varepsilon})$. The result follows, since $f(\tilde{x}_T) - f(\tilde{x}^*) = F(x_T) - F(x^*)$. ∎

We recall a few concepts that were assumed during Section 2 to better interpret Theorem 6. We work in the hyperbolic space, or in an open hemisphere. The aim is to minimize a smooth and g-convex function defined on any of these manifolds, or a submanifold of them. The existence of a point x^* that satisfies $\nabla F(x^*) = 0$ is assumed. Starting from an arbitrary point x_0, we let R be a bound of the distance between x_0 and x^*, that is, $R \geq d(x_0, x^*)$. We perform constrained optimization over $\mathcal{B}_R = \mathrm{Exp}_{x_0}(\bar{B}(0, R))$. Note $x^* \in \mathcal{B}_R$. We assume $F : \mathcal{M}_K \to \mathbb{R}$ is a differentiable function, $\mathcal{B}_R \subset \mathcal{M}_K$, and $\mathcal{M}_K$ has constant sectional curvature K. If K is positive, we restrict $R < \pi/2\sqrt{K}$ so $\mathcal{B}_R$ is contained in an open hemisphere and it is uniquely geodesic. We define a geodesic map $h : \mathcal{M}_K \to M$, where $M \subset \mathbb{R}^n$ and define the function $f : h(\mathcal{M}_K) \to \mathbb{R}$ as $f = F \circ h^{-1}$. We perform constrained optimization over this function f in $\mathcal{X} = h(\mathcal{B}_R)$ in an accelerated way, up to constants and log factors, where the constants appear as an effect of the deformation of the geometry and depend on R and K only.

We note that we assumed the existence of $x^* \in \mathcal{B}_R$, such that $\nabla F(x^*) = 0$ for simplicity only. The algorithm can find $\hat{x}^* = \arg\min_{x \in \mathcal{B}_R}\{F(x)\}$ instead. This point always satisfies the first

32

inequality in tilted-convexity (3)—the same proof used for x^* works— so the lower bounds L_i can be defined in the same way as we did, and all the proofs work in the same way, except that in this case, then we would need to bound the Lipschitz constant of F when restricted to B_R, which for the case $\nabla F(\hat{x}^*) = 0$, it was $O(L)$. This Lipschitz constant would be used for the analysis of the binary search and for the computation of $\tilde{L}$, the smoothness constant of f.

Note that there are some cases in which one can perform the aforementioned constrained optimization to adapt to the value of $d(x_0, x^*)$, if an estimate $R > d(x_0, x^*)$ is not available. For instance, in the negative curvature case, if we know the value of $F(x^*)$, we can run the algorithm with a small value of R and with precision ε. We can check if we obtained a global ε-minimizer and if we did not, we increase R by a factor of 2 and repeat. As soon as our R upper bounds $d(x_0, x^*)$, we will obtain a global ε-minimizer, and the total complexity is of the order of the one of the last run. We always have the lower bound $d(x_0, x^*) \geq \|\nabla F(x_0)\|/L$ so we can choose the right hand side as the initial value of R.

A.1. Auxiliary lemmas

The following are classical lemmas of convex optimization that we used in this section and that we add for completeness.

Fact 16 *Let $\psi : Q \to \mathbb{R}$ be a differentiable strongly-convex function. Then*

$$\nabla \psi^*(\tilde{z}) = \arg\min_{\tilde{x} \in Q}\{\langle -\tilde{z}, \tilde{x}\rangle + \psi(\tilde{x})\}.$$

See Bertsekas and Nedic (2003) for a proof. Note that by the definition of Bregman divergence $\arg\min_{\tilde{x} \in Q}\{\langle \tilde{y} - \nabla \psi(\tilde{x}_0), \tilde{x}\rangle + \psi(\tilde{x})\} = \arg\min_{\tilde{x} \in Q}\{\langle \tilde{y}, \tilde{x}\rangle + D_\psi(\tilde{x}, \tilde{x}_0)\}$.

Lemma 17 (Triangle equality of Bregman Divergences) *For all $\tilde{x}, \tilde{y}, \tilde{z} \in Q$ we have*

$$D_{\psi^*}(\tilde{x}, \tilde{y}) = D_{\psi^*}(\tilde{z}, \tilde{y}) + D_{\psi^*}(\tilde{x}, \tilde{z}) + \langle \nabla \psi^*(\tilde{z}) - \nabla \psi^*(\tilde{y}), \tilde{x} - \tilde{z}\rangle.$$

Proof

$$\begin{aligned}
D_{\psi^*}(\tilde{z}, \tilde{y}) &+ D_{\psi^*}(\tilde{x}, \tilde{z}) + \langle \nabla \psi^*(\tilde{z}) - \nabla \psi^*(\tilde{y}), \tilde{x} - \tilde{z}\rangle \\
&= (\psi^*(\tilde{z}) - \psi^*(\tilde{y}) - \langle \nabla \psi^*(\tilde{y}), \tilde{z} - \tilde{y}\rangle) \\
&\quad + (\psi^*(\tilde{x}) - \psi^*(\tilde{z}) - \langle \nabla \psi^*(\tilde{z}), \tilde{x} - \tilde{z}\rangle) \\
&\quad + \langle \nabla \psi^*(\tilde{z}) - \nabla \psi^*(\tilde{y}), \tilde{x} - \tilde{z}\rangle \\
&= \psi^*(\tilde{x}) - \psi^*(\tilde{y}) - \langle \nabla \psi^*(\tilde{y}), \tilde{z} - \tilde{y}\rangle + \langle -\nabla \psi^*(\tilde{y}), \tilde{x} - \tilde{z}\rangle \\
&= D_{\psi^*}(\tilde{x}, \tilde{y}).
\end{aligned}$$

■

Lemma 18 *Given a σ-strongly convex function $\psi(\cdot)$ the following holds:*

$$D_{\psi^*}(\tilde{z}_1, \tilde{z}_2) \geq \frac{\sigma}{2}\|\nabla \psi^*(\tilde{z}_1) - \nabla \psi^*(\tilde{z}_2)\|^2.$$

Proof Using the first order optimality condition of the Fenchel dual and (16) we obtain

$$\langle \nabla \psi(\nabla \psi^*(\tilde{z}_1)) - \tilde{z}_1, \nabla \psi^*(\tilde{z}_2) - \nabla \psi^*(\tilde{z}_1) \rangle \geq 0$$

Using σ-strong convexity of ψ and the previous inequality we have

$$
\begin{aligned}
D_{\psi^*}(\tilde{z}_1, \tilde{z}_2) &= \psi(\nabla \psi^*(\tilde{z}_2)) - \psi(\nabla \psi^*(\tilde{z}_1)) - \langle \tilde{z}_1, \nabla \psi^*(\tilde{z}_2) - \nabla \psi^*(\tilde{z}_1) \rangle \\
&\geq \frac{\sigma}{2} \|\nabla \psi^*(\tilde{z}_1) - \nabla \psi^*(\tilde{z}_2)\|^2 + \langle \nabla \psi(\nabla \psi^*(\tilde{z}_1)) - \tilde{z}_1, \nabla \psi^*(\tilde{z}_2) - \nabla \psi^*(\tilde{z}_1) \rangle \\
&\geq \frac{\sigma}{2} \|\nabla \psi^*(\tilde{z}_1) - \nabla \psi^*(\tilde{z}_2)\|^2.
\end{aligned}
$$

∎

Appendix B. Reductions. Proofs of results in Section 3

Proof of Theorem 7. Let $\mathcal{A}_{\text{ns}}$ be the algorithm in the statement of the theorem. By strong g-convexity of F and the assumptions on $\mathcal{A}_{\text{ns}}$ we have that $\hat{x}_T$, the point computed by $\mathcal{A}_{\text{ns}}$, satisfies

$$\frac{\mu}{2} d(\hat{x}_T, x^*)^2 \leq F(\hat{x}_T) - F(x^*) \leq \frac{\mu}{2} \frac{R^2}{2},$$

after $T = \text{Time}_{\text{ns}}(L, \mu, R)$ queries to the gradient oracle. This implies $d(\hat{x}_T, x^*)^2 \leq R^2/2$. We perform this process $r \stackrel{\text{def}}{=} \lceil \log(\mu R^2/\varepsilon) - 1 \rceil$ times. We use the previous output as input for the next round. The distance bound to x^* can be updated to a lower value. We denote R_i the distance bound from the input to x^* at stage i and we set its value to $R_i = R_{i-1}/\sqrt{2}$, for $R_1 = R$. Thus, after r stages we obtain a point $\hat{x}_T^r$ that satisfies

$$F(\hat{x}_T^r) - F(x^*) \leq \frac{\mu \cdot R_r^2}{4} = \frac{\mu \cdot R_1^2}{4 \cdot 2^{r-1}} \leq \varepsilon.$$

And the total running time is $\text{Time}_{\text{ns}}(L, \mu, R) \cdot r = O(\text{Time}_{\text{ns}}(L, \mu, R) \log(\mu/\varepsilon))$. ∎

Proof of Corollary 8. We can assume without loss of generality $K \in \{-1, 1\}$ as we did in Section 2. Let R be an upper bound on the distance between the initial point x_0 and an optimizer x^*, i.e. $d(x_0, x^*) \leq R$. Note that $\|\tilde{x}_0 - \tilde{x}^*\|/R$ is bounded by a constant depending on R by Lemma 2.a. Note that γ_{n} and γ_{p} are constants depending on R by Lemma 3. As any g-strongly convex function is g-convex, by using Theorem 14 and Lemma 15 with $\varepsilon = \mu \frac{R^2}{4}$ we obtain that Algorithm 1 obtains a $\mu \frac{R^2}{4}$-minimizer in at most

$$T = O\left(\frac{\|\tilde{x}_0 - \tilde{x}^*\|}{R} \sqrt{\frac{4L}{\mu \gamma_{\text{n}}^2 \gamma_{\text{p}}}} \log \left(\frac{\|\tilde{x}_0 - \tilde{x}^*\|}{R} \sqrt{\frac{4L}{\mu \gamma_{\text{n}}^2 \gamma_{\text{p}}}} \right) \right) = O\left(\sqrt{L/\mu} \log(L/\mu) \right)$$

queries to the gradient oracle. Subsequent stages, i.e. calls to Algorithm 1, use the point computed at the previous stage as its input. The distance bound to x^* is updated, following the proof of Theorem 7. Because the constant depending on R in the running time of the subroutine decreases when R has a lower value, subsequent stages need a time which is $O(\sqrt{L/\mu} \log(L/\mu))$ as well. So

we satisfy the assumption of Theorem 7 for $\text{Time}_{\text{ns}}(L, \mu, R) = O(\sqrt{L/\mu} \log(L/\mu))$. We conclude that given $\varepsilon > 0$ and running Algorithm 1 in stages, we obtain an ε-minimizer of F in

$$O(\sqrt{L/\mu} \log(L/\mu) \log(\mu/\varepsilon)) = O^*(\sqrt{L/\mu} \log(\mu/\varepsilon))$$

queries to the gradient oracle.

Note that each time we call Algorithm 1 we recenter the geodesic map. In order to perform the method with these recentering steps, we need the function F to be defined over at least $\text{Exp}_{x_0}(\bar{B}(0, R \cdot (1 + 2^{-1/2})))$, since subsequent centers are only guaranteed to be $\leq R/\sqrt{2}$ close to x^*, and they could get slightly farther than R from x_0. But they are no farther than $R + R/\sqrt{2}$ since $d(x_0, \hat{x}_T^i) \leq d(x_0, x^*) + d(x^*, \hat{x}_T^i) \leq R + R/\sqrt{2}$, where $\hat{x}_T^i$ is the center at stage i, and where $i > 1$. ∎

B.1. Proof of Theorem 9

We provide here the full statement of the theorem.

Theorem 19 *Let $\mathcal{M}$ be a Riemannian manifold of bounded sectional curvature, let $F : \mathcal{M} \to \mathbb{R}$ be an L-smooth and g-convex function, and assume there is a point $x^* \in \mathcal{M}$ such that $\nabla F(x^*) = 0$. Let x_0 be a starting point such that $d(x_0, x^*) \leq R$ and let Δ satisfy $F(x_0) - F(x^*) \leq \Delta$. Assume Exp_{x_0} is a diffeomorphism when restricted to $\bar{B}(0, R)$ and that we have an algorithm $\mathcal{A}$ that given an L-smooth and μ-strongly g-convex function $\hat{F} : \mathcal{M} \to \mathbb{R}$, with minimizer in $\text{Exp}_{x_0}(\bar{B}(0, R))$, and any initial point $\hat{x}_0 \in \mathcal{M}$, produces a point $\hat{x} \in \text{Exp}_{x_0}(\bar{B}(0, R))$ in time $\hat{T} = \text{Time}(L, \mu, \mathcal{M}, R)$ satisfying $\hat{F}(\hat{x}) - \min_{x \in \mathcal{M}} \hat{F}(x) \leq (\hat{F}(\hat{x}_0) - \min_{x \in \mathcal{M}} \hat{F}(x))/4$. Let $T = \lceil \log_2(\Delta/\varepsilon) \rceil + 1$. Then, we can compute an ε-minimizer in time $\sum_{t=0}^{T-1} \text{Time}(L + 2^{-t} \Delta \mathcal{K}_R^-/R^2, 2^{-t} \Delta \mathcal{K}_R^+/R^2, \mathcal{M}, R)$, where $\mathcal{K}_R^+$ and $\mathcal{K}_R^-$ are constants that depend on R and the bounds on the sectional curvature of $\mathcal{M}$.*

The algorithm is the following. We successively regularize the function with strongly g-convex regularizers in this way $F^{(\mu_i)}(x) \overset{\text{def}}{=} F(x) + \frac{\mu_i}{2} d(x, x_0)^2$ for $i \geq 0$. For each $i \geq 0$, we use the algorithm $\mathcal{A}$ on the function $F^{(\mu_i)}$ for the time in the statement of the theorem and obtain a point $\hat{x}_{i+1}$, starting from point $\hat{x}_i$, where $\hat{x}_0 = x_0$. The regularizers are decreased exponentially $\mu_{i+1} = \mu_i/2$ from $\mu_0 = \Delta/R^2$, until we reach roughly $\mu_T = \varepsilon/R^2$, see below for the precise value. Let's see how this algorithm works. We first state the following fact, that says that indeed $\frac{\mu_i}{2} d(x, x_0)^2$ is a strongly g-convex regularizer. Recall that $\text{Exp}_{x_0}(\bar{B}(0, R)) \subset \mathcal{M}$. We define the following quantities

$$\mathcal{K}_R^+ \overset{\text{def}}{=} \begin{cases} 1 & \text{if } K_{\max} \leq 0 \\ \sqrt{K_{\max}} R \cot(\sqrt{K_{\max}} R) & \text{if } K_{\max} > 0 \end{cases}$$

$$\mathcal{K}_R^- \overset{\text{def}}{=} \begin{cases} \sqrt{-K_{\min}} R \coth(\sqrt{-K_{\min}} R) & \text{if } K_{\min} < 0 \\ 1 & \text{if } K_{\min} \geq 0 \end{cases}$$

Here $K_{\max}$ and $K_{\min}$ are the upper and lower bounds on the sectional curvature of the manifold $\mathcal{M}$.

Fact 20 *Let $\mathcal{M}$ be a manifold with sectional curvature bounded below and above by $K_{\min}$ and $K_{\max}$, respectively. For a point $x_0 \in \mathcal{M}$, assume Exp_{x_0} is a diffeomorphism when restricted to $\bar{B}(0, R)$. The function $f : \mathcal{M} \to \mathbb{R}$ defined as $f(x) = \frac{1}{2}d(x, x_0)^2$ is $\mathcal{K}_R^+$-g-strongly convex and $\mathcal{K}_R^-$-smooth in $\mathrm{Exp}_{x_0}(\bar{B}(0, R))$.*

The result regarding strong convexity can be found, for instance, in (Alimisis et al., 2019) and it is a direct consequence of the following inequality, which can also be found in (Alimisis et al., 2019):

$$d(y, x_0)^2 \geq d(x, x_0)^2 - 2\langle \mathrm{Exp}_x^{-1}(x_0), y - x \rangle + \mathcal{K}_R^+ d(x, y)^2,$$

along with the fact that $\mathrm{grad}\, f(x) = -\mathrm{Exp}_x^{-1}(x_0)$. The result regarding smoothness is, similarly, obtained from the following inequality:

$$d(y, x_0)^2 \leq d(x, x_0)^2 - 2\langle \mathrm{Exp}_x^{-1}(x_0), y - x \rangle + \mathcal{K}_R^- d(x, y)^2,$$

which can be found in (Zhang and Sra, 2016) (Lemma 6). These inequalities are tight in spaces of constant sectional curvature. Alternatively, one can derive these inequalities from upper and lower bounds on the Hessian of $f(x) = \frac{1}{2}d(x, x_0)^2$, as it was done in (Lezcano-Casado, 2020) (Theorem 3.15).

We prove now that the regularization makes the minimum to be closer to x_0, so the assumption of the theorem on $\hat{F}$ holds for the functions we use. Define x_{i+1} as the minimizer of $F^{(\mu_i)}$.

Lemma 21 *We have $d(x_{i+1}, x_0) \leq d(x^*, x_0)$.*

Proof By the fact that x_{i+1} is the minimizer of $F^{(\mu_i)}$ we have $F^{(\mu_i)}(x_{i+1}) - F^{(\mu_i)}(x^*) \leq 0$. Note that by g-strong convexity, equality only holds if $x_{i+1} = x^*$ which only happens if $x_0 = x_{i+1} = x^*$. By using the definition of $F^{(\mu_i)}(x) = F(x) + \frac{\mu_i}{2}d(x, x_0)^2$ we have:

$$F(x_{i+1}) + \frac{\mu_i}{2}d(x_{i+1}, x_0)^2 - F(x^*) - \frac{\mu_i}{2}d(x^*, x_0)^2 \leq 0$$

$$\Rightarrow\ d(x_{i+1}, x_0) \leq d(x^*, x_0),$$

where in the last step we used the fact $F(x_{i+1}) - F(x^*) \geq 0$ that holds because x^* is the minimizer of F. ∎

We note that previous techniques proved and used the fact that $d(x_{i+1}, x^*) \leq d(x_0, x^*)$ instead (Allen Zhu and Hazan, 2016). But crucially, we need our former lemma in order to prove the bound for our non-Euclidean case. Our variant can be applied to (Allen Zhu and Hazan, 2016) to decrease the constants of their Euclidean reduction. Now we are ready to prove the theorem.

Proof of Theorem 9. We recall the definitions above. $F^{(\mu_i)}(x) = F(x) + \frac{\mu_i}{2}d(x, x_0)^2$. We start with $\hat{x}_0 = x_0$ and compute $\hat{x}_{i+1}$ using algorithm $\mathcal{A}$ with starting point $\hat{x}_i$ and function $F^{(\mu_i)}$ for time $\mathrm{Time}(L^{(i)}, \mu^{(i)}, \mathcal{M}, R)$, where $L^{(i)}$ and $\mu^{(i)}$ are the smoothness and strong g-convexity parameters of $F^{(\mu_i)}$. We denote by x_{i+1} the minimizer of $F^{(\mu_i)}$. We pick $\mu_i = \mu_{i-1}/2$ and we will choose later the value of μ_0 and the total number of stages. By the assumption of the theorem on $\mathcal{A}$, we have that

$$F^{(\mu_i)}(\hat{x}_{i+1}) - \min_{x \in \mathcal{M}} F^{(\mu_i)}(x) = F^{(\mu_i)}(\hat{x}_{i+1}) - F^{(\mu_i)}(x_{i+1}) \leq \frac{F^{(\mu_i)}(\hat{x}_i) - F^{(\mu_i)}(x_{i+1})}{4}. \quad (27)$$

36

Define $D_i \overset{\text{def}}{=} F^{(\mu_i)}(\hat{x}_i) - F^{(\mu_i)}(x_{i+1})$ to be the initial objective distance to the minimum on function $F^{(\mu_i)}$ before we call $\mathcal{A}$ for the $(i+1)$-th time. At the beginning, we have the upper bound $D_0 = F^{(\mu_0)}(\hat{x}_0) - \min_x F^{(\mu_0)}(x) \le F(x_0) - F(x^*)$. For each stage $i \ge 1$, we compute that

$$D_i = F^{(\mu_i)}(\hat{x}_i) - F^{(\mu_i)}(x_{i+1})$$

$$\overset{①}{=} F^{(\mu_{i-1})}(\hat{x}_i) - \frac{\mu_{i-1} - \mu_i}{2}d(x_0, \hat{x}_i)^2 - F^{(\mu_{i-1})}(x_{i+1}) + \frac{\mu_{i-1} - \mu_i}{2}d(x_0, x_{i+1})^2$$

$$\overset{②}{\le} F^{(\mu_{i-1})}(\hat{x}_i) - F^{(\mu_{i-1})}(x_i) + \frac{\mu_{i-1} - \mu_i}{2}d(x_0, x_{i+1})^2$$

$$\overset{③}{\le} \frac{D_{i-1}}{4} + \frac{\mu_i}{2}d(x_0, x_{i+1})^2$$

$$\overset{④}{\le} \frac{D_{i-1}}{4} + \frac{\mu_i}{2}d(x_0, x^*)^2.$$

Above, ① follows from the definition of $F^{(\mu_i)}(\cdot)$ and $F^{(\mu_{i-1})}(\cdot)$; ② follows from the fact that x_i is the minimizer of $F^{(\mu_{i-1})}(\cdot)$. And we dropped the negative term $-(\mu_{i-1} - \mu_i)d(x_0, \hat{x}_i)/2$. ③ follows from the definition of D_{i-1}, the assumption on $\mathcal{A}$, and the choice $\mu_i = \mu_{i-1}/2$ for $i \ge 1$; and ④ follows from Lemma 21. Now applying the above inequality recursively, we have

$$D_T \le \frac{D_0}{4^T} + d(x_0, x^*)^2 \cdot \left(\frac{\mu_T}{2} + \frac{\mu_{T-1}}{8} + \cdots\right) \le \frac{F(x_0) - F(x^*)}{4^T} + \mu_T \cdot d(x_0, x^*)^2. \qquad (28)$$

We have used the choice $\mu_i = \mu_{i-1}/2$ for the second inequality. Lastly, we can prove that $\hat{x}_T$, the last point computed, satisfies

$$F(\hat{x}_T) - F(x^*) \overset{①}{\le} F^{(\mu_T)}(\hat{x}_T) - F^{(\mu_T)}(x^*) + \frac{\mu_T}{2}d(x_0, x^*)^2$$

$$\overset{②}{\le} F^{(\mu_T)}(\hat{x}_T) - F^{(\mu_T)}(x_{T+1}) + \frac{\mu_T}{2}d(x_0, x^*)^2$$

$$\overset{③}{=} D_T + \frac{\mu_T}{2}d(x_0, x^*)^2$$

$$\overset{④}{\le} \frac{F(x_0) - F(x^*)}{4^T} + \frac{3\mu_T}{2}d(x_0, x^*)^2.$$

We use the definition of $F^{(\mu_T)}$ in ① and drop $-\frac{\mu_T}{2}d(x_0, \hat{x}_T)^2$. In ② we use the fact that x_{T+1} is the minimizer of $F^{(\mu_T)}$. The definition of D_T is used in ③. We use inequality (28) for step ④. Recall the assumption of the theorem $F(x_0) - F(x^*) \le \Delta$. Finally, by choosing $T = \lceil \log_2(\Delta/\varepsilon) \rceil + 1$ and $\mu_0 = \Delta/R^2$ we obtain that the point $\hat{x}_T$ satisfies

$$F(\hat{x}_T) - F(x^*) \le \frac{F(x_0) - F(x^*)}{4\Delta/\varepsilon} + \frac{3\mu_0}{8\Delta/\varepsilon}d(x_0, x^*)^2 \le \frac{\varepsilon}{4} + \frac{3\varepsilon}{8} < \varepsilon,$$

and can be computed in time $\sum_{t=0}^{T-1} \text{Time}(L + 2^{-t}\mu_0 \mathcal{K}_R^-, 2^{-t}\mu_0 \mathcal{K}_R^+, \mathcal{M}, R)$, since by Fact 20 the function $F^{(\mu_t)}$ is $L + 2^{-t}\mu_0 \mathcal{K}_R^-$ smooth and $2^{-t}\mu_0 \mathcal{K}_R^+$ g-strongly convex. ∎

B.2. Example 1

We use the algorithm in Corollary 8 as the algorithm $\mathcal{A}$ of the reduction of Theorem 9. Given a manifold under consideration $\mathcal{M}_K$, the assumption on $\mathcal{A}$ is satisfied for $\text{Time}(L, \mu, \mathcal{M}_K, R) = O(\sqrt{L/\mu} \log(L/\mu))$. Indeed, if Δ is a bound on the gap $\hat{F}(x_0) - \hat{F}(x^*) = \hat{F}(x_0) - \min_{x \in \mathcal{M}_K} \hat{F}(x) = \hat{F}(x_0) - \min_{x \in \text{Exp}_{x_0}(\bar{B}(0,R))} \hat{F}(x)$ for some μ-strongly g-convex $\hat{F}$, then we know that $d(x_0, x^*)^2 \leq \frac{2\Delta}{\mu}$ by μ-strong g-convexity. By calling the algorithm in Corollary 8 with $\varepsilon = \frac{\Delta}{4}$ we require a time that is

$$O(\sqrt{L/\mu} \log(L/\mu) \log(\mu \cdot d(x_0, x^*)^2/(\Delta/4))) = O(\sqrt{L/\mu} \log(L/\mu) \log(\mu \cdot (2\Delta/\mu)/(\Delta/4)))$$
$$= O(\sqrt{L/\mu} \log(L/\mu)).$$

Let $T = \lceil \log_2(\Delta/\varepsilon) \rceil + 1$. The reduction of Theorem 9 gives an algorithm with rates

$$\sum_{t=0}^{T-1} \text{Time}(L + 2^{-t}\mu_0 \mathcal{K}_R^-, 2^{-t}\mu_0 \mathcal{K}_R^+, \mathcal{M}_K, R)$$

$$\overset{\text{①}}{=} O\left(\sum_{t=0}^{T-1} \sqrt{\frac{\mathcal{K}_R^-}{\mathcal{K}_R^+} + \frac{L}{2^{-t}\mathcal{K}_R^+ \Delta/R^2}} \cdot \log\left(\frac{\mathcal{K}_R^-}{\mathcal{K}_R^+} + \frac{L}{2^{-t}\mathcal{K}_R^+ \Delta/R^2} \right) \right)$$

$$\overset{\text{②}}{=} O\left(\left(\sqrt{\frac{\mathcal{K}_R^-}{\mathcal{K}_R^+}} \log(\Delta/\varepsilon) + \sum_{t=0}^{T-1} \sqrt{\frac{L}{2^{-t}\mathcal{K}_R^+ \Delta/R^2}} \right) \log\left(\frac{\mathcal{K}_R^-}{\mathcal{K}_R^+} + \frac{L}{\mathcal{K}_R^+ \varepsilon} \right) \right)$$

$$\overset{\text{③}}{=} O\left(\left(\sqrt{\frac{\mathcal{K}_R^-}{\mathcal{K}_R^+}} \log(\Delta/\varepsilon) + \sqrt{\frac{L}{\mathcal{K}_R^+ \varepsilon}} \right) \log\left(\frac{\mathcal{K}_R^-}{\mathcal{K}_R^+} + \frac{L}{\mathcal{K}_R^+ \varepsilon} \right) \right)$$

$$\overset{\text{④}}{=} \widetilde{O}(\sqrt{L/\varepsilon})$$

In ① we write down the definition and use the value $\mu_0 = \Delta/R^2$. In ② we have used Minkowski's inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$. We added up the first group of summands. For the $\log$ factor, we upper bounded $L/(2^{-t}\mathcal{K}_R^+ \Delta/R^2) = O(L/\mathcal{K}_R^+ \varepsilon)$, for $t < T$. In ③ we used the fact that $\sqrt{1/\varepsilon} + \sqrt{1/2\varepsilon} + \cdots = O(\sqrt{1/\varepsilon})$, along with the fact $\varepsilon/2R^2 \leq 2^{-(T-1)}\mu_0 \leq \varepsilon/R^2$. Note that by L-smoothness and the diameter being $2R$, we have $\Delta \leq 2LR^2$ so $\sqrt{\mathcal{K}_R^-/\mathcal{K}_R^+} \log(\Delta/\varepsilon) = \widetilde{O}(1)$. We applied this in ④.

Appendix C. Geometric results. Proofs of Lemmas 2, 3 and 4

In this section we prove the lemmas that take into account the deformations of the geometry and the geodesic map h to obtain relationships between F and f. Namely Lemma 2, Lemma 3 and Lemma 4. First, we recall the characterizations of the geodesic map and some consequences. Then in Appendix C.2, Appendix C.3 and Appendix C.4, we prove the results related to distances angles and gradient deformations, respectively. That is, each of the three parts of Lemma 2. In Appendix C.4 we also prove Lemma 4, which comes naturally after the proof of Lemma 2.c. In Appendix C.5 we prove Lemma 3 and finish with a proof on lower bounds for the condition number of strongly g-convex functions and an intuitive comment on its implications. We conclude in Appendix D.1 with a remark about a hardness result proved in a previous work.

Before this, we note that we can assume without loss of generality that the curvature of our manifolds of interest can be taken to be $K \in \{1, -1\}$. One can see that the final rates depend on K through R, L and μ.

Remark 22 *For a function $F : \mathcal{M}_K \to \mathbb{R}$ where $\mathcal{M}_K$ is a manifold of constant sectional curvature $K \notin \{1, -1, 0\}$, we can apply a rescaling to the Gnomonic or Beltrami-Klein projection to define a function on a manifold of constant sectional curvature $K \in \{1, -1\}$. Namely, we can map $\mathcal{M}_K$ to M via the geodesic map $h : \mathcal{M}_K \to M$, then we can rescale M by multiplying each vector in M by the factor $\sqrt{|K|}$ and then we can apply the inverse geodesic map for the manifold of curvature $K \in \{1, -1\}$. If R is the original bound of the initial distance to an optimum, and F is L-smooth and μ-strongly g-convex (possibly with $\mu = 0$) then the initial distance bound becomes $\sqrt{|K|}R$ and the induced function becomes $L/|K|$-smooth and $\mu/|K|$-strongly g-convex. This is a consequence of the transformation rescaling distances by a factor of $\sqrt{|K|}$, i.e. if $r : \mathcal{M}_K \to \mathcal{M}_{K/|K|}$ is the rescaling function, then $d_K(x,y)\sqrt{|K|} = d_{K/|K|}(r(x), r(y))$, where $d_c(\cdot, \cdot)$ denotes the distance on the manifold of constant sectional curvature c.*

C.1. Preliminaries

We recall our characterization of the geodesic map. Given two points $\tilde{x}, \tilde{y} \in \mathcal{X}$, we have that $d(x, y)$, the distance between x and y with the metric of $\mathcal{M}_K$, satisfies

$$\mathrm{C}_K(d(x,y)) = \frac{1 + K\langle \tilde{x}, \tilde{y} \rangle}{\sqrt{1 + K\|\tilde{x}\|^2} \cdot \sqrt{1 + K\|\tilde{y}\|^2}}. \tag{29}$$

And since the expression is symmetric with respect to rotations, $\mathcal{X} = h(\mathcal{B}_R)$ is a closed ball of radius $\tilde{R}$, with $\mathrm{C}_K(R) = (1 + K\tilde{R}^2)^{-1/2}$. Equivalently,

$$\begin{aligned}
\tilde{R} &= \tan(R) \quad \text{if } K = 1, \\
\tilde{R} &= \tanh(R) \quad \text{if } K = -1.
\end{aligned} \tag{30}$$

Similarly, we can write the distances as

$$\begin{aligned}
d(x,y) &= \arccos\left(\frac{1 + \langle \tilde{x}, \tilde{y} \rangle}{\sqrt{1 + \|\tilde{x}^2\|}\sqrt{1 + \|\tilde{y}^2\|}} \right) \quad \text{if } K = 1, \\
d(x,y) &= \mathrm{arccosh}\left(\frac{1 - \langle \tilde{x}, \tilde{y} \rangle}{\sqrt{1 - \|\tilde{x}^2\|}\sqrt{1 - \|\tilde{y}^2\|}} \right) \quad \text{if } K = -1,
\end{aligned} \tag{31}$$

Alternatively, we have the following expression for the distance $d(x, y)$ when $K = -1$. Let $\tilde{a}, \tilde{b}$ be the two points of intersection of the ball $B(0, 1) \supseteq M$ with the line joining $\tilde{x}, \tilde{y}$, so the order of the points in the line is $\tilde{a}, \tilde{x}, \tilde{y}, \tilde{b}$. Then

$$d(x,y) = \frac{1}{2} \log\left(\frac{\|\tilde{a} - \tilde{y}\|\|\tilde{x} - \tilde{b}\|}{\|\tilde{a} - \tilde{x}\|\|\tilde{b} - \tilde{y}\|} \right) \text{ if } K = -1. \tag{32}$$

We will use this expression when working with the hyperbolic space. A simple elementary proof of the equivalence of the expressions in (31) and (32) when $K = -1$ is the following. We can

assume without loss of generality that we work with the hyperbolic plane, i.e. $n = 2$. By rotational symmetry, we can also assume that $\tilde{x} = (x_1, x_2)$ and $\tilde{y} = (y_1, y_2)$, for $x_1 = y_1$. In fact, it is enough to prove it in the case $x_2 = 0$ because we can split a general segment into two, each with one endpoint at $(x_1, 0)$, and then add their lengths up. So according to (31) and (32), respectively, we have

$$\frac{1}{\cosh^2(d(x,y))} = \frac{(1-x_1^2)(1-y_1^2-y_2^2)}{(1-x_1^2)^2} = \frac{(1-x_1^2-y_2^2)}{1-x_1^2},$$

$$d(x,y) = \frac{1}{2} \log \left(\frac{(\sqrt{1-y_1^2}+y_2)(\sqrt{1-x_1^2})}{(\sqrt{1-x_1^2})(\sqrt{1-y_1^2}-y_2)} \right) = \frac{1}{2} \log \left(\frac{1 + y_2/\sqrt{1-x_1^2}}{1 - y_2/\sqrt{1-x_1^2}} \right)$$

$$= \operatorname{arctanh} \left(\frac{y_2}{\sqrt{1-x_1^2}} \right).$$

where we have used the equality $\operatorname{arctanh}(t) = \frac{1}{2} \log(\frac{1+t}{1-t})$. Now, using the trigonometric identity $\frac{1}{\cosh^2(t)} = 1 - \tanh^2(t)$, for $t = d(x,y)$, we obtain that the two expressions above are equal. See Theorem 7.4 in (Greenberg, 1993) (p. 268) for more details about the distance formula under this geodesic map.

The spherical case is of a remarkable simplicity. If we have an (n)-sphere of radius 1 centered at 0, we can see the transformation of the geodesic map as the projection onto the plane $x_{n+1} = 1$. Given two points $\mathbf{x} = (\tilde{x}, 1)$, $\mathbf{y} = (\tilde{y}, 1)$ then the angle between these two vectors is the distance of the projected points on the sphere so we have $\cos(d(x,y)) = \langle \mathbf{x}, \mathbf{y} \rangle / \|\mathbf{x}\| \|\mathbf{y}\|$ which is equivalent to the corresponding formula in (31).

C.2. Distance deformation

Lemma 23 *Let $x, y \in \mathcal{B}_R = \operatorname{Exp}_{x_0}(\bar{B}(0, R)) \subseteq \mathcal{M}_K$ be two different points, where $\mathcal{M}_K$ is the hyperbolic space with constant sectional curvature $K = -1$. Then, we have*

$$1 \le \frac{d(x,y)}{\|\tilde{x} - \tilde{y}\|} \le \cosh^2(R).$$

Proof We can assume without loss of generality that the dimension is $n = 2$. As in (30), let $\tilde{R} = \tanh(R)$, so any point $\tilde{x} \in \mathcal{X}$ satisfies $\|\tilde{x}\| \le \tilde{R}$, or equivalently $d(x, x_0) \le R$. Recall $\tilde{x}_0 = h(x_0) = 0$. Without loss of generality, we parametrize an arbitrary segment of length ℓ in $\mathcal{X}$ by two endpoints $\tilde{x}, \tilde{y}$ with coordinates $\tilde{x} = (x_1, x_2)$ and $\tilde{y} = (x_1 - \ell, x_2)$, for $0 \le x_2 \le \tilde{R}$, $0 \le x_1 \le \sqrt{\tilde{R}^2 - x_2^2}$ and $0 < \ell \le x_1 + \sqrt{\tilde{R}^2 - x_2^2}$. Let $\mathfrak{d}(x_1, x_2, \ell) \stackrel{\text{def}}{=} \frac{d(x,y)}{\ell}$, the quantity we aim to bound. We will prove the upper bound on $\mathfrak{d}(x_1, x_2, \ell)$ in three steps.

1. If $x_1 = \ell$ then $\mathfrak{d}(\cdot)$ is larger the larger x_1 is. This allows to prove that it is enough to consider points with the extra constraint $\ell \le x_1$.

2. The partial derivative of $\mathfrak{d}(\cdot)$ with respect to x_1, whenever $\ell \le x_1$, is non-negative. So we can just look at the points for which $x_1 = \sqrt{\tilde{R}^2 - x_2^2}$.

3. With the constraints above, $\mathfrak{d}(\cdot)$ is larger the smaller ℓ is. So we have

$$\mathfrak{d}(x_1, x_2, \ell) \leq \lim_{\ell \to 0} \mathfrak{d}(\sqrt{\tilde{R}^2 - x_2^2}, x_2, \ell) = \sqrt{1 - x_2^2}/(1 - \tilde{R}^2).$$

This expression is maximized at $x_2 = 0$ and evaluates to $1/(1 - \tanh^2(R)) = \cosh^2(R)$.

We proceed now to prove the steps above. For the first step, we note

$$\mathfrak{d}(x_1, x_2, x_1) = \frac{1}{2x_1} \log\left(\frac{\sqrt{1 - x_2^2}(\sqrt{1 - x_2^2} + x_1)}{\sqrt{1 - x_2^2}(\sqrt{1 - x_2^2} - x_1)}\right) = \frac{1}{2x_1} \log\left(1 + \frac{2x_1}{\sqrt{1 - x_2^2} - x_1}\right).$$

We prove that the inverse of this expression is not increasing with respect to x_1. By taking a partial derivative:

$$\frac{\partial(1/\mathfrak{d}(x_1, x_2, x_1))}{\partial x_1} = 2\frac{\frac{-2x_1\sqrt{1-x_2^2}}{1-x_2^2-x_1^2} + \log(1 + 2x_1/(\sqrt{1 - x_2^2} - x_1))}{\log(1 + 2x_1/(\sqrt{1 - x_2^2} - x_1))^2} \overset{?}{\leq} 0$$

$$\iff \frac{2x_1\sqrt{1 - x_2^2}}{1 - x_2^2 - x_1^2} - \log(1 + (2x_1\sqrt{1 - x_2^2} + 2x_1^2)/(1 - x_2^2 - x_1^2)) \overset{?}{\geq} 0.$$

In order to see that the last inequality is true, note that the expression on the left hand side is 0 when $x_1 = x_2 = 0$. And the partial derivatives of this with respect to x_1 and x_2, respectively, are:

$$\frac{4\sqrt{1 - x_2^2}x_1^2}{(1 - x_2^2 - x_1^2)^2} \quad \text{and} \quad \frac{4x_2 x_1^3}{\sqrt{1 - x_2^2}(1 - x_2^2 - x_1^2)^2}.$$

Both are greater than 0 in the interior of the domain $0 \leq x_2 \leq \tilde{R}, 0 \leq x_1 \leq \sqrt{\tilde{R}^2 - x_2^2}$ and at least 0 in the border. Now we use this monotonicity to prove that we can consider $\ell \leq x_1$ only. Suppose $\ell > x_1$. The segment ℓ is divided into two parts by the e_2 axis and we can assume without loss of generality that the negative part is no greater than the other, i.e. $x_1 \geq \ell - x_1$. Otherwise, we can perform the computations after a symmetry over the e_2 axis. Let $\tilde{r}$ be the point $(0, x_2)$. We want to see that the segment from $\tilde{x}$ to $\tilde{r}$ gives a greater value of $\mathfrak{d}(\cdot)$:

$$\frac{d(x, r)}{x_1} \geq \frac{d(x, y)}{\ell} \iff d(x, r)(x_1 + (\ell - x_1)) \geq x_1(d(x, r) + d(r, y))$$

$$\iff d(x, r)/x_1 \geq d(r, y)/(\ell - x_1),$$

and the last inequality holds true by the monotonicity we just proved.

In order to prove the second step, we take the partial derivative of $\mathfrak{d}(x_1, x_2, \ell)$ with respect to x_1. We have

$$\mathfrak{d}(x_1, x_2, \ell) = \frac{1}{2\ell} \log\left(\frac{1 + \ell/(\sqrt{1 - x_2^2} - x_1)}{1 - \ell/\sqrt{1 - x_2^2} + x_1}\right),$$

$$\frac{\partial \mathfrak{d}(x_1, x_2, \ell)}{\partial x_1} = \frac{\sqrt{1 - d^2}(2x_1 - \ell)}{2(1 - x_2^2 - x_1^2)(1 - x_2^2 - (x_1 - \ell)^2)}.$$

And the derivative is positive in the domain we are considering.

41

We now prove step 3. We want to show that $\mathfrak{d}(\sqrt{\tilde{R}^2 - x_2^2}, x_2, \ell\cdot)$ decreases with ℓ, within our constraints $\ell \le x_1 = \sqrt{\tilde{R}^2 - x_2^2}$, $0 \le x_2 \le \tilde{R}$. If we split the segment joining $\tilde{x}$ and $\tilde{y}$ in half, with respect to the metric in $\mathcal{X}$, we see that due to the monotonicity proved in step 1, the segment that is farther to the origin is longer in $\mathcal{M}$ than the other one and so $\mathfrak{d}(\cdot)$ is greater for this half of the segment than for the original one. In symbols, let $\tilde{r}$ be the middle point of the segment joining $\tilde{x}$ and $\tilde{y}$. We have by monotonicity that $\mathfrak{d}(x_1, x_2, \ell/2) \ge \mathfrak{d}(x_1, x_2 - \ell/2, \ell/2)$. So $\mathfrak{d}(x_1, x_2, \ell/2) = \frac{d(\tilde{x}, \tilde{r})}{\ell/2} \ge \frac{d(\tilde{x}, \tilde{r}) + d(\tilde{r}, \tilde{y})}{\ell} = \mathfrak{d}(x_1, x_2, \ell)$. Thus,

$$\mathfrak{d}(x_1, x_2, \ell) \le \lim_{\ell \to 0} \mathfrak{d}(\sqrt{\tilde{R}^2 - x_2^2}, x_2, \ell)$$

$$= \lim_{\ell \to 0} \frac{1}{2\ell} \log \left(\frac{1 + \ell / \left(\sqrt{1 - x_2^2} - \sqrt{\tilde{R}^2 - x_2^2} \right)}{1 - \ell / \left(\sqrt{1 - x_2^2} + \sqrt{\tilde{R}^2 - x_2^2} \right)} \right)$$

$$\overset{①}{=} \lim_{\ell \to 0} \frac{\sqrt{1 - x_2^2}}{1 - \tilde{R}^2 - 2\ell \sqrt{\tilde{R}^2 - x_2^2} + \ell^2}$$

$$= \frac{\sqrt{1 - x_2^2}}{1 - \tilde{R}^2}.$$

We used L'Hôpital's rule for ①. We can maximize the last the result of the limit by setting $x_2 = 0$ and obtain that for any two different $\tilde{x}, \tilde{y} \in \mathcal{X}$

$$\frac{d(x, y)}{\|\tilde{x} - \tilde{y}\|} \le \frac{1}{1 - \tilde{R}^2} = \frac{1}{1 - \tanh^2(R)} = \cosh^2(R).$$

The lower bound is similar, assume that $\ell > x_1$ and define $\tilde{r}$ as above. We assume again without loss of generality that $x_1 \ge \ell - x_1$. Then

$$\frac{d(x, r) + d(r, y)}{\ell} \ge \frac{d(x, r)}{\ell - x_1} \iff \frac{d(r, y)}{x_1} \ge \frac{d(x, r)}{\ell - x_1}$$

and the latter is true by the monotonicity proved in step 1. This means that we can also consider $\ell \le x_1$. But this time, according to step 2, we want x_1 to be the lowest possible, so it is enough to consider $x_1 = \ell$. Using step 1 again, we obtain that the lowest value of $\mathfrak{d}(\cdot)$ can be bounded by the limit $\lim_{\ell \to 0} \mathfrak{d}(\ell, x_2, \ell)$ which using L'Hôpital's rule in ① is

$$\mathfrak{d}(x_1, x_2, \ell) \ge \lim_{\ell \to 0} \mathfrak{d}(\ell, x_2, \ell)$$

$$= \lim_{\ell \to 0} \frac{1}{2\ell} \log \left(1 + \frac{2\ell}{\sqrt{1 - x_2^2} - \ell} \right)$$

$$\overset{①}{=} \lim_{\ell \to 0} \frac{\frac{2(\sqrt{1 - x_2^2} - \ell) + 2\ell}{(\sqrt{1 - x_2^2} - \ell)^2}}{2(1 + 2\ell / (\sqrt{1 - x_2^2} - \ell))}$$

$$= \frac{1}{\sqrt{1 - x_2^2}}.$$

The expression is minimized at $x_2 = 0$ and evaluates to 1. ∎

The proof of the corresponding lemma for the sphere is analogous, we add it for completeness.

Lemma 24 *Let $x, y \in B_R = \mathrm{Exp}_{x_0}(\bar{B}(0, R)) \subseteq \mathcal{M}_K$ be two different points, where $\mathcal{M}_K$ is the spherical space with constant sectional curvature $K = 1$, and $R < \pi/2$. Then, we have*

$$\cos^2(R) \leq \frac{d(x, y)}{\|\tilde{x} - \tilde{y}\|} \leq 1.$$

Proof We proceed in a similar way than with the hyperbolic case. We can also work with $d = 2$ only, since $\tilde{x}, \tilde{y}$ and $\tilde{x}_0$ lie on a plane. We parametrize a general pair of points as $\tilde{x} = (x_1, x_2) \in \mathcal{X}$ and $y = (x_1 - \ell, x_2) \in \mathcal{X}$, so $x_1^2 + x_2^2 \leq \tilde{R}^2$, for $\tilde{R} = \tan(R)$ and by definition $\ell = \|\tilde{x} - \tilde{y}\|$.

Let $\mathfrak{d}(x_1, x_2, \ell) \overset{\text{def}}{=} d(x, y)/\|\tilde{x} - \tilde{y}\|$. We proceed to prove the result in three steps, similarly to the hyperbolic case.

1. If $x_1 = \ell$ then $\mathfrak{d}(x_1, x_2, \ell)$ decreases whenever x_1 increases. This allows to prove that it is enough to consider points in which $\ell \leq x_1$.

2. $\frac{\partial \mathfrak{d}(\cdot)}{\partial x_1} \leq 0$, whenever $\ell \leq x_1$. So we can consider $x_1 = \sqrt{\tilde{R}^2 - x_2^2}$ only.

3. With the constraints above, $\mathfrak{d}(\cdot)$ increases with ℓ, so in order to lower bound $\mathfrak{d}(\cdot)$ we can consider $\lim_{\ell \to 0} \mathfrak{d}(\sqrt{\tilde{R} - x_2}, x_2, \ell) = \sqrt{1 + x_2^2}/(1 + \tilde{R}^2)$. This is minimized at $x_2 = 0$ and evaluates to $1/(1 + \tilde{R}^2)$.

For the first step, we compute the partial derivative:

$$\frac{\partial \mathfrak{d}(x_1, x_2, x_1)}{\partial x_1} = \frac{x_1 \sqrt{1 + x_2^2}/(1 + x_1^2 + x_2^2) - \arccos\left(\sqrt{(1 + x_2^2)/(1 + x_1^2 + x_2^2)}\right)}{x_1^2}. \tag{33}$$

In order to see that it is non-positive, we compute the partial derivative of the denominator with respect to x_2 and obtain

$$\frac{2x_1^3 x_2}{\sqrt{1 + x_2^2}(1 + x_1^2 + x_2^2)} \geq 0.$$

so in order to maximize (33) we set $x_2 = \sqrt{\tilde{R} - x_1^2}$. In that case, the numerator is

$$\frac{x_1 \sqrt{1 + R^2 - x_1^2}}{1 + R^2} - \arccos\left(\sqrt{\frac{1 + R^2 - x_1^2}{1 + R^2}}\right), \tag{34}$$

and its derivative with respect to x_1 is

$$-\frac{2x_1^2}{(1 + R^2)\sqrt{1 + R^2 - x_1^2}} \leq 0.$$

and given that (34) with $x_1 = 0$ evaluates to 0 we conclude that (33) is non-positive. Similarly to Lemma 23, suppose the horizontal segment that joins $\tilde{x}$ and $\tilde{y}$ passes through $\tilde{r} \overset{\text{def}}{=} (0, x_2)$. And

43

suppose without loss of generality that $d(x, r) \geq d(r, y)$, i.e. $x_1 \geq \ell - x_1$. Then by the monotonicity we just proved, we have

$$\frac{d(x, r)}{\|\tilde{x} - \tilde{r}\|} = \mathfrak{d}(x_1, x_2, x_1) \leq \mathfrak{d}(\ell - x_1, x_2, \ell - x_1) = \frac{d(r, y)}{\|\tilde{r} - \tilde{y}\|}. \tag{35}$$

And this implies $\mathfrak{d}(x_1, x_2, x_1) \leq \mathfrak{d}(x_1, x_2, \ell)$. Indeed, that is equivalent to show

$$\frac{d(x, r)}{\|\tilde{x} - \tilde{r}\|} \leq \frac{d(x, y)}{\|\tilde{x} - \tilde{y}\|} = \frac{d(x, r) + d(r + y)}{\|\tilde{x} - \tilde{r}\| + \|\tilde{r} - \tilde{y}\|}.$$

Which is true, since after simplifying we arrive to (35). So in order to lower bound $\mathfrak{d}(\cdot)$, it is enough to consider $\ell \leq x_1$.

We focus on step 2 now. We have

$$\frac{\partial \mathfrak{d}(x_1, x_2, \ell)}{\partial x_1} = \frac{\sqrt{1 + x_2^2}(\ell - 2x_1)}{(1 + x_2^2 + (\ell - x_1)^2)(1 + x_2^2 + x_1^2)}.$$

which is non-positive given the restrictions we imposed after step 1. So in order to lower bound $\mathfrak{d}(\cdot)$ we can consider $x_1 = \sqrt{\tilde{R} - x_2^2}$ only.

Finally, in order to complete step 3 we compute

$$\frac{\partial \mathfrak{d}(\sqrt{\tilde{R} - x_2^2}, x_2, \ell)}{\partial \ell} = \frac{\sqrt{1 + x_2^2}}{\ell(1 + \tilde{R}^2) + \ell^3 - 2\ell^2 \sqrt{\tilde{R}^2 - x_2^2}}$$

$$- \frac{1}{\ell^2} \arccos \left(\frac{1 + \tilde{R}^2 - \ell\sqrt{\tilde{R}^2 - x_2^2}}{\sqrt{(1 + \tilde{R}^2)(1 + \tilde{R}^2 + \ell^2 - 2\ell\sqrt{\tilde{R}^2 - x_2^2})}} \right)$$

And in order to prove that this is non-negative, we will prove that the same expression is non-negative, when multiplied by ℓ^2. We compute the partial derivative of the aforementioned expression with respect to ℓ:

$$\frac{\partial}{\partial \ell} \left(\frac{\partial \mathfrak{d}(\sqrt{\tilde{R} - x_2^2}, x_2, \ell)}{\partial \ell} \ell^2 \right) = \frac{2\ell\sqrt{1 + x_2^2}(\sqrt{\tilde{R}^2 - x_2^2} - \ell)}{(1 + \tilde{R}^2 + \ell^2 - 2\ell\sqrt{\tilde{R}^2 - x_2^2})^2} \geq 0.$$

And $\ell^2(\partial \mathfrak{d}(\sqrt{\tilde{R} - x_2^2}, x_2, \ell)/\partial \ell)$ evaluated at 0 is 0 for all choices of parameters R and x_2 in the domain. So we conclude that $\partial \mathfrak{d}(\sqrt{\tilde{R} - x_2^2}, x_2, \ell)/\partial \ell \geq 0$.

Thus, we can consider the limit when $\ell \to 0$ in order to lower bound $\mathfrak{d}(\cdot)$. In the defined domain, we have

$$
\begin{aligned}
\lim_{\ell \to 0} \mathfrak{d}(\sqrt{\tilde{R} - x_2}, x_2, \ell) &= \lim_{\ell \to 0} \frac{1}{\ell} \arccos \left(\frac{1 + \tilde{R}^2 - x\sqrt{\tilde{R}^2 - x_2^2}}{\sqrt{1 + \tilde{R}^2}\sqrt{1 + x_2^2 + (\ell - \sqrt{\tilde{R}^2 - x_2^2})^2}} \right) \\
&\overset{\textcircled{1}}{=} \lim_{\ell \to 0} \frac{\sqrt{1 + x_2^2}}{1 + \tilde{R}^2 + \ell^2 - 2\ell\sqrt{\tilde{R}^2 - x_2^2}} \\
&= \frac{\sqrt{1 + x_2^2}}{1 + \tilde{R}^2}.
\end{aligned}
$$

We used L'Hôpital's rule for $\textcircled{1}$. Now, the right hand side of the previous expression is minimized at $x_2 = 0$ so we conclude that we have

$$
\cos^2(R) = \frac{1}{1 + \tan^2(R)} = \frac{1}{1 + \tilde{R}^2} \leq \mathfrak{d}(x_1, x_2, \ell) = \frac{d(p, q)}{\|\tilde{p} - \tilde{q}\|}.
$$

The upper bound uses again a similar argument. Assume that $\ell > x_1$ and define $\tilde{r}$ as above. We assume again without loss of generality that $x_1 \geq \ell - x_1$. Then

$$
\frac{d(x, r) + d(r, y)}{\ell} \leq \frac{d(x, r)}{\ell - x_1} \iff \frac{d(r, y)}{x_1} \leq \frac{d(x, r)}{\ell - x_1}
$$

and the latter is true by the monotonicity proved in step 1. Consequently we can just consider the points that satisfy $\ell \leq x_1$. By step 2, $\mathfrak{d}(\cdot)$ is maximal whenever x_1 is the lowest possible, so it is enough to consider $x_1 = \ell$. Using step 1 again, we obtain that the greatest value of $\mathfrak{d}(\cdot)$ can be bounded by the limit $\lim_{\ell \to 0} \mathfrak{d}(\ell, x_2, \ell)$ which using L'Hôpital's rule in $\textcircled{1}$ and simplifying is

$$
\begin{aligned}
\mathfrak{d}(x_1, x_2, \ell) &\leq \lim_{\ell \to 0} \mathfrak{d}(\ell, x_2, \ell) \\
&= \lim_{\ell \to 0} \frac{1}{\ell} \arccos \left(\sqrt{\frac{1 + x_2^2}{1 + \ell^2 + x_2^2}} \right) \\
&\overset{\textcircled{1}}{=} \frac{1}{\sqrt{1 + x_2^2}}.
\end{aligned}
$$

The expression is maximized at $x_2 = 0$ and evaluates to 1. ∎

C.3. Angle deformation

Lemma 25 *Let $x, y \in B_R = \operatorname{Exp}_{x_0}(\bar{B}(0, R)) \subseteq \mathcal{M}_K$ be two different points and different from x_0, where $\mathcal{M}_K$ is a manifold constant sectional curvature $K \in \{1, -1\}$, and if $K = 1$, then $R < \pi/2$. Let $\tilde{\alpha}$ be the angle $\angle x_0 x y$, formed by the vectors $x_0 - x$ and $y - x$. Let α be the corresponding angle between the vectors $\operatorname{Exp}_x^{-1}(x_0)$ and $\operatorname{Exp}_x^{-1}(y)$. The following holds:*

$$
\sin(\alpha) = \sin(\tilde{\alpha})\sqrt{\frac{1 + K\|\tilde{x}\|^2}{1 + K\|\tilde{x}\|^2 \sin^2(\tilde{\alpha})}}, \qquad \cos(\alpha) = \cos(\tilde{\alpha})\sqrt{\frac{1}{1 + K\|\tilde{x}\|^2 \sin^2(\tilde{\alpha})}}.
$$

Proof Note that we can restrict ourselves to $\alpha \in [0, \pi]$ because we have $\widetilde{(-w)} = -\tilde{w}$ (recall our notation about vectors with tilde). This means that the result for the range $\alpha \in [-\pi, 0]$ can be deduced from the result for $-\alpha$.

We start with the case $K = -1$. We can assume without loss of generality that the dimension is $n = 2$, and that the coordinates of $\tilde{x}$ are $(0, x_2)$, for $x_2 \le \tanh(R)$ that $\tilde{y} = (y_1, y_2)$, for some $y_1 \le 0$ and $\tilde{\delta} \stackrel{\text{def}}{=} \angle \tilde{y} \tilde{x}_0 \tilde{x} \in [0, \pi/2]$, since we can make the distance $\|\tilde{x} - \tilde{y}\|$ as small as we want. Recall $\tilde{x}_0 = \mathbb{0}_n$. We recall that $d(x, x_0) = \operatorname{arctanh}(\|\tilde{x}\|)$ and we note that $\sinh(\operatorname{arctanh}(t)) = \frac{t}{1-t^2}$, so that $\sinh(d(x, x_0)) = \|\tilde{x}\|/\sqrt{1 - \|\tilde{x}\|^2}$, for any $\tilde{x} \in \mathcal{X}$. We will apply the hyperbolic and Euclidean law of sines Fact 26 in order to compute the value of $\sin(\alpha)$ with respect to $\tilde{\alpha}$. Let $\tilde{a}$ and $\tilde{b}$ be points in the border of $B(0, 1)$ such the segment joining $\tilde{a}$ and $\tilde{b}$ is a chord that contains $\tilde{x}$ and $\tilde{y}$ and $\|\tilde{a} - \tilde{x}\| \le \|\tilde{b} - \tilde{x}\|$. So $\|\tilde{a} - \tilde{x}\|$ and $\|\tilde{b} - \tilde{x}\|$ are $\sqrt{1 - \|\tilde{x}\|^2 \sin^2(\tilde{\alpha})} - \|\tilde{x}\| \cos(\tilde{\alpha})$ and $\sqrt{1 - \|\tilde{x}\|^2 \sin^2(\tilde{\alpha})} + \|\tilde{x}\| \cos(\tilde{\alpha})$, respectively. We have

$$\sin(\alpha) \stackrel{\text{\textcircled{1}}}{=} \frac{\sinh(d(x_0, y)) \sin(\tilde{\delta})}{\sinh(d(x, y))} \stackrel{\text{\textcircled{2}}}{=} \frac{\|\tilde{x}_0 - \tilde{y}\|}{\sqrt{1 - \|\tilde{x}_0 - \tilde{y}\|^2}} \cdot \frac{\|\tilde{x} - \tilde{y}\| \sin(\tilde{\alpha})}{\|\tilde{x}_0 - \tilde{y}\|} \cdot \frac{1}{\sinh(d(x, y))}$$

$$\stackrel{\text{\textcircled{3}}}{=} \frac{\sin(\tilde{\alpha})}{\sqrt{1 - \|\tilde{x}\|^2 + \|\tilde{x} - \tilde{y}\|(-2\|\tilde{x}\| \cos(\tilde{\alpha}) + \|\tilde{x} - \tilde{y}\|)}} \cdot \frac{\|\tilde{x} - \tilde{y}\|}{\sinh(d(x, y))}$$

$$\stackrel{\text{\textcircled{4}}}{=} \frac{\sin(\tilde{\alpha})}{\sqrt{1 - \|\tilde{x}\|^2}} \lim_{d(x,y) \to 0} \|\tilde{x} - \tilde{y}\| \frac{1}{\sinh(d(x, y))}$$

$$\stackrel{\text{\textcircled{5}}}{=} \frac{\sin(\tilde{\alpha})}{\sqrt{1 - \|\tilde{x}\|^2}} \lim_{d(x,y) \to 0} \frac{(e^{2d(x,y)} - 1)(\|\tilde{a} - \tilde{x}\| \cdot \|\tilde{b} - \tilde{x}\|)}{e^{2d(x,y)}(\|\tilde{a} - \tilde{x}\| + \|\tilde{b} - \tilde{x}\|)} \cdot \frac{2e^{d(x,y)}}{e^{2d(x,y)} - 1}$$

$$= \frac{\sin(\tilde{\alpha})}{\sqrt{1 - \|\tilde{x}\|^2}} \cdot \frac{2\|\tilde{a} - \tilde{x}\| \cdot \|\tilde{b} - \tilde{x}\|}{\|\tilde{a} - \tilde{x}\| + \|\tilde{b} - \tilde{x}\|}$$

$$\stackrel{\text{\textcircled{6}}}{=} \frac{\sin(\tilde{\alpha})}{\sqrt{1 - \|\tilde{x}\|^2}} \cdot \frac{2(1 - \|\tilde{x}\|^2 \sin^2(\tilde{\alpha}) - \|\tilde{x}\|^2 \cos^2(\tilde{\alpha}))}{2\sqrt{1 - \|\tilde{x}\|^2 \sin^2(\tilde{\alpha})}} = \sin(\tilde{\alpha}) \sqrt{\frac{1 - \|\tilde{x}\|^2}{1 - \|\tilde{x}\|^2 \sin^2(\tilde{\alpha})}}.$$

In \textcircled{1} we used the hyperbolic sine theorem. In \textcircled{2} we used the expression above regarding segments that pass through the origin, and the Euclidean sine theorem. In \textcircled{3}, we simplify and use that the coordinates of $\tilde{y}$ are $(-\sin(\tilde{\alpha})\|\tilde{x} - \tilde{y}\|, \|\tilde{x}\| - \cos(\tilde{\alpha})\|\tilde{x} - \tilde{y}\|)$. Then, in \textcircled{4}, since $\sin(\alpha)$ does not depend on $\|\tilde{x} - \tilde{y}\|$, we can take the limit when $d(x, y) \to 0$, by which we mean we take the limit $\tilde{y} \to \tilde{x}$ by keeping the angle $\tilde{\alpha}$ constant. Since a posteriori the limit of each fraction exists, we compute them one at a time. \textcircled{5} uses (32) and the definition of $\sinh(d(x, y))$ for the last factor. The equality for the other factor can be checked with (32). In \textcircled{6} we substitute $\|\tilde{a} - \tilde{x}\|$ and $\|\tilde{b} - \tilde{x}\|$ by their values.

The spherical case is similar to the hyperbolic case. We also assume without loss of generality that the dimension is $n = 2$. Define $\tilde{y}$ as a point such that $\angle \tilde{x}_0 \tilde{x} \tilde{y} = \tilde{\alpha}$. We can assume without loss of generality that the coordinates of $\tilde{x}$ are $(0, x_2)$, that $\tilde{y} = (y_1, y_2)$, for $y_1 \le 0$, and $\tilde{\delta} \stackrel{\text{def}}{=} \angle \tilde{y} \tilde{x}_0 \tilde{x} \in [0, \pi/2]$, since we can make the distance $\|\tilde{x} - \tilde{y}\|$ as small as we want. We recall that by (30) we have $d(x_0, x) = \arctan(\|\tilde{x}_0 - \tilde{x}\|)$ and we note that $\sin(\arctan(t)) = \frac{t}{1+t^2}$, so that $\sin(d(x_0, x)) = \|\tilde{x}_0 - \tilde{x}\|/\sqrt{1 + \|\tilde{x}_0 - \tilde{x}\|^2}$, for any $\tilde{x} \in \mathcal{X}$. We will apply the spherical and Euclidean law of sines Fact 26 in order to compute the value of $\sin(\alpha)$ with respect to $\tilde{\alpha}$. We have

$$\sin(\alpha) \stackrel{\text{①}}{=} \frac{\sin(d(x_0,y))\sin(\tilde{\delta})}{\sin(d(x,y))} \stackrel{\text{②}}{=} \frac{\|\tilde{x}_0 - \tilde{y}\|}{\sqrt{1 + \|\tilde{x}_0 - \tilde{y}\|^2}} \cdot \frac{\|\tilde{x} - \tilde{y}\|\sin(\tilde{\alpha})}{\|\tilde{x}_0 - \tilde{y}\|} \frac{1}{\sin(d(x,y))}$$

$$\stackrel{\text{③}}{=} \frac{\sin(\tilde{\alpha})\|\tilde{x} - \tilde{y}\|}{\sqrt{1 + \|\tilde{x}_0 - \tilde{y}\|^2}\sqrt{1 - \frac{(1 - \|x\|\cos(\tilde{\alpha})\|\tilde{x} - \tilde{y}\| + \|\tilde{x}\|^2)^2}{(1 + \|\tilde{x}\|^2)(1 + \|\tilde{x}_0 - \tilde{y}\|^2)}}}$$

$$\stackrel{\text{④}}{=} \frac{\sin(\tilde{\alpha})\|\tilde{x} - \tilde{y}\|}{\sqrt{\|\tilde{x} - \tilde{y}\|^2(1 + \|\tilde{x}\|^2 - \|\tilde{x}\|^2\cos^2(\tilde{\alpha}))/(1 + \|\tilde{x}\|^2)}} \stackrel{\text{⑤}}{=} \sin(\tilde{\alpha})\sqrt{\frac{1 + \|\tilde{x}\|^2}{1 + \|\tilde{x}\|^2\sin^2(\tilde{\alpha})}}.$$

In ① we used the spherical sine theorem. In ② we used the expression above regarding segments that pass through the origin, and the Euclidean sine theorem. In ③, we use the fact that the coordinates of $\tilde{y}$ are $(-\sin(\tilde{\alpha})\|\tilde{x} - \tilde{y}\|, \|\tilde{x}\|^2 - \cos(\tilde{\alpha})\|\tilde{x} - \tilde{y}\|)$, use the distance formula (31) and the trigonometric equality $\sin(\arccos(x)) = \sqrt{1 - x^2}$. Then, in ④ and ⑤, we multiply and simplify.

Finally, in both cases, the cosine formula is derived from the identity $\sin^2(\alpha) + \cos^2(\alpha) = 1$ after noticing that the sign of $\cos(\alpha)$ and the sign of $\cos(\tilde{\alpha})$ are the same. The latter fact can be seen to hold true by noticing that α is monotonous with respect to $\tilde{\alpha}$ and the fact that $\tilde{\alpha} = \pi/2$ implies $\sin(\alpha) = 0$. ∎

Fact 26 (Constant Curvature non-Euclidean Law of Sines and Law of Cosines) *Let $K \neq 0$ and let $\mathrm{S}_K(\cdot)$ and $\mathrm{C}_K(\cdot)$ denote the special sine and cosine, respectively, defined as $\mathrm{S}_K(t) = \sin(\sqrt{K}t)$ and $\mathrm{C}_K(t) = \cos(\sqrt{K}t)$ if $K > 0$, and as $\mathrm{S}_K(t) = \sinh(\sqrt{-K}t)$ and $\mathrm{C}_K(t) = \cosh(\sqrt{-K}t)$ if $K < 0$. Let a, b, c be the lengths of the sides of a geodesic triangle defined on a manifold of constant sectional curvature K. Let α, β, γ be the angles of the geodesic triangle, that are opposite to the sides a, b, c. The following holds:*

- *Law of sines:*

$$\frac{\sin(\alpha)}{\mathrm{S}_K(a)} = \frac{\sin(\beta)}{\mathrm{S}_K(b)} = \frac{\sin(\gamma)}{\mathrm{S}_K(c)}.$$

- *Law of cosines:*

$$\mathrm{C}_K(a) = \mathrm{C}_K(b)\,\mathrm{C}_K(c) + \cos(\alpha)\,\mathrm{S}_K(b)\,\mathrm{S}_K(c).$$

We refer to (Greenberg, 1993) for a proof of these classical theorems.

C.4. Gradient deformation and smoothness of f

Lemma 25, with $\tilde{\alpha} = \pi/2$, shows that $e_1 \perp e_j$, for $j \neq 1$. The rotational symmetry implies $e_i \perp e_j$ for $i \neq j$ and $i, j > 1$. As in Lemma 2, let $x \in \mathcal{B}_R$ be a point and assume without loss of generality that $\tilde{x} \in \mathrm{span}\{\tilde{e}_1\}$ and $\nabla f(\tilde{x}) \in \mathrm{span}\{\tilde{e}_1, \tilde{e}_2\}$. It can be assumed without loss of generality because of the symmetries. So we can assume the dimension is $n = 2$. Using Lemma 2 we obtain that $\tilde{\alpha} = 0$ implies $\alpha = 0$. Also $\tilde{\alpha} = \pi/2$ implies $\alpha = \pi/2$, so the adjoint of the differential of h^{-1} at x, $(\mathrm{d}h^{-1})_x^*$ diagonalizes and has e_1 and e_2 as eigenvectors. We only need to compute the eigenvalues. The computation of the first one uses that the geodesic passing from x_0 and x can be parametrized as $h^{-1}(\tilde{x}_0 + \arctan(\tilde{\lambda}\tilde{e}_1))$ if $K = 1$ and $h^{-1}(\tilde{x}_0 + \mathrm{arctanh}(\tilde{\lambda}\tilde{e}_1))$ if $K = -1$, by (29).

The derivative of $\arctan(\cdot)$ or $\operatorname{arctanh}(\cdot)$ reveals that the first eigenvector, the one corresponding to e_1, is $1/(1 + K\|\tilde{x}^2\|)$, i.e. $\nabla f(\tilde{x})_1 = \nabla F(x)_1/(1 + K\|\tilde{x}^2\|)$. For the second one, let $x = (x_1, 0)$ and $y = (y_1, y_2)$, with $y_1 = x_1$ the second eigenvector results from the computation, for $K = -1$:

$$\lim_{y_2 \to 0} \frac{d(x,y)}{y_2} = \lim_{y_2 \to 0} \frac{1}{2y_2} \log\left(1 + \frac{2y_2}{\sqrt{1 - x_1^2 - y_2}}\right)$$

$$\overset{\textcircled{1}}{=} \lim_{y_2 \to 0} \frac{\frac{2}{\sqrt{1-x_1^2-y_2}} + \frac{2y_2}{(\sqrt{1-x_1^2-y_2})^2}}{2 + \frac{4y_2}{\sqrt{1-x_1^2-y_2}}}$$

$$= \frac{1}{\sqrt{1 - x_1^2}},$$

and for $K = 1$:

$$\lim_{y_2 \to 0} \frac{d(x,y)}{y_2} = \lim_{y_2 \to 0} \frac{1}{y_2} \arccos\left(\frac{\sqrt{1 + x_1^2}}{\sqrt{1 + x_1^2 + y_2^2}}\right)$$

$$\overset{\textcircled{2}}{=} \lim_{y_2 \to 0} \frac{\sqrt{1 + x_1^2}}{1 + x_1^2 + y_2^2}$$

$$= \frac{1}{\sqrt{1 + x_1^2}}.$$

So, since $x_1 = \|\tilde{x}\|$, we have $\nabla f(\tilde{x})_2 = \nabla F(x)_2/\sqrt{1 + K\|\tilde{x}\|^2}$ for $K \in \{1, -1\}$. We used L'Hôpital's rule in $\textcircled{1}$ and $\textcircled{2}$.

Also note that if $v \in T_x\mathcal{M}_K$ is a vector normal to $\nabla F(x)$, then $\tilde{v}$ is normal to $\nabla f(x)$. It is easy to see this geometrically: Indeed, no matter how h changes the geometry, since it is a geodesic map, a geodesic in the direction of first-order constant increase of F is mapped via h to a geodesic in the direction of first-order constant increase of f. And the respective gradients must be perpendicular to all these directions. Alternatively, this can be seen algebraically. Suppose first $n = 2$, then v is proportional to $(\nabla F(x)_2, -\nabla F(x)_1) = (\sqrt{1 + K\|\tilde{x}\|^2}\nabla f(\tilde{x})_2, -(1 + K\|\tilde{x}\|^2)\nabla f(\tilde{x}_1))$. And a vector $\tilde{v}'$ normal to $\nabla f(x)$ must be proportional to $(-\nabla f(x)_2, \nabla f(x)_1)$. Let α be the angle formed by v and $-e_1$, $\tilde{\alpha}$ the corresponding angle formed between $\tilde{v}$ and $-\tilde{e}_1$, and $\tilde{\alpha}'$ the angle formed by $\tilde{v}'$ and $-\tilde{e}_1$. Then we have, using the expression for the vectors proportional to v and $\tilde{v}'$:

$$\sin(\alpha) = \frac{-f(x)_2}{\sqrt{\nabla f(x)_2^2 + (1 + \|x\|^2)\nabla f(x)_1^2}} \text{ and } \sin(\tilde{\alpha}') = \frac{-f(x)_2}{\sqrt{\nabla f(x)_2^2 + \nabla f(x)_1^2}}$$

and using one equation on the other yields $\sin(\alpha) = \sin(\tilde{\alpha}')\sqrt{(1 + K\|\tilde{x}^2\|)/(1 + K\|\tilde{x}^2\|\sin^2(\tilde{\alpha}'))}$, which after applying Lemma 25 we obtain $\sin(\tilde{\alpha}') = \sin(\tilde{\alpha})$ from which we conclude that $\tilde{\alpha}' = \tilde{\alpha}$ given that the angles are in the same quadrant. So $\tilde{v} \perp \nabla f(x)$. In order to prove this for $n \geq 3$ one can apply the reduction (42) to the case $n = 2$ that we obtain in the next section.

Combining the results obtained so far in Appendix C, we can prove Lemma 2. We continue by proving Lemma 4, which will generalize the computations we just performed, in order to analyze the Hessian of f and provide smoothness. Then, in the next section, we combine the results in Lemma 2 to prove Lemma 3.

Proof of Lemma 2. The lemma follows from Lemmas 23, 24, 25 and the previous reasoning in this Section C.4. ∎

Proof of Lemma 4. Recall $F : \mathcal{M}_K \to \mathbb{R}$ is a function defined on a manifold of constant sectional curvature with a point $x^* \in \mathcal{B}_R$ such that $\nabla F(x^*) = 0$ and we call R an upper bound on $d(x_0, x^*)$, for an initial point x_0.

We will compute the Hessian of $f = F \circ h^{-1}$ and we will bound its spectral norm for any point $\tilde{x} \in \mathcal{X}$. We can assume without loss of generality that $n = 2$ and $\tilde{x} = (\tilde{\ell}, 0)$, for $\tilde{\ell} > 0$ (the case $\tilde{\ell} = 0$ is trivial), since there is a rotational symmetry with e_1 as axis. This means that by rotating we could align the top eigenvector of the Hessian at a point so that it is in $\mathrm{span}\{e_1, e_2\}$. Let $\tilde{y} = (y_1, y_2) \in \mathcal{X}$ be another point, with $y_1 = \tilde{\ell}$. We can also assume that $y_2 > 0$ without loss of generality, because of our symmetry. Our approach will be the following. We know by Lemma 2.b and by the beginning of this Section C.4 that the adjoint of the differential of h^{-1} at y, $(\mathrm{d}h^{-1})^*_y$ has $\mathrm{Exp}_y^{-1}(x_0)$ and a normal vector to it as eigenvectors. Their corresponding eigenvalues are $1/(1 + K\|\tilde{y}\|^2)$ and $1/\sqrt{1 + K\|\tilde{y}\|^2}$, respectively. Consider the basis $\{e_1, e_2\}$ of $T_x \mathcal{M}_K$ as defined at the beginning of this section, i.e. where e_1 is a unit vector proportional to $-\mathrm{Exp}_x^{-1}(x_0)$ and e_2 is the normal vector to e_1 that makes the basis orthonormal. Consider this basis being transported to y using parallel transport and denote the result $\{v_y, v_y^\perp\}$. Assume we have the gradient $\nabla F(y)$ written in this basis. Then we can compute the gradient of f at y by applying $(\mathrm{d}h^{-1})^*_y$ to $\nabla F(y)$. In order to do that, we compose the change of basis from $\{v_y, v_y^\perp\}$ to the basis of eigenvectors of $(\mathrm{d}h^{-1})^*_y$, then we apply a diagonal transformation given by the eigenvalues and finally we change the basis to $\{\tilde{e}_1, \tilde{e}_2\}$. Once this is done, we can differentiate with respect to y_2 in order to compute a column of the Hessian. Let $\tilde{\alpha}$ be the angle formed by the vectors $\tilde{y}$ and $\tilde{x}$. Note that $\tilde{\alpha} = \arctan(y_2/y_1)$. Let $\tilde{\gamma}$ be the angle formed by the vectors $(\tilde{y} - \tilde{x})$ and $-\tilde{y}$. That is, the angle $\tilde{\gamma} = \pi - \angle \tilde{x}\tilde{y}\tilde{x}_0$. Since $v_y^\perp$ is the parallel transport of $e_2^\perp$, the angle between $v_y^\perp$ and the vector $\mathrm{Exp}_y^{-1}(x_0)$ is γ. Note we use the same convention as before for the angles, i.e. γ is the corresponding angle to $\tilde{\gamma}$, meaning that if γ is the angle between two intersecting geodesics in $\mathcal{B}_R$, then $\tilde{\gamma}$ is the angle between the respective corresponding geodesics in $\mathcal{X}$. Note the first change of basis is a rotation and that the angle of rotation is $\gamma - \pi/2$. The last change of basis is a rotation with angle equal to the angle formed by a vector $\tilde{v}$ normal to $-\tilde{y}$ ($\tilde{v}$ is the one such that $-\tilde{y} \times \tilde{v} > 0$) and the vector $\tilde{e}_2$. This vector is equal to $\tilde{\alpha}$. So we have

$$\nabla f(y) = \begin{pmatrix} \cos(\tilde{\alpha}) & -\sin(\tilde{\alpha}) \\ \sin(\tilde{\alpha}) & \cos(\tilde{\alpha}) \end{pmatrix} \begin{pmatrix} \frac{1}{1+K(y_1^2+y_2^2)} & 0 \\ 0 & \frac{1}{\sqrt{1+K(y_1^2+y_2^2)}} \end{pmatrix} \begin{pmatrix} \sin(\gamma) & -\cos(\gamma) \\ \cos(\gamma) & \sin(\gamma) \end{pmatrix} \nabla F(y) \tag{36}$$

We want to take the derivative of this expression with respect to y_2 and we want to evaluate it at $y_2 = 0$. Let the matrices above be A, B and C so that $\nabla f(y) = ABC\nabla F(y)$. Using Lemma 2.b we have

$$\sin(\gamma) = \sin(\tilde{\gamma})\sqrt{\frac{1+K(y_1^2+y_2^2)}{1+K(y_1^2+y_2^2)\sin^2(\tilde{\gamma})}} \overset{①}{=} \cos(\tilde{\alpha})\sqrt{\frac{1+K(y_1^2+y_2^2)}{1+K(y_1^2+y_2^2)\cos^2(\tilde{\alpha})}},$$

$$\cos(\gamma) = -\sin(\tilde{\alpha})\sqrt{\frac{1}{1+K(y_1^2+y_2^2)\cos^2(\tilde{\alpha})}}, \tag{37}$$

where ① follows from $\sin(\tilde{\gamma}) = \sin(\tilde{\alpha} + \pi/2) = \cos(\tilde{\alpha})$. Now we compute some quantities

$$A|_{y_2=0} = I, \ B|_{y_2=0} = \begin{pmatrix} \frac{1}{1+Ky_1^2} & 0 \\ 0 & \frac{1}{\sqrt{1+Ky_1^2}} \end{pmatrix}, \ C|_{y_2=0} = I,$$

$$\frac{\partial A}{\partial y_2}\bigg|_{y_2=0} = \frac{\partial \tilde{\alpha}}{\partial y_2}\bigg|_{y_2=0} \cdot \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \overset{①}{=} \begin{pmatrix} 0 & \frac{-1}{y_1} \\ \frac{1}{y_1} & 0 \end{pmatrix},$$

$$\frac{\partial B}{\partial y_2}\bigg|_{y_2=0} = \begin{pmatrix} \frac{2Ky_2}{(1+K(y_1^2+y_2^2))^2} & 0 \\ 0 & \frac{2Ky_2}{2(1+K(y_1^2+y_2^2))^{3/2}} \end{pmatrix}\bigg|_{y_2=0} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix},$$

$$\frac{\partial C}{\partial y_2}\bigg|_{y_2=0} \overset{②}{=} \begin{pmatrix} 0 & \frac{1}{y_1\sqrt{1+Ky_1^2}} \\ \frac{-1}{y_1\sqrt{1+Ky_1^2}} & 0 \end{pmatrix}.$$

Equalities ① and ② follow after using (37), $\tilde{\alpha} = \arctan(\frac{y_2}{y_1})$ and taking derivatives. Now we differentiate (36) with respect to y_2 and evaluate to $y_2 = 0$ using the chain rule. The result is

$$\begin{pmatrix} \nabla^2 f(\tilde{x})_{12} \\ \nabla^2 f(\tilde{x})_{22} \end{pmatrix} = \left(\frac{\partial A}{\partial y_2} BC\nabla F(x) + A\frac{\partial B}{\partial y_2} C\nabla F(x) + AB\frac{\partial C}{\partial y_2}\nabla F(x) + ABC\frac{\partial \nabla F(x)}{\partial y_2} \right)\bigg|_{y_2=0}$$

$$= \begin{pmatrix} \frac{-\nabla f(\tilde{x})_2}{y_1\sqrt{1+Ky_1^2}} \\ \frac{\nabla f(\tilde{x})_1}{y_1(1+Ky_1^2)} \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \end{pmatrix} + \begin{pmatrix} \frac{\nabla f(\tilde{x})_2}{y_1(1+Ky_1^2)^{3/2}} \\ \frac{-\nabla f(\tilde{x})_1}{y_1(1+Ky_1^2)} \end{pmatrix} + \begin{pmatrix} \frac{\nabla^2 F(x)_{12}}{(1+Ky_1^2)^{3/2}} \\ \frac{\nabla^2 F(x)_{22}}{1+Ky_1^2} \end{pmatrix}$$

Computing the other column of the Hessian is easier. We can just consider (36) with $\tilde{\alpha} = 0$, $\gamma = \pi/2$ and vary y_1. Taking derivatives with respect to y_1 gives

$$\begin{pmatrix} \nabla^2 f(\tilde{x})_{11} \\ \nabla^2 f(\tilde{x})_{21} \end{pmatrix} = \begin{pmatrix} \frac{-2Ky_1\nabla f(\tilde{x})_1}{(1+Ky_1^2)^2} \\ \frac{-Ky_1\nabla f(\tilde{x})_2}{(1+Ky_1^2)^{3/2}} \end{pmatrix} + \begin{pmatrix} \frac{\nabla^2 F(x)_{11}}{(1+Ky_1^2)^2} \\ \frac{\nabla^2 F(x)_{21}}{(1+Ky_1^2)^{3/2}} \end{pmatrix}.$$

Note in the computations of both of the columns of the Hessian we have used

$$\frac{\partial \nabla F(y)_i}{\partial y_1} = \nabla F(x)_{i1} \cdot \frac{1}{1+Ky_1^2} \quad \text{and} \quad \frac{\partial \nabla F(y)_i}{\partial y_2}\bigg|_{y_2=0} = \nabla F(x)_{i2} \cdot \frac{1}{\sqrt{1+Ky_1^2}},$$

for $i = 1, 2$. The eigenvalues of the adjoint of the differential of h^{-1} appear as a factor because we are differentiating with respect to the geodesic in $\mathcal{X}$ which moves at a different speed than the corresponding geodesic in $\mathcal{B}_R$. Note as well, as a sanity check, that the cross derivatives are equal, since

$$-\frac{1}{y_1\sqrt{1+Ky_1^2}} + \frac{1}{y_1(1+Ky_1^2)^{3/2}} = \frac{1}{y_1\sqrt{1+Ky_1^2}}\left(-1 + \frac{1}{1+Ky_1^2}\right) = \frac{-Ky_1}{(1+Ky_1^2)^{3/2}}.$$

Finally, we bound the new smoothness constant $\tilde{L}$ by bounding the spectral norm of this Hessian. First note that using $y_1 = \tilde{\ell}$ we have that $\frac{1}{\sqrt{1+K\tilde{\ell}^2}} = \mathrm{C}_K(\ell)$ and then for $K = -1$ it is $\tilde{\ell} = \tanh(\ell)$ and for $K = 1$ it is $\tilde{\ell} = \tan(\ell)$, where $\ell = d(x, x_0) < R$. We have that since there is a point

$x^* \in \mathcal{B}_R$ such that $\nabla F(x^*) = 0$ and F is L-smooth, then it is $\|\nabla F(x)\| \leq 2LR$. Similarly, by L-smoothness $|\nabla^2 F(x)_{ij}| \leq L$. We are now ready to prove $\tilde{L}$-smoothness of f.

$$
\begin{aligned}
\tilde{L}^2 &\leq \max_{\tilde{x} \in \mathcal{X}} \|\nabla^2 f(\tilde{x})\|_2^2 \\
&\leq \max_{\tilde{x} \in \mathcal{X}} \|\nabla^2 f(\tilde{x})\|_F^2 = \max_{\tilde{x} \in \mathcal{X}} \{\nabla^2 f(\tilde{x})_{11} + 2\nabla^2 f(\tilde{x})_{12} + \nabla^2 f(\tilde{x})_{22}\} \\
&\leq L^2([\mathrm{C}_K^4(R) + 4R\,\mathrm{S}_K(R)\,\mathrm{C}_K^3(R)]^2 + 2[\mathrm{C}_K^3(R) + 2R\,\mathrm{S}_K(R)\,\mathrm{C}_K^2(R)]^2 + \mathrm{C}_K^4(R))
\end{aligned}
$$

and this can be bounded by $44L^2 \max\{1, R^2\}$ if $K = 1$ and $44L^2 \max\{1, R^2\}\,\mathrm{C}_K^8(R)$ if $K = -1$. In any case, it is $O(L^2)$. We note that for tilted convex functions we have that gradient Lipschitzness, smoothness and bounded Hessian are equivalent properties. Indeed, this is a classical result for convex functions and the proof of the implication that does not hold for general differentiable functions (showing that smoothness implies gradient Lipschitzness) only requires that a point with zero gradient is a global minimizer, cf. (Zhou, 2018) for instance. This property is true for tilted convex functions. ■

C.5. Proof of Lemma 3

Proof Assume for the moment that the dimension is $n = 2$. We can assume without loss of generality that $\tilde{x} = (\tilde{\ell}, 0)$. We are given two vectors, that are the gradients $\nabla F(x)$, $\nabla f(\tilde{x})$ and a vector $w \in T_x \mathcal{M}_K$. Let $\tilde{\delta}$ be the angle between $\tilde{w}$ and $-\tilde{x}$. Let δ be the corresponding angle, i.e. the angle between w and $u \overset{\text{def}}{=} \mathrm{Exp}_x^{-1}(x_0)$. Let α be the angle in between $\nabla F(x)$ and u. Let $\tilde{\beta}$ be the angle in between $\nabla f(\tilde{x})$ and $-x$. $\tilde{\alpha}$ and β are defined similarly. We claim

$$
\frac{\left\langle \frac{\nabla F(x)}{\|\nabla F(x)\|}, \frac{w}{\|w\|} \right\rangle}{\left\langle \frac{\nabla f(\tilde{x})}{\|\nabla f(\tilde{x})\|}, \frac{\tilde{w}}{\|\tilde{w}\|} \right\rangle} = \sqrt{\frac{1 + K\tilde{\ell}^2}{(1 + K\tilde{\ell}^2 \sin^2(\tilde{\delta}))(1 + K\tilde{\ell}^2 \cos^2(\tilde{\beta}))}}. \tag{38}
$$

Let's see how to arrive to this expression. By Lemma 2.c we have

$$
\tan(\alpha) = \frac{\tan(\tilde{\beta})}{\sqrt{1 + K\tilde{\ell}^2}}. \tag{39}
$$

From this relationship we deduce

$$
\cos(\alpha) = \cos(\tilde{\beta}) \sqrt{\frac{1 + K\tilde{\ell}^2}{1 + K\tilde{\ell}^2 \cos^2(\tilde{\beta})}}, \tag{40}
$$

that comes from squaring (39), reorganizing terms and noting that $\text{sign}(\cos(\alpha)) = \text{sign}(\cos(\tilde{\beta}))$ which is implied by Lemma 2.c. We are now ready to prove the claim (38) (for $n = 2$). We have

$$
\frac{\left\langle \frac{\nabla F(x)}{\|\nabla F(x)\|}, \frac{w}{\|w\|} \right\rangle}{\left\langle \frac{\nabla f(\tilde{x})}{\|\nabla f(\tilde{x})\|}, \frac{\tilde{w}}{\|\tilde{w}\|} \right\rangle} = \frac{\cos(\alpha - \delta)}{\cos(\tilde{\beta} - \tilde{\delta})}
$$

$$
\stackrel{②}{=} \frac{\cos(\delta) + \tan(\alpha)\sin(\delta)}{\cos(\tilde{\beta})\cos(\tilde{\delta}) + \sin(\tilde{\beta})\sin(\tilde{\delta})}\cos(\alpha)
$$

$$
\stackrel{③}{=} \frac{\frac{\cos(\tilde{\delta})}{\sqrt{1+K\tilde{\ell}^2\sin^2(\tilde{\delta})}} + \frac{\tan(\tilde{\beta})}{\sqrt{1+K\tilde{\ell}^2}}\frac{\sin(\tilde{\delta})\sqrt{1+K\tilde{\ell}^2}}{\sqrt{1+K\tilde{\ell}^2\sin^2(\tilde{\delta})}}}{\cos(\tilde{\beta})\cos(\tilde{\delta}) + \sin(\tilde{\beta})\sin(\tilde{\delta})}\cos(\tilde{\beta})\sqrt{\frac{1+K\tilde{\ell}^2}{1+K\tilde{\ell}^2\cos^2(\tilde{\beta})}}
$$

$$
\stackrel{④}{=} \sqrt{\frac{1+K\tilde{\ell}^2}{(1+K\tilde{\ell}^2\sin^2(\tilde{\delta}))(1+K\tilde{\ell}^2\cos^2(\tilde{\beta}))}}.
$$

Equality ① follows by the definition of $\alpha, \delta, \tilde{\delta}$, and $\tilde{\beta}$. In ②, we used trigonometric identities. In ③ we used Lemma 2.b, (39) and (40). By reordering the expression, the denominator cancels out with a factor of the numerator in ④.

In order to work with arbitrary dimension, we note it is enough to prove it for $n = 3$, since in order to bound

$$
\frac{\left\langle \frac{\nabla F(x)}{\|\nabla F(x)\|}, \frac{v}{\|v\|} \right\rangle}{\left\langle \frac{\nabla f(\tilde{x})}{\|\nabla f(\tilde{x})\|}, \frac{\tilde{v}}{\|\tilde{v}\|} \right\rangle},
$$

it is enough to consider the following submanifold

$$
\mathcal{M}'_K \stackrel{\text{def}}{=} \text{Exp}_x(\text{span}\{v, \text{Exp}_x^{-1}(x_0), \nabla F(x)\}).
$$

for an arbitrary vector $v \in T_x\mathcal{M}_K$ and a point x defined as above. The case $n = 3$ can be further reduced to the case $n = 2$ in the following way. Suppose $\mathcal{M}'_K$ is a three dimensional manifold (if it is one or two dimensional there is nothing to do). Define the following orthonormal basis of $T_x\mathcal{M}_K$, $\{e_1, e_2, e_3\}$ where $e_1 = -\text{Exp}_x^{-1}(x_0)/\|\text{Exp}_x^{-1}(x_0)\|$, e_2 is a unit vector, normal to e_1 such that $e_2 \in \text{span}\{e_1, \nabla F(x)\}$. And e_3 is a vector that completes the orthonormal basis. In this basis, let v be parametrized by $\|v\|(\sin(\delta), \cos(\nu)\cos(\delta), \sin(\nu)\cos(\delta))$, where δ can be thought as the angle between the vector v and its projection onto the plane $\text{span}\{e_2, e_3\}$ and ν can be thought as the angle between this projection and its projection onto e_2. Similarly we parametrize $\tilde{v}$ by $\|\tilde{v}\|(\sin(\tilde{\delta}), \cos(\tilde{\nu})\cos(\tilde{\delta}), \sin(\tilde{\nu})\cos(\tilde{\delta}))$, where the base used is the analogous base to the previous one, i.e. the vectors $\{\tilde{e}_1, \tilde{e}_2, \tilde{e}_3\}$. Taking into account that $e_2 \perp e_1$, $e_3 \perp e_1$, $\tilde{e}_2 \perp \tilde{e}_1$, $\tilde{e}_3 \perp \tilde{e}_1$, and the fact that e_1 is parallel to $-\text{Exp}_x(x_0)$, by the radial symmetry of the geodesic map we have that $\nu = \tilde{\nu}$. Also, by looking at the submanifold $\text{Exp}_x(\text{span}\{e_1, v\})$ and using Lemma 2.b we have

$$
\sin(\delta) = \sin(\tilde{\delta})\sqrt{\frac{1+K\tilde{\ell}^2}{1+K\tilde{\ell}^2\sin(\tilde{\delta})}}.
$$

If we want to compare $\langle \nabla F(x), v \rangle$ with $\langle \nabla f(\tilde{x}), \tilde{v} \rangle$ we should be able to just zero out the third components of v and $\tilde{v}$ and work in $n = 2$. But in order to completely obtain a reduction to the two-dimensional case we studied a few paragraphs above, we would need to prove that if we call

$w \stackrel{\text{def}}{=} (\sin(\delta), \cos(\nu)\cos(\delta), 0)$ the vector v with the third component made 0, then $\tilde{w}$ is in the same direction of the vector $\tilde{v}$, when the third component is made 0. The norm of these two vectors will not be the same, however. Let $\tilde{w}' = (\sin(\tilde{\delta}), \cos(\nu)\cos(\tilde{\delta}), 0)$ be the vector $\tilde{v}$ when the third component is made 0. Then

$$\|w\| = \|v\|\sqrt{\sin^2(\delta) + \cos^2(\delta)\cos^2(\nu)} \text{ and } \|\tilde{w}'\| = \|\tilde{v}\|\sqrt{\sin^2(\tilde{\delta}) + \cos^2(\tilde{\delta})\cos^2(\nu)}. \quad (41)$$

But indeed, we claim

$$\tilde{w} \text{ and } \tilde{w}' \text{ have the same direction.} \quad (42)$$

This is easy to see geometrically: since we are working with a geodesic map, the submanifolds $\text{Exp}_x(\text{span}\{v, e_3\})$ and $\text{Exp}_x(\text{span}\{e_1, e_2\})$ contain w. Similarly the submanifolds $x + \text{span}\{\tilde{v}, \tilde{e}_3\}$ and $x + \text{span}\{\tilde{e}_1, \tilde{e}_2\}$ contain $\tilde{w}'$. If the intersections of each of these pair of manifolds is a geodesic then the geodesic map must map one intersection to the other one, implying $\tilde{w}$ is proportional to $\tilde{w}'$. If the intersections are degenerate the case is trivial. Alternatively, one can prove this fact algebraically after some computations. It will be convenient for the rest of the proof so we will also include it here. If we call δ^* and $\tilde{\delta}'$ the angles formed by, respectively, the vectors e_2 and w, and the vectors $\tilde{e}_2$ and $\tilde{w}'$, then we have $\tilde{w}'$ is proportional to $\tilde{w}$ if $\tilde{\delta}' = \tilde{\delta}^*$, or equivalently $\delta' = \delta^*$. Using the definitions of w and $\tilde{w}'$ we have

$$\sin(\delta^*) = \sin\left(\arctan\left(\frac{\sin(\delta)}{\cos(\nu)\cos(\delta)}\right)\right) = \frac{\tan(\delta)/\cos(\nu)}{(\tan(\delta)/\cos(\nu))^2 + 1}$$
$$= \frac{\sin(\delta)}{\sqrt{\sin^2(\delta) + \cos^2(\nu)\cos^2(\delta)}},$$

and analogously

$$\sin(\tilde{\delta}') = \sin\left(\arctan\left(\frac{\sin(\tilde{\delta})}{\cos(\nu)\cos(\tilde{\delta})}\right)\right) = \frac{\tan(\tilde{\delta})/\cos(\nu)}{(\tan(\tilde{\delta})/\cos(\nu))^2 + 1}$$
$$= \frac{\sin(\tilde{\delta})}{\sqrt{\sin^2(\tilde{\delta}) + \cos^2(\nu)\cos^2(\tilde{\delta})}}. \quad (43)$$

Using Lemma 2.b for the pairs δ', $\tilde{\delta}'$ and δ^*, $\tilde{\delta}^*$, and the equations above we obtain

$$\sin(\delta^*) = \frac{\sin(\tilde{\delta})\sqrt{\frac{1+K\tilde{\ell}^2}{1+K\tilde{\ell}^2\sin^2(\tilde{\delta})}}}{\sqrt{\sin^2(\tilde{\delta})\frac{1+K\tilde{\ell}^2}{1+K\tilde{\ell}^2\sin^2(\tilde{\delta})} + \cos^2(\nu)\frac{\cos^2(\tilde{\delta})}{1+K\tilde{\ell}^2\sin^2(\tilde{\delta})}}} = \frac{\sin(\tilde{\delta})\sqrt{1+K\tilde{\ell}^2}}{\sqrt{\sin^2(\tilde{\delta})(1+K\tilde{\ell}^2) + \cos^2(\nu)\cos^2(\tilde{\delta})}},$$

and

$$\sin(\delta') = \frac{\sin(\tilde{\delta})}{\sqrt{\sin^2(\tilde{\delta}) + \cos^2(\nu)\cos^2(\tilde{\delta})}}\sqrt{\frac{1+K\tilde{\ell}^2}{1 + K\tilde{\ell}^2\left(\frac{\sin^2(\tilde{\delta})}{\sin^2(\tilde{\delta})+\cos^2(\nu)\cos^2(\tilde{\delta})}\right)}},$$

The two expressions on the right hand side are equal. This implies $\sin(\delta') = \sin(\delta^*)$. Since the angles were in the same quadrant we have $\delta' = \delta^*$.

We can now come back to the study of $\frac{\langle \nabla F(x), v \rangle}{\langle \nabla f(\tilde{x}), \tilde{v} \rangle}$. By (41) we have

$$\frac{\langle \nabla F(x), v \rangle}{\langle \nabla f(\tilde{x}), \tilde{v} \rangle} = \frac{\|\nabla F(x)\|}{\|\nabla f(\tilde{x})\|} \frac{\|v\|}{\|\tilde{v}\|} \frac{\langle \frac{\nabla F(x)}{\|\nabla F(x)\|}, \frac{w}{\|w\|} \rangle}{\langle \frac{\nabla f(\tilde{x})}{\|\nabla f(\tilde{x})\|}, \frac{\tilde{w}}{\|\tilde{w}\|} \rangle} \frac{\sqrt{\sin^2(\delta) + \cos^2(\delta)\cos^2(\nu)}}{\sqrt{\sin^2(\tilde{\delta}) + \cos^2(\tilde{\delta})\cos^2(\nu)}}$$

We now operate the last two fractions. Using (38) and (41) we get that the product of the last two fractions above is equal to

$$\sqrt{\frac{1 + K\tilde{\ell}^2}{(1 + K\tilde{\ell}^2 \sin^2(\tilde{\delta}^*))(1 + K\tilde{\ell}^2 \cos^2(\tilde{\beta}))}} \frac{\sqrt{\sin^2(\tilde{\delta})\frac{1+K\tilde{\ell}^2}{(1+K\tilde{\ell}^2 \sin^2(\tilde{\delta}))} + \cos^2(\nu)\frac{\cos^2(\tilde{\delta})}{1+K\tilde{\ell}^2 \sin(\tilde{\delta})}}}{\sin^2(\tilde{\delta}) + \cos^2(\tilde{\delta})\cos^2(\nu)}$$

which after using (43) (recall $\tilde{\delta}^* = \tilde{\delta}'$), and simplifying it yields

$$\sqrt{\frac{1 + K\tilde{\ell}^2}{(1 + K\tilde{\ell}^2 \sin^2(\tilde{\delta}))(1 + K\tilde{\ell}^2 \cos^2(\tilde{\beta}))}}.$$

So finally we have

$$\frac{\langle \nabla F(x), v \rangle}{\langle \nabla f(\tilde{x}), \tilde{v} \rangle} = \frac{\|\nabla F(x)\|}{\|\nabla f(\tilde{x})\|} \frac{\|v\|}{\|\tilde{v}\|} \sqrt{\frac{1 + K\tilde{\ell}^2}{(1 + K\tilde{\ell}^2 \sin^2(\tilde{\delta}))(1 + K\tilde{\ell}^2 \cos^2(\tilde{\beta}))}}.$$

In order to bound the previous expression, we now use Lemma 2.c and Lemma 2.a, and bound $\sin^2(\tilde{\delta})$ and $\cos^2(\tilde{\beta})$ by 0 or 1 depending on the inequality. Recall that, by (30) we have $1/\sqrt{1 + K\tilde{\ell}^2} = \mathrm{C}_K(\ell)$, for $\ell = d(x, x_0) \le R$. And $\tilde{\ell} = \|\tilde{x}\|$. Let's proceed. We obtain, for $K = -1$

$$\cosh^{-3}(R) \le \frac{1}{\cosh^2(\ell)} \cdot 1 \cdot \frac{1}{\cosh(\ell)} \le \frac{\langle \nabla F(x), v \rangle}{\langle \nabla f(\tilde{x}), \tilde{v} \rangle} \le \frac{1}{\cosh(\ell)} \cdot \cosh^2(\ell) \cdot \cosh(\ell) \le \cosh^2(R).$$

and for $K = 1$ it is

$$\cos^2(R) \le \frac{1}{\cos(\ell)} \cdot \cos^2(\ell) \cdot \cos(\ell) \le \frac{\langle \nabla F(x), v \rangle}{\langle \nabla f(\tilde{x}), \tilde{v} \rangle} \le \frac{1}{\cos^2(\ell)} \cdot 1 \cdot \frac{1}{\cos(\ell)} \le \cos^{-3}(R).$$

The first part of Lemma 3 follows, for $\gamma_{\mathrm{p}} = \cosh^{-3}(R)$ and $\gamma_{\mathrm{n}} = \cosh^{-2}(R)$ when $K = -1$, and $\gamma_{\mathrm{p}} = \cos^2(R)$ and $\gamma_{\mathrm{n}} = \cos^3(R)$ when $K = 1$.

The second part of Lemma 3 follows readily from the first one and g-convexity of F, as in the following. It holds

$$f(\tilde{x}) + \frac{1}{\gamma_{\mathrm{n}}} \langle \nabla f(\tilde{x}), \tilde{y} - \tilde{x} \rangle \overset{①}{\le} F(x) + \langle \nabla F(x), y - x \rangle \overset{②}{\le} F(y) = f(\tilde{y}),$$

and

$$f(\tilde{x}) + \gamma_{\mathrm{p}} \langle \nabla f(\tilde{x}), \tilde{y} - \tilde{x} \rangle \overset{③}{\le} F(x) + \langle \nabla F(x), y - x \rangle \overset{④}{\le} F(y) = f(\tilde{y}),$$

where ① and ③ hold if $\langle \nabla f(\tilde{x}), \tilde{y} - \tilde{x} \rangle \le 0$ and $\langle \nabla f(\tilde{x}), \tilde{y} - \tilde{x} \rangle \ge 0$, respectively, by the first part of this theorem. Inequalities ② and ④ hold by g-convexity of F. ∎

Appendix D. Constants depending on R and K, and comparisons

We discuss the value of the constants of our algorithms in Remark 28 and discuss recent hardness results in Remark 29. But we start by proving a relevant result that says that the condition number of an L-smooth and μ-strongly g-convex function $F : \mathcal{B}_R \to \mathbb{R}$ is lower bounded by a term depending on R and K, where the condition number is defined by L/μ. This is unlike in the Euclidean case, for which there are functions with condition number 1.

In particular, we show that the function $x \mapsto \frac{1}{2}d(x, x_0)^2$ has minimum condition number on $\mathcal{B}_R$, and is $(R\sqrt{|K|}\cot_K(R))^{-\operatorname{sign}(K)}$, where $\cot_K(R)$ is the special cotangent that is $\cot(\sqrt{|K|}R)$ if $K > 0$ and $\coth(\sqrt{|K|}R)$ if $K < 0$. And $\operatorname{sign}(K)$ is $K/|K|$ for $K \neq 0$. The fact about the condition number of $\frac{1}{2}d(x, x_0)^2$ can be obtained from the proof of Fact 20, and actually the fact per se as a comparison geometry theorem that uses that the inequality there is satisfied with equality in the constant curvature case. However, we recover the computation of this condition number while proving the proposition.

Proposition 27 *Let $F : \mathcal{M}_K \to \mathbb{R}$ be an L-smooth and μ-strongly convex function on $\mathcal{B}_R \subset \mathcal{M}_K$. Assume F is twice differentiable with continuous Hessian. Then, the condition number L/μ of F on $\mathcal{B}_R$ is at least the condition number of the function $\frac{1}{2}d(x, x_0)^2$ on $\mathcal{B}_R$.*

Proof As we have done before, we can assume $K \in \{1, -1\}$ because the other cases can be reduced to this one by a rescaling, cf. Remark 22. Recall that by definition of $\mathcal{M}_K$ and $\mathcal{B}_R$, for $K > 0$, we have that $R < \pi/2\sqrt{K}$.

We start by noting that given F, we can obtain another function G whose condition number is at most the one of F and such that it is symmetric with respect to every rotation whose axis goes through x_0. Formally, $G = G \circ \exp_{x_0} \circ \sigma \circ \exp_{x_0}^{-1}$ for a rotation $\sigma \in \mathrm{SO}(n)$. Equivalently, the function $G(x)$ depends on $\|\exp_{x_0}^{-1}(x)\|$ only. Indeed, an average of F and itself after performing an arbitrary rotation σ, that is $(F + F \circ \exp_{x_0} \circ \sigma \circ \exp_{x_0}^{-1})/2$, has a condition number that is at most the condition number of F. This is due to the Hessian being linear and its maximum and minimum eigenvalues over the domain determining the condition number. That is, the smoothness constant can only decrease or stay the same after performing the average. It would only be the same if, at some point, the Hessian matrices of each of the two added functions both have the same eigenvector with maximum eigenvalue and it equals the smoothness constant. The argument for the minimum is analogous. This argument extends to the case in which we integrate the function, pointwise, over $\mathrm{SO}(n)$ after applying a rotation. That is, defining $g(x) = \int_{SO(n)} F \circ \exp_{x_0} \circ \sigma \circ \exp_{x_0}^{-1}(x)d\sigma$ we obtain a symmetric function with condition number that is at most the condition number of F. So without loss of generality we can solely study symmetric functions G and in fact, due to the symmetries we do not lose generality if we work in dimension $n = 2$.

Denote $y_x = \|\exp_{x_0}^{-1}(x)\| \in \mathbb{R}$. We will express the condition number of G by using the function $g : \mathbb{R} \to \mathbb{R}$, defined as $g(y_x) = G(x)$ for any point $x \in \mathcal{M}_K$. Note the function is well defined by the symmetry property on G. A basis formed by (two) eigenvectors of $\nabla^2 G(x)$ can be chosen to have vectors in the direction of $\exp_x^{-1}(x_0)$ and its normal. Indeed, either every vector is an eigenvector associated to the same eigenvalue, which satisfies the above, or by the symmetry of $\nabla^2 G(x)$, there exists a base $\{v_1, v_2\}$ of orthonormal eigenvectors, associated with different eigenvalues $\lambda_1 > \lambda_2$. By the symmetry of G we have that $\lambda_1 = v_1^\top \nabla^2 G(x)v_1 = v_1'^\top \nabla^2 G(x)v_1'$, where v_1' is the symmetric vector to v_1 with respect to $\exp_x^{-1}(x_0)$. However, since $\lambda_1 \neq \lambda_2$ then the only unit vectors v that can satisfy $\lambda_1 = v^\top \nabla^2 G(x)v$ are $\pm v_1$, so $v_1 = v_1'$ and therefore v_1 and v_2

can be taken to be in the direction of $\exp_x^{-1}(x_0)$ and its normal. Consequently, one eigenvalue of $\nabla^2 G(x)$ is $g''(y_x)$. We can compute the other eigenvalue by using the non-Euclidean cosine theorem, cf. Fact 26. In order to do this, first note that $\nabla G(x)$ must be in the direction of $\exp_x^{-1}(x_0)$ by the symmetry of G and it must be $\|\nabla G(x)\| = g'(y_x)$. Now given $x \in \mathcal{M}$ and small enough $\eta \in \mathbb{R}$, we consider a right geodesic triangle with vertices x_0, x and z_η, where $z_\eta = \exp_x(\eta v_2)$ for v_2 defined above. Recall it is a unit vector that is normal to $\exp_x^{-1}(x_0)$ and it is an eigenvector of $\nabla^2 G(x)$. The definition of z_η implies that the angle between $\exp_x^{-1}(x_0)$ and $\exp_x^{-1}(x_2)$ is $\pi/2$ and $d(x, z_\eta) = \eta$. Let $\alpha(\eta)$ be the angle between $\exp_{z_\eta}^{-1}(x_0)$ and $\exp_{z_\eta}^{-1}(x)$. Since we are only interested about the eigenvalue of $\nabla^2 G(x)$ associated to the eigenvector v_2 we can project $\nabla G(z_\eta)$ onto $\exp_{z_\eta}^{-1}(x)$, which has norm $\|\nabla G(z_\eta)\cos(\alpha(\eta))\|$. We compute the eigenvalue as

$$
\begin{aligned}
\lim_{\eta \to 0} \frac{\|\nabla G(z_\eta)\|\cos(\alpha(\eta))}{\eta} &= \|\nabla G(x)\| \lim_{\eta \to 0} \frac{\cos(\alpha(\eta))}{\eta} \\
&\overset{\text{\textcircled{1}}}{=} g'(y_x) \lim_{\eta \to 0} \frac{\mathrm{C}_K(d(x_0,x)) - \mathrm{C}_K(\eta)\,\mathrm{C}_K(d(x_0,z_\eta))}{K\,\mathrm{S}_K(d(x_0,z_\eta))\eta\,\mathrm{S}_K(\eta)} \\
&\overset{\text{\textcircled{2}}}{=} g'(y_x) \lim_{\eta \to 0} \frac{\mathrm{C}_K(d(x_0,x))(1 - \mathrm{C}_K^2(\eta))/K}{\mathrm{S}_K(d(x_0,z_\eta))\eta\,\mathrm{S}_K(\eta)} \\
&\overset{\text{\textcircled{3}}}{=} g'(y_x)\cot_K(d(x_0,x)) = g'(y_x)\cot_K(y_x).
\end{aligned}
$$

Above, ① uses the cosine theorem, cf. Fact 26, applied as

$$
\mathrm{C}_K(d(x_0,x)) = \mathrm{C}_K(d(x,z_\eta))\,\mathrm{C}_K(d(x_0,z_\eta)) + K\cos(\alpha(\eta))\,\mathrm{S}_K(d(x_0,z_\eta))\,\mathrm{S}_K(d(x,z_\eta)).
$$

Recall that we have $\eta = d(x, z_\eta)$ by definition. Equality ② uses the cosine theorem again, with a different ordering of the sides so we obtain

$$
\mathrm{C}_K(d(x_0,z_\eta)) = \mathrm{C}_K(d(x_0,x))\,\mathrm{C}_K(d(x,z_\eta)),
$$

by using the right angle of the geodesic triangle. Finally ③ simplifies some terms, since $(1 - \mathrm{C}_K^2(\eta))/K = \mathrm{S}_K^2(\eta)$ and uses that $d(x_0, z_\eta)$ and $\mathrm{S}_K(\eta)/\eta$ tend to $d(x_0, x)$ and 1, respectively, when $\eta \to 0$. We conclude that the condition number of G is

$$
\kappa_G = \frac{\max_{y \in [0,R]}\{g''(y), g'(y)\cot_K(y)\}}{\min_{y \in [0,R]}\{g''(y), g'(y)\cot_K(y)\}}. \tag{44}
$$

We only need to prove that for any twice differentiable function $g : [0,R] \to \mathbb{R}$ with continuous second derivative, the quotient above is at least the value of the quotient that we obtain for $g(y) = y^2/2$, which is $(R\cot_K(R))^{-K}$. This is computed by noticing that, for that choice of g, we have that $g''(y) = 1$, that if $K = 1$ then $g'(y)\cot_K(y) \le 1$ and it reaches its minimum at $y = R$. If $K = -1$ then $g'(y)\cot_K(y) \ge 1$ and it is maximum at $y = R$. Note this implies that the condition number of $\frac{1}{2}d(x_0,x)^2$ on $\mathcal{B}_R$ is $(R\cot_K(R))^{-K}$, as it was advanced before.

Given g, let a, b be tight constants such that $g''(y) \in [a,b]$ for $y \in [0,R]$. Such constants must exist since g'' is a continuous function defined on a compact. We have $g'(y) \le by$, since by the symmetry and differentiability of G it must be $g'(0) = 0$. We obtain a lower bound on κ_G

if we lower bound the numerator of (44) by $\max_{y \in [0,R]}\{g''(y)\} = b$ and if we upper bound the denominator by $g'(R) \cot_K(R)$. We obtain

$$\kappa_G \geq \frac{b}{g'(R) \cot_K(R)} \geq \frac{b}{Rb \cot_K(R)} = \frac{1}{R \cot_K(R)}.$$

Similarly, if we lower bound the denominator of (44) by $aR \cot_K(R) \leq \max_{y \in [0,R]}\{g'(y) \cot_K(y)\}$ and upper bound the denominator by $a = \min_{y \in [0,R]}\{g''(y)\}$ we obtain

$$\kappa_G \geq \frac{aR \cot_K(R)}{a} = R \cot_K(R).$$

For each case $K \in \{1, -1\}$, there is only one of the lower bounds above such that its right hand side is greater than 1 and it precisely matches the value of the condition number of $\frac{1}{2}d(x_0, x)^2$ we computed above. ∎

Remark 28 *The previous proposition intuitively suggests that it could be unavoidable to have some particular constants depending on R in the rates of any optimization algorithm. For starters, optimizing a g-convex function by adding a strongly g-convex regularizer and optimizing the resulting strongly g-convex problem would entail rates containing a factor depending on the condition number of the regularizer, which the proposition proves it is at least the value $(R\sqrt{|K|} \cot_K(R))^{-\operatorname{sign}(K)}$. This implies that in the case of positive curvature $K = 1$, a μ-strongly g-convex and L-smooth function defined on the ball $\mathcal{B}_R$ must have condition number that is at least $\tan(R)/R \in [\frac{2}{\pi \cos(R)}, \frac{1}{\cos(R)}]$. This grows fast with R, but it is only natural if one takes into account that no strongly g-convex function exists if $R \geq \frac{\pi}{2}$, due to the space containing a full geodesic circle (so the constant function is the only g-convex function in this domain). Optimization in manifolds of positive curvature only makes sense in spaces of low diameter.*

The classical domain of application of accelerated methods for strongly convex functions consists of functions with large condition number κ, due to the $\sqrt{\kappa}$-dependence of the rates. For $K = 1$, the constants of our algorithm $1/\gamma_{\mathrm{p}} = \cos^{-2}(R)$ and $1/\gamma_{\mathrm{n}} = \cos^{-3}(R)$ (we also have the constant $\sqrt{44 \max\{1, R^2\}}$ coming from $\tilde{L}$) might seem large but they are a small polynomial of the minimum attainable condition number. If the condition number is large or, in its limit to infinity, whenever the function is g-convex, then acceleration is beneficial. For the case $K = -1$ the previous proposition shows that the minimum condition number is $R \tanh(R)$. In this case, our constants are $1/\gamma_{\mathrm{p}} = \cosh^3(R)$ and $1/\gamma_{\mathrm{n}} = \cosh^2(R)$, and a constant of a similar nature coming from $\tilde{L}$ (cf, [proof of Lemma 4](#)), which do not present an analogous dependency with respect to the minimum attainable condition number as in the previous case. This exponential dependence could be due to the exponential volume that a ball contains in the hyperbolic space. Studying if these constants are necessary for a global full accelerated method is interesting open problem and future direction of research. Regardless, the essence of our results for μ-strongly g-convex functions is that we can optimize at a full accelerated rate globally as opposed to essentially fully accelerating in a small neighborhood of radius $O((\mu/L)^{3/4})$ around the minimizer (this is explicit in ([Zhang and Sra, 2018](#)) and implicit in ([Ahn and Sra, 2020](#)) since the rates of AGD are nearly achieved only after a number of steps that is what RGD needs to reach the neighborhood). Note that, additionally, we can achieve acceleration in the g-convex case, which was not possible before. In any case, we note that in machine learning applications, it has been observed that the iterates do not get far

from initialization (Nagarajan and Kolter, 2019), especially in overparametrized models. Conse-quently, in such regime, R being a small constant is not a strong assumption and the constants of our algorithms do not become significant.

In the sequel, we comment on the work of Hamilton and Moitra (2021); Criscitiello and Boumal (2021), that show a hardness result in this direction. Our intuition is that, due to the geometry, it is necessary to have an additive and/or multiplicative constant depending on R on the optimal rates of convergence, similarly to the multiplicative constant R that one has in the lower bound for the class of smooth and convex functions in the Euclidean space. And for *easy* strongly g-convex functions (low condition number), this hardness could dominate convergence. However, when the condition number is large, which is the traditional regime of application of accelerated methods, or in its limit to infinity, that is in the case of g-convexity, acceleration becomes again a very useful tool.

D.1. Comment on hardness results

Remark 29 *After this work was publicly available, two lower bounds have been constructed, (Hamilton and Moitra, 2021; Criscitiello and Boumal, 2021). For the first one (Hamilton and Moitra, 2021), in the noisy setting, in the hyperbolic plane, the authors claim "[to have] dashed these hopes [of having Nesterov-like accelerated algorithms] by showing that acceleration is impos-sible even in the simplest of settings where we want to minimize a smooth and strongly geodesically convex function over the hyperbolic plane". We argue here that this is not the case.*

Hamilton and Moitra (2021) essentially argue that, in their setting with a noisy oracle in the hyperbolic plane, one needs $\gtrsim R/\log(R)$ noisy queries to the gradient or function value for opti-mizing functions of the form $d(x, x^*)^2$, while their condition numbers are $L/\mu \approx R$ so obtaining rates $\lesssim \sqrt{L/\mu}$ is impossible in general. But it does not preclude to have an algorithm with rates that are, for instance, $\lesssim R + \sqrt{L/\mu}\log(1/\varepsilon)$. Or a similar expression that involves some other additive or multiplicative constants depending on R. In fact, they are able to show that "acceler-ation is impossible even in the simplest of settings" precisely because they study the simplest of settings! That is, they show there is some hardness depending on the geometry. In particular, when the condition number is low this hardness can dominate the convergence. For instance, for rates $R + \sqrt{L/\mu}\log(1/\varepsilon)$ the R can dominate convergence unless $L/\mu \geq R^2$ or ε is small enough. The lower bound does not mean that acceleration is doomed to fail. In fact, the problems for which ac-celeration gets the most improvement are ill-conditioned problems and for those one would expect to still have acceleration in their noisy setting. In particular, acceleration is of importance when L/μ is large or in the limit to infinity, that is, when the function is g-convex.

We note that Hamilton and Moitra (2021) independently proved a similar result as our Propo-sition 27, limited to the hyperbolic plane. In particular they show that the condition number for an L-smooth μ-strongly g-convex function F defined on the hyperbolic disk of curvature $K = -1$ must be $\kappa_F = L/\mu \geq \Omega(R)$, which is similar to the precise result that we found that had optimal constant $R\cot(R)$.

The lower bound of Criscitiello and Boumal (2021) is a generalization of the previous one, in which they show that the $\widetilde{\Omega}(R)$ lower bound still holds for optimization with a deterministic first-order oracle, in hyperbolic spaces and some other more general negatively curved manifolds. Similarly, this lower bound applies when the condition number is small enough ($\approx R$), close to the lower bound on the condition number that we proved in Proposition 27. They also provide a lower bound in the case of smooth and only g-convex functions. Their lower bound requires to have R

growing as $\widetilde{\Theta}(\frac{1}{\varepsilon})$ and in that case they show that RGD is optimal. This result is in the same spirit as the previous one: The geometry causes some hardness, and if we allow to grow the feasible space enough, this hardness can dominate convergence. On the other hand, this does not mean that one cannot accelerate when, for instance, R is fixed and ε is small enough.

D.2. Comment on the rates of eventually accelerated algorithms (Ahn and Sra, 2020)

Remark 30 *The local algorithm in (Zhang and Sra, 2018) requires starting $O((L/\mu)^{-3/4})$ close to the optimum and it finds an ε-minimizer in $O(\sqrt{L/\mu}\log(1/\varepsilon))$. On the other hand RGD has a convergence rate of $O(L/\mu\log(\mu/\varepsilon))$. Hence, we could run both algorithms in parallel and restart them every few iterations from the best of the two points that both algorithms yielded. In that case we would obtain the convergence rate $O^*(L/\mu + \sqrt{L/\mu}\log(\mu/\varepsilon))$. Indeed, note that we would just compute twice as many gradients as if we run RGD but we perform as well as if we first run RGD until it gets into the desired neighborhood and then we run the local accelerated algorithm. And by μ-strong g-convexity we can guarantee we are $\mu\bar{\varepsilon}^2/2$-close to a minimizer if we are at an $\bar{\varepsilon}$-minimizer so if we set $\bar{\varepsilon}$ so that $\mu\bar{\varepsilon}^2/2 = O((L/\mu)^{-3/4})$ and run RGD we reach the neighborhood after $O((L/\mu)\log(\mu(L/\mu)^{3/4}))$ iterations.*

We note that this mix of RGD and the local algorithm in (Zhang and Sra, 2018) enjoys the same worse case guarantee of (Ahn and Sra, 2020). This latter work is a generalization of (Zhang and Sra, 2018) that eventually accelerates. The proofs of this paper reveal that in order for their bound to reach accelerated rates the algorithm needs as much time as RGD takes to reach the accelerating neighborhood of (Zhang and Sra, 2018). Indeed, they can guarantee that for their iterates y_t, their algorithm converges at an accelerated rate $f(y_t) - f(x^*) \leq O(f(y_{t-1}) - f(x^*))(1 - \sqrt{\mu/L})$ when $t = \Omega^*(\frac{1}{\log(1/\lambda)}) = \Omega^*(L/\mu)$, where $\lambda = \Omega(1 - \mu/L)$. A summary of rates is presented in Table 1, including this fact.