Negative curvature obstructs acceleration for strongly geodesically convex optimization, even with exact first-order oracles

Christopher Criscitiello CHRISTOPHER.CRISCITIELLO@EPFL.CH

Nicolas Boumal NICOLAS.BOUMAL@EPFL.CH

Ecole Polytechnique Fédérale de Lausanne (EPFL), Institute of Mathematics

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

Hamilton and Moitra (2021) showed that, in certain regimes, it is not possible to accelerate Riemannian gradient descent in the hyperbolic plane if we restrict ourselves to algorithms which make queries in a (large) bounded domain and which receive gradients and function values corrupted by a (small) amount of noise. We show that acceleration remains unachievable for any deterministic algorithm which receives exact gradient and function-value information (unbounded queries, no noise). Our results hold for a large class of Hadamard manifolds including hyperbolic spaces and the symmetric space $\mathrm{SL}(n)/\mathrm{SO}(n)$ of positive definite $n \times n$ matrices of determinant one. This cements a surprising gap between the complexity of convex optimization and geodesically convex optimization: for hyperbolic spaces, Riemannian gradient descent is optimal on the class of smooth and strongly geodesically convex functions (in the regime where the condition number scales with the radius of the optimization domain). The key idea for proving the lower bound consists of perturbing squared distance functions with sums of bump functions chosen by a resisting oracle.

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

1 Introduction

We consider optimization problems of the form

$$\min_{x \in \mathcal{M}} f(x) \tag{P}$$

where $\mathcal{M}$ is a Riemannian manifold and $f \colon \mathcal{M} \to \mathbb{R}$ is a smooth strongly geodesically convex (g-convex) function (we review technical geometric terms in Section 2). When $\mathcal{M}$ is a Euclidean space, problem (P) amounts to smooth strongly convex optimization.

Several problems of interest are non-convex but can be recast as g-convex optimization problems, which means global solutions can be found efficiently. Examples and applications in data science, statistics and machine learning include: computing intrinsic means or medians on curved spaces (Karcher, 1977; Yuan et al., 2020) such as for computational anatomy (Fletcher et al., 2009) or phylogenetics (Bacák, 2014, Ch. 8), metric learning (Zadeh et al., 2016), computing optimistic likelihoods (Nguyen et al., 2019), parameter estimation for mixture models (Hosseini and Sra, 2015), robust covariance estimation and subspace recovery (Auderset et al., 2005; Wiesel, 2012; Zhang, 2012; Wiesel and Zhang, 2014; Sra and Hosseini, 2015; Ciobotaru and Mazza, 2020; Franks and Moitra, 2020), estimation for matrix normal models (Tang and Allen, 2021; Amendola et al., 2021; Franks et al., 2021), sampling on Riemannian manifolds (Goyal and Shetty, 2019), and landscape analysis such as for matrix completion (Ahn and Suarez, 2021). In mathematics and theoretical computer science, applications of g-convex optimization include computing Brascamp–Lieb constants (Sra et al., 2018), the null cone membership problem and polynomial identity testing—see (Allen-Zhu et al., 2018; Bürgisser et al., 2019; Franks and Reichenbach, 2021) and references therein. More generally, optimization on manifolds also has applications in scientific computing, imaging, communications and robotics (Absil et al., 2008; Hu et al., 2020; Boumal, 2022).

Given these applications, it is natural to ask for fast algorithms for the g-convex optimization problem (P). We consider algorithms which have access to an oracle providing first-order information (function values and gradients), and consider the following computational task:

> *Let $f\colon \mathcal{M} \to \mathbb{R}$ be a μ-strongly g-convex function which is L-smooth in a geodesic ball B of radius r and whose minimizer x^* lies in B (see Sections 1.2 and 2.3). Find a point $x \in \mathcal{M}$ within distance $\frac{r}{5}$ of x^*.*[1]

The radius r represents our initial uncertainty about the location of the minimizer of f. Thus we ask: how many queries are required to reduce our uncertainty by a constant factor (five in this case)? When $\mathcal{M} = \mathbb{R}^d$ is Euclidean space, g-convexity is equivalent to convexity, and it is well known that (projected) gradient descent (GD) uses at most $O(\kappa)$ queries to solve this computational task, where the condition number $\kappa = \frac{L}{\mu}$ represents the conditioning of the problem. In contrast, Nesterov's accelerated gradient method (NAG) (adapted to the ball domain) uses $\tilde{O}\big(\sqrt{\kappa}\big)$ queries (Nesterov, 2013, Thm. 6 and Sec. 5.1), and that is optimal (Nesterov, 2004, Ch. 2).[2]

For the moment, let us consider the case where $\mathcal{M}$ is a hyperbolic space, meaning it has constant negative curvature. Zhang and Sra (2016) show that (projected) Riemannian gradient descent (RGD) uses at most $\tilde{O}(\kappa)$ gradient queries to solve the computational task described above.[3] This matches the rate of gradient descent in Euclidean spaces. We are led to the following question:

> *Is there an algorithm for g-convex optimization on hyperbolic spaces which solves the above computational task in $\tilde{O}(\sqrt{\kappa})$ queries?*

In this paper, we show that *no such accelerated algorithm exists*, and in fact *RGD is optimal* for smooth strongly g-convex optimization on hyperbolic spaces, in the regime $r = \Theta(\kappa)$ (see Section 1.3).[4] Indeed, a number of algorithms have been developed to address this question (Liu et al., 2017; Zhang and Sra, 2018; Ahn and Sra, 2020; Jin and Sra, 2021; Martínez-Rubio, 2021; Lezcano-Casado, 2020; Alimisis et al., 2020, 2021; Huang and Wei, 2021; Duruisseaux and Leok, 2021; Franca et al., 2021a,b), but none of them are proven to achieve the fully accelerated rate of $\tilde{O}(\sqrt{\kappa})$.

Our analysis builds on the recent work of Hamilton and Moitra (2021), who show that acceleration on the hyperbolic plane is impossible when function values and gradients are corrupted by noise, even when this noise is very small. Their argument introduces a number of important ideas, most notably a key geometric property of the hyperbolic plane which we call the "ball-packing property": for $r > 0$ sufficiently large, any geodesic ball of radius r in the hyperbolic plane contains $N = e^{\Theta(r)}$ disjoint open geodesic balls of radius $\frac{r}{4}$. Using several of Hamilton and Moitra (2021)'s ideas, plus additional ideas we introduce, we prove that acceleration is impossible even when function values and gradients are known exactly, i.e., not corrupted by noise.

[1] It suffices to ask how many queries are required to reduce the uncertainty radius r by a factor ϵ, for any fixed $\epsilon \in (0, 1)$. Throughout we take $\epsilon = \frac{1}{5}$, following Hamilton and Moitra (2021).

[2] Throughout, O and Ω do *not* hide the parameter r. Also, $\tilde{O}$ and $\tilde{\Omega}$ hide logarithmic factors in $\kappa = \frac{L}{\mu}$ and r.

[3] The rate given by Zhang and Sra (2016) depends on r, but see Appendix K for how to remove this dependence on r for hyperbolic spaces.

[4] To establish lower bounds when $\kappa \gg r \gg 1$, further ideas seem to be necessary.

In addition to proving lower bounds for queries yielding exact information (our main contribution), we improve upon the results of Hamilton and Moitra (2021) in several ways. In Section 2.2, we establish the ball-packing property for a large class of Hadamard manifolds, including the symmetric space $\mathrm{SL}(n)/\mathrm{SO}(n)$ of positive definite matrices with determinant one which is important in applications, at least in part because the Riemannian metric on $\mathrm{SL}(n)/\mathrm{SO}(n)$ is the Fisher–Rao information metric for covariance matrices of Gaussian distributions (Skovgaard, 1984; Lenglet et al., 2006). In turn, we show that acceleration is impossible on this large class of Hadamard manifolds. Hamilton and Moitra (2021) also restrict all algorithms to query in a bounded domain. We remove this assumption using a reduction which, starting from hard functions designed for algorithms making bounded queries, produces hard functions for algorithms which can make unbounded queries.

1.1 Key ideas: building hard g-convex functions

Hamilton and Moitra (2021) establish their lower bound by exhibiting a distribution on strongly g-convex functions which is challenging for any algorithm receiving function information corrupted by noise. Amazingly, the hard distribution they consider is simply a uniform distribution over a finite number of Riemannian squared distance functions. Intuitively, this distribution is difficult for algorithms because geodesics diverge rapidly in hyperbolic space (see Lemma 6), so a small amount of noise in a gradient is magnified. However, like in Euclidean spaces, the Riemannian gradient of a squared distance function points directly towards the function's minimizer. Therefore, squared distance functions are not enough to go beyond noisy oracles.

The key idea we introduce is to use squared distance functions *perturbed by a resisting oracle*, i.e., functions $f(x) = \frac{1}{2}\mathrm{dist}(x, x^*)^2 + H(x)$ with $\|\mathrm{Hess}H(x)\|$ small. The perturbations H are not g-convex, but since their Hessian is small, the perturbed functions f retain strong g-convexity. Each perturbation is constructed as a *sum of bump functions*, that is, C^∞ functions with compact support.

1.2 Algorithm and problem classes

It is crucial to define the class of functions for which we prove lower bounds. A natural function class to consider is the set of functions $f\colon \mathcal{M} \to \mathbb{R}$ which are L-smooth[5] and μ-strongly g-convex on all of $\mathcal{M}$ (see Section 2.3). Yet, if $\mathcal{M}$ has sectional curvatures upper bounded by some $K_{\mathrm{up}} < 0$ or if $\mathcal{M} = \mathrm{SL}(n)/\mathrm{SO}(n)$, then this class is empty. It is impossible for a function to be both L-smooth and strongly g-convex on all of $\mathcal{M}$ if $K_{\mathrm{up}} < 0$ or if $\mathcal{M} = \mathrm{SL}(n)/\mathrm{SO}(n)$.[6]

A simple remedy for this issue is to consider minimizing μ-strongly g-convex functions which are L-smooth in a ball of finite radius r. This is especially natural since whether acceleration is possible depends on how r compares with κ—this will become clearer in Section 1.4. Let $\mathcal{M}$ be a Hadamard manifold, and let $B(x_{\mathrm{ref}}, r) \subseteq \mathcal{M}$ denote the closed geodesic ball centered at $x_{\mathrm{ref}} \in \mathcal{M}$ of radius r (see Section 2.1). We consider the following class of real-valued functions $f\colon \mathcal{M} \to \mathbb{R}$.

Definition 1 *For $\kappa \geq 1, r > 0, x_{\mathrm{ref}} \in \mathcal{M}$, let $\mathcal{F}_{\kappa,r}^{x_{\mathrm{ref}}}(\mathcal{M})$ be the set of C^∞ functions on $\mathcal{M}$ which*

- *are μ-strongly g-convex in all of $\mathcal{M}$ with $\mu > 0$;*
- *are L-smooth in $B(x_{\mathrm{ref}}, r)$ with $\kappa = \frac{L}{\mu}$; and*

[5] We say a function is L-smooth if it has L-Lipschitz Riemannian gradient (Definition 9). When we say a function is smooth, we mean that it is L-smooth for some $L \geq 0$. We say a function is C^∞ if it is infinitely differentiable.

[6] See Proposition 28 in Appendix I, which is an extension of a result due to Hamilton and Moitra (2021).

- *have a unique global minimizer x^* which lies in the ball $B(x_{\mathrm{ref}}, \frac{3}{4}r)$.*

In the third item of Definition 1, we require $\frac{3}{4}r$ instead of r to ensure that the ball $B(x^*, \frac{r}{5})$ is contained in the interior of $B(x_{\mathrm{ref}}, r)$.

We impose no restrictions on the algorithm except that it is deterministic. A deterministic first-order algorithm $\mathcal{A}$ on $\mathcal{M}$ is an initial point x_0 and a sequence of maps $(\mathcal{A}_k \colon (\mathbb{R}\times \mathrm{T}\mathcal{M})^k \to \mathcal{M})_{k\geq 1}$. Running an algorithm $\mathcal{A}$ on a cost function $f \colon \mathcal{M} \to \mathbb{R}$ produces iterates $x_0, x_1, x_2, \ldots$ given by $x_k = \mathcal{A}_k((f_0, (x_0, g_0)), \ldots, (f_{k-1}, (x_{k-1}, g_{k-1})))$, where $f_\ell = f(x_\ell)$ and $g_\ell = \mathrm{grad} f(x_\ell)$ constitute the past function value and gradient information gathered thus far. It is an open question whether the lower bounds in this paper can be extended to randomized algorithms.

1.3 Main results

We now state our main results about the impossibility of acceleration for the function class $\mathcal{F}_{\kappa,r}^{x_{\mathrm{ref}}}(\mathcal{M})$ in Definition 1. There is some leeway in choosing the constants below.

Theorem 2 *Let $\mathcal{M}$ be a Hadamard manifold of dimension $d \geq 2$ whose sectional curvatures are in the interval $[K_{\mathrm{lo}}, K_{\mathrm{up}}]$ with $K_{\mathrm{up}} < 0$. Let $x_{\mathrm{ref}} \in \mathcal{M}$, $\kappa \geq 1000\sqrt{\frac{K_{\mathrm{lo}}}{K_{\mathrm{up}}}}$ and define $r > 0$ such that $\kappa = 12r\sqrt{-K_{\mathrm{lo}}} + 9$. For every deterministic first-order algorithm $\mathcal{A}$, there is a function $f \in \mathcal{F}_{\kappa,r}^{x_{\mathrm{ref}}}(\mathcal{M})$ such that algorithm $\mathcal{A}$ requires at least*

$$\left\lfloor \sqrt{\frac{K_{\mathrm{up}}}{K_{\mathrm{lo}}}} \cdot \frac{\kappa}{1000\log\left(10\kappa\right)} \right\rfloor = \tilde{\Omega}\left(\sqrt{\frac{K_{\mathrm{up}}}{K_{\mathrm{lo}}}} \cdot \kappa\right)$$

queries in order to find a point $x \in \mathcal{M}$ within distance $\frac{r}{5}$ of the minimizer of f.

Corollary 3 *Let $\mathcal{M}$ be a hyperbolic space ($K_{\mathrm{lo}} = K_{\mathrm{up}} = K < 0$), $x_{\mathrm{ref}} \in \mathcal{M}$, $\kappa \geq 1000$ and define $r > 0$ such that $\kappa = 12r\sqrt{-K} + 9$. Among deterministic first-order algorithms, Riemannian gradient descent is optimal (up to log factors) on the function class $\mathcal{F}_{\kappa,r}^{x_{\mathrm{ref}}}(\mathcal{M})$.*

The symmetric space $\mathrm{SL}(n)/\mathrm{SO}(n)$ does not have strictly negative curvature as required by Theorem 2, but we can still show that acceleration is unachievable if $n \geq 2$ is held fixed as κ grows.

Theorem 4 *Let $x_{\mathrm{ref}} \in \mathrm{SL}(n)/\mathrm{SO}(n)$, $\kappa \geq 1000n$ and define $r > 0$ such that $\kappa = 6r\sqrt{2} + 9$. For every deterministic first-order algorithm $\mathcal{A}$, there is a function $f \in \mathcal{F}_{\kappa,r}^{x_{\mathrm{ref}}}(\mathrm{SL}(n)/\mathrm{SO}(n))$ such that the algorithm $\mathcal{A}$ requires at least $\left\lfloor \frac{1}{n} \cdot \frac{\kappa}{1000\log(10\kappa)} \right\rfloor = \tilde{\Omega}\left(\frac{1}{n} \cdot \kappa\right)$ queries in order to find a point x within distance $\frac{r}{5}$ of the minimizer of f.*

The lower bound $\tilde{\Omega}(\frac{\kappa}{n})$ also holds for the symmetric space $\mathcal{P}_n$ of positive definite matrices with affine-invariant metric because it is isometric to $\mathbb{R} \times \mathrm{SL}(n)/\mathrm{SO}(n)$ (see Appendix J). It is an open question whether one can remove the factors $\sqrt{\frac{K_{\mathrm{up}}}{K_{\mathrm{lo}}}}$ and $\frac{1}{n}$ in the lower bounds in Theorems 2 and 4.

1.4 Comparison to literature: best known upper bounds

Let us review the best known upper bounds for smooth g-convex optimization (see Appendix A for a more complete discussion of the literature). Ahn and Sra (2020) provide an algorithm which is strictly faster than RGD, and requires only $\tilde{O}(\sqrt{\kappa})$ queries for the computational task described in the introduction when $r \leq O(\frac{1}{\kappa^{3/4}})$. Intuitively this makes sense because Riemannian manifolds are locally Euclidean, so in a small enough ball the effects of curvature are negligible. When r is not small, the algorithm of Ahn and Sra (2020) requires $\tilde{O}(\kappa)$ gradient queries.

The guarantees for the algorithm provided by Ahn and Sra (2020) hold for Hadamard manifolds of bounded curvature. For hyperbolic spaces in particular, Martínez-Rubio (2021) improves upon these guarantees by providing an algorithm requiring $e^{\tilde{O}(r)}\sqrt{\kappa}$ queries to solve the computational task; in particular, this algorithm is accelerated when $r \leq O(1)$.

2 Preliminaries and the ball-packing property

We introduce the tools used to prove the main results. For an introduction to Riemannian manifolds see (Lee, 2012, 2018), or (Absil et al., 2008; Boumal, 2022) for an optimization perspective.

2.1 Hadamard manifolds

Throughout, $\mathcal{M}$ denotes a smooth manifold which has tangent bundle $\mathrm{T}\mathcal{M}$ and tangent spaces $\mathrm{T}_x\mathcal{M}$. We equip $\mathcal{M}$ with a Riemannian metric: a smoothly-varying inner product $\langle \cdot, \cdot \rangle_x$ on each tangent space $\mathrm{T}_x\mathcal{M}$. Throughout, we drop the subscript and denote these inner products by $\langle \cdot, \cdot \rangle$. The metric allows us to define the gradient $\mathrm{grad} f(x) \in \mathrm{T}_x\mathcal{M}$ and Hessian $\mathrm{Hess} f(x) \colon \mathrm{T}_x\mathcal{M} \to \mathrm{T}_x\mathcal{M}$ of the cost function f at each point x (Boumal, 2022, Ch. 3, 5). We write $\|v\| = \sqrt{\langle v, v \rangle}$ for $v \in \mathrm{T}_x\mathcal{M}$ and $\|A\|$ for the operator norm of a linear operator $A \colon \mathrm{T}_x\mathcal{M} \to \mathrm{T}_y\mathcal{M}$. We use I to denote the identity linear operator from $\mathrm{T}_x\mathcal{M}$ to $\mathrm{T}_x\mathcal{M}$.

The Riemannian metric gives $\mathcal{M}$ a notion of distance dist and geodesics. The closed (geodesic) ball of radius r centered at $x \in \mathcal{M}$ is $B(x, r) = \{y \in \mathcal{M} : \mathrm{dist}(y, x) \leq r\}$. The closed ball in $\mathrm{T}_x\mathcal{M}$ centered at $g \in \mathrm{T}_x\mathcal{M}$ with radius r is $B_x(g, r) = \{s \in \mathrm{T}_x\mathcal{M} : \|s - g\| \leq r\}$.

The metric also provides a notion of *intrinsic curvature*. We focus on Hadamard manifolds:

Definition 5 *A Riemannian manifold $\mathcal{M}$ is a Hadamard manifold if $\mathcal{M}$ is complete, simply connected and has nonpositive sectional curvature everywhere.*

By the Cartan–Hadamard Theorem, all d-dimensional Hadamard manifolds $\mathcal{M}$ are diffeomorphic to $\mathbb{R}^d$ (Lee, 2018, Thm. 12.8). The Hopf–Rinow Theorem implies that the exponential map $\exp \colon \mathrm{T}\mathcal{M} \to \mathcal{M}$ is well defined on the entire tangent bundle, and moreover every pair of points can be connected by a unique geodesic and this geodesic is minimal (Lee, 2018, Prop. 12.9). This means that the inverse of the exponential map $\exp_x^{-1} \colon \mathcal{M} \to \mathrm{T}_x\mathcal{M}$ is well defined for all $x \in \mathcal{M}$. We use $P_{x \to y} \colon \mathrm{T}_x\mathcal{M} \to \mathrm{T}_y\mathcal{M}$ to denote parallel transport along the geodesic connecting x and y.

The next lemma is a direct consequence of the hyperbolic law of cosines and Toponogov's triangle comparison theorem—see Appendix B. It expresses the fact that when the underlying space is negatively curved, geodesics diverge quickly. Lemma 6 forms the basis of Lemma 7 (spaces with sufficient negative curvature satisfy the ball-packing property), which is the most important geometric fact underlying Theorems 2 and 4. A proof of Lemma 6 can be found in Appendix H.1.

Lemma 6 (Geodesics diverge) *Let v_1, v_2 be two tangent vectors at x_{ref} on a Hadamard manifold $\mathcal{M}$ with identical norms $s = \|v_1\| = \|v_2\|$ and forming an angle at least θ. If the sectional curvatures of $\mathcal{M}$ are upper bounded by $K_{\mathrm{up}} < 0$ and $\theta = e^{1 - \frac{2}{3}s\sqrt{-K_{\mathrm{up}}}}$, then $\mathrm{dist}(z_1, z_2) \geq \frac{2}{3}s$ where $z_i = \exp_{x_{\mathrm{ref}}}(v_i)$ for $i = 1, 2$.*

It is instructive to compare this lemma to the Euclidean case, where the law of cosines implies $\|z_1 - z_2\|^2 \leq 2s^2 - 2s^2\cos(\theta) = O(s^2\theta^2)$. Therefore, if $\theta = \Theta(e^{-s})$, then $\|z_1 - z_2\| = O(se^{-s})$.

2.2 The ball-packing property

To prove the lower bound, we require our space to satisfy the following geometric property.

A 1 (Ball-packing property) *There is a point $x_{\text{ref}} \in \mathcal{M}$, an $\tilde{r} > 0$ and a $\tilde{c} > 0$ such that for all $r \geq \tilde{r}$, there exist $N \geq e^{\tilde{c}r}$ points $z_1, \ldots, z_N$ in the ball $B(x_{\text{ref}}, \frac{3}{4}r)$ so that all pairs of points are separated by a distance of at least $\frac{r}{2}$: $\text{dist}(z_i, z_j) \geq \frac{r}{2}$ for all $i \neq j$. We say $\mathcal{M}$ satisfies the ball-packing property with $\tilde{r}, \tilde{c}$ and $x_{\text{ref}} \in \mathcal{M}$.*

We think of the points $z_1, \ldots, z_N$ as centers of disjoint open balls of radius $\frac{r}{4}$ contained in $B(x_{\text{ref}}, r)$. No Euclidean space $\mathbb{R}^d$ satisfies a ball-packing property as the volume of a ball of radius r scales polynomially as r^d, not exponentially.

A 2 (Strong ball-packing property) *There is an $\tilde{r} > 0$ and a $\tilde{c} > 0$ such that $\mathcal{M}$ satisfies the ball-packing property with $\tilde{r}, \tilde{c}$ and every $x_{\text{ref}} \in \mathcal{M}$.*

Lemma 7 *Let $d \geq 2$ and $\mathcal{M}$ be a d-dimensional Hadamard manifold whose sectional curvatures are in the interval $(-\infty, K_{\text{up}}]$ with $K_{\text{up}} < 0$. Then $\mathcal{M}$ satisfies the strong ball-packing property A2 for $\tilde{r} = \frac{4}{\sqrt{-K_{\text{up}}}}$ and $\tilde{c} = d\frac{\sqrt{-K_{\text{up}}}}{8}$.*

Proof Let $x_{\text{ref}} \in \mathcal{M}$. Let $r \geq \tilde{r}$ and let $s = \frac{3}{4}r$. Let $\theta = e^{1-\frac{2}{3}s\sqrt{-K_{\text{up}}}}$. Consider the sphere $\mathbb{S}^{d-1}_{x_{\text{ref}}}(s) = \{v \in \mathrm{T}_{x_{\text{ref}}}\mathcal{M} : \|v\| = s\}$. We have $\theta \leq \frac{\pi}{2}$ because $s\sqrt{-K_{\text{up}}} \geq 3$. Therefore using $d \geq 2$ and a standard covering number argument adapted to our setting (see Lemma 27 in Appendix H.2), we find there exist

$$N \geq \theta^{-(d-1)} = e^{(d-1)(\frac{2}{3}s\sqrt{-K_{\text{up}}}-1)} \geq e^{\frac{1}{2}d(\frac{1}{2}r\sqrt{-K_{\text{up}}}-1)} \geq e^{\frac{1}{8}dr\sqrt{-K_{\text{up}}}}$$

tangent vectors $v_1, \ldots, v_N \in \mathbb{S}^{d-1}_{x_{\text{ref}}}(s)$ such that the angle between vectors v_i and v_j is at least θ for all $i \neq j$. Define $z_j = \exp_{x_{\text{ref}}}(v_j)$ for $j = 1, 2, \ldots, N$. Therefore, $z_j \in B(x_{\text{ref}}, \frac{3}{4}r)$ for all j. Moreover, $\text{dist}(z_i, z_j) \geq \frac{2}{3}s = \frac{r}{2}$ for all $i \neq j$ owing to Lemma 6 (geodesics diverge). ■

If $\mathcal{M}$ is a hyperbolic space, we can instead argue Lemma 7 using a simple volume argument, combined with the fact that the covering number is less than the packing number (Vershynin, 2018, Lem. 4.2.8). However, that argument does not hold for spaces with nonconstant curvature because we would need a bound on the ratio of K_{lo} to K_{up}. Lemma 7 does not make such an assumption.

The symmetric spaces $\mathcal{SLP}_n = \text{SL}(n)/\text{SO}(n)$ and $\mathcal{P}_n = \mathbb{R} \times \mathcal{SLP}_n$ are Hadamard manifolds which do not have strictly negative curvature, so we cannot apply Lemma 7. However, $\mathcal{SLP}_n$ contains an $(n-1)$-dimensional totally geodesic submanifold isometric to a hyperbolic space (Bridson and Haefliger, 1999, Ch. II.10). This allows us to prove Lemma 8 in Appendix J, where we also argue the best possible $\tilde{c}$ for $\mathcal{SLP}_n$ satisfies $\tilde{c} \leq O(n^{3/2}) = o(\dim(\mathcal{SLP}_n))$.

Lemma 8 *For $n \geq 3$, both $\mathcal{SLP}_n$ and $\mathcal{P}_n$ satisfy the strong ball-packing property A2 with $\tilde{r} = 8\sqrt{2}, \tilde{c} = \frac{n-1}{16\sqrt{2}}$, and $\mathcal{SLP}_2$ and $\mathcal{P}_2$ satisfy the strong ball-packing property with $\tilde{r} = 4\sqrt{2}, \tilde{c} = \frac{1}{4\sqrt{2}}$.*

2.3 Geodesic convexity

A subset D of a Hadamard manifold $\mathcal{M}$ is g-convex if the geodesic segment connecting each pair of points in D is contained in D (Udrişte, 1994). Geodesic balls are g-convex. We study special functions on g-convex sets (Lemma 16 has other characterizations of g-convexity and smoothness).

Definition 9 *Let $f\colon \mathcal{M} \to \mathbb{R}$ be a twice continuously differentiable function on a Hadamard manifold $\mathcal{M}$, and let D be a g-convex subset of $\mathcal{M}$. We say (somewhat restrictively) that:*

- *f is μ-strongly g-convex in D if $\mathrm{Hess}f(x) \succeq \mu I$ for all $x \in D$.*
- *f is L-smooth in D if $\|\mathrm{Hess}f(x)\| \leq L$ for all $x \in D$.*

(If f is L-smooth in D, then $\|\mathrm{grad}f(x) - P_{y\to x}\mathrm{grad}f(y)\| \leq L\mathrm{dist}(x, y)$ for all $x, y \in D$.)

Lemma 10 *(Alimisis et al., 2020, Lem. 2 in App. B) Let $\mathcal{M}$ be a Hadamard manifold with sectional curvatures in $[K_{\mathrm{lo}}, 0]$. Fix $z \in \mathcal{M}$, and let $f\colon \mathcal{M} \to \mathbb{R}, f(x) = \frac{1}{2}\mathrm{dist}(x, z)^2$. Then f is C^∞, $\mathrm{grad}f(x) = -\exp_x^{-1}(z)$, and f is 1-strongly g-convex in $\mathcal{M}$ and L-smooth in $B(z, r)$ for any $r > 0$ with $L = \frac{r\sqrt{-K_{\mathrm{lo}}}}{\tanh(r\sqrt{-K_{\mathrm{lo}}})} \leq 1 + r\sqrt{-K_{\mathrm{lo}}}$.*

3 Technical version of the main theorem and proof of key lemma

We are now ready to prove our main technical theorem, from which Theorems 2 and 4 in the introduction follow. For ease of exposition, we state and prove the following slightly simpler theorem in the main part of the paper. For this theorem, we assume the algorithm only receives gradient information (no function values), and the algorithm always makes queries in a bounded domain. For the statement below, recall the definition of the function class $\mathcal{F}_{\kappa,r}^{x_{\mathrm{ref}}}(\mathcal{M})$ from Section 1.2.

Theorem 11 *Let $\mathcal{M}$ be a Hadamard manifold of dimension $d \geq 2$ which satisfies the ball-packing property A1 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 $r \geq \max\left\{\tilde{r}, \frac{8}{\sqrt{-K_{\mathrm{lo}}}}, \frac{4(d+2)}{\tilde{c}}\right\}$. Define $\kappa = 4r\sqrt{-K_{\mathrm{lo}}} + 3$. Let $\mathcal{A}$ be any deterministic algorithm which only makes gradient queries, and assume that $\mathcal{A}$ always queries in $B(x_{\mathrm{ref}}, \mathscr{R})$, with $\mathscr{R} \geq r$.*

Then there is a function $f \in \mathcal{F}_{\kappa,r}^{x_{\mathrm{ref}}}(\mathcal{M})$ with minimizer x^* such that running $\mathcal{A}$ on f yields iterates $x_0, x_1, x_2, \ldots$ satisfying $\mathrm{dist}(x_k, x^*) \geq \frac{r}{4}$ for all $k = 0, 1, \ldots, T-1$, where

$$T = \left\lfloor \frac{\frac{1}{2}\tilde{c}d^{-1}r}{\log\left(2000 \cdot \frac{1}{2}\tilde{c}d^{-1}r(3\mathscr{R}\sqrt{-K_{\mathrm{lo}}} + 2)\right)} \right\rfloor. \tag{1}$$

In the theorem, observe that $\kappa = \Theta(r)$ and so $T = \tilde{\Theta}(r) = \tilde{\Theta}(\kappa)$ (assuming $\mathscr{R} = \mathrm{poly}(r)$).

In Appendices F and G, we state and prove Theorem 24: an extension of Theorem 11 which provides a lower bound for algorithms which can also make function-value queries as well as unbounded queries. Theorems 2 and 4 follow directly from Theorem 24 and the ball-packing properties established in Lemmas 7 and 8—see Appendices M.1 and M.2 for the details.

To allow for algorithms which make unbounded queries, the high-level idea is to modify all hard instances f from Theorem 11 so that $f(x) = \frac{1}{2}\mathrm{dist}(x, x_{\mathrm{ref}})^2$ for $x \notin B(x_{\mathrm{ref}}, \mathscr{R})$ (recall $\mathscr{R} \geq r$). This way, the algorithm gains no information by querying outside the ball $B(x_{\mathrm{ref}}, \mathscr{R})$. On the other hand, we still want the hard functions f to remain untouched in the ball $B(x_{\mathrm{ref}}, r)$. In the region between radii r and $\mathscr{R}$, we smoothly interpolate between these two choices of functions. We show that we can choose $\mathscr{R}$ appropriately so that the lower bound $\tilde{\Omega}(r)$ still holds and the modified functions are still strongly g-convex. Technically, we do this via a reduction, which is depicted in Figure 1 in Appendix G with additional details.

3.1 Key lemma: Pièce de résistance

The main ingredient to prove Theorem 11 is Lemma 12 stated below. At a high-level, we show that as long as the algorithm has made at most T queries, the oracle can always answer these queries in such a way that there exist two cost functions consistent with these queries and yet whose minimizers are significantly far away from each other. Let us make this more precise.

Consider a Hadamard manifold $\mathcal{M}$ satisfying the ball-packing property A1 with $\tilde{r}, \tilde{c} > 0$ and $x_{\mathrm{ref}} \in \mathcal{M}$. Let $z_1, z_2, \ldots, z_N$, with $N \geq e^{\tilde{c}r}$, be points in $B(x_{\mathrm{ref}}, \frac{3}{4}r), r \geq \tilde{r}$, so that all pairs of points are separated by a distance of at least $\frac{r}{2}$. Let $\mathcal{A}$ be a first-order optimization algorithm. One can even give the list of points $z_1, \ldots, z_N$ to the algorithm designer. The algorithm $\mathcal{A}$ queries points $x_0, x_1, \ldots$ and our job (as the resisting oracle) is to choose gradients $g_0, g_1, \ldots$ to return to $\mathcal{A}$.

At each iteration $k \geq 0$, we maintain a list of "active candidate functions" $f_{j,k} \colon \mathcal{M} \to \mathbb{R}$ indexed by $j \in A_k \subseteq \{1, \ldots, N\}$. The notation A_k stands for "active" set at iteration k. Each of the functions $f_{j,k}, j \in A_k$, is differentiable, strongly g-convex, and has minimizer at z_j with z_j a distance of at least $\frac{r}{4}$ from all queried points. Additionally, the functions $f_{j,k}, j \in A_k$, are consistent with the k gradient queries $(x_0, g_0), \ldots, (x_{k-1}, g_{k-1})$ answered so far, meaning $\mathrm{grad} f_{j,k}(x_m) = g_m$ for all $m < k$ and $j \in A_k$. Therefore, any of the functions $f_{j,k}$, with $j \in A_k$, can be the actual function being optimized. Hence, any of the minimizers z_j, with $j \in A_k$, can be the minimizer of the actual function being optimized. As long as A_k is nonempty, we can conclude that the algorithm $\mathcal{A}$ has not queried a point within distance $\frac{r}{4}$ of the minimizer up to iteration k.

The next set of active candidate functions $\{f_{j,k+1} : j \in A_{k+1}\}$ is chosen by modifying the current set of active candidate functions: $f_{j,k+1} = f_{j,k} + h_{j,k}$. The modifications $h_{j,k}$ and the set $A_{k+1} \subseteq A_k$ are chosen so that $\mathrm{grad} f_{j,k+1}(x_k) = g_k$, where g_k is the gradient chosen by the resisting oracle to return to the algorithm in response to the query x_k. Given the queries $x_0, x_1, \ldots, x_k$ made by the algorithm and the current active set A_k, the resisting oracle chooses $g_k \in \mathrm{T}_{x_k}\mathcal{M}$ in such a way that the algorithm gains as little information about the location of x^* as possible. This amounts to choosing g_k so that the cardinality of A_{k+1} is as large as possible. For example, if $\mathcal{M}$ is a d-dimensional hyperbolic space of curvature -1, we show $|A_{k+1}| \geq \tilde{\Omega}(|A_k| / r^d)$. Since $|A_0| \geq e^{\Omega(dr)}$ due to the ball-packing lemma 7, this allows us to conclude the desired lower bound.

Lemma 12 *Let $\mathcal{M}$ be a Hadamard manifold of dimension $d \geq 2$ with sectional curvatures in the interval $[K_{\mathrm{lo}}, 0]$ and $K_{\mathrm{lo}} < 0$. Let $x_{\mathrm{ref}} \in \mathcal{M}$, $r \geq \frac{8}{\sqrt{-K_{\mathrm{lo}}}}$, $\mathscr{R} \geq r$. Let $z_1, \ldots, z_N \in B(x_{\mathrm{ref}}, \frac{3}{4}r)$ be distinct points in $\mathcal{M}$ such that $\mathrm{dist}(z_i, z_j) \geq \frac{r}{2}$ for all $i \neq j$. Define $A_0 = \{1, 2, \ldots, N\}$. Let $\mathcal{A}$ be any first-order algorithm which only makes gradient queries and only queries points in $B(x_{\mathrm{ref}}, \mathscr{R})$. Finally, let $w \geq 1$ (this is a tuning parameter we will set later).*

For every $k = 0, 1, 2, \ldots, \lfloor 2w \rfloor$, algorithm $\mathcal{A}$ queries $x_k = \mathcal{A}_k((x_0, g_0), \ldots, (x_{k-1}, g_{k-1}))$ and there exists a tangent vector $g_k \in \mathrm{T}_{x_k}\mathcal{M}$ and a set $A_{k+1} \subseteq A_k$ satisfying

$$|A_{k+1}| \geq \frac{|A_k| - 1}{(2000w(3\mathscr{R}\sqrt{-K_{\mathrm{lo}}} + 2))^d} \tag{2}$$

such that for each $j \in A_{k+1}$ there is a C^∞ function $f_{j,k+1} \colon \mathcal{M} \to \mathbb{R}$ of the form

$$f_{j,k+1}(x) = \frac{1}{2}\mathrm{dist}(x, z_j)^2 + H_{j,k+1}(x) \tag{3}$$

satisfying:

- ***L1*** *$f_{j,k+1}$ is $(1 - \frac{k+1}{4w})$-strongly g-convex in $\mathcal{M}$ and $[2r\sqrt{-K_{\mathrm{lo}}} + 1 + \frac{k+1}{4w}]$-smooth in $B(x_{\mathrm{ref}}, r)$;*
- ***L2*** *$\mathrm{grad} f_{j,k+1}(z_j) = 0$ (hence in particular, the minimizer of $f_{j,k+1}$ is z_j);*
- ***L3*** *$\mathrm{grad} f_{j,k+1}(x_m) = g_m$ for $m = 0, 1, \ldots, k$ ($f_{j,k+1}$ is compatible with all queries);*
- ***L4*** *$\mathrm{dist}(x_m, z_j) \geq \frac{r}{4}$ for all $m = 0, 1, \ldots, k$;*
- ***L5*** *$\|\mathrm{grad} H_{j,k+1}(x)\| \leq \frac{k+1}{4w\sqrt{-K_{\mathrm{lo}}}}$ and $\|\mathrm{Hess} H_{j,k+1}(x)\| \leq \frac{k+1}{4w}$ for all $x \in \mathcal{M}$.*[7]

Proof [Proof of Theorem 11] Let us apply Lemma 12 to $\mathcal{M}$ and $\mathcal{A}$. Let the points $z_1, \ldots, z_N$ be provided by the ball-packing property so that $N \geq e^{\tilde{c}r}$. Set $w = \tilde{c}d^{-1}r/4$ in Lemma 12.

It is easy to show by induction that inequality (2) along with $|A_0| \geq e^{\tilde{c}r}$ and $r \geq \frac{4(d+2)}{\tilde{c}}$ imply $|A_k| \geq 2$ for all $k \leq \min\{T, \lfloor 2w \rfloor\} = T$. For completeness, we give a short proof in Appendix L.

Since A_T is nonempty, we can choose $j \in A_T$ and let $f = f_{j,T}$. By property **L1** and $T \leq 2w$, f is $\frac{1}{2}$-strongly g-convex in $\mathcal{M}$ and $[2r\sqrt{-K_{\mathrm{lo}}} + \frac{3}{2}]$-smooth in $B(x_{\mathrm{ref}}, r)$. Property **L2** implies f has minimizer z_j which is contained in $B(x_{\mathrm{ref}}, \frac{3}{4}r)$. Thus, f is in $\mathcal{F}_{\kappa,r}^{x_{\mathrm{ref}}}(\mathcal{M})$ with $\kappa = 4r\sqrt{-K_{\mathrm{lo}}} + 3$.

On the other hand, properties **L3** and **L4** imply running $\mathcal{A}$ on f produces iterates $x_0, \ldots, x_{T-1}$ satisfying $\mathrm{dist}(x_k, z_j) \geq \frac{r}{4}$ for all $k = 0, 1, \ldots, T-1$: all are far from the minimizer. ■

3.2 Proof of the key lemma 12

We prove by induction on k that there is a set A_{k+1} and functions $f_{j,k+1}, j \in A_{k+1}$, satisfying the properties **L1**, **L2**, **L3**, **L4**, **L5**. As we do this, we also construct the gradients $g_0, g_1, \ldots$, and we show that inequality (2) holds for each $k \geq 0$.

Our **base case** is $k + 1 = 0$. At the start (no queries), we simply define

$$f_{j,0}(x) = \frac{1}{2}\mathrm{dist}(x, z_j)^2, \quad \forall j \in A_0 = \{1, \ldots, N\}. \tag{4}$$

Clearly $\mathrm{grad} f_{j,0}(z_j) = 0$ (**L2** is satisfied), and $f_{j,0}$ is 1-strongly g-convex and $[2r\sqrt{-K_{\mathrm{lo}}} + 1]$-smooth in $B(x_{\mathrm{ref}}, r)$ by Lemma 10 (**L1** is satisfied). At iteration 0, all the functions $f_{j,0}$ are trivially consistent with the set of past queries because there are no past queries (**L3** and **L4** are satisfied). Finally, $H_{j,0}$ is identically zero (**L5** is satisfied).

Now let us move on to the **inductive step**. The remainder of this section is devoted to the inductive step. We are at iteration $k \in [0, \lfloor 2w \rfloor)$, and there have been k past queries at the points $x_0, \ldots, x_{k-1}$, along with the k gradients $g_0, \ldots, g_{k-1}$ returned by the oracle.

The algorithm queries a point x_k. If $k \geq 1$, let x_ℓ be a previous query point closest to x_k, i.e.,

$$x_\ell \in \arg\min_{x \in \{x_0, x_1, \ldots, x_{k-1}\}} \mathrm{dist}(x_k, x). \tag{5}$$

[7] This last property is helpful for the induction used to prove Lemma 12. It is not explicitly used to prove Theorem 11.

If $x_k = x_\ell$ (i.e., the algorithm repeats a query), just return $g_k = g_\ell$, take $A_{k+1} = A_k$, and we are done. Otherwise, we can assume $x_k \neq x_\ell$.

By the inductive hypothesis (**IH**), we have a set A_k such that for each $j \in A_k$ there is an infinitely differentable function $f_{j,k}$ for which:

IH1 $f_{j,k}$ is $(1 - \frac{k}{4w})$-strongly g-convex in $\mathcal{M}$ and $[2r\sqrt{-K_{\text{lo}}} + 1 + \frac{k}{4w}]$-smooth in $B(x_{\text{ref}}, r)$;

IH2 $\text{grad} f_{j,k}(z_j) = 0$;

IH3 $\text{grad} f_{j,k}(x_m) = g_m$ for $m = 0, \ldots, k-1$;

IH4 $\text{dist}(x_m, z_j) \geq \frac{r}{4}$ for all $m = 0, 1, \ldots, k-1$;

IH5 $\|\text{grad} H_{j,k}(x)\| \leq \frac{k}{4w\sqrt{-K_{\text{lo}}}}$ and $\|\text{Hess} H_{j,k}(x)\| \leq \frac{k}{4w}$ for all $x \in \mathcal{M}$.

We want to choose a large set $A_{k+1} \subseteq A_k$, and for each $j \in A_{k+1}$ we must construct $f_{j,k+1}$ as

$$f_{j,k+1} = f_{j,k} + h_{j,k}, \tag{6}$$

where $h_{j,k} \colon \mathcal{M} \to \mathbb{R}$ is an appropriately chosen function. What properties do we want the functions $h_{j,k}$ to satisfy? Compare the properties **IH1**, **IH2**, **IH3**, **IH4**, **IH5** satisfied by $f_{j,k}$, with the properties **L1**, **L2**, **L3**, **L4**, **L5** we want $f_{j,k+1}$ to satisfy. Let us look at each property.

- In order to ensure $f_{j,k+1}$ satisfies **L4**, we simply need to choose A_{k+1} so that $\text{dist}(x_k, z_j) \geq \frac{r}{4}$ for all $j \in A_{k+1}$. Define

  $$\tilde{A}_k = \left\{ j \in A_k : \text{dist}(x_k, z_j) \geq \frac{r}{4} \right\}. \tag{7}$$

  Since any pair of minimizers z_i, z_j are separated by a distance of at least $r/2$, there is at most one $j \in A_k$ such that $\text{dist}(x_k, z_j) < r/4$. Therefore $|\tilde{A}_k| \geq |A_k| - 1$. Below we define A_{k+1} as a particular subset of $\tilde{A}_k$.

- Let us look at property **L3**. If $k = 0$, then $f_{j,k+1}$ is trivially consistent with the past queries (because there are none). Assume $k \geq 1$. In order for $f_{j,k+1}$ to remain consistent with the past queries $x_0, \ldots, x_{k-1}$, it is sufficient to enforce that the closed support[8] $\overline{\text{supp}(h_{j,k})}$ does not contain $x_0, \ldots, x_{k-1}$. Of course $h_{j,k}$ vanishes identically on the complement of its closed support $\mathcal{M} \setminus \overline{\text{supp}(h_{j,k})}$. Further, $\mathcal{M} \setminus \overline{\text{supp}(h_{j,k})}$ is an open set, so

  $$\text{grad} h_{j,k}(x) = 0, \quad \text{Hess} h_{j,k}(x) = 0 \quad \forall x \in \mathcal{M} \setminus \overline{\text{supp}(h_{j,k})}.$$

  Using $\text{grad} f_{j,k+1} = \text{grad} f_{j,k} + \text{grad} h_{j,k}$ and the inductive hypothesis **IH3**, this ensures that $f_{j,k+1}$ is consistent with the queries $x_0, \ldots, x_{k-1}$.

  In order to gain control of the gradient of $f_{j,k+1}$ at x_k, we also want the support of $h_{j,k}$ to contain x_k. So using that x_ℓ (5) is a past query point closest to x_k, it is enough to enforce that the support of $h_{j,k}$ remains in the ball $B(x_k, \frac{1}{4}\text{dist}(x_k, x_\ell))$. We are not done with **L3** but let us move on for now.

[8] We use the notation $\text{supp}(f) = \{x \in \mathcal{M} : f(x) \neq 0\}$ to denote the support of a function $f \colon \mathcal{M} \to \mathbb{R}$, and $\overline{S}$ to denote the closure of the set S.

- In the previous item we saw that if $k \geq 1$, we want the support of $h_{j,k}$ to be contained in a ball centered at x_k and whose radius is no more than $\frac{1}{4}\mathrm{dist}(x_k, x_\ell)$. For **L2**, it is convenient to require that this radius is no more than $\frac{r}{8}$. Precisely, we shall ensure that the support of $h_{j,k}$ is contained in the ball $B(x_k, R_{\mathrm{ball}}^{(k)})$ where
$$R_{\mathrm{ball}}^{(0)} = \frac{r}{8}, \qquad R_{\mathrm{ball}}^{(k)} = \min\left\{\frac{1}{4}\mathrm{dist}(x_k, x_\ell), \frac{r}{8}\right\} \quad \text{if } k \geq 1. \tag{8}$$
Let us now show that this choice of $R_{\mathrm{ball}}^{(k)}$ guarantees **L2**, i.e., $\mathrm{grad} f_{j,k+1}(z_j) = 0$. We know $\mathrm{grad} f_{j,k}(z_j) = 0$. Therefore, to satisfy **L2** it is sufficient to impose that the closed support of $h_{j,k}$ does not contain z_j. We know that $\mathrm{dist}(x_k, z_j) \geq \frac{r}{4}$ for all $j \in \tilde{A}_k$. Since $R_{\mathrm{ball}}^{(k)} \leq \frac{r}{8}$, we indeed have $z_j \notin B(x_k, R_{\mathrm{ball}}^{(k)})$ for all $j \in \tilde{A}_k$.
- Since $f_{j,k+1} = f_{j,k} + h_{j,k}$ (6) and $H_{j,k+1} = H_{j,k} + h_{j,k}$ due to (3),
$$\mathrm{Hess} f_{j,k} - \|\mathrm{Hess} h_{j,k}\|\, I \preceq \mathrm{Hess} f_{j,k+1} \preceq \mathrm{Hess} f_{j,k} + \|\mathrm{Hess} h_{j,k}\|\, I,$$
$$\|\mathrm{grad} H_{j,k+1}\| \leq \|\mathrm{grad} H_{j,k}\| + \|\mathrm{grad} h_{j,k}\|, \quad \|\mathrm{Hess} H_{j,k+1}\| \leq \|\mathrm{Hess} H_{j,k}\| + \|\mathrm{Hess} h_{j,k}\|.$$
Therefore (using Definition 9), for $f_{j,k+1}$ to satisfy **L1** and $H_{j,k+1}$ to satisfy **L5**, it is enough to require $\|\mathrm{Hess} h_{j,k}(x)\| \leq \frac{1}{4w}$ and $\|\mathrm{grad} h_{j,k}(x)\| \leq \frac{1}{4w\sqrt{-K_{\mathrm{lo}}}}$ for all $x \in \mathcal{M}$.

Since we are looking for a function $h_{j,k}$ with support contained in a ball, we are looking to construct a *bump function*, that is a C^∞ function on $\mathcal{M}$ whose closed support is compact. In Appendix E, we state and prove Lemma 17 which, given a point $x_k \in \mathcal{M}$ and a radius $R_{\mathrm{ball}} > 0$, provides a family of bump functions
$$\{h_g \colon \mathcal{M} \to \mathbb{R}\}_{g \in B_{x_k}(0, w^{-1} g_{\mathrm{norm}}(R_{\mathrm{ball}}))}, \tag{9}$$
such that for each g with $\|g\| \leq w^{-1} g_{\mathrm{norm}}(R_{\mathrm{ball}})$, the function h_g is supported in $B(x_k, R_{\mathrm{ball}})$ (property **BF2** in Lemma 17), has $\mathrm{grad} h_g(x_k) = g$ (property **BF1**), and in addition satisfies the bounds $\|\mathrm{Hess} h_g(x)\| \leq \frac{1}{4w}$ and $\|\mathrm{grad} h_g(x)\| \leq \frac{1}{4w\sqrt{-K_{\mathrm{lo}}}}$ for all $x \in \mathcal{M}$ (property **BF3**). Here $g_{\mathrm{norm}} \colon [0, \infty) \to \mathbb{R}$ is a certain univariate function with a simple explicit formula (equation (17)). (Remember that $B_{x_k}(0, w^{-1} g_{\mathrm{norm}}(R_{\mathrm{ball}})) \subseteq \mathrm{T}_{x_k}\mathcal{M}$ denotes a Euclidean ball, see Section 2.1.)

So far we have not chosen $g_k \in \mathrm{T}_{x_k}\mathcal{M}$. However, using the family of bump functions (9), we have shown for any choice of $g \in B_{x_k}(0, w^{-1} g_{\mathrm{norm}}(R_{\mathrm{ball}}^{(k)}))$ and $j \in \tilde{A}_k$, the function $f_{j,k+1} = f_{j,k} + h_g$ satisfies properties **L1**, **L2**, **L4**, **L5**, as well as $\mathrm{grad} f_{j,k+1}(x_m) = g_m$ for $m = 0, 1, \ldots k-1$ (this is part of property **L3**). It remains to choose $g_k \in \mathrm{T}_{x_k}\mathcal{M}$ and $A_{k+1} \subseteq \tilde{A}_k$ so that $\mathrm{grad} f_{j,k+1}(x_k) = g_k$ for all $j \in A_{k+1}$, and so that inequality (2) is satisfied.

Around each gradient $\mathrm{grad} f_{j,k}(x_k), j \in \tilde{A}_k$, there is a small ball $B_{j,k}$ which is the set of possible gradients of $f_{j,k+1} = f_{j,k} + h_g$ at x_k. More precisely, the balls $B_{j,k} \subseteq \mathrm{T}_{x_k}\mathcal{M}$ are defined by
$$B_{j,k} = B_{x_k}(\mathrm{grad} f_{j,k}(x_k), w^{-1} g_{\mathrm{norm}}(R_{\mathrm{ball}}^{(k)})), \qquad \forall j \in \tilde{A}_k. \tag{10}$$
Some of these balls overlap, and we want to choose $g_k \in \mathrm{T}_{x_k} M$ which simultaneously lies in as many of the balls as possible. This is because the oracle only gets to pick one g_k and we would like g_k to be compatible with as many $f_{j,k+1}, j \in \tilde{A}_k$, as possible. Therefore, choose
$$g_k \in \arg\max_{g \in \mathrm{T}_{x_k}\mathcal{M}} \left|\left\{j \in \tilde{A}_k : g \in B_{j,k}\right\}\right|. \tag{11}$$

Define $A_{k+1} = \{j \in \tilde{A}_k : g_k \in B_{j,k}\}$. The number of balls $\{B_{j,k}\}$ which intersect at the common vector g_k equals $|A_{k+1}|$. For each $j \in A_{k+1}$ the vector $g_{j,k} = g_k - \mathrm{grad} f_{j,k}(x_k)$ is in $B_{x_k}(0, w^{-1} g_{\mathrm{norm}}(R_{\mathrm{ball}}^{(k)}))$ and so

$$\mathrm{grad}(f_{j,k} + h_{g_{j,k}})(x_k) = \mathrm{grad} f_{j,k}(x_k) + \mathrm{grad} h_{g_{j,k}}(x_k) = \mathrm{grad} f_{j,k}(x_k) + g_{j,k} = g_k$$

using property **BF1** of Lemma 17. Therefore, defining $h_{j,k} = h_{g_{j,k}}$, we have

$$\mathrm{grad} f_{j,k+1}(x_k) = \mathrm{grad}(f_{j,k} + h_{j,k})(x_k) = g_k, \qquad \text{for all } j \in A_{k+1}.$$

It remains to show inequality (2), i.e., a large enough subset of the balls $B_{j,k}$ do indeed intersect at a common point. For this, we use a geometric lemma (short proof in Appendix C).

Lemma 13 *Consider n closed balls $B_1, \ldots, B_n \subseteq \mathbb{R}^d$ of radius q each, and assume each of the balls is also contained in a larger closed ball B of radius r: $B_j \subseteq B$ for all $j = 1, \ldots, n$. Choose $g \in \arg\max_{y \in B} |\{j \in \{1, \ldots, n\} : y \in B_j\}|$ and let $A = \{j \in \{1, \ldots, n\} : g \in B_j\}$. Then*

$$|A| \geq n \mathrm{Vol}(B_1)/\mathrm{Vol}(B) = nq^d/r^d.$$

To use Lemma 13, we need to find a ball $B_k \subseteq \mathrm{T}_{x_k}\mathcal{M}$ containing all the balls $B_{j,k}, j \in \tilde{A}_k$, and we want the radius of B_k to be small. This is the last step of the proof. Care has to be taken in bounding the radius of B_k because the distance between x_k and x_ℓ (5) can be arbitrarily small. If $\mathrm{dist}(x_k, x_\ell)$ is very small, then the radius of the balls $B_{j,k}$ is very small, and so we must show that the radius of B_k is also sufficiently small for Lemma 13 to be useful. We upper bound the radius of the ball B_k in the following two cases, showing that

$$\frac{\mathrm{Vol}(B_{j,k})}{\mathrm{Vol}(B_k)} \geq \frac{1}{(2000w(3\mathscr{R}\sqrt{-K_{\mathrm{lo}}} + 2))^d} \tag{12}$$

holds in each case. The most important part is to choose a good center for B_k:

Case 1: either $k = 0$, or $k \geq 1$ and $\sqrt{-K_{\mathrm{lo}}}\mathrm{dist}(x_k, x_\ell) > 4$. This captures the scenario where either there are no previous query points, or the algorithm queries x_k not close to any previous query. In this case, the overarching idea to upper bound the radius of B_k is as follows: $f_{j,k}$ is a perturbed version of the squared distance function $x \mapsto \frac{1}{2}\mathrm{dist}(x, z_j)^2$. Therefore, the gradient of $f_{j,k}$ at x_k (which is the center of $B_{j,k}$) is approximately $-\exp_{x_k}^{-1}(z_j)$ (see Lemma 10). The points $\{z_j\}_{j \in \tilde{A}_k}$ are clustered around x_{ref} in a ball of radius r. Therefore the vectors $\{-\exp_{x_k}^{-1}(z_j)\}_{j \in \tilde{A}_k}$ are clustered around $-\exp_{x_k}^{-1}(x_{\mathrm{ref}})$. Consequently the same is true for the gradients $\{\mathrm{grad} f_{j,k}(x_k)\}_{j \in \tilde{A}_k}$. We use this intuition to work out the details in Appendix D.1.

Case 2: $k \geq 1$ and $\sqrt{-K_{\mathrm{lo}}}\mathrm{dist}(x_k, x_\ell) \leq 4$. This captures the scenario where the algorithm queries x_k close to a previous query. In this case, the overarching idea to bound the radius of a ball B_k is as follows. All the functions $f_{j,k}, j \in \tilde{A}_k$, have the same gradient at x_ℓ, namely g_ℓ. Therefore, since $\mathrm{dist}(x_k, x_\ell)$ is small, gradient-Lipschitzness of the functions $f_{j,k}$ (see Definition 9) implies that the gradients $\{\mathrm{grad} f_{j,k}(x_k)\}_{j \in \tilde{A}_k}$ are all clustered around $P_{x_\ell \to x_k} g_\ell$ (the parallel transport of g_ℓ to $\mathrm{T}_{x_k}\mathcal{M}$). This intuition guides the details given in Appendix D.2.

After establishing inequality (12), we can use Lemma 13 to show that g_k (11) is contained in

$$|A_{k+1}| \geq |\tilde{A}_k| \frac{\mathrm{Vol}(B_{j,k})}{\mathrm{Vol}(B_k)} \geq \frac{|A_k| - 1}{(2000w(3\mathscr{R}\sqrt{-K_{\mathrm{lo}}} + 2))^d}$$

of the balls $B_{j,k}, j \in \tilde{A}_k$. This concludes the inductive step, proving Lemma 12.

Acknowledgments

We thank David Martínez-Rubio for helpful discussions and feedback on a version of this paper.

References

P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, Princeton, NJ, 2008. ISBN 978-0-691-13298-3.

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.

K. Ahn and F. Suarez. Riemannian perspective on matrix factorization. *arXiv: 2102.00937*, 2021.

S. Alexander, V. Kapovitch, and A. Petrunin. Alexandrov geometry: preliminary version no. 1. *arXiv: 1903.08539*, 2019.

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.

F. Alimisis, A. Orvieto, G. Becigneul, and A. Lucchi. Momentum improves optimization on riemannian manifolds. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 1351–1359. PMLR, 13–15 Apr 2021. URL http://proceedings.mlr.press/v130/alimisis21a.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.

C. Amendola, K. Kohn, P. Reichenbach, and A. Seigal. Invariant theory and scaling algorithms for maximum likelihood estimation. *SIAM Journal on Applied Algebra and Geometry*, 5(2):304–337, 2021.

C. Auderset, C. Mazza, and E.A. Ruh. Angular Gaussian and Cauchy estimation. *Journal of Multivariate Analysis*, 93(1):180–197, 2005. ISSN 0047-259X.

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.

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

R. Bhatia. *Positive definite matrices*. Princeton University Press, 2007.

N. Boumal. An introduction to optimization on smooth manifolds. To appear with Cambridge University Press, Jan 2022. URL http://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.

D. Burago, Y. Burago, and S. Ivanov. *A course in metric geometry*. Graduate studies in mathematics. American Mathematical Society, Providence (R.I.), 2001. ISBN 0-8218-2129-6.

P. Bürgisser, C. Franks, A. Garg, R. Oliveira, M. Walter, and A. Wigderson. Towards a theory of non-commutative optimization: Geodesic 1st and 2nd order methods for moment maps and polytopes. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 845–861, 2019. doi: 10.1109/FOCS.2019.00055.

Y. Carmon, J.C. Duchi, O. Hinder, and A. Sidford. Lower bounds for finding stationary points I. *Mathematical Programming*, 2019. doi: 10.1007/s10107-019-01406-y.

P. Chossat and O. Faugeras. Hyperbolic planforms in relation to visual edges and textures perception. *PLoS Computational Biology*, 5(12):e1000625, Dec 2009. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1000625. URL http://dx.doi.org/10.1371/journal.pcbi.1000625.

C. Ciobotaru and C. Mazza. Geometrical and statistical properties of m-estimates of scatter on grassmann manifolds. *arXiv: 1812.11605*, 2020.

C. Criscitiello and N. Boumal. An accelerated first-order method for non-convex optimization on manifolds. *arXiv 2008.02252*, 2020.

A. Dolcetti and D. Pertici. Differential properties of spaces of symmetric real matrices. *arXiv: 1807.01113*, 2018.

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.

V. Duruisseaux and M. Leok. A variational formulation of accelerated optimization on riemannian manifolds. *arXiv: 2101.06552*, 2021.

P. T. Fletcher and S. Joshi. Riemannian geometry for the statistical analysis of diffusion tensor data. *Signal Process.*, 87(2):250–262, feb 2007. ISSN 0165-1684. doi: 10.1016/j.sigpro.2005.12.018. URL https://doi.org/10.1016/j.sigpro.2005.12.018.

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.

G. Franca, A. Barp, M. Girolami, and M. I. Jordan. Optimization on manifolds: A symplectic approach. *arXiv: 2107.11231*, 2021a.

G. Franca, M. I. Jordan, and R. Vidal. On dissipative symplectic integration with applications to gradient-based optimization. *arXiv: 2004.06840*, 2021b.

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.

C. Franks and P. Reichenbach. Barriers for recent methods in geodesic optimization. In *Proceedings of the 36th Computational Complexity Conference (CCC 2021)*, pages 13:1–13:54, 2021.

C. Franks, R. Oliveira, A. Ramachandran, and M. Walter. Near optimal sample complexity for matrix and tensor normal models via geodesic convexity. *arXiv: 2110.07583*, 2021.

N. Goyal and A. Shetty. Sampling and optimization on convex sets in riemannian manifolds of non-negative curvature. In *32nd Annual Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 1–43. PMLR, 2019. URL http://proceedings.mlr.press/v99/goyal19a/goyal19a.pdf.

X. Gual-Aenau and A. M. Naveira. Volume of tubes in noncompact symmetric spaces. *Publ. Math. Debrecen*, 54:313–320, 1999.

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

R. Hosseini and S. Sra. Matrix manifold optimization for Gaussian mixtures. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 910–918. Curran Associates, Inc., 2015. URL http://papers.nips.cc/paper/5812-matrix-manifold-optimization-for-gaussian-mixtures.pdf.

J. Hu, X. Liu, Z.-W. Wen, and Y.-X. Yuan. A brief introduction to manifold optimization. *Journal of the Operations Research Society of China*, 8(2):199–248, April 2020. doi: 10.1007/s40305-020-00295-9.

W. Huang and K. Wei. An extension of fista to riemannian optimization for sparse pca. *arXiv: 1909.05485*, 2021.

J. Jin and S. Sra. A riemannian accelerated proximal extragradient framework and its implications. *arXiv: 2111.02763*, 2021.

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

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.

C. Lenglet, M. Rousson, R. Deriche, and O. Faugeras. Statistics on the manifold of multivariate normal distributions: Theory and application to diffusion tensor mri processing. *J. Math. Imaging Vis.*, 25(3):423–444, oct 2006. ISSN 0924-9907. doi: 10.1007/s10851-006-6897-z. URL https://doi.org/10.1007/s10851-006-6897-z.

M. Lezcano-Casado. Adaptive and momentum methods on manifolds through trivializations. *arXiv: 2010.04617*, 2020.

M. Lezcano-Casado and D. Martínez-Rubio. Cheap orthogonal constraints in neural networks: A simple parametrization of the orthogonal and unitary group. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3794–3803. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/lezcano-casado19a.html.

Y. Liu, F. Shang, J. Cheng, H. Cheng, and L. Jiao. Accelerated first-order methods for geodesically convex optimization on Riemannian manifolds. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/6ef80bb237adf4b6f77d0700e1255907-Paper.pdf.

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

M. Moakher. A differential geometric approach to the geometric mean of symmetric positive-definite matrices. *SIAM J. Matrix Anal. Appl.*, 26(3):735–747, March 2005. doi: 10.1137/S0895479803436937.

M. Moakher and P.G. Batchelor. *Symmetric Positive-Definite Matrices: From Geometry to Applications and Visualization*, pages 285–298. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006. doi: 10.1007/3-540-31272-217.

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.

V. A. Nguyen, S. Shafieezadeh-Abadeh, M.-C. Yue, D. Kuhn, and W. Wiesemann. Calculating optimistic likelihoods using (geodesically) convex optimization. 2019.

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

L. T. Skovgaard. A riemannian geometry of the multivariate normal model. *Scandinavian Journal of Statistics*, 11(4):211–223, 1984. URL https://www.jstor.org/stable/4615960.

S. Sra and R. Hosseini. Conic geometric optimization on the manifold of positive definite matrices. *SIAM Journal on Optimization*, 25(1):713–739, 2015. doi: 10.1137/140978168.

S. Sra, N. K. Vishnoi, and O. Yildiz. On geodesically convex formulations for the brascamp-lieb constant. In Eric Blais, Klaus Jansen, José D. P. Rolim, and David Steurer, editors, *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, APPROX/RANDOM 2018, August 20-22, 2018 - Princeton, NJ, USA*, volume 116 of *LIPIcs*, pages 25:1–25:15. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2018. doi: 10.4230/LIPIcs.APPROX-RANDOM.2018.25. URL https://doi.org/10.4230/LIPIcs.APPROX-RANDOM.2018.25.

T. M. Tang and G. I. Allen. Integrated principal components analysis. *arXiv: 1810.00832*, 2021.

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.

R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018. doi: 10.1017/9781108231596.

A. Wiesel. Geodesic convexity and covariance estimation. *IEEE Transactions on Signal Processing*, 60(12):6182–6189, 2012. doi: 10.1109/TSP.2012.2218241.

A. Wiesel and T. Zhang. Structured robust covariance estimation. *Foundations and Trends in Signal Processing*, 8(3):127–216, 2014. doi: 10.1561/2000000053.

X. Yuan, W. Huang, P.-A. Absil, and K. A. Gallivan. Computing the matrix geometric mean: Riemannian versus euclidean conditioning, implementation techniques, and a riemannian BFGS method. *Numer. Linear Algebra Appl.*, 27(5), 2020. doi: 10.1002/nla.2321. URL https://doi.org/10.1002/nla.2321.

P. H. Zadeh, R. Hosseini, and S. Sra. Geometric mean metric learning. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning*, ICML, pages 2464–2471. JMLR.org, 2016. URL http://dl.acm.org/citation.cfm?id=3045390.3045650.

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.

T. Zhang. Robust subspace recovery by geodesically convex optimization. 2012. doi: 10.1.1.759.6576.

A Further literature review

Liu et al. (2017) were the first to claim acceleration on Riemannian manifolds. However, their algorithm requires solving nonlinear equations at each iteration which a priori might be as difficult as the optimization problem itself.

The results of Ahn and Sra (2020) mentioned previously are an improvement on the results of Zhang and Sra (2018) who also show that acceleration is possible if the initial iterate is sufficiently close to the minimizer x^*. Ahn and Sra (2020)'s algorithm additionally converges globally. Jin and Sra (2021) propose a framework for generating and analyzing eventually-accelerated algorithms; the algorithm of Ahn and Sra (2020) is an instance of this framework.

Martínez-Rubio (2021) presents algorithms for acceleration on spheres and hyperbolic spaces. For hyperbolic spaces, Martínez-Rubio (2021) proves the rates $e^{\tilde{O}(r)}\sqrt{\kappa}$ for the strongly g-convex case, and $e^{\tilde{O}(r)}\frac{1}{\sqrt{\epsilon}}$ for the nonstrongly g-convex case (to find a point x satisfying $f(x) - f(x^*) \leq \epsilon \cdot \frac{1}{2}Lr^2$). The key idea consists of pulling back the optimization problem to a vector space via a geodesic map; the pullback satisfies a relaxed notion of convexity. This idea is similar to the method of trivializations, introduced in (Lezcano-Casado and Martínez-Rubio, 2019) and applied to momentum methods in (Lezcano-Casado, 2020).

Alimisis et al. (2021) tackle the problem of acceleration on the class of smooth nonstrongly g-convex functions. In certain scenarios (when r and curvature are sufficiently small), their algorithm outperforms RGD; however, in general their algorithm requires $O(\frac{r^2}{\epsilon})$ iterations to solve the problem, which is not better than RGD. Nevertheless, the experimental results of Alimisis et al. (2021) show promise. Huang and Wei (2021) develop an algorithm for Riemannian optimization based on FISTA (Beck and Teboulle, 2009) which also demonstrates promising experimental results.

Alimisis et al. (2020) construct an ordinary differential equation (ODE) to model a Riemmanian version of Nesterov's accelerated gradient method. They prove that this ODE achieves an accelerated rate. It is unclear whether the discretization of this ODE preserves a similar acceleration. Recently, techniques from dynamical systems and symplectic geometry have been used to derive ODEs (Duruisseaux and Leok, 2021) and discretize such ODEs to obtain algorithms on Riemannian manifolds (Franca et al., 2021a,b). It is also unclear whether such algorithms achieve acceleration.

In stark contrast to the results presented in this paper, on non-convex functions it is possible to achieve acceleration *for finding (first- and second-order) critical points* on Riemannian manifolds, even negatively curved manifolds (Criscitiello and Boumal, 2020).

One can also sometimes model non-Euclidean geometries using a Bregman distance function; such geometries have different properties than Riemannian geometries. For example, the functions in consideration are still convex in the Euclidean sense (unlike g-convex functions); the Bregman geometry alters the notion smoothness and conditioning of these functions. Dragomir et al. (2021) recently showed that acceleration is not possible in this setting. The techniques they use and the key geometric obstructions to acceleration are significantly different from the Riemannian setting.

B Useful geometric propositions, and characterizations of g-convexity and smoothness

In the appendices, we use the following geometric propositions, which are consequences of the Euclidean law of cosines, the hyperbolic law of cosines (Ratcliffe, 2019, Thm. 3.5.3), and Toponogov's triangle comparison theorem (see (Lee, 2018, Thm. 11.10), (Burago et al., 2001, Sec. 6.5), or (Alexander et al., 2019, Thm. 8.13.3)). In the appendices, we also use the equivalent characterizations of μ-strong g-convexity and L-smoothness given in Lemma 16.

Proposition 14 *Let $\mathcal{M}$ be a Hadamard manifold. Let xyz be a geodesic triangle of $\mathcal{M}$ with vertices $x, y, z \in \mathcal{M}$ and side lengths* $\mathrm{dist}(y,z) = a, \mathrm{dist}(x,z) = b, \mathrm{dist}(x,y) = c$. *Also let the angle at x be α, i.e., $\alpha = \arccos\left(\frac{\langle \exp_x^{-1}(y), \exp_x^{-1}(z)\rangle}{\mathrm{dist}(x,y)\mathrm{dist}(x,z)}\right)$. Then $a^2 \geq b^2 + c^2 - 2bc\cos(\alpha)$ (Lee, 2018,* *Prop. 12.10). Equivalently,*

$$\mathrm{dist}(y,z)^2 \geq \mathrm{dist}(x,z)^2 + \mathrm{dist}(x,y)^2 - 2\left\langle \exp_x^{-1}(y), \exp_x^{-1}(z)\right\rangle = \left\|\exp_x^{-1}(y) - \exp_x^{-1}(z)\right\|^2.$$

Proposition 15 *Consider the same setting as Proposition 14. In addition, assume the sectional curvatures of $\mathcal{M}$ are in the interval $(-\infty, K_{\mathrm{up}}]$ with $K_{\mathrm{up}} < 0$. Then*

$$\cosh(a\sqrt{-K_{\mathrm{up}}}) \geq \cosh(b\sqrt{-K_{\mathrm{up}}})\cosh(c\sqrt{-K_{\mathrm{up}}}) - \sinh(b\sqrt{-K_{\mathrm{up}}})\sinh(c\sqrt{-K_{\mathrm{up}}})\cos(\alpha).$$

Lemma 16 *Let $\mathcal{M}$ be a Hadamard manifold, and $D \subseteq \mathcal{M}$ be a g-convex set. Let $f\colon \mathcal{M} \to \mathbb{R}$ be twice continuously differentiable. With reference to Definition 9:*

- *If f is μ-strongly g-convex in D then $f(y) \geq f(x) + \left\langle \mathrm{grad} f(x), \exp_x^{-1}(y)\right\rangle + \frac{\mu}{2}\mathrm{dist}(x,y)^2$ for all $x, y \in D$.*
- *If f is L-smooth in D then $\left|f(y) - f(x) - \left\langle \mathrm{grad} f(x), \exp_x^{-1}(y)\right\rangle\right| \leq \frac{L}{2}\mathrm{dist}(x,y)^2$ for all $x, y \in D$.*

C Proof of the simple geometric lemma 13

For each $x \in B$, let $N(x)$ be the number of smaller balls which contain x:

$$N(x) = \left|\{j \in \{1, \ldots, n\} : x \in B_j\}\right|.$$

Therefore, $g \in \arg\max_{y \in B} N(y)$. The sum of the volumes of the smaller balls is

$$n\mathrm{Vol}(B_1) = \int_{x \in B} N(x) dV(x) \leq \int_{x\in B}\left(\max_{y \in B} N(y)\right) dV(x) = \left(\max_{y\in B} N(y)\right)\mathrm{Vol}(B).$$

So $|A| = \max_{y\in B} N(y) \geq n\mathrm{Vol}(B_1)/\mathrm{Vol}(B) = nq^d/r^d$.

D Details for Cases 1 and 2 in proof of the key lemma 12

D.1 Case 1: x_k is not close to any previous query point

By Proposition 14, nonpositive curvature yields $\left\|\exp_{x_k}^{-1}(z_j) - \exp_{x_k}^{-1}(x_{\mathrm{ref}})\right\| \leq \mathrm{dist}(z_j, x_{\mathrm{ref}}) \leq r$. By the inductive hypothesis and the assumptions $k \leq 2w$ and $r\sqrt{-K_{\mathrm{lo}}} \geq 8$,

$$\|\mathrm{grad} H_{j,k}(x_k)\| \leq k\frac{1}{4w\sqrt{-K_{\mathrm{lo}}}} \leq \frac{1}{2\sqrt{-K_{\mathrm{lo}}}} \leq \frac{r}{2}.$$

Therefore, using the definition of $H_{j,k}$ (equation (3)),

$$\left\|\mathrm{grad} f_{j,k}(x_k) - \exp_{x_k}^{-1}(x_{\mathrm{ref}})\right\| = \left\|\exp_{x_k}^{-1}(z_j) + \mathrm{grad} H_{j,k}(x_k) - \exp_{x_k}^{-1}(x_{\mathrm{ref}})\right\| \leq \frac{3r}{2}.$$

Since $g_{\mathrm{norm}}(\cdot)$ (17) from Lemma 17 is increasing on $[0, \infty)$,

$$w^{-1} g_{\mathrm{norm}}(R_{\mathrm{ball}}^{(k)}) \leq w^{-1}\lim_{t\to\infty} g_{\mathrm{norm}}(t) = \frac{w^{-1}}{9e^{1/3}\sqrt{-K_{\mathrm{lo}}}} \leq \frac{1}{9e^{1/3}\sqrt{-K_{\mathrm{lo}}}} \leq \frac{r}{9e^{1/3}} \leq \frac{r}{2}.$$

So each $B_{j,k}$ is contained in the ball

$$B_{x_k}\left(\exp_{x_k}^{-1}(x_{\mathrm{ref}}), \frac{3r}{2} + w^{-1}g_{\mathrm{norm}}(R_{\mathrm{ball}}^{(k)})\right) \subseteq B_{x_k}(\exp_{x_k}^{-1}(x_{\mathrm{ref}}), 2r).$$

Defining, $B_k = B_{x_k}(\exp_{x_k}^{-1}(x_{\mathrm{ref}}), 2r)$, we have $\frac{\mathrm{Vol}(B_{j,k})}{\mathrm{Vol}(B_k)} = \frac{w^{-d}g_{\mathrm{norm}}(R_{\mathrm{ball}}^{(k)})^d}{(2r)^d}$.

Now assume $k = 0$ or $\sqrt{-K_{\mathrm{lo}}}\mathrm{dist}(x_k, x_\ell) > 4$, where x_ℓ is defined in equation (5). Using $r\sqrt{-K_{\mathrm{lo}}} \geq 8$ and the definition of $R_{\mathrm{ball}}^{(k)}$ (8), this assumption implies $R_{\mathrm{ball}}^{(k)} \geq 1/\sqrt{-K_{\mathrm{lo}}}$. Thus,

$$g_{\mathrm{norm}}(R_{\mathrm{ball}}^{(k)}) \geq g_{\mathrm{norm}}(1/\sqrt{-K_{\mathrm{lo}}}) \geq \frac{8/\sqrt{-K_{\mathrm{lo}}}}{e^{1/3}(1485+72)} \geq \frac{1}{300\sqrt{-K_{\mathrm{lo}}}}.$$

We conclude

$$\frac{\mathrm{Vol}(B_{j,k})}{\mathrm{Vol}(B_k)} = \frac{w^{-d}g_{\mathrm{norm}}(R_{\mathrm{ball}}^{(k)})^d}{(2r)^d} \geq \frac{w^{-d}}{(300\sqrt{-K_{\mathrm{lo}}})^d \cdot (2r)^d} = \frac{1}{(600wr\sqrt{-K_{\mathrm{lo}}})^d}. \tag{13}$$

D.2 Case 2: x_k is close to a previous query point

Assume $\sqrt{-K_{\mathrm{lo}}}\mathrm{dist}(x_k, x_\ell) \leq 4$, where x_ℓ is defined in equation (5). Since $r \geq 8/\sqrt{-K_{\mathrm{lo}}}$, we have $R_{\mathrm{ball}}^{(k)} = \frac{1}{4}\mathrm{dist}(x_k, x_\ell)$.

By assumption, $\mathrm{dist}(x_k, x_{\mathrm{ref}}) \leq \mathscr{R}$. Since $\mathrm{dist}(z_j, x_{\mathrm{ref}}) \leq r$ and $\mathrm{dist}(x_k, x_\ell) \leq 4/\sqrt{-K_{\mathrm{lo}}} \leq r/2$, the triangle inequality implies $\mathrm{dist}(x_k, z_j)$ and $\mathrm{dist}(x_\ell, z_j)$ are both at most $3\mathscr{R}$. The inductive hypothesis **IH5** implies $f_{j,k}(x) = \frac{1}{2}\mathrm{dist}(x, z_j)^2 + H_{j,k}(x)$ and $\|\mathrm{Hess}H_{j,k}(x)\| \leq \frac{k}{4w} \leq \frac{1}{2}$ for all $x \in \mathcal{M}$, using $k \leq 2w$. So by Lemma 10,

$$\|\mathrm{Hess}f_{j,k}(x)\| \leq \max\{\mathrm{dist}(x_k, z_j), \mathrm{dist}(x_\ell, z_j)\}\sqrt{-K_{\mathrm{lo}}} + \frac{3}{2} \leq 3\mathscr{R}\sqrt{-K_{\mathrm{lo}}} + \frac{3}{2} \tag{14}$$

for all $x \in B(z_j, \max\{\mathrm{dist}(x_k, z_j), \mathrm{dist}(x_\ell, z_j)\})$. Additionally, the inductive hypothesis **IH3** implies $\mathrm{grad}f_{j,k}(x_\ell) = g_\ell$ for all $j \in \tilde{A}_k$. Therefore, by Definition 9:

$$\|\mathrm{grad}f_{j,k}(x_k) - P_{x_\ell \to x_k}g_\ell\| \leq \left(3\mathscr{R}\sqrt{-K_{\mathrm{lo}}} + \frac{3}{2}\right)\mathrm{dist}(x_k, x_\ell) \quad \text{for all } j \in \tilde{A}_k.$$

We have shown that all the gradients $\mathrm{grad}f_{j,k}(x_k), j \in \tilde{A}_k$, are contained in a ball in $\mathrm{T}_{x_k}\mathcal{M}$ centered at $P_{x_\ell \to x_k}g_\ell$ with radius $(3\mathscr{R}\sqrt{-K_{\mathrm{lo}}} + \frac{3}{2})\mathrm{dist}(x_k, x_\ell)$.

Recall the definition of the balls $B_{j,k}$ defined in equation (10). We conclude all the balls $B_{j,k}, j \in \tilde{A}_k$, are contained in a ball $B_k \subseteq \mathrm{T}_{x_k}\mathcal{M}$ centered at $P_{x_\ell \to x_k}g_\ell$ with radius

$$\left(3\mathscr{R}\sqrt{-K_{\mathrm{lo}}} + \frac{3}{2}\right)\mathrm{dist}(x_k, x_\ell) + w^{-1}g_{\mathrm{norm}}\left(\frac{1}{4}\mathrm{dist}(x_k, x_\ell)\right) \leq (3\mathscr{R}\sqrt{-K_{\mathrm{lo}}} + 2)\mathrm{dist}(x_k, x_\ell),$$

using that $w \geq 1$ and $g_{\mathrm{norm}}\left(\frac{1}{4}\mathrm{dist}(x_k, x_\ell)\right) \leq \frac{8\mathrm{dist}(x_k,x_\ell)/4}{e^{1/3}(1485)} \leq \frac{\mathrm{dist}(x_k,x_\ell)}{1000}$. Therefore,

$$\begin{aligned}\frac{\mathrm{Vol}(B_{j,k})}{\mathrm{Vol}(B_k)} &= \frac{w^{-d}g_{\mathrm{norm}}(\mathrm{dist}(x_k, x_\ell)/4)^d}{((3\mathscr{R}\sqrt{-K_{\mathrm{lo}}} + 2)\mathrm{dist}(x_k, x_\ell))^d} \geq \frac{w^{-d}\mathrm{dist}(x_k, x_\ell)^d}{(2000(3\mathscr{R}\sqrt{-K_{\mathrm{lo}}} + 2)\mathrm{dist}(x_k, x_\ell))^d} \\ &= \frac{1}{(2000w(3\mathscr{R}\sqrt{-K_{\mathrm{lo}}} + 2))^d},\end{aligned} \tag{15}$$

using $g_{\mathrm{norm}}\left(\frac{1}{4}\mathrm{dist}(x_k, x_\ell)\right) \geq \frac{2\mathrm{dist}(x_k,x_\ell)}{e^{1/3}(1485+72)} \geq \frac{\mathrm{dist}(x_k,x_\ell)}{2000}$, which is due to $\sqrt{-K_{\mathrm{lo}}}\mathrm{dist}(x_k, x_\ell) \leq 4$.

E Bump functions

Lemma 17 (Family of bump functions) *Let $\mathcal{M}$ be a Hadamard manifold with sectional curvatures in the interval $[K_{\mathrm{lo}}, 0]$ with $K_{\mathrm{lo}} < 0$. Let $R_{\mathrm{ball}} > 0, w > 0$, $x_k \in \mathcal{M}$. Define*

$$a\colon [0,\infty) \to \mathbb{R}, \qquad a(R) = \frac{R}{4(4\sqrt{-K_{\mathrm{lo}}} + 55/R)}, \tag{16}$$

$$g_{\mathrm{norm}}\colon [0,\infty) \to \mathbb{R}, \qquad g_{\mathrm{norm}}(R) = \frac{8R}{e^{1/3}(1485 + 72R\sqrt{-K_{\mathrm{lo}}})}. \tag{17}$$

There is a family of functions $\{h_g\colon \mathcal{M} \to \mathbb{R}\}$, indexed by $g \in B_{x_k}(0, w^{-1}g_{\mathrm{norm}}(R_{\mathrm{ball}}))$, satisfying for each $g \in B_{x_k}(0, w^{-1}g_{\mathrm{norm}}(R_{\mathrm{ball}})) \subseteq \mathrm{T}_{x_k}\mathcal{M}$:

BF1 $\mathrm{grad} h_g(x_k) = g$*;*

BF2 *the support of each h_g is contained in $B(x_k, R_{\mathrm{ball}})$;*

BF3 $\|\mathrm{grad} h_g(x)\| \leq \frac{1}{4w\sqrt{-K_{\mathrm{lo}}}}$, $\|\mathrm{Hess} h_g(x)\| \leq \frac{1}{4w}$ *for all $x \in \mathcal{M}$;*

BF4 $h_g(x_k) = \frac{3}{8}\|g\|\, g_{\mathrm{norm}}^{-1}(w\|g\|)$, *and $|h_g(x)| \leq w^{-1}a(R_{\mathrm{ball}})$ for all $x \in \mathcal{M}$.*

E.1 Proof of Lemma 17

Define $\phi_R\colon \mathbb{R} \to \mathbb{R}$ by

$$\phi_R(t) = e \cdot \exp\left(-1\Big/\left(1 - \frac{2t}{R^2}\right)\right) = \exp(2t/(2t - R^2)) \qquad \text{for } t \in (-\infty, R^2/2),$$

and $\phi_R(t) = 0$ elsewhere. The function $\phi_R\colon \mathbb{R} \to \mathbb{R}$ is C^∞ (Lee, 2012, Lem. 2.20). We consider bump functions supported in $B(p, R)$ of the form $h(x) = a \cdot \phi_R(\mathrm{dist}(x,p)^2/2)$ for $a \in \mathbb{R}$. As a composition of C^∞ functions, these bump functions are also C^∞.

Remark 18 *Since $\mathrm{dist}(x,p)^2/2 \geq 0$, the values of ϕ_R on $(-\infty, 0)$ are irrelevant. All that matters is that ϕ_R is infinitely differentiable in a neighborhood of the origin.*

We have $\phi(R^2/8) = e^{-1/3}$, and for $t \in (-\infty, R^2/2)$

$$\phi_R'(t) = -\phi_R(t) \cdot 2R^2/(R^2 - 2t)^2, \qquad \phi_R''(t) = \phi_R(t) \cdot 4R^2(4t - R^2)/(R^2 - 2t)^4.$$

We have partitioned the proof of Lemma 17 into several subsections: E.1.1, E.1.2 and E.1.3.

E.1.1 BUMP FUNCTION CONSTRUCTION

For each $p \in B(x_k, R_{\mathrm{ball}}/3)$: let $R = 2\mathrm{dist}(x_k, p)$ and define $\tilde{h}_{x_k,p} : \mathcal{M} \to \mathbb{R}$ by

$$\tilde{h}_{x_k,p}(x) = w^{-1}a(R)\phi_R(\mathrm{dist}(x,p)^2/2).$$

By construction, $\tilde{h}_{x_k,p}$ is supported in the closed ball $B(p,R) \subseteq B(x_k, R_{\mathrm{ball}})$.

We have

$$\mathrm{grad}\tilde{h}_{x_k,p}(x) = -w^{-1}a(R)\phi_R'(\mathrm{dist}(x,p)^2/2)\exp_x^{-1}(p).$$

So using that $\left\|\exp_{x_k}^{-1}(p)\right\| = \mathrm{dist}(x_k, p) = R/2$,

$$\begin{aligned}\left\|\mathrm{grad}\tilde{h}_{x_k,p}(x_k)\right\| &= w^{-1}a(R)|\phi_R'(R^2/8)|R/2 \\ &= w^{-1}a(R)\phi(R^2/8)\frac{32}{9R^2}R/2 = w^{-1}e^{-1/3}a(R)\frac{16}{9R} \\ &= w^{-1}\frac{4e^{-1/3}}{9R(4\sqrt{-K_{\mathrm{lo}}}/R + 55/R^2)} = w^{-1}g_{\mathrm{norm}}\left(\frac{3}{2}R\right)\end{aligned} \tag{18}$$

where R can take any value in $[0, 2R_{\mathrm{ball}}/3]$. Since the function g_{norm} is strictly increasing on $[0, \infty)$ and $g_{\mathrm{norm}}(0) = 0$, we see that $\left\|\mathrm{grad}\tilde{h}_{x_k,p}(x_k)\right\|$ takes all values in the interval $[0, w^{-1}g_{\mathrm{norm}}(R_{\mathrm{ball}})]$ (as p varies).

On the other hand, $\mathrm{grad}\tilde{h}_{x_k,p}(x_k) = \left\|\mathrm{grad}\tilde{h}_{x_k,p}(x_k)\right\| \frac{\exp_{x_k}^{-1}(p)}{\left\|\exp_{x_k}^{-1}(p)\right\|}$. Therefore,

$$\{\mathrm{grad}\tilde{h}_{x_k,p}(x_k) : p \in B(x_k, R_{\mathrm{ball}}/3)\} = B_{x_k}(0, w^{-1}g_{\mathrm{norm}}(R_{\mathrm{ball}})).$$

More precisely, for each $g \in B_{x_k}(0, w^{-1}g_{\mathrm{norm}}(R_{\mathrm{ball}}))$ there is exactly one $p \in B(x_k, R_{\mathrm{ball}}/3)$ such that $g = \mathrm{grad}\tilde{h}_{x_k,p}(x_k)$, and vice versa. Finally, define

$$h_{\mathrm{grad}\tilde{h}_{x_k,p}(x_k)} := \tilde{h}_{x_k,p} \quad \forall p \in B(x_k, R_{\mathrm{ball}}/3).$$

This defines the family of functions in Lemma 17, and also establishes property **BF1**. By construction, property **BF2** is also satisfied.

We calculate

$$\begin{aligned}h_{\mathrm{grad}\tilde{h}_{x_k,p}(x_k)}(x_k) = \tilde{h}_{x_k,p}(x_k) &= w^{-1}a(R)\phi_R(\mathrm{dist}(x_k,p)^2/2) = w^{-1}a(R)\phi_R(R^2/8) \\ &= w^{-1}a(R)e^{-1/3} = \left\|\mathrm{grad}\tilde{h}_{x_k,p}(x_k)\right\|\frac{9R}{16} \\ &= \left\|\mathrm{grad}\tilde{h}_{x_k,p}(x_k)\right\|\frac{9}{16}\cdot\frac{2}{3}g_{\mathrm{norm}}^{-1}\left(w\left\|\mathrm{grad}\tilde{h}_{x_k,p}(x_k)\right\|\right)\end{aligned}$$

where we used equation (18) for the last two equalities. This shows the first part of property **BF4**.

E.1.2 BOUNDING THE FUNCTION VALUES AND GRADIENTS OF THE BUMP FUNCTIONS

The maximum of $\left|\tilde{h}_{x_k,p}(x)\right|$ is attained when $x = p$ and equals $w^{-1}a(R)$, which is at most $w^{-1}a(2R_{\mathrm{ball}}/3) \leq w^{-1}a(R_{\mathrm{ball}})$. This shows the second part of property **BF4**.

By Section E.1.1, we know that

$$\mathrm{grad}\tilde{h}_{x_k,p}(x) = -w^{-1}a(R)\phi_R'(\mathrm{dist}(x,p)^2/2)\exp_x^{-1}(p)$$

if $\mathrm{dist}(x,p) \leq R$ and $\mathrm{grad}\tilde{h}_{x_k,p}(x) = 0$ otherwise. Therefore, for any $x \in \mathcal{M}$,

$$\begin{aligned}\left\|\mathrm{grad}\tilde{h}_{x_k,p}(x)\right\| &\leq w^{-1}a(R)\max_{t\in[0,R]}\left|\phi_R'(t^2/2)\right|t \\ &= w^{-1}a(R)\max_{t\in[0,R]} t\cdot\phi_R(t^2/2)\cdot 2R^2/(R^2-t^2)^2\end{aligned}$$

It is easy to see that the maximizer of this problem is $t = R/3^{1/4}$, which yields

$$\begin{aligned}\left\|\mathrm{grad}\tilde{h}_{x_k,p}(x)\right\| &\leq w^{-1}a(R)R^{-1}\sqrt{36+21\sqrt{3}e^{-1/2-\sqrt{3}/2}} \leq 3w^{-1}a(R)R^{-1}\\ &\leq w^{-1}\frac{R}{55+4R\sqrt{-K_{\mathrm{lo}}}} \leq \frac{1}{4w\sqrt{-K_{\mathrm{lo}}}}.\end{aligned}$$

This proves the first part of property **BF3**.

E.1.3 BOUNDING THE HESSIAN OF THE BUMP FUNCTIONS

For $v \in \mathrm{T}_xM$ and $\mathrm{dist}(x,p) \leq R$, we have

$$\begin{aligned}\langle v, \mathrm{Hess}\tilde{h}_{x_k,p}(x)v\rangle &= w^{-1}a(R)\phi_R''(t^2/2)\langle v, -\exp_x^{-1}(p)\rangle^2 &&+ w^{-1}a(R)\phi_R'(t^2/2)\langle v, \mathscr{H}(x)v\rangle\\ &= (\text{term } 1) &&+ (\text{term } 2)\end{aligned}$$

where $t = \mathrm{dist}(x,p)$ and $\mathscr{H}(x)$ is the Hessian of the function $x \mapsto \mathrm{dist}(x,p)^2/2$.

We have $t = \mathrm{dist}(x,p) = \left\|\exp_x^{-1}(p)\right\|$ and

$$\|\mathscr{H}(x)\| \leq 1 + \mathrm{dist}(x,p)\sqrt{-K_{\mathrm{lo}}} = 1 + t\sqrt{-K_{\mathrm{lo}}},$$

by Lemma 10. So for $t \in [0,R]$ we have that

$$\begin{aligned}|\text{term } 1| &\leq (w^{-1}a(R)\|v\|^2)\left[\phi(t^2/2)\frac{4R^2\left|2t^2-R^2\right|}{(R^2-t^2)^4}t^2\right] \leq (w^{-1}a(R)\|v\|^2)\left[4R^6\frac{\phi(t^2/2)}{(R^2-t^2)^4}\right]\\ &\leq (w^{-1}a(R)\|v\|^2)\left[4R^6\cdot 256e^{-3}/R^8\right] \leq (w^{-1}a(R)\|v\|^2)\left[51/R^2\right], \text{and}\end{aligned}$$

$$\begin{aligned}|\text{term } 2| &\leq (w^{-1}a(R)\|v\|^2)\left[\phi(t^2/2)(1+t\sqrt{-K_{\mathrm{lo}}})\frac{2R^2}{(R^2-t^2)^2}\right]\\ &\leq (w^{-1}a(R)\|v\|^2)\left[2(1+R\sqrt{-K_{\mathrm{lo}}})R^2\frac{\phi(t^2/2)}{(R^2-t^2)^2}\right]\\ &\leq (w^{-1}a(R)\|v\|^2)\left[2(1+R\sqrt{-K_{\mathrm{lo}}})R^2\cdot 4e^{-1}/R^4\right]\\ &\leq (w^{-1}a(R)\|v\|^2)\left[4\sqrt{-K_{\mathrm{lo}}}/R + 4/R^2\right].\end{aligned}$$

Of course $\|\mathrm{Hess}\tilde{h}_{x_k,p}(x)\| = 0$ for $x \notin B(p,R)$. So for all x, we have $\|\mathrm{Hess}\tilde{h}_{x_k,p}(x)\| \leq w^{-1}a(R)(4\sqrt{-K_{\mathrm{lo}}}/R + 55/R^2) = \frac{1}{4w}$. We have proven the second part of property **BF3**.

E.2 Bump functions for function-value queries

In this section, we construct a family of bump functions parametrized by both function value and gradient, as explained in Appendix F. To do this, we first prove Lemma 19. Then we prove Lemma 22 (stated in Appendix F).

Lemma 19 *Let $\mathcal{M}$ be a Hadamard manifold with sectional curvatures bounded below by $K_{\mathrm{lo}} < 0$. Let $R_{\mathrm{ball}} > 0, w > 0, x_k \in \mathcal{M}$. Define $a\colon [0, \infty) \to \mathbb{R}$ as in equation (16). There is a family of bump functions*

$$\{\hat{h}_f \colon \mathcal{M} \to R\}_{f \in [-w^{-1}a(R_{\mathrm{ball}}), w^{-1}a(R_{\mathrm{ball}})]},$$

satisfying for each $f \in [-w^{-1}a(R_{\mathrm{ball}}), w^{-1}a(R_{\mathrm{ball}})]$:

- $\hat{h}_f(x_k) = f$*;*
- $\mathrm{grad}\hat{h}_f(x_k) = 0$*;*
- *the support of each $\hat{h}_f$ is contained in $B(x_k, R_{\mathrm{ball}})$;*
- $\left|\hat{h}_f(x)\right| \leq w^{-1}a(R_{\mathrm{ball}})$, $\left\|\mathrm{grad}\hat{h}_f(x)\right\| \leq \frac{w^{-1}}{4\sqrt{-K_{\mathrm{lo}}}}$, $\left\|\mathrm{Hess}\hat{h}_f(x)\right\| \leq \frac{1}{4w}$ *for all $x \in \mathcal{M}$.*

Proof We use the notation established in Section E.1. For $f \in [-w^{-1}a(R_{\mathrm{ball}}), w^{-1}a(R_{\mathrm{ball}})]$, define the smooth functions $\hat{h}_f \colon \mathcal{M} \to \mathbb{R}$ as follows:

$$\hat{h}_{cw^{-1}a(R_{\mathrm{ball}})}(x) = cw^{-1}a(R_{\mathrm{ball}})\phi_{R_{\mathrm{ball}}}(\mathrm{dist}(x, x_k)^2/2), \qquad \forall c \in [-1, 1], \forall x \in \mathcal{M}.$$

For each $f \in [-w^{-1}a(R_{\mathrm{ball}}), w^{-1}a(R_{\mathrm{ball}})]$, we know:

- $\hat{h}_f$ is supported in $B(x_k, R_{\mathrm{ball}})$;
- $\hat{h}_f(x_k) = cw^{-1}a(R_{\mathrm{ball}}) = f$;
- $\mathrm{grad}\hat{h}_f(x_k) = 0$, since $x \mapsto \mathrm{dist}(x, x_k)^2/2$ has zero gradient at $x = x_k$;
- using the calculations from Section E.1.2 and $|c| \leq 1$,

$$\begin{aligned}\left\|\mathrm{grad}\hat{h}_f(x)\right\| &= |c|\, w^{-1}a(R_{\mathrm{ball}}) \left|\phi'_{R_{\mathrm{ball}}}(\mathrm{dist}(x, x_k)^2/2)\right| \cdot \left\|\exp_x^{-1}(x_k)\right\| \\ &\leq |c|\, w^{-1}a(R_{\mathrm{ball}}) \max_{t \in [0, R_{\mathrm{ball}}]} \left|\phi'_{R_{\mathrm{ball}}}(t^2/2)\right| \cdot t \leq w^{-1}\frac{1}{4\sqrt{-K_{\mathrm{lo}}}};\end{aligned}$$

- using the calculations from Section E.1.3 and $|c| \leq 1$,

$$\left\|\mathrm{Hess}\hat{h}_f(x)\right\| \leq |c|\, w^{-1}a(R_{\mathrm{ball}})(4\sqrt{-K_{\mathrm{lo}}}/R_{\mathrm{ball}} + 55/R_{\mathrm{ball}}^2) \leq \frac{1}{4w}.$$

■

Proof [Proof of Lemma 22] For all

$$\hat{f} \in [-w^{-1}a(R_{\mathrm{ball}}), w^{-1}a(R_{\mathrm{ball}})], \qquad g \in B_{x_k}(0, w^{-1}g_{\mathrm{norm}}(R_{\mathrm{ball}})),$$

Lemmas 17 (property **BF4**) and 19 imply $(\hat{h}_{\hat{f}} + h_g)(x_k) = \hat{f} + \frac{3}{8}\|g\|\, g_{\mathrm{norm}}^{-1}\Big(w\,\|g\|\Big)$ and $\mathrm{grad}(\hat{h}_{\hat{f}} + h_g)(x_k) = 0 + g = g$.

We know that $\|g\| \in [0, w^{-1} g_{\mathrm{norm}}(R_{\mathrm{ball}})]$. Additionally, $g_{\mathrm{norm}}(0) = 0$ and g_{norm} is strictly increasing. Therefore (introducing the change of variables $g_{\mathrm{norm}}(t) = w\,\|g\|$)

$$\min_{t\in[0,R_{\mathrm{ball}}]} \frac{3}{8} w^{-1} g_{\mathrm{norm}}(t)\cdot t \leq \frac{3}{8}\|g\|\, g_{\mathrm{norm}}^{-1}\Big(w\,\|g\|\Big) \leq \max_{t\in[0,R_{\mathrm{ball}}]} \frac{3}{8} w^{-1} g_{\mathrm{norm}}(t)\cdot t.$$

Using that $t \mapsto g_{\mathrm{norm}}(t)$ is increasing,

$$0 \leq \frac{3}{8}\|g\|\, g_{\mathrm{norm}}^{-1}\Big(w\,\|g\|\Big) \leq \frac{3}{8} w^{-1} R_{\mathrm{ball}} g_{\mathrm{norm}}(R_{\mathrm{ball}}).$$

Therefore, for any $f \in [-w^{-1}a(R_{\mathrm{ball}}) + \frac{3}{8}w^{-1}R_{\mathrm{ball}}g_{\mathrm{norm}}(R_{\mathrm{ball}}), w^{-1}a(R_{\mathrm{ball}})]$ and for any $g \in B_{x_k}(0, w^{-1}g_{\mathrm{norm}}(R_{\mathrm{ball}}))$, we can define $h_{f,g} = \hat{h}_{f-\frac{3}{8}\|g\|g_{\mathrm{norm}}^{-1}(w\|g\|)} + h_g$. By construction, we have $h_{f,g}(x_k) = f$ and $\mathrm{grad} h_{f,g}(x_k) = g$. Moreover, Lemmas 17 and 19 imply

- the support of each $h_{f,g}$ is contained in $B(x_k, R_{\mathrm{ball}})$;
- $\|\mathrm{grad} h_{f,g}(x)\| \leq \frac{1}{4w\sqrt{-K_{\mathrm{lo}}}} + \frac{1}{4w\sqrt{-K_{\mathrm{lo}}}} \leq \frac{1}{2w\sqrt{-K_{\mathrm{lo}}}}$ for all $x \in \mathcal{M}$;
- $\|\mathrm{Hess} h_{f,g}(x)\| \leq \frac{1}{4w} + \frac{1}{4w} \leq \frac{1}{2w}$ for all $x \in \mathcal{M}$.

■

F Incorporating function-value queries

We want to extend the lower bound in Theorem 11 to algorithms which can make function-value and unbounded queries. We do this in two steps. First in Appendix F (this section), we prove Theorem 20 below, an extension of Theorem 11 providing a lower bound for algorithms using function-values but making bounded queries. Second in Appendix G, we prove Theorem 24, an extension of Theorem 20 providing a lower bound for algorithms making function-value and unbounded queries.

Theorem 20 *Let $\mathcal{M}$ be a Hadamard manifold of dimension $d \geq 2$ which satisfies the ball-packing property A1 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 $r \geq \max\left\{\tilde{r}, \frac{8}{\sqrt{-K_{\mathrm{lo}}}}, \frac{4(d+2)}{\tilde{c}}\right\}$. Define $\kappa = 4r\sqrt{-K_{\mathrm{lo}}}+3$. Let $\mathcal{A}$ be any deterministic algorithm, and assume that $\mathcal{A}$ always queries in $B(x_{\mathrm{ref}}, \mathscr{R})$, with $\mathscr{R} \geq r$.*

There is a function $f \in \mathcal{F}_{\kappa,r}^{x_{\mathrm{ref}}}(\mathcal{M})$ with minimizer x^* such that running $\mathcal{A}$ on f yields iterates $x_0, x_1, x_2, \ldots$ satisfying $\mathrm{dist}(x_k, x^*) \geq \frac{r}{4}$ for all $k = 0, 1, \ldots, T-1$, where

$$T = \left\lfloor \frac{\tilde{c}(d+2)^{-1} r}{\log\left(2000\tilde{c}(d+2)^{-1} r(3\mathscr{R}\sqrt{-K_{\mathrm{lo}}}+2)\right)} \right\rfloor. \tag{19}$$

Moreover, f is of the form $f(x) = \frac{1}{2}\mathrm{dist}(x, x^*)^2 + H(x)$ where $x^* \in B(x_{\mathrm{ref}}, \frac{3}{4}r)$ and $H\colon \mathcal{M} \to \mathbb{R}$ is a C^∞ function satisfying

$$|H(x)| \leq \frac{r}{64\sqrt{-K_{\mathrm{lo}}}}, \qquad \|\mathrm{grad} H(x)\| \leq \frac{1}{2\sqrt{-K_{\mathrm{lo}}}}, \qquad \|\mathrm{Hess} H(x)\| \leq \frac{1}{2}, \quad \forall x \in \mathcal{M}. \tag{20}$$

Inequalities (20) are included because they are useful for the proof for Theorem 24 (see Section G.2).

Before continuing to the details, let us first sketch the main ideas needed for this proof. In the case where the oracle only returns a gradient, Lemma 17 gives us a family of bump functions indexed by vectors g such that for each g in a ball of $\mathrm{T}_{x_k}\mathcal{M}$ there is a bump function $h_g \colon \mathcal{M} \to \mathbb{R}$ satisfying $\mathrm{grad} h_g(x_k) = g$ (and a number of other properties). This allowed us to use Lemma 13 to choose the vector g_k. In the case where the oracle also returns function values, we need a lemma which gives us a family of bump functions indexed by pairs (f, g) lying in a *cylinder* $I \times B$, where I is a closed interval of the real line and B is a closed ball in $\mathrm{T}_{x_k}\mathcal{M}$. For each pair (f, g), the lemma should provide a bump function $h_{f,g} \colon \mathcal{M} \to \mathbb{R}$ satisfying $h_{f,g}(x_k) = f$ and $\mathrm{grad} h_{f,g}(x_k) = g$. Lemma 22 in Appendix F does exactly this.

Lemma 22 works by constructing a bump function $h_{f,g}$ as a sum of two bump functions $\hat{h}_f$ and h_g (the latter from Lemma 17), the first controlling the function value of $h_{f,g}$, the second controlling its gradient. We then use Lemma 23, which is analogous to Lemma 13, to choose a pair (f_k, g_k) to return to the algorithm. Since we are now comparing the volumes of sets of the form $I \times B$ which live in a space of dimension larger than d, we end up showing that $|A_{k+1}| \geq \Omega(|A_k| / (\mathscr{R}\sqrt{-K_{\mathrm{lo}}})^{d+2})$ instead of $|A_{k+1}| \geq \Omega(|A_k| / (\mathscr{R}\sqrt{-K_{\mathrm{lo}}})^{d})$. This is not an issue because $d + 2 = \Theta(d)$.

F.1 Proof of Theorem 20

Lemma 21, which is analogous to Lemma 12, forms the backbone of the proof of Theorem 20.

Lemma 21 *Let $\mathcal{M}$ be a Hadamard manifold of dimension $d \geq 2$ with sectional curvatures in the interval $[K_{\mathrm{lo}}, 0]$ with $K_{\mathrm{lo}} < 0$. Let $x_{\mathrm{ref}} \in \mathcal{M}$, $r \geq 8/\sqrt{-K_{\mathrm{lo}}}$, $\mathscr{R} \geq r$. Let $z_1, \ldots, z_N \in B(x_{\mathrm{ref}}, \frac{3}{4}r)$ be distinct points in $\mathcal{M}$ such that $\mathrm{dist}(z_i, z_j) \geq \frac{r}{2}$ for all $i \neq j$. Define $A_0 = \{1, 2, \ldots, N\}$. Let $\mathcal{A}$ be any first-order algorithm which only queries points in $B(x_{\mathrm{ref}}, \mathscr{R})$. Finally, let $w \geq 1$ (this is a tuning parameter we will set later).*

For every nonnegative integer $k = 0, 1, 2, \ldots, \lfloor w \rfloor$, the algorithm $\mathcal{A}$ makes the query $x_k = \mathcal{A}_k((f_0, (x_0, g_0)), \ldots, (f_{k-1}, (x_{k-1}, g_{k-1})))$ and there exists $f_k \in \mathbb{R}$, $g_k \in \mathrm{T}_{x_k}\mathcal{M}$ and a set $A_{k+1} \subseteq \{1, \ldots, N\}$ satisfying

$$|A_{k+1}| \geq \frac{12000w(|A_k| - 1)}{\left(2000w(3\mathscr{R}\sqrt{-K_{\mathrm{lo}}} + 2)\right)^{d+2}} \tag{21}$$

such that for each $j \in A_{k+1}$ there is a C^∞ function $f_{j,k+1} \colon \mathcal{M} \to \mathbb{R}$ of the form

$$f_{j,k+1}(x) = \frac{1}{2}\mathrm{dist}(x, z_j)^2 + H_{j,k+1}(x)$$

satisfying:

Lfv1 *$f_{j,k+1}$ is $(1 - \frac{k+1}{2w})$-strongly g-convex in $\mathcal{M}$ and $[2r\sqrt{-K_{\mathrm{lo}}} + 1 + \frac{k+1}{2w}]$-smooth in $B(x_{\mathrm{ref}}, r)$;*

Lfv2 *$\mathrm{grad} f_{j,k+1}(z_j) = 0$ (hence in particular the minimizer of $f_{j,k+1}$ is z_j);*

Lfv3 *$f_{j,k+1}(x_m) = f_m$ and $\mathrm{grad} f_{j,k+1}(x_m) = g_m$ for $m = 0, 1, \ldots, k$;*

Lfv4 *$\mathrm{dist}(x_m, z_j) \geq \frac{r}{4}$ for all $m = 0, 1, \ldots, k$.*

Lfv5 $|H_{j,k+1}(x)| \leq \frac{(k+1)r}{64w\sqrt{-K_{\mathrm{lo}}}}$, $\|\mathrm{grad} H_{j,k+1}(x)\| \leq \frac{(k+1)}{2w\sqrt{-K_{\mathrm{lo}}}}$, $\|\mathrm{Hess} H_{j,k+1}(x)\| \leq \frac{k+1}{2w}$ *for all* $x \in \mathcal{M}$.

Proof [Proof of Theorem 20] Let us apply Lemma 21 to the manifold $\mathcal{M}$ and algorithm $\mathcal{A}$. Let the points $z_1, \dots, z_N$ be provided by the ball-packing property so that $N \geq e^{\tilde{c}r}$.

Set $w = \tilde{c}r(d+2)^{-1}$ in Lemma 21, and observe that

$$\min\{\lfloor w \rfloor, T\} = \min\left\{\lfloor \tilde{c}r(d+2)^{-1} \rfloor, \left\lfloor \frac{\tilde{c}r(d+2)^{-1}}{\log\left(2000w(3\mathscr{R}\sqrt{-K_{\mathrm{lo}}}+2)\right)} \right\rfloor \right\} = T$$

because $w = \tilde{c}r(d+2)^{-1} \geq 4$ and $\mathscr{R}\sqrt{-K_{\mathrm{lo}}} \geq 8$.

For the same reasons as in the proof of Theorem 11, it suffices to show that $|A_k| \geq 2$ for all $k \leq T$. We induct on k. (**Base case**) By the ball-packing property, $|A_0| \geq e^{\tilde{c}r} \geq 2$ since $r \geq 4(d+2)/\tilde{c}$. (**Inductive hypothesis**) Assume $|A_m| \geq 2$ for all $m \leq k$ and $k+1 \leq T$. Therefore, $|A_m| - 1 \geq |A_m|/2$ for all $m \leq k$.

Lemma 21 implies

$$|A_{m+1}| \geq \frac{6000w\,|A_m|}{\left(2000w(3R\sqrt{-K_{\mathrm{lo}}}+2)\right)^{d+2}} \geq \frac{2\,|A_m|}{\left(2000w(3\mathscr{R}\sqrt{-K_{\mathrm{lo}}}+2)\right)^{d+2}} \quad \forall m \leq k.$$

Unrolling these inequalities and using $|A_0| \geq e^{\tilde{c}r}$, we get

$$|A_{k+1}| \geq \frac{e^{\tilde{c}r}2^{k+1}}{\left(2000w(3\mathscr{R}\sqrt{-K_{\mathrm{lo}}}+2)\right)^{(k+1)(d+2)}}.$$

On the other hand, $k+1 \leq T$ implies

$$\frac{e^{\tilde{c}r}}{\left(2000w(3\mathscr{R}\sqrt{-K_{\mathrm{lo}}}+2)\right)^{(k+1)(d+2)}} \geq 1.$$

So $|A_{k+1}| \geq 2$. Lastly, note that Lemma 21 and our choice of T implies for all $x \in \mathcal{M}$

$$|H_{j,T}(x)| \leq \frac{r}{64\sqrt{-K_{\mathrm{lo}}}}, \qquad \|\mathrm{grad} H_{j,T}(x)\| \leq \frac{1}{2\sqrt{-K_{\mathrm{lo}}}}, \qquad \|\mathrm{Hess} H_{j,T}(x)\| \leq \frac{1}{2}.$$

■

F.2 Proof of Lemma 21

The proof approach for Lemma 21 is very similar to the proof presented in Section 3.2, so we are more succinct, focusing on the additional analysis needed to handle function-value queries. Before we prove this lemma, we state two lemmas which we use. The following lemma is analogous to Lemma 17, and its proof can be found in Appendix E.2.

Lemma 22 *Let* $\mathcal{M}$ *be a Hadamard manifold with sectional curvatures in the interval* $[K_{\mathrm{lo}}, 0]$ *with* $K_{\mathrm{lo}} < 0$. *Let* $R_{\mathrm{ball}} > 0, w > 0, x_k \in \mathcal{M}$. *Define* $a\colon [0,\infty) \to \mathbb{R}$ *and* $g_{\mathrm{norm}}\colon [0,\infty) \to \mathbb{R}$ *as in Lemma E. There is a family of bump functions* $\{h_{f,g}\colon \mathcal{M} \to R\}$ *indexed by* (f,g) *satisfying*

$$f \in \left[-w^{-1}a(R_{\mathrm{ball}}) + \frac{3}{8}w^{-1}R_{\mathrm{ball}}g_{\mathrm{norm}}(R_{\mathrm{ball}}), w^{-1}a(R_{\mathrm{ball}})\right], \quad g \in B_{x_k}(0, w^{-1}g_{\mathrm{norm}}(R_{\mathrm{ball}}))$$

such that for each such (f,g):

- $h_{f,g}(x_k) = f$;
- $\mathrm{grad} h_{f,g}(x_k) = g$;
- *the support of each* $h_{f,g}$ *is contained in* $B(x_k, R_{\mathrm{ball}})$;
- $|h_{f,g}(x)| \leq 2w^{-1}a(R_{\mathrm{ball}})$, $\|\mathrm{grad} h_{f,g}(x)\| \leq \frac{1}{2w\sqrt{-K_{\mathrm{lo}}}}$, $\|\mathrm{Hess} h_{f,g}(x)\| \leq \frac{1}{2w}$ *for all* $x \in \mathcal{M}$.

The interval $[-w^{-1}a(R_{\mathrm{ball}}) + \frac{3}{8}w^{-1}R_{\mathrm{ball}}g_{\mathrm{norm}}(R_{\mathrm{ball}}), w^{-1}a(R_{\mathrm{ball}})]$ *has length at least* $w^{-1}a(R_{\mathrm{ball}})$.

By *cylinder* we mean any subset C of a Euclidean space of the form $C = I \times B$, where I is a closed interval of $\mathbb{R}$ and B is a closed Euclidean ball. Note that these cylinders include their interior—there is a distinction between a cylinder and the surface of a cylinder. The height of a cylinder $C = I \times B$ is the length of I; the radius of C is the radius of the ball B.

Lemma 23 *In the following* I, I_j *denote closed intervals of* $\mathbb{R}$*, and* B, B_j *denote closed* d*-dimensional balls of Euclidean space.*

Let $C_1 = I_1 \times B_1, \ldots, C_n = I_n \times B_n$ *be* n *cylinders of radius* q *and height* a *each. Assume each of the cylinders is also contained in a larger cylinder* $C = I \times B$ *of radius* r *and height* b*:* $C_j \subseteq C$ *for all* $j = 1, \ldots, n$*. Choose*

$$(f, g) \in \arg\max_{(t,y)\in C} |\{j \in \{1, \ldots, n\} : (t, y) \in C_j\}|,$$

and let $A = \{j \in \{1, \ldots, n\} : (f, g) \in C_j\}$*. Then*

$$|A| \geq n\frac{\mathrm{Vol}(C_1)}{\mathrm{Vol}(C)} = n\frac{\mathrm{Vol}(B_1)}{\mathrm{Vol}(B)} \cdot \frac{a}{b} = n\frac{q^d}{r^d} \cdot \frac{a}{b}.$$

The proof of Lemma 23 is essentially identical to the proof of the analogous Lemma 13.

Let us now prove Lemma 21. We construct the function values $f_0, f_1, \ldots$, gradients $g_0, g_1, \ldots$, sets $A_0, A_1, \ldots$, and functions $f_{j,0}, f_{j,1}, \ldots$ inductively. We prove the claim by induction on k. The **base case** is the same as in Section 3.2. In particular, we define $f_{j,0}(x) = \frac{1}{2}\mathrm{dist}(x, z_j)^2$ for all $j \in A_0 = \{1, \ldots, N\}$.

Let's consider the **inductive step**. We are at iteration $k \geq 0$, and we assume properties **Lfv1**, **Lfv2**, **Lfv3**, **Lfv4**, **Lfv5** hold with k replacing $k+1$ in all expressions (the inductive hypothesis). The algorithm queries a point x_k. If $k \geq 1$, let $x_\ell, \ell < k$, be a previous query point closest to x_k. We can assume $x_\ell \neq x_k$. Define $\tilde{A}_k$ as in equation (7).

We shall define $f_{j,k+1} = f_{j,k} + h_{j,k}$ where $h_{j,k}$ is an appropriately chosen bump function. We want $h_{j,k}$ to be a bump function whose support is contained in $B(x_k, R_{\mathrm{ball}}^{(k)})$ where $R_{\mathrm{ball}}^{(k)}$ is defined by equation (8). With this choice for $R_{\mathrm{ball}}^{(k)}$, we set $h_{j,k}$ to be one of the bump functions $h_{f,g}$ supplied by Lemma 22 (which one remains to be determined). With this setup, we immediately know the function $f_{j,k+1} = f_{j,k} + h_{f,g}$ satisfies properties **Lfv2** and **Lfv4**, as well as $f_{j,k+1}(x_m) = f_m$ and $\mathrm{grad} f_{j,k+1}(x_m) = g_m$ for $m = 0, 1, \ldots, k-1$, for the reasons given in Section 3.2. Additionally, using

$$2w^{-1}a(R_{\mathrm{ball}}^{(k)}) \leq 2w^{-1}a\left(\frac{r}{8}\right) \leq 2w^{-1}\frac{r/8}{4(4\sqrt{-K_{\mathrm{lo}}})} = \frac{r}{64w\sqrt{-K_{\mathrm{lo}}}},$$

we see the function $f_{j,k+1} = f_{j,k} + h_{f,g}$ satisfies properties **Lfv1** and **Lfv5**.

It remains to choose $A_{k+1} \subseteq \tilde{A}_k$, $f_k \in \mathbb{R}$ and $g_k \in \mathrm{T}_{x_k}\mathcal{M}$ so that $f_{j,k+1}(x_k) = f_k, \mathrm{grad} f_{j,k+1}(x_k) = g_k$ for all $j \in A_{k+1}$, and inequality (21) is satisfied.

Consider the cylinders in $\mathbb{R} \times \mathrm{T}_{x_k}\mathcal{M}$ defined by $C_{j,k} = I_{j,k} \times B_{j,k}$ where we define

$$I_{j,k} = \Big[f_{j,k}(x_k) - w^{-1}a(R_{\text{ball}}^{(k)}) + \frac{3}{8}w^{-1}R_{\text{ball}}^{(k)}g_{\text{norm}}(R_{\text{ball}}^{(k)}), f_{j,k}(x_k) + w^{-1}a(R_{\text{ball}}^{(k)})\Big] \tag{22}$$

and recall that $B_{j,k} = B_{x_k}(\mathrm{grad} f_{j,k}(x_k), w^{-1}g_{\text{norm}}(R_{\text{ball}}^{(k)}))$. Let

$$(f_k, g_k) \in \arg\max_{(f,g)\in\mathbb{R}\times\mathrm{T}_{x_k}\mathcal{M}} \left|\{j \in \tilde{A}_k : (f,g) \in C_{j,k}\}\right|.$$

Define $A_{k+1} = \{j \in \tilde{A}_k : (f_k, g_k) \in C_{j,k}\}$.

For each $j \in A_{k+1}$ define $g_{j,k} = g_k - \mathrm{grad} f_{j,k}(x_k)$ and $f_{j,k}^{(\text{f.v.})} = f_k - f_{j,k}(x_k)$. Then Lemma 22 implies for each $j \in A_{k+1}$ there is a bump function $h_{j,k} := h_{f_{j,k}^{(\text{f.v.})}, g_{j,k}}$ satisfying

$$h_{j,k}(x_k) = f_{j,k}^{(\text{f.v.})} = f_k - f_{j,k}(x_k) \quad \text{and} \quad \mathrm{grad} h_{j,k}(x_k) = g_{j,k} = g_k - \mathrm{grad} f_{j,k}(x_k).$$

Therefore for all $j \in A_{k+1}$, $f_{j,k+1}(x_k) = f_{j,k}(x_k) + h_{j,k}(x_k) = f_k$ and $\mathrm{grad} f_{j,k+1}(x_k) = \mathrm{grad} f_{j,k}(x_k) + \mathrm{grad} h_{j,k}(x_k) = g_k$.

It remains to verify inequality (21). To do so, we use Lemma 23. To use this lemma, we need (a) a good upper bound on the radius of a ball $B_k \subseteq \mathrm{T}_{x_k}\mathcal{M}$ containing the balls $B_{j,k}, j \in \tilde{A}_k$, and (b) a good upper bound for the length of an interval $I \subseteq \mathbb{R}$ containing the intervals $I_{j,k}, j \in \tilde{A}_k$. We've already done (a) in the proof from Section 3.2. Recall that we showed (using lines (13) and (15))

$$\frac{\mathrm{Vol}(B_{j,k})}{\mathrm{Vol}(B_k)} \geq \frac{1}{(2000w(3\mathscr{R}\sqrt{-K_{\text{lo}}}+2))^d}.$$

For (b), we upper bound the length of an interval I_k containing $I_{j,k}, j \in \tilde{A}_k$ in two cases, as in Section 3.2:

Case 1: either $k = 0$, or $k \geq 1$ and $\sqrt{-K_{\text{lo}}}\mathrm{dist}(x_k, x_\ell) > 4$.

Case 2: $k \geq 1$ and $\sqrt{-K_{\text{lo}}}\mathrm{dist}(x_k, x_\ell) \leq 4$.

In each case, we show that

$$\frac{\mathrm{Length}(I_{j,k})}{\mathrm{Length}(I_k)} \geq \frac{12000w}{(2000w(3\mathscr{R}\sqrt{-K_{\text{lo}}}+2))^2}.$$

Therefore, using Lemma 23, (f_k, g_k) is contained in

$$|A_{k+1}| \geq \left|\tilde{A}_k\right| \frac{\mathrm{Vol}(B_{j,k})}{\mathrm{Vol}(B_k)} \cdot \frac{\mathrm{Length}(I_{j,k})}{\mathrm{Length}(I_k)} \geq \frac{12000w(|A_k|-1)}{\Big(2000w(3\mathscr{R}\sqrt{-K_{\text{lo}}}+2)\Big)^{d+2}}$$

of the cylinders $C_{j,k}, j \in \tilde{A}_k$. This concludes the inductive step, proving Lemma 21.

F.2.1 CASE 1 (FOR FUNCTION-VALUE QUERIES)

We have

$$\begin{aligned}&\left|\mathrm{dist}(x_k, z_j)^2 - \mathrm{dist}(x_k, x_{\mathrm{ref}})^2\right| \\ &= (\mathrm{dist}(x_k, z_j) + \mathrm{dist}(x_k, x_{\mathrm{ref}}))\left|\mathrm{dist}(x_k, z_j) - \mathrm{dist}(x_k, x_{\mathrm{ref}})\right| \\ &\leq (\mathrm{dist}(x_k, z_j) + \mathrm{dist}(x_k, x_{\mathrm{ref}}))\mathrm{dist}(z_j, x_{\mathrm{ref}}) \\ &\leq (2\mathrm{dist}(x_k, x_{\mathrm{ref}}) + \mathrm{dist}(z_j, x_{\mathrm{ref}}))\mathrm{dist}(z_j, x_{\mathrm{ref}}) \leq (2\mathrm{dist}(x_k, x_{\mathrm{ref}}) + r)r.\end{aligned} \tag{23}$$

By the inductive hypothesis and $k \leq w$, $|H_{j,k}(x_k)| \leq \frac{kw^{-1}r}{64\sqrt{-K_{\mathrm{lo}}}} \leq \frac{r}{64\sqrt{-K_{\mathrm{lo}}}}$. Combining this with inequality (23), we conclude

$$\begin{aligned}&\left|f_{j,k}(x_k) - \frac{1}{2}\mathrm{dist}(x_k, x_{\mathrm{ref}})^2\right| = \left|H_{j,k}(x_k) + \frac{1}{2}\mathrm{dist}(x_k, z_j)^2 - \frac{1}{2}\mathrm{dist}(x_k, x_{\mathrm{ref}})^2\right| \\ &\leq \frac{r}{64\sqrt{-K_{\mathrm{lo}}}} + \frac{1}{2}(2\mathrm{dist}(x_k, x_{\mathrm{ref}}) + r)r \leq (\mathrm{dist}(x_k, x_{\mathrm{ref}}) + r)r \leq (\mathscr{R} + r)r\end{aligned} \tag{24}$$

using $r\sqrt{-K_{\mathrm{lo}}} \geq 8$ for the penultimate inequality. Therefore, all function values $f_{j,k}(x_k), j \in \tilde{A}_k$, are contained in an interval centered at $\frac{1}{2}\mathrm{dist}(x_k, x_{\mathrm{ref}})^2$ of length at most $2(\mathscr{R} + r)r$. This implies that all the intervals $I_{j,k}, j \in \tilde{A}_k$, are contained in an interval I_k centered at $\frac{1}{2}\mathrm{dist}(x_k, x_{\mathrm{ref}})^2$ of length at most

$$\begin{aligned}2(\mathscr{R} + r)r + 2w^{-1}a(R_{\mathrm{ball}}^{(k)}) &\leq 2(\mathscr{R} + r)r + 2w^{-1}a\left(\frac{r}{8}\right) \leq 2(\mathscr{R} + r)r + \frac{r}{64\sqrt{-K_{\mathrm{lo}}}} \\ &\leq 2(\mathscr{R} + 2r)r \leq 6\mathscr{R}r\end{aligned}$$

using that $w \geq 1$ and $r\sqrt{-K_{\mathrm{lo}}} \geq 8$.

Now assume $k = 0$ or $\sqrt{-K_{\mathrm{lo}}}\mathrm{dist}(x_k, x_\ell) > 4$, where x_ℓ is defined in equation (5). Using $r\sqrt{-K_{\mathrm{lo}}} \geq 8$ and the definition of $R_{\mathrm{ball}}^{(k)}$ (8), this assumption implies $R_{\mathrm{ball}}^{(k)} \geq 1/\sqrt{-K_{\mathrm{lo}}}$. Therefore,

$$\frac{\mathrm{Length}(I_{j,k})}{\mathrm{Length}(I_k)} \geq \frac{w^{-1}a(R_{\mathrm{ball}}^{(k)})}{6\mathscr{R}r} \geq \frac{w^{-1}a(1/\sqrt{-K_{\mathrm{lo}}})}{6\mathscr{R}r} \geq \frac{1}{2000w(\mathscr{R}\sqrt{-K_{\mathrm{lo}}})(r\sqrt{-K_{\mathrm{lo}}})}.$$

F.2.2 CASE 2 (FOR FUNCTION-VALUE QUERIES)

Assume $k \geq 1$ and $\sqrt{-K_{\mathrm{lo}}}\mathrm{dist}(x_k, x_\ell) \leq 4$. Since $r\sqrt{-K_{\mathrm{lo}}} \geq 8$, $R_{\mathrm{ball}}^{(k)} = \mathrm{dist}(x_k, x_\ell)/4$. The analysis for this case is similar to Case 2 in Section D.2.

The inductive hypothesis implies $\|\mathrm{Hess}H_{j,k}(x)\| \leq \frac{k}{2w} \leq \frac{1}{2}$ for all $x \in \mathcal{M}$, using $k \leq w$. So as in equation (14), we know

$$\|\mathrm{Hess}f_{j,k}(x)\| \leq 3\mathscr{R}\sqrt{-K_{\mathrm{lo}}} + \frac{3}{2}, \qquad \forall x \in B(z_j, \max\{\mathrm{dist}(x_k, z_j), \mathrm{dist}(x_\ell, z_j)\}).$$

Additionally, the inductive hypothesis implies $f_{j,k}(x_\ell) = f_\ell$ and $\mathrm{grad}f_{j,k}(x_\ell) = g_\ell$ for all $j \in \tilde{A}_k$. Therefore, by Lemma 16

$$\left\|f_{j,k}(x_k) - f_\ell - \langle g_\ell, \exp_{x_\ell}^{-1}(x_k)\rangle\right\| \leq \left(3\mathscr{R}\sqrt{-K_{\mathrm{lo}}} + \frac{3}{2}\right)\mathrm{dist}(x_k, x_\ell)^2, \qquad \forall j \in \tilde{A}_k.$$

We have shown that all the function-values $f_{j,k}(x_k), j \in \tilde{A}_k$, are contained in an interval centered at $f_\ell + \langle g_\ell, \exp_{x_\ell}^{-1}(x_k)\rangle$ with length at most $2(3\mathscr{R}\sqrt{-K_{\mathrm{lo}}} + \frac{3}{2})\mathrm{dist}(x_k, x_\ell)^2$.

Therefore, all the intervals $I_{j,k}, j \in \tilde{A}_k$, are contained in an interval $I_k \subseteq \mathbb{R}$ of length

$$2\Big(3\mathscr{R}\sqrt{-K_{\mathrm{lo}}} + \frac{3}{2}\Big)\mathrm{dist}(x_k, x_\ell)^2 + 2w^{-1}a(\mathrm{dist}(x_k, x_\ell)/4) \leq 2\Big(3\mathscr{R}\sqrt{-K_{\mathrm{lo}}} + 2\Big)\mathrm{dist}(x_k, x_\ell)^2$$

using $w \geq 1$ and $a(\mathrm{dist}(x_k, x_\ell)/4) = \frac{(\mathrm{dist}(x_k,x_\ell)/4)^2}{4(\sqrt{-K_{\mathrm{lo}}}\mathrm{dist}(x_k,x_\ell)+55)} \leq \frac{\mathrm{dist}(x_k,x_\ell)^2}{64(55)}$. Therefore,

$$\begin{aligned}\frac{\mathrm{Length}(I_{j,k})}{\mathrm{Length}(I_k)} &\geq \frac{w^{-1}a(\mathrm{dist}(x_k, x_\ell)/4)}{2\Big(3\mathscr{R}\sqrt{-K_{\mathrm{lo}}} + 2\Big)\mathrm{dist}(x_k, x_\ell)^2} \geq \frac{w^{-1}\mathrm{dist}(x_k, x_\ell)^2}{8000\Big(3\mathscr{R}\sqrt{-K_{\mathrm{lo}}} + 2\Big)\mathrm{dist}(x_k, x_\ell)^2} \\ &= \frac{3\mathscr{R}\sqrt{-K_{\mathrm{lo}}} + 2}{24} \cdot \frac{12000w}{\Big(2000w\Big(3\mathscr{R}\sqrt{-K_{\mathrm{lo}}} + 2\Big)\Big)^2} \geq \frac{12000w}{\Big(2000w\Big(3\mathscr{R}\sqrt{-K_{\mathrm{lo}}} + 2\Big)\Big)^2}\end{aligned}$$

using $R_{\mathrm{ball}}^{(k)} = \mathrm{dist}(x_k, x_\ell)/4$ and

$$a(\mathrm{dist}(x_k, x_\ell)/4) = \frac{(\mathrm{dist}(x_k, x_\ell)/4)^2}{4(\sqrt{-K_{\mathrm{lo}}}\mathrm{dist}(x_k, x_\ell) + 55)} \geq \frac{(\mathrm{dist}(x_k, x_\ell)/4)^2}{4(4+55)} \geq \frac{\mathrm{dist}(x_k, x_\ell)^2}{4000},$$

which itself follows from $\mathrm{dist}(x_k, x_\ell) \leq 4/\sqrt{-K_{\mathrm{lo}}}$.

G Unbounded queries

We now want to extend the lower bound from Theorem 20, which holds for algorithms querying only in $B(x_{\mathrm{ref}}, \mathscr{R})$, to algorithms which can query anywhere. That is, we want to prove Theorem 24.

Theorem 24 *Let $\mathcal{M}$ be a Hadamard manifold of dimension $d \geq 2$ which satisfies the ball-packing property A1 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 $r \geq \max\{\tilde{r}, \frac{8}{\sqrt{-K_{\mathrm{lo}}}}, \frac{4(d+2)}{\tilde{c}}\}$. Define $\kappa = 4r\sqrt{-K_{\mathrm{lo}}} + 3$. Let $\mathcal{A}$ be any deterministic algorithm.*

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

$$T = \left\lfloor \frac{\tilde{c}(d+2)^{-1}r}{\log\big(2000\tilde{c}(d+2)^{-1}r(3\mathscr{R}\sqrt{-K_{\mathrm{lo}}} + 2)\big)} \right\rfloor \geq \left\lfloor \frac{\tilde{c}(d+2)^{-1}r}{\log\big(2 \cdot 10^6 \cdot \tilde{c}(d+2)^{-1}r(r\sqrt{-K_{\mathrm{lo}}})^2\big)} \right\rfloor$$

with $\mathscr{R} = 2^9 r \log(r\sqrt{-K_{\mathrm{lo}}})^2$.

To prove Theorem 24, the high-level idea is to modify all hard instances f from Theorem 20 so that $f(x) = \frac{1}{2}\mathrm{dist}(x, x_{\mathrm{ref}})^2$ for $x \notin B(x_{\mathrm{ref}}, \mathscr{R})$ (recall $\mathscr{R} \geq r$). This way, the algorithm gains no information by querying outside the ball $B(x_{\mathrm{ref}}, \mathscr{R})$. On the other hand, we still want the hard functions f to remain untouched in the ball $B(x_{\mathrm{ref}}, r)$. In the region between radii r and $\mathscr{R}$, we smoothly interpolate between these two choices of functions. We show that we can choose $\mathscr{R}$ appropriately so that the lower bound $\tilde{\Omega}(r)$ still holds and the modified functions are still strongly g-convex. Technically, we do this via a reduction, which is depicted in Figure 1. This argument was inspired by (Carmon et al., 2019, Sec. 5.2).

G.1 Proof of Theorem 24: a reduction from Theorem 20

Define $\mathscr{D}\colon \mathcal{M} \to \mathbb{R}$ by $\mathscr{D}(x) = \frac{1}{2}\mathrm{dist}(x, x_{\mathrm{ref}})^2$. Given any $f\colon \mathcal{M} \to \mathbb{R}$ (think from Theorem 20), define the function $f_{r,\mathscr{R}}\colon \mathcal{M} \to \mathbb{R}$ by

$$f_{r,\mathscr{R}}(x) = s_{r,\mathscr{R}}\big(\mathscr{D}(x)\big) f(x) + \Big[1 - s_{r,\mathscr{R}}\big(\mathscr{D}(x)\big)\Big] \mathscr{D}(x) \tag{25}$$

where $s_{r,\mathscr{R}}\colon \mathbb{R} \to \mathbb{R}$ is a C^∞ function which is 1 on $(-\infty, \frac{1}{2}r^2]$ and 0 on $[\frac{1}{2}\mathscr{R}^2, \infty)$. More precisely, following Lee (2012, Lem. 2.20, 2.21) we define the C^∞ function $t\colon \mathbb{R} \to \mathbb{R}$ by

$$t(\tau) = \begin{cases} 1 & \text{for } \tau \in (-\infty, 0]; \\ \frac{e^{-\frac{1}{1-\tau}}}{e^{-\frac{1}{1-\tau}} + e^{-\frac{1}{\tau}}} & \text{for } \tau \in (0, 1); \\ 0 & \text{for } \tau \in [1, \infty) \end{cases}$$

and define $s_{r,\mathscr{R}}\colon \mathbb{R} \to \mathbb{R}$ by

$$s_{r,\mathscr{R}}(\mathscr{D}) = t\left(\frac{\mathscr{D} - \frac{1}{2}r^2}{\frac{1}{2}\mathscr{R}^2 - \frac{1}{2}r^2}\right), \qquad \text{for all } \mathscr{D} \in \mathbb{R}.$$

In Appendix G.2, we show that if we set $\mathscr{R} = 2^9 r \log(r\sqrt{-K_{\mathrm{lo}}})^2$ and if $f \in \mathcal{F}_{\kappa,r}^{x_{\mathrm{ref}}}(\mathcal{M})$ is from Theorem 20, then $f_{r,\mathscr{R}} \in \mathcal{F}_{3\kappa,r}^{x_{\mathrm{ref}}}(\mathcal{M})$.

Definition 25 *The (first-order) oracle for a differentiable function $f\colon \mathcal{M} \to \mathbb{R}$ is the map $\mathcal{O}_f\colon \mathcal{M} \to \mathbb{R} \times \mathrm{T}\mathcal{M}$ given by $\mathcal{O}_f(x) = (f(x), (x, \mathrm{grad} f(x)))$.*

Given the oracle $\mathcal{O}_f$ of any function f, we can use $\mathcal{O}_f$ to emulate the oracle $\mathcal{O}_{f_{r,\mathscr{R}}}$ using equation (25), and the following formula for $\mathrm{grad} f_{r,\mathscr{R}}$:

$$\mathrm{grad} f_{r,\mathscr{R}}(x) = \begin{cases} -\exp_x^{-1}(x_{\mathrm{ref}}) & \text{if } \mathrm{d}(x, x_{\mathrm{ref}}) > \mathscr{R}; \\ \mathrm{grad} f(x) & \text{if } \mathrm{d}(x, x_{\mathrm{ref}}) \leq r; \\ -s'_{r,\mathscr{R}}\big(\mathscr{D}(x)\big)\big(f(x) - \mathscr{D}(x)\big) \exp_x^{-1}(x_{\mathrm{ref}}) & \\ \quad -\big(1 - s_{r,\mathscr{R}}\big(\mathscr{D}(x)\big)\big) \exp_x^{-1}(x_{\mathrm{ref}}) & \\ \quad + s_{r,\mathscr{R}}\big(\mathscr{D}(x)\big) \mathrm{grad} f(x) & \text{otherwise.} \end{cases}$$

See Appendix G.2 for the derivation of this formula for $\mathrm{grad} f_{r,\mathscr{R}}$.

To prove a lower bound for an algorithm $\mathcal{B}$ querying anywhere, we make $\mathcal{B}$ interact with the oracle $\mathcal{O}_{f_{r,\mathscr{R}}}$ (which we simulate using $\mathcal{O}_f$). This implicitly defines an algorithm $\mathcal{A}$ which interacts with $\mathcal{O}_f$—see Figure 1. Explicitly, the algorithm $\mathcal{A}$ internally runs the algorithm $\mathcal{B}$ as a subroutine as follows:

- if $\mathcal{B}$ outputs $y_k \notin B(x_{\mathrm{ref}}, \mathscr{R})$, $\mathcal{A}$ does not query the oracle $\mathcal{O}_f$, but simply passes

$$(f_{r,\mathscr{R}}(y_k), \mathrm{grad} f_{r,\mathscr{R}}(y_k)) = \Big(\frac{1}{2}\mathrm{dist}(y_k, x_{\mathrm{ref}})^2, -\exp_{y_k}^{-1}(x_{\mathrm{ref}})\Big)$$

to $\mathcal{B}$; this corresponds to path 1'-2'-3' in Figure 1;

- if $\mathcal{B}$ outputs $y_k \in B(x_{\mathrm{ref}}, \mathscr{R})$, $\mathcal{A}$ queries $\mathcal{O}_f$ at $x_i = y_k$, receives $(f(x_i), \mathrm{grad} f(x_i))$ from $\mathcal{O}_f$, and passes $(f_{r,\mathscr{R}}(x_i), \mathrm{grad} f_{r,\mathscr{R}}(x_i))$ to $\mathcal{B}$ (which it computes using $(f(x_i), \mathrm{grad} f(x_i))$); this corresponds to path 1-2-3-4-5 in Figure 1.

Inside of $\mathcal{A}$, the algorithm $\mathcal{B}$ outputs the sequence $y_0, y_1, y_2, \ldots$. The algorithm $\mathcal{A}$ produces the sequence of queries $x_0 = y_{k_0}, x_1 = y_{k_1}, \ldots \in B(x_{\mathrm{ref}}, \mathscr{R})$ where $0 \leq k_0 < k_1 < k_2 < \ldots$ are integers. Let $\mathcal{K}_T = \{k_0, k_1, \ldots, k_{T-1}\}$.

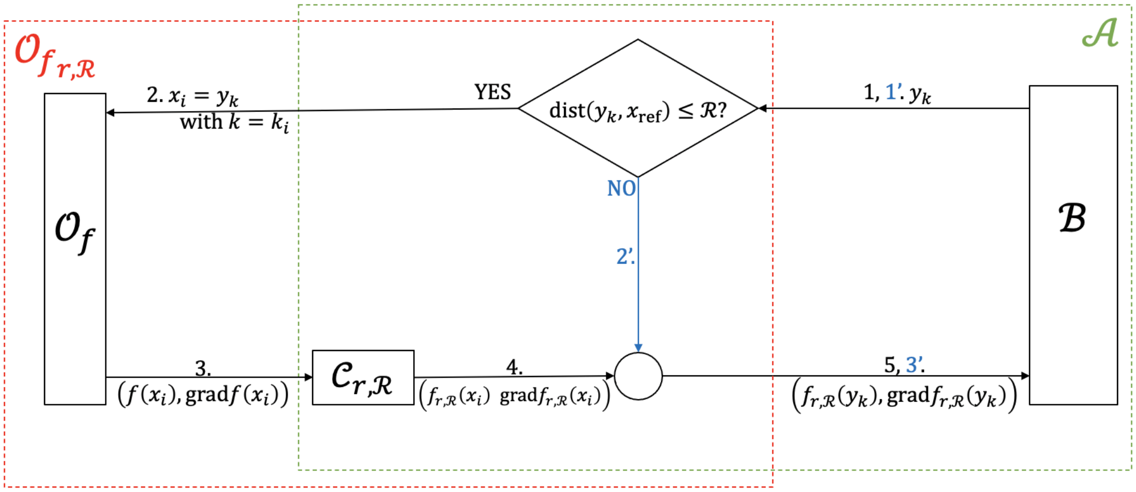
$\mathcal{O}_{f_{r,\mathcal{R}}}$
$\mathcal{A}$
2. $x_i = y_k$
with $k = k_i$
YES
$\mathrm{dist}(y_k, x_{\mathrm{ref}}) \leq \mathcal{R}$?
1, 1'. y_k
$\mathcal{O}_f$
NO
2'.
$\mathcal{B}$
3.
$(f(x_i), \mathrm{grad} f(x_i))$
$\mathcal{C}_{r,\mathcal{R}}$
4.
$(f_{r,\mathcal{R}}(x_i)\ \mathrm{grad} f_{r,\mathcal{R}}(x_i))$
5, 3'.
$(f_{r,\mathcal{R}}(y_k), \mathrm{grad} f_{r,\mathcal{R}}(y_k))$

Figure 1: A diagram of the reduction used in Section G.1. The algorithm $\mathcal{A}$ first internally runs $\mathcal{B}$ to get an iterate y_k. Then, depending on the distance between y_k and x_{ref}, $\mathcal{A}$ either queries the oracle $\mathcal{O}_f$ (path 1-2-3-4-5) or does not query the oracle (path 1'-2'-3'). The box $\mathscr{C}_{r,\mathscr{R}}$ represents a map which when given a pair $(f(x), \mathrm{grad} f(x))$, outputs $(f_{r,\mathscr{R}}(x), \mathrm{grad} f_{r,\mathscr{R}}(x))$. At step 2', $\mathcal{A}$ computes $(\frac{1}{2}\mathrm{dist}(y_k, x_{\mathrm{ref}})^2, -\exp_{y_k}(x_{\mathrm{ref}}))$ (not shown for clarity) which it then returns to $\mathcal{B}$.

By design, algorithm $\mathcal{A}$ makes queries only in $B(x_{\mathrm{ref}}, \mathscr{R})$. Therefore, we can apply Theorem 20 to $\mathcal{A}$. That theorem implies there is a function $f \in \mathcal{F}_{\kappa,r}^{x_{\mathrm{ref}}}(\mathcal{M})$ with minimizer x^* for which running $\mathcal{A}$ on f yields $x_0 = y_{k_0}, x_1 = y_{k_1}, \ldots, x_{T-1} = y_{k_{T-1}} \in B(x_{\mathrm{ref}}, \mathscr{R})$ satisfying $\mathrm{dist}(x^*, x_k) \geq \frac{r}{4}$ for all $k = 0, 1, \ldots T-1$. Here, $\kappa = 4r\sqrt{-K_{\mathrm{lo}}} + 3$, and T is given by Theorem 20 with $\mathscr{R} = 2^9 r \log(r\sqrt{-K_{\mathrm{lo}}})^2$, that is,

$$T = \left\lfloor \frac{\tilde{c}(d+2)^{-1}r}{\log\left(2000\tilde{c}(d+2)^{-1}r(3\mathscr{R}\sqrt{-K_{\mathrm{lo}}}+2)\right)} \right\rfloor \geq \left\lfloor \frac{\tilde{c}(d+2)^{-1}r}{\log\left(2 \cdot 10^6 \cdot \tilde{c}(d+2)^{-1}r(r\sqrt{-K_{\mathrm{lo}}})^2\right)} \right\rfloor,$$

using that $r\sqrt{-K_{\mathrm{lo}}} \geq 8$. In other words, $\mathrm{dist}(x^*, y_k) \geq \frac{r}{4}$ for all $k \in \mathcal{K}_T$.

On the other hand, we know that $\mathrm{dist}(x_{\mathrm{ref}}, y_k) \geq \mathscr{R}$ for all $k \in \{0, 1, \ldots, T-1\} \setminus \mathcal{K}_T$. Therefore, using that $\mathscr{R} \geq r$ and $x^* \in B(x_{\mathrm{ref}}, \frac{3}{4}r)$,

$$\mathrm{dist}(x^*, y_k) \geq \frac{r}{4} \text{ for all } k \in \{0, 1, \ldots, T-1\} \setminus \mathcal{K}_T.$$

We conclude $\mathrm{dist}(x^*, y_k) \geq \frac{r}{4}$ for all $k = 0, 1, ..., T-1$. Finally, observe that (by our construction of $\mathcal{A}$) if we run $\mathcal{B}$ on the function $f_{r,\mathscr{R}}$ then we get exactly the sequence $y_0, y_1, \ldots, y_{T-1}$. Since $f_{r,\mathscr{R}} \in \mathcal{F}_{3\kappa,r}^{x_{\mathrm{ref}}}(\mathcal{M})$ if $\mathscr{R} = 2^9 r \log(r\sqrt{-K_{\mathrm{lo}}})^2$, as stated above, this proves Theorem 24.

G.2 Verifying $f_{r,\mathscr{R}}$ is in the function class

In this section, we abbreviate $s = s_{r,\mathscr{R}}$. To finish the proof of Theorem 24 (from Section G.1), it remains to show that if $f \in \mathcal{F}_{\kappa,r}^{x_{\mathrm{ref}}}(\mathcal{M})$ is a hard function from Theorem 20, then $f_{r,\mathscr{R}} \in \mathcal{F}_{3\kappa,r}^{x_{\mathrm{ref}}}(\mathcal{M})$ for a suitable choice of $\mathscr{R}$. To do this, we use that a hard function f from Theorem 20 is $\frac{1}{2}$-strongly g-convex in $\mathcal{M}$ and $[2r\sqrt{-K_{\mathrm{lo}}} + \frac{3}{2}]$-smooth in $B(x_{\mathrm{ref}}, r)$. We also use that f has the form

$$f(x) = \frac{1}{2}\mathrm{dist}(x, x^*)^2 + H(x), \quad \text{with } x^* \in B\Big(x_{\mathrm{ref}}, \frac{3}{4}r\Big),$$

and, from inequalities (20), for all $x \in \mathcal{M}$ we have

- $\|\mathrm{grad} H(x)\| \leq \frac{1}{2\sqrt{-K_{\mathrm{lo}}}} \leq \frac{r}{16}$ (since we assume $r\sqrt{-K_{\mathrm{lo}}} \geq 8$); and
- $|H(x)| \leq \frac{r}{64\sqrt{-K_{\mathrm{lo}}}} \leq \frac{r^2}{512}$ (again since we assume $r\sqrt{-K_{\mathrm{lo}}} \geq 8$).

Recall Definition 1 for the function classes $\mathcal{F}_{\kappa,r}^{x_{\mathrm{ref}}}(\mathcal{M})$ and $\mathcal{F}_{3\kappa,r}^{x_{\mathrm{ref}}}(\mathcal{M})$. Since $f(x) = f_{r,\mathscr{R}}(x)$ for all $x \in B(x_{\mathrm{ref}}, r)$, it suffices to show that $\mathrm{Hess} f_{r,\mathscr{R}}(x) \succeq \frac{1}{6}I$ for all $x \in \mathcal{M}$. We just need to check that this is true when $r \leq \mathrm{dist}(x, x_{\mathrm{ref}}) \leq \mathscr{R}$. Let's compute $\mathrm{grad} f_{r,\mathscr{R}}(x)$ and $\mathrm{Hess} f_{r,\mathscr{R}}(x)$ when $r \leq \mathrm{dist}(x, x_{\mathrm{ref}}) \leq \mathscr{R}$.

Let $\gamma(t)$ be a geodesic with $\gamma(0) = x, \gamma'(0) = v$ and $\|v\| = 1$. For the moment, define $\mathscr{D}(x) = \frac{1}{2}\mathrm{dist}(x, x_{\mathrm{ref}})^2$, keeping in mind that $\mathscr{D} \colon \mathcal{M} \to \mathbb{R}$ depends on x_{ref}. Additionally, define $\tau(x) = \frac{\mathscr{D}(x) - \frac{1}{2}r^2}{\frac{1}{2}\mathscr{R}^2 - \frac{1}{2}r^2}$ so that

$$s(\mathscr{D}(x)) = t(\tau(x)), \quad s'(\mathscr{D}(x)) = \frac{1}{\frac{1}{2}\mathscr{R}^2 - \frac{1}{2}r^2} t'(\tau(x)), \quad s''(\mathscr{D}(x)) = \frac{1}{(\frac{1}{2}\mathscr{R}^2 - \frac{1}{2}r^2)^2} t''(\tau(x)).$$

For the gradient, we have:

$$\begin{aligned}
\langle v, \mathrm{grad} f_{r,\mathscr{R}}(x) \rangle =& \frac{d}{dt}\Big[f_{r,\mathscr{R}}(\gamma(t))\Big]_{t=0} = \frac{d}{dt}\Big[s(\mathscr{D}(\gamma(t)))f(\gamma(t)) + [1 - s(\mathscr{D}(\gamma(t)))]\mathscr{D}(\gamma(t))\Big]_{t=0} \\
=& \frac{d}{dt}\Big[s(\mathscr{D}(\gamma(t)))\Big]_{t=0}(f(x) - \mathscr{D}(x)) + s(\mathscr{D}(x))\frac{d}{dt}\Big[f(\gamma(t))\Big]_{t=0} \\
&+ [1 - s(\mathscr{D}(x))]\frac{d}{dt}\Big[\mathscr{D}(\gamma(t))\Big]_{t=0} \\
=& s'(\mathscr{D}(x))\left\langle v, -\exp_x^{-1}(x_{\mathrm{ref}})\right\rangle(f(x) - \mathscr{D}(x)) + s(\mathscr{D}(x))\left\langle v, \mathrm{grad} f(x)\right\rangle \\
&+ [1 - s(\mathscr{D}(x))]\left\langle v, -\exp_x^{-1}(x_{\mathrm{ref}})\right\rangle.
\end{aligned}$$

For the Hessian, we have:

$$\begin{aligned}
\left\langle v, \mathrm{Hess} f_{r,\mathscr{R}}(x) v\right\rangle &= \frac{d^2}{dt^2}\Big[f_{r,\mathscr{R}}(\gamma(t))\Big]_{t=0} \\
&= \frac{d^2}{dt^2}\Big[s(\mathscr{D}(\gamma(t)))f(\gamma(t)) + [1 - s(\mathscr{D}(\gamma(t)))]\mathscr{D}(\gamma(t))\Big]_{t=0} \\
&= \frac{d^2}{dt^2}\Big[s(\mathscr{D}(\gamma(t)))\Big]_{t=0} f(x) + 2\frac{d}{dt}\Big[s(\mathscr{D}(\gamma(t)))\Big]_{t=0}\frac{d}{dt}\Big[f(\gamma(t))\Big]_{t=0} \\
&\quad + s(\mathscr{D}(x))\frac{d^2}{dt^2}\Big[f(\gamma(t))\Big]_{t=0} - \frac{d^2}{dt^2}\Big[s(\mathscr{D}(\gamma(t)))\Big]_{t=0}\mathscr{D}(x) \\
&\quad - \frac{d}{dt}\Big[s(\mathscr{D}(\gamma(t)))\Big]_{t=0}\frac{d}{dt}\Big[\mathscr{D}(\gamma(t))\Big]_{t=0} + [1 - s(\mathscr{D}(x))]\frac{d^2}{dt^2}\Big[\mathscr{D}(\gamma(t))\Big]_{t=0}.
\end{aligned}$$

Further simplifying yields:

$$\begin{aligned}
\left\langle v, \mathrm{Hess} f_{r,\mathscr{R}}(x) v\right\rangle &= s(\mathscr{D}(x))\frac{d^2}{dt^2}\Big[f(\gamma(t))\Big]_{t=0} + [1 - s(\mathscr{D}(x))]\frac{d^2}{dt^2}\Big[\mathscr{D}(\gamma(t))\Big]_{t=0} \\
&\quad + \frac{d^2}{dt^2}\Big[s(\mathscr{D}(\gamma(t)))\Big]_{t=0}(f(x) - \mathscr{D}(x)) \\
&\quad + 2\frac{d}{dt}\Big[s(\mathscr{D}(\gamma(t)))\Big]_{t=0}\bigg(\frac{d}{dt}\Big[f(\gamma(t))\Big]_{t=0} - \frac{d}{dt}\Big[\mathscr{D}(\gamma(t))\Big]_{t=0}\bigg).
\end{aligned}$$

Using that $f(x) = \frac{1}{2}\mathrm{dist}(x, x^*)^2 + H(x)$,

$$\begin{aligned}
\left\langle v, \mathrm{Hess} f_{r,\mathscr{R}}(x) v\right\rangle &= s(\mathscr{D}(x))\left\langle v, \mathrm{Hess} f(x) v\right\rangle + [1 - s(\mathscr{D}(x))]\left\langle v, \mathrm{Hess}\mathscr{D}(x) v\right\rangle \\
&+ \frac{d^2}{dt^2}\Big[s(\mathscr{D}(\gamma(t)))\Big]_{t=0}\bigg(\frac{1}{2}\mathrm{dist}(x, x^*)^2 + H(x) - \mathscr{D}(x)\bigg) \\
&- 2\frac{d}{dt}\Big[s(\mathscr{D}(\gamma(t)))\Big]_{t=0}\bigg(\left\langle v, \exp_x^{-1}(x^*) - \exp_x^{-1}(x_{\mathrm{ref}})\right\rangle - \left\langle v, \mathrm{grad} H(x)\right\rangle\bigg).
\end{aligned}$$

Rearranging yields

$$\begin{aligned}
\left\langle v, \mathrm{Hess} f_{r,\mathscr{R}}(x) v\right\rangle &= t(\tau(x))\left\langle v, \mathrm{Hess} f(x) v\right\rangle + [1 - t(\tau(x))]\left\langle v, \mathrm{Hess}\mathscr{D}(x) v\right\rangle \\
&+ \frac{d^2}{dt^2}\Big[s(\mathscr{D}(\gamma(t)))\Big]_{t=0}\frac{1}{2}\bigg(\mathrm{dist}(x, x^*) - \mathrm{dist}(x, x_{\mathrm{ref}})\bigg)\bigg(\mathrm{dist}(x, x^*) + \mathrm{dist}(x, x_{\mathrm{ref}})\bigg) \\
&- 2\frac{d}{dt}\Big[s(\mathscr{D}(\gamma(t)))\Big]_{t=0}\left\langle v, \exp_x^{-1}(x^*) - \exp_x^{-1}(x_{\mathrm{ref}})\right\rangle \\
&+ \frac{d^2}{dt^2}\Big[s(\mathscr{D}(\gamma(t)))\Big]_{t=0} H(x) + 2\frac{d}{dt}\Big[s(\mathscr{D}(\gamma(t)))\Big]_{t=0}\left\langle v, \mathrm{grad} H(x)\right\rangle.
\end{aligned}$$

Using $r \le \mathrm{dist}(x, x_{\mathrm{ref}}) \le \mathscr{R}$, we have $\left|\left\langle v, \exp_x^{-1}(x_{\mathrm{ref}})\right\rangle\right| \le \mathrm{dist}(x, x_{\mathrm{ref}}) \le \mathscr{R}$. Using Proposition 14,

$$\left|\left\langle v, \exp_x^{-1}(x^*) - \exp_x^{-1}(x_{\mathrm{ref}})\right\rangle\right| \le \left\|\exp_x^{-1}(x^*) - \exp_x^{-1}(x_{\mathrm{ref}})\right\| \le \mathrm{dist}(x^*, x_{\mathrm{ref}}) \le r.$$

Using the triangle inequality,

$$\begin{aligned}
&\left|\bigg(\mathrm{dist}(x, x^*) - \mathrm{dist}(x, x_{\mathrm{ref}})\bigg)\bigg(\mathrm{dist}(x, x^*) + \mathrm{dist}(x, x_{\mathrm{ref}})\bigg)\right| \\
&\le \mathrm{dist}(x^*, x_{\mathrm{ref}})\bigg(\mathrm{dist}(x_{\mathrm{ref}}, x^*) + 2\mathrm{dist}(x, x_{\mathrm{ref}})\bigg) \le r(r + 2\mathscr{R}) \le 3r\mathscr{R}.
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
\left\langle v, \mathrm{Hess} f_{r,\mathscr{R}}(x) v\right\rangle &\geq t(\tau(x)) \left\langle v, \mathrm{Hess} f(x) v\right\rangle + [1 - t(\tau(x))] \left\langle v, \mathrm{Hess}\mathscr{D}(x) v\right\rangle \\
&\quad - \frac{3}{2} r\mathscr{R} \left| \frac{d^2}{dt^2}\Big[s(\mathscr{D}(\gamma(t)))\Big]_{t=0}\right| - 2r \left| \frac{d}{dt}\Big[s(\mathscr{D}(\gamma(t)))\Big]_{t=0}\right| \\
&\quad - \frac{r^2}{512} \left| \frac{d^2}{dt^2}\Big[s(\mathscr{D}(\gamma(t)))\Big]_{t=0}\right| - \frac{r}{8} \left| \frac{d}{dt}\Big[s(\mathscr{D}(\gamma(t)))\Big]_{t=0}\right| \\
&\geq t(\tau(x)) \left\langle v, \mathrm{Hess} f(x) v\right\rangle + [1 - t(\tau(x))] \left\langle v, \mathrm{Hess}\mathscr{D}(x) v\right\rangle \\
&\quad - \frac{8}{5} r\mathscr{R} \left| \frac{d^2}{dt^2}\Big[s(\mathscr{D}(\gamma(t)))\Big]_{t=0}\right| - 3r \left| \frac{d}{dt}\Big[s(\mathscr{D}(\gamma(t)))\Big]_{t=0}\right|.
\end{aligned}$$

Additionally, we have

$$\begin{aligned}
\left| \frac{d}{dt}\Big[s(\mathscr{D}(\gamma(t)))\Big]_{t=0}\right| &= \left| s'(\mathscr{D}(x)) \frac{d}{dt}\Big[\mathscr{D}(\gamma(t))\Big]_{t=0}\right| = \frac{1}{\frac{1}{2}\mathscr{R}^2 - \frac{1}{2}r^2} \left| t'(\tau(x)) \left\langle v, -\exp_x^{-1}(x_{\mathrm{ref}})\right\rangle\right| \\
&\leq \frac{\mathscr{R}}{\frac{1}{2}\mathscr{R}^2 - \frac{1}{2}r^2} \left| t'(\tau(x))\right|,
\end{aligned}$$

and

$$\begin{aligned}
&\left| \frac{d^2}{dt^2}\Big[s(\mathscr{D}(\gamma(t)))\Big]_{t=0}\right| = \left| s''(\mathscr{D}(x)) \left(\frac{d}{dt}\Big[\mathscr{D}(\gamma(t))\Big]_{t=0}\right)^2 + s'(\mathscr{D}(x)) \frac{d^2}{dt^2}\Big[\mathscr{D}(\gamma(t))\Big]_{t=0}\right| \\
&\leq \frac{\mathscr{R}^2}{(\frac{1}{2}\mathscr{R}^2 - \frac{1}{2}r^2)^2} \left| t''(\tau(x))\right| + \frac{1}{\frac{1}{2}\mathscr{R}^2 - \frac{1}{2}r^2} \left| t'(\tau(x))\right| \left\langle v, \mathrm{Hess}\mathscr{D}(x) v\right\rangle.
\end{aligned}$$

In the following, we set $\mathscr{R} = 2^9 r \log(r\sqrt{-K_{\mathrm{lo}}})^2)$. This choice of $\mathscr{R}$ and $r\sqrt{-K_{\mathrm{lo}}} \geq 8$ implies $\mathscr{R} = 2^9 r \log(r\sqrt{-K_{\mathrm{lo}}})^2) \geq 2^9 \log(8)^2 r \geq 2^{11} r$. Since $\mathscr{R} \geq 2^{11} r$, we conclude

$$\begin{aligned}
\left\langle v, \mathrm{Hess} f_{r,\mathscr{R}}(x) v\right\rangle &\geq t(\tau(x)) \left\langle v, \mathrm{Hess} f(x) v\right\rangle \\
&\quad + \left[1 - t(\tau(x)) - \frac{2r\mathscr{R}}{\frac{1}{2}\mathscr{R}^2 - \frac{1}{2}r^2} \left| t'(\tau(x))\right| \right] \left\langle v, \mathrm{Hess}\mathscr{D}(x) v\right\rangle \\
&\quad - \frac{\frac{8}{5} r\mathscr{R}^3}{(\frac{1}{2}\mathscr{R}^2 - \frac{1}{2}r^2)^2} \left| t''(\tau(x))\right| - \frac{3r\mathscr{R}}{\frac{1}{2}\mathscr{R}^2 - \frac{1}{2}r^2} \left| t'(\tau(x))\right| \\
&\geq t(\tau(x)) \left\langle v, \mathrm{Hess} f(x) v\right\rangle \\
&\quad + \left[1 - t(\tau(x)) - \frac{2r\mathscr{R}}{\frac{1}{2}(1 - 2^{-22})\mathscr{R}^2} \left| t'(\tau(x))\right| \right] \left\langle v, \mathrm{Hess}\mathscr{D}(x) v\right\rangle \\
&\quad - \frac{\frac{8}{5} r\mathscr{R}^3}{(\frac{1}{2}(1 - 2^{-22})\mathscr{R}^2)^2} \left| t''(\tau(x))\right| - \frac{3r\mathscr{R}}{\frac{1}{2}(1 - 2^{-22})\mathscr{R}^2} \left| t'(\tau(x))\right| \\
&= t(\tau(x)) \left\langle v, \mathrm{Hess} f(x) v\right\rangle \\
&\quad + \left[1 - t(\tau(x)) - \frac{4.5r}{\mathscr{R}} \left| t'(\tau(x))\right| \right] \left\langle v, \mathrm{Hess}\mathscr{D}(x) v\right\rangle \\
&\quad - \frac{\frac{32}{5} r}{(1 - 2^{-22})^2\mathscr{R}} \left| t''(\tau(x))\right| - \frac{7r}{\mathscr{R}} \left| t'(\tau(x))\right|.
\end{aligned}$$

One can check that $-2 \le t'(\tau) \le 0$ and $|t(\tau)''(0)| \le 16$ for all $\tau \in (0,1)$. So using $\mathscr{R} \ge 2^{11}r$,

$$\left\langle v, \mathrm{Hess} f_{r,\mathscr{R}}(x)v \right\rangle \ge t(\tau(x))\left\langle v, \mathrm{Hess} f(x)v \right\rangle + \left[1 - t(\tau(x)) - \frac{4.5r}{\mathscr{R}}\left|t'(\tau(x))\right|\right]\left\langle v, \mathrm{Hess}\mathscr{D}(x)v \right\rangle - \frac{1}{16}.$$

Next we make two observations about the univariate function t:

- if $\tau \in [\frac{1}{2}, 1)$, then $1 - t(\tau) - \frac{4.5}{2^{11}}\left|t'(\tau)\right| \ge \frac{1}{2} - \frac{4.5}{2^{10}}$;
- if $\tau \in (0, \frac{1}{2})$ and $\mathscr{R} \ge 9 \cdot 4.5r$, then $1 - t(\tau) - \frac{4.5r}{\mathscr{R}}\left|t'(\tau)\right| \ge -2e^{-\sqrt{\frac{\mathscr{R}}{9r}}}$. We prove this fact in the next Section G.3.

Using these facts, if $\tau(x) \in [\frac{1}{2}, 1)$ then (using $\langle v, \mathrm{Hess}\mathscr{D}(x)v\rangle \ge 1$ and $\mathscr{R} \ge 2^{11}r$)

$$\left\langle v, \mathrm{Hess} f_{r,\mathscr{R}}(x)v \right\rangle \ge \frac{1}{2} - \frac{4.5}{2^{10}} - \frac{1}{16} \ge \frac{1}{6}.$$

If $\tau(x) \in (0, \frac{1}{2})$, then (using $\langle v, \mathrm{Hess} f(x)v\rangle \ge \frac{1}{2}$ and $\langle v, \mathrm{Hess}\mathscr{D}(x)v\rangle \le 2\mathscr{R}\sqrt{-K_{\mathrm{lo}}}$)

$$\left\langle v, \mathrm{Hess} f_{r,\mathscr{R}}(x)v \right\rangle \ge \frac{1}{2} \cdot \frac{1}{2} - 2e^{-\sqrt{\frac{\mathscr{R}}{9r}}} \cdot 2\mathscr{R}\sqrt{-K_{\mathrm{lo}}} - \frac{1}{16} \ge \frac{1}{2} \cdot \frac{1}{2} - \frac{1}{2^6} - \frac{1}{16} \ge \frac{1}{6},$$

where the last inequality follows from choosing $\mathscr{R} = 2^9 r \log(r\sqrt{-K_{\mathrm{lo}}})^2$ and $r\sqrt{-K_{\mathrm{lo}}} \ge 8$:

$$\begin{aligned} 2e^{-\sqrt{\frac{\mathscr{R}}{9r}}} \cdot 2\mathscr{R}\sqrt{-K_{\mathrm{lo}}} &= 4e^{-\sqrt{\frac{2^9}{9}}\log(r\sqrt{-K_{\mathrm{lo}}})} \cdot 2^9 r\sqrt{-K_{\mathrm{lo}}}\log(r\sqrt{-K_{\mathrm{lo}}})^2 \\ &= 2^{11}(r\sqrt{-K_{\mathrm{lo}}})^{-\sqrt{\frac{2^9}{9}}} r\sqrt{-K_{\mathrm{lo}}}\log(r\sqrt{-K_{\mathrm{lo}}})^2 \le \frac{1}{2^6}. \end{aligned}$$

G.3 Technical fact about the function $t\colon \mathbb{R} \to \mathbb{R}$ in the interval $(0,1)$

Lemma 26 *If $\tau \in (0, \frac{1}{2})$ and $c \ge 9$, then $1 - t(\tau) - \frac{1}{c}\left|t'(\tau)\right| \ge -2e^{-\sqrt{c/2}}$.*

Proof We have $t'(\tau) = \frac{e^{\frac{1}{\tau-\tau^2}}}{\left(e^{\frac{1}{1-\tau}} + e^{\frac{1}{\tau}}\right)^2} \cdot \frac{\left(-2\tau^2+2\tau-1\right)}{(\tau-1)^2\tau^2} \le 0$, so

$$\left|t'(\tau)\right| = \frac{e^{\frac{1}{\tau-\tau^2}}}{\left(e^{\frac{1}{1-\tau}} + e^{\frac{1}{\tau}}\right)^2} \cdot \frac{\left(2\tau^2 - 2\tau + 1\right)}{(\tau-1)^2\tau^2}. \tag{26}$$

Consider $\tau \in (0, \frac{1}{2})$ and also take $c \ge 9$. Using (26) we find

$$\begin{aligned} 1 - t(\tau) - \frac{1}{c}\left|t'(\tau)\right| &= \frac{e^{\frac{1}{\tau-\tau^2}}}{\left(e^{\frac{1}{1-\tau}} + e^{\frac{1}{\tau}}\right)^2} \frac{\left(c\tau^4 - 2c\tau^3 + c\tau^2 - 2\tau^2 + 2\tau - 1\right)}{c(\tau-1)^2\tau^2} + \frac{e^{\frac{2}{1-\tau}}}{\left(e^{\frac{1}{1-\tau}} + e^{\frac{1}{\tau}}\right)^2} \\ &\ge \frac{e^{\frac{1}{\tau-\tau^2}}}{\left(e^{\frac{1}{1-\tau}} + e^{\frac{1}{\tau}}\right)^2} \frac{\left(c\tau^4 - 2c\tau^3 + c\tau^2 - 2\tau^2 + 2\tau - 1\right)}{c(\tau-1)^2\tau^2} \\ &\ge \frac{e^{\frac{1}{\tau-\tau^2}}}{\left(e^{\frac{1}{1-\tau}} + e^{\frac{1}{\tau}}\right)^2}\left(\frac{c-1}{2c} - \frac{1}{c\tau^2}\right) \ge \frac{e^{\frac{1}{\tau-\tau^2}}}{\left(e^{\frac{1}{1-\tau}} + e^{\frac{1}{\tau}}\right)^2} \min\left\{\frac{c-1}{2c} - \frac{1}{c\tau^2}, 0\right\} \end{aligned}$$

where for the penultimate inequality we used the fact

$$\frac{c\tau^4 - 2c\tau^3 + c\tau^2 - 2\tau^2 + 2\tau - 1}{c(\tau-1)^2\tau^2} \ge \frac{c-1}{2c} - \frac{1}{c\tau^2} \qquad \forall \tau \le \frac{1}{2}, c \ge 7.$$

This algebraic inequality can be verified by a computer algebra system, such as Mathematica.

For $\tau \in (0, \frac{1}{2}]$, we have

$$\frac{e^{\frac{1}{\tau-\tau^2}}}{\left(e^{\frac{1}{1-\tau}} + e^{\frac{1}{\tau}}\right)^2} \le \frac{e^{\frac{1}{\tau-\tau^2}}}{\left(e^{\frac{1}{\tau}}\right)^2} = e^{\frac{1}{\tau-\tau^2} - \frac{2}{\tau}} \le e^{2-\frac{1}{\tau}} \tag{27}$$

where for the last inequality we used that $\frac{1}{\tau-\tau^2} - \frac{2}{\tau} \le 2 - \frac{1}{\tau}$ for $\tau \in [0, \frac{1}{2}]$. Therefore, using (27),

$$1 - t(\tau) - \frac{1}{c}\left|t'(\tau)\right| \ge e^{2-\frac{1}{\tau}} \min\left\{\frac{c-1}{2c} - \frac{1}{c\tau^2}, 0\right\} = \min\left\{e^{2-\frac{1}{\tau}}\left(\frac{c-1}{2c} - \frac{1}{c\tau^2}\right), 0\right\}.$$

We know $\tau \in (0, \frac{1}{2}]$ and $\frac{c-1}{2c} - \frac{1}{c\tau^2} \le 0$ if and only if $0 < \tau \le \frac{\sqrt{2}}{\sqrt{c-1}} \le \frac{1}{2}$. Additionally, $\lim_{\tau \to 0^+} e^{2-\frac{1}{\tau}}\left(\frac{c-1}{2c} - \frac{1}{c\tau^2}\right) = 0$. Therefore the minimum of $\tau \mapsto \min\left\{e^{2-\frac{1}{\tau}}\left(\frac{c-1}{2c} - \frac{1}{c\tau^2}\right), 0\right\}$ for $\tau \in (0, \frac{1}{2})$ must occur at a critical point of $\tau \mapsto e^{2-\frac{1}{\tau}}\left(\frac{c-1}{2c} - \frac{1}{c\tau^2}\right)$. Let's compute that point:

$$0 = \frac{d}{d\tau}\left[e^{2-\frac{1}{\tau}}\left(\frac{c-1}{2c} - \frac{1}{c\tau^2}\right)\right] = \frac{e^{2-\frac{1}{\tau}}\left(c\tau^2 - \tau^2 + 4\tau - 2\right)}{2c\tau^4} \implies \tau = \frac{1}{\frac{\sqrt{c+1}}{\sqrt{2}} + 1}.$$

Therefore,

$$\begin{aligned} 1 - t(\tau) - \frac{1}{c}\left|t'(\tau)\right| &\ge \min\left\{\left[e^{2-\frac{1}{\tau}}\left(\frac{c-1}{2c} - \frac{1}{c\tau^2}\right)\right]_{\tau = \frac{1}{\frac{\sqrt{c+1}}{\sqrt{2}}+1}}, 0\right\} = -\frac{2e^2\left(1 + \frac{\sqrt{c+1}}{\sqrt{2}}\right)}{ce^{1+\frac{\sqrt{c+1}}{\sqrt{2}}}} \\ &\ge -\frac{2}{e^{\frac{\sqrt{c+1}}{\sqrt{2}}}} \ge -\frac{2}{e^{\sqrt{c/2}}}. \end{aligned}$$

■

H Technical details for the ball-packing property

H.1 Geodesics diverge: Proof of Lemma 6

The angle between v_1 and v_2 is in the interval $[0, \pi]$; therefore, the statement of the lemma requires a proof only for $\theta \in [0, \pi]$. We split this into two cases. For both we use the following consequence of Proposition 15:

$$\cosh(\mathrm{dist}(z_1, z_2)\sqrt{-K_{\mathrm{up}}}) \ge \cosh(s\sqrt{-K_{\mathrm{up}}})^2 - \sinh(s\sqrt{-K_{\mathrm{up}}})^2 \cos(\theta). \tag{28}$$

If $\theta > \frac{\pi}{2}$, then

$$\cosh(\mathrm{dist}(z_1, z_2)\sqrt{-K_{\mathrm{up}}}) \ge \cosh(s\sqrt{-K_{\mathrm{up}}})^2 \ge \cosh(s\sqrt{-K_{\mathrm{up}}}),$$

and so $\mathrm{dist}(z_1, z_2) \geq s \geq \frac{2}{3}s$. So we can assume that $\theta \leq \frac{\pi}{2}$.

Note that $e^{1-\frac{2}{3}t} \geq \sqrt{3\left(1 - \frac{\cosh(t)^2 - \cosh(2t/3)}{\sinh(t)^2}\right)}$ for all $t \geq 0$. Therefore,

$$\theta = e^{1-\frac{2}{3}s\sqrt{-K_{\mathrm{up}}}} \geq \sqrt{3\left(1 - \frac{\cosh(s\sqrt{-K_{\mathrm{up}}})^2 - \cosh(2s\sqrt{-K_{\mathrm{up}}}/3)}{\sinh(s\sqrt{-K_{\mathrm{up}}})^2}\right)}$$

which implies $\frac{\cosh(s\sqrt{-K_{\mathrm{up}}})^2 - \cosh(2s\sqrt{-K_{\mathrm{up}}}/3)}{\sinh(s\sqrt{-K_{\mathrm{up}}})^2} \geq 1 - \frac{1}{3}\theta^2 \geq \cos(\theta)$ (since $\theta \in (0, \frac{\pi}{2}]$). Rearranging this inequality and applying inequality (28),

$$\cosh(2s\sqrt{-K_{\mathrm{up}}}/3) \leq \cosh(s\sqrt{-K_{\mathrm{up}}})^2 - \sinh(s\sqrt{-K_{\mathrm{up}}})^2\cos(\theta) \leq \cosh(\mathrm{dist}(z_1, z_2)\sqrt{-K_{\mathrm{up}}}).$$

We conclude $\mathrm{dist}(z_1, z_2) \geq \frac{2}{3}s$.

H.2 Placing well-separated points on the unit sphere

To prove Lemma 7, we used the following lemma about placing well-separated points on the unit sphere $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d : \|x\| = 1\}$. For $x, y \in \mathbb{S}^{d-1}$, $\mathrm{dist}_{\mathbb{S}^{d-1}}(x, y)$ equals the angle between the vectors x and y: $\mathrm{dist}_{\mathbb{S}^{d-1}}(x, y) = \arccos(x^\top y)$.

Below, we use $\mathrm{Vol}(\mathbb{S}^{d-1})$ to denote the volume of the "surface" of the sphere (with the usual metric). Note that $\mathrm{Vol}(\mathbb{S}^{d-1})$ does *not* denote the volume of the unit Euclidean ball in $\mathbb{R}^d$.

Lemma 27 *For any $d \geq 2$ and $\theta \in \left(0, \frac{\pi}{2}\right]$, there are $N \geq \frac{1}{\theta^{d-1}}$ vectors $v_1, \ldots, v_N$ on the $d-1$-dimensional unit sphere $\mathbb{S}^{d-1}$ satisfying $\mathrm{dist}_{\mathbb{S}^{d-1}}(v_i, v_j) \geq \theta \quad \forall i \neq j$.*

Proof The sphere $\mathbb{S}^{d-1}$ is a metric space. The packing number on any metric space is lower bounded by the covering number (Vershynin, 2018, Lem. 4.2.8). More precisely (Vershynin, 2018, Lem. 4.2.8) imples there exist N distinct vectors $\mathcal{V} = \{v_1, v_2, \ldots, v_N\}$ with $v_j \in \mathbb{S}^d$ such that

1. $\mathrm{dist}_{\mathbb{S}^d}(v_i, v_j) \geq \theta$ for all $i \neq j$;
2. and moreover the geodesic balls on the sphere (spherical caps) of radius θ centered at $v \in \mathcal{V}$ cover $\mathbb{S}^{d-1}$, i.e., $\bigcup_{v\in\mathcal{V}} B_v^{\mathbb{S}^{d-1}}(\theta) \supseteq \mathbb{S}^{d-1}$.

(The set $\mathcal{V}$ is said to be a maximally θ-separated net.) Therefore, the sum of the volumes of the balls $\{B_v^{\mathbb{S}^{d-1}}(\theta)\}_{v\in\mathcal{V}}$ must at least be the volume of the unit sphere, i.e.,

$$N \cdot \mathrm{Vol}\Big(B^{\mathbb{S}^{d-1}}(\theta)\Big) = \sum_{v\in\mathcal{V}} \mathrm{Vol}\Big(B_v^{\mathbb{S}^{d-1}}(\theta))\Big) \geq \mathrm{Vol}(\mathbb{S}^{d-1}) = \frac{2\pi^{d/2}}{\Gamma(\frac{d}{2})}.$$

The last equality is the standard formula for the surface area of a sphere in Euclidean space. The volume of a geodesic ball of radius θ on $\mathbb{S}^{d-1}$ is

$$\mathrm{Vol}\Big(B^{\mathbb{S}^{d-1}}(\theta)\Big) = \mathrm{Vol}(\mathbb{S}^{d-2}) \int_0^\theta \sin^{d-2}(\eta)d\eta = \frac{2\pi^{\frac{d-1}{2}}}{\Gamma(\frac{d-1}{2})} \int_0^\theta \sin^{d-2}(\eta)d\eta,$$

see (Lee, 2018, Cor. 10.17) or (Gual-Aenau and Naveira, 1999, p. 314). Using that $\theta \leq \frac{\pi}{2}$ and $\sin(\eta) \leq \eta$ for all $\eta \in [0, \frac{\pi}{2}]$, $\mathrm{Vol}\Big(B^{\mathbb{S}^{d-1}}(\theta)\Big) \leq \frac{2\pi^{\frac{d-1}{2}}}{\Gamma(\frac{d-1}{2})}\frac{1}{d-1}\theta^{d-1}$. Therefore,

$$N \geq \pi^{1/2}\frac{(d-1)\Gamma(\frac{d-1}{2})}{\Gamma(\frac{d}{2})}\frac{1}{\theta^{d-1}} \geq \pi^{1/2}\frac{(d-1)\Gamma(\frac{d-1}{2})}{\Gamma(\frac{d}{2})}\frac{1}{\theta^{d-1}} \geq \frac{1}{\theta^{d-1}}.$$

■

I Geometry influences the objective function

Hamilton and Moitra (2021) show that there is no strongly g-convex function which has bounded condition number on all of the hyperbolic plane. This statement is of course not true in Euclidean space. Using a different technique, Martínez-Rubio (2021) proves a similar result (see Proposition C.6 therein). We extend the result of Hamilton and Moitra (2021) to Hadamard spaces with sectional curvature upper bounded by $K_{\mathrm{up}} < 0$.

Proposition 28 *Let $\mathcal{M}$ be a Hadamard manifold whose sectional curvatures are in the interval $(-\infty, K_{\mathrm{up}}]$ with $K_{\mathrm{up}} < 0$. Let $f\colon \mathcal{M} \to \mathbb{R}$ be L-smooth and μ-strongly g-convex in a ball $B(x_{\mathrm{ref}}, r)$. Then $\frac{L}{\mu} \geq \frac{1}{8}\Big(r\sqrt{-K_{\mathrm{up}}} - 1\Big)$ provided $r \geq \frac{4}{\mu}\left\|\mathrm{grad} f(x_{\mathrm{ref}})\right\| + \frac{1}{\sqrt{-K_{\mathrm{up}}}}$.*

Before proceeding to the proof of Proposition 28, we note that the bound $\kappa \geq \Omega(r)$ also applies to the symmetric spaces $\mathcal{SLP}_n$ and $\mathcal{P}_n$, even though neither have strictly negative curvature. This is an immediate corollary of the result of Hamilton and Moitra (2021) because for every point x in those spaces, there is always a totally geodesic submanifold containing x and which is isometric to a hyperbolic plane (see Appendix J).

Proof The proof is very similar to the proof of Hamilton and Moitra (2021). The main difference is we have to be a little careful because the manifold no longer necessarily has the same symmetries as a hyperbolic space. Denote $\partial B(x_{\mathrm{ref}}, r) = \{x \in \mathcal{M} : \mathrm{dist}(x, x_{\mathrm{ref}}) = r\}$.

Let $c = \frac{1}{\sqrt{-K_{\mathrm{up}}}}$. Let $x \in \arg\min_{y \in \partial B(x_{\mathrm{ref}}, r-c)} f(y)$. Geodesic convexity of f implies

$$\frac{r-c}{r}f(y) + \frac{c}{r}f(x_{\mathrm{ref}}) \geq f\bigg(\exp_{x_{\mathrm{ref}}}\Big(\frac{r-c}{r}\exp_{x_{\mathrm{ref}}}^{-1}(y)\Big)\bigg) \geq f(x) \quad \forall y \in \partial B(x_{\mathrm{ref}}, r).$$

Therefore,

$$f(y) - f(x_{\mathrm{ref}}) \geq \frac{r}{r-c}(f(x) - f(x_{\mathrm{ref}})) \quad \forall y \in \partial B(x_{\mathrm{ref}}, r). \tag{29}$$

On the other hand, μ-strong g-convexity of f implies

$$\begin{aligned} f(x) - f(x_{\mathrm{ref}}) &\geq \left\langle \mathrm{grad} f(x_{\mathrm{ref}}), \exp_{x_{\mathrm{ref}}}^{-1}(x)\right\rangle + \frac{\mu}{2}(r-c)^2 \\ &\geq -\left\|\mathrm{grad} f(x_{\mathrm{ref}})\right\|(r-c) + \frac{\mu}{2}(r-c)^2 \geq \frac{\mu}{4}(r-c)^2 \end{aligned} \tag{30}$$

provided $r - c \geq \frac{4}{\mu}\left\|\mathrm{grad} f(x_{\mathrm{ref}})\right\|$.

Consider any geodesic $\gamma \colon \mathbb{R} \to \mathcal{M}$ with $\gamma(0) = x$, $\|\gamma'(0)\| = 1$ and $\left\langle \gamma'(0), \exp_{x_{\mathrm{ref}}}^{-1}(x) \right\rangle = 0$. We claim $\gamma(\mathbb{R})$ intersects $\partial B(x_{\mathrm{ref}}, r)$ in at least two distinct points y_+, y_-. By Proposition 14,

$$\begin{aligned}\mathrm{dist}(\gamma(t), x_{\mathrm{ref}})^2 &\geq \mathrm{dist}(\gamma(t), x)^2 + (r-c)^2 - 2\left\langle \exp_x^{-1}(\gamma(t)), \exp_x^{-1}(x_{\mathrm{ref}}) \right\rangle \\ &= \mathrm{dist}(\gamma(t), x)^2 + (r-c)^2 = t^2 + (r-c)^2.\end{aligned}$$

Choosing t so that $t^2 + (r-c)^2 > r$, continuity of γ implies that we must have $\gamma(t_+), \gamma(t_-) \in \partial B(x_{\mathrm{ref}}, r)$ for some $t_+ > 0$ and $t_- < 0$. Let $y_+ = \gamma(t_+)$ and $y_- = \gamma(t_-)$. It is clear that $y_+ \neq y_-$ as geodesics do not form closed loops in Hadamard manifolds (Lee, 2018, Prop. 12.9). Observe

$$\exp_x^{-1}(y_+) = t_+\gamma'(0), \qquad \text{and} \qquad \exp_x^{-1}(y_+) = t_-\gamma'(0). \tag{31}$$

By L-smoothness of f,

$$\begin{aligned}f(y_+) &\leq f(x) + \left\langle \mathrm{grad} f(x), \exp_x^{-1}(y_+) \right\rangle + \frac{L}{2}\mathrm{dist}(x, y_+)^2, \\ f(y_-) &\leq f(x) + \left\langle \mathrm{grad} f(x), \exp_x^{-1}(y_-) \right\rangle + \frac{L}{2}\mathrm{dist}(x, y_-)^2\end{aligned}$$

which summed yield

$$\begin{aligned}\frac{-t_-}{t_+ - t_-} f(y_+) + \frac{t_+}{t_+ - t_-} f(y_-) &\leq f(x) + \frac{L}{2}\left(\frac{-t_-}{t_+ - t_-}\mathrm{dist}(x, y_+)^2 + \frac{t_+}{t_+ - t_-}\mathrm{dist}(x, y_-)^2 \right) \\ &\leq f(x) + \frac{L}{2}\Big(\mathrm{dist}(x, y_+)^2 + \mathrm{dist}(x, y_-)^2 \Big)\end{aligned}$$

where we have used (31) to cancel the terms $\left\langle \mathrm{grad} f(x), \exp_x^{-1}(y_\pm) \right\rangle$. Using inequality (29),

$$\begin{aligned}\frac{r}{r-c}(f(x) - f(x_{\mathrm{ref}})) &\leq \frac{-t_-}{t_+ - t_-}(f(y_+) - f(x_{\mathrm{ref}})) + \frac{t_+}{t_+ - t_-}(f(y_-) - f(x_{\mathrm{ref}})) \\ &\leq f(x) - f(x_{\mathrm{ref}}) + \frac{L}{2}\Big(\mathrm{dist}(x, y_+)^2 + \mathrm{dist}(x, y_-)^2 \Big),\end{aligned}$$

which rearranging and applying inequality (30) becomes

$$\frac{\mu}{4}c(r-c) = \frac{c}{r-c} \cdot \frac{\mu}{4}(r-c)^2 \leq \frac{c}{r-c}(f(x) - f(x_{\mathrm{ref}})) \leq \frac{L}{2}\Big(\mathrm{dist}(x, y_+)^2 + \mathrm{dist}(x, y_-)^2 \Big)$$

provided $r - c \geq \frac{4}{\mu}\left\| \mathrm{grad} f(x_{\mathrm{ref}}) \right\|$.

For the last step we shall upper bound $\mathrm{dist}(y_+, x)^2$ and $\mathrm{dist}(y_-, x)^2$. Let us focus on $\mathrm{dist}(y_+, x)^2$ since the exact same reasoning applies to $\mathrm{dist}(y_-, x)^2$. Consider the geodesic triangle $x_{\mathrm{ref}} x y_+$. Again, note that the angle at x is $\frac{\pi}{2}$. So by Proposition 15,

$$\begin{aligned}\cosh(r\sqrt{-K_{\mathrm{up}}}) &= \cosh(\mathrm{dist}(x_{\mathrm{ref}}, y_+)\sqrt{-K_{\mathrm{up}}}) \\ &\geq \cosh(\mathrm{dist}(x, y_+)\sqrt{-K_{\mathrm{up}}}) \cosh(\mathrm{dist}(x_{\mathrm{ref}}, x)\sqrt{-K_{\mathrm{up}}}) \\ &= \cosh(\mathrm{dist}(x, y_+)\sqrt{-K_{\mathrm{up}}}) \cosh((r-c)\sqrt{-K_{\mathrm{up}}}).\end{aligned}$$

Using $e^q \cosh(t-q) = \frac{1}{2}(e^{2q-t}+e^t) \geq \frac{1}{2}(e^{-t}+e^t) = \cosh(t)$ for any $t \in \mathbb{R}$ and $q \geq 0$,

$$\cosh(\mathrm{dist}(x, y_+)\sqrt{-K_{\mathrm{up}}}) \leq \frac{\cosh(r\sqrt{-K_{\mathrm{up}}})}{\cosh((r-c)\sqrt{-K_{\mathrm{up}}})} \leq e^{c\sqrt{-K_{\mathrm{up}}}} = e$$

i.e., $\mathrm{dist}(x, y_+) \leq \frac{1}{\sqrt{-K_{\mathrm{up}}}} \operatorname{arccosh}(e)$.

We conclude that if $r-c \geq \frac{4}{\mu} \|\mathrm{grad} f(x_{\mathrm{ref}})\|$, then $\frac{\mu}{4}c(r-c) \leq \frac{L}{-K_{\mathrm{up}}} \operatorname{arccosh}(e)^2$. Rearranging,

$$\frac{\sqrt{-K_{\mathrm{up}}}(r-\frac{1}{\sqrt{-K_{\mathrm{up}}}})}{8} \leq \frac{\sqrt{-K_{\mathrm{up}}}(r-\frac{1}{\sqrt{-K_{\mathrm{up}}}})}{4\operatorname{arccosh}(e)^2} \leq \frac{L}{\mu},$$

provided $r - \frac{1}{\sqrt{-K_{\mathrm{up}}}} \geq \frac{4}{\mu} \|\mathrm{grad} f(x_{\mathrm{ref}})\|$. ■

J Positive definite matrices

Lemma 29 *Let $\mathcal{M}$ be a Hadamard manifold of dimension d which contains a totally geodesic submanifold $\mathcal{N}$ of dimension d_1. Assume that all the sectional curvatures of the submanifold $\mathcal{N}$ are upper bounded by K_{up}, with $K_{\mathrm{up}} < 0$. Then, $\mathcal{M}$ satisfies the ball-packing property for $\tilde{r} = \frac{4}{\sqrt{-K_{\mathrm{up}}}}, \tilde{c} = \frac{d_1\sqrt{-K_{\mathrm{up}}}}{8}$ and any $x_{\mathrm{ref}} \in \mathcal{N}$. If in addition $\mathcal{M}$ is a homogeneous manifold, then $\mathcal{M}$ satisfies the strong ball-packing property with the same constants $\tilde{r}$ and $\tilde{c}$.*

Proof Let $x_{\mathrm{ref}} \in \mathcal{N}$. By Lemma 7, there are at least $e^{\frac{d_1\sqrt{-K_{\mathrm{up}}}}{8}r}$ points in $B_{\mathcal{M}}(x_{\mathrm{ref}}, \frac{3}{4}r)$ which are pairwise separated by a distance of $\frac{r}{2}$, provided $r \geq \tilde{r}$. Note that here we have used that distance on $\mathcal{N}$ is equal to distance on $\mathcal{M}$ because $\mathcal{N}$ is totally geodesic.

If $\mathcal{M}$ is homogenous, then by definition for all $x, y \in \mathcal{M}$ there is an isometry $\phi\colon \mathcal{M} \to \mathcal{M}$ such that $\phi(x) = y$. In particular, this implies that every $x \in \mathcal{M}$ is an element of a totally geodesic submanifold isometric to $\mathcal{N}$. ■

Lemma 30 *Let $\mathcal{M}_1$ be a d_1-dimensional Hadamard manifold whose sectional curvatures are upper bounded by K_{up} everywhere, with $K_{\mathrm{up}} < 0$. Let $\mathcal{M}_2$ be a Hadamard manifold. Then $\mathcal{M} = \mathcal{M}_1 \times \mathcal{M}_2$ satisfies the strong ball-packing property for $\tilde{r} = \frac{4}{\sqrt{-K_{\mathrm{up}}}}, \tilde{c} = \frac{d_1\sqrt{-K_{\mathrm{up}}}}{8}$ and x_{ref} any point in $\mathcal{M}$.*

Proof This follows from Lemma 29 by noting that for every $x_2 \in \mathcal{M}_2$, $\mathcal{M}_1 \times \{x_2\}$ is a totally geodesic submanifold, so we can apply the same logic from Lemma 29. ■

Let $\mathcal{P}_n = \{P \in \mathbb{R}^{n\times n} : P^\top = P, P \succ 0\}$ be the Riemannian manifold of $n \times n$ positive definite matrices (with real entries), endowed with the so-called affine-invariant metric

$$\langle X, Y\rangle_P = \mathrm{Tr}(P^{-1}XP^{-1}Y) \quad \text{for } P \in \mathcal{P}_n, \text{ and } X, Y \in \mathrm{T}_P\mathcal{P}_n \cong \mathrm{Sym}(n),$$

where $\mathrm{Sym}(n)$ is the set of $n \times n$ real symmetric matrices. Let $\mathcal{SLP}_n = \mathrm{SL}(n)/\mathrm{SO}(n)$ be the totally geodesic submanifold of $\mathcal{P}_n$ consisting of those matrices of determinant one. Both $\mathcal{P}_n$ and

$\mathcal{SLP}_n$ are important in applications (Skovgaard, 1984; Bhatia, 2007; Fletcher and Joshi, 2007; Lenglet et al., 2006; Sra and Hosseini, 2015; Moakher, 2005; Moakher and Batchelor, 2006; Allen-Zhu et al., 2018; Ciobotaru and Mazza, 2020). We know $\mathcal{SLP}_n$ and $\mathcal{P}_n$ are symmetric spaces and Hadamard manifolds (Dolcetti and Pertici, 2018, Prop. 3.1) whose sectional curvatures are each between $-\frac{1}{2}$ and 0 (Criscitiello and Boumal, 2020, Prop. I.1). Since they are symmetric, $\mathcal{SLP}_n$ and $\mathcal{P}_n$ are also a homogeneous manifolds (Lee, 2018, prob. 6-19).

It is well-known that $\mathcal{SLP}_2$ is isomorphic to the hyperbolic plane of curvature $-\frac{1}{2}$ (Chossat and Faugeras, 2009; Dolcetti and Pertici, 2018), and thus satisfies a strong ball property by Lemma 7. For $n \geq 3$, Bridson and Haefliger (1999, Ch. II.10) show that $\mathcal{SLP}_n$ contains a totally geodesic submanifold containing the identity matrix I which is isomorphic to an $(n-1)$-dimensional hyperbolic space for some $K < 0$. We show that $K = -\frac{1}{8}$, see Lemma 31. Therefore applying Lemma 29, $\mathcal{SLP}_n$ satisfies the strong ball-packing property with:

- $\tilde{r} = \frac{4}{\sqrt{1/2}} = 4\sqrt{2}, \tilde{c} = 2\frac{\sqrt{1/2}}{8} = \frac{1}{4\sqrt{2}}$ if $n = 2$;
- $\tilde{r} = \frac{4}{\sqrt{1/8}} = 8\sqrt{2}, \tilde{c} = \frac{(n-1)\sqrt{1/8}}{8} = \frac{n-1}{16\sqrt{2}}$ if $n \geq 3$.

Since $\mathcal{P}_n$ is isometric to $\mathbb{R} \times \mathcal{SLP}_n$ (Dolcetti and Pertici, 2018), Lemma 30 implies the strong ball packing property holds for $\mathcal{P}_n$ with the same constants $\tilde{r}, \tilde{c}$ just given. This proves Lemma 8. We note that Franks and Reichenbach (2021) independently use the observation that $\mathcal{SLP}_n$ contains a hyperbolic *plane* for a similar purpose.

Lemma 31 *For $n \geq 3$, $\mathcal{SLP}_n$ contains a totally geodesic submanifold containing I which is isomorphic to the $(n-1)$-dimensional hyperbolic space of curvature $-\frac{1}{8}$.*

Proof Theorem 10.58 and Remark 10.60(4) of (Bridson and Haefliger, 1999) state that $\mathcal{N} = \mathcal{P}_n \cap O(n-1,1)$ is a totally geodesic submanifold of $\mathcal{P}_n$ which is isometric to a $(n-1)$-dimensional hyperbolic space of some constant sectional curvature $K < 0$. Here, $O(n-1,1) = \{A \in \mathbb{R}^{n\times n} : A^\top J A = J\}$ is an indefinite orthogonal group (symmetries of the $(n-1)$-dimensional hyperboloid model in Minkowski space), where $J = \mathrm{diag}(1,1,\ldots,1,-1)$ (Bridson and Haefliger, 1999, Ex. 10.20(4)).

Note that $\mathcal{N} \subset \mathcal{SLP}_n \cap O(n-1,1)$ since $A^\top J A = J \implies \det(A)^2 = 1$, and any positive definite matrix has positive determinant. Thus, $\mathcal{N}$ is also a totally geodesic submanifold of $\mathcal{SLP}_n$.

We have $\mathrm{T}_I O(n-1,1) = \{X \in \mathbb{R}^{n\times n} : X^\top J = -JX\}$. Therefore,

$$\mathrm{T}_I\mathcal{N} = \mathrm{Sym}_0(n) \cap \mathrm{T}_I O(n-1,1) = \left\{ \begin{pmatrix} 0_{(n-1)\times(n-1)} & s \\ s^\top & 0 \end{pmatrix} : s \in \mathbb{R}^{n-1} \right\}$$

where $\mathrm{Sym}_0(n)$ is the set of $n \times n$ real symmetric matrices with vanishing trace.

Let $s_1, s_2 \in \mathbb{R}^{n-1}$ with $\|s_1\|^2 = \|s_2\|^2 = 1/2, s_1^\top s_2 = 0$. Let

$$X_1 = \begin{pmatrix} 0_{(n-1)\times(n-1)} & s_1 \\ s_1^\top & 0 \end{pmatrix}, \quad X_2 = \begin{pmatrix} 0_{(n-1)\times(n-1)} & s_2 \\ s_2^\top & 0 \end{pmatrix}.$$

Therefore $\langle X_1, X_2\rangle = 0, \|X_1\|^2 = \|X_2\|^2 = 1$, and $[X_1, X_2] = \begin{pmatrix} s_1 s_2^\top - s_2 s_1^\top & 0 \\ 0 & 0 \end{pmatrix}$.

By Proposition 2.3 of (Dolcetti and Pertici, 2018), the curvature tensor of $\mathcal{SLP}_d$ is

$$\mathrm{Rm}(W, X, Y, Z)(P) = -\frac{1}{4}\mathrm{Tr}([P^{-1}W, P^{-1}X][P^{-1}Y, P^{-1}Z]), \quad \text{for } W, X, Y, Z \in \mathrm{Sym}_0(n)$$

where $[X, Y] = XY - YX$ is the matrix commutator of X, Y. Therefore,

$$\begin{aligned} K &= \mathrm{Rm}(X_1, X_2, X_2, X_1)(I) = -\frac{1}{4}\mathrm{Tr}([X_1, X_2][X_2, X_1]) = \frac{1}{4}\mathrm{Tr}([X_1, X_2]^2) \\ &= \frac{1}{4}\mathrm{Tr}((s_1 s_2^\top - s_2 s_1^\top)^2) = \frac{1}{4}\cdot\frac{1}{2}\mathrm{Tr}(-s_1 s_1^\top - s_2 s_2^\top) = -\frac{1}{8}. \end{aligned}$$

■

For positive definite matrices, $\tilde{c} \leq O(n^{3/2})$

We do not know if the constant $\tilde{c}$ stated in Lemma 8 is the best possible constant (i.e., is as large as possible). Dolcetti and Pertici (2018) show that $\mathcal{SLP}_n$ is an Einstein manifold with constant Ricci curvature $-\frac{n}{4}$. Therefore, by the Bishop-Gromov volume comparison theorem (Lee, 2018, Thm. 11.19), the volume of a geodesic ball of radius r in $\mathcal{SLP}_n$ is at most the volume of a geodesic ball in a $\dim(\mathcal{SLP}_n)$-dimensional hyperbolic space of sectional curvature $-\frac{n}{4(\dim(\mathcal{SLP}_n)-1)}$. Hence, the volume of a geodesic ball of radius r in $\mathcal{SLP}_n$ is at most

$$\exp\left(\Theta\left(\dim(\mathcal{SLP}_n) r\sqrt{\frac{n}{4(\dim(\mathcal{SLP}_n)-1)}}\right)\right) = \exp(\Theta(rn^{3/2})).$$

On the other hand, for r sufficiently large, the volume of a geodesic ball of radius $\frac{r}{4}$ in $\mathcal{SLP}_n$ is at least 1. So the number of disjoint balls of radius $\frac{r}{4}$ we can pack into a ball of radius r in $\mathcal{SLP}_n$ is at most $\exp(\Theta(rn^{3/2}))$, which implies $\tilde{c} \leq \Theta(n^{3/2})$.

K Comparison to Riemannian Gradient Descent

Zhang and Sra (2016) show that, for a bounded g-convex domain D with diameter $2r$, projected RGD initialized in D finds a point x within $\frac{r}{5}$ of the minimizer of f in no more than $\tilde{O}(\max\{\kappa, r\sqrt{-K_{\mathrm{lo}}}\})$ queries. This rate depends on curvature. However, if $\mathcal{M}$ is a hyperbolic space of curvature $K < 0$, Proposition 28 implies $\kappa \geq \Omega(r\sqrt{-K})$. Hence, RGD uses at most $\tilde{O}(\kappa)$ queries when $\mathcal{M}$ is a hyperbolic space—this is a curvature-independent rate. We have the following proposition.

Proposition 32 *Let $\mathcal{M}$ be a hyperbolic space of curvature $K < 0$, and let $x_{\mathrm{ref}} \in \mathcal{M}$. Let $L \geq \mu > 0$, $\kappa = \frac{L}{\mu}$, and $r > 0$. Let $f \in \mathcal{F}_{\kappa,r}^{x_{\mathrm{ref}}}$ be L-smooth and have minimizer x^*. Then projected RGD*

$$x_{k+1} = \mathrm{Proj}_D\Big(\exp_{x_k}\Big(-\frac{1}{L}\mathrm{grad}f(x_k)\Big)\Big), \qquad x_0 = x_{\mathrm{ref}}, \qquad D = B(x_{\mathrm{ref}}, r)$$

satisfies $\mathrm{dist}(x_k, x^*)^2 \leq 4\Big(1 - \frac{1}{100}\cdot\frac{1}{\kappa}\Big)^{k-2}\kappa r^2$, *for all* $k \geq 2$. *Here,* Proj_D *denotes metric projection on to the geodesic ball* D.

Proof Zhang and Sra (2016) prove

$$f(x_k) - f(x^*) \le (1-\delta)^{k-2}\frac{1}{2}L(2r)^2 = 2(1-\delta)^{k-2}Lr^2 \qquad \forall k \ge 2$$

where $\delta^{-1} = \max\{\frac{L}{\mu}, \frac{r\sqrt{-K_{\mathrm{lo}}}}{\tanh(r\sqrt{-K_{\mathrm{lo}}})}\}$. By μ-strong g-convexity, $\frac{\mu}{2}\mathrm{dist}(x_k, x^*)^2 \le f(x_k) - f(x^*)$.

First, assume $r\sqrt{-K} < 8$. Then, $\frac{r\sqrt{-K}}{\tanh(r\sqrt{-K})} \le 1 + r\sqrt{-K} \le 9 \le 9\frac{L}{\mu}$. Hence, $\delta^{-1} \le 9\frac{L}{\mu}$.

Second, assume $r\sqrt{-K} \ge 8$. This implies $\frac{r}{4}\sqrt{-K} - 1 \ge \frac{r}{8}\sqrt{-K}$. Proposition 28 applied to the ball $B = B(x^*, \frac{1}{4}r)$ implies that the condition number of f in B is at least $\frac{1}{8}(\frac{r}{4}\sqrt{-K} - 1) \ge \frac{1}{8}(\frac{r}{8}\sqrt{-K}) = \frac{r}{64}\sqrt{-K}$. On the other hand, we know $x^* \in B(x_{\mathrm{ref}}, \frac{3}{4}r)$ because $f \in \mathcal{F}_{\kappa,r}^{x_{\mathrm{ref}}}$. Therefore, $B \subset B(x_{\mathrm{ref}}, r)$, which implies $\frac{L}{\mu} \ge \frac{r}{64}\sqrt{-K}$. We conclude $\frac{r\sqrt{-K}}{\tanh(r\sqrt{-K})} \le 1 + r\sqrt{-K} \le 1 + 64\frac{L}{\mu} \le 65\frac{L}{\mu} \le 100\frac{L}{\mu}$, and so $\delta^{-1} \le 100\frac{L}{\mu}$. ■

L Technical fact from proof of Theorem 11

We show that the inequality (2) implies $|A_k| \ge 2$ for all $k \le T$, where T is given by (1). We do this by induction on $k \ge 0$. (**Base case**) By the ball-packing property, $|A_0| \ge e^{\tilde{c}r} \ge 2$ since $r \ge \frac{4(d+2)}{\tilde{c}} \ge \frac{4}{\tilde{c}}$. (**Inductive hypothesis**) Assume $k+1 \le T$, and $|A_m| \ge 2$ for all $m \le k$. Therefore, $|A_m| - 1 \ge |A_m|/2$ for all $m \le k$.

The bounds $r \ge \frac{4(d+2)}{\tilde{c}}$ and $k+1 \le T$ imply that $k+1 \le \lfloor 2w \rfloor$ (recall $w = \tilde{c}d^{-1}r/4$). So we can apply Lemma 12 to get

$$|A_{m+1}| \ge \frac{|A_m| - 1}{(2000w(3\mathscr{R}\sqrt{-K_{\mathrm{lo}}} + 2))^d} \ge \frac{|A_m|/2}{(2000w(3\mathscr{R}\sqrt{-K_{\mathrm{lo}}} + 2))^d}, \qquad \forall m \le k.$$

Unrolling these inequalities and using $|A_0| \ge e^{\tilde{c}r}$, we get

$$|A_{k+1}| \ge \frac{e^{\tilde{c}r}/2^{k+1}}{\left(2000w(3\mathscr{R}\sqrt{-K_{\mathrm{lo}}} + 2)\right)^{(k+1)d}} \ge \frac{e^{\tilde{c}r}/2^{(k+1)d}}{\left(2000w(3\mathscr{R}\sqrt{-K_{\mathrm{lo}}} + 2)\right)^{(k+1)d}}. \tag{32}$$

On the other hand, using the formula (1) for T, $k+1 \le T$ implies

$$\frac{e^{\tilde{c}r/2}/2^{(k+1)d}}{\left(2000w(3\mathscr{R}\sqrt{-K_{\mathrm{lo}}} + 2)\right)^{(k+1)d}} \ge 1. \tag{33}$$

Combining inequalities (33) and (32) (and using that $e^{\tilde{c}r/2} \le e^{\tilde{c}r}/2$), we determine that $|A_{k+1}| \ge 2$.

M Deriving Theorems 2 and 4 from Theorem 24

Theorem 2 from the introduction follows from Theorem 24 and Lemma 7 (see Appendix M.1). Theorem 4 follows from Theorem 24, Lemma 8, and the fact that $\mathcal{SLP}_n$ has sectional curvatures in the interval $[-\frac{1}{2}, 0]$ (Criscitiello and Boumal, 2020, Prop. I.1) (see Appendix M.2).

M.1 Deriving Theorem 2 from Theorem 24 and Lemma 7

We use the values for $\tilde{r}$ and $\tilde{c}$ given by Lemma 7. First, we have to check that the assumptions of Theorem 2 imply $r \geq \max\left\{\tilde{r}, \frac{8}{\sqrt{-K_{\mathrm{lo}}}}, \frac{4(d+2)}{\tilde{c}}\right\}$. Indeed, the bound $\kappa \geq 1000\sqrt{\frac{K_{\mathrm{lo}}}{K_{\mathrm{up}}}}$ implies $\kappa \geq 1000$, and so

$$\begin{aligned} r &= \frac{\kappa - 9}{12\sqrt{-K_{\mathrm{lo}}}} \geq \frac{\frac{99}{100}\kappa}{12\sqrt{-K_{\mathrm{lo}}}} \geq \frac{990}{12\sqrt{-K_{\mathrm{up}}}} \geq \frac{64}{\sqrt{-K_{\mathrm{up}}}} \\ &\geq \max\left\{\tilde{r}, \frac{8}{\sqrt{-K_{\mathrm{lo}}}}, \frac{4 \cdot 8(d+2)}{d\sqrt{-K_{\mathrm{up}}}}\right\} = \max\left\{\tilde{r}, \frac{8}{\sqrt{-K_{\mathrm{lo}}}}, \frac{4(d+2)}{\tilde{c}}\right\}. \end{aligned}$$

Second, we have to verify the lower bound in Theorem 2 follows from the lower bound for T given in Theorem 24. We have

$$\frac{\sqrt{-K_{\mathrm{up}}}}{8}\frac{\kappa}{12\sqrt{-K_{\mathrm{lo}}}} \geq \tilde{c}(d+2)^{-1}r = \frac{d\sqrt{-K_{\mathrm{up}}}}{8(d+2)}\frac{\kappa - 9}{12\sqrt{-K_{\mathrm{lo}}}} \geq \frac{\sqrt{-K_{\mathrm{up}}}}{16}\frac{\frac{99}{100}\kappa}{12\sqrt{-K_{\mathrm{lo}}}}.$$

Therefore,

$$\begin{aligned} T &\geq \left\lfloor \frac{\sqrt{-K_{\mathrm{up}}}}{16}\frac{\frac{99}{100}\kappa}{12\sqrt{-K_{\mathrm{lo}}}} \cdot \frac{1}{\log(2 \cdot 10^6 \cdot \frac{\sqrt{-K_{\mathrm{up}}}}{8}\frac{\kappa}{12\sqrt{-K_{\mathrm{lo}}}}(r\sqrt{-K_{\mathrm{lo}}})^2)} \right\rfloor \\ &\geq \left\lfloor \frac{\sqrt{-K_{\mathrm{up}}}}{16}\frac{\frac{99}{100}\kappa}{12\sqrt{-K_{\mathrm{lo}}}} \cdot \frac{1}{\log(2^{-2} \cdot 10^6 \cdot \frac{\kappa}{12}(\frac{\kappa}{12})^2)} \right\rfloor \\ &\geq \left\lfloor \frac{\sqrt{-K_{\mathrm{up}}}}{16}\frac{\frac{99}{100}\kappa}{12\sqrt{-K_{\mathrm{lo}}}} \cdot \frac{1}{3\log(10\kappa)} \right\rfloor \geq \left\lfloor \sqrt{\frac{K_{\mathrm{up}}}{K_{\mathrm{lo}}}} \cdot \frac{\kappa}{1000\log(10\kappa)} \right\rfloor. \end{aligned}$$

M.2 Deriving Theorem 4 from Theorem 24 and Lemma 8

We use the values for $\tilde{r}$ and $\tilde{c}$ given by Lemma 8, and $K_{\mathrm{lo}} = -\frac{1}{2}$. First, we have to check that the assumptions of Theorem 4 imply $r \geq \max\left\{\tilde{r}, \frac{8}{\sqrt{-K_{\mathrm{lo}}}}, \frac{4(d+2)}{\tilde{c}}\right\}$. Indeed, the bound $\kappa \geq 1000n$ implies $\kappa \geq 1000$, and so

$$\begin{aligned} r &= \frac{\kappa - 9}{6\sqrt{2}} \geq \frac{\frac{99}{100}\kappa}{6\sqrt{2}} \geq \frac{990n}{6\sqrt{2}} \geq \max\left\{8\sqrt{2}, \frac{8}{\sqrt{1/2}}, \frac{2(n(n+1)+2)}{\tilde{c}}\right\} \\ &\geq \max\left\{\tilde{r}, \frac{8}{\sqrt{-K_{\mathrm{lo}}}}, \frac{4(d+2)}{\tilde{c}}\right\}. \end{aligned}$$

For the second to last inequality, we used (a) $\frac{990n}{6\sqrt{2}} \geq \frac{2(n(n+1)+2) \cdot 16\sqrt{2}}{n-1}$ for all $n \geq 3$, and (b) $\frac{990n}{6\sqrt{2}} \geq 2(n(n+1)+2)4\sqrt{2}$ if $n = 2$. For the last inequality, we used $d = \dim(\mathcal{SLP}_n) = \frac{n(n+1)}{2} - 1$.

Second, we have to verify the lower bound in Theorem 4 follows from the lower bound for T given in Theorem 24. We have for $n \geq 2$

$$\begin{aligned} \frac{n-1}{4\sqrt{2}}\frac{2}{n(n+1)+2} \cdot \frac{\kappa}{6\sqrt{2}} \geq \tilde{c}(d+2)^{-1}r &= \frac{n-1}{c_n\sqrt{2}}\frac{2}{n(n+1)+2} \cdot \frac{\kappa - 9}{6\sqrt{2}} \\ &\geq \frac{n-1}{16c_n\sqrt{2}}\frac{2}{n(n+1)+2} \cdot \frac{\frac{99}{100}\kappa}{6\sqrt{2}}, \end{aligned}$$

where $c_n = 1$ if $n \geq 3$ and $c_2 = 1/4$. Therefore,

$$\begin{aligned} T &\geq \left\lfloor \frac{n-1}{16 c_n \sqrt{2}} \frac{2}{n(n+1)+2} \cdot \frac{\frac{99}{100}\kappa}{6\sqrt{2}} \cdot \frac{1}{\log(2 \cdot 10^6 \cdot \frac{n-1}{4\sqrt{2}} \frac{2}{n(n+1)+2} \cdot \frac{\kappa}{6\sqrt{2}} (\frac{\kappa}{6\sqrt{2}})^2)} \right\rfloor \\ &\geq \left\lfloor \frac{1}{16\sqrt{2}} \frac{2}{\frac{7}{3}n} \cdot \frac{\frac{99}{100}\kappa}{6\sqrt{2}} \cdot \frac{1}{\log(2 \cdot 10^6 \cdot \frac{1}{4\sqrt{2}} \frac{2}{7} \cdot \frac{\kappa}{6\sqrt{2}} (\frac{\kappa}{6\sqrt{2}})^2)} \right\rfloor \\ &\geq \left\lfloor \frac{1}{16\sqrt{2}} \frac{2}{\frac{7}{3}n} \cdot \frac{\frac{99}{100}\kappa}{6\sqrt{2}} \cdot \frac{1}{3\log(10\kappa)} \right\rfloor \geq \left\lfloor \frac{1}{n} \cdot \frac{1}{1000 \log(10\kappa)} \right\rfloor. \end{aligned}$$

For the third to last inequality, we used that $\frac{3c_n}{7n} \leq \frac{n-1}{n(n+1)+2} \leq 3$ for all $n \geq 2$.