Curvature and complexity:
Better lower bounds for geodesically convex optimization

Christopher Criscitiello   CHRISTOPHER.CRISCITIELLO@EPFL.CH

Nicolas Boumal   NICOLAS.BOUMAL@EPFL.CH
Ecole Polytechnique Fédérale de Lausanne (EPFL), Institute of Mathematics

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

We study the query complexity of geodesically convex (g-convex) optimization on a manifold. To isolate the effect of that manifold's curvature, we primarily focus on hyperbolic spaces. In a variety of settings (smooth or not; strongly g-convex or not; high- or low-dimensional), known upper bounds worsen with curvature. It is natural to ask whether this is warranted, or an artifact.

For many such settings, we propose a first set of lower bounds which indeed confirm that (negative) curvature is detrimental to complexity. To do so, we build on recent lower bounds (Hamilton and Moitra, 2021; Criscitiello and Boumal, 2022a) for the particular case of smooth, strongly g-convex optimization. Using a number of techniques, we also secure lower bounds which capture dependence on condition number and optimality gap, which was not previously the case.

We suspect these bounds are not optimal. We conjecture optimal ones, and support them with a matching lower bound for a class of algorithms which includes subgradient descent, and a lower bound for a related game. Lastly, to pinpoint the difficulty of proving lower bounds, we study how negative curvature influences (and sometimes obstructs) interpolation with g-convex functions.

Keywords: geodesic convexity; Riemannian optimization; curvature; lower bounds; hyperbolic

1 Introduction and contributions

Let $\mathcal{M}$ be a d-dimensional Riemannian manifold. We consider optimization problems of the form

$$\min_{x \in \mathcal{M}} f(x), \quad \text{knowing that } f \text{ attains its minimum in } B(x_{\text{ref}}, r) = \{x : \text{dist}(x, x_{\text{ref}}) \le r\}. \quad \text{(P)}$$

Here, $f \colon \mathcal{M} \to \mathbb{R}$ is geodesically convex (g-convex) and dist denotes Riemannian distance. When $\mathcal{M}$ is a Euclidean space, problem (P) amounts to convex optimization. Motivated by applications in statistics, machine learning and computer science (see below), it is natural to study algorithms for (P) (upper bounds), and to ask whether they are optimal (lower bounds).

Both in smooth and nonsmooth g-convex optimization, known upper bounds *worsen* with the curvature of $\mathcal{M}$ (Zhang and Sra, 2016; Bento et al., 2017). *Is this effect of curvature warranted?*

To address this question, we focus on Hadamard manifolds, which have nonpositive curvature, because they are the most natural setting for geodesic convexity. Non-Hadamard manifolds, even simple ones, often do not globally carry interesting g-convex functions.[1] Accordingly, most applications of geodesic convexity are in Hadamard manifolds. To isolate the effect of curvature, we further focus on the case where $\mathcal{M}$ has constant negative curvature, i.e., $\mathcal{M}$ is a *hyperbolic*

---

1. For example, on compact manifolds, global g-convex functions are constant. Moreover, if a complete Riemannian manifold admits a smooth strongly g-convex function, then it must be diffeomorphic to $\mathbb{R}^d$ (Sakai, 1996, Prop. 5.10).

space of curvature $K < 0$.[2] We prove lower bounds for algorithms with access to first-order information (function values and subgradients). Dependence on curvature is captured by the quantity $\zeta = \zeta_{r\sqrt{-K}} := \frac{r\sqrt{-K}}{\tanh(r\sqrt{-K})}$, as it appears in known upper bounds.[3]

Lower bounds in the Euclidean case are well understood. For curved spaces, the only known result is due to Hamilton and Moitra (2021) and Criscitiello and Boumal (2022a), who prove the lower bound[4] $\tilde{\Omega}(\zeta)$ for smooth strongly g-convex optimization, showing that in this setting the effect of curvature is unavoidable. However, *this known lower bound has two major limitations*:

(a) It only holds for smooth, strongly g-convex optimization (as opposed to other settings);

(b) It does not depend on the condition number κ of the strongly g-convex function.

These are real limitations. Many g-convex problems in applications fail to be smooth or strongly g-convex, e.g., Fréchet medians (Fletcher et al., 2009) or operator scaling (Allen-Zhu et al., 2018). Moreover, most problems that are smooth and strongly g-convex have condition number κ which is much greater than ζ (e.g., robust covariance estimation (Franks and Moitra, 2020)), so the complexity of the problem is determined by κ rather than ζ. (We note however that the problem of Fréchet means (Karcher, 1977) is strongly g-convex with condition number that naturally scales as $\Theta(\zeta)$.)

In this paper, we address these two limitations. We provide *four main contributions*. First, to address limitation (a), in Section 3 we extend the $\tilde{\Omega}(\zeta)$ lower bound to the Lipschitz and the smooth (nonstrongly) g-convex settings. This shows that "full" acceleration $O(\frac{1}{\sqrt{\epsilon}})$ (independent of curvature) for smooth g-convex optimization is impossible, and also justifies the presence of curvature terms in the upper bounds for subgradient descent proven by Zhang and Sra (2016).

Second, to address limitation (b), in Sections 4 and 5 we provide new techniques for proving lower bounds which depend on the problem class parameters (e.g., κ or ϵ). In addition, in Appendix D we show how upper and lower bounds from Euclidean space carry over to the Riemannian setting when r is very small. In the first four rows of Table 1, we present these lower bounds, along with the best known upper bounds. See Section 1.1 for the precise function classes and complexity measures. For any fixed ζ, the scalings of these lower bounds match those found in Euclidean space.

Our lower bounds do not match the best known upper bounds in Table 1, and we strongly suspect the lower bounds can be improved. Our third contribution is to provide a roadmap for improving on our lower bounds. We *conjecture* that the "multiplicative" upper bounds in the second, third and fourth rows of the table are optimal, and that the optimal upper bound for the first row is $\Theta(\zeta d)$.[5] We do not know if those are the correct bounds, but we believe they are reasonable targets.

We provide two pieces of evidence for this conjecture. First, in Section 6 *we prove the multiplicative lower bound* $\Omega(\frac{\zeta}{\epsilon^2})$ *for a class of (intuitively reasonable) algorithms which includes subgradient descent* (with Polyak step size). In particular, this shows that the analysis of subgradient descent cannot be improved. Second, in Section 7 we prove the $\tilde{\Omega}(\zeta d)$ lower bound for a "cutting-planes game" which serves as a proxy for low-dimensional Lipschitz g-convex optimization, and also provides a lower bound for g-convex "cutting-planes schemes" (Nesterov, 2004, Sec. 3.2.6).[6]

---

2. We expect our results extend to the manifold of positive definite matrices with affine-invariant metric, because it contains a large totally geodesic submanifold isometric to hyperbolic space (Criscitiello and Boumal, 2022a, App. J).

3. If $r\sqrt{-K} \ll 1$, then $\zeta \approx 1$ and bounds match those from Euclidean space. If $r\sqrt{-K} \gg 1$, then $\zeta \sim r\sqrt{-K}$ is large.

4. We write $\tilde{\Omega}(\cdot)$, $\tilde{O}(\cdot)$ or $\tilde{\Theta}(\cdot)$ when we omit logarithmic terms.

5. The lower bounds we prove depend *additively* in terms of ζ and the other parameters ($\tilde{\Omega}(\zeta + \sqrt{\kappa})$ for the strongly g-convex case). On the other hand, the best known upper bounds depend *multiplicatively* on the parameters ($O(\sqrt{\zeta\kappa})$).

6. Nesterov (2004, Sec. 3.2.5) considers the analogous game as a proxy for Lipschitz convex optimization.

To build lower bounds for a function class (e.g., g-convex functions), one (implicitly or explicitly) *interpolates* a collection of function values and gradients with a function from that class. The recent focus on *necessary and sufficient interpolation conditions* for convex functions (Taylor et al., 2016), has led to a number of insights into the complexity of convex optimization (Taylor et al., 2017; Dragomir et al., 2021; Taylor and Drori, 2022). Our fourth contribution is a study of interpolation by g-convex functions in Section 8. We show that (unlike in the convex case), the naive necessary conditions for interpolation by g-convex functions are *not* sufficient. This geometric obstruction partly explains the difficulties of showing lower bounds for g-convex optimization.

Complexity of g-convex optimization					
g-convex setting	Function class	Lower bound	Upper bound	Algorithm	Result
Lipschitz, low-dimensional, **P1**	$\mathcal{F}_{r,M,0,\infty}^{d}$	$\tilde{\Omega}(\zeta + d)$	$O(\zeta d^2)$	center of gravity (Rusciano, 2019)	Thm. 7, Rmk. 9
Lipschitz, **P2**	$\mathcal{F}_{r,M,0,\infty}$	$\tilde{\Omega}(\zeta + \frac{1}{\zeta^2\epsilon^2})$	$O(\frac{\zeta}{\epsilon^2})$	subgradient descent (Zhang and Sra, 2016)	Thm. 7, 8
smooth, **P3**	$\mathcal{F}_{r,\infty,0,L}$	$\tilde{\Omega}(\zeta + \frac{1}{\zeta\sqrt{\epsilon}})$	$\tilde{O}\left(\sqrt{\frac{\zeta}{\epsilon}}\right)$ [a]	RNAG-C (Kim and Yang, 2022)	Thm. 7, 11
smooth, strongly g-convex, **P4**	$\mathcal{F}_{r,\infty,\mu,L}$	$\tilde{\Omega}(\zeta + \sqrt{\kappa})$	$O(\sqrt{\zeta\kappa})$ [b]	RNAG-SC (Kim and Yang, 2022)	Cor. 12
cutting-planes game, Sec. 7	N/A	$\tilde{\Omega}(\zeta d)$	$O(\zeta d^2)$	center of gravity (Rusciano, 2019)	Thm. 15

Table 1: $\mathcal{M}$ is a hyperbolic space of curvature $K < 0$, and $\zeta = \zeta_{r\sqrt{-K}} = \frac{r\sqrt{-K}}{\tanh(r\sqrt{-K})}$. All entries in the column "Lower bound" are novel, except for the term ζ in the fourth row.

---

a. The upper bounds for smooth g-convex and smooth strongly g-convex minimization have an important caveat. Those upper bounds come from Kim and Yang (2022) who *assume* the iterates produced by their algorithm stay in the optimization domain $B(x_{\mathrm{ref}}, r)$. Recent work by Martínez-Rubio and Pokutta (2022) shows how to remove this assumption at the expense of somewhat worse complexity guarantees: $\tilde{O}(\zeta\sqrt{\frac{1}{\epsilon}})$ and $\tilde{O}(\zeta\sqrt{\kappa})$.

b. Kim and Yang (2022) report the bounds $O(\zeta\sqrt{\frac{1}{\epsilon}})$ and $O(\zeta\sqrt{\kappa})$. We observe that the rate $O(\zeta\sqrt{\kappa})$ can be improved to $O(\sqrt{\zeta\kappa})$ by choosing the step size in their algorithm as $\Theta(\frac{1}{L})$ instead of $\Theta(\frac{1}{\zeta L})$. More precisely, using that $\kappa \geq \zeta$ and the step size $\Theta(\frac{1}{L})$, one can check that their analysis (see Corollary F.1 in their paper) follows through unchanged. Moreover, the rate $\tilde{O}(\sqrt{\frac{\zeta}{\epsilon}})$ can be achieved by a reduction from the $O(\sqrt{\zeta\kappa})$ strongly g-convex algorithm (see Proposition 57), using that $\epsilon \leq \frac{8}{\zeta}$ (see Proposition 13).

1.1 Problem classes and algorithms

Let $\mathbb{H}^d$ denote a d-dimensional hyperbolic space. Without loss of generality, we can assume $\mathbb{H}^d$ has curvature $K = -1$, and so we do this from now on.[7] We usually denote the underlying manifold

---

7. Let $\mathcal{M}_1 = (\mathcal{M}, g)$ be a hyperbolic space of curvature $K_1 < 0$, and scale the metric g on $\mathcal{M}$ to get a hyperbolic space $\mathcal{M}_2 = (\mathcal{M}, \frac{K_1}{K_2}g)$ of curvature K_2. Then it is easy to see that the function classes $\mathcal{F}_{r_1,M,\mu,L}^{x_{\mathrm{ref}}}(\mathcal{M}_1)$ and $\mathcal{F}_{r_2,M,\mu,L}^{x_{\mathrm{ref}}}(\mathcal{M}_2)$ are identical provided $r_1\sqrt{-K_1} = r_2\sqrt{-K_2}$. Stated differently: what matters is $r\sqrt{-K}$, not r and K separately. We thus fix $K = -1$.

by $\mathcal{M}$ (i.e., $\mathcal{M} = \mathbb{H}^d$), and its tangent bundle by $T\mathcal{M}$ (see Section 2). For each $d \geq 2$, fix a point $x_{\text{ref}}^d \in \mathbb{H}^d$. Throughout we usually drop the superscript on x_{ref}^d and write x_{ref}. For a function $f \colon \mathcal{M} \to \mathbb{R}$, we always denote $f^* = \min_{x \in \mathcal{M}} f(x)$.

For $r > 0$, $M \geq 0$, and $L \geq \mu \geq 0$, let $\mathcal{F}_{r,M,\mu,L}^d$ be the class of functions $f \colon \mathbb{H}^d \to \mathbb{R}$ which (a) are globally M-Lipschitz, (b) are globally μ-strongly g-convex, (c) are L-smooth (globally if $\mu = 0$, or in the ball $B(x_{\text{ref}}, r)$ if $\mu > 0$),[8] and (d) have a global minimizer x^* which is in $B(x_{\text{ref}}, r)$. Lastly, define $\mathcal{F}_{r,M,\mu,L} = \bigcup_{d=2}^{\infty} \mathcal{F}_{r,M,\mu,L}^d$.

A deterministic first-order algorithm $\mathcal{A}$ on $\mathcal{M}$ is a sequence of functions $(\mathcal{A}_k \colon (\mathbb{R} \times T\mathcal{M})^k \to \mathcal{M})_{k \geq 0}$; in particular, $\mathcal{A}_0$ returns an initial point x_0. Such an algorithm has access to an oracle $\mathcal{O}_f \colon \mathcal{M} \to \mathbb{R} \times T\mathcal{M}$ which for each query $x \in \mathcal{M}$ returns the function value $f(x)$ and a subgradient $g \in \partial f(x)$. Running $\mathcal{A}$ with an oracle $\mathcal{O}_f$ produces iterates $x_0, x_1, \ldots$ as follows.[9] Let $\mathcal{H}_0 = \emptyset$. After already making $k \geq 0$ queries $x_0, \ldots, x_{k-1}$, the algorithm uses the known information $\mathcal{H}_k$ to compute the next query $x_k = \mathcal{A}_k(\mathcal{H}_k)$. The oracle $\mathcal{O}_f$ gives the algorithm $F_k = f(x_k)$ and $g_k \in \partial f(x_k)$, and we update the known information $\mathcal{H}_{k+1} = (F_\ell, x_\ell, g_\ell)_{\ell=0}^k$.

Given a tolerance $\delta > 0$, we define $T_\delta(\mathcal{A}, f)$ to be the first k for which $f(x_k) - f^* \leq \delta$, where $(x_k)_{k \geq 0}$ is the sequence produced by running $\mathcal{A}$ on f. Given a function class $\mathcal{F}$, the complexity of optimization on that function class is defined by infimizing over deterministic first-order algorithms:

$$T_\delta(\mathcal{F}) = \inf_{\mathcal{A}} \sup_{f \in \mathcal{F}} T_\delta(\mathcal{A}, f).$$

We focus on four function classes. In the following, $M > 0, L > 0, d \geq 2$.

P1 (*low-dimensional Lipschitz*) For $0 < \epsilon < 1$, define $\delta = \epsilon \cdot Mr$, and $T_{\epsilon,r,d} = T_\delta(\mathcal{F}_{r,M,0,\infty}^d)$.[10]

P2 (*high-dimensional Lipschitz*) For $0 < \epsilon < 1$, define $\delta = \epsilon \cdot Mr$, and $T_{\epsilon,r} = T_\delta(\mathcal{F}_{r,M,0,\infty})$.

P3 (*high-dimensional smooth*) For $0 < \epsilon < \min\{1, \frac{8}{\zeta_r}\}$, define $\delta = \epsilon \cdot \frac{1}{2}Lr^2$, $T_{\epsilon,r} = T_\delta(\mathcal{F}_{r,\infty,0,L})$.[11]

P4 (*high-dimensional smooth strongly g-convex*) For $0 < \epsilon < \min\{1, \frac{8}{\zeta_r}\}, \mu > 0, L \geq \mu\zeta_r$, define $\delta = \epsilon \cdot \frac{1}{2}Lr^2$, $\kappa = \frac{L}{\mu}$, and $T_{\epsilon,r,\kappa} = T_\delta(\mathcal{F}_{r,\infty,\mu,L})$.

We expect complexity of problems **P2** and **P3** (both denoted by $T_{\epsilon,r}$) to scale *polynomially* in ϵ^{-1}. For those, Table 1 reports bounds on $T_{\epsilon,r}$. On the other hand, due to known upper bounds, we can solve problems **P1** and **P4** at least at a linear rate, that is, scaling *logarithmically* in ϵ as $\log(\epsilon^{-1})$. Therefore, it is reasonable to at least initially focus on the factors in front of the $\log(\epsilon^{-1})$: this is the approach we take. Said differently, for problem **P1** we expect there exists $q > 0$ so that $f(x_k) - f^* \leq Mr \cdot e^{-k/q}$ after k queries. *The quantity q measures the number of queries needed to reduce the optimality gap by a constant factor.* To suppress the dependence on ϵ for **P1** and **P4**, we define $q_{d,r} = \sup_{\epsilon \in (0,1)} \{ \frac{T_{\epsilon,r,d}}{\log(\epsilon^{-1})} \}$ and define $q_{\kappa,r}$ similarly. Table 1 reports bounds on $q_{d,r}$ and $q_{\kappa,r}$.

---

8. For (c): if $\mu > 0$ and we require f to be L-smooth globally, then the function class is empty, see Section 5.1.

9. We often implicitly assume there is an oracle associated to f, and simply say "running $\mathcal{A}$ on f produces iterates . . . ."

10. With this choice of δ, by scaling the functions in the class, one can check that $T_\delta(\mathcal{F}_{r,M,0,\infty}^d)$ is independent of M, and only depends on ϵ, r, d. Similar considerations hold for the other function classes.

11. This is the right scaling for ϵ – see Section 5.1. Indeed, in Proposition 13 we prove if $f \in \mathcal{F}_{r,\infty,0,L}$ then $f(x_{\text{ref}}) - f^* \leq \frac{1}{2}Lr^2 \cdot \frac{8}{\zeta_r}$. This is analogous to the observation that $\kappa \geq \zeta_r$, although a different proof is required to show this.

2 Preliminaries: Hadamard manifolds, hyperbolic spaces and geodesic convexity

Throughout, $\mathcal{M}$ denotes a Hadamard manifold which has tangent bundle $T\mathcal{M}$ and tangent spaces $T_x\mathcal{M}$. A Hadamard manifold is a complete, simply connected Riemannian manifold with nonpositive curvature, see (Bridson and Haefliger, 1999) and (Lee, 2018, Ch. 12). The Riemannian metric on $\mathcal{M}$ is denoted $\langle \cdot, \cdot \rangle$ and $\|v\| = \sqrt{\langle v, v \rangle}$ for $v \in T_x\mathcal{M}$. The Riemannian metric gives $\mathcal{M}$ a notion of distance dist, volume Vol, geodesics, and intrinsic curvature. The exponential map on $\mathcal{M}$ at $x \in \mathcal{M}$ is denoted by $\exp_x : T_x\mathcal{M} \to \mathcal{M}$, and its inverse map by $\log_x : \mathcal{M} \to T_x\mathcal{M}$. As $\mathcal{M}$ is Hadamard, the exponential map and its inverse are global diffeomorphisms. For $x, y \in \mathcal{M}$, $P_{x \to y} : T_x\mathcal{M} \to T_y\mathcal{M}$ denotes parallel transport along the geodesic connecting x and y. The boundary of a set $D \subseteq \mathcal{M}$ is denoted ∂D, and its interior is $\text{int}D = D \setminus \partial D$.

We usually take $\mathcal{M}$ to be a d-dimensional hyperbolic space of curvature -1, denoted by $\mathbb{H}^d$. Certain submanifolds of $\mathbb{H}^d$ feature prominently in our lower bound constructions. A connected and complete Riemannian submanifold $S \subseteq \mathcal{M}$ is called *totally geodesic* if a geodesic in S is also a geodesic in $\mathcal{M}$ (see Appendix A.1). Totally geodesic submanifolds are abundant in $\mathbb{H}^d$ (but are rare in general Hadamard manifolds (Chen, 2000, Sec. 11.1)).

A set $D \subseteq \mathcal{M}$ is *g-convex* if for all $x, y \in D$ the geodesic segment connecting x, y is contained in D. Balls $B(x, r)$, $r \geq 0$, and totally geodesic submanifolds are g-convex sets. The following lemma can be proven using the hyperboloid or Beltrami-Klein models of $\mathbb{H}^d$ (Appendix A.1).

Lemma 1 *Let $z \in \mathcal{M} = \mathbb{H}^d$, $g \in T_z\mathcal{M}$, $L = \exp_z(\{v \in T_z\mathcal{M} : \langle g, v \rangle \geq 0\})$, and $S = \partial L = \exp_z(\{v \in T_z\mathcal{M} : \langle g, v \rangle = 0\})$. The half-space L is g-convex. Its boundary $S = \partial L$ is a $(d-1)$-dimensional totally geodesic submanifold of $\mathbb{H}^d$.*

References on g-convex optimization include (Udriște, 1994), (Rapcsák, 1997), (Bacák, 2014), and (Boumal, 2023, Ch. 11). For $f : \mathcal{M} \to \mathbb{R}$, we denote its Riemannian gradient and Hessian by ∇f and $\nabla^2 f$, respectively (see, for example, Chapters 3 and 5 of (Boumal, 2023)).

Definition 2 *Let $\mathcal{M}$ be a Hadamard manifold, and let $D \subseteq \mathcal{M}$ be g-convex. A function $f : D \to \mathbb{R}$ is μ-strongly g-convex if $f(\gamma(t)) \leq (1-t)f(x) + tf(y) - \frac{\mu}{2}t(1-t)\text{dist}(x,y)^2$ for all $x, y \in D$, where $\gamma : [0,1] \to \mathbb{R}$ is the geodesic with $\gamma(0) = x, \gamma(1) = y$. If $\mu = 0$, we say that f is g-convex or "nonstrongly" g-convex if we wish to emphasize $\mu = 0$. If $\mu > 0$, we say f is strongly g-convex.*

If f is differentiable, f is μ-strongly g-convex in $\mathcal{M}$ if and only if

$$f(y) \geq f(x) + \langle \nabla f(x), \log_x(y) \rangle + \frac{\mu}{2}\text{dist}(x,y)^2 \qquad \forall x, y \in \mathcal{M}.$$

If f is twice differentiable, f is μ-strongly g-convex in $\mathcal{M}$ if and only if $\nabla^2 f(x) \succeq \mu I \ \forall x \in \mathcal{M}$.

Definition 3 *A vector $g \in T_x\mathcal{M}$ is a subgradient of $f : \mathcal{M} \to \mathbb{R}$ at x if $f(y) \geq f(x) + \langle g, \log_x(y) \rangle$ for all $y \in \mathcal{M}$. The subdifferential $\partial f(x)$ of f at x is the set of all subgradients of f at x.*

For a g-convex function, subdifferentials are never empty. A g-convex function is differentiable if and only if all its subdifferentials contain exactly one vector (the gradient) (Udriște, 1994, Sec. 3.4).

Definition 4 *Let $D \subseteq \mathcal{M}$ be connected. A function $f : \mathcal{M} \to \mathbb{R}$ is M-Lipschitz in D if $|f(x) - f(y)| \leq M\text{dist}(x,y)$ for all $x, y \in D$. A differentiable function $f : \mathcal{M} \to \mathbb{R}$ is L-smooth in D if $\|\nabla f(x) - P_{y \to x}\nabla f(y)\| \leq L\text{dist}(x,y)$ for all $x, y \in D$.*

If f is C^1, then f is M-Lipschitz in $\mathcal{M}$ if and only if $\|\nabla f(x)\| \leq M$ for all $x \in \mathcal{M}$. Let D be g-convex. If f is C^1 and L-smooth in D then $|f(y) - f(x) - \langle \nabla f(x), \log_x(y)\rangle| \leq \frac{L}{2}\text{dist}(x,y)^2$ for all $x, y \in D$. If f is C^2, f is L-smooth in D if $\|\nabla^2 f(x)\| \leq L$ for all $x \in D$ (operator norm).

The supremum of g-convex functions is always g-convex (Udrişte, 1994, Cor. 2.7). Distance functions and squared distance functions to g-convex sets are always g-convex. More precisely, we have the following (see (Alimisis et al., 2020, Lem. 2 in App. B) or (Jost, 2011, Thm. 5.6.1)).

Lemma 5 *Let $\mathcal{M}$ have sectional curvatures in the interval $[K, 0]$. If D is a closed g-convex set, then $x \mapsto \text{dist}(x, D)$ is g-convex and 1-Lipschitz globally, and C^∞ at points $x \notin D$.*

Fix $z \in \mathcal{M}$, and let $f(x) = \frac{1}{2}\text{dist}(x, z)^2$. Then f is C^∞, $\nabla f(x) = -\log_x(z)$ and $\nabla^2 f(x) \succeq I$ for all $x \in \mathcal{M}$, and $\|\nabla^2 f(x)\| \leq \zeta_{r\sqrt{-K}} \leq 1 + r\sqrt{-K}$ for all $x \in B(z, r)$.

3 Extending the $\tilde{\Omega}(\zeta)$ lower bound: a reduction argument

Building on (Hamilton and Moitra, 2021), Criscitiello and Boumal (2022a) prove the lower bound $\tilde{\Omega}(\zeta_r)$ for the strongly g-convex problem **P4**. We extend this result to the problems **P1**, **P2** and **P3**. For simplicity, in this section we only state the results for hyperbolic space. The more general result for Hadamard manifolds of bounded curvature (Theorem 44) can be found in Appendix F.1.

We have the following simple consequence of Theorem 24 of (Criscitiello and Boumal, 2022a).

Lemma 6 *Let $d \geq 2$, $L > 0$, $r \geq 64$, and $\mathcal{M} = \mathbb{H}^d$. Define $\hat{\epsilon} = \frac{1}{2^{10}r}$ and $\mu = 64\hat{\epsilon}L = \frac{L}{2^4 r}$. Let $\mathcal{A}$ be a deterministic first-order algorithm. There is a C^∞ function $f \colon \mathbb{H}^d \to \mathbb{R}$ with minimizer $x^* \in B(x_{\text{ref}}, \frac{3}{4}r)$ such that running $\mathcal{A}$ on f yields iterates $x_0, x_1, \ldots x_{T-1}$ satisfying $f(x_k) - f^* \geq 2\hat{\epsilon}Lr^2$ for all $k = 0, 1, \ldots T - 1$, where $T = \lfloor \frac{\zeta_r}{50\log(64\zeta_r)} \rfloor$.*

Moreover, f is μ-strongly g-convex in $\mathcal{M}$, and $\mu(12\mathscr{R} + 3)$-Lipschitz and $\mu(12\mathscr{R} + 9)$-smooth in the ball $B(x_{\text{ref}}, \mathscr{R})$, where $\mathscr{R} = 2^9 r \log^2(r)$. For all $x \notin B(x_{\text{ref}}, \mathscr{R})$, $f(x) = 3\mu\text{dist}(x, x_{\text{ref}})^2$.

The second paragraph in Lemma 6 is not stated explicitly by Criscitiello and Boumal (2022a) but is apparent from their proof. The proof of Lemma 6 can be found in Appendix F.1.

Using a reduction, we next prove lower bounds for the function class $\mathcal{F}^d_{r,L/2,0,L}$. The idea is that, given a hard function f from Lemma 6, we modify f so that it remains the same inside $B(x_{\text{ref}}, \mathscr{R})$, and outside $B(x_{\text{ref}}, \mathscr{R})$ it is not strongly g-convex but is strictly g-convex, $\frac{L}{2}$-Lipschitz and L-smooth. We know that $f(x)$ is proportional to $\text{dist}(x, x_{\text{ref}})^2$ outside $B(x_{\text{ref}}, \mathscr{R})$. We modify f there so that it behaves similarly to $\text{dist}(x, x_{\text{ref}})$ instead. This works because the function $x \mapsto \text{dist}(x, x_{\text{ref}})$ is g-convex and (outside a sufficiently large ball surrounding x_{ref}) that same function is 1-Lipschitz and 2-smooth on $\mathcal{M} = \mathbb{H}^d$ (Lee, 2018, Thm. 11.7).

Given any C^∞ function f, define the C^∞ function $\tilde{f}_{\mathscr{R}} \colon \mathcal{M} \to \mathbb{R}$ by

$$\tilde{f}_{\mathscr{R}}(x) = u_{\mathscr{R}}\left(\frac{1}{2}\text{dist}(x, x_{\text{ref}})^2\right)f(x), \tag{1}$$

where $u_{\mathscr{R}} \colon \mathbb{R} \to \mathbb{R}$ is a C^∞ function which is 1 on $(-\infty, \frac{1}{2}\mathscr{R}^2]$ and scales as $\sqrt{\frac{1}{2}\mathscr{R}^2 \cdot t^{-1}}$ for $t > \frac{1}{2}\mathscr{R}^2$ sufficiently large. See Appendix F.2 for the definition of $u_{\mathscr{R}}$. Suppose f is a hard function from Lemma 6. Since $\tilde{f}_{\mathscr{R}} = f$ in $B(x_{\text{ref}}, \mathscr{R})$, we know $\tilde{f}_{\mathscr{R}}$ is μ-strongly g-convex, and $\mu(12\mathscr{R}+3)$-Lipschitz and $\mu(12\mathscr{R}+9)$-smooth in $B(x_{\text{ref}}, \mathscr{R})$. Lemma 6 also guarantees $f(x) = 3\mu\text{dist}(x, x_{\text{ref}})^2$

for all $x \notin B(x_{\text{ref}}, \mathscr{R})$. In Appendix F.3, we use this fact to show that $\tilde{f}_{\mathscr{R}}$ is $12\mu\mathscr{R}$-Lipschitz, $24\mu\mathscr{R}$-smooth and strictly g-convex outside of $B(x_{\text{ref}}, \mathscr{R})$. Using the definitions of $\mu, \mathscr{R}$ in Lemma 6 as well as the bound $r \geq 64$, we conclude $\tilde{f}_{\mathscr{R}}$ is in $\mathcal{F}^d_{r,\tilde{L}/2,0,\tilde{L}}$ where $\tilde{L} = L \cdot 2^{10} \log^2(r)$.

Given the oracle $\mathcal{O}_f$ of any function f, we can use $\mathcal{O}_f$ to emulate the oracle $\mathcal{O}_{\tilde{f}_{\mathscr{R}}}$ using equation (1) and the formula (25) for $\nabla \tilde{f}_{\mathscr{R}}$ (given in Appendix F.3). To prove a lower bound for an algorithm $\tilde{\mathcal{A}}$ designed to minimize g-convex functions, we make $\tilde{\mathcal{A}}$ interact with the oracle $\mathcal{O}_{\tilde{f}_{\mathscr{R}}}$ (which we simulate using $\mathcal{O}_f$). This implicitly defines an algorithm $\mathcal{A}$ which interacts with $\mathcal{O}_f$. Explicitly, the algorithm $\mathcal{A}$ runs $\tilde{\mathcal{A}}$ as a subroutine as follows. If $\tilde{\mathcal{A}}$ outputs x_k, then $\mathcal{A}$ queries $\mathcal{O}_f$ at x_k, receives $(f(x_k), \nabla f(x_k))$ from $\mathcal{O}_f$, and then passes $(\tilde{f}_{\mathscr{R}}(x_k), \nabla \tilde{f}_{\mathscr{R}}(x_k))$ to $\tilde{\mathcal{A}}$, which it computes using $(f(x_k), \nabla f(x_k))$ and equations (1), (25).

Applying Lemma 6 to the algorithm $\mathcal{A}$, we know there is a function f with minimizer x^* so that $f(x_k) - f(x^*) \geq 2\hat{\epsilon}Lr^2$ for all $k \leq T - 1$. Since $\tilde{f}_{\mathscr{R}} = f$ in $B(x_{\text{ref}}, \mathscr{R})$ and $\tilde{f}_{\mathscr{R}}$ is strictly g-convex on $\mathcal{M}$, we know the minimizer of $\tilde{f}_{\mathscr{R}}$ is also x^* and that $\tilde{f}_{\mathscr{R}}(x_k) - \tilde{f}_{\mathscr{R}}(x^*) = f(x_k) - f(x^*)$ if $x_k \in B(x_{\text{ref}}, \mathscr{R})$. In Appendix F.5, we show that $\tilde{f}_{\mathscr{R}}(x_k) - \tilde{f}_{\mathscr{R}}(x^*) \geq 2\hat{\epsilon}Lr^2$ if $x_k \notin B(x_{\text{ref}}, \mathscr{R})$. Lastly, observe that (by design) if we run $\tilde{\mathcal{A}}$ on the function $\tilde{f}_{\mathscr{R}}$ then we get exactly the sequence $x_0, x_1, \ldots, x_{T-1}$. We have proven the following theorem.

Theorem 7 *Let $d \geq 2$, $\tilde{L} > 0$, $r \geq 64$, and $\mathcal{M} = \mathbb{H}^d$. Define $\hat{\epsilon} = \frac{1}{2^{10}r}$, $\epsilon = \frac{1}{2^{18}\zeta_r \log^2(\zeta_r)}$, $\epsilon' = \frac{1}{2^{18} \log^2(\zeta_r)}$ and $L = \frac{\tilde{L}}{2^{10} \log^2(r)}$. Let $\mathcal{A}$ be any deterministic first-order algorithm.*

There is a C^∞ function $\tilde{f} \in \mathcal{F}^d_{r,\tilde{L}/2,0,\tilde{L}}$ with unique minimizer x^* such that running $\mathcal{A}$ on $\tilde{f}$ yields iterates $x_0, x_1, \ldots$ satisfying

$$\tilde{f}(x_k) - \tilde{f}(x^*) \geq 2\hat{\epsilon}Lr^2 \geq \epsilon' \cdot \frac{\tilde{L}}{2}r \geq \epsilon \cdot \frac{1}{2}\tilde{L}r^2$$

for all $k = 0, 1, \ldots, T - 1$ where $T = \lfloor \frac{\zeta_r}{50 \log(64\zeta_r)} \rfloor$.

Theorem 7 shows that the $\tilde{\Omega}(\zeta_r)$ bound holds for problems **P1**, **P2** and **P3**, provided ϵ is not too big. Note that $\epsilon' = \tilde{\Theta}(1)$ and $\epsilon = \tilde{\Theta}(\frac{1}{\zeta_r})$ in Theorem 7 (recall that $\epsilon \leq \Theta(\frac{1}{\zeta_r})$ for problem **P3**).

In Table 1, the best known upper bounds all depend on ζ_r. Theorem 7 shows that this dependence is unavoidable. We know that a variant of Riemannian gradient descent for the smooth g-convex problem **P3** has complexity $O(\frac{1}{\epsilon})$.[12] Theorem 7 shows that in the regime $\epsilon = \tilde{\Theta}(\frac{1}{\zeta_r})$, *this variant of RGD is optimal* (up to log factors). Contrast this with the Euclidean case, where gradient descent is *never* optimal and acceleration is used to achieve the optimal complexity of $O(\frac{1}{\sqrt{\epsilon}})$.

4. Lower bounds for Lipschitz g-convex optimization: a resisting oracle argument

In this section we prove a lower bound for the Lipschitz g-convex problems **P1** and **P2**. For Euclidean space $\mathbb{R}^d$, such lower bounds are constructed using a maximum of affine functions $x \mapsto \max_{i=1,\ldots,d}\{\langle s_i e_i, x \rangle\}$, where $(e_i)_{i=1}^d$ are the standard basis vectors and $s_i \in \{+1, -1\}$ (Nemirovski, 1994, Sec 7.4). Before the algorithm makes any queries, there are 2^d possible minimizers,

---

12. Zhang and Sra (2016) prove the complexity $O(\frac{\zeta_r}{\epsilon})$ for RGD for problem **P3**. This can be improved to $O(\frac{1}{\epsilon})$ as follows. In Appendix D of their paper, Martínez-Rubio and Pokutta (2022) provide a variant on RGD for constrained optimization, which has complexity $O(\kappa)$ for problem **P4**. Using the reduction provided in Proposition 57, we find that a regularized version of that method has complexity $O(\frac{1}{\epsilon})$.

$\frac{r}{\sqrt{d}}(s_1, \ldots, s_d)$, corresponding to the vertices of a d-dimensional hypercube inscribed in a sphere of radius r. After each query by the algorithm, the set of possible minimizers is halved.

We use a similar technique here. However, we need a replacement for the affine functions used in Euclidean space. Affine functions are both g-convex and g-concave. On a Riemannian manifold, a function is affine if its Hessian vanishes identically. Innami (1982) shows that on most manifolds (including hyperbolic space), non-constant affine functions do not exist.[13] Moreover, the functions $x \mapsto \langle g, \log_y(x) \rangle$ do not work because they are not g-convex (see Appendix C).

Our main idea is to replace the linear functions by *distance functions to $(d-1)$-dimensional totally geodesic submanifolds, and suitably arrange those totally geodesic submanifolds.*

Theorem 8 *Let $r > 0$, $M > 0$ and T a positive integer. For every deterministic algorithm $\mathcal{A}$, there exists an M-Lipschitz g-convex function $f \in \mathcal{F}_{r,M,0,\infty}$ such that $\mathcal{A}$ requires at least T queries to find a point $x \in \mathcal{M}$ with $f(x) - f^* \leq Mr \cdot \frac{1}{2\zeta_r \sqrt{T}}$.*

Proof Without loss of generality, we can assume $M = 1$. Let $d = T$ and $(e_i)_{i=1}^d$ be an orthonormal basis of $T_{x_{\text{ref}}}\mathcal{M}$. Let $a > 0$ be such that $\tanh(a)/\tanh(r) = 1/\sqrt{d}$, and define $\delta = \frac{a}{2T}$. For $s \in \{+1, -1\}$ and $i = 1, \ldots, d$ define

$$z_i^s = \exp_{x_{\text{ref}}}(ase_i), \quad H_i^s = \{v \in T_{z_i^s}\mathcal{M} : \langle \log_{z_i^s}(x_{\text{ref}}), v \rangle = 0\}, \quad S_i^s = \exp_{z_i^s}(H_i^s)$$

and define the 1-Lipschitz g-convex functions

$$h_i^s(x) = \text{dist}(x, S_i^s) - \text{dist}(x_{\text{ref}}, S_i^s) = \text{dist}(x, S_i^s) - a.$$

These functions are g-convex by Lemmas 1 and 5.

We build the hard function for $\mathcal{A}$ iteratively as a max of functions h_i^s (shifted appropriately so that the subdifferential at x_k always consists of exactly one vector). Let $\mathcal{I}_0 = \{1, \ldots, d\}$ and $\mathcal{H}_0 = \emptyset$. For $k = 0, \ldots, T-1$, inductively define

$$
\begin{aligned}
\mathcal{H}_k &= ((F_\ell, x_\ell, g_\ell))_{\ell=0}^{k-1}, & x_k &= \mathcal{A}_k(\mathcal{H}_k), \\
(i_k, s_k) &\in \arg\max_{i \in \mathcal{I}_k, s \in \{+1,-1\}} h_i^s(x_k), & \mathcal{I}_{k+1} &= \mathcal{I}_k \setminus \{i_k\}, \\
f_k(x) &= \max_{\ell \in \{0,\ldots,k\}} \{h_{i_\ell}^{s_\ell}(x) - \ell\delta\}, & F_k &= f_k(x_k), g_k \in \partial f_k(x_k).
\end{aligned}
\tag{2}
$$

Let f be the 1-Lipschitz g-convex function f_{T-1}. We claim that running $\mathcal{A}$ on f produces exactly the sequence of iterates $x_0, \ldots, x_{T-1}$. It suffices to show $f_{T-1}(x) = f_k(x)$ for all k and all x in $B(x_k, \frac{\delta}{2})$. For all $\ell = k+1, \ldots, T-1$ we know $f_k(x_k) \geq h_{i_k}^{s_k}(x_k) - k\delta \geq h_{i_\ell}^{s_\ell}(x_k) - \ell\delta + \delta$, using the definition of (i_k, s_k). Therefore, $f_k(x_k) \geq \delta + \max_{\ell \in \{k+1,\ldots,T-1\}} \{h_{i_\ell}^{s_\ell}(x_k) - \ell\delta\}$. We conclude $f_k(x) \geq \max_{\ell \in \{k+1,\ldots,T-1\}} \{h_{i_\ell}^{s_\ell}(x) - \ell\delta\}$ for all $x \in B(x_k, \frac{\delta}{2})$, since f_k is 1-Lipschitz. Hence, $f_{T-1}(x) = \max\{f_k(x), \max_{\ell \in \{k+1,\ldots,T-1\}} \{h_{i_\ell}^{s_\ell}(x) - \ell\delta\}\} = f_k(x)$ for all $x \in B(x_k, \delta/2)$.

It remains to lower bound $f(x_k) - f^*$ for all k. Define $x^* = \exp_{x_{\text{ref}}}(\frac{r}{\sqrt{d}} \sum_{k=0}^{T-1} \nabla h_{i_k}^{s_k}(x_{\text{ref}}))$. We know that $(\nabla h_{i_k}^{s_k}(x_{\text{ref}}))_{k=0}^{T-1}$ forms an orthonormal basis of $T_{x_{\text{ref}}}\mathcal{M}$, because we defined $\mathcal{I}_{k+1} = \mathcal{I}_k \setminus \{i_k\}$. Therefore, $\text{dist}(x_{\text{ref}}, x^*) = r$. In Appendix E, we show that $x^* \in \bigcap_{k=0}^{T-1} S_{i_k}^{s_k}$ using the

---

13. In fact, more can be said: on a hyperbolic space, if f is a C^3 g-convex function whose Hessian vanishes at even just a single point x then necessarily $\nabla f(x) = 0$, and so f cannot be affine unless it is constant – see Proposition 55.

hyperbolic law of cosines (Ratcliffe, 2019, Ch. 3.5). From there, we know that $\text{dist}(x^*, S_{i_k}^{s_k}) = 0$ for all k, so $f(x^*) = \max_{k \in \{0,\dots,T-1\}} \{-a - k\delta\} = -a$. Also, $f_{T-1}(x) \geq h_{i_0}^{s_0}(x) \geq -a$ for all $x \in \mathcal{M}$. We conclude that $x^* \in B(x_{\text{ref}}, r)$ is a global minimizer of f on $\mathcal{M}$ with $f^* = f(x^*) = -a$.

Next, we lower bound $f(x_k)$. For $i \in \{1, \dots, d\}$ and $s \in \{+1, -1\}$, g-convexity of h_i^s and $\nabla h_i^s(x_{\text{ref}}) = -se_i$ imply

$$\{x \in \mathcal{M} : h_i^s(x) < h_i^s(x_{\text{ref}}) = 0\} \subseteq \exp_{x_{\text{ref}}}(\{v \in \mathrm{T}_{x_{\text{ref}}}\mathcal{M} : \langle -se_i, v \rangle < 0\}).$$

Since $\bigcup_{s \in \{+1,-1\}}(\exp_{x_{\text{ref}}}(\{v \in \mathrm{T}_{x_{\text{ref}}}\mathcal{M} : \langle -se_i, v \rangle \geq 0\})) = \mathcal{M}$, we find that $\bigcup_{s \in \{+1,-1\}}\{x \in \mathcal{M} : h_i^s(x) \geq 0\} = \mathcal{M}$. Therefore, $h_{i_k}^{s_k}(x_k) \geq \max_{s \in \{+1,-1\}} h_{i_k}^s(x_k) \geq 0$ for all $k = 0, \dots, T-1$.

Therefore, $f(x_k) = f_{T-1}(x_k) \geq h_{i_k}^{s_k}(x_k) - k\delta \geq -(T-1)\delta \geq -\frac{a}{2}$ for all $k = 0, \dots, T-1$. Combining this upper bound on f^* and lower bound on $f(x_k)$, we conclude $f(x_k) - f^* \geq -\frac{a}{2} - (-a) = \frac{a}{2} = \frac{1}{2}r \cdot r^{-1} \text{arctanh}(\tanh(r)/\sqrt{T}) \geq r\frac{1}{2\zeta_r\sqrt{T}}$. ∎

Remark 9 *Theorem 8 says that if $\epsilon = \frac{1}{2\zeta_r\sqrt{d}}$, then $T_{\epsilon,r,d} \geq d = \frac{d}{\log(2\zeta_r\sqrt{d})}\log(\epsilon^{-1})$. The lower bound in Theorem 8 implies the lower bound $q_{d,r} \geq \frac{d}{\log(2\zeta_r\sqrt{d})}$ for the low-dimensional problem P1.*

5 Lower bounds for smooth g-convex optimization: a Moreau smoothing argument

In this section we prove lower bounds for the smooth g-convex problems **P3** and **P4**. In smooth convex optimization, the most well-known technique for proving a lower bound is building a so-called "worst function in the world" (Nesterov, 2004, 2.1.2). However, there is perhaps the less well-known technique of smoothing convex Lipschitz functions, usually via *Moreau envelopes* (Guzmán and Nemirovski, 2015). We adopt the smoothing approach because this technique works far better than the other on manifolds, and the smoothing technique seems more general. That is, given a lower bound construction for Lipschitz g-convex optimization which improves over the one we present in Section 4, smoothing is likely to produce an improved lower bound in the smooth setting.

The (Riemannian) Moreau envelope of a function $f: \mathcal{M} \to \mathbb{R}$ is the function defined as follows:

$$f_\lambda: \mathcal{M} \to \mathbb{R}, \qquad f_\lambda(x) = \inf_{y \in \mathcal{M}} \left\{ f(y) + \frac{1}{2\lambda}\text{dist}(x,y)^2 \right\}. \tag{3}$$

In convex analysis, the Moreau envelope is closely related to Fenchel duality. There is no especially satisfying theory of Fenchel duality on Hadamard manifolds (but see (Silva Louzeiro et al., 2022; Hirai, 2022)). However, the Riemannian Moreau envelope can be studied directly without reference to duality. It satisfies the following properties, proved in (Azagra and Ferrera, 2005, 2015). For completeness, we provide a brief (simpler) proof in Appendix G.

Lemma 10 *Let $\mathcal{M}$ be a Hadamard manifold with curvature lower bounded by -1. Let $f: \mathcal{M} \to \mathbb{R}$ be 1-Lipschitz and g-convex. Then, $f_\lambda: \mathcal{M} \to \mathbb{R}$ with $\lambda > 0$ is g-convex, 1-Lipschitz, and $\frac{1}{\tanh(\lambda)}$-smooth. Moreover, $f(x) \geq f_\lambda(x) \geq f(x) - \lambda$ for all $x \in \mathcal{M}$, and the value of f_λ at $x \in \mathcal{M}$ is determined by the values of f in the ball $B(x, \lambda)$: $f_\lambda(x) = \min_{y \in B(x,\lambda)} \left\{ f(y) + \frac{1}{2\lambda}\text{dist}(x,y)^2 \right\}$.*

We now smooth the construction from Section 4, following (Guzmán and Nemirovski, 2015).

9

Theorem 11 *Let $r > 0$, $L > 0$ and T a positive integer. For every deterministic algorithm $\mathcal{A}$, there exists an L-smooth g-convex function $f \in \mathcal{F}_{r,\infty,0,L}$ such that $\mathcal{A}$ requires at least T queries to find a point $x \in \mathcal{M}$ with $f(x) - f^* \leq \frac{1}{2}Lr^2 \cdot \frac{1}{8\zeta_r^2 T^2}$.*

Proof Define $d, e_i, a, \delta, z_i^s, S_i^s, h_i^s$ as in the proof of Theorem 8. Define $\lambda = \frac{\delta}{4}$, $\mathcal{I}_0 = \{1, \ldots, d\}$ and $\mathcal{H}_0 = \emptyset$. For $k = 0, \ldots, T - 1$, inductively define $\mathcal{H}_k, x_k, (i_k, s_k), \mathcal{I}_{k+1}, f_k$ as in equation (2), and let $F_k = f_{\lambda,k}(x_k), g_k = \nabla f_{\lambda,k}(x_k)$ where $f_{\lambda,k}$ is the Moreau envelope of f_k with parameter λ.

Let f equal the 1-Lipschitz g-convex function $f_{\lambda,T-1}$. We claim that running $\mathcal{A}$ on f produces exactly the sequence of iterates $x_0, \ldots, x_{T-1}$. It suffices to show for each k that $f_{\lambda,T-1}(x) = f_{\lambda,k}(x)$ for all $x \in B(x_k, \frac{\delta}{4})$. Lemma 10 implies that the values of $f_{\lambda,T-1}$ and $f_{\lambda,k}$ in $B(x_k, \frac{\delta}{4})$ are determined, respectively, by the values of f_{T-1}, f_k in $B(x_k, \lambda + \frac{\delta}{4}) = B(x_k, \delta/2)$. However, in the proof of Theorem 8, we already showed for each k that $f_{T-1}(x) = f_k(x)$ for all $x \in B(x_k, \delta/2)$. We conclude $f_{\lambda,T-1}(x) = f_{\lambda,k}(x)$ for all $x \in B(x_k, \frac{\delta}{4})$, as desired.

Let us lower bound the suboptimality gap $f(x_k) - f^*$ for each k. Take $x^* \in B(x_{\text{ref}}, r)$ as defined in Theorem 8. From the proof of Theorem 8, we know x^* is a global minimizer of f_{T-1} with $f_{T-1}(x^*) = -a$. It is immediate from equation (3) that x^* is therefore a global minimizer of $f = f_{\lambda,T-1}$ with $f^* = f_{\lambda,T-1}(x^*) = -a$. From the proof of Theorem 8, $f_{T-1}(x_k) \geq -(T-1)\delta$. So $f(x_k) \geq f_{T-1}(x_k) - \lambda \geq -T\delta \geq -\frac{a}{2}$ by Lemma 10. Hence for all k,

$$f(x_k) - f^* \geq \frac{a}{2} = \frac{1}{2}\frac{Lr^2}{T^2}\frac{a}{Lr^2/T^2} = \frac{1}{2}\frac{Lr^2}{T^2}\frac{T^2}{r^2}a\tanh\left(\frac{a}{8T}\right)$$

$$= \frac{1}{2}\frac{Lr^2}{T^2}\left[\frac{T^2}{r^2}\operatorname{arctanh}\left(\frac{\tanh(r)}{\sqrt{T}}\right)\tanh\left(\frac{1}{8T}\operatorname{arctanh}\left(\frac{\tanh(r)}{\sqrt{T}}\right)\right)\right] \geq \frac{1}{2}\frac{Lr^2}{T^2}\left[\frac{1}{8\zeta_r^2}\right],$$

where the smoothness constant of f equals $L = \frac{1}{\tanh(\lambda)} = \frac{1}{\tanh(\frac{a}{8T})}$, by Lemma 10. ∎

Proposition 57 in Appendix I provides a reduction between the smooth nonstrongly and strongly g-convex settings (see also (Martínez-Rubio, 2021)). Using this proposition, the lower bound in Theorem 11 extends to an $\Omega(\frac{\sqrt{\kappa}}{\log(\zeta_r\kappa)})$ lower bound for the smooth strongly g-convex problem **P4**.

Corollary 12 *Let $r > 0$ and $\kappa > \zeta_r$. For $\epsilon = \frac{2}{\kappa - \zeta_r}$, $T_{\epsilon,r,\kappa} \geq \frac{\sqrt{\kappa - \zeta_r}}{8\log(\kappa[3+\zeta_r^2])}\log(\epsilon^{-1})$. In particular, $q_{\kappa,r} = \sup_{\epsilon \in (0,1)}\{\frac{T_{\epsilon,r,\kappa}}{\log(\epsilon^{-1})}\} \geq \frac{\sqrt{\kappa - \zeta_r}}{8\log(\kappa[3+\zeta_r^2])}$.*

Lastly, we mention that if the optimization is carried out in a very small region, e.g., $r \leq O(\sqrt{\epsilon})$, then smooth g-convex optimization reduces to Euclidean convex optimization, and in particular upper and lower bounds from Euclidean space carry over. See Appendix D for details.

5.1 Why assume $\epsilon \leq O(\zeta^{-1})$?

In problem **P3** we assumed $\epsilon \leq \frac{8}{\zeta_r}$. The following proposition explains why this is justified.

Proposition 13 *Let $f: \mathbb{H}^d \to \mathbb{R}$ be differentiable, globally g-convex, and L-smooth in $B(x_{\text{ref}}, r)$. Suppose $\nabla f(x^*) = 0$ and $x^* \in B(x_{\text{ref}}, r)$. Then $f(x_{\text{ref}}) - f^* \leq \frac{1}{2}Lr^2 \cdot \frac{8}{\zeta_r}$.*

Proposition 13 follows from the following fact expressed in Proposition 51: if $f \colon \mathbb{H}^d \to \mathbb{R}$ (not necessarily g-convex) is L-smooth (in a large enough region) then it is also $\frac{\pi\sqrt{5}}{2}L$-Lipschitz.[14] We can also use this fact to prove that $\kappa \geq \Omega(\zeta_r)$ even holds for a g-convex domain of diameter $2r$ which is *not* a ball, as long as the domain is not too *eccentric*: see Proposition 53.

6 Lower bound $\Omega(\frac{\zeta}{\epsilon^2})$ for subgradient descent: a worst-function-in-the-world argument

For problem **P2**, we showed the lower bound $\tilde{\Omega}(\zeta_r + \frac{1}{\zeta_r^2\epsilon^2})$ in Section 4. This lower bound does not match the upper bound $O(\frac{\zeta_r}{\epsilon^2})$. We conjecture that this upper bound is optimal. To support this conjecture, in Appendix A we prove an $\Omega(\frac{\zeta_r}{\epsilon^2})$ lower bound for subgradient descent for problem **P2**. For simplicity, we focus on subgradient descent *with Polyak step size* (see Appendix A.3), although we expect the proof extends to other similar step size choices, like the one in (Zhang and Sra, 2018).

Corollary 14 *Let $\epsilon \leq \frac{1}{4\sqrt{2}}$ and $M > 0$. Let $T = d = \lfloor \frac{1}{32}\frac{\zeta_r}{\epsilon^2} \rfloor$. There is a globally g-convex and M-Lipschitz function $f \colon \mathcal{M} = \mathbb{H}^d \to \mathbb{R}$ so that the following holds: There is an oracle for f so that running subgradient descent with Polyak step size with that oracle produces iterates $x_0, x_1, \dots$ satisfying $f(x_k) - f^* \geq \epsilon M r$ for all $k = 0, \dots, T-1$.*

We actually prove the $\Omega(\frac{\zeta_r}{\epsilon^2})$ lower bound for a *large class of (intuitively reasonable) algorithms*, not just subgradient descent, see Theorem 29. Corollary 14 is a consequence of this result.

Corollary 14 and Theorem 29 are proven using a novel "worst function in the world"

$$f(x) = \text{dist}(x, x^*) + \max_{i=0,\dots,d-2}\left\{\frac{1}{4\epsilon}\text{dist}(x, L_i)\right\},$$

for an appropriately chosen $x^* \in \partial B(x_{\text{ref}}, r)$, and a particular choice of (g-convex) half-spaces $L_k = \exp_{y_k}(\{v \in \mathrm{T}_{y_k}\mathcal{M} : \langle V_k, v \rangle \geq 0\})$. If an algorithm were simply minimizing the "base function" $x \mapsto \text{dist}(x, x^*)$, then one subgradient query would reveal x^*. The $\max$-term in f is carefully constructed to perform a correction to the subgradients of this base function so that one query of f reveals little information about x^*.

The key geometric fact leading to the $\Omega(\frac{\zeta_r}{\epsilon^2})$ lower bound is the following. Consider a right hyperbolic triangle with vertices x^*, x_0, x_1, where the right angle is at x_1, and the angle at x_0 is $\theta \in (0, \frac{\pi}{2})$. Define the sidelengths $\text{dist}(x_0, x^*) = r_0$ and $\text{dist}(x_1, x^*) = r_1$. Let $\epsilon = \cos(\theta)$. Then, in hyperbolic space $r_1^2 \approx (1 - \frac{\epsilon^2}{\zeta_{r_0}})r_0^2$ for ϵ sufficiently small, whereas in Euclidean space $r_1^2 = (1 - \epsilon^2)r_0^2$. The side r_1 is significantly longer in hyperbolic space than in Euclidean space.

7 Cutting-planes game: a width-bounded-separators argument

We conjecture that the optimal complexity for the low-dimensional Lipschitz g-convex problem **P1** is $\Theta(\zeta_r d)$. In this section we consider a "cutting-planes game" which serves as a proxy for problem **P1**, and we prove an $\tilde{\Omega}(\zeta_r d)$ lower bound for this game. We take $\mathcal{M} = \mathbb{H}^d$ throughout. Let us motivate the game. Consider an algorithm $\mathcal{A}$ minimizing a g-convex 1-Lipschitz function f with a

---

14. The proof of Proposition 51 is due to Petrunin (2023, pgs. 25, 33). We fill in the details in Appendix H. This fact closely mirrors the following result: if f is ρ-Hessian-Lipschitz (in a large enough region) then it is 2ρ-Lipschitz and $\sqrt{8}\rho$-smooth (Criscitiello and Boumal, 2022b, Remark 3.2).

unique minimizer x^* in $B(x_{\mathrm{ref}}, r)$, whose goal is to find a point x with $f(x) - f^* \leq \epsilon r$, $\epsilon \in (0, 1)$. If $\mathcal{A}$ queries $x_k \in \mathcal{M}$, it learns $F_k = f(x_k)$ and a subgradient $g_k \in \partial f(x_k)$ satisfying $\|g_k\| \leq 1$. Since f is g-convex, we know $f(x) \geq F_k + \langle g_k, \log_{x_k}(x) \rangle$ for all x. In particular, taking $x = x^*$ we find $\langle g_k, \log_{x_k}(x^*) \rangle \leq 0$. Moreover, $0 \geq F_k - f^* + \langle g_k, \log_{x_k}(x^*) \rangle \geq F_k - f^* - \mathrm{dist}(x^*, S_k)$, where $S_k = \exp_{x_k}(\{v \in \mathrm{T}_{x_k} \mathcal{M} : \langle v, g_k \rangle = 0\})$. The last inequality follows from the g-convexity of $x \mapsto \mathrm{dist}(x, S_k)$ and $-g_k$ is a subgradient of $x \mapsto \mathrm{dist}(x, S_k)$. Therefore, the algorithm solves this g-convex problem if $\mathrm{dist}(x^*, S_k) \leq \epsilon r$ for some k.

In the cutting-planes game the algorithm is given a ball $B(x_{\mathrm{ref}}, r)$ which contains a point x^*. The algorithm has access to the following (weaker) oracle: when the algorithm queries $x_k \in \mathcal{M}$ it receives a unit vector $g_k \in \mathrm{T}_{x_k} \mathcal{M}$ such that $\langle \log_{x_k}(x^*), g_k \rangle \leq 0$. The algorithm's goal is to find x_k so that $\mathrm{dist}(x^*, S_k) \leq \epsilon r$ where $S_k = \exp_{x_k}(\{v \in \mathrm{T}_{x_k} \mathcal{M} : \langle v, g_k \rangle = 0\})$. Equivalently, the algorithm's goal is to find x_k such that the ball $B(x^*, \epsilon r)$ *intersects* S_k.

Evidently, an algorithm which can solve the cutting-planes game can also solve the Lipschitz g-convex optimization problem. Nesterov (2004, §3.2.6) calls such methods "cutting-plane schemes," and these include the center of gravity and ellipsoid methods. A cutting-plane scheme "forgets" that it is minimizing a g-convex function, and simply plays the cutting-planes game. In this sense, our lower bound below serves as a lower bound for cutting-plane schemes.

Before we proceed, we recall that the complexity of the cutting-planes game is $\Theta(d)$ for Euclidean space. For hyperbolic space, the best known upper bound is $O(\zeta_r d^2)$ (Rusciano, 2019), and the best known lower bound was $\tilde{\Omega}(\zeta_r)$ (from Section 3). We improve this lower bound to $\tilde{\Omega}(\zeta_r d)$.

Theorem 15 *Let $d \geq 3, r \geq 1280(d-1)\log(d)$ and $\epsilon = \frac{1}{320(d-1)}$. For every algorithm, there is an $x^* \in B(x_{\mathrm{ref}}, r)$ and a sequence of oracle responses $(g_k)_{k \geq 0}$ such that the algorithm needs at least $T = \frac{1}{32}(d-1)r$ queries to find a point x_k so that $\mathrm{dist}(x^*, S_k) \leq \epsilon r$ (or equivalently $S_k \cap B(x^*, \epsilon r) \neq \emptyset$), where $S_k = \exp_{x_k}(\{v \in \mathrm{T}_{x_k} \mathcal{M} : \langle v, g_k \rangle = 0\})$.*

In the Euclidean case, the lower bound is constructed by tesselating space by hypercubes and choosing oracle hyperplanes which are parallel to the sides of the cube (Nesterov, 2004, §3.2.5). This construction does not generalize, so we propose a new strategy relying on "width-bounded separators" which first appeared in complexity theory literature (Fu, 2011), and whose properties were worked out by Kisfaludi-Bak (2020) for hyperbolic space. We use these to show Theorem 15.

Let us sketch the idea behind the proof of Theorem 15. We start with a packing of $B(x_{\mathrm{ref}}, r)$ with $e^{\Theta(dr)}$ balls of radii ϵr. The initial set of candidates for x^*, denoted A_0, consists of the centers of these balls. At each iteration k, we maintain a set of possible remaining candidates for x^*, denoted A_k. When the algorithm queries x_k, we show that there exists a unit vector $\tilde{g}_k \in \mathrm{T}_{x_k} \mathcal{M}$ such that $S_k = \exp_{x_k}(\{v \in \mathrm{T}_{x_k} \mathcal{M} : \langle v, \tilde{g}_k \rangle = 0\})$ intersects at most $\frac{1}{2}|A_k|$ candidate balls (see Lemma 34). We then choose $g_k \in \{\tilde{g}_k, -\tilde{g}_k\}$ so that the next set of candidate minimizers $A_{k+1} \subseteq A_k$ satisfies $|A_{k+1}| \geq \frac{1}{2}(|A_k| - \frac{1}{2}|A_k|) = \frac{1}{4}|A_k|$. Hence, after $T = \Theta(\log(|A_0|)) = \Theta(\log(e^{dr})) = \Theta(dr)$ iterations A_T will be nonempty. We take x^* to be an element of A_T. The complete proof of Theorem 15 can be found in Appendix B.

8 Geodesically convex interpolation: geometric obstructions

To build lower bounds for g-convex optimization, we must choose a collection of function values and gradients $(F_i, x_i, g_i)_{i=1}^N$, and be able to build g-convex functions which interpolate those function values and gradients. We say the data $(F_i, x_i, g_i)_{i=1}^N$ is interpolated by a g-convex function

$f \colon \mathcal{M} \to \mathbb{R}$ if $f(x_i) = F_i$ and $g_i \in \partial f(x_i)$ for all i. If the data $(F_i, x_i, g_i)_{i=1}^{N}$ is interpolated by the μ-strongly g-convex function f, $\mu \geq 0$, they must satisfy the "necessary conditions"

$$F_j \geq F_i + \langle g_i, \log_{x_i}(x_j) \rangle + \frac{\mu}{2}\mathrm{dist}(x_i, x_j)^2 \quad \forall i, j \in \{1, \ldots, N\}. \tag{4}$$

If $\mathcal{M}$ is a Euclidean space these necessary conditions are sufficient for interpolation by a μ-strongly convex function. This forms the basis for new insights into the complexity of convex optimization (Taylor et al., 2016). Unfortunately, in the g-convex case these conditions are *not sufficient* even for just three points.

Proposition 16 *Let $\mathcal{M} = \mathbb{H}^2$. There exists data $(F_i, x_i, g_i)_{i=1}^{3}$ such that $F_j \geq F_i + \langle g_i, \log_{x_i}(x_j) \rangle$ for all $i, j \in \{1, 2, 3\}$, yet this data cannot be interpolated by any g-convex function.*

The geometric fact underlying Proposition 16 is that the altitude of an isoceles hyperbolic triangle is shorter than the altitude in a corresponding Euclidean triangle (see Proposition 37 in Appendix C).

We do, however, have the following sufficient conditions, whose proof relies on the fact that the function $x \mapsto \langle g, \log_y(x) \rangle$ is $\|g\|$-smooth on $\mathbb{H}^d$ (see Lemma 40 in Appendix C).

Proposition 17 *Let $\mathcal{M} = \mathbb{H}^d$. Consider the data $(F_i, x_i, g_i)_{i=1}^{N}$. Assume inequalities (4) hold, and $\|g_i\| \leq \frac{\mu}{2}$ for all i. Then $(F_i, x_i, g_i)_{i=1}^{N}$ is interpolated by a $\frac{\mu}{2}$-strongly g-convex function.*

In Appendix C, we take a more careful look at g-convex interpolation. For example, in Proposition 38, we give necessary and sufficient conditions for interpolation (but they seem to be of limited use), and link the difficulty of g-convex interpolation to the fact that g-convexity does not naturally fall into the framework of abstract convexity (Rubinov, 2010).

References

N. Agarwal, N. Boumal, B. Bullins, and C. Cartis. Adaptive regularization with cubics on manifolds. *Mathematical Programming*, 188(1):85–134, 2020. doi: 10.1007/s10107-020-01505-1.

K. Ahn and S. Sra. From nesterov's estimate sequence to riemannian acceleration. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 84–118. PMLR, 09–12 Jul 2020. URL https://proceedings.mlr.press/v125/ahn20a.html.

F. Alimisis, A. Orvieto, G. Becigneul, and A. Lucchi. A continuous-time perspective for modeling acceleration in riemannian optimization. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 1297–1307. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/alimisis20a.html.

Z. Allen-Zhu, A. Garg, Y. Li, R. Oliveira, and A. Wigderson. Operator scaling via geodesically convex optimization, invariant theory and polynomial identity testing. In *Proceedings of the 50th Annual ACM SIGACT Symposium on the Theory of Computing (STOC 2018)*, 2018. doi: 10.1145/3188745.3188942.

D. Azagra and J. Ferrera. Inf-convolution and regularization of convex functions on riemannian manifolds of nonpositive curvature. *Revista Matematica Complutense*, 19:323–345, 2005.

D. Azagra and J. Ferrera. Regularization by sup–inf convolutions on riemannian manifolds: An extension of lasry–lions theorem to manifolds of bounded curvature. *Journal of Mathematical Analysis and Applications*, 423(2):994–1024, 2015. ISSN 0022-247X. doi: https://doi.org/10.1016/j.jmaa.2014.10.022. URL https://www.sciencedirect.com/science/article/pii/S0022247X14009457.

M. Bacák. *Convex analysis and optimization in Hadamard spaces*, volume 22 of *De Gruyter Series in Nonlinear Analysis and Applications*. Walter de Gruyter GmbH & Co KG, 2014.

G.C. Bento, O.P. Ferreira, and J.G. Melo. Iteration-complexity of gradient, subgradient and proximal point methods on Riemannian manifolds. *Journal of Optimization Theory and Applications*, 173(2):548–562, 2017. doi: 10.1007/s10957-017-1093-4.

Nicolas Boumal. *An introduction to optimization on smooth manifolds*. Cambridge University Press, 2023. doi: 10.1017/9781009166164. URL https://www.nicolasboumal.net/book.

M. R. Bridson and A. Haefliger. *Metric Spaces of Non-Positive Curvature*. Springer-Verlag Berlin Heidelberg, 1999. doi: 10.1007/978-3-662-12494-9.

Bang-Yen Chen. Chapter 3 - riemannian submanifolds. In Franki J.E. Dillen and Leopold C.A. Verstraelen, editors, *Handbook of Differential Geometry*, volume 1 of *Handbook of Differential Geometry*, pages 187–418. North-Holland, 2000. doi: https://doi.org/10.1016/S1874-5741(00)80006-0. URL https://www.sciencedirect.com/science/article/pii/S1874574100800060.

C. Criscitiello and N. Boumal. Negative curvature obstructs acceleration for strongly geodesically convex optimization, even with exact first-order oracles. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 496–542. PMLR, 02–05 Jul 2022a. URL https://proceedings.mlr.press/v178/criscitiello22a.html.

C. Criscitiello and N. Boumal. An accelerated first-order method for non-convex optimization on manifolds. *Journal of Foundations of Computational Mathematics*, 2022b. doi: 10.1007/s10208-022-09573-9.

R.-A. Dragomir, A. Taylor, J. Bolte, and A. d'Aspremont. Optimal Complexity and Certification of Bregman First-Order Methods. *Mathematical Programming*, April 2021. doi: 10.1007/s10107-021-01618-1. URL https://hal.inria.fr/hal-02384167.

F. Dragomirescu and C. Ivan. The smallest convex extensions of a convex function. *Optimization*, 24(3-4):193–206, 1992. doi: 10.1080/02331939208843789. URL https://doi.org/10.1080/02331939208843789.

P. T. Fletcher, S. Venkatasubramanian, and S. C. Joshi. The geometric median on Riemannian manifolds with application to robust atlas estimation. *NeuroImage*, 45:S143–S152, 2009.

C. Franks and A. Moitra. Rigorous guarantees for tyler's m-estimator via quantum expansion. In *33rd Annual Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 1–32. PMLR, 2020. URL http://proceedings.mlr.press/v125/franks20a/franks20a.pdf.

Bin Fu. Theory and application of width bounded geometric separators. *Journal of Computer and System Sciences*, 77(2):379–392, 2011. ISSN 0022-0000. doi: https://doi.org/10.1016/j.jcss.2010.05.003. URL https://www.sciencedirect.com/science/article/pii/S0022000010000577. Adaptivity in Heterogeneous Environments.

Cristóbal Guzmán and Arkadi Nemirovski. On lower complexity bounds for large-scale smooth convex optimization. *Journal of Complexity*, 31(1):1–14, 2015. ISSN 0885-064X. doi: https://doi.org/10.1016/j.jco.2014.08.003. URL https://www.sciencedirect.com/science/article/pii/S0885064X14000831.

L. Hamilton and A. Moitra. No-go theorem for acceleration in the hyperbolic plane. *arXiv: 2101.05657*, 2021.

Hiroshi Hirai. Convex analysis on hadamard spaces and scaling problems, 2022. URL https://arxiv.org/abs/2203.03193.

Nobuhiro Innami. Splitting theorems of riemannian manifolds. *Compositio Mathematica*, 47(3): 237–247, 1982.

J. Jost. *Riemannian Geometry and Geometric Analysis*. Universitext. Springer Berlin, Heidelberg, 6 edition, 2011. ISBN 978-3-642-21298-7.

H. Karcher. Riemannian center of mass and mollifier smoothing. *Communications on pure and applied mathematics*, 30(5):509–541, 1977.

Jungbin Kim and Insoon Yang. Accelerated gradient methods for geodesically convex optimization: Tractable algorithms and convergence analysis. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 11255–11282. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/kim22k.html.

S. Kisfaludi-Bak. *Hyperbolic intersection graphs and (quasi)-polynomial time*, pages 1621–1638. 2020. doi: 10.1137/1.9781611975994.100.

S. Lang. *Fundamentals of differential geometry*, volume 191. Springer Science & Business Media, 1999. doi: 10.1007/978-1-4612-0541-8.

J. M. Lee. *Introduction to Smooth Manifolds*, volume 218 of *Graduate Texts in Mathematics*. Springer-Verlag New York, 2nd edition, 2012. doi: 10.1007/978-1-4419-9982-5.

J. M. Lee. *Introduction to Riemannian Manifolds*, volume 176 of *Graduate Texts in Mathematics*. Springer, 2nd edition, 2018. doi: 10.1007/978-3-319-91755-9.

Alexander Lytchak and Anton Petrunin. About every convex set in any generic riemannian manifold. *Journal für die reine und angewandte Mathematik (Crelles Journal)*, 2022(782):235–245, 2022. doi: doi:10.1515/crelle-2021-0058. URL https://doi.org/10.1515/crelle-2021-0058.

D. Martínez-Rubio. Global Riemannian acceleration in hyperbolic and spherical spaces. *arXiv: 2012.03618*, 2021.

David Martínez-Rubio and Sebastian Pokutta. Accelerated riemannian optimization: Handling constraints with a prox to bound geometric penalties, 2022. URL https://arxiv.org/abs/2211.14645.

A. Nemirovski. *Information-based complexity of convex programming*. 1994. URL https://www2.isye.gatech.edu/~nemirovs/Lec_EMCO.pdf.

A. S. Nemirovskii and D.B. Yudin. *Problem complexity and method efficiency in optimization.* Wiley-Interscience series in discrete mathematics. Wiley, Chichester ;, 1983. ISBN 0471103454.

Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87 of *Applied optimization*. Springer, 2004. ISBN 978-1-4020-7553-7.

Y. E. Nesterov. Gradient methods for minimizing composite functions. *Math. Program.*, 140 (1):125–161, 2013. doi: 10.1007/s10107-012-0629-5. URL https://doi.org/10.1007/s10107-012-0629-5.

B. O'Neill. *Semi-Riemannian geometry: with applications to relativity*, volume 103. Academic Press, 1983.

Anton Petrunin. PIGTIKAL (puzzles in geometry that i know and love), 2023. URL https://arxiv.org/abs/0906.0290.

B.T. Polyak. *Introduction to optimization*. Optimization Software, 1987.

T. Rapcsák. *Smooth Nonlinear Optimization in $\mathbb{R}^n$*, volume 19 of *Nonconvex Optimization and Its Applications*. Springer, 1997. doi: 10.1007/978-1-4615-6357-0.

J. G. Ratcliffe. *Foundations of Hyperbolic Manifolds*. Springer International Publishing, 2019. doi: 10.1007/978-3-030-31597-9.

A. Rubinov. *Abstract Convexity and Global Optimization*. Springer Series in Nonconvex Optimization and Its Applications. Springer, 2010. ISBN 978-1-4757-3200-9.

A. Rusciano. A riemannian corollary of helly's theorem. *arXiv: 1804.10738*, 2019.

T. Sakai. *Riemannian geometry*, volume 149. American Mathematical Society, 1996.

Maurício Silva Louzeiro, Ronny Bergmann, and Roland Herzog. Fenchel duality and a separation theorem on hadamard manifolds. *SIAM Journal on Optimization*, 32(2):854–873, 2022. doi: 10.1137/21M1400699. URL https://doi.org/10.1137/21M1400699.

A. Taylor. *Convex interpolation and performance estimation of first-order methods for convex optimization*. PhD thesis, 2018. URL https://dial.uclouvain.be/pr/boreal/object/boreal:182881.

A. B. Taylor, J. M. Hendrickx, and F. Glineur. Smooth strongly convex interpolation and exact worst-case performance of first-order methods. *Mathematical Programming*, 161:307–345, 2016. doi: 10.1007/s10107-016-1009-3.

A.B. Taylor, J.M. Hendrickx, and F. Glineur. Exact worst-case performance of first-order methods for composite convex optimization. *SIAM Journal on Optimization*, 27(3):1283–1313, 2017. doi: 10.1137/16M108104X. URL https://doi.org/10.1137/16M108104X.

Adrien Taylor and Yoel Drori. An optimal gradient method for smooth strongly convex minimization. *Mathematical Programming*, June 2022. doi: 10.1007/s10107-022-01839-y. URL https://doi.org/10.1007/s10107-022-01839-y.

C. Udrişte. *Convex functions and optimization methods on Riemannian manifolds*, volume 297 of *Mathematics and its applications*. Kluwer Academic Publishers, 1994. doi: 10.1007/978-94-015-8390-9.

H. Zhang and S. Sra. First-order methods for geodesically convex optimization. In *Conference on Learning Theory*, pages 1617–1638, 2016.

H. Zhang and S. Sra. An estimate sequence for geodesically convex optimization. In S. Bubeck, V. Perchet, and P. Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 1703–1723. PMLR, 06–09 Jul 2018. URL http://proceedings.mlr.press/v75/zhang18a.html.

A Proving an $\Omega(\frac{\zeta}{\epsilon^2})$ lower bound for a class of algorithms including subgradient descent

This section provides a proof of Corollary 14 from Section 6.

A.1 Totally geodesic submanifolds and the hyperboloid model

Let $\mathcal{M}$ be a Hadamard manifold and $S \subset \mathcal{M}$ a Riemannian submanifold. If $x \in S$ and $v \in \mathrm{T}_x\mathcal{M}$, we say v is tangent to S if $v \in \mathrm{T}_xS$, and v is orthogonal to S if v is orthogonal to T_xS, i.e., $\langle v, w \rangle = 0$ for all $w \in \mathrm{T}_xS$. Below, parallel transport P^γ along a curve γ, and the exponential map $\exp_x$ both denote those operations *in the ambient space* $\mathcal{M}$.

Definition 18 *A Riemannian submanifold S is totally geodesic if it is connected, complete, and all geodesics in S are geodesics in the ambient manifold $\mathcal{M}$.*

We have the following equivalent characterizations, see (Lee, 2018, Prop. 8.12), (O'Neill, 1983, Ch. 4), (Lang, 1999, Ch. XI, XIV).

Lemma 19 *Let S be a complete, connected Riemannian submanifold of a Hadamard manifold $\mathcal{M}$. The following are equivalent.*

- *S is totally geodesic.*

- *The second fundamental form of S vanishes.*

- *If $(x, v) \in \mathrm{T}S$, then $\exp_x(tv) \in S$ for all $t \in \mathbb{R}$.*

- *If $x, y \in S$, then the geodesic through x, y is contained in S.*

Lemma 20 *Let S be a (complete, connected) totally geodesic submanifold of a Hadamard manifold $\mathcal{M}$. Then:*

- *If $x, y \in S$, then $\log_x(y) \in \mathrm{T}_x S$.*

- *Let $\gamma \colon \mathbb{R} \to S$ be a smooth curve in S with $\gamma(0) = x$, and $v \in T_x S$. Then the parallel transport of v along γ is the same in S and $\mathcal{M}$.*

- *Let $\gamma \colon \mathbb{R} \to S$ be a smooth curve in S with $\gamma(0) = x$, and let $v \in \mathrm{T}_x S$ and $w \in \mathrm{T}_x \mathcal{M}$ orthogonal to $\mathrm{T}_x S$. Then $P_{0 \to t}^{\gamma} v \in \mathrm{T}_{\gamma(t)} S$, and $P_{0 \to t}^{\gamma} w$ is orthogonal to $\mathrm{T}_{\gamma(t)} S$.*

Now let us consider $\mathcal{M} = \mathbb{H}^d$. To make our lower bound construction as concrete as possible, in the next section we shall give explicit formulae for certain geometric objects using the hyperboloid model of hyperbolic space. We introduce the relevant definitions and formulae in this section (see also (Ratcliffe, 2019, Ch. 3)).

Let $e_0, \ldots, e_d$ denote the standard basis vectors of $\mathbb{R}^{d+1}$. Let $J = \mathrm{diag}(-1, 1, \ldots, 1)$ be the $(d+1) \times (d+1)$ diagonal matrix whose diagonal consists of all ones, except its first entry is minus one. For $x, y \in \mathbb{R}^{d+1}$, let $\langle x, y \rangle = x^\top J y$. Define the submanifold $\mathcal{M} = \{x \in \mathbb{R}^{d+1} : \langle x, x \rangle = -1, \langle x, e_0 \rangle > 0\} \subset \mathbb{R}^{d+1}$, which has tangent spaces $\mathrm{T}_x \mathcal{M} = \{v \in \mathbb{R}^{d+1} : \langle x, v \rangle = 0\}$. Hyperbolic space $\mathbb{H}^d$ is *identified* with $\mathcal{M}$ endowed with the inner product $\langle u, v \rangle = u^\top J v$ on its tangent spaces.

We have $\mathrm{dist}(x, y) = \mathrm{arccosh}(-\langle x, y \rangle)$. For $x, y \in \mathcal{M}$ and $v \in \mathrm{T}_x \mathcal{M}$,

$$\exp_x(v) = \cosh(\|v\|)x + \frac{\sinh(\|v\|)}{\|v\|}v, \quad \log_x(y) = \frac{\mathrm{dist}(x, y)}{\sinh(\mathrm{dist}(x, y))}(y - \cosh(\mathrm{dist}(x, y))x).$$

If $v \in \mathrm{T}_x \mathcal{M}$, $y = \exp_x(v)$, and $u, w \in \mathrm{T}_x \mathcal{M}$ with u orthogonal to v and w parallel to v, then

$$P_{x \to y} w = \sinh(\|v\|)x + \cosh(\|v\|)w, \quad P_{x \to y} u = u. \tag{5}$$

Let us consider totally geodesic submanifolds of $\mathcal{M} = \mathbb{H}^d$. We have the following classical characterization in the hyperboloid model.

Lemma 21 *With $\mathcal{M} = \mathbb{H}^d$:*

- *If P is a subspace of $\mathbb{R}^{d+1}$ and $\mathcal{M} \cap P$ is nonempty, then $\mathcal{M} \cap P$ is a totally geodesic submanifold of $\mathcal{M}$. Moreover, $\mathrm{T}_x S = \mathrm{T}_x \mathcal{M} \cap P$ for all $x \in S$*

- *Conversely, if S is a k-dimensional totally geodesic submanifold of $\mathcal{M}$, then there is a unique $k+1$-dimensional subspace $P \subseteq \mathbb{R}^{d+1}$ such that $S = \mathcal{M} \cap P$.*

- *If $x \in \mathcal{M}$ and $\mathscr{S}$ is a k-dimensional linear subspace of $\mathrm{T}_x \mathcal{M}$, then $S = \exp_x(\mathscr{S}) = \mathcal{M} \cap \mathrm{span}(x, \mathscr{S})$ is a k-dimensional totally geodesic submanifold whose tangent space at x equals $\mathscr{S}$.*

- *A k-dimensional totally geodesic submanifold of $\mathbb{H}^d$ is isometric to $\mathbb{H}^k$.*

We define the following analogue of the Euclidean span. This object plays an important role in assumption A1 in the subsequent section.

Definition 22 *For $(x_1, v_1), \ldots, (x_m, v_m) \in \mathrm{T}\mathcal{M}$, define*

$$\mathrm{gspan}(x_1, \ldots, x_m, v_1, \ldots, v_m) = \mathcal{M} \cap \mathrm{span}(x_1, \ldots, x_m, v_1, \ldots, v_m).$$

If any of the points x_i are repeated, then we usually do not repeat them in the gspan. If any of the v_i equal zero, we usually omit them from the gspan.

Proposition 23 *Let $(x_1, v_1), \ldots, (x_m, v_m) \in \mathrm{T}\mathcal{M}$. Let $S = \mathrm{gspan}(x_1, \ldots, x_m, v_1, \ldots, v_m)$. Then S is the minimal totally geodesic submanifold of $\mathcal{M}$ which passes through each x_i and is tangent to each v_i, i.e.,*

$$(x_i, v_i) \in \mathrm{T}S \quad \text{for all } i = 1, \ldots, m. \tag{6}$$

By "S is minimal," we mean that for any totally geodesic submanifold S' satisfying (6), then we must have $S \subset S'$. This proposition tells us that the gspan is independent of the model we use to represent hyperbolic space.

Proof [Proof of Proposition 23] S is a totally geodesic submanifold. Moreover, it is clear that S satisfies (6) because $x_i \in S$ and $v_i \in \mathrm{T}_{x_i}\mathcal{M} \cap \mathrm{span}(x_1, \ldots, x_m, v_1, \ldots, v_m) = \mathrm{T}_{x_i}S$.

Consider any totally geodesic submanifold S' satisfying (6). We know $S' = \mathcal{M} \cap P'$ for some subspace P' of $\mathbb{R}^{d+1}$. Moreover, for each $i = 1, \ldots, m$, there is a curve $\gamma_i \subset S' \subset P'$ with $\gamma_i(0) = x_i, \gamma_i'(0) = v_i$. Therefore $x_i \in P'$ and $v_i \in \mathrm{T}_{x_i}S' = P' \cap \mathrm{T}_{x_i}\mathcal{M} \subset P'$. So $\mathrm{span}(x_1, \ldots, x_m, v_1, \ldots, v_m) \subset P'$. Therefore $S \subset S'$. ∎

A.2 An $\Omega(\frac{\zeta}{\epsilon^2})$ lower bound for a class of algorithms

In this section, we prove an $\Omega(\frac{\zeta}{\epsilon^2})$ lower bound for a class of algorithms (satisfying assumptions A1 and A2 below) solving problem **P2**.

Let $\epsilon \leq \frac{1}{4\sqrt{2}}$. Let $T = d = \lfloor \frac{1}{32}\frac{\zeta_r}{\epsilon^2} \rfloor$. Fix $\theta \in (0, \frac{\pi}{2})$ so that $\epsilon = \frac{1}{4}\cos(\theta)$ (and so $\cos^2(\theta) \leq \frac{1}{2}$). Without loss of generality, we consider functions with Lipschitz constant $M = \frac{2}{\cos(\theta)}$. We divide the proof into two parts: (I) the **geometric setup**, and (II) the actual **lower bound proof**.

(I) Geometric setup: We inductively define a series of geometric objects.

Let us first initialize ($k = 0$). Let $y_0 = x_{\mathrm{ref}}$ and $I_0 = \{y_0\}$. Let $S_0 = \partial B(y_0, r)$ (S_0 is a hyperbolic sphere of dimension $d - 1$). S_0 is the boundary of the ball $B_0 = B(y_0, r)$. Let r_0 be the radius of S_0, i.e., $r_0 = r$. S_0 and B_0 are contained in the d-dimensional totally geodesic submanifold $H_0 = \mathcal{M}$. Choose an orthonormal basis $e_1, \ldots, e_d$ for $\mathrm{T}_{y_0}\mathcal{M}$. Define $e_i^{(0)} = e_i$ for $i = 1, \ldots d$. *Working in the hyperboloid model*, we take $x_{\mathrm{ref}} = e_0$, and let the orthonormal basis for $\mathrm{T}_{y_0}\mathcal{M}$ consist of the coordinate basis vectors, also denoted $e_1, \ldots, e_d$.

Next, for $k = 1, \ldots, d - 1$:

- Define $\Delta_{k-1} > 0$ and $r_k > 0$ by

$$\cos(\theta) = \frac{\tanh(\Delta_{k-1})}{\tanh(r_{k-1})}, \;\; \sin(\theta) = \frac{\sinh(r_k)}{\sinh(r_{k-1})}, \;\; \cosh(r_{k-1}) = \cosh(r_k)\cosh(\Delta_{k-1}). \tag{7}$$

(The third equation is a consequence of the first two equations.)

Lemma 24 *The sequence $r_k > 0$ given by $r_0 = r, \sin(\theta) = \frac{\sinh(r_k)}{\sinh(r_{k-1})}$ satisfies $r_k \geq \frac{r}{2}$ for all $k \leq T = d = \lfloor \frac{1}{32}\frac{\zeta_r}{\epsilon^2} \rfloor = \lfloor \frac{\zeta_r}{2\cos^2(\theta)} \rfloor$.*

Proof Indeed, $\sinh(r_k) = \sin(\theta)^k \sinh(r) \geq \sinh(r/2)$ provided

$$k \leq \frac{\log\left(\frac{\sinh(r/2)}{\sinh(r)}\right)}{\log(\sin(\theta))} = 2\frac{\log\left(\frac{\sinh(r/2)}{\sinh(r)}\right)}{\log(1 - \cos^2(\theta))}.$$

We know $2\frac{\log\left(\frac{\sinh(r/2)}{\sinh(r)}\right)}{\log(1-x)} \geq \frac{\zeta_r}{2x}$ for all $x \in (0, 1/2]$ and $r > 0$. Therefore, since $k \leq \frac{\zeta_r}{2\cos^2(\theta)}$ and $\cos^2(\theta) \leq 1/2$, we conclude $r_k \geq r/2$. ∎

- Define $\tilde{g}_{k-1} = -\frac{1}{\cos(\theta)}e_k^{(k-1)}$. Define $y_k = \exp_{y_{k-1}}(\Delta_{k-1}e_k^{(k-1)}) = \exp_{y_{k-1}}(-\Delta_{k-1}\frac{\tilde{g}_{k-1}}{\|\tilde{g}_{k-1}\|})$. Define $e_i^{(k)} = P_{y_{k-1}\to y_k}e_i^{(k-1)}$ for all $i = 1, \ldots, d$.

  In the hyperboloid model, observe that $e_i^{(k)} = e_i$ for all $i \geq k+1$, and $e_i^{(k)} \in \mathrm{span}(e_0, \ldots, e_k)$ for all $1 \leq i \leq k$ (both of which can be verified by induction and formula (5)). Moreover,

  $$y_k = \cosh(\Delta_{k-1})y_{k-1} + \sinh(\Delta_{k-1})e_k, \quad e_k^{(k)} = \sinh(\Delta_{k-1})y_{k-1} + \cosh(\Delta_{k-1})e_k. \quad (8)$$

  In particular, $y_k \in \mathcal{M} \cap \mathrm{span}(e_0, e_1, \ldots, e_k)$.

- Define the k-dimensional totally geodesic submanifold $I_k = \mathrm{gspan}(y_0, \ldots, y_k)$. Note that

  $$I_k = \mathcal{M} \cap \mathrm{span}(e_0, e_1, \ldots, e_k) = \mathrm{gspan}(y_0, e_1, \ldots, e_k) = \mathrm{gspan}(y_0, \ldots, y_{k-1}, e_k^{(k-1)}).$$

  Using Proposition 23 and that $e_i^{(k)} \in \mathrm{span}(e_0, \ldots, e_k)$ for all $1 \leq i \leq k$, we determine that $\mathrm{T}_{y_k}I_k = \mathrm{span}(e_1^{(k)}, \ldots, e_k^{(k)})$, and $I_k = \mathrm{gspan}(y_k, e_1^{(k)}, \ldots, e_k^{(k)})$. Lastly, note that $I_0 \subset I_1 \subset, \ldots$.

- Define $H_k = \{x \in \mathcal{M} : \langle x, e_i^{(i)}\rangle = 0, \forall 1 \leq i \leq k\}$. As an intersection of $\mathcal{M}$ and a subspace, H_k is a totally geodesic submanifold. Note that $H_0 \supset H_1 \supset \ldots$.

Lemma 25 *H_k is a $d-k$-dimensional totally geodesic submanifold which (a) passes through y_k, (b) is orthogonal to $e_i^{(k)}$ for $i = 1, \ldots, k$, and (c) is tangent to $e_i^{(k)}$ for $i = k+1, \ldots, d$. In particular, $H_k = \mathrm{gspan}(y_k, e_{k+1}^{(k)}, \ldots, e_d^{(k)}) = \mathcal{M} \cap \mathrm{span}(y_k, e_{k+1}, \ldots, e_d)$.*

Proof We can prove (a) and (b) by induction on k. The base case $k = 0$ is clear. By the inductive hypothesis $y_{k-1} \in H_{k-1}$ and $e_k^{(k-1)} \in \mathrm{T}_{y_{k-1}}H_{k-1}$, so we have that $y_k \in H_{k-1}$ (H_{k-1} is totally geodesic). Equation (8) implies $y_k \in \{x \in \mathcal{M} : \langle x, e_k^{(k)}\rangle = 0\}$. Therefore $y_k \in H_k$.

Again by the inductive hypothesis, H_{k-1} is orthogonal to $e_i^{(k-1)}$ for all $i \leq k - 1$. Since $y_k \in H_{k-1}$, H_{k-1} is orthogonal to $e_i^{(k)} = P_{y_{k-1}\to y_k}e_i^{(k-1)}$ for all $i \leq k - 1$ (Lemma 20). As $H_k \subset H_{k-1}$, H_k is orthogonal to $e_i^{(k)}$ for all $i \leq k - 1$. By definition, H_k is orthogonal to $e_k^{(k)}$, and so we conclude H_k is orthogonal to $e_i^{(k)}$ for all $i \leq k$.

For (c): by Lemma 21, we know that

$$\mathrm{T}_{y_k}H_k = \{u \in \mathbb{R}^{d+1} : \langle y_k, u\rangle = 0, \langle e_i^{(i)}, u\rangle = 0 \,\forall 1 \leq i \leq k\}.$$

Since $y_k \in \mathrm{span}(e_0, \ldots, e_k)$ and $e_i^{(i)} \in \mathrm{span}(e_0, \ldots, e_i)$, we conclude that $e_i = e_i^{(k)} \in \mathrm{T}_{y_k} H_k$ for $i \geq k + 1$. Using (b), $\mathrm{T}_{y_k} H_k = \mathrm{span}(e_{k+1}^{(k)}, \ldots, e_d^{(k)})$, and so H_k has dimension $d - k$ and $H_k = \exp_{y_k}(\mathrm{span}(e_{k+1}^{(k)}, \ldots, e_d^{(k)})) = \mathcal{M} \cap \mathrm{span}(y_k, e_{k+1}, \ldots, e_d)$ (Lemma 21). ∎

- Let S_k be the set of all points $x^* \in S_{k-1}$ such that the angle between $e_k^{(k-1)}$ and $\log_{y_{k-1}}(x^*)$ equals θ. Note that $S_0 \supset S_1 \supset \ldots$.

Lemma 26 *(a) S_k is the hyperbolic sphere of dimension $d - k - 1$ contained in H_k which has center y_k and radius r_k.*

(b) If $x^* \in S_k$, then $\log_{y_k}(x^*)$ is orthogonal to I_k. In particular, y_k is the closest point in I_k to x^*.

Proof We can verify (a) by induction using the hyperboloid model. It is clearly true for $k = 0$. By the inductive hypothesis, $\mathrm{dist}(x^*, y_{k-1}) = r_{k-1}$ for all $x^* \in S_{k-1}$. Therefore,

$$
\begin{aligned}
S_k &= \{x^* \in S_{k-1} : r_{k-1}\cos(\theta) = \langle \log_{y_{k-1}}(x^*), e_k^{(k-1)}\rangle = \langle \log_{y_{k-1}}(x^*), e_k\rangle\} \\
&= \{x^* \in S_{k-1} : r_{k-1}\cos(\theta) = \frac{r_{k-1}}{\sinh(r_{k-1})}\langle x^*, e_k\rangle\}.
\end{aligned}
\tag{9}
$$

So using equation (8) and that $\cosh(r_{k-1}) = -\langle x^*, y_{k-1}\rangle$, if $x^* \in S_k$ then

$$
\langle x^*, e_k^{(k)}\rangle = -\cosh(r_{k-1})\sinh(\Delta_{k-1}) + \sinh(r_{k-1})\cos(\theta)\cosh(\Delta_{k-1}) = 0,
$$

where the last equality follows from equation (7). We conclude that S_k is contained in $\{x \in \mathcal{M} : \langle x, e_k^{(k)}\rangle = 0\}$. Because $S_k \subset S_{k-1} \subset H_{k-1}$, we conclude

$$
S_k \subset H_{k-1} \cap \{x \in \mathcal{M} : \langle x, e_k^{(k)}\rangle = 0\} = \{x \in \mathcal{M} : \langle x, e_i^{(i)}\rangle = 0 \; \forall i \leq k\} = H_k.
$$

A calculation using equations (7), (8) and (9) shows that $\mathrm{dist}(y_k, x^*) = r_k$ for all $x^* \in S_k$. So we have shown S_k is a subset of the sphere of radius r_k centered at y_k in H_k.

By the inductive hypothesis, S_{k-1} is a hyperbolic sphere in H_{k-1} with center y_{k-1} and radius r_{k-1}. Now let x^* be a point in H_k with $\mathrm{dist}(x^*, y_k) = r_k$. By Lemma 25, H_k contains y_k and is orthogonal to $e_k^{(k)}$. Since H_k is totally geodesic, $\log_{y_k}(x^*)$ is orthogonal to $e_k^{(k)}$. Therefore, the hyperbolic triangle x^*, y_k, y_{k-1} has a right angle at y_k. Using $\mathrm{dist}(x^*, y_k) = r_k$, $\mathrm{dist}(y_k, y_{k-1}) = \Delta_{k-1}$, equations (7) and hyperbolic trigonometry, we find that $\mathrm{dist}(y_{k-1}, x^*) = r_{k-1}$ and the angle between $e_k^{(k-1)}$ and $\log_{y_{k-1}}(x^*)$ equals θ. Therefore, $x^* \in S_k$. We conclude that S_k contains (and so equals) the sphere of radius r_k centered at y_k in H_k. We have shown (a).

For (b): we know that $y_k \in H_k$ by Lemma 25 and $x^* \in S_k \subset H_k$. Since H_k is totally geodesic, $\log_{y_k}(x^*) \in \mathrm{T}_{y_k} H_k$. However, Lemma 25 implies $\log_{y_k}(x^*) \in \mathrm{T}_{y_k} H_k$ is orthogonal to $\mathrm{span}(e_1^{(k)}, \ldots, e_k^{(k)}) = \mathrm{T}_{y_k} I_k$. ∎

- Let $B_k \subset H_k$ be the $d - k$-dimensional hyperbolic ball whose boundary is S_k. Since $B_0 \supset B_1 \supset \ldots$, we find that $y_\ell \in B_k$ for all $\ell \geq k$.

For $k = 0, \ldots, d - 2$, consider the following additional geometric objects:

- For each $x^* \in S_{k+1}$, define the tangent vector V_{k,x^*} and corresponding half-space L_{k,x^*}:

$$V_{k,x^*} = -\tilde{g}_k - \frac{\log_{y_k}(x^*)}{\text{dist}(y_k, x^*)}, \qquad L_{k,x^*} = \exp_{y_k}(\{v \in \mathrm{T}_{y_k}\mathcal{M} : \langle V_{k,x^*}, v \rangle \geq 0\}).$$

- Define $C_k = \bigcap_{x^* \in S_{k+1}} L_{k,x^*}$. As an intersection of (g-convex) half-spaces, C_k is also g-convex.

Lemma 27 *We have the following:*

- *For $x^* \in S_{k+1}$, $\partial L_{k,x^*}$ is a $(d - 1)$-dimensional totally geodesic submanifold containing x^* and I_k.*

- *For $x^* \in S_{k+1}$, $y_{k+1} \in B_{k+1} \subset C_k \subset L_{k,x^*}$ for all $k \in \{0, \ldots, d - 2\}$.*

- *Let $x^* \in S_{d-1}$. Then $y_0, \ldots, y_{d-1} \in C_k \subset L_{k,x^*}$ for all $k \leq d - 2$.*

Proof *For the first bullet*: if $x^* \in S_{k+1}$, we know $\langle \log_{y_k}(x^*), e_{k+1}^{(k)} \rangle = \text{dist}(y_k, x^*) \cos(\theta)$, and so

$$\langle \frac{1}{\cos(\theta)} e_{k+1}^{(k)} - \frac{\log_{y_k}(x^*)}{\text{dist}(y_k, x^*)}, \log_{y_k}(x^*) \rangle = \text{dist}(y_k, x^*) - \frac{\text{dist}(y_k, x^*)^2}{\text{dist}(y_k, x^*)} = 0.$$

That is, $\partial L_{k,x^*}$ contains x^*. Using that $e_{k+1}^{(k)} = e_{k+1}$ and $y_k \in I_k = \mathcal{M} \cap \text{span}(e_0, \ldots, e_k)$ in the hyperboloid model, we see that $\langle \frac{1}{\cos(\theta)} e_{k+1}^{(k)} - \frac{\log_{y_k}(x^*)}{\text{dist}(y_k, x^*)}, \log_{y_k}(x) \rangle = 0$ for all $x \in I_k$. Therefore, $\partial L_{k,x^*}$ also contains I_k.

For the second bullet: we know angle between $e_{k+1}^{(k)}$ and $\log_{y_k}(x^*)$ equals θ, and so the angle between $e_{k+1}^{(k)}$ and V_{k,x^*} equals $\frac{\pi}{2} - \theta$ (V_{k,x^*} is a linear combination of $e_{k+1}^{(k)}$ and $\log_{y_k}(x^*)$). Therefore, $y_{k+1} \in L_{k,x^*}$, and since $x^* \in S_{k+1}$ was arbitrary we have $y_{k+1} \in C_k$.

We know V_{k,x^*} is orthogonal to $\partial L_{k,x^*}$. Since $y_k, x^* \in \partial L_{k,x^*}$ and $\partial L_{k,x^*}$ is totally geodesic (by the first bullet), we know that $P_{y_k \to x^*} V_{k,x^*}$ is orthogonal to $\partial L_{k,x^*}$ (Lemma 20). Therefore, $L_{k,x^*} = \exp_{x^*}(\{u \in \mathrm{T}_{x^*}\mathcal{M} : \langle P_{y_k \to x^*} V_{k,x^*}, u \rangle \geq 0\})$. Using that $H_{k+1} = \exp_{x^*}(\mathrm{T}_{x^*} H_{k+1})$ since H_{k+1} is totally geodesic, we find

$$H_{k+1} \cap L_{k,x^*} = \exp_{x^*}(\{u \in \mathrm{T}_{x^*} H_{k+1} : \langle \mathcal{P}_{k+1} P_{y_k \to x^*} V_{k,x^*}, u \rangle \geq 0\}), \tag{10}$$

where $\mathcal{P}_{k+1}$ denotes orthogonal projection onto $\mathrm{T}_{x^*} H_{k+1}$. Therefore, $H_{k+1} \cap L_{k,x^*}$ is a half-space in H_{k+1}. We seek to compute $\mathcal{P}_{k+1} P_{y_k \to x^*} V_{k,x^*}$. We can do this using two-dimensional hyperbolic geometry.

Define the two-dimensional totally geodesic submanifold

$$\tilde{H} = \text{gspan}(y_k, y_{k+1}, x^*) = \text{gspan}(y_k, e_{k+1}^{(k)}, \log_{y_k}(x^*)) = \text{gspan}(y_{k+1}, e_{k+1}^{(k)}, \log_{y_{k+1}}(x^*)).$$

By Lemma 25, $e_{k+1}^{(k+1)}$ is orthogonal to H_k, and so in particular to $\log_{y_{k+1}}(x^*)$ as well. Therefore, $e_{k+1}^{(k+1)}$ and $\log_{y_{k+1}}(x^*)$ form an orthogonal basis for $\mathrm{T}_{y_{k+1}}\tilde{H}$, and so $P_{y_{k+1}\to x^*}e_{k+1}^{(k+1)}$ and $\log_{x^*}(y_{k+1})$ form an orthogonal basis for $\mathrm{T}_{x^*}\tilde{H}$ (Lemma 20). By its definition, V_{k,x^*} is in

$$\mathrm{span}(e_{k+1}^{(k)}, \log_{y_k}(x^*)) = \mathrm{T}_{y_k}\tilde{H}$$

and is orthogonal to $\log_{y_k}(x^*)$. Therefore, $P_{y_k\to x^*}V_{k,x^*}$ is in

$$\mathrm{T}_{x^*}\tilde{H} = \mathrm{span}(P_{y_{k+1}\to x^*}e_{k+1}^{(k+1)}, \log_{x^*}(y_{k+1}))$$

and is orthogonal to $\log_{x^*}(y_k)$. Let $\alpha \in (0, \frac{\pi}{2})$ be the angle between $\log_{x^*}(y_k)$ and $\log_{x^*}(y_{k+1})$ (y_k, y_{k+1}, x^* is a right triangle). We conclude the angle between $P_{y_k\to x^*}V_{k,x^*}$ and $\log_{x^*}(y_k)$ is $\frac{\pi}{2} - \alpha$, and so we can decompose $P_{y_k\to x^*}V_{k,x^*}$ in the orthogonal basis for $\mathrm{T}_{x^*}\tilde{H}$:

$$P_{y_k\to x^*}V_{k,x^*} = \|P_{y_k\to x^*}V_{k,x^*}\| \left[\sin\left(\frac{\pi}{2} - \alpha\right) \frac{P_{y_{k+1}\to x^*}e_{k+1}^{(k+1)}}{\|P_{y_{k+1}\to x^*}e_{k+1}^{(k+1)}\|} + \cos\left(\frac{\pi}{2} - \alpha\right) \frac{\log_{x^*}(y_{k+1})}{\|\log_{x^*}(y_{k+1})\|} \right].$$

On the other hand, $P_{y_{k+1}\to x^*}e_{k+1}^{(k+1)}$ is orthogonal to H_{k+1} and $\log_{x^*}(y_{k+1})$ is tangent to H_{k+1}, since H_{k+1} is totally geodesic, is orthogonal to $e_{k+1}^{(k+1)}$, and contains x^* and y_{k+1}. We conclude that

$$\mathcal{P}_{k+1}P_{y_k\to x^*}V_{k,x^*} = \|P_{y_k\to x^*}V_{k,x^*}\| \left[\cos\left(\frac{\pi}{2} - \alpha\right) \frac{\log_{x^*}(y_{k+1})}{\|\log_{x^*}(y_{k+1})\|} \right].$$

Using equation (10), we get

$$H_{k+1} \cap L_{k,x^*} = \exp_{x^*}(\{u \in \mathrm{T}_{x^*}H_{k+1} : \langle\log_{x^*}(y_{k+1}), u\rangle \geq 0\}),$$

Since $B_{k+1} \subset H_{k+1}$ is g-convex, has center y_{k+1} and contains x^* on its boundary (S_{k+1}), we conclude that $B_{k+1} \subset H_{k+1} \cap L_{k,x^*} \subset L_{k,x^*}$. Since $x^* \in S_{k+1}$ was arbitrary, $B_{k+1} \subset C_k$.

For the third bullet: for all $k = 0, \ldots, d-2$ we know that $x^* \in C_k$ by the second bullet, and $I_k \subset C_k$ by the first bullet. In particular, $x^*, y_0, \ldots, y_k \in C_k$ for all $k = 0, \ldots, d-2$. Since $y_\ell \in B_\ell \subset B_{k+1}$ for all $\ell \geq k+1$, we conclude (using the second bullet) that $y_0, \ldots, y_{d-1} \in C_k$ for all $k \leq d-2$. ∎

Lemma 28 *Let $k \in \{0, \ldots, d-2\}$, $x^* \in S_{k+1}$ and $y \in I_k$. The g-convex function $x \mapsto \frac{1}{\cos(\theta)}\mathrm{dist}(x, L_{k,x^*})$ has a subgradient $\hat{g}$ at y such that $-\hat{g} + \frac{\log_y(x^*)}{\mathrm{dist}(y,x^*)}$ is tangent to I_{k+1}. If $y = y_k$, then we can take $\hat{g}$ such that $-\hat{g} + \frac{\log_{y_k}(x^*)}{\mathrm{dist}(y_k,x^*)} = \frac{1}{\cos(\theta)}e_{k+1}^{(k)} = -\tilde{g}_k$.*

Proof First, let us consider the particular case $y = y_k$. The subdifferential of $x \mapsto \frac{1}{\cos(\theta)}\mathrm{dist}(x, L_{k,x^*})$ at y_k contains all vectors of the form $t\frac{V_{k,x^*}}{\|V_{k,x^*}\|}$ with $t \in [-\frac{1}{\cos(\theta)}, 0]$. Take $\hat{g} = -\tan(\theta)\frac{V_{k,x^*}}{\|V_{k,x^*}\|} = -\tan(\theta)\frac{V_{k,x^*}}{\sqrt{\frac{1}{\cos^2(\theta)}-1}} = -V_{k,x^*} = \tilde{g}_k + \frac{\log_{y_k}(x^*)}{\mathrm{dist}(y_k,x^*)}$.

23

Now consider the general case $y \in I_k$. By Lemma 26(b), $\partial L_{k,x^*}$ contains y and y_k. Therefore, $P_{y_k \to y} V_{k,x^*}$ is orthogonal to $\partial L_{k,x^*}$ (Lemma 20), and the subdifferential of $x \mapsto \frac{1}{\cos(\theta)} \mathrm{dist}(x, L_{k,x^*})$ at y contains all vectors of the form $t P_{y_k \to y} \frac{V_{k,x^*}}{\|V_{k,x^*}\|}$ with $t \in [-\frac{1}{\cos(\theta)}, 0]$.

Let $\theta_y \in (0, \frac{\pi}{2})$ be the angle between $\log_y(y_{k+1})$ and $\log_y(x^*)$. By Lemma 27, y_k is the closest point in I_k to x^*, so $\mathrm{dist}(y, x^*) \geq \mathrm{dist}(y_k, x^*)$. Therefore, comparing the two right triangles y_k, y_{k+1}, x^* and y, y_{k+1}, x^* (the triangles are right by part (b) of Lemma 26), we determine $\theta_y \leq \theta$. In particular, $\hat{g} = -\frac{\sin(\theta_y)}{\cos(\theta)} P_{y_k \to y} \frac{V_{k,x^*}}{\|V_{k,x^*}\|}$ is a subgradient at y.

Consider the two-dimensional totally geodesic submanifold $\tilde{H} = \mathrm{gspan}(y, y_{k+1}, x^*)$. We know that $\log_{y_{k+1}}(x^*)$ is tangent to $\tilde{H}$ and orthogonal to $\log_{y_{k+1}}(y)$. Therefore, $P_{y_{k+1} \to y} \log_{y_{k+1}}(x^*)$ is tangent to $\tilde{H}$ and orthogonal to $\log_y(y_{k+1})$. That is, $P_{y_{k+1} \to y} \log_{y_{k+1}}(x^*)$ and $\log_y(y_{k+1})$ form an orthogonal basis for $\mathrm{T}_y \tilde{H}$. However, $\log_y(x^*)$ is contained in $\mathrm{T}_y \tilde{H}$, and so we can decompose in this orthogonal basis:

$$\frac{\log_y(x^*)}{\mathrm{dist}(y, x^*)} = \cos(\theta_y) \frac{\log_y(y_{k+1})}{\mathrm{dist}(y, y_{k+1})} + \sin(\theta_y) \frac{P_{y_{k+1} \to y} \log_{y_{k+1}}(x^*)}{\|P_{y_{k+1} \to y} \log_{y_{k+1}}(x^*)\|}. \tag{11}$$

Taking $y = y_k$ and using $\frac{\log_{y_k}(y_{k+1})}{\mathrm{dist}(y_k, y_{k+1})} = e_{k+1}^{(k)}$,

$$\frac{\log_{y_k}(x^*)}{\mathrm{dist}(y_k, x^*)} = \cos(\theta) e_{k+1}^{(k)} + \sin(\theta) \frac{P_{y_{k+1} \to y_k} \log_{y_{k+1}}(x^*)}{\|P_{y_{k+1} \to y_k} \log_{y_{k+1}}(x^*)\|}. \tag{12}$$

Recalling that $\log_{y_{k+1}}(x^*)$ is orthogonal to I_{k+1}, and using formula (5) and Lemma 20,

$$P_{y_{k+1} \to y} \log_{y_{k+1}}(x^*) = P_{y_k \to y} P_{y_{k+1} \to y_k} \log_{y_{k+1}}(x^*). \tag{13}$$

Combining the definition of V_{k,x^*}, equations (12) and (13), and $\|V_{k,x^*}\| = \tan(\theta)$, we find

$$\hat{g} = -\frac{\sin(\theta_y)}{\cos(\theta)} \frac{1}{\|V_{k,x^*}\|} P_{y_k \to y} V_{k,x^*} = -\frac{\sin(\theta_y)}{\cos(\theta)} \frac{1}{\|V_{k,x^*}\|} P_{y_k \to y} \left[\frac{1}{\cos(\theta)} e_{k+1}^{(k)} - \frac{\log_{y_k}(x^*)}{\mathrm{dist}(y_k, x^*)} \right]$$

$$= c_1 P_{y_k \to y} e_{k+1}^{(k)} + \frac{\sin(\theta_y)}{\cos(\theta)} \frac{1}{\|V_{k,x^*}\|} \sin(\theta) P_{y_k \to y} \frac{P_{y_{k+1} \to y_k} \log_{y_{k+1}}(x^*)}{\|P_{y_{k+1} \to y_k} \log_{y_{k+1}}(x^*)\|}$$

$$= c_1 P_{y_k \to y} e_{k+1}^{(k)} + \sin(\theta_y) P_{y_{k+1} \to y} \frac{\log_{y_{k+1}}(x^*)}{\|P_{y_{k+1} \to y_k} \log_{y_{k+1}}(x^*)\|}$$

for some constant c_1 which we are omitting. Using equation (11), we find

$$-\hat{g} + \frac{\log_y(x^*)}{\mathrm{dist}(y, x^*)} = -c_1 P_{y_k \to y} e_{k+1}^{(k)} + \cos(\theta_y) \frac{\log_y(y_{k+1})}{\mathrm{dist}(y, y_{k+1})}$$

which is indeed tangent to I_{k+1} ∎

(II) The lower bound: We know that S_{d-1} is a 0-dimensional hyperbolic sphere of radius $r_{d-1} \geq r/2$ (i.e., a set consisting of two points which are at least $2 \cdot r/2 = r$ apart). From now on we *fix* an $x^* \in S_{d-1}$. Consider the following g-convex, M-Lipschitz function:

$$f(x) = \mathrm{dist}(x, x^*) + \max_{i=0,\dots,d-2} \left\{ \frac{1}{\cos(\theta)} \mathrm{dist}(x, L_{i,x^*}) \right\}. \tag{14}$$

We know $x^* \in S_{d-1} \subset S_k$ for all $k \in \{0, \ldots, d-1\}$ and $S_{k+1} \subset L_{k,x^*}$ for all $k \in \{0, \ldots, d-2\}$ (Lemma 27). Therefore, $x^* \in L_{k,x^*}$ for all $k \in \{0, \ldots, d-2\}$, and we conclude $f^* = f(x^*) = \mathrm{dist}(x^*, x^*) + 0 = 0$.

Let $\mathcal{A}$ be a deterministic first-order algorithm. We make the following mild assumption on $\mathcal{A}$.

A 1 *The algorithm $\mathcal{A}$ chooses its queries x_k in the* gspan *of past queries and subgradients. That is, $x_0 = x_{\mathrm{ref}}$, and $x_k \in \mathrm{gspan}(x_0, \ldots, x_{k-1}, g_0, \ldots, g_{k-1})$ for all $k \geq 1$.*

The assumption A1 is mild, and can probably be removed using similar techniques as in the Euclidean case (Nemirovskii and Yudin, 1983, Ch. 7). All algorithms we are familiar with and which use the exponential map $\exp$ as the retraction (and parallel transport or the differential of $\exp$ as the transporter) satisfy this assumption.

A 2 *When running on the oracle for f defined below, for each $k \leq d-1$ the algorithm $\mathcal{A}$ queries $x_k \in \cap_{\ell=0}^{k-1} L_{\ell,x^*}$.*

The assumption A2 is *more restrictive*, and it is an open question how it can be removed. We verify it for a reasonable algorithm in the next section.

We know that $\mathcal{A}$ initially queries $x_0 = x_{\mathrm{ref}} = y_0 \in I_0$ by assumption A1, that $x_0 = y_0 \in L_{\ell,x^*}$ for all $\ell \in \{0, \ldots, d-2\}$ (Lemma 27), and $f(x_0) - f^* = r_0 = r \geq r/2$. We show by induction that for each $k = 1, \ldots, d-1$, (a) there is a subgradient $g_{k-1} \in \partial f(x_{k-1})$ which is tangent to I_k, (b) the algorithm $\mathcal{A}$ produces the next query x_k in I_k, (c) $x_k \in L_{\ell,x^*}$ for all $\ell \in \{0, \ldots, d-2\}$, and (d) $f(x_k) - f^* = r_k \geq \frac{r}{2}$.

Indeed, given $x_{k-1} \in I_{k-1}$, Lemma 28 implies $x \mapsto \frac{1}{\cos(\theta)} \mathrm{dist}(x, L_{k-1,x^*})$ has a subgradient $\hat{g}_{k-1}$ such that $g_{k-1} = \hat{g}_{k-1} - \frac{\log_{x_{k-1}}(x^*)}{\mathrm{dist}(x_{k-1},x^*)}$ is tangent to I_k. By the inductive hypothesis we know $x_{k-1} \in L_{\ell,x^*}$ for all $\ell \in \{0, \ldots, d-2\}$, and so g_{k-1} is a subgradient of f at x_{k-1} which is tangent to I_k. Giving this subgradient to the algorithm, the next query x_k lies in $\mathrm{gspan}(x_0, \ldots, x_{k-1}, g_0, \ldots, g_{k-1}) = I_k$ by assumption A1. This shows (a) and (b).

By assumption A2, we know that $x_k \in L_{\ell,x^*}$ for $\ell < k$. By Lemma 27, $I_\ell \subset \partial L_{\ell,x^*}$ for all ℓ, so in particular $x_k \in I_k \subset I_\ell \subset L_{\ell,x^*}$ for all $\ell \geq k$. From this, we conclude (c). Since $x_k \in L_{\ell,x^*}$ for all $\ell \in \{0, \ldots, d-2\}$, $f(x_k) = \mathrm{dist}(x_k, x^*) \geq \mathrm{dist}(y_k, x^*) = r_k \geq r/2$ by Lemmas 24 and 26. This shows (d), and concludes the induction. We showed that for all $k = 0, \ldots, d-1$, $f(x_k) - f^* \geq \frac{r}{2} = \frac{1}{4} \cos(\theta) \frac{2r}{\cos(\theta)} = \epsilon M r$. That is, we have proven:

Theorem 29 *Let $\epsilon \leq \frac{1}{4\sqrt{2}}$. Let $T = d = \lfloor \frac{1}{32} \frac{\zeta_r}{\epsilon^2} \rfloor$. Let $f \colon \mathbb{H}^d \to \mathbb{R}$ be the globally g-convex and M-Lipschitz function given by equation (14). For any algorithm $\mathcal{A}$ satisfying assumptions A1 and A2, there is an oracle for f so that running $\mathcal{A}$ with that oracle produces iterates $x_0, x_1, \ldots$ satisfying $f(x_k) - f^* \geq \epsilon M r$ for all $k = 0, \ldots, T-1$.*

Remark 30 *Let us consider the role of assumption A2, and how strong that assumption is.*

Observe that if $x_k \in I_k$ then $f(x_k) - f^* \geq r_k \geq \frac{r}{2}$ by Lemma 26. So assumption A2 is not needed to guarantee that the function gap is large. However, if $x_k \in I_k$ is not in $\cap_{\ell=0}^{k-1} L_{\ell,x^*}$, then it is possible that there is no subgradient in $\partial f(x_k)$ which is tangent to I_{k+1}. Hence, a subgradient at $x_k \notin \cap_{\ell=0}^{k-1} L_{\ell,x^*}$ might reveal too much information about x^*. Assumption A2 helps to guarantee that $\partial f(x_k)$ contains a subgradient which is tangent to I_{k+1}.

Consider an algorithm satisfying assumptions A1 and A2. Then the algorithm queries x_1 in the geodesic $I_1 = \{\exp_{x_0}(-tg_0) : t \in \mathbb{R}\}$. Assumption A2 further restricts that x_1 is in $\{\exp_{x_0}(-tg_0) : t \geq 0\}$, which seems to be a reasonable restriction (it is intuitively unclear what advantage in general an algorithm could have by querying outside this region). In general, we know that the algorithm queries x_k in $Q_k := I_k \cap \bigcap_{\ell=0}^{k-1} L_{k,x^*}$ which is a (g-convex) subset of

$$\tilde{Q}_k := I_k \cap \bigcap_{\ell=0}^{k-1} \exp_{x_\ell}(\{v \in \mathrm{T}_{x_\ell}\mathcal{M} : \langle -g_\ell, u \rangle \geq 0\}).$$

Intuitively, Q_k seems to occupy most of $\tilde{Q}_k$. So the assumption $x_k \in Q_k$ also seems relatively mild.

Remark 31 *Under the assumption that $x_k \in \cap_{\ell=0}^{k-1} C_k$ (in place of A2), it is not hard to see that the following alternative function also works in place of the function f defined in equation (14):*

$$f(x) = \mathrm{dist}(x, x^*) + \max_{i=0,\ldots,d-2} \left\{ \frac{1}{\cos(\theta)} \mathrm{dist}(x, C_k) \right\}$$

This function has the added benefit that it is more symmetric than f from equation (14).

A.3 An $\Omega(\frac{\zeta}{\epsilon^2})$ lower bound for subgradient descent with Polyak step size

Our goal is to prove Corollary 14. To this end, let us first describe and analyze subgradient descent with Polyak step size for arbitrary f in $\mathcal{F}_{r,M,0,\infty}$. It takes the form $x_{k+1} = \exp_{x_k}(-\eta_k g_k)$ with a particular choice of η_k depending on r, $f(x_k) - f^*$ and $\|g_k\|$. To motivate the choice of step size, we take a geometric approach, which is also useful in the lower bound proof. We initially know that $x^* \in B(x_{\mathrm{ref}}, r)$, and define $s_0 = r$ and $x_0 = x_{\mathrm{ref}}$. At iteration $k \geq 0$, suppose we know that $x^* \in B(x_k, s_k)$. After querying x_k, we receive a subgradient $g_k \in \partial f(x_k)$ with $\|g_k\| \leq M$. Since g_k is a subgradient, we know that $x^* \in U_k = B(x_k, s_k) \cap \{x \in \mathcal{M} : \langle g_k, \log_{x_k}(x) \rangle \leq f^* - f(x_k)\}$.

We define x_{k+1} to be the center of the *minimal ball* containing U_k, and let its radius be s_{k+1}. A computation using hyperbolic trigonometry (see Appendix A.3.1) shows that $x_{k+1} = \exp_{x_k}(-\eta_k g_k)$, where $\eta_k > 0$ and s_{k+1} satisfy

$$\cos(\theta_k) = \frac{f(x_k) - f^*}{s_k \|g_k\|} = \tanh(\eta_k \|g_k\|)/\tanh(s_k), \quad \sin(\theta_k) = \sinh(s_{k+1})/\sinh(s_k),$$

$$\cosh(s_{k+1}) = \cosh(s_k)\sqrt{1 - \frac{1}{\zeta_{s_k}^2} \cdot \frac{(f(x_k) - f^*)^2}{\|g_k\|^2}}, \tag{15}$$

with $\theta_k \in (0, \frac{\pi}{2})$ (see the next subsection for the definition of θ_k).

Defining $\overline{\Delta}_T = \min_{k=0,\ldots,T-1}\{f(x_k) - f^*\}$, we have

$$\cosh(s_{k+1}) \leq \cosh(s_k)\sqrt{1 - \frac{1}{\zeta_r^2} \cdot \frac{(f(x_k) - f^*)^2}{M^2}} \leq \cosh(s_k)\sqrt{1 - \frac{1}{\zeta_r^2} \cdot \frac{\overline{\Delta}_T^2}{M^2}}.$$

Therefore, $1 \leq \cosh(s_T) \leq \cosh(r)\left(\sqrt{1 - \frac{1}{\zeta_r^2} \cdot \frac{\overline{\Delta}_T^2}{M^2}}\right)^T$. After taking logs and rearranging,

$$\log(\cosh(r)) \geq -T \log\left(\sqrt{1 - \frac{1}{\zeta_r^2} \cdot \frac{\overline{\Delta}_T^2}{M^2}}\right) \geq \tanh(r)^2 \cdot \frac{T}{2} \cdot \frac{\overline{\Delta}_T^2}{r^2 M^2},$$

and we conclude that $\overline{\Delta}_T^2 \leq \frac{2\log(\cosh(r))}{\tanh(r)^2} \cdot \frac{r^2 M^2}{T} \leq 2\zeta_r \frac{r^2 M^2}{T}$. So subgradient descent with Polyak step size solves problem **P2** in $O(\frac{\zeta_r}{\epsilon^2})$ queries. We argue below that this is tight.

Remark 32 *In Euclidean space, the Polyak step size can be derived in an analogous way, and leads to $\eta_k = \frac{f(x_k)-f^*}{\|g_k\|^2}$ and $s_{k+1}^2 = s_k^2 - \frac{(f(x_k)-f^*)^2}{\|g_k\|^2}$ (see (Polyak, 1987, Sec. 5.3.2)).*

Proof [Proof of Corollary 14] Consider applying subgradient descent to the function f defined in the proof of Theorem 29. We show by induction on k that the iterates produced by subgradient descent with Polyak step size are exactly $x_k = y_k$, and $s_k = r_k$. This is clearly true for $k = 0$. Suppose $x_k = y_k$ and $s_k = r_k$. Then $f(x_k) - f^* = f(y_k) - f^* = \text{dist}(x_k, x^*) + 0 = r_k$ (Lemma 27), and so

$$\cos(\theta_k) = \frac{f(x_k) - f^*}{s_k \|g_k\|} = \frac{r_k}{r_k \|g_k\|} = \frac{1}{\|g_k\|} = \cos(\theta)$$

using equation (15), and that the subgradient at y_k returned by the oracle from Theorem 29 equals $g_k = \tilde{g}_k = -\frac{1}{\cos(\theta)}e_{k+1}^{(k)}$ (Lemma 28). We conclude that $\theta_k = \theta$, and so $r_{k+1} = s_{k+1}$ and $\eta_k \|g_k\| = \Delta_k$ by equations (15) and (7). Therefore, $x_{k+1} = \exp_{x_k}(-\Delta_k \frac{g_k}{\|g_k\|}) = \exp_{y_k}(\Delta_k e_{k+1}^{(k)}) = y_{k+1}$, which concludes the induction.

By Lemma 27, $y_0, \ldots, y_{d-1} \in L_{k,x^*}$ for all k, so subgradient descent with Polyak step size satisfies assumption A2. Also, clearly subgradient descent satisfies A1. We can therefore, apply Theorem 29. ∎

A.3.1 MINIMAL BALL CONTAINING $B(x_k, r_k) \cap \{x \in \mathcal{M} : \langle g_k, \log_{x_k}(x)\rangle \leq f^* - f(x_k)\}$

In this section we derive explicit formulae for the center x_{k+1} of the minimal ball and its radius s_{k+1}. By symmetry $x_{k+1} = \exp_{x_k}(-\eta_k g_k)$ for some $\eta_k \in \mathbb{R}$.

Take $v \in T_{x_k}\mathcal{M}$ so that $\|v\| = s_k$ and $\langle g_k, v\rangle = f^* - f(x_k)$. Let $\theta_k \in [0, \frac{\pi}{2}]$ denote the angle between $-g_k$ and v. It satisfies $\cos(\theta_k) = \frac{f(x_k)-f^*}{s_k \|g_k\|}$. Let z be the point on the geodesic $t \mapsto \exp_{x_k}(-tg_k)$ such that the hyperbolic triangle $x_k, \exp_{x_k}(v), z$ has right angle at z. Define $s'_{k+1} = \text{dist}(\exp_{x_k}(v), z)$. A computation (details omitted) shows that in fact $B(z, s'_{k+1})$ contains $B(x_k, s_k) \cap \{x \in \mathcal{M} : \langle g_k, \log_{x_k}(x)\rangle \leq f^* - f(x_k)\}$ and is the unique minimal ball, and so we find that $x_{k+1} = z$ and $s_{k+1} = s'_{k+1}$.

To determine s_{k+1}, hyperbolic trigonometry on the right triangle $x_k, \exp_{x_k}(v), z = x_{k+1}$ implies $\frac{f(x_k)-f^*}{s_k \|g_k\|} = \cos(\theta_k) = \tanh(\eta_k \|g_k\|)/\tanh(s_k)$, $\sin(\theta_k) = \sinh(s_{k+1})/\sinh(s_k)$, and

$$\cosh(s_{k+1}) = \cosh(s_k)/\cosh(\eta_k \|g_k\|) = \cosh(s_k)\sqrt{1 - \tanh(\eta_k\|g_k\|)^2}$$

$$= \cosh(s_k)\sqrt{1 - \frac{1}{\zeta_{s_k}^2} \cdot \frac{(f(x_k)-f^*)^2}{\|g_k\|^2}},$$

where the last equality follows from $\frac{f(x_k)-f^*}{s_k\|g_k\|} = \tanh(\eta_k\|g_k\|)/\tanh(s_k)$.

B Lower bound for the cutting-planes game: Proof of Theorem 15

We first need the following bounds on the volume of a hyperbolic ball. Let $V_d(r)$ be the hyperbolic volume of a ball of radius r in $\mathbb{H}^d$. Let ω_d, σ_d be the volume and surface area of the d-dimensional ball, sphere, respectively (as a subset of Euclidean space).

Lemma 33 *We have $V_d(r) = \omega_d \int_0^r \sinh^{d-1}(t)dt \leq \frac{\omega_d}{(d-1)2^{d-1}}e^{r(d-1)}$ for all $r > 0$. If $r \geq 4\log(d)$ then $V_d(r) \geq \frac{\omega_d}{4(d-1)2^{d-1}}e^{r(d-1)}$.*

Proof The formula $V_d(r) = \omega_d \int_0^r \sinh^{d-1}(t)dt$ is given by Ratcliffe (2019, Ex. 3.4.6). The upper bound follows from $\sinh(t) \leq e^t/2$ for all $t \geq 0$. The lower bound follows from the inequality

$$\sinh(t) \geq \frac{1}{2 \cdot 2^{1/(d-1)}}e^t \quad \forall t \geq 2\log(d).$$

Taking $a = 2\log(d)$ and using that $r \geq 2a$, we have

$$V_d(r) \geq \omega_d \int_a^r \sinh(t)^{d-1}dt \geq \frac{\omega_d}{2 \cdot 2^{d-1}} \int_a^r e^{t(d-1)}dt = \frac{\omega_d}{2(d-1)2^{d-1}}[e^{r(d-1)} - e^{a(d-1)}]$$

$$\geq \frac{\omega_d}{4(d-1)2^{d-1}}e^{r(d-1)}.$$

∎

The following Lemma 34 is due to Kisfaludi-Bak (2020). Following the definitions from (Kisfaludi-Bak, 2020, Sec. 2), a tiling of $\mathcal{M}$ is a collection of interior disjoint compact subsets of $\mathcal{M}$ which cover $\mathcal{M}$. A tiling is called (ρ_1, ρ_2)-nice if each tile $T \in \mathcal{T}$ contains a ball of radius ρ_1 with center denoted c_T, and is contained in a ball of radius ρ_2. For $p \in \mathcal{M}$, we define a "random totally geodesic submanifold" through p as follows: $S = \exp_p(\{v \in T_p\mathcal{M} : \langle g, v \rangle = 0\})$ is the totally geodesic submanifold whose unit normal vector $g \in T_p\mathcal{M}$ is chosen uniformly at random from the $(d-1)$-dimensional unit sphere. Kisfaludi-Bak (2020) does not work out the dependence on the dimension d in Lemma 34, so we do this in Appendix J.

Lemma 34 *(Kisfaludi-Bak, 2020, Lem. 9) Let $d \geq 3$, $c_0 \in (0, 1]$ and $\tau \geq 8c_0^{-1}\log(d)$. Let $\mathcal{T}$ be a $(c_0\tau/2, \tau/2)$-nice tiling, and let $\mathcal{I}$ be a finite subset of $\mathcal{T}$. Let $p \in M$ and S be a random totally geodesic submanifold through p. Then $\mathbb{E}[|\mathcal{S}|] \leq \frac{1}{2}e^{5(d-1)\tau}|\mathcal{I}|^{\frac{d-2}{d-1}}$, where $\mathcal{S} = \{T \in \mathcal{I} : T \cap S \neq \emptyset\}$.*

Proof [Proof of Theorem 15] Define $c_0 = 1/4$, $\tau = \frac{r}{40(d-1)}$. Then $\tau \geq 8c_0^{-1}\log(d)$, and there exists a $(\epsilon r = c_0\tau/2, \tau/2)$-nice tiling $\mathcal{T}$ of $\mathcal{M}$, as explained in Appendix J. Let $\mathcal{I}_0$ be the set of tiles in $\mathcal{T}$ contained in $B(x_{\mathrm{ref}}, r)$, and define $A_0 = \{c_T : T \in \mathcal{I}_0\}$. A_0 is the set of initial possible candidates for x^*. We know that the tiles $\mathcal{I}_0$ cover $B(x_{\mathrm{ref}}, r/2)$. Using the bounds on $V_d(r)$ given by Lemma 33, we have $|A_0| = |\mathcal{I}_0| \geq \frac{V_d(r/2)}{V_d(\tau/2)} \geq \frac{1}{4}e^{\frac{1}{4}(d-1)r}$.

For every $k = 0, 1, \ldots, T - 1$, the algorithm makes the query x_k, and we show (by induction) that there is a unit vector $g_k \in T_{x_k}\mathcal{M}$ and a set $\mathcal{I}_{k+1} \subseteq \mathcal{I}_k$ such that $|\mathcal{I}_{k+1}| \geq |\mathcal{I}_k|/4$, and each $x^* \in A_{k+1} = \{c_T : T \in \mathcal{I}_{k+1}\}$ is consistent with the past queries and such that the algorithm has not won the game, i.e.,

$$\langle g_\ell, \log_{x_\ell}(x^*) \rangle \leq 0, \quad \text{and} \quad S_\ell \cap B(x^*, \epsilon r) = \emptyset, \quad \forall \ell \in \{0, \ldots, k\} \text{ and } \forall x^* \in A_{k+1}.$$

Indeed, fix $k \in \{0, 1, \ldots, T - 1\}$. By Lemma 34 there exists $\tilde{g}_k \in T_{x_k}\mathcal{M}$ such that

$$|\{T \in \mathcal{I}_k : B(c_T, \epsilon r) \cap S_k \neq \emptyset\}| \leq |\{T \in \mathcal{I}_k : T \cap S_k \neq \emptyset\}| \leq \frac{1}{2}e^{5(d-1)\tau}|\mathcal{I}|^{\frac{d-2}{d-1}}$$

28

where as usual $S_k = \exp_{x_k}(\{v \in \mathrm{T}_{x_k}\mathcal{M} : \langle g_k, v \rangle = 0\})$. For $s \in \{+1, -1\} = \{+, -\}$, define

$$\mathcal{I}_{k+1}^s = \{T \in \mathcal{I}_k : s\langle g_k, \log_{x_k}(y) \rangle \geq 0, \ \forall y \in B(c_T, \epsilon r)\}.$$

If $|\mathcal{I}_{k+1}^+| \geq |\mathcal{I}_{k+1}^-|$, define $g_k = \tilde{g}_k, \mathcal{I}_{k+1} = \mathcal{I}_{k+1}^+$. Otherwise, define $g_k = -\tilde{g}_k, \mathcal{I}_{k+1} = \mathcal{I}_{k+1}^-$. So,

$$|\mathcal{I}_{k+1}| \geq \frac{1}{2}\left(|\mathcal{I}_k| - |\{T \in \mathcal{I}_k : B(c_T, \epsilon r) \cap S_k \neq \emptyset\}|\right) \geq \frac{1}{2}|\mathcal{I}_k|\left(1 - \frac{1}{2}e^{5(d-1)\tau}|\mathcal{I}_k|^{-1/(d-1)}\right).$$

By induction, we know that

$$|\mathcal{I}_k| \geq \frac{1}{4^k}|\mathcal{I}_0| \geq \frac{1}{4^{T-1}}|\mathcal{I}_0| \geq \frac{1}{4^T}e^{\frac{1}{4}(d-1)r} \geq \frac{1}{e^{2T}}e^{\frac{1}{4}(d-1)r} \geq e^{\frac{1}{8}(d-1)r} \geq e^{5(d-1)^2\tau}$$

where the penultimate inequality follows from $T = \frac{1}{16}(d-1)r$, and the last inequality follows from $\tau = \frac{r}{40(d-1)}$. Therefore $|\mathcal{I}_k|^{-1/(d-1)} \leq e^{-5(d-1)\tau}$, and so $|\mathcal{I}_{k+1}| \geq \frac{1}{4}|\mathcal{I}_k|$, finishing the induction.

We know A_T is nonempty because $|A_0| \geq \frac{1}{4}e^{\frac{1}{4}(d-1)r}$ and $|A_T| \geq \frac{1}{4^T}|A_0| \geq \frac{1}{4^{T+1}}e^{\frac{1}{4}(d-1)r} \geq 1$. So there exists $x^* \in A_T$ such that $\langle g_\ell, \log_{x_\ell}(x^*) \leq 0$ and $\mathrm{dist}(S_\ell, x^*) > \epsilon r$ for all $\ell = 0, \ldots, T$. ∎

C Geodesically convex interpolation: a closer look than Section 8

In this section, we initiate a study of interpolation by g-convex functions. We leave a more complete study of interpolation to future work. Let $\mathcal{M}$ be a Hadamard manifold. We call a collection $(F_i, x_i, g_i)_{i=1}^N$ the "data" of an interpolation problem if $x_i \in \mathcal{M}, F_i \in \mathbb{R}$ and $g_i \in \mathrm{T}_{x_i}\mathcal{M}$ for all $i = 1, \ldots, N$. We say the data is *interpolated* by a g-convex function $f \colon \mathcal{M} \to \mathbb{R}$ if $f(x_i) = F_i$ and $g_i \in \partial f(x_i)$ for all i. We say a function $f \colon \mathcal{M} \to \mathbb{R}$ is the *minimal* (*maximal*) g-convex function interpolating the data $(F_i, x_i, g_i)_{i=1}^N$ if for all other g-convex functions $\tilde{f} \colon \mathcal{M} \to \mathbb{R}$ we have $\tilde{f}(x) \geq f(x)$ ($\tilde{f}(x) \leq f(x)$) for all $x \in \mathcal{M}$.

Consider an algorithm running on a g-convex function $f \colon \mathcal{M} \to \mathbb{R}$ querying points $x_0, x_1, \ldots$. Given $F_k = f(x_k)$ and $g_k \in \partial f(x_k)$, the algorithm learns that f is lower bounded as

$$f(x) \geq F_k + \langle g_k, \log_{x_k}(x) \rangle \quad \forall x \in \mathcal{M}. \tag{16}$$

Since the data $(F_\ell, x_\ell, g_\ell)_{\ell=0}^k$ is interpolated by the g-convex function f, it must satisfy

$$F_j \geq F_i + \langle g_i, \log_{x_i}(x_j) \rangle \quad \forall i, j \in \{0, \ldots, k\}. \tag{17}$$

Aggregating all the lower bounds (16) from past queries, the algorithm knows that f satisfies

$$f(x) \geq f_k(x) \quad \forall x \in \mathcal{M}, \quad \text{where} \quad f_k(x) = \max_{\ell \in \{0,\ldots,k\}}\{F_\ell + \langle g_\ell, \log_{x_\ell}(x) \rangle\}.$$

Using the conditions (17), one sees that this lower bound f_k does interpolate the data $(F_\ell, x_\ell, g_\ell)_{\ell=0}^k$.

If $\mathcal{M}$ is a Euclidean space, the functions $x \mapsto F_i + \langle g_i, \log_{x_i}(x) \rangle$ are affine, and so the lower bound f_k is *convex*: f_k is the minimal convex function interpolating the data learned, and in particular the algorithm knows $f^* \geq \min_{x \in B(x_{\mathrm{ref}}, r)}\{f_k(x)\}$. However for general $\mathcal{M}$ (e.g., if $\mathcal{M}$ is a hyperbolic space), the functions $x \mapsto F_i + \langle g_i, \log_{x_i}(x) \rangle$ are not g-convex, and the lower bound f_k is *not g-convex*. This points to a primary difficulty inherent in proving complexity lower bounds for g-convex optimization. Perhaps when the algorithm learns that $F = f(y)$ and $g \in \partial f(y)$, actually lower bound $f(x) \geq F + \langle g, \log_y(x) \rangle$ is loose and can be improved? This is not the case.

Proposition 35 *Let $\mathcal{M}$ be a Hadamard manifold, $F \in \mathbb{R}$, and $(y, g) \in T\mathcal{M}$. Let $\mathcal{F}_{F,y,g}$ denote the set of g-convex functions $f \colon \mathcal{M} \to \mathbb{R}$ with $f(y) = F$ and $g \in \partial f(y)$. Then for all $x \in \mathcal{M}$*

$$\min_{f \in \mathcal{F}_{F,y,g}} \{f(x)\} = F + \langle g, \log_y(x) \rangle.$$

Proof Fix $x \in \mathcal{M} \setminus \{y\}$. Define $g_{||} = \frac{\langle g, \log_y(x) \rangle}{\mathrm{dist}(x,y)^2} \log_y(x)$ and $g_\perp = g - g_{||}$. Let S be the geodesic line $\{\exp_y(t \log_y(x)) : t \in \mathbb{R}\}$. Note that S is a g-convex set, and so $z \mapsto \mathrm{dist}(z, S)$ is g-convex.

Define the point x' as $x' = x$ if $\langle g, \log_y(x) \rangle \leq 0$, and $x' = \exp_y(-\log_y(x))$ if $\langle g, \log_y(x) \rangle > 0$. Define the functions

$$f_{||}(z) = F - \|g_{||}\|\mathrm{dist}(x', y) + \|g_{||}\|\mathrm{dist}(x', z), \quad f_\perp(z) = \|g_\perp\|\mathrm{dist}(z, S), \quad f(z) = f_{||}(z) + f_\perp(z).$$

We know that $g_\perp$ is a subgradient of $f_\perp$ at y, since $g_\perp$ is orthogonal to S, and

$$\nabla f_{||}(y) = -\frac{\|g_{||}\|}{\mathrm{dist}(x, y)} \log_y(x') = -\frac{|\langle g, \log_y(x) \rangle|}{\mathrm{dist}(x, y)^2} \log_y(x') = \frac{\langle g, \log_y(x) \rangle}{\mathrm{dist}(x, y)^2} \log_y(x) = g_{||}.$$

Therefore, $\nabla f_{||}(y) + g_\perp = g_{||} + g_\perp = g$ is a subgradient of f at y. Also, $f(y) = F + \|g_\perp\|\mathrm{dist}(y, S) = F$, and f is g-convex (as the sum of two g-convex functions). So $f \in \mathcal{F}_{F,y,g}$.

Lastly, if $\langle g, \log_y(x) \rangle \leq 0$, then $f(x) = F - \|g_{||}\|\mathrm{dist}(x, y) = F + \langle \log_y(x), g \rangle$, using $\mathrm{dist}(x, S) = 0$ and $x = x'$. If $\langle g, \log_y(x) \rangle > 0$, then $f(x) = F + \|g_{||}\|\mathrm{dist}(x, y) = F + \langle \log_y(x), g \rangle$, using $\mathrm{dist}(x, S) = 0$ and $\mathrm{dist}(x, x') = 2\mathrm{dist}(x, y)$. ∎

Remark 36 *Observe that the worst function in the world given by equation (14) is essentially a maximum of functions from the proof of Proposition 35. This is not a coincidence.*

We know that f_k is not g-convex. However, perhaps the necessary conditions (17) for interpolation by a g-convex function are also sufficient? The necessary conditions are sufficient if $\mathcal{M}$ is a Euclidean space; they are *not sufficient* for negatively curved spaces even for just three points.

Proposition 37 (Restatement of Proposition 16) *Let $\mathcal{M} = \mathbb{H}^2$. There exists data $(F_i, x_i, g_i)_{i=1}^3$ such that*

$$F_j \geq F_i + \langle g_i, \log_{x_i}(x_j) \rangle \qquad \forall i, j \in \{1, 2, 3\}, \tag{18}$$

yet this data cannot be interpolated by any g-convex function.

Proof We denote the geodesic segment $\{\exp_x(t \log_x(y)) : t \in [0, 1]\}$ between $x, y \in \mathcal{M}$ by $[x, y]$.

Consider three points $x_1, x_2, x_3 \in \mathcal{M}$ such that $\mathrm{dist}(x_1, x_2) = \mathrm{dist}(x_1, x_3) = 1$ and the angle (at x_1) between the geodesic segments $[x_1, x_2]$ and $[x_1, x_3]$ equals 2θ. Let h be the length of the altitude from x_1 to the geodesic segment $[x_2, x_3]$. Assume the altitude intersects the segment $[x_2, x_3]$ at point p. Of course $h = \mathrm{dist}(x_1, p)$, p is the midpoint of $[x_2, x_3]$, the altitude bisects the angle at x_1, and the altitude intersects $[x_2, x_3]$ at a right angle.

Consider the data $F_1 = 1, F_2 = F_3 = 0$, and let $g_1 = -\frac{1}{\cos(\theta)} \cdot \frac{\log_{x_1}(p)}{\mathrm{dist}(x_1, p)}$. Assume, for the sake of contradiction, that the data can be interpolated by a g-convex function f. Then, $f(p) \leq$

$\max\{f_1, f_2\} = 0$. Additionally, the function $t \mapsto f(\exp_{x_1}(t \exp_{x_1}^{-1}(p)))$ is convex, so $f(p) \geq f(x_1) + \langle g_1, \log_{x_1}(p) \rangle = 1 - \frac{h}{\cos(\theta)}$.

We have not used the non-Euclidean geometry of $\mathcal{M}$. We do so now. By hyperbolic trigonometry, $\cos(\theta) = \frac{\tanh(h)}{\tanh(1)} < h$ (since $h < 1$). Hence, $f(p) \geq 1 - \frac{h}{\cos(\theta)} > 0$. This is a contradiction.

Lastly, it is easy to choose g_2 and g_3 so that the data $(F_i, x_i, g_i)_{i=1}^3$ satisfies (18). For example, one may choose $g_2 = g_3 = 0$. Alternatively, if we let α be the angle (at x_2) between the segments $[x_2, x_1]$ and $[x_2, x_3]$, then we can take g_2, g_3 to be tangent vectors of length $\frac{1}{\sin(\alpha)}$ which are perpendicular to the geodesic segment $[x_2, x_3]$. ∎

Let us take another perspective. The set of convex functions on $\mathbb{R}^d$ equals the set of functions which are each a supremum of affine functions. More generally, abstract convexity (Rubinov, 2010) considers classes of functions (the "convex" functions) which can be written as supremums of a base class of functions (the "base" functions). Geodesic convexity does not naturally fit into this framework. Indeed, it is easy to check that every g-convex function can be written as a supremum of functions of the form $x \mapsto F + \langle g, \log_y(x) \rangle$. However, it is not true that every supremum of functions of this form are g-convex (e.g., consider finite supremums like f_k above).

Despite the negative result Proposition 16, we do, however, have the following necessary and sufficient conditions for interpolation.

Proposition 38 *Consider the data $(F_i, x_i, g_i)_{i=1}^N$, let $\mathcal{D} = (F_i, x_i)_{i=1}^N$, and let $\mathcal{F}_\mathcal{D}$ denote the set of g-convex functions f satisfying $f(x_i) \leq F_i$ for $i = 1, \ldots, N$. Define the function*

$$f_\mathcal{D} \colon \mathcal{M} \to \mathbb{R} \cup \{+\infty\}, \quad f_\mathcal{D}(x) = \sup_{f \in \mathcal{F}_\mathcal{D}} \{f(x)\}.$$

Then the data $(F_i, x_i, g_i)_{i=1}^N$ can be interpolated by a g-convex function if and only if

$$f_\mathcal{D}(x) \geq F_i + \langle g_i, \log_{x_i}(x) \rangle, \quad \forall x \in \mathcal{M} \text{ and } \forall i = 1, \ldots, N. \tag{19}$$

If $(F_i, x_i, g_i)_{i=1}^N$ can be interpolated by a g-convex function, $f_\mathcal{D}$ is the maximal such function.

Proof (Condition (19) is sufficient) Assume (19) holds. Taking $x = x_i$, $f_\mathcal{D}(x_i) \geq F_i$. But by definition, $f_\mathcal{D} \leq F_i$. Hence, $f_\mathcal{D}(x_i) = F_i$, and so condition (19) implies g_i is a subgradient of $f_\mathcal{D}$ at x_i. So $f_\mathcal{D}$ interpolates the data.

(Condition (19) is necessary) Assume the data is interpolated by a g-convex function f. Then $f(x) \leq f_\mathcal{D}(x)$ for all $x \in \mathcal{M}$. Therefore for each i and x, $f_\mathcal{D}(x) \geq f(x) \geq F_i + \langle g_i, \log_{x_i}(x) \rangle$. So condition (19) holds, and the previous paragraph implies $f_\mathcal{D}$ interpolates the data. Therefore, $f_\mathcal{D}$ is the maximal g-convex function interpolating the data. ∎

Remark 39 *One can relate $f_\mathcal{D}$ to a g-convex hull (Lytchak and Petrunin, 2022) in the Hadamard manifold $\mathbb{R} \times \mathcal{M}$. For simplicity, we omit the details. One can verify that if $\mathcal{M}$ is a hyperbolic space and x is not in the g-convex hull of $\{x_1, \ldots, x_N\}$, then $f_\mathcal{D}(x) = +\infty$.*

If $\mathcal{M}$ is a Euclidean space, then $f_\mathcal{D}$ has the explicit formula

$$f_\mathcal{D}(x) = \min_{\lambda_i \geq 0, \sum_{i=1}^N \lambda_i = 1, \sum_{i=1}^N \lambda_i x_i = x} \left\{ \sum_{i=1}^N \lambda_i F_i \right\},$$

and the set of necessary and sufficient conditions in Proposition 38 can shown to be equivalent to the conditions (17) (Taylor, 2018, Ch. 3). However, it is unclear that there is an analogous formula even for hyperbolic space.

Let us conclude this section with a positive result about g-convex interpolation. For simplicity we focus on hyperbolic spaces. We have the following lemma which may be useful in other contexts.

Lemma 40 *Let $\mathcal{M} = \mathbb{H}^d$. The function $x \mapsto \langle g, \log_y(x) \rangle$ is globally $\|g\|$-Lipschitz and $\|g\|$-smooth.*

Proof For $s, \dot{s} \in \mathrm{T}_y\mathcal{M}$, let $P_{y,s} = P_{y \to \exp_y(s)}$, $T_{y,s} = d\exp_y(s)$ and $c_{y,s,\dot{s}}(t) = \exp_y(s + t\dot{s})$. Given $s, \dot{s} \in \mathrm{T}_y\mathcal{M}$, let $\dot{s}_{||} = \frac{\langle s, \dot{s} \rangle}{\|s\|^2} s$ and $\dot{s}_\perp = \dot{s} - \dot{s}_{||}$. Moreover, given $\ddot{s} \in \mathrm{T}_{\exp_y(s)}\mathcal{M}$, let $\ddot{s}_{||} = \frac{\langle P_{y,s}s, \ddot{s} \rangle}{\|s\|^2} s$ and $\ddot{s}_\perp = \ddot{s} - \ddot{s}_{||}$. It is well-known (see for example (Criscitiello and Boumal, 2022b, App. B)) that

$$T_{y,s}\dot{s} = P_{y,s}\dot{s}_{||} + \frac{\sinh(\|s\|)}{\|s\|} P_{y,s}\dot{s}_\perp, \qquad T_{y,s}^{-1}\ddot{s} = P_{y,s}^{-1}\ddot{s}_{||} + \frac{\|s\|}{\sinh(\|s\|)} P_{y,s}^{-1}\ddot{s}_\perp. \qquad (20)$$

Define $f(x) = \langle g, \log_y(x) \rangle$, and $\hat{f} = f \circ \exp_y \colon \mathrm{T}_y\mathcal{M} \to \mathbb{R}$. Note that $\hat{f}(s) = \langle g, s \rangle$. Let $x \in \mathcal{M}$, $s = \log_y(x)$, and $\ddot{s} \in \mathrm{T}_x\mathcal{M}$ with $\|\ddot{s}\| = 1$. By Lemma 5 of (Agarwal et al., 2020),

$$\nabla f(x) = (T_{y,s}^{-1})^* \nabla \hat{f}(s) = (T_{y,s}^{-1})^* g \qquad (21)$$

and, defining $\dot{s} = T_{y,s}^{-1}\ddot{s}$,

$$\langle \ddot{s}, \nabla^2 f(x)\ddot{s} \rangle = \langle T_{y,s}^{-1}\ddot{s}, \nabla^2 \hat{f}(s) T_{y,s}^{-1}\ddot{s} \rangle - \langle \nabla f(x), c_{y,s,\dot{s}}''(0) \rangle = -\langle g, T_{y,s}^{-1} c_{y,s,\dot{s}}''(0) \rangle. \qquad (22)$$

Equations (20) and (21) imply $\|\nabla f(x)\| \leq \|T_{y,s}^{-1}\| \cdot \|g\| \leq \|g\|$, so f is $\|g\|$-Lipschitz.

By equation (73) from (Criscitiello and Boumal, 2022b, App. B), we know that

$$c_{y,s,\dot{s}}''(0) = -\frac{\sinh(\|s\|)\cosh(\|s\|) - \|s\|}{\|s\|^3} \|\dot{s}_\perp\|^2 \cdot P_{y,s}s + 2\frac{\cosh(\|s\|)\|s\| - \sinh(\|s\|)}{\|s\|^3} \langle s, \dot{s} \rangle \cdot P_{y,s}\dot{s}_\perp.$$

Therefore, using equation (20) and $\dot{s} = T_{y,s}^{-1}\ddot{s}$,

$$\begin{aligned}
T_{y,s}^{-1} c_{y,s,\dot{s}}''(0) = {}& -\frac{\sinh(\|s\|)\cosh(\|s\|) - \|s\|}{\|s\|^3} \|\dot{s}_\perp\|^2 \cdot s \\
& + 2\frac{\|s\|}{\sinh(\|s\|)} \cdot \frac{\cosh(\|s\|)\|s\| - \sinh(\|s\|)}{\|s\|^3} \langle s, \dot{s} \rangle \cdot \dot{s}_\perp \\
= {}& -\frac{\sinh(\|s\|)\cosh(\|s\|) - \|s\|}{\|s\|^3} \frac{\|s\|^2}{\sinh^2(\|s\|)} \|\ddot{s}_\perp\|^2 \cdot s \\
& + 2\frac{\|s\|^2}{\sinh^2(\|s\|)} \cdot \frac{\cosh(\|s\|)\|s\| - \sinh(\|s\|)}{\|s\|^3} \langle P_{y,s}s, \ddot{s} \rangle \cdot P_{y,s}^{-1}\ddot{s}_\perp.
\end{aligned}$$

Using $\|\ddot{s}\| = 1$, we find

$$\|T_{y,s}^{-1} c_{y,s,\dot{s}}''(0)\| \leq \frac{\sinh(\|s\|)\cosh(\|s\|) - \|s\|}{\|s\|^3} \frac{\|s\|^2}{\sinh^2(\|s\|)} \|s\| \leq 1.$$

By (22), we conclude $|\langle \ddot{s}, \nabla^2 f(x)\ddot{s}\rangle| \leq \|g\| \cdot \|T_{y,s}^{-1} c''_{y,s,\dot{s}}(0)\| \leq \|g\|$. ∎

Using this lemma, we can derive sufficient conditions for g-convex interpolation (Proposition 41) and extension (Proposition 42).

Proposition 41 (Restatement of Proposition 17) *Consider the data $(F_i, x_i, g_i)_{i=1}^{N}$. Assume that $\|g_i\| \leq \frac{\mu}{2}$ for all $i = 1, \ldots, N$, and*

$$F_j \geq F_i + \langle g_i, \log_{x_i}(x_j)\rangle + \frac{\mu}{2}\mathrm{dist}(x_i, x_j)^2 \quad \forall i, j \in \{1, \ldots, N\}. \tag{23}$$

Then $(F_i, x_i, g_i)_{i=1}^{N}$ is interpolated by a $\frac{\mu}{2}$-strongly g-convex function.

Proof Let $f(x) = \max_{i \in \{1, \ldots, N\}} \left\{ F_i + \langle g_i, \log_{x_i}(x)\rangle + \frac{\mu}{2}\mathrm{dist}(x_i, x)^2 \right\}$. Taking $i = j$, we know $f(x_j) \geq F_j$. Inequality (23) implies $f(x_j) \leq F_j$, so we conclude $f(x_j) = F_j$, for all j. Moreover,

$$f(x) \geq F_i + \langle g_i, \log_{x_i}(x_j)\rangle + \frac{\mu}{2}\mathrm{dist}(x_i, x)^2 \geq f(x) + \langle g_i, \log_{x_i}(x_j)\rangle, \quad \forall x \in \mathcal{M}, \forall i \in \{1, \ldots, N\},$$

so $g_i \in \partial f(x_i)$. So f interpolates the data.

By Lemma 40, each function $x \mapsto F_i + \langle g_i, \log_{x_i}(x)\rangle + \frac{\mu}{2}\mathrm{dist}(x_i, x)^2$ is $\mu - \|g_i\| \geq \mu - \frac{\mu}{2} = \frac{\mu}{2}$-strongly g-convex. Therefore, f is $\frac{\mu}{2}$-strongly g-convex. ∎

We have the following result (compare with the convex case (Dragomirescu and Ivan, 1992)):

Proposition 42 *Let D be an open g-convex subset of $\mathcal{M} = \mathbb{H}^d$ and let $f \colon D \to \mathbb{R}$ be differentiable and μ-strongly g-convex in D. If $\|\nabla f(x)\| \leq \mu$ for all $x \in D$, then there is a globally g-convex function $\tilde{f} \colon \mathcal{M} \to \mathbb{R}$ such that $\tilde{f}(x) = f(x)$ for all $x \in D$.*

Proof Define

$$\tilde{f}(x) = \sup_{y \in D} \left\{ f(y) + \langle \nabla f(y), \log_y(x)\rangle + \frac{\mu}{2}\mathrm{dist}(x, y)^2 \right\}.$$

If $x \in D$, then $\tilde{f}(x) \geq f(x)$ (taking $y = x$). On the other hand, μ-strong g-convexity of f in D implies $f(x) \geq f(y) + \langle \nabla f(y), \log_y(x)\rangle + \frac{\mu}{2}\mathrm{dist}(x, y)^2$ for all $y \in D$. Therefore, $f(x) \geq \tilde{f}(x)$. We conclude $f(x) = \tilde{f}(x)$ if $x \in D$.

By Lemma 40, the function $x \mapsto f(y) + \langle \nabla f(y), \log_y(x)\rangle$ is $\|\nabla f(y)\|$-smooth. Hence, the function $x \mapsto f(y) + \langle \nabla f(y), \log_y(x)\rangle + \frac{\mu}{2}\mathrm{dist}(x, y)^2$ is $\mu - \|\nabla f(y)\| \geq \mu - \mu = 0$-strongly g-convex, if $y \in D$. Therefore, $\tilde{f}$ is g-convex. ∎

D Carrying over the convex upper and lower bounds when r is very small

It is important to mention that if the optimization is carried out in a very small region then smooth g-convex optimization reduces to Euclidean convex optimization, and in particular upper and lower bounds from Euclidean space carry over. However, we shall see that the resulting lower bounds are worse than those given in Section 5. This technique does, however, allow us to easily construct an "eventually accelerated" algorithm for smooth strongly g-convex optimization, like the algorithm of Ahn and Sra (2020). See below.

To state the following proposition, we need properties of the Riemannian curvature tensor R and its covariant derivative ∇R which are worked out in (Criscitiello and Boumal, 2022b, Sec. 2). The assumption $\|\nabla R\| \leq F$ holds with $F = 0$ for symmetric spaces, including hyperbolic spaces.

Proposition 43 *Let $\mathcal{M}$ be a complete Riemannian manifold which has sectional curvatures in the interval $[-K, K]$ with $K \geq 0$ and also satisfies $\|\nabla R(x)\| \leq F$ for all $x \in \mathcal{M}$.*

Let $x_{\mathrm{ref}} \in \mathcal{M}$. Let $f \colon \mathcal{M} \to \mathbb{R}$ be a twice continuously differentiable function which has a minimizer x^*. Assume f is μ-strongly g-convex and L-smooth in $B(x_{\mathrm{ref}}, r)$, $\kappa = \frac{L}{\mu}$, and x^* is contained in $B(x_{\mathrm{ref}}, r)$, where $r \in (0, \frac{1}{\sqrt{\kappa}} \cdot \min\{\frac{1}{4\sqrt{K}}, \frac{K}{4F}\}]$. Define the pullback $\hat{f} = f \circ \exp_{x_{\mathrm{ref}}}$. Then, $\hat{f}$ is $\frac{\mu}{2}$-strongly convex and $\frac{3}{2}L$-smooth in the Euclidean sense in $B_{x_{\mathrm{ref}}}(0, r) \subseteq \mathrm{T}_{x_{\mathrm{ref}}}\mathcal{M}$.

Proof Following (Criscitiello and Boumal, 2022b, Sec. 2), if $\|s\| \leq \min\{\frac{1}{4\sqrt{K}}, \frac{K}{4F}\}$ then

$$\left\| \nabla^2 \hat{f}(s) - P^*_{x_{\mathrm{ref}},s} \nabla^2 f(y) P_{x_{\mathrm{ref}},s} \right\| \leq \frac{7}{9} LK \|s\|^2 + \frac{3}{2} K \|s\| \|\nabla f(y)\|$$

where $y = \exp_{x_{\mathrm{ref}}}(s)$ and $P_{x,s}$ denotes parallel transport along the curve $c(t) = \exp_x(ts)$ from $t = 0$ to $t = 1$. Using L-smoothness of f and $x^* \in B(x_{\mathrm{ref}}, r)$,

$$\|\nabla f(y)\| = \|\nabla f(y) - P_{x^* \to y} \nabla f(x^*)\| \leq L \mathrm{dist}(y, x^*) \leq 2Lr.$$

We determine $\left\| \nabla^2 \hat{f}(s) - P^*_{x_{\mathrm{ref}},s} \nabla^2 f(y) P_{x_{\mathrm{ref}},s} \right\| \leq 4LKr^2 \leq \frac{\mu}{2}$ if $\|s\| \leq r$, by our choice of r.

By μ-strong g-convexity and L-smoothness of f, we know all eigenvalues of the symmetric operator $P^*_{x_{\mathrm{ref}},s} \nabla^2 f(y) P_{x_{\mathrm{ref}},s}$ are in $[\mu, L]$. Hence, all eigenvalues of $\nabla^2 \hat{f}(s)$ are in the interval $[\frac{\mu}{2}, L + \frac{\mu}{2}]$ for all $\|s\| \leq r$. ∎

We can use this proposition to prove an $\Omega(\sqrt{\kappa})$ lower bound when minimizing a κ-conditioned smooth strongly g-convex function in a ball of radius $r \leq O(\frac{1}{\sqrt{\kappa}})$. Indeed, any algorithm $\mathcal{A}$ querying points x on $\mathcal{M}$ can be converted into an algorithm $\tilde{\mathcal{A}}$ querying points in a fixed tangent space $\mathrm{T}_{x_{\mathrm{ref}}}\mathcal{M}$ using the exponential map in the obvious way. The lower bound for Euclidean spaces provides a hard function $\tilde{f} \colon \mathrm{T}_{x_{\mathrm{ref}}}\mathcal{M} \to \mathbb{R}$ for $\tilde{\mathcal{A}}$. In turn, the function $f = \tilde{f} \circ \log_{x_{\mathrm{ref}}}$ is a hard function for $\mathcal{A}$. Proposition 43 guarantees f is g-convex on $\mathcal{M}$ in a small enough ball of radius r. Likewise, we can prove an $\Omega(\frac{1}{\sqrt{\epsilon}})$ lower bound for minimizing smooth g-convex functions in a ball of radius $r \leq O(\sqrt{\epsilon})$.

However, *these lower bounds have two major downsides*. First, the functions constructed are not globally g-convex (or L-smooth) and so do not actually belong to the function classes we defined in Section 1.1. Second, these lower bounds only hold when r is very small, e.g., $r \leq O(\sqrt{\epsilon})$. Ignoring the first downside, one can still try extend these bounds to large r; however, this results in unsatisfactory bounds. For example, for smooth g-convex optimization, roughly we have that $\Omega(\frac{1}{\sqrt{\epsilon}})$ queries are required to find a point x such that $f(x) - f^* \leq \epsilon \cdot \frac{1}{2} L(\sqrt{\epsilon})^2 = \frac{\epsilon^2}{r^2} \cdot \frac{1}{2} Lr^2$. Letting $\epsilon' = \frac{\epsilon^2}{r^2}$, we find that $\Omega(\frac{1}{\sqrt{r}(\epsilon')^{1/4}})$ queries are needed to find a point x such that $f(x) - f^* \leq \epsilon' \cdot \frac{1}{2} Lr^2$. This bound does not even have the correct scaling in ϵ'!

Lastly, we mention that we can use Proposition 43 to construct an "eventually accelerated" algorithm for smooth strongly g-convex optimization, like the algorithm of Ahn and Sra (2020). Let $\mathcal{M}$ be a complete Riemannian manifold with constants K and F as in Proposition 43. Consider a function $f \colon \mathcal{M} \to \mathbb{R}$ with a minimizer x^*. Assume f is globally μ-strongly g-convex and L-smooth in some g-convex set D containing the ball $B(x^*, r)$ where $r = \frac{1}{\sqrt{\kappa}} \cdot \min\{\frac{1}{4\sqrt{K}}, \frac{K}{4F}\}$. Suppose $x_0 \in B(x^*, r)$ and $B(x_0, r) \subseteq D$. Since $\hat{f}$ has condition number at most 3κ in $B_{x_0}(0, r) \subseteq \mathrm{T}_{x_0}\mathcal{M}$, we can minimize f by simply running a Euclidean algorithm on $\hat{f} = f \circ \exp_{x_0}$ in the ball $B_{x_0}(0, r)$.

For example, running an accelerated algorithm for constrained optimization (like from (Nesterov, 2013, Sec. 5.1)) produces a *locally* accelerated algorithm for g-convex optimization.

This gives a simple proof that when $\text{dist}(x_0, x^*) \leq O(\frac{1}{\sqrt{\kappa}})$, there is an accelerated algorithm. We can combine this algorithm with Riemannian gradient descent (RGD) to get a globally convergent algorithm which performs at least as well as RGD and which is eventually accelerated. Consider running $T = \tilde{\Theta}(\kappa)$ steps of the algorithm we just described (with $r = \frac{1}{\sqrt{\kappa}} \cdot \min\{\frac{1}{4\sqrt{K}}, \frac{K}{4F}\}$) in parallel with T steps of projected RGD on f (in D), both initialized at the same point $x_0 \in D$. After these T steps, we set $x_1 \in D$ to be the last iterate produced by one of the algorithms based on function value and we repeat starting from x_1, etc. By design, this algorithm performs at least as well as RGD in terms of function value. By our choice of $T = \tilde{\Theta}(\kappa)$, once an iterate x_k of the algorithm enters $B(x^*, r)$, then the distance to the minimizer of the subsequent outer iterates $x_{k+1}, x_{k+2}, \ldots$ is decreasing. So once this method is restarted in $B(x^*, r)$, it stays in this ball, and the NAG sequence in a tangent space provides cost function value decrease at an accelerated rate. Therefore this algorithm is eventually accelerated. We note that the algorithms of Zhang and Sra (2018) and Ahn and Sra (2020) do not require a bound on $\|\nabla R\|$ while the algorithm we presented does.

E Proof of Theorem 8: showing $x^* \in \bigcap_{k=0}^{T-1} S_{i_k}^{s_k}$

Consider the hyperbolic triangle formed by $x^*, x_{\text{ref}}, z_{i_k}^{s_k}$. Let $\beta \in [0, \pi]$ be the angle between $\log_{z_{i_k}^{s_k}}(x^*)$ and $\log_{z_{i_k}^{s_k}}(x_{\text{ref}})$. Let $b = \text{dist}(x^*, z_{i_k}^{s_k})$. The angle between $\log_{x_{\text{ref}}}(x^*)$ and $\log_{x_{\text{ref}}}(z_{i_k}^{s_k}) = as_k e_{i_k}$ equals $\theta = \arccos(1/\sqrt{d})$ for all $k = 0, \ldots, T-1$. The hyperbolic law of cosines (Ratcliffe, 2019, Ch. 3.5) and $\cos(\theta) = \frac{1}{\sqrt{d}} = \frac{\tanh(a)}{\tanh(r)}$ imply

$$\cosh(b) = \cosh(a)\cosh(r) - \sinh(a)\sinh(r)\cos(\theta)$$
$$= \cosh(a)\cosh(r) - \sinh(a)\cosh(r)\tanh(a)$$
$$= \frac{\cosh(r)}{\cosh(a)}[\cosh(a)^2 - \sinh^2(a)] = \frac{\cosh(r)}{\cosh(a)}.$$

One more application of the hyperbolic law of cosines implies

$$\cosh(r) = \cosh(a)\cosh(b) - \sinh(a)\sinh(b)\cos(\beta) = \cosh(r) - \sinh(a)\sinh(b)\cos(\beta).$$

Therefore, $\cos(\beta) = 0$, i.e., $\langle \log_{z_{i_k}^{s_k}}(x_{\text{ref}}), \log_{z_{i_k}^{s_k}}(x^*) \rangle = 0$. So, $x^* \in S_{i_k}^{s_k}$ for each k, as claimed.

F Extending the $\tilde{\Omega}(\zeta)$ lower bound: details from Section 3

F.1 Hadamard manifolds of bounded curvature

In this section, we prove the following generalization of Theorem 7.

Theorem 44 (Generalization of Theorem 7) *Let $\mathcal{M}$ be a Hadamard manifold of dimension $d \geq 2$ whose sectional curvatures are in the interval $[K_{\text{lo}}, K_{\text{up}}]$ with $K_{\text{up}} < 0$. Let $\tilde{L} > 0$ and $r \geq \frac{64}{\sqrt{-K_{\text{up}}}}$. Define*

$$\hat{\epsilon} = \frac{1}{2^{10} r \sqrt{-K_{\text{lo}}}}, \quad \epsilon = \frac{1}{2^{18} \zeta_{r\sqrt{-K_{\text{lo}}}} \log^2(\zeta_{r\sqrt{-K_{\text{lo}}}})}, \quad \epsilon' = \frac{1}{2^{18} \log^2(\zeta_{r\sqrt{-K_{\text{lo}}}})},$$

and $L = \frac{\tilde{L}}{2^{10}\log^2(r\sqrt{-K_{\mathrm{lo}}})}, \tilde{M} = \frac{\tilde{L}}{2\sqrt{-K_{\mathrm{lo}}}}$. *Let $\mathcal{A}$ be any deterministic algorithm.*

There is a C^∞ function $\tilde{f} \in \mathcal{F}^{x_{\mathrm{ref}}}_{r,\tilde{M},0,\tilde{L}}$ with unique minimizer x^* such that running $\mathcal{A}$ on $\tilde{f}$ yields iterates $x_0, x_1, x_2, \ldots$ satisfying

$$\tilde{f}(x_k) - \tilde{f}(x^*) \geq 2\hat{\epsilon}Lr^2 \geq \epsilon' \cdot \tilde{M}r \geq \epsilon \cdot \frac{1}{2}\tilde{L}r^2$$

for all $k = 0, 1, \ldots, T-1$, where $T = \left\lfloor \frac{\zeta_r\sqrt{-K_{\mathrm{up}}}}{50\log\left(64\zeta_r\sqrt{-K_{\mathrm{lo}}}\right)} \right\rfloor$.

Towards this end, we have the following generalization of Lemma 6.

Lemma 45 (Generalization of Lemma 6) *Let $\mathcal{M}$ be a Hadamard manifold of dimension $d \geq 2$ which satisfies the ball-packing property (see A1 in (Criscitiello and Boumal, 2022a)) with constants $\tilde{r}, \tilde{c}$ and point $x_{\mathrm{ref}} \in \mathcal{M}$. Also assume $\mathcal{M}$ has sectional curvatures in the interval $[K_{\mathrm{lo}}, 0]$ with $K_{\mathrm{lo}} < 0$. Let $L > 0$ and $r \geq \max\left\{\tilde{r}, \frac{8}{\sqrt{-K_{\mathrm{lo}}}}, \frac{4(d+2)}{\tilde{c}}\right\}$. Define $\hat{\epsilon} = \frac{1}{2^{10}r\sqrt{-K_{\mathrm{lo}}}}$ and $\mu = 64\hat{\epsilon}L = \frac{L}{2^4 r\sqrt{-K_{\mathrm{lo}}}}$. Let $\mathcal{A}$ be any deterministic algorithm.*

There is a C^∞ function f with minimizer $x^* \in B(x_{\mathrm{ref}}, \frac{3}{4}r)$ such that running $\mathcal{A}$ on f yields iterates $x_0, x_1, x_2, \ldots$ satisfying $f(x_k) - f(x^*) \geq 2\hat{\epsilon}Lr^2$ for all $k = 0, 1, \ldots, T-1$, where

$$T = \left\lfloor \frac{\tilde{c}(d+2)^{-1}r}{\log\left(2 \cdot 10^6 \cdot \tilde{c}(d+2)^{-1}r(r\sqrt{-K_{\mathrm{lo}}})^2\right)} \right\rfloor. \tag{24}$$

Moreover, f is μ-strongly g-convex in $\mathcal{M}$, and $\mu(12\mathscr{R} + \frac{3}{\sqrt{-K_{\mathrm{lo}}}})$-Lipschitz and $\mu(12\mathscr{R}\sqrt{-K_{\mathrm{lo}}} + 9)$-smooth in the ball $B(x_{\mathrm{ref}}, \mathscr{R})$, where $\mathscr{R} = 2^9 r\log(r\sqrt{-K_{\mathrm{lo}}})^2$. Outside of the ball $B(x_{\mathrm{ref}}, \mathscr{R})$, $f(x) = 3\mu\mathrm{dist}(x, x_{\mathrm{ref}})^2$ for all $x \notin B(x_{\mathrm{ref}}, \mathscr{R})$.

Proof Due to the assumed lower bound on r, we can apply Theorem 24 of (Criscitiello and Boumal, 2022a) to $\mathcal{A}$. After scaling by 6μ,[15] that theorem provides a function f which is μ-strongly g-convex in $\mathcal{M}$, and $\mu(12\mathscr{R} + \frac{3}{\sqrt{-K_{\mathrm{lo}}}})$-Lipschitz and $\mu(12\mathscr{R}\sqrt{-K_{\mathrm{lo}}} + 9)$-smooth in $B(x_{\mathrm{ref}}, \mathscr{R})$ with minimizer $x^* \in B(x_{\mathrm{ref}}, \frac{3}{4}r)$ such that

$$\mathrm{dist}(x_k, x^*) \geq \frac{r}{4} \qquad \forall k \leq T-1,$$

where $x_0, x_1, \ldots$ denote the queries produced by running $\mathcal{A}$ on f. Using $\mathrm{dist}(x_k, x^*) \geq \frac{r}{4}$,

$$2\hat{\epsilon} \cdot Lr^2 = \frac{1}{32}\mu r^2 \leq \frac{\mu}{2}\mathrm{dist}(x_k, x^*)^2 \leq f(x_k) - f(x^*), \qquad \forall k \leq T-1.$$

For the last inequality we used the μ-strong g-convexity of f. ∎

Applying the exact same proof given in Section 3 for Theorem 7 to Lemma 45, we conclude the following theorem (which is also a generalization of Theorem 7).

---

15. Scaling by 6μ is legitimate by a simple reduction argument: given an algorithm $\mathcal{A}$, we can create an algorithm $\hat{\mathcal{A}}$ which internally runs $\mathcal{A}$ to query an oracle $\mathcal{O}_{\hat{f}}$ and multiplies the outputs of $\mathcal{O}_{\hat{f}}$ by 6μ before forwarding them to $\mathcal{A}$. We call upon Theorem 24 of (Criscitiello and Boumal, 2022a) to claim there exists a hard function $\hat{f}$ for $\hat{\mathcal{A}}$. Since $\mathcal{A}$ is effectively interacting with $\mathcal{O}_{6\mu\hat{f}}$, we find $f = 6\mu\hat{f}$ is a hard function for $\mathcal{A}$.

Theorem 46 *Let $\mathcal{M}$ be a Hadamard manifold of dimension $d \geq 2$ which satisfies the ball-packing property with constants $\tilde{r}, \tilde{c}$ and point $x_{\mathrm{ref}} \in \mathcal{M}$. Also assume $\mathcal{M}$ has sectional curvatures in the interval $[K_{\mathrm{lo}}, 0]$ with $K_{\mathrm{lo}} < 0$. Let $\tilde{L} > 0$ and $r \geq \max\left\{\tilde{r}, \frac{8}{\sqrt{-K_{\mathrm{lo}}}}, \frac{4(d+2)}{\tilde{c}}\right\}$. Define*

$$\hat{\epsilon} = \frac{1}{2^{10} r \sqrt{-K_{\mathrm{lo}}}}, \quad \epsilon = \frac{1}{2^{18} \zeta_{r\sqrt{-K_{\mathrm{lo}}}} \log^2(\zeta_{r\sqrt{-K_{\mathrm{lo}}}})}, \quad \epsilon' = \frac{1}{2^{18} \log^2(\zeta_{r\sqrt{-K_{\mathrm{lo}}}})},$$

and $L = \frac{\tilde{L}}{2^{10} \log^2(r\sqrt{-K_{\mathrm{lo}}})}, \tilde{M} = \frac{\tilde{L}}{2\sqrt{-K_{\mathrm{lo}}}}$. Let $\mathcal{A}$ be any deterministic algorithm.

There is a C^∞ function $\tilde{f} \in \mathcal{F}_{r,\tilde{M},0,\tilde{L}}$ with unique minimizer x^* such that running $\mathcal{A}$ on $\tilde{f}$ yields iterates $x_0, x_1, x_2, \ldots$ satisfying

$$\tilde{f}(x_k) - \tilde{f}(x^*) \geq 2\hat{\epsilon} L r^2 \geq \epsilon' \cdot \tilde{M} r \geq \epsilon \cdot \frac{1}{2}\tilde{L} r^2$$

for all $k = 0, 1, \ldots, T - 1$, where T is given by equation (24).

Theorem 44 now follows directly from Theorem 46 by using Lemma 7 of (Criscitiello and Boumal, 2022a) for the values of $\tilde{r}$ and $\tilde{c}$ in the ball-packing property.

F.2 Defining the function $u_{\mathscr{R}}\colon \mathbb{R} \to \mathbb{R}$

Define the C^∞ function $u_{\mathscr{R}}\colon \mathbb{R} \to \mathbb{R}$ by $u_{\mathscr{R}}(\mathscr{D}) = 1$ for all $\mathscr{D} \leq \frac{1}{2}\mathscr{R}^2$ and $u_{\mathscr{R}}(\mathscr{D}) = 1 - e^{-4/\sqrt{2\mathscr{D}/\mathscr{R}^2 - 1}}$ for all $\mathscr{D} > \frac{1}{2}\mathscr{R}^2$. That $u_{\mathscr{R}}$ is C^∞ can be verified using the same method used in the proof of Lemma 2.20 from (Lee, 2012, Ch. 2).

F.3 Verifying $\tilde{f}_{\mathscr{R}}$ is Lipschitz, smooth and strictly g-convex in $\mathcal{M}$

Let $\mathcal{M}$ be Hadamard manifold with sectional curvatures in $[K_{\mathrm{lo}}, 0]$. To complete the proof of Theorem 7 from Section 3 (or its generalization Theorem 46), we show that given f with $f(x) = 3\mu\mathrm{dist}(x, x_{\mathrm{ref}})^2$ for $x \notin B(x_{\mathrm{ref}}, \mathscr{R})$, the function $\tilde{f}_{\mathscr{R}}$ defined by (1) is strictly g-convex, $12\mu\mathscr{R}$-Lipschitz, and $24\mu\mathscr{R}\sqrt{-K_{\mathrm{lo}}}$-smooth outside of the ball $B(x_{\mathrm{ref}}, \mathscr{R})$.

Let $\gamma(t)$ be a geodesic with $\gamma(0) = x, \gamma'(0) = v$ and $\|v\| = 1$. For the moment, define $\mathscr{D}(x) = \frac{1}{2}\mathrm{dist}(x, x_{\mathrm{ref}})^2$. For the gradient, we have

$$\begin{aligned}
\left\langle \nabla \tilde{f}_{\mathscr{R}}(x), v \right\rangle &= \frac{d}{dt}[\tilde{f}(\gamma(t))]_{t=0} = \frac{d}{dt}[u_{\mathscr{R}}(\mathscr{D}(\gamma(t)))f(\gamma(t))]_{t=0} \\
&= u_{\mathscr{R}}(\mathscr{D}(x))\frac{d}{dt}[f(\gamma(t))]_{t=0} + \frac{d}{dt}[u_{\mathscr{R}}(\mathscr{D}(\gamma(t)))]_{t=0}f(x) \\
&= u_{\mathscr{R}}(\mathscr{D}(x))\left\langle \nabla f(x), v \right\rangle + f(x)u_{\mathscr{R}}'(\mathscr{D}(x))\frac{d}{dt}[\mathscr{D}(\gamma(t))]_{t=0} \\
&= u_{\mathscr{R}}(\mathscr{D}(x))\left\langle \nabla f(x), v \right\rangle - f(x)u_{\mathscr{R}}'(\mathscr{D}(x))\left\langle \log_x(x_{\mathrm{ref}}), v \right\rangle
\end{aligned}$$

which gives us the formula

$$\nabla \tilde{f}_{\mathscr{R}}(x) = \begin{cases} \nabla f(x) & \text{if } \mathrm{dist}(x, x_{\mathrm{ref}}) \leq \mathscr{R}; \\ u_{\mathscr{R}}(\mathscr{D}(x))\nabla f(x) \\ \quad - f(x)u_{\mathscr{R}}'(\mathscr{D}(x))\exp_x^{-1}(x_{\mathrm{ref}}) & \text{otherwise.} \end{cases} \tag{25}$$

37

By Lemma 50, we also see that $\|\nabla \tilde{f}_{\mathscr{R}}(x)\| \le 12\mu\mathscr{R}$ if $\mathscr{D}(x) > \frac{1}{2}\mathscr{R}^2$.

For the Hessian we have:

$$
\begin{aligned}
\left\langle v, \nabla^2 \tilde{f}_{\mathscr{R}}(x)v \right\rangle &= \frac{d^2}{dt^2}[\tilde{f}_{\mathscr{R}}(\gamma(t))]_{t=0} = \frac{d^2}{dt^2}[u_{\mathscr{R}}(\mathscr{D}(\gamma(t)))f(\gamma(t))]_{t=0} \\
&= \frac{d^2}{dt^2}[u_{\mathscr{R}}(\mathscr{D}(\gamma(t)))]_{t=0}f(x) + u_{\mathscr{R}}(\mathscr{D}(x))\frac{d^2}{dt^2}[f(\gamma(t))]_{t=0} \\
&\quad + 2\frac{d}{dt}[u_{\mathscr{R}}(\mathscr{D}(\gamma(t)))]_{t=0}\frac{d}{dt}[f(\gamma(t))]_{t=0} \\
&= f(x)u''_{\mathscr{R}}(\mathscr{D}(x))\left(\frac{d}{dt}[\mathscr{D}(\gamma(t))]_{t=0}\right)^2 + f(x)u'_{\mathscr{R}}(\mathscr{D}(x))\frac{d^2}{dt^2}[\mathscr{D}(\gamma(t))]_{t=0} \\
&\quad + u_{\mathscr{R}}(\mathscr{D}(x))\frac{d^2}{dt^2}[f(\gamma(t))]_{t=0} + 2\frac{d}{dt}[u_{\mathscr{R}}(\mathscr{D}(\gamma(t)))]_{t=0}\frac{d}{dt}[f(\gamma(t))]_{t=0}.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\left\langle v, \nabla^2 \tilde{f}_{\mathscr{R}}(x)v \right\rangle &= f(x)u''_{\mathscr{R}}(\mathscr{D}(x))\left\langle \log_x(x_{\text{ref}}), v \right\rangle^2 + f(x)u'_{\mathscr{R}}(\mathscr{D}(x))\left\langle v, \nabla^2\mathscr{D}(x)v \right\rangle \\
&\quad + u_{\mathscr{R}}(\mathscr{D}(x))\frac{d^2}{dt^2}[f(\gamma(t))]_{t=0} - 2\left\langle \log_x(x_{\text{ref}}), v \right\rangle \left\langle \nabla f(x), v \right\rangle \\
&= 6\mu\Bigg([u_{\mathscr{R}}(\mathscr{D}(x)) + \mathscr{D}(x)u'_{\mathscr{R}}(\mathscr{D}(x))]\left\langle v, \nabla^2\mathscr{D}(x)v \right\rangle \\
&\qquad + [2u'_{\mathscr{R}}(\mathscr{D}(x)) + \mathscr{D}(x)u''_{\mathscr{R}}(\mathscr{D}(x))]\left\langle \log_x(x_{\text{ref}}), v \right\rangle^2 \Bigg).
\end{aligned}
$$

Lemmas 47 and 48 from Appendix F.4 shows that $u_{\mathscr{R}}(\mathscr{D}) + \mathscr{D}u'_{\mathscr{R}}(\mathscr{D}) \ge 0$ and $2u'_{\mathscr{R}}(\mathscr{D}) + \mathscr{D}u''_{\mathscr{R}}(\mathscr{D}) \le 0$ for all $\mathscr{D} > \frac{1}{2}\mathscr{R}^2$. Therefore, $\left\langle v, \nabla^2 \tilde{f}_{\mathscr{R}}(x)v \right\rangle$ is at least

$$
6\mu\Bigg([u_{\mathscr{R}}(\mathscr{D}(x)) + \mathscr{D}(x)u'_{\mathscr{R}}(\mathscr{D}(x))] + [2u'_{\mathscr{R}}(\mathscr{D}(x)) + \mathscr{D}(x)u''_{\mathscr{R}}(\mathscr{D}(x))]2\mathscr{D} \Bigg).
$$

Lemma 49 shows that this is strictly greater than zero, verifying strict g-convexity.

Likewise, Lemma 50 along with Lemma 5 imply $\left\langle v, \nabla^2 \tilde{f}_{\mathscr{R}}(x)v \right\rangle$ is at most

$$
6\mu\Bigg([u_{\mathscr{R}}(\mathscr{D}(x)) + \mathscr{D}(x)u'_{\mathscr{R}}(\mathscr{D}(x))]2\sqrt{-K_{\text{lo}}}\sqrt{2\mathscr{D}} \Bigg) \le 24\mu\sqrt{-K_{\text{lo}}}\mathscr{R}.
$$

F.4 Technical facts about the function $u_{\mathscr{R}} \colon \mathbb{R} \to \mathbb{R}$

In the following lemmas, we use the change of variables

$$
\tau = 1/\sqrt{\frac{2\mathscr{D}}{\mathscr{R}^2} - 1} \iff \mathscr{D} = \mathscr{R}^2\frac{1+\tau^2}{2\tau^2}.
$$

for $\mathscr{D} \in (\frac{1}{2}\mathscr{R}^2, \infty)$ and $\tau \in (0, \infty)$.

Lemma 47 $u_{\mathscr{R}}(\mathscr{D}) + \mathscr{D}u'_{\mathscr{R}}(\mathscr{D}) \geq 0, \quad \forall \mathscr{D} \in (\frac{1}{2}\mathscr{R}^2, \infty).$

Proof Computing we find:

$$u_{\mathscr{R}}(\mathscr{D}) + \mathscr{D}u'_{\mathscr{R}}(\mathscr{D}) = 1 - e^{-4\tau}(1 + 2\tau + 2\tau^3).$$

On the other hand, we know $1 + 4\tau + 8\tau^2 + \frac{32}{3}\tau^3 \geq 1 + 2\tau + 2\tau^3$ for all $\tau \geq 0$ (which can be verified with Mathematica), so

$$e^{4\tau} = \sum_{j=0}^{\infty} \frac{1}{j!}(4\tau)^j \geq \sum_{j=0}^{3} \frac{1}{j!}(4\tau)^j = 1 + 4\tau + 8\tau^2 + \frac{32}{3}\tau^3 \geq 1 + 2\tau + 2\tau^3, \quad \forall \tau \geq 0,$$

which allows us to conclude. ∎

Lemma 48 $2u'_{\mathscr{R}}(\mathscr{D}) + \mathscr{D}u''_{\mathscr{R}}(\mathscr{D}) \leq 0, \quad \forall \mathscr{D} \in (\frac{1}{2}\mathscr{R}^2, \infty).$

Proof Computing we find:

$$2u'_{\mathscr{R}}(\mathscr{D}) + \mathscr{D}u''_{\mathscr{R}}(\mathscr{D}) = 2\mathscr{R}^{-2}e^{-4\tau}\tau^3(-1 - 4\tau + 3\tau^2 - 4\tau^3)$$

and one can check that $-1 - 4\tau + 3\tau^2 - 4\tau^3 \leq 0$ for all $\tau \geq 0$. ∎

Lemma 49 $[u_{\mathscr{R}}(\mathscr{D}) + \mathscr{D}u'_{\mathscr{R}}(\mathscr{D})] + [2u'_{\mathscr{R}}(\mathscr{D}) + \mathscr{D}u''_{\mathscr{R}}(\mathscr{D})]2\mathscr{D} > 0, \quad \forall \mathscr{D} \in (\frac{1}{2}\mathscr{R}^2, \infty).$

Proof Computing we find:

$$[u_{\mathscr{R}}(\mathscr{D}) + \mathscr{D}u'_{\mathscr{R}}(\mathscr{D})] + [2u'_{\mathscr{R}}(\mathscr{D}) + \mathscr{D}u''_{\mathscr{R}}(\mathscr{D})]2\mathscr{D} = 1 - e^{-4\tau}(1 + 4\tau + 8\tau^2 - 2\tau^3 + 16\tau^4 - 6\tau^5 + 8\tau^6).$$

On the other hand, we know $\sum_{j=0}^{7} \frac{1}{j!}(4\tau)^j > 1 + 4\tau + 8\tau^2 - 2\tau^3 + 16\tau^4 - 6\tau^5 + 8\tau^6$ for all $\tau > 0$ (which can be verified with Mathematica), which allows us to conclude. ∎

Lemma 50 $(u_{\mathscr{R}}(\mathscr{D}) + \mathscr{D}u'_{\mathscr{R}}(\mathscr{D}))2\sqrt{2\mathscr{D}} \leq 4\mathscr{R}, \quad \forall \mathscr{D} \in (\frac{1}{2}\mathscr{R}^2, \infty).$

Proof This follows from a calculation similar to the proof of Lemma 47. ∎

F.5   Technical fact from Section 3: $x_k \notin B(x_{\text{ref}}, \mathscr{R})$ implies $\tilde{f}_{\mathscr{R}}(x_k) - \tilde{f}_{\mathscr{R}}(x^*) \geq 2\hat{\epsilon}Lr^2$

Recall Lemma 6 provides a $\mu = 64\hat{\epsilon}L$-strongly g-convex function f, with minimizer x^*, satisfying $f(x_k) - f(x^*) \geq 2\hat{\epsilon}Lr^2$ for all $k \leq T - 1$.

If $x_k \notin B(x_{\text{ref}}, \mathscr{R})$, then consider the geodesic segment $\gamma_{x^* \to x_k} : [0, 1] \to \mathcal{M}$ given by $\gamma_{x^* \to x_k}(t) = \exp_{x^*}(t \log_{x^*}(x_k))$. Continuity of $t \mapsto \text{dist}(\gamma_{x^* \to x_k}(t), x_{\text{ref}})$ implies there exists a $t \in (0, 1)$ so that $\mathscr{R} = \text{dist}(\gamma_{x^* \to x_k}(t), x_{\text{ref}})$. Let $y = \gamma(t)$. Geodesic convexity of $\tilde{f}_{\mathscr{R}}$ implies

$$\tilde{f}_{\mathscr{R}}(y) \leq (1 - t)\tilde{f}_{\mathscr{R}}(x^*) + t\tilde{f}_{\mathscr{R}}(x_k) \leq (1 - t)\tilde{f}_{\mathscr{R}}(y) + t\tilde{f}_{\mathscr{R}}(x_k),$$

which implies $\tilde{f}_{\mathscr{R}}(y) \leq \tilde{f}_{\mathscr{R}}(x_k)$. On the other hand, we know $\tilde{f}_{\mathscr{R}} = f$ in $B(x_{\text{ref}}, \mathscr{R})$. Therefore $\tilde{f}_{\mathscr{R}}$ is μ-strongly g-convex in $B(x_{\text{ref}}, \mathscr{R})$, so

$$\tilde{f}_{\mathscr{R}}(x_k) - \tilde{f}_{\mathscr{R}}(x^*) \geq \tilde{f}_{\mathscr{R}}(y) - \tilde{f}_{\mathscr{R}}(x^*) \geq \frac{\mu}{2}\text{dist}(y, x^*)^2 \geq \frac{\mu}{2}(\mathscr{R} - r)^2 \geq 32\hat{\epsilon}L(2^{11}r)^2 \geq 2\hat{\epsilon}Lr^2.$$

G Properties of the Moreau envelope: Proof of Lemma 10

Azagra and Ferrera (2005, Cor. 4.5) shows that f_λ is g-convex and C^1. The function $y \mapsto f(y) + \frac{1}{2\lambda}\mathrm{dist}(x,y)^2$ is strongly g-convex and so has a unique minimizer which we denote

$$y_\lambda(x) = \arg\min_{y\in\mathcal{M}}\{f(y) + \frac{1}{2\lambda}\mathrm{dist}(x,y)^2\}.$$

First-order optimality conditions imply $\nabla f(y_\lambda(x)) = \frac{1}{\lambda}\log_{y_\lambda(x)}(x)$. Since, f is 1-Lipschitz, this implies $\mathrm{dist}(x,y_\lambda(x)) = \|\log_{y_\lambda(x)}(x)\| \leq \lambda$. This proves $f_\lambda(x) = \min_{y\in B(x,\lambda)}\{f(y) + \frac{1}{2\lambda}\mathrm{dist}(x,y)^2\}$. Next, using the g-convexity and 1-Lipschitzness of f, we have

$$f(x) \geq f_\lambda(x) \geq f(y_\lambda(x)) \geq f(x) + \langle\nabla f(x), \log_x(y_\lambda(x))\rangle \geq f(x) - \mathrm{dist}(x,y_\lambda(x)) \geq f(x) - \lambda.$$

To compute the Lipschitz constant of ∇f_λ, we follow (but slightly simplify) the proof of (Azagra and Ferrera, 2015, Prop. 7.1). Let $\delta > 0$. By Theorem 1.5 of (Azagra and Ferrera, 2015) (which provides a useful characterization of Lipschitz gradient), it suffices to show that for each $x_0 \in \mathcal{M}$ the function $x \mapsto f_\lambda(x) - \frac{\zeta_{\lambda+2\delta}}{2\lambda}\mathrm{dist}(x,x_0)^2$ is g-concave when restricted to $B(x_0,\delta)$. For $x \in B(x_0,\delta)$,

$$f_\lambda(x) - \frac{\zeta_{\lambda+2\delta}}{2\lambda}\mathrm{dist}(x,x_0)^2 = \min_{y\in B(x,\lambda)}\{f(y) + \frac{1}{2\lambda}\mathrm{dist}(x,y)^2\} - \frac{\zeta_{\lambda+2\delta}}{2\lambda}\mathrm{dist}(x,x_0)^2$$

$$= \min_{y\in B(x_0,\delta+\lambda)}\{f(y) + \frac{1}{2\lambda}\mathrm{dist}(x,y)^2 - \frac{\zeta_{\lambda+2\delta}}{2\lambda}\mathrm{dist}(x,x_0)^2\}.$$

As the minimum of g-concave functions is g-concave, it suffices to show that for each $y \in B(x_0,\delta+\lambda)$, the function $\tilde{d}(x) = \frac{1}{2}\mathrm{dist}(x,y)^2 - \frac{\zeta_{\lambda+2\delta}}{2}\mathrm{dist}(x,x_0)^2$ is g-concave when restricted to $B(x_0,\delta)$. This is easy to see by looking at the maximum eigenvalue of its Hessian. For $x \in B(x_0,\delta)$,

$$\lambda_{max}(\nabla^2\tilde{\mathrm{dist}}(x)) \leq \zeta_{\mathrm{dist}(x,y)} - \zeta_{\lambda+2\delta} \leq 0$$

using that $\mathrm{dist}(x,y) \leq \delta + \delta + \lambda = 2\delta + \lambda$. This concludes the proof of Lipschitz gradient.

Finally, to see that f_λ is 1-Lipschitz, note that $\nabla f_\lambda(x) = -\frac{1}{\lambda}\log_x(y_\lambda(x))$ for all $x \in \mathcal{M}$ (Azagra and Ferrera, 2005, Prop. 3.7), so $\|\nabla f_\lambda(x)\| = \frac{1}{\lambda}\mathrm{dist}(x,y_\lambda(x)) \leq 1$.

H The relation between ζ, ϵ and κ: proofs from Section 5.1

The following proposition states if f is gradient-Lipschitz then it is Lipschitz. The proof is due to (Petrunin, 2023). We just fill in the details.

Proposition 51 *Let $p \in \mathcal{M} = \mathbb{H}^d$, and assume $f\colon \mathcal{M} \to \mathbb{R}$ is differentiable and L-smooth in $B(p,2)$. Then $\|\nabla f(p)\| \leq \frac{\pi\sqrt{5}L}{2} \leq 4L$.*

Remark 52 *It is necessary that we assume f is L-smooth in a region which is not too small, because we can construct functions with small Hessian and arbitrarily large gradient in a sufficiently small region by pushing forward functions defined in a tangent space via the exponential map, as in Appendix D.*

Proof [Proof of Proposition 51] Take a 2-dimensional totally geodesic submanifold S which passes through p and is tangent to $\nabla f(p)$, i.e., $\nabla f(p) \in T_p S$. Let $\hat{f}$ be the restriction of f to S. Note that $\hat{f}$ is L-smooth in $S \cap B(p, 2)$.[16] Consider a circle $C \subseteq S \cap B(p, 2)$ of radius $R = \operatorname{arccosh}(3/2) < 1$ containing p. Let $c \colon [0, T] \to C$, $T = 2\pi \sinh(R)$, be an arc-length parameterization of C with $c(0) = c(T) = p$, and let $P_{t \to s}^c$ denote parallel transport (in S) along c from $c(t)$ to $c(s)$.

For the moment, assume that f is twice differentiable. Observe that

$$\frac{d}{dt}[P_{t \to 0}^c \nabla \hat{f}(c(t))] = P_{t \to 0}^c D_t \nabla \hat{f}(c(t)) = P_{t \to 0}^c \nabla^2 \hat{f}(c(t)) c'(t),$$

where D_t is the covariant derivative along γ. Therefore,

$$\|P_{T \to 0}^c \nabla \hat{f}(p) - \nabla \hat{f}(p)\| = \|\int_0^T P_{t \to 0}^c \nabla^2 \hat{f}(c(t)) c'(t) dt\| \le LT. \tag{26}$$

On the other hand, $P_{T \to 0}^c \nabla \hat{f}(p)$ is a rotation of $\nabla \hat{f}(p)$ by some angle θ, and the Gauss-Bonnet Formula (Lee, 2018, Ch. 9) states that $\theta = \int K dA = 2\pi - \int_C k_g = -2\pi(\cosh(R) - 1)$, where $\int K dA$ is the integral of the curvature $K = -1$ over the interior of C (a disk), and $k_g = \coth(R)$ is the geodesic curvature of C. Since $R = \operatorname{arccosh}(3/2)$, we know $\theta = -\pi$, and $\|P_{T \to 0}^c \nabla \hat{f}(p) - \nabla \hat{f}(p)\| = 2\|\nabla \hat{f}(p)\|$. We conclude that $\|\nabla f(p)\| = \|\nabla \hat{f}(p)\| \le LT/2 = L\pi \sinh(R) = L\pi \sqrt{5}/2 \le 4L$.

If f is not twice differentiable, we can still conclude the result in a similar way. Place N equally spaced points around C, $p_n = c(t_n), t_n = \frac{nT}{N}$ for $n = 0, \ldots, N$. Let $\alpha = \frac{\pi}{N}$. Let γ_n denote the geodesic segment between p_n, p_{n+1}. Let Γ_N denote the geodesic polygon formed by the segments γ_n. Let $P_{n \to m}^{\Gamma}, m \le n$, denote parallel transport (in S) from p_n to p_m along the geodesic segments $\gamma_n, \gamma_{n-1}, \ldots, \gamma_m$. Note that $P_{n \to 0}^{\Gamma} = P_{n-1 \to 0}^{\Gamma} P_{n \to n-1}^{\Gamma}$. The L-smoothness of $\hat{f}$ implies, for $n = 0, \ldots, N$,

$$\|P_{n \to 0}^{\Gamma} \nabla \hat{f}(p_n) - \nabla \hat{f}(p)\|$$
$$= \|P_{n-1 \to 0}^{\Gamma} P_{n \to n-1}^{\Gamma} \nabla \hat{f}(p_n) - P_{n-1 \to 0}^{\Gamma} \nabla \hat{f}(p_{n-1}) + P_{n-1 \to 0}^{\Gamma} \nabla \hat{f}(p_{n-1}) - \nabla \hat{f}(p)\|$$
$$\le \|P_{n \to n-1}^{\Gamma} \nabla \hat{f}(p_n) - \nabla \hat{f}(p_{n-1})\| + \|P_{n-1 \to 0}^{\Gamma} \nabla \hat{f}(p_{n-1}) - \nabla \hat{f}(p)\|$$
$$\le L\operatorname{dist}(p_n, p_{n-1}) + \|P_{n-1 \to 0}^{\Gamma} \nabla \hat{f}(p_{n-1}) - \nabla \hat{f}(p)\|.$$

Therefore, $\|P_{n \to 0}^{\Gamma} \nabla \hat{f}(p_n) - \nabla \hat{f}(p)\| \le \sum_{n=0}^{n-1} L\operatorname{dist}(p_n, p_{n-1}) \le LT$, and so for all N,

$$\|P_{N \to 0}^{\Gamma} \nabla \hat{f}(p) - \nabla \hat{f}(p)\| \le LT.$$

On the other hand, $P_{N \to 0}^{\Gamma} \nabla \hat{f}(p)$ is a rotation of $\nabla \hat{f}(p)$ by some angle θ_N, and the Gauss-Bonnet Formula (Lee, 2018, Ch. 9) states that $\theta_N = \int K dA$, where $\int K dA$ is the integral of the curvature $K = -1$ over the interior of the polygon Γ. The area of that polygon equals $2N = \frac{2}{\pi \alpha}$ times the area of a right hyperbolic triangle with angle α and hypotenuse R. Let $\beta \in (0, \frac{\pi}{2})$ be the other angle in this triangle which is not a right angle. Using hyperbolic trigonometry, we find

$$\beta = \arcsin\left(\frac{1}{\sinh(R)} \sinh\left(\operatorname{arctanh}(\tanh(R) \cos(\alpha))\right)\right).$$

---

16. This follows from: if $\mathcal{P}_x$ denotes orthogonal projection from $T_x \mathcal{M}$ to $T_x S$, one can check that $\mathcal{P}_y P_{x \to y} = P_{x \to y} \mathcal{P}_x$.

The area of this triangle equals $\pi - \frac{\pi}{2} - \alpha - \beta$, and so we find that

$$| - \pi - \theta_N| = |-\pi - K \cdot \frac{2\pi}{\alpha} \cdot (\pi - \frac{\pi}{2} - \alpha - \beta)| \leq \frac{5\pi}{4}\alpha^2, \quad \forall \alpha \in (0, \frac{\pi}{2}].$$

Hence, taking N large enough (i.e., α small enough), $\|P^\Gamma_{N\to0}\nabla\hat{f}(p) - \nabla\hat{f}(p)\|$ is arbitrarily close to $2\|\nabla\hat{f}(p)\|$. Yet, $\|P^\Gamma_{N\to0}\nabla\hat{f}(p) - \nabla\hat{f}(p)\| \leq LT$, and we conclude as in the case where f is twice differentiable. ∎

Proof [Proof of Proposition 13] Assume $r > 2$. By Proposition 51, $\|\nabla f(p)\| \leq \frac{\pi\sqrt{5}L}{2}$ for all $p \in B(x_{\text{ref}}, r - 2)$. If $\text{dist}(x^*, x_{\text{ref}}) \leq r - 2$, then we know that $f(x_{\text{ref}}) - f^* \leq \frac{\pi\sqrt{5}L}{2}(r - 2)$. Otherwise, define $x' = \exp_{x_{\text{ref}}}\left(\frac{r-2}{\text{dist}(x^*,x_{\text{ref}})}\log_{x_{\text{ref}}}(x^*)\right)$. Then $f(x_{\text{ref}}) - f(x') \leq \frac{\pi\sqrt{5}L}{2}(r-2)$ and $\|\nabla f(x')\| \leq \frac{\pi\sqrt{5}L}{2}$. So by L-smoothness of f,

$$f^* = f(x^*) \geq f(x') + \langle\nabla f(x'), \log_{x'}(x^*)\rangle - \frac{L}{2}\text{dist}(x', x^*)^2 \geq f(x') - \frac{\pi\sqrt{5}L}{2}\cdot 2 - \frac{L}{2}\cdot 4$$

and so, if $r > 2$,

$$f(x_{\text{ref}}) - f^* \leq f(x_{\text{ref}}) - f(x') + f(x') - f^* \leq \frac{\pi\sqrt{5}L}{2}(r-2) + \pi\sqrt{5}L + 2L \leq \frac{\pi\sqrt{5}L}{2}r + 2L.$$

We conclude, for all $r > 0$, $f(x_{\text{ref}}) - f^* \leq \min\left\{\frac{\pi\sqrt{5}L}{2}r + 2L, \frac{1}{2}Lr^2\right\} \leq \frac{8}{\zeta_r}\cdot\frac{1}{2}Lr^2$. ∎

To prove the bound $\kappa \geq \zeta_r$ Martínez-Rubio (2021) and Hamilton and Moitra (2021) use that the domain is a ball. We can use Proposition 51 to prove $\kappa \geq \Omega(\zeta_r)$ if the domain is not too *eccentric*.

Proposition 53 *Let $f\colon \mathbb{H}^d \to \mathbb{R}$ be differentiable, μ-strongly g-convex globally, and suppose $\nabla f(x^*) = 0$. Let $r > 2$, and consider a g-convex subset D containing x^* such that*

1. there exists $\tilde{D} \subseteq D$ such that $B(x, 2) \subseteq D$ for all $x \in \tilde{D}$, and

2. there is a $\tilde{x} \in \tilde{D}$ so that $\text{dist}(\tilde{x}, x^*) \geq r - 2$.

(For example, $D = B(x^*, r)$ satisfies these assumptions.)
Let L be the Lipschitz constant of ∇f in D. Then $\kappa = \frac{L}{\mu} \geq \frac{1}{10}\zeta_r$.

Proof [Proof of Proposition 53] We know that $\|\nabla f(x)\| \leq \frac{\pi\sqrt{5}}{2}L$ for all $x \in \tilde{D}$, and so

$$\frac{\pi\sqrt{5}}{2}L\text{dist}(x, x^*) \geq f(x) - f^* \geq \frac{\mu}{2}\text{dist}(x, x^*)^2$$

for all $x \in \tilde{D}$. Using $\text{dist}(\tilde{x}, x^*) \geq r - 2$, $\frac{L}{\mu} \geq \max\{1, \frac{r-2}{\pi\sqrt{5}}\} \geq \frac{1}{10}\zeta_r$. ∎

Remark 54 *From the results in this section, the most natural scalings for smooth g-convex optimization on hyperbolic spaces are arguably not the ones given in P3 and P4. Instead, for P3 one should ask for x such that $f(x) - f^* \leq \epsilon \cdot \frac{1}{2\zeta_r}Lr^2$, and for P4 one should define a new condition number $\kappa' = \frac{L}{\mu\zeta_r}$. We do not do this here in order to maintain consistency with previous literature.*

Proposition 55 *If f is a C^3 g-convex function on a hyperbolic space whose Hessian vanishes at a point x then necessarily $\nabla f(x) = 0$.*

Proof Let $\nabla^3 f(x) \colon \mathrm{T}_x \mathcal{M} \times \mathrm{T}_x \mathcal{M} \times \mathrm{T}_x \mathcal{M} \to \mathbb{R}$ denote the third derivative of f at x (see (Boumal, 2023, Ex. 10.78) for a definition). As observed in (Criscitiello and Boumal, 2022b, Remark 3.2), the Ricci identity applied to ∇f implies

$$\|\nabla f(x)\| \leq 2\|\nabla^3 f(x)\|,$$

since hyperbolic space has constant sectional curvature -1.

Let $x \in \mathcal{M}$, $u, v \in \mathrm{T}_x \mathcal{M}$ and $\gamma(t) = \exp_x(tv)$. Define the parallel vector field U along γ as $U(t) = P_{x \to \gamma(t)} u$. By the definition of $\nabla^3 f$ (see (Boumal, 2023, Ex. 10.78)),

$$\nabla^3 f(u, u, v) = \frac{d}{dt}\Big[\langle U(t), \nabla^2 f(\gamma(t)) U(t)\rangle\Big]_{t=0}.$$

Since f is g-convex, we know that $\langle U(t), \nabla^2 f(\gamma(t)) U(t)\rangle \geq 0$ for all t. Since $\langle U(0), \nabla^2 f(\gamma(0)) U(0)\rangle = 0$, we know $\frac{d}{dt}\Big[\langle U(t), \nabla^2 f(\gamma(t)) U(t)\rangle\Big]_{t=0} = 0$. We conclude that $\nabla^3 f(u, u, v) = 0$ for all $u, v \in \mathrm{T}_x \mathcal{M}$. By symmetry of the first two arguments of $\nabla^3 f$, we conclude $\nabla^3 f(x) = 0$. Hence, $\|\nabla f(x)\| \leq 0$. ∎

Remark 56 *The assumption that f is three times differentiable in Proposition 55 is necessary. Indeed, let $y \in \mathcal{M}$, $g \in \mathrm{T}_y \mathcal{M}$ with $\|g\| = 1$, and consider the function*

$$f(x) = \langle g, \log_y(x)\rangle + \frac{\tau}{6}\mathrm{dist}(x, y)^3.$$

If τ is large enough (e.g., $\tau = 2$), then one can check, like in the proof of Proposition 40, that this function is globally g-convex. However, $\nabla^2 f(x) = 0$ and $\nabla f(x) = g$.

I Reduction between strongly and nonstrongly g-convex settings

Proposition 57 *Fix $\epsilon \in (0, 1)$, $L > 0$ and $r > 0$. Define*

$$\delta = \epsilon \cdot \frac{1}{2} L r^2, \qquad \tilde{\delta} = \frac{1}{2} r^2, \qquad \sigma = \frac{\epsilon L}{2}.$$

Let $\mathcal{A}$ be a first-order deterministic algorithm. There is a first-order deterministic algorithm $\mathcal{A}'$ such that: for all $f \in \mathcal{F}_{r,\infty,0,L}$, if we define

$$\tilde{f}(x) = \frac{1}{\sigma} f(x) + \frac{1}{2}\mathrm{dist}(x, x_{\mathrm{ref}})^2,$$

then $T_\delta(\mathcal{A}', f) \leq T_{\tilde{\delta}}(\mathcal{A}, \tilde{f})$ (recall Section 1.1).

In particular, if $r' \geq 2r$ and $\kappa = \frac{L}{\sigma} + \zeta_{r'} = \frac{2}{\epsilon} + \zeta_{r'}$, then $T_{\epsilon,r} \leq T_{\epsilon',r',\kappa}$ where $\epsilon' = \frac{r^2}{\kappa(r')^2}$.

Proof Let $f \in \mathcal{F}_{r,\infty,0,L}$ be an L-smooth g-convex function with minimizer $x^* \in B(x_{\text{ref}}, r)$ satisfying $\nabla f(x^*) = 0$. The function $\tilde{f}$ is 1-strongly g-convex in $\mathcal{M}$. Thus, it has a unique minimizer $\tilde{x}^* \in \mathcal{M}$ satisfying $\nabla \tilde{f}(\tilde{x}^*) = 0$ and $\text{dist}(\tilde{x}^*, x_{\text{ref}}) \leq 2\text{dist}(x^*, x_{\text{ref}}) \leq 2r$ because

$$\sigma(\text{dist}(\tilde{x}^*, x_{\text{ref}}) - \text{dist}(x^*, x_{\text{ref}})) \leq \sigma\text{dist}(x^*, \tilde{x}^*) \leq \left\| \nabla \tilde{f}(x^*) \right\|$$
$$= \left\| -\sigma \exp_{x^*}^{-1}(x_{\text{ref}}) \right\| = \sigma\text{dist}(x^*, x_{\text{ref}}).$$

We conclude that $\tilde{x}^* \in B(x_{\text{ref}}, 2r)$.

Let $\mathcal{A}'$ be the algorithm which runs $\mathcal{A}$ on the function $\tilde{f}$ (and so outputs the same queries as $\mathcal{A}$). Let $T = T_{\tilde{\delta}}(\mathcal{A}, \tilde{f})$ (we can assume T is finite otherwise there is nothing to prove). By assumption, algorithm $\mathcal{A}$ uses at most T queries to find a point $x \in \mathcal{M}$ with $\sigma(\tilde{f}(x) - \tilde{f}^*) \leq \sigma\tilde{\delta} = \sigma\frac{1}{2}r^2 = \frac{\epsilon}{4}Lr^2$, where $\tilde{f}^* = \tilde{f}(\tilde{x}^*)$. Let us show that in fact $f(x) - f^* \leq \epsilon \cdot \frac{1}{2}Lr^2$. Since

$$\sigma\tilde{f}^* \leq \sigma\tilde{f}(x^*) = f^* + \frac{\epsilon L}{4}\text{dist}(x^*, x_{\text{ref}})^2 \leq f^* + \frac{\epsilon L}{4}r^2$$

we have

$$f(x) - f^* - \frac{\epsilon L}{4}r^2 \leq f(x) - \sigma\tilde{f}^* \leq \sigma(\tilde{f}(x) - \tilde{f}^*) \leq \frac{\epsilon}{4}Lr^2,$$

that is, $f(x) - f^* \leq \epsilon \cdot \frac{1}{2}Lr^2 = \delta$. Hence, $T_\delta(\mathcal{A}', f) \leq T = T_{\tilde{\delta}}(\mathcal{A}, \tilde{f})$, as claimed.

Given $r' \geq 2r$, we know that $\tilde{x}^* \in B(x_{\text{ref}}, r')$ and $\tilde{f}$ is $\frac{L}{\sigma} + \zeta_{r'} = \kappa$-smooth in $B(x_{\text{ref}}, r')$ by Lemma 5. Hence, $\tilde{f} \in \mathcal{F}_{r',\infty,1,\kappa}$, and $T_\delta(\mathcal{A}', f) \leq T_{\tilde{\delta}}(\mathcal{A}, \tilde{f}) \leq \sup_{f' \in \mathcal{F}_{r',\infty,1,\kappa}} T_{\tilde{\delta}}(\mathcal{A}, f')$. Recall that this holds for all $f \in \mathcal{F}_{r,\infty,0,L}$, so $\sup_{f \in \mathcal{F}_{r,\infty,0,L}} T_\delta(\mathcal{A}', f) \leq \sup_{f' \in \mathcal{F}_{r',\infty,1,\kappa}} T_{\tilde{\delta}}(\mathcal{A}, f')$, so

$$T_{\epsilon,r} = \inf_{\tilde{\mathcal{A}}} \sup_{f \in \mathcal{F}_{r,\infty,0,L}} T_\delta(\tilde{\mathcal{A}}, f) \leq \sup_{f' \in \mathcal{F}_{r',\infty,1,\kappa}} T_{\tilde{\delta}}(\mathcal{A}, f')$$

where the infimum is over all algorithms $\tilde{\mathcal{A}}$. This holds for all algorithms $\mathcal{A}$, so we conclude

$$T_{\epsilon,r} = \inf_{\tilde{\mathcal{A}}} \sup_{f \in \mathcal{F}_{r,\infty,0,L}} T_\delta(\tilde{\mathcal{A}}, f) \leq \inf_{\mathcal{A}} \sup_{f' \in \mathcal{F}_{r',\infty,1,\kappa}} T_{\tilde{\delta}}(\mathcal{A}, f') = T_{\epsilon',r',\kappa}$$

using that $\tilde{\delta} = \frac{1}{2}r^2 = \frac{r^2}{\kappa(r')^2} \cdot \frac{1}{2}\kappa(r')^2 = \epsilon' \cdot \frac{1}{2}\kappa(r')^2$. ∎

We then have the following proof of Theorem 12.

Proof [Proof of Theorem 12] We have two cases:

Case I: $r' \geq 2$. Taking $r = 1$ and $\epsilon = \frac{2}{\kappa - \zeta_{r'}}$, Proposition 57 and Theorem 11 imply

$$q_{\kappa,r'} \geq \frac{T_{\epsilon',r',\kappa}}{\log(1/\epsilon')} \geq \frac{T_{\epsilon,r}}{\log(1/\epsilon')} \geq \frac{1}{\log(\kappa(r')^2)\sqrt{8\zeta_r^2\epsilon}} = \frac{\sqrt{\kappa - \zeta_{r'}}}{4\zeta_r\log(\kappa(r')^2)} \geq \frac{\sqrt{\kappa - \zeta_{r'}}}{8\log(\kappa(r')^2)}.$$

Case II: $r' < 2$. Taking $r = r'/2$ and $\epsilon = \frac{2}{\kappa - \zeta_{r'}}$, Proposition 57 and Theorem 11 imply

$$q_{\kappa,r'} \geq \frac{T_{\epsilon',r',\kappa}}{\log(1/\epsilon')} \geq \frac{T_{\epsilon,r}}{\log(1/\epsilon')} \geq \frac{1}{\log(4\kappa)\sqrt{8\zeta_r^2\epsilon}} = \frac{\sqrt{\kappa - \zeta_{r'}}}{4\zeta_r\log(4\kappa)} \geq \frac{\sqrt{\kappa - \zeta_{r'}}}{8\log(4\kappa)}.$$

∎

J Width-bounded separators: Working out dependence on d in Lemma 34

Kisfaludi-Bak (2020, Lemma 9 (ii)), shows that $\mathbb{E}[\mathcal{S}] \leq \sum_{j=1}^{J} \mathbb{P}[T \cap H \neq \emptyset | T \in L_j]|L_j|$ where

- $L_j = \{T \in \mathcal{T} : \text{dist}(p, T) \in [(j-1)\tau, j\tau)\}$. In particular, using that $c_0\tau/2 \geq 4\log(d)$ and Lemma 33,
$$|L_j| \leq V_d((j+1)\tau)/V_d(c_0\tau/2) \leq 4e^{(d-1)\tau(j+1-c_0/2)}.$$

- $J \geq 2$ is such that the number of tiles T which intersect the interior of $B(p, (J-1)\tau)$ is at most $|\mathcal{I}|$. In particular, this implies that $V_d((J-1)\tau)/V_d(\tau/2) \geq |\mathcal{I}|$. Using 33, we know that
$$V_d((J-1)\tau)/V_d(\tau/2) \geq e^{(d-1)(J-1)\tau}e^{-\tau(d-1)/2-2}$$
and so $(J-1)\tau \leq \frac{1}{d-1}(\log(|\mathcal{I}|) + \tau(d-1)/2 + 2)$.

Kisfaludi-Bak (2020, Lemma 8), also shows that

$$\mathbb{P}[T \cap H \neq \emptyset | T \in L_j] \leq 4\frac{\sinh(2\tau)}{\sinh(j\tau)} \cdot \frac{\sigma_{d-2}}{\sigma_{d-1}} \leq \frac{4\sqrt{d}}{2\pi} \cdot \frac{\sinh(2\tau)}{\sinh(j\tau)} \leq \frac{8\sqrt{d}}{2\pi}e^{-\tau(j-2)}.$$

Plugging in these bounds and using $\tau \geq \frac{2}{3(d-1)}\log(\frac{64e^2}{\sqrt{2\pi}}\sqrt{d})$ for the last inequality below, we find

$$
\begin{aligned}
\mathbb{E}[\mathcal{S}] &\leq \frac{32\sqrt{d}}{\sqrt{2\pi}}e^{(d-1)\tau(3-c_0/2)} \sum_{j=1}^{J} e^{(d-2)\tau(j-2)} \\
&= \frac{32\sqrt{d}}{\sqrt{2\pi}}e^{(d-1)\tau(3-c_0/2)}[-e^{-(d-2)\tau} + \frac{e^{(d-2)\tau(J-1)}-1}{e^{(d-2)\tau}-1}] \\
&\leq \frac{64\sqrt{d}}{\sqrt{2\pi}}e^{(d-1)\tau(3)}e^{(d-2)\tau(J-1)} \\
&\leq \frac{64\sqrt{d}}{\sqrt{2\pi}}e^{3(d-1)\tau}e^{\frac{d-2}{d-1}(\tau(d-1)/2+2)}|\mathcal{I}|^{\frac{d-2}{d-1}} \leq \frac{1}{2}e^{5(d-1)\tau}|\mathcal{I}|^{\frac{d-2}{d-1}}.
\end{aligned}
$$

Next, we need to workout c_0. Kisfaludi-Bak (2020, Lemma 5(ii)), shows there exists a $(c_0\tau/2, \tau/2)$-nice tiling if $c_0 \leq \tau^{-1}\log\left(1 + \frac{2}{d}(\cosh(\tau/2)-1)\right)$. Using that $\cosh(\tau/2) - 1 \geq e^{\tau/2}/4$ if $\tau \geq 3$ and $\frac{1}{2d}e^{\tau/2} \geq e^{\tau/4}$ if $\tau \geq 8\log(d)$, we find that

$$\tau^{-1}\log\left(1 + \frac{2}{d}(\cosh(\tau/2)-1)\right) \geq \tau^{-1}\log\left(\frac{1}{2d}e^{\tau/2}\right) \geq \frac{1}{4}.$$

Hence, we can take $c_0 = 1/4$.

Acknowledgments