JMLR: Workshop and Conference Proceedings vol 49:1–22, 2016

First-order Methods for Geodesically Convex Optimization

Hongyi Zhang HONGYIZ@MIT.EDU
Suvrit Sra SUVRIT@MIT.EDU
Laboratory for Information & Decision Systems, Massachusetts Institute of Technology

Abstract

Geodesic convexity generalizes the notion of (vector space) convexity to nonlinear metric spaces. But unlike convex optimization, geodesically convex (g-convex) optimization is much less developed. In this paper we contribute to the understanding of g-convex optimization by developing iteration complexity analysis for several first-order algorithms on Hadamard manifolds. Specifically, we prove upper bounds for the global complexity of deterministic and stochastic (sub)gradient methods for optimizing smooth and nonsmooth g-convex functions, both with and without strong g-convexity. Our analysis also reveals how the manifold geometry, especially *sectional curvature*, impacts convergence rates. To the best of our knowledge, our work is the first to provide global complexity analysis for first-order algorithms for general g-convex optimization.

Keywords: first-order methods; geodesic convexity; manifold optimization; nonpositively curved spaces; iteration complexity

1. Introduction

Convex optimization is fundamental to numerous areas including machine learning. Convexity often helps guarantee polynomial runtimes and enables robust, more stable numerical methods. But almost invariably, the use of convexity in machine learning is limited to vector spaces, even though convexity *per se* is not limited to vector spaces. Most notably, it generalizes to *geodesically convex* metric spaces (Gromov, 1978; Bridson and Haefliger, 1999; Burago et al., 2001), through which it offers a much richer setting for developing mathematical models amenable to global optimization.

Our broader aim is to increase awareness about g-convexity (see Definition 1); while our specific focus in this paper is on contributing to the understanding of geodesically convex (g-convex) optimization. In particular, we study first-order algorithms for smooth and nonsmooth g-convex optimization, for which we prove iteration complexity upper bounds. Except for a fundamental lemma that applies to general g-convex metric spaces, we limit our discussion to Hadamard manifolds (Riemannian manifolds with global nonpositive curvature), as they offer the most convenient grounds for generalization[1] while also being relevant to numerous applications (see e.g., Section 1.1).

Specifically, we study optimization problems of the form

$$\min \quad f(x) \qquad \text{such that} \quad x \in \mathcal{X} \subset \mathcal{M}, \tag{1}$$

where $f : \mathcal{M} \to \mathbb{R} \cup \{\infty\}$ is a proper g-convex function, $\mathcal{X}$ is a geodesically convex set and $\mathcal{M}$ is a Hadamard manifold (Bishop and O'Neill, 1969; Gromov, 1978). We solve (1) via first-

1. Hadamard manifolds have unique geodesics between any two points. This key property ensures that the exponential map is a global diffeomorphism. Unique geodesics also make it possible to generalize notions such as convex sets and convex functions. (Compact manifolds such as spheres, do not admit globally geodesically convex functions other than the constant function; local g-convexity is possible, but that is a separate topic).

order methods under a variety of settings analogous to the Euclidean case: nonsmooth, Lipschitz-smooth, and strongly g-convex. We present results for both deterministic and stochastic (where $f(x) = \mathbb{E}[F(x, \xi)]$) g-convex optimization.

Although Riemannian geometry provides tools that enable generalization of Euclidean algorithms (Udriste, 1994; Absil et al., 2009), to obtain iteration complexity bounds we must overcome some fundamental geometric hurdles. We introduce key results that overcome some of these hurdles, and pave the way to analyzing first-order g-convex optimization algorithms.

1.1. Related work and motivating examples

We recollect below a few items of related work and some examples relevant to machine learning, where g-convexity and more generally Riemannian optimization play an important role.

Standard references on Riemannian optimization are (Udriste, 1994; Absil et al., 2009), but these primarily consider problems on manifolds without necessarily assuming g-convexity. Consequently, their analysis is limited to asymptotic convergence (except for (Theorem 4.2, Udriste, 1994) that proves linear convergence for functions with positive-definite and bounded Riemannian Hessians). The recent monograph (Bacák, 2014) is devoted to g-convexity and g-convex optimization on geodesic metric spaces, though without any attention to global complexity analysis. Bacák (2014) also details a noteworthy application: averaging trees in the geodesic metric space of phylogenetic trees (Billera et al., 2001).

At a more familiar level, implicitly the topic of "geometric programming" (Boyd et al., 2007) may be viewed as a special case of g-convex optimization (Sra and Hosseini, 2015). For instance, computing stationary states of Markov chains (e.g., while computing PageRank) may be viewed as g-convex optimization problems by placing suitable geometry on the positive orthant; this idea has a fascinating extension to nonlinear iterations on convex cones (in Banach spaces) endowed with the structure of a geodesic metric space (Lemmens and Nussbaum, 2012).

Perhaps the most important example of such metric spaces is the set of positive definite matrices viewed as a Riemannian or Finsler manifold; a careful study of this setup was undertaken by Sra and Hosseini (2015). They also highlighted applications to maximum likelihood estimation for certain non-Gaussian (heavy- or light-tailed) distributions, resulting in various g-convex and nonconvex likelihood problems; see also (Wiesel, 2012; Zhang et al., 2013). However, none of these three works presents a global convergence rate analysis for their algorithms.

There exist several nonconvex problems where Riemannian optimization has proved quite useful, e.g., low-rank matrix and tensor factorization (Vandereycken, 2013; Ishteva et al., 2011; Mishra et al., 2013); dictionary learning (Sun et al., 2015; Harandi et al., 2012); optimization under orthogonality constraints (Edelman et al., 1998; Moakher, 2002; Shen et al., 2009; Liu et al., 2015); and Gaussian mixture models (Hosseini and Sra, 2015), for which g-convexity helps accelerate manifold optimization to substantially outperform the Expectation Maximization (EM) algorithm.

1.2. Contributions

We summarize the main contributions of this paper below.

– We develop a new inequality (Lemma 5) useful for analyzing the behavior of optimization algorithms for functions in Alexandrov space with curvature bounded below, which can be applied to (not necessarily g-convex) optimization problems on Riemannian manifolds and beyond.

– For g-convex optimization problems on Hadamard manifolds (Riemannian manifolds with global nonpositive sectional curvature), we prove iteration complexity upper bounds for several existing algorithms (Table 1). For the special case of smooth geodesically strongly convex optimization, a prior linear convergence result that uses line-search is known (Udriste, 1994); our results do not require line search. Moreover, as far as we are aware, ours are the first global complexity results for general g-convex optimization.

f	Algorithm	Stepsize	Rate[2]	Averaging[3]	Theorem
g-convex, Lipschitz	projected subgradient	$\frac{D}{L_f\sqrt{ct}}$	$O\left(\sqrt{\frac{c}{t}}\right)$	Yes	9
g-convex, bounded subgradient	projected stochastic subgradient	$\frac{D}{G\sqrt{ct}}$	$O\left(\sqrt{\frac{c}{t}}\right)$	Yes	10
g-strongly convex, Lipschitz	projected subgradient	$\frac{2}{\mu(s+1)}$	$O\left(\frac{c}{t}\right)$	Yes	11
g-strongly convex, bounded subgradient	projected stochastic subgradient	$\frac{2}{\mu(s+1)}$	$O\left(\frac{c}{t}\right)$	Yes	12
g-convex, smooth	projected gradient	$\frac{1}{L_g}$	$O\left(\frac{c}{c+t}\right)$	No	13
g-convex, smooth bounded variance	projected stochastic gradient	$\frac{1}{L_g+\frac{\sigma}{D}\sqrt{ct}}$	$O\left(\frac{c+\sqrt{ct}}{c+t}\right)$	Yes	14
g-strongly convex, smooth	projected gradient	$\frac{1}{L_g}$	$O\left(\left(1-\min\left\{\frac{1}{c},\frac{\mu}{L_g}\right\}\right)^t\right)$	No	15

Table 1: **Summary of results.** This table summarizes the non-asymptotic convergence rates we have proved for various geodesically convex optimization algorithms. s: iterate index; t: total number of iterates; D: diameter of domain; L_f: Lipschitz constant of f; c: a constant dependent on D and on the sectional curvature lower bound κ; G: upper bound of gradient norms; μ: strong convexity constant of f; L_g: Lipschitz constant of the gradient; σ: square root variance of the gradient.

2. Here for simplicity only the dependencies on c and t are shown, while other factors are considered constant and thus omitted. Please refer to the theorems for complete results.
3. "Yes": result holds for proper averaging of the iterates; "No": result holds for the last iterate. Please refer to the theorems for complete results.

2. Background

Before we describe the algorithms and analyze their properties, we would like to introduce some concepts in metric geometry and Riemannian geometry that generalize concepts in Euclidean space.

2.1. Metric Geometry

For generalization of nonlinear optimization methods to metric space, we now recall some basic concepts in metric geometry, which cover vector spaces and Riemannian manifolds as special cases. A *metric space* is a pair (X, d) of set X and distance function d that satisfies positivity, symmetry, and the triangle inequality (Burago et al., 2001). A continuous mapping from the interval $[0, 1]$ to X is called a *path*. The *length* of a path $\gamma : [0, 1] \to X$ is defined as $\mathrm{length}(\gamma) := \sup \sum_{i=1}^{n} d(\gamma(t_{i-1}), \gamma(t_i))$, where the supremum is taken over the set of all partitions $0 = t_0 < \cdots < t_n = 1$ of the interval $[0, 1]$, with an arbitrary $n \in \mathbb{N}$. A metric space is a *length space* if for any $x, y \in X$ and $\epsilon > 0$ there exists a path $\gamma : [0, 1] \to X$ joining x and y such that $\mathrm{length}(\gamma) \leq d(x, y) + \epsilon$. A path $\gamma : [0, 1] \to X$ is called a *geodesic* if it is parametrized by the arc length. If every two points $x, y \in X$ are connected by a geodesic, we say (X, d) is a *geodesic space*. If the geodesic connecting every $x, y \in X$ is unique, the space is called *uniquely geodesic* (Bacák, 2014).

The properties of *geodesic triangles* will be central to our analysis of optimization algorithms. A geodesic triangle $\triangle pqr$ with vertices $p, q, r \in X$ consists of three geodesics $\overline{pq}, \overline{qr}, \overline{rp}$. Given $\triangle pqr \in X$, a *comparison triangle* $\triangle \bar{p}\bar{q}\bar{r}$ in k-plane is a corresponding triangle with the same side lengths in two-dimensional space of constant Gaussian curvature k. A length space with curvature bound is called an Alexandrov space. The notion of angle is defined in the following sense. Let $\gamma : [0, 1] \to X$ and $\eta : [0, 1] \to X$ be two geodesics in (X, d) with $\gamma_0 = \eta_0$, we define the *angle* between γ and η as $\alpha(\gamma, \eta) := \limsup_{s,t \to 0_+} \measuredangle \bar{\gamma}_s \bar{\gamma}_0 \bar{\eta}_t$ where $\measuredangle \bar{\gamma}_s \bar{\gamma}_0 \bar{\eta}_t$ is the angle at $\bar{\gamma}_0$ of the corresponding triangle $\triangle \bar{\gamma}_s \bar{\gamma}_0 \bar{\eta}_t$. We use Toponogov's theorem to relate the angles and lengths of any geodesic triangle in a geodesic space to those of a comparison triangle in a space of constant curvature (Burago et al., 1992, 2001).

2.2. Riemannian Geometry

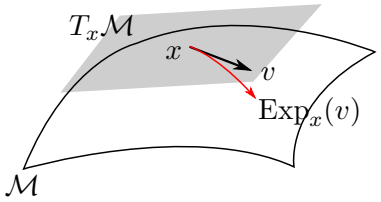

Figure 1: Illustration of a manifold. Also shown are tangent space, geodesic and exponential map.

An n-dimensional *manifold* is a topological space where each point has a neighborhood that is homeomorphic to the n-dimensional Euclidean space. At any point x on a manifold, tangent vectors are defined as the tangents of parametrized curves passing through x. The *tangent space* $T_x\mathcal{M}$ of a manifold $\mathcal{M}$ at x is defined as the set of all tangent vectors at the point x. An exponential map at $x \in \mathcal{M}$ is a mapping from the tangent space $T_x\mathcal{M}$ to $\mathcal{M}$ with the requirement that a

vector $v \in T_x\mathcal{M}$ is mapped to the point $y := \mathrm{Exp}_x(v) \in \mathcal{M}$ such that there exists a geodesic $\gamma : [0, 1] \to \mathcal{M}$ satisfying $\gamma(0) = x, \gamma(1) = y$ and $\gamma'(0) = v$.

As tangent vectors at two different points $x, y \in \mathcal{M}$ lie in different tangent spaces, we cannot compare them directly. To meaningfully compare vectors in different tangent spaces, one needs to define a way to move a tangent vector along the geodesics, while 'preserving' its length and orientation. We thus need to use an inner product structure on tangent spaces, which is called a Riemannian metric. A *Riemannian manifold* $(\mathcal{M}, \mathfrak{g})$ is a real smooth manifold equipped with an inner product $\mathfrak{g}_x$ on the tangent space $T_x\mathcal{M}$ of every point x, such that if u, v are two vector fields on $\mathcal{M}$ then $x \mapsto \langle u, v\rangle_x := \mathfrak{g}_x(u, v)$ is a smooth function. On a Riemannian manifold, the notion of *parallel transport* (parallel displacement) provides a sensible way to transport a vector along a geodesic. Intuitively, a tangent vector $v \in T_x\mathcal{M}$ at x of a geodesic γ is still a tangent vector $\Gamma(\gamma)_x^y v$ of γ after being transported to a point y along γ. Furthermore, parallel transport preserves inner products, i.e. $\langle u, v\rangle_x = \langle \Gamma(\gamma)_x^y u, \Gamma(\gamma)_x^y v\rangle_y$.

The curvature of a Riemannian manifold is characterized by its Riemannian metric tensor at each point. For worst-case analysis, it is sufficient to consider the trigonometry of geodesic triangles. *Sectional curvature* is the Gauss curvature of a two dimensional submanifold formed as the image of a two dimensional subspace of a tangent space after exponential mapping. A two dimensional submanifold with positive, zero or negative sectional curvature is locally isometric to a two dimensional sphere, a Euclidean plane, or a hyperbolic plane with the same Gauss curvature.

2.3. Function Classes on a Riemannian Manifold

We first define some key terms. $\mathcal{X} \subset \mathcal{M}$ is called a geodesically convex set if for any $x, y \in \mathcal{X}$ the minimal distance geodesic γ connecting x, y lies within $\mathcal{X}$. Throughout the paper, we assume that the function f is defined on a Riemannian manifold $\mathcal{M}$, f assumes at least a global minimum point within $\mathcal{X}$, and $x^* \in \mathcal{X}$ is a minimizer of f, unless stated otherwise.

Definition 1 (Geodesic convexity) *A function $f : \mathcal{M} \to \mathbb{R}$ is said to be geodesically convex if for any $x, y \in \mathcal{M}$, a geodesic γ such that $\gamma(0) = x$ and $\gamma(1) = y$, and $t \in [0, 1]$, it holds that*

$$f(\gamma(t)) \leq (1 - t)f(x) + tf(y).$$

It can be shown that an equivalent definition for geodesic convexity is that for any $x, y \in \mathcal{M}$, there exists a tangent vector $g_x \in T_x\mathcal{M}$ such that

$$f(y) \geq f(x) + \langle g_x, \mathrm{Exp}_x^{-1}(y)\rangle_x,$$

and g_x is called a *subgradient* of f at x, or the *gradient* if f is differentiable, and $\langle \cdot, \cdot\rangle_x$ denotes the inner product in the tangent space of x induced by the Riemannian metric. In the rest of the paper we will omit the index of tangent space when it is clear from the context.

Definition 2 (Strong convexity) *A function $f : \mathcal{M} \to \mathbb{R}$ is said to be geodesically μ-strongly convex if for any $x, y \in \mathcal{M}$,*

$$f(y) \geq f(x) + \langle g_x, \mathrm{Exp}_x^{-1}(y)\rangle_x + \frac{\mu}{2}d^2(x, y).$$

or, equivalently, for any geodesic γ such that $\gamma(0) = x, \gamma(1) = y$ and $t \in [0, 1]$,

$$f(\gamma(t)) \leq (1 - t)f(x) + tf(y) - \frac{\mu}{2}t(1 - t)d^2(x, y).$$

Definition 3 (Lipschitzness) *A function $f : \mathcal{M} \to \mathbb{R}$ is said to be geodesically L_f-Lipschitz if for any $x, y \in \mathcal{M}$,*

$$|f(x) - f(y)| \leq L_f d(x, y).$$

Definition 4 (Smoothness) *A differentiable function $f : \mathcal{M} \to \mathbb{R}$ is said to be geodesically L_g-smooth if its gradient is L_g-Lipschitz, i.e. for any $x, y \in \mathcal{M}$,*

$$\|g_x - \Gamma_y^x g_y\| \leq L_g d(x, y)$$

where Γ_y^x is the parallel transport from y to x.

Observe that compared to the Euclidean setup, the above definition requires a parallel transport operation to "transport" g_y to g_x. It can be proved that if f is L_g-smooth, then for any $x, y \in \mathcal{M}$,

$$f(y) \leq f(x) + \langle g_x, \operatorname{Exp}_x^{-1}(y)\rangle_x + \frac{L_g}{2} d^2(x, y).$$

3. Convergence Rates of First-order Methods

In this section, we analyze the global complexity of deterministic and stochastic gradient methods for optimizing various classes of g-convex functions on Hadamard manifolds. We assume access to a projection oracle $\mathcal{P}_\mathcal{X}$ that maps a point $x \in \mathcal{M}$ to $\mathcal{P}_\mathcal{X}(x) \in \mathcal{X} \subset \mathcal{M}$ such that

$$d(x, \mathcal{P}_\mathcal{X}(x)) < d(x, y), \quad \forall y \in \mathcal{X} \setminus \{\mathcal{P}_\mathcal{X}(x)\},$$

where $\mathcal{X}$ is a geodesically convex set and $\max_{y,z\in\mathcal{X}} d(y, z) \leq D$. General projected subgradient / gradient algorithms on Riemannian manifolds take the form

$$x_{s+1} = \mathcal{P}_\mathcal{X}(\operatorname{Exp}_{x_s}(-\eta_s g_s)), \tag{2}$$

where s is the iterate index, g_s is a subgradient of the objective function, and η_s is a step-size. For brevity, we will use the word 'gradient' to refer to both subgradient and gradient, deterministic or stochastic; the meaning should be apparent from the context.

While it is easy to translate first-order optimization algorithms from Euclidean space to Riemannian manifolds, and similarly to prove asymptotic convergence rates (since locally Riemannian manifolds resemble Euclidean space), it is much harder to carry out *non-asymptotic* analysis, at least due to the following two difficulties:

- **Non-Euclidean trigonometry is difficult to use.** Trigonometric geometry in nonlinear spaces is fundamentally different from Euclidean space. In particular, for analyzing optimization algorithms, the law of cosines in Euclidean space

  $$a^2 = b^2 + c^2 - 2bc\cos(A), \tag{3}$$

  where a, b, c are the sides of a Euclidean triangle with A the angle between sides b and c, is an essential tool for bounding the squared distance between the iterates and the minimizer(s). Indeed, consider the Euclidean update $x_{s+1} = x_s - \eta_s g_s$. Applying (3) to the triangle $\triangle x_s x x_{s+1}$, with $a = \overline{x x_{s+1}}$, $b = \overline{x_s x_{s+1}}$, $c = \overline{x x_s}$, and $A = \measuredangle x x_s x_{s+1}$, we get the frequently used formula

  $$\|x_{s+1} - x\|^2 = \|x_s - x\|^2 - 2\eta_s\langle g_s, x_s - x\rangle + \eta_s^2\|g_s\|^2$$

  However, this nice equality does not exist for nonlinear spaces.

- **Linearization does not work.** Another key technique used in bounding squared distances is inspired by the proximal algorithms. Here, gradient-like updates are seen as proximal steps for minimizing a series of *linearizations* of the objective function. Specifically, let $\psi(x; x_s) = f(x_s) + \langle g_s, x - x_s \rangle$ be the linearization of the convex function f, and let $g_s \in \partial f(x_s)$. Then, $x_{s+1} = x_s - \eta_s g_s$ is the unique solution to the following minimization problem

$$\min_x \left\{ \psi(x; x_s) + \frac{1}{2\eta_s} \|x - x_s\|^2 \right\}.$$

  Since $\psi(x; x_s)$ is convex, we thus have (see e.g. Tseng (2009)) the recursively useful bound

$$\psi(x_{s+1}; x_s) + \frac{1}{2\eta_s} \|x_{s+1} - x\|^2 \leq \psi(x; x_s) + \frac{1}{2\eta_s} \|x_s - x\|^2 - \frac{\eta_s}{2} \|g_s\|^2.$$

  But in nonlinear space there is no trivial analogy of a linear function. For example, for any given $y \in \mathcal{M}$ and $g_y \in T_y\mathcal{M}$, the function

$$\psi(x; y) = f(y) + \langle g_y, \operatorname{Exp}_y^{-1}(x) \rangle,$$

  is geodesically both star-concave and star-convex in y, but neither convex nor concave in general. Thus a nonlinear analogue of the above result does not hold.

We address the first difficulty by developing an easy-to-use trigonometric distance bound for Alexandrov space with curvature bounded below. When specialized to Hadamard manifolds, our result reduces to the analysis in (Bonnabel, 2013), which in turn relies on (Cordero-Erausquin et al., 2001, Lemma 3.12). However, unlike (Cordero-Erausquin et al., 2001), our proof assumes no manifold structure on the geodesic space of interest, and is fundamentally different in its techniques.

3.1. Trigonometric Distance Bound

As noted above, a main hurdle in analyzing non-asymptotic convergence of first-order methods in geodesic spaces is that the Euclidean law of cosines does not hold any more. For general nonlinear spaces, there are no corresponding analytical expressions. Even for the (hyperbolic) space of constant negative curvature -1, perhaps the simplest and most studied nonlinear space, the law of cosines is replaced by the *hyperbolic law of cosines*:

$$\cosh a = \cosh b \cosh c - \sinh b \sinh c \cos(A), \tag{4}$$

which does not align well with the standard techniques of convergence rate analysis. With the goal of developing analysis for nonlinear space optimization algorithms, our first contribution is the following trigonometric distance bound for Alexandrov space with curvature bounded below. Owing to its fundamental nature, we believe that this lemma may be of broader interest too.

Lemma 5 *If a, b, c are the sides (i.e., side lengths) of a geodesic triangle in an Alexandrov space with curvature lower bounded by κ, and A is the angle between sides b and c, then*

$$a^2 \leq \frac{\sqrt{|\kappa|}c}{\tanh(\sqrt{|\kappa|}c)} b^2 + c^2 - 2bc\cos(A). \tag{5}$$

Proof sketch. The complete proof contains technical details that digress from the main focus of this paper, so we leave them in the appendix. Below we sketch the main steps.

Our first observation is that by the famous Toponogov's theorem (Burago et al., 1992, 2001), we can upper bound the side lengths of a geodesic triangle in an Alexandrov space with curvature bounded below by the side lengths of a comparison triangle in the hyperbolic plane, which satisfies (*cf.* (4)):

$$\cosh(\sqrt{|\kappa|}a) = \cosh(\sqrt{|\kappa|}b)\cosh(\sqrt{|\kappa|}c) - \sinh(\sqrt{|\kappa|}b)\sinh(\sqrt{|\kappa|}c)\cos(A). \tag{6}$$

Second, we observe that it suffices to study $\kappa = -1$, which corresponds to (4), since Eqn. (6) can be seen as Eqn. (4) with side lengths $a = \sqrt{|\kappa|}a', b = \sqrt{|\kappa|}b', c = \sqrt{|\kappa|}c'$ (see Lemma 19).

Finally, we observe that in (4), $\frac{\partial^2}{\partial b^2}\cosh(a) = \cosh(a)$. Letting $g(b, c, A) := \cosh(\sqrt{\text{rhs}(b, c, A)})$, where $\text{rhs}(b, c, A)$ is the right hand side of (5), we then see that it is sufficient to prove the following:

1. $\cosh(a)$ and $g(b, c, A)$ are equal at $b = 0$.
2. the first partial derivatives of $\cosh(a)$ and $g(b, c, A)$ w.r.t. b agree at $b = 0$.
3. $\frac{\partial^2}{\partial b^2} g(b, c, A) \geq g(b, c, A)$ for $b, c \geq 0$ (Lemma 16).

These three steps, if true, lead to the proof of $\cosh(a) \leq g(b, c, A)$ for $b, c \geq 0$, thus proving a special case of Lemma 5 for space with constant sectional curvature -1 as shown in Lemma 17, 18. Combing this special case with our first two observations concludes the proof of the lemma. ■

Remark 6 *Inequality (5) provides an upper bound on the side lengths of a geodesic triangle in an Alexandrov space with curvature bounded below. Some examples of such spaces are Riemannian manifolds, including hyperbolic space, Euclidean space, sphere, orthogonal groups, and compact sets on a PSD manifold. However, our derivation does not rely on any manifold structure, thus it also applies to certain cones and convex hypersurfaces (Burago et al., 2001).*

We now recall a lemma showing that metric projection in Hadamard manifold is nonexpansive.

Lemma 7 (Bacák (2014)) *Let $(\mathcal{M}, \mathfrak{g})$ be a Hadamard manifold. Let $\mathcal{X} \subset \mathcal{X}$ be a closed convex set. Then the mapping $\mathcal{P}_{\mathcal{X}}(x) := \{y \in \mathcal{X} : d(x, y) = \inf_{z \in \mathcal{X}} d(x, z)\}$ is single-valued and nonexpansive, that is, we have for every $x, y \in \mathcal{M}$*

$$d(\mathcal{P}_{\mathcal{X}}(x), \mathcal{P}_{\mathcal{X}}(y)) \leq d(x, y).$$

In the sequel, we use the notation $\zeta(\kappa, c) \triangleq \frac{\sqrt{|\kappa|}c}{\tanh(\sqrt{|\kappa|}c)}$ for the curvature dependent quantity from inequality (5). From Lemma 5 and Lemma 7 it is straightforward to prove the following corollary, which characterizes an important relation between two consecutive updates of an iterative optimization algorithm on Riemannian manfiold with curvature bounded below.

Corollary 8 *For any Riemannian manifold $\mathcal{M}$ where the sectional curvature is lower bounded by κ and any point x, $x_s \in \mathcal{X}$, the update $x_{s+1} = \mathcal{P}_{\mathcal{X}}(\text{Exp}_{x_s}(-\eta_s g_s))$ satisfies*

$$\langle -g_s, \text{Exp}_{x_s}^{-1}(x)\rangle \leq \frac{1}{2\eta_s}\left(d^2(x_s, x) - d^2(x_{s+1}, x)\right) + \frac{\zeta(\kappa, d(x_s, x))\eta_s}{2}\|g_s\|^2. \tag{7}$$

Proof Denote $\tilde{x}_{s+1} := \operatorname{Exp}_{x_s}(-\eta_s g_s)$. Note that for the geodesic triangle $\triangle x_s \tilde{x}_{s+1} x$, we have $d(x_s, \tilde{x}_{s+1}) = \eta_s \|g_s\|$, while $d(x_s, \tilde{x}_{s+1}) d(x_s, x) \cos(\measuredangle \tilde{x}_{s+1} x_s x) = \langle -\eta_s g_s, \operatorname{Exp}_{x_s}^{-1}(x) \rangle$. Now let $a = \overline{\tilde{x}_{s+1} x}, b = \overline{\tilde{x}_{s+1} x_s}, c = \overline{x_s x}, A = \measuredangle \tilde{x}_{s+1} x_s x$, apply Lemma 5 and simplify to obtain

$$\langle -g_s, \operatorname{Exp}_{x_s}^{-1}(x) \rangle \leq \frac{1}{2\eta_s} \left(d^2(x_s, x) - d^2(\tilde{x}_{s+1}, x) \right) + \frac{\zeta(\kappa, d(x_s, x)) \eta_s}{2} \|g_s\|^2,$$

whereas by Lemma 7 we have $d^2(\tilde{x}_{s+1}, x) \geq d^2(x_{s+1}, x)$, which then implies (7). ■

It is instructive to compare (7) with its Euclidean counterpart (for which actually $\zeta = 1$):

$$\langle -g_s, x - x_s \rangle = \tfrac{1}{2\eta_s} \left(\|x_s - x\|^2 - \|x_{s+1} - x\|^2 \right) + \tfrac{\eta_s}{2} \|g_s\|^2.$$

Note that as long as the diameter of the domain and the sectional curvature remain bounded, ζ is bounded, and we get a meaningful bound in a form similar to the law of cosines in Euclidean space.

Corollary 8 furnishes the missing tool for analyzing non-asymptotic convergence rates of manifold optimization algorithms. We now move to the analysis of several such first-order algorithms.

3.2. Convergence Rate Analysis

Nonsmooth convex optimization. The following two theorems show that both deterministic and stochastic subgradient methods achieve a *curvature-dependent* $O(1/\sqrt{t})$ rate of convergence for g-convex on Hadamard manifolds.

Theorem 9 *Let f be g-convex and L_f-Lipschitz, the diameter of domain be bounded by D, and the sectional curvature lower-bounded by $\kappa \leq 0$. Then, the projected subgradient method with a constant stepsize $\eta_s = \eta = \frac{D}{L_f \sqrt{\zeta(\kappa, D) t}}$ and $\overline{x}_1 = x_1, \overline{x}_{s+1} = \operatorname{Exp}_{\overline{x}_s} \left(\frac{1}{s+1} \operatorname{Exp}_{\overline{x}_s}^{-1}(x_{s+1}) \right)$ satisfies*

$$f(\overline{x}_t) - f(x^*) \leq D L_f \sqrt{\frac{\zeta(\kappa, D)}{t}}.$$

Proof Since f is g-convex, it satisfies $f(x_s) - f(x^*) \leq \langle -g_s, \operatorname{Exp}_{x_s}^{-1}(x^*) \rangle$, which combined with Corollary 8 and the L_f-Lipschitz condition yields the upper bound

$$f(x_s) - f(x^*) \leq \frac{1}{2\eta} \left(d^2(x_s, x^*) - d^2(x_{s+1}, x^*) \right) + \frac{\zeta(\kappa, D) L_f^2 \eta}{2}. \tag{8}$$

Summing over s from 1 to t and dividing by t, we obtain

$$\frac{1}{t} \sum_{s=1}^{t} f(x_s) - f(x^*) \leq \frac{1}{2t\eta} \left(d^2(x_1, x^*) - d^2(x_{t+1}, x^*) \right) + \frac{\zeta(\kappa, D) L_f^2 \eta}{2}. \tag{9}$$

Plugging in $d(x_1, x^*) \leq D$ and $\eta = \frac{D}{L_f \sqrt{\zeta(\kappa, D) t}}$ we further obtain

$$\frac{1}{t} \sum_{s=1}^{t} f(x_s) - f(x^*) \leq D L_f \sqrt{\frac{\zeta(\kappa, D)}{t}}.$$

It remains to show that $f(\overline{x}_t) \leq \frac{1}{t}\sum_{s=1}^{t} f(x_s)$, which can be proved by an easy induction. ■

We note that Theorem 9 and our following results are all generalizations of known results in Euclidean space. Indeed, setting curvature $\kappa = 0$ we can recover the Euclidean convergence rates (in some cases up to a difference in small constant factors). However, for Hadamard manifolds $\kappa < 0$ and the theorem implies that the algorithms may converge more slowly. Also worth noting is that we must be careful in how we obtain the "average" iterate $\overline{x}_t$ on the manifold.

Theorem 10 *If f is g-convex, the diameter of the domain is bounded by D, the sectional curvature of the manifold is lower bounded by $\kappa \leq 0$, and the stochastic subgradient oracle satisfies $\mathbb{E}[\widetilde{g}(x)] = g(x) \in \partial f(x), \mathbb{E}[\|\widetilde{g}_s\|^2] \leq G^2$, then the projected stochastic subgradient method with stepsize $\eta_s = \eta = \frac{D}{G\sqrt{\zeta(\kappa,D)t}}$, and $\overline{x}_1 = x_1, \overline{x}_{s+1} = \operatorname{Exp}_{\overline{x}_s}\left(\frac{1}{s+1}\operatorname{Exp}^{-1}_{\overline{x}_s}(x_{s+1})\right)$ satisfies the upper bound*

$$\mathbb{E}[f(\overline{x}_t) - f(x^*)] \leq DG\sqrt{\frac{\zeta(\kappa, D)}{t}}.$$

Proof The proof structure is very similar, except that for each equation we take expectation with respect to the sequence $\{x_s\}_{s=1}^t$. Since f is g-convex, we have

$$\mathbb{E}[f(x_s) - f(x^*)] \leq \langle -\mathbb{E}[\widetilde{g}_s], \operatorname{Exp}^{-1}_{x_s}(x^*)\rangle,$$

which combined with Corollary 8 and $\mathbb{E}[\|\widetilde{g}_s\|^2] \leq G^2$ yields

$$\mathbb{E}[f(x_s) - f(x^*)] \leq \frac{1}{2\eta}\mathbb{E}\left[d^2(x_s, x^*) - d^2(x_{s+1}, x^*)\right] + \frac{\zeta(\kappa, D)G^2\eta}{2}. \tag{10}$$

Now arguing as in Theorem 9 the proof follows. ■

Strongly convex nonsmooth functions. The following two theorems show that both subgradient method and stochastic subgradient method achieve a curvature dependent $O(1/t)$ rate of convergence for g-strongly convex functions on Hadamard manifolds.

Theorem 11 *If f is geodesically μ-strongly convex and L_f-Lipschitz, and the sectional curvature of the manifold is lower bounded by $\kappa \leq 0$, then the projected subgradient method with $\eta_s = \frac{2}{\mu(s+1)}$ satisfies*

$$f(\overline{x}_t) - f(x^*) \leq \frac{2\zeta(\kappa, D)L_f^2}{\mu(t+1)},$$

where $\overline{x}_1 = x_1$, and $\overline{x}_{s+1} = \operatorname{Exp}_{\overline{x}_s}\left(\frac{2}{s+1}\operatorname{Exp}^{-1}_{\overline{x}_s}(x_{s+1})\right)$.

Proof Since f is geodesically μ-strongly convex, we have

$$f(x_s) - f(x^*) \leq \langle -g_s, \operatorname{Exp}^{-1}_{x_s}(x^*)\rangle - \frac{\mu}{2}d^2(x_s, x^*),$$

which combined with Corollary 8 and L_f-Lipschitz condition yields

$$f(x_s) - f(x^*) \leq \left(\frac{1}{2\eta_s} - \frac{\mu}{2}\right) d^2(x_s, x^*) - \frac{1}{2\eta_s} d^2(x_{s+1}, x^*) + \frac{\zeta(\kappa, D) L_f^2 \eta_s}{2} \tag{11}$$

$$= \frac{\mu(s-1)}{4} d^2(x_s, x^*) - \frac{\mu(s+1)}{4} d^2(x_{s+1}, x^*) + \frac{\zeta(\kappa, D) L_f^2}{\mu(s+1)}. \tag{12}$$

Multiply (12) by s and sum over s from 1 to t; then divide the result by $\frac{t(t+1)}{2}$ to obtain

$$\frac{2}{t(t+1)} \sum_{s=1}^{t} s f(x_s) - f(x^*) \leq \frac{2\zeta(\kappa, D) L_f^2}{\mu(t+1)}. \tag{13}$$

The final step is to show $f(\overline{x}_t) \leq \frac{2}{t(t+1)} \sum_{s=1}^{t} s f(x_s)$, which again follows by an easy induction. ■

Theorem 12 *If f is geodesically μ-strongly convex, the sectional curvature of the manifold is lower bounded by $\kappa \leq 0$, and the stochastic subgradient oracle satisfies $\mathbb{E}[\widetilde{g}(x)] = g(x) \in \partial f(x), \mathbb{E}[\|\widetilde{g}_s\|^2] \leq G^2$, then the projected subgradient method with $\eta_s = \frac{2}{\mu(s+1)}$ satisfies*

$$\mathbb{E}[f(\overline{x}_t) - f(x^*)] \leq \frac{2\zeta(\kappa, D) G^2}{\mu(t+1)}$$

where $\overline{x}_1 = x_1$, and $\overline{x}_{s+1} = \operatorname{Exp}_{\overline{x}_s}\left(\frac{2}{s+1} \operatorname{Exp}_{\overline{x}_s}^{-1}(x_{s+1})\right)$.

Proof The proof structure is very similar to the previous theorem, except that now we take expectations over the sequence $\{x_s\}_{s=1}^t$. We leave the details to the appendix for brevity. ■

Theorems 11 and 12 are generalizations of their Euclidean counterparts (Lacoste-Julien et al., 2012), and follow the same proof structures. Our upper bounds depend linearly on $\zeta(\kappa, D)$, which implies that with $\kappa < 0$ the algorithms may converge more slowly. However, note that for strongly convex problems, the distances from iterates to the minimizer are shrinking, thus the inequality (11) (or its stochastic version) may be too pessimistic, and better dependency on κ may be obtained with a more refined analysis. We leave this as an open problem for the future.

Smooth convex optimization. The following two theorems show that gradient descent algorithm achieves a curvature dependent $O(1/t)$ rate of convergence, whereas stochastic gradient achieves a curvature dependent $O(1/t + \sqrt{1/t})$ rate for smooth g-convex functions on Hadamard manifolds.

Theorem 13 *If $f : \mathcal{M} \to \mathbb{R}$ is g-convex with an L_g-Lipschitz gradient, the diameter of domain is bounded by D, and the sectional curvature of the manifold is bounded below by κ, then projected gradient descent with $\eta_s = \eta = \frac{1}{L_g}$ satisfies for $t > 1$ the upper bound*

$$f(x_t) - f(x^*) \leq \frac{\zeta(\kappa, D) L_g D^2}{2(\zeta(\kappa, D) + t - 2)}.$$

Proof For simplicity we denote $\Delta_s = f(x_s) - f(x^*)$. First observe that with $\eta = \frac{1}{L_g}$ the algorithm is a descent method. Indeed, we have

$$\Delta_{s+1} - \Delta_s \leq \langle g_s, \operatorname{Exp}_{x_s}^{-1}(x_{s+1})\rangle + \frac{L_g}{2} d^2(x_{s+1}, x_s) = -\frac{\|g_s\|^2}{2L_g}. \tag{14}$$

On the other hand, by the convexity of f and Corollary 8 we obtain

$$\Delta_s \leq \langle -g_s, \operatorname{Exp}_{x_s}^{-1}(x^*)\rangle \leq \frac{L_g}{2}\left(d^2(x_s, x^*) - d^2(x_{s+1}, x^*)\right) + \frac{\zeta(\kappa, D)\|g_s\|^2}{2L_g}. \tag{15}$$

Multiplying (14) by $\zeta(\kappa, D)$ and adding to (15), we get

$$\zeta(\kappa, D)\Delta_{s+1} - (\zeta(\kappa, D) - 1)\Delta_s \leq \frac{L_g}{2}\left(d^2(x_s, x^*) - d^2(x_{s+1}, x^*)\right). \tag{16}$$

Now summing over s from 1 to $t-1$, a brief manipulation shows that

$$\zeta(\kappa, D)\Delta_t + \sum_{s=2}^{t-1} \Delta_s \leq (\zeta(\kappa, D) - 1)\Delta_1 + \tfrac{L_g D^2}{2}. \tag{17}$$

Since for $s \leq t$ we proved $\Delta_t \leq \Delta_s$, and by assumption $\Delta_1 \leq \frac{L_g D^2}{2}$, for $t > 1$ we get

$$\Delta_t \leq \frac{\zeta(\kappa, D) L_g D^2}{2(\zeta(\kappa, D) + t - 2)},$$

yielding the desired bound. ■

Theorem 14 *If $f : \mathcal{M} \to \mathbb{R}$ is g-convex with L_g-Lipschitz gradient, the diameter of domain is bounded by D, the sectional curvature of the manifold is bounded below by κ, and the stochastic gradient oracle satisfies $\mathbb{E}[\widetilde{g}(x)] = g(x) = \nabla f(x), \mathbb{E}[\|\nabla f(x) - \widetilde{g}_s\|^2] \leq \sigma^2$, then the projected stochastic gradient algorithm with $\eta_s = \eta = \frac{1}{L_g + 1/\alpha}$ where $\alpha = \frac{D}{\sigma}\sqrt{\frac{1}{\zeta(\kappa, D)t}}$ satisfies for $t > 1$*

$$\mathbb{E}[f(\overline{x}_t) - f(x^*)] \leq \frac{\zeta(\kappa, D) L_g D^2 + 2D\sigma\sqrt{\zeta(\kappa, D)t}}{2(\zeta(\kappa, D) + t - 2)},$$

where $\overline{x}_2 = x_2$, $\overline{x}_{s+1} = \operatorname{Exp}_{\overline{x}_s}\left(\frac{1}{s}\operatorname{Exp}_{\overline{x}_s}^{-1}(x_{s+1})\right)$ for $2 \leq s \leq t-2$, $\overline{x}_t = \operatorname{Exp}_{\overline{x}_{t-1}}\left(\frac{\zeta(\kappa, D)}{\zeta(\kappa, D) + t - 2}\operatorname{Exp}_{\overline{x}_{t-1}}^{-1}(x_t)\right)$.

Proof As before we write $\Delta_s = f(x_s) - f(x^*)$. First we observe that

$$\Delta_{s+1} - \Delta_s \leq \langle g_s, \operatorname{Exp}_{x_s}^{-1}(x_{s+1})\rangle + \frac{L_g}{2} d^2(x_{s+1}, x_s) \tag{18}$$

$$= \langle \widetilde{g}_s, \operatorname{Exp}_{x_s}^{-1}(x_{s+1}\rangle + \langle g_s - \widetilde{g}_s, \operatorname{Exp}_{x_s}^{-1}(x_{s+1})\rangle + \frac{L_g}{2} d^2(x_{s+1}, x_s) \tag{19}$$

$$\leq \langle \widetilde{g}_s, \operatorname{Exp}_{x_s}^{-1}(x_{s+1})\rangle + \frac{\alpha}{2}\|g_s - \widetilde{g}_s\|^2 + \frac{1}{2}\left(L_g + \frac{1}{\alpha}\right) d^2(x_{s+1}, x_s) \tag{20}$$

Taking expectation, and letting $\eta = \frac{1}{L_g+1/\alpha}$, we obtain

$$\mathbb{E}[\Delta_{s+1} - \Delta_s] \leq \frac{\alpha\sigma^2}{2} - \frac{\mathbb{E}[\|\widetilde{g}_s\|^2]}{2\left(L_g + \frac{1}{\alpha}\right)}. \tag{21}$$

On the other hand, using convexity of f and Corollary 8 we get

$$\Delta_s \leq \langle -g_s, \operatorname{Exp}_{x_s}^{-1}(x^*)\rangle \leq \frac{L_g + \frac{1}{\alpha}}{2}\mathbb{E}\left[d^2(x_s, x^*) - d^2(x_{s+1}, x^*)\right] + \frac{\zeta(\kappa, D)\mathbb{E}[\|\widetilde{g}_s\|^2]}{2\left(L_g + \frac{1}{\alpha}\right)}. \tag{22}$$

Multiply (21) by $\zeta(\kappa, D)$ and add to (22), we get

$$\mathbb{E}[\zeta(\kappa, D)\Delta_{s+1} - (\zeta(\kappa, D) - 1)\Delta_s] \leq \frac{L_g + \frac{1}{\alpha}}{2}\mathbb{E}\left[d^2(x_s, x^*) - d^2(x_{s+1}, x^*)\right] + \frac{\alpha\zeta(\kappa, D)\sigma^2}{2}.$$

Summing over s from 1 to $t-1$ and simplifying, we obtain

$$\mathbb{E}[\zeta(\kappa, D)\Delta_t + \sum_{s=2}^{t-1}\Delta_s] \leq \mathbb{E}[(\zeta(\kappa, D) - 1)\Delta_1] + \frac{L_g D^2}{2} + \frac{1}{2}\left(\frac{D^2}{\alpha} + \alpha\zeta(\kappa, D)t\sigma^2\right). \tag{23}$$

Now set $\alpha = \frac{D}{\sigma\sqrt{\zeta(\kappa,D)t}}$, and note that $\Delta_1 \leq \frac{L_g D^2}{2}$; thus, for $t > 1$ we get

$$\mathbb{E}\big[\zeta(\kappa, D)\Delta_t + \textstyle\sum_{s=2}^{t-1}\Delta_s\big] \leq \frac{\zeta(\kappa, D)L_g D^2}{2} + D\sigma\sqrt{\zeta(\kappa, D)t}.$$

Finally, due to g-convexity of f it is easy to verify by induction that

$$\mathbb{E}[f(\overline{x}_t) - f(x^*)] \leq \frac{\mathbb{E}[\zeta(\kappa, D)\Delta_t + \sum_{s=2}^{t-1}\Delta_s]}{\zeta(\kappa, D) + t - 2}.$$

■

Smooth and strongly convex functions. Next we prove that gradient descent achieves a curvature dependent linear rate of convergence for geodesically strongly convex and smooth functions on Hadamard manifolds.

Theorem 15 *If $f : \mathcal{M} \to \mathbb{R}$ is geodesically μ-strongly convex with L_g-Lipschitz gradient, and the sectional curvature of the manifold is bounded below by κ, then the projected gradient descent algorithm with $\eta_s = \eta = \frac{1}{L_g}$, $\epsilon = \min\{\frac{1}{\zeta(\kappa,D)}, \frac{\mu}{L_g}\}$ satisfies for $t > 1$*

$$f(x_t) - f(x^*) \leq \frac{(1-\epsilon)^{t-2}L_g D^2}{2}.$$

Proof As before we use $\Delta_s = f(x_s) - f(x^*)$. Observe that with $\eta = \frac{1}{L_g}$ we have descent:

$$\Delta_{s+1} - \Delta_s \leq \langle g_s, \operatorname{Exp}_{x_s}^{-1}(x_{s+1})\rangle + \frac{L_g}{2}d^2(x_{s+1}, x_s) = -\frac{\|g_s\|^2}{2L_g}. \tag{24}$$

On the other hand, by the strong convexity of f and Corollary 8 we obtain the bounds

$$\Delta_s \leq \langle -g_s, \operatorname{Exp}_{x_s}^{-1}(x^*) \rangle - \frac{\mu}{2} d^2(x_t, x^*) \tag{25}$$

$$\leq \frac{L_g - \mu}{2} d^2(x_s, x^*) - \frac{L_g}{2} d^2(x_{s+1}, x^*) + \frac{\zeta(\kappa, D) \|g_s\|^2}{2L_g}. \tag{26}$$

Multiply (24) by $\zeta(\kappa, D)$ and add to (25) to obtain

$$\zeta(\kappa, D)\Delta_{s+1} - (\zeta(\kappa, D) - 1)\Delta_s \leq \frac{L_g - \mu}{2} d^2(x_s, x^*) - \frac{L_g}{2} d^2(x_{s+1}, x^*) \tag{27}$$

Let $\epsilon = \min\{\frac{1}{\zeta(\kappa, D)}, \frac{\mu}{L_g}\}$, multiply (27) by $(1-\epsilon)^{-(s-1)}$ and sum over s from 1 to $t-1$, we get

$$\zeta(\kappa, D)(1-\epsilon)^{-(t-2)}\Delta_t \leq (\zeta(\kappa, D) - 1)\Delta_1 + \frac{L_g - \mu}{2} d^2(x_1, x^*). \tag{28}$$

Observe that since $\Delta_1 \leq \frac{L_g D^2}{2}$, it follows that $\Delta_t \leq \frac{(1-\epsilon)^{t-2} L_g D^2}{2}$, as desired. ■

It must be emphasized that the proofs of Theorems 13, 14, and 15 contain some additional difficulties beyond their Euclidean counterparts. In particular, the term Δ_s does not cancel nicely due to the presence of the curvature term $\zeta(\kappa, D)$, which necessitates use of a different Lyapunov function to ensure convergence. Consequently, the stochastic gradient algorithm in Theorem 14 requires some unusual looking averaging scheme. In Theorem 15, since the distance between iterates and the minimizer is shrinking, better dependency on κ may also be possible if one replaces $\zeta(\kappa, D)$ by a tighter constant.

4. Experiments

To empirically validate our results, we compare the performance of a stochastic gradient algorithm with a full gradient descent algorithm on the matrix Karcher mean problem. Averaging PSD matrices have applications in averaging data of anisotropic symmetric positive-definite tensors, such as in diffusion tensor imaging (Pennec et al., 2006; Fletcher and Joshi, 2007) and elasticity theory (Cowin and Yang, 1997). The computation and properties of various notions of geometric means have been studied by many (e.g. Moakher (2005); Bini and Iannazzo (2013); Sra and Hosseini (2015)). Specifically, the Karcher mean of a set of N symmetric positive definite matrices $\{A_i\}_{i=1}^N$ is defined as the PSD matrix that minimizes the sum of squared distance induced by the Riemannian metric:

$$d(X, Y) = \|\log(X^{-1/2} Y X^{-1/2})\|_F$$

The loss function

$$f(X; \{A_i\}_{i=1}^N) = \sum_{i=1}^N \|\log(X^{-1/2} A_i X^{-1/2})\|_F^2$$

is known to be nonconvex in Euclidean space but geometrically $2N$-strongly convex , enabling the use of geometrically convex optimization algorithms. The full gradient update step is

$$X_{s+1} = X_s^{1/2} \exp\left(-\eta_s \sum_{i=1}^N \log(X_s^{1/2} A_i^{-1} X_s^{1/2})\right) X_s^{1/2}$$

For stochastic gradient update, we set

$$X_{s+1} = X_s^{1/2} \exp\left(-\eta_s N \log(X_s^{1/2} A_{i(s)}^{-1} X_s^{1/2})\right) X_s^{1/2}$$

where each index $i(s)$ is drawn uniformly at random from $\{1, \ldots, N\}$. The step-sizes η_s for gradient descent and stochastic gradient method have to be chosen according to the smoothness constant or the strongly-convex constant of the loss function. Unfortunately, unlike the Euclidean square loss, there is no cheap way to compute the smoothness constant exactly. In (Bini and Iannazzo, 2013) the authors proposed an adaptive procedure to estimate the optimal step-size. Empirically, however, we observe that an L_g estimate of $5N$ always guarantees convergence. We compare the performance of three algorithms that can be applied to this problem:

- Gradient descent (GD) with $\eta_s = \frac{1}{5N}$ set according to the estimate of the smoothness constant (Theorem 15).
- Stochastic gradient method for smooth functions (SGD-sm) $\eta_s = \frac{1}{L_g + 1/\alpha}$ with $\alpha = \frac{D}{\sigma}\sqrt{\frac{1}{\zeta(\kappa, D)t}}$, where the parameters are set according to the estimates of the smoothness constant, domain diameter and gradient variance (Theorem 14).
- Stochastic subgradient method for strongly convex functions (SGD-st) with $\eta_s = \frac{1}{N(s+1)}$ set according to the $2N$-strong convexity of the loss function (Theorem 12).

Our data are 100×100 random PSD matrices generated using the Matrix Mean Toolbox (Bini and Iannazzo, 2013). All matrices are explicitly normalized so that their norms all equal 1. We compare the algorithms on four datasets with $N \in \{10^2, 10^3\}$ matrices to average and the condition number Q of each matrix being either 10^2 or 10^8. For all experiments we initialize X using the arithmetic mean of the dataset. Figure 4 shows $f(X) - f(X^*)$ as a function of number of passes through the dataset. We observe that the full gradient algorithm with fixed step-size achieves linear convergence, whereas the stochastic gradient algorithms have a sublinear convergence rate, but is much faster during the initial steps.

5. Discussion

In this paper, we make contributions to the understanding of geodesically convex optimization on Hadamard manifolds. Our contributions are twofold: first, we develop a user-friendly trigonometric distance bound for Alexandrov space with curvature bounded below, which includes several commonly known Riemannian manifolds as special cases; second, we prove iteration complexity upper bounds for several first-order algorithms on Hadamard manifolds, which are the first such analyses up to the best of our knowledge. We believe that our analysis is a small step, yet in the right direction, towards understanding and realizing the power of optimization in nonlinear spaces.

5.1. Future Directions

Many questions are not yet answered. We summarize some important ones in the following:

- A long-time question is whether the famous Nesterov's accelerated gradient descent algorithms have nonlinear space counterparts. The analysis of Nesterov's algorithms typically

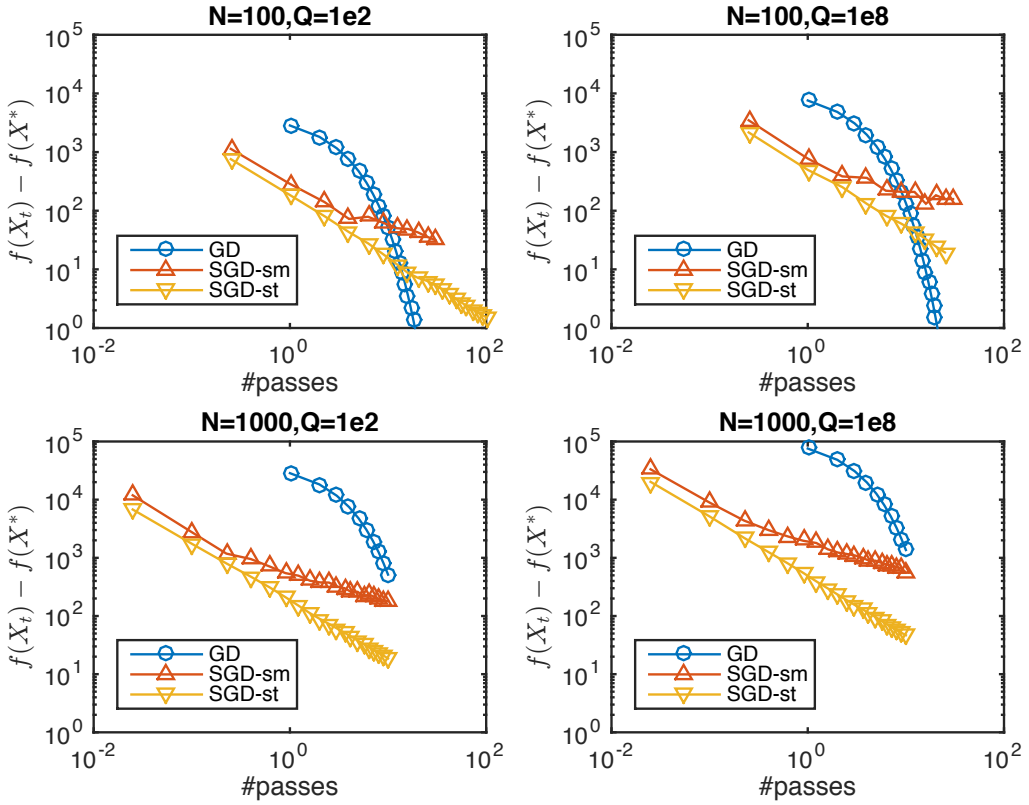

Figure 2: Comparing gradient descent and stochastic gradient methods in matrix Karcher mean problems. Shown are loglog plots of three algorithms on different datasets. GD: gradient descent (Theorem 15); SGD-sm: stochastic gradient method for smooth functions (Theorem 14); SGD-st: stochastic (sub)gradient method for strongly convex functions (Theorem 12). We varied two parameters: size of the dataset $n \in \{10^2, 10^3\}$ and conditional number $Q \in \{10^2, 10^8\}$. Data generating process, initialization and step-size are described in the main text. It is validated from the figures that GD converges at a linear rate, SGD-sm converges asymptotically at the $O(1/\sqrt{t})$ rate, and SGD-st converges at the $O(1/t)$ rate.

relies on a proximal gradient projection interpretation. In nonlinear space, we have not been able to find an analogy to such a projection. Further study is needed to see if similar analysis can be developed, or a different approach is required, or Nesterov's algorithms have no nonlinear space counterparts.

- Another interesting direction is variance reduced stochastic gradient methods for geodesically convex functions. For smooth and convex optimization in Euclidean space, these methods have recently drawn great interests and enjoyed remarkable empirical success. We hypothesize that similar algorithms can achieve faster convergence over naive incremental gradient methods on Hadamard manifolds.

- Finally, since in applications it is often favorable to replace exponential mapping with computationally cheap retractions, it is important to understand the effect of this approximation on convergence rate. Analyzing this effect is of both theoretical and practical interests.

Acknowledgments

We thank the anonymous reviewers for helpful suggestions. HZ is generously supported by the Leventhal Graduate Student Fellowship. SS acknowledges partial support from NSF IIS–1409802.

References

P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009.

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

Louis J Billera, Susan P Holmes, and Karen Vogtmann. Geometry of the space of phylogenetic trees. *Advances in Applied Mathematics*, 27(4):733–767, 2001.

Dario A Bini and Bruno Iannazzo. Computing the Karcher mean of symmetric positive definite matrices. *Linear Algebra and its Applications*, 438(4):1700–1710, 2013.

Richard L Bishop and Barrett O'Neill. Manifolds of negative curvature. *Transactions of the American Mathematical Society*, 145:1–49, 1969.

Silvere Bonnabel. Stochastic gradient descent on Riemannian manifolds. *Automatic Control, IEEE Transactions on*, 58(9):2217–2229, 2013.

Stephen Boyd, Seung-Jean Kim, Lieven Vandenberghe, and Arash Hassibi. A tutorial on geometric programming. *Optimization and engineering*, 8(1):67–127, 2007.

Martin R Bridson and André Haefliger. *Metric spaces of non-positive curvature*, volume 319. Springer, 1999.

Dmitri Burago, Yuri Burago, and Sergei Ivanov. *A course in metric geometry*, volume 33. American Mathematical Society Providence, 2001.

Yu Burago, Mikhail Gromov, and Gregory Perel'man. A.D. Alexandrov spaces with curvature bounded below. *Russian mathematical surveys*, 47(2):1, 1992.

Dario Cordero-Erausquin, Robert J McCann, and Michael Schmuckenschläger. A Riemannian interpolation inequality à la Borell, Brascamp and Lieb. *Inventiones mathematicae*, 146(2):219–257, 2001.

Stephen C Cowin and Guoyu Yang. Averaging anisotropic elastic constant data. *Journal of Elasticity*, 46(2):151–180, 1997.

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

Thomas Fletcher and Sarang Joshi. Riemannian geometry for the statistical analysis of diffusion tensor data. *Signal Processing*, 87(2):250–262, 2007.

Mikhail Gromov. Manifolds of negative curvature. *J. Differential Geom*, 13(2):223–230, 1978.

Mehrtash T Harandi, Conrad Sanderson, Richard Hartley, and Brian C Lovell. Sparse coding and dictionary learning for symmetric positive definite matrices: A kernel approach. In *ECCV 2012*, pages 216–229. Springer, 2012.

Reshad Hosseini and Suvrit Sra. Matrix manifold optimization for Gaussian mixtures. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

Mariya Ishteva, P-A Absil, Sabine Van Huffel, and Lieven De Lathauwer. Best low multilinear rank approximation of higher-order tensors, based on the Riemannian trust-region scheme. *SIAM Journal on Matrix Analysis and Applications*, 32(1):115–135, 2011.

Simon Lacoste-Julien, Mark Schmidt, and Francis Bach. A simpler approach to obtaining an O(1/t) convergence rate for the projected stochastic subgradient method. *arXiv preprint arXiv:1212.2002*, 2012.

Bas Lemmens and Roger Nussbaum. *Nonlinear Perron-Frobenius Theory*, volume 189. Cambridge University Press, 2012.

Xin-Guo Liu, Xue-Feng Wang, and Wei-Guo Wang. Maximization of matrix trace function of product Stiefel manifolds. *SIAM Journal on Matrix Analysis and Applications*, 36(4):1489–1506, 2015.

Bamdev Mishra, Gilles Meyer, Francis Bach, and Rodolphe Sepulchre. Low-rank optimization with trace norm penalty. *SIAM Journal on Optimization*, 23(4):2124–2149, 2013.

Maher Moakher. Means and averaging in the group of rotations. *SIAM journal on matrix analysis and applications*, 24(1):1–16, 2002.

Maher Moakher. A differential geometric approach to the geometric mean of symmetric positive-definite matrices. *SIAM Journal on Matrix Analysis and Applications*, 26(3):735–747, 2005.

Xavier Pennec, Pierre Fillard, and Nicholas Ayache. A Riemannian framework for tensor computing. *International Journal of Computer Vision*, 66(1):41–66, 2006.

Hao Shen, Stefanie Jegelka, and Arthur Gretton. Fast kernel-based independent component analysis. *Signal Processing, IEEE Transactions on*, 57(9):3498–3511, 2009.

Suvrit Sra and Reshad Hosseini. Conic Geometric Optimization on the Manifold of Positive Definite Matrices. *SIAM J. Optimization (SIOPT)*, 25(1):713–739, 2015.

Ju Sun, Qing Qu, and John Wright. Complete dictionary recovery over the sphere II: Recovery by Riemannian trust-region method. *arXiv:1511.04777*, 2015.

Paul Tseng. On accelerated proximal gradient methods for convex-concave optimization. *Submitted to SIAM J. Optim*, 2009.

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

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

Ami Wiesel. Geodesic convexity and covariance estimation. *Signal Processing, IEEE Transactions on*, 60(12):6182–6189, 2012.

Teng Zhang, Ami Wiesel, and Maria S Greco. Multivariate generalized Gaussian distribution: Convexity and graphical models. *Signal Processing, IEEE Transactions on*, 61(16):4141–4148, 2013.

Appendix A. Proof of Lemma 1

Lemma 16 *Let*

$$g(b,c) = \cosh\sqrt{\frac{c}{\tanh(c)}b^2 + c^2 - 2bc\cos(A)}$$

then

$$\frac{\partial^2}{\partial b^2}g(b,c) \geq g(b,c), \quad b,c \geq 0$$

Proof If $c = 0$, $g(b,c) = \cosh(b) = \frac{\partial^2}{\partial b^2}g(b,c)$. Now we focus on the case when $c > 0$. If $c > 0$, Let $u = \sqrt{(1+x)b^2 + c^2 - 2bc\cos(A)}$ where $x = x(c)$. We have

$$u^2 = (1+x)b^2 - 2bc\cos(A) + c^2 \geq \frac{c^2(x+\sin^2 A)}{1+x} = u_{\min}^2 > 0$$

$$\frac{\partial^2}{\partial b^2}g(b,c) = \left(1 + x - c^2\left(x+\sin^2 A\right)\frac{1}{u^2}\right)\cosh(u) + c^2\left(x+\sin^2 A\right)\frac{\sinh(u)}{u^3}$$

Since $g(b,c) = \cosh(u) > 0$, it suffices to prove

$$\frac{\frac{\partial^2}{\partial b^2}g(b,c)}{g(b,c)} - 1 = x\left(1 - \frac{c^2}{x}\left(x+\sin^2 A\right)\frac{1}{u^2} + \frac{c^2}{x}\left(x+\sin^2 A\right)\frac{\tanh(u)}{u^3}\right) \geq 0$$

so it suffices to prove

$$h_1(u) = \frac{u^3}{u - \tanh(u)} \geq \frac{c^2}{x}(x + \sin^2 A)$$

Solving for $h_1'(u) = 0$, we get $u = 0$. Since $\lim_{u \to 0_+} h_1(u) = 0$ and $h_1(u) > 0, \forall u > 0$, $h_1(u)$ is monotonically increasing on $u > 0$. Thus $h_1(u) \geq h_1(u_{\min}), \forall u > 0$. Note that $\frac{c^2}{x}(x + \sin^2 A) = \frac{1+x}{x} u_{\min}^2$, thus it suffices to prove

$$h_1(u_{\min}) = \frac{u_{\min}^3}{u_{\min} - \tanh(u_{\min})} \geq \frac{(1+x)u_{\min}^2}{x}$$

or equivalently

$$\frac{\tanh(u_{\min})}{u_{\min}} \geq \frac{1}{1+x}$$

Now fix c and notice that $\frac{\tanh(u_{\min})}{u_{\min}}$ as a function of $\sin^2 A$ is monotonically decreasing. Therefore its minimum is obtained at $\sin^2 A = 1$, where $u_{\min}^2 = u_*^2 = c^2$, i.e. $u_* = c$. So it only remains to show

$$\frac{\tanh(u_*)}{u_*} = \frac{\tanh(c)}{c} \geq \frac{1}{1+x}, \forall c > 0$$

or equivalently

$$1 + x \geq \frac{c}{\tanh(c)}, \forall c > 0$$

which is true by our definition of g. ■

Lemma 17 *Suppose $h(x)$ is twice differentiable on $[r, +\infty)$ with three further assumptions:*

1. $h(r) \leq 0$,
2. $h'(r) \leq 0$,
3. $h''(x) \leq h(x), \forall x \in [r, +\infty)$,

then $h(x) \leq 0, \forall x \in [r, +\infty)$

Proof It suffices to prove $h'(x) \leq 0, \forall x \in [r, +\infty)$. We prove this claim by contradiction.

Suppose the claim doesn't hold, then there exist some $t > s \geq r$ so that $h'(x) \leq 0$ for any x in $[r, s]$, $h'(s) = 0$ and $h'(x) > 0$ is monotonically increasing in $(s, t]$. It follows that for any $x \in [s, t]$ we have

$$h''(x) \leq h(x) \leq \int_r^x h'(u)du \leq \int_s^x h'(u)du \leq (x - s)h'(x) \leq (t - s)h'(x)$$

Thus by Grönwall's inequality,

$$h'(t) \leq h'(s)e^{(t-s)^2} = 0$$

which leads to a contradiction with our assumption $h'(t) > 0$. ■

Lemma 18 *If a, b, c are the sides of a (geodesic) triangle in a hyperbolic space of constant curvature -1, and A is the angle between b and c, then*

$$a^2 \leq \frac{c}{\tanh(c)} b^2 + c^2 - 2bc\cos(A)$$

Proof For a fixed but arbitrary $c \geq 0$, define $h_c(x) = f(x, c) - g(x, c)$. By Lemma 16 it is easy to verify that $h_c(x)$ satisfies the assumptions of Lemma 17. Apply Lemma 17 to h_c with $r = 0$ to show $h_c \leq 0$ in $[0, +\infty)$. Therefore $f(b, c) \leq g(b, c)$ for any $b, c \geq 0$. Finally use the fact that $\cosh(x)$ is monotonically increasing on $[0, +\infty)$. ■

Corollary 19 *If a, b, c are the sides of a (geodesic) triangle in a hyperbolic space of constant curvature κ, and A is the angle between b and c, then*

$$a^2 \leq \frac{\sqrt{|\kappa|}c}{\tanh(\sqrt{|\kappa|}c)} b^2 + c^2 - 2bc\cos(A)$$

Proof For hyperbolic space of constant curvature $\kappa < 0$, the law of cosines is

$$\cosh(\sqrt{|\kappa|}a) = \cosh(\sqrt{|\kappa|}b)\cosh(\sqrt{|\kappa|}c) - \sinh(\sqrt{|\kappa|}b)\sinh(\sqrt{|\kappa|}c)\cos A$$

which corresponds to the law of cosines of a geodesic triangle in hyperbolic space of curvature -1 with side lengths $\sqrt{|\kappa|}a$, $\sqrt{|\kappa|}b$, $\sqrt{|\kappa|}c$. Applying Lemma 18 we thus get

$$|\kappa|a^2 \leq \frac{\sqrt{|\kappa|}c}{\tanh(\sqrt{|\kappa|}c)} |\kappa|b^2 + |\kappa|c^2 - 2|\kappa|bc\cos(A)$$

and the corollary follows directly. ■

Appendix B. Proof of Theorem 12

Theorem 12 *If f is geodesically μ-strongly convex, the sectional curvature of the manifold is lower bounded by $\kappa \leq 0$, and the stochastic subgradient oracle satisfies $\mathbb{E}[\widetilde{g}(x)] = g(x) \in \partial f(x), \mathbb{E}[\|\widetilde{g}_s\|^2] \leq G^2$, then the projected subgradient method with $\eta_s = \frac{2}{\mu(s+1)}$ satisfies*

$$\mathbb{E}[f(\overline{x}_t) - f(x^*)] \leq \frac{2\zeta(\kappa, D)G^2}{\mu(t+1)}$$

where $\overline{x}_1 = x_1$, and $\overline{x}_{s+1} = \mathrm{Exp}_{\overline{x}_s}\left(\frac{2}{s+1}\mathrm{Exp}_{\overline{x}_s}^{-1}(x_{s+1})\right)$.

Proof Since f is geodesically μ-strongly convex, we have

$$\mathbb{E}[f(x_s) - f(x^*)] \leq \langle -\mathbb{E}[\widetilde{g}_s], \mathrm{Exp}_{x_s}^{-1}(x^*)\rangle - \frac{\mu}{2}\mathbb{E}[d^2(x_s, x^*)]$$

which combined with Corollary 8 and $\mathbb{E}[\|\widetilde{g}_s\|^2] \leq G^2$ yields

$$\begin{aligned}\mathbb{E}[f(x_s) - f(x^*)] &\leq \left(\frac{1}{2\eta_s} - \frac{\mu}{2}\right)\mathbb{E}\left[d^2(x_s, x^*)\right] - \frac{1}{2\eta_s}\mathbb{E}\left[d^2(x_{s+1}, x^*)\right] + \frac{\zeta(\kappa, D)G^2\eta_s}{2} \\ &= \frac{\mu(s-1)}{4}\mathbb{E}\left[d^2(x_s, x^*)\right] - \frac{\mu(s+1)}{4}\mathbb{E}\left[d^2(x_{s+1}, x^*)\right] + \frac{\zeta(\kappa, D)G^2}{\mu(s+1)} \qquad (29)\end{aligned}$$

Multiply (29) by s and sum over s from 1 to t, then divide the result by $\frac{t(t+1)}{2}$ we get

$$\mathbb{E}\left[\frac{2}{t(t+1)}\sum_{s=1}^{t} sf(x_s) - f(x^*)\right] \leq \frac{2\zeta(\kappa, D)G^2}{\mu(t+1)} \qquad (30)$$

The final step is to show $f(\overline{x}_t) \leq \frac{2}{t(t+1)}\sum_{s=1}^{t} sf(x_s)$ by induction. ■