Proceedings of Machine Learning Research vol 75:1–21, 2018 31st Annual Conference on Learning Theory

An Estimate Sequence for Geodesically Convex Optimization

Hongyi Zhang HONGYIZ@MIT.EDU
BCS and LIDS, Massachusetts Institute of Technology

Suvrit Sra SUVRIT@MIT.EDU
EECS and LIDS, Massachusetts Institute of Technology

Editors: Sebastien Bubeck, Vianney Perchet and Philippe Rigollet

Abstract

We propose a Riemannian version of Nesterov's Accelerated Gradient algorithm (RAGD), and show that for *geodesically* smooth and strongly convex problems, within a neighborhood of the minimizer whose radius depends on the condition number as well as the sectional curvature of the manifold, RAGD converges to the minimizer with acceleration. Unlike the algorithm in (Liu et al., 2017) that requires the exact solution to a nonlinear equation which in turn may be intractable, our algorithm is constructive and computationally tractable[1]. Our proof exploits a new estimate sequence and a novel bound on the nonlinear metric distortion, both ideas may be of independent interest.

Keywords: Riemannian optimization; geodesically convex optimization; Nesterov's accelerated gradient method; nonlinear optimization

1. Introduction

Convex optimization theory has been a fruitful area of research for decades, with classic work such as the ellipsoid algorithm (Khachiyan, 1980) and the interior point methods (Karmarkar, 1984). However, with the rise of machine learning and data science, growing problem sizes have shifted the community's focus to first-order methods such as gradient descent and stochastic gradient descent. Over the years, impressive theoretical progress has also been made here, helping elucidate problem characteristics and bringing insights that drive the discovery of provably faster algorithms, notably Nesterov's accelerated gradient descent (Nesterov, 1983) and variance reduced incremental gradient methods (e.g., Johnson and Zhang, 2013; Schmidt et al., 2013; Defazio et al., 2014).

Outside convex optimization, however, despite some recent progress on nonconvex optimization our theoretical understanding remains limited. Nonetheless, nonconvexity pervades machine learning applications and motivates identification and study of specialized structure that enables sharper theoretical analysis, e.g., optimality bounds, global complexity, or faster algorithms. Some examples include, problems with low-rank structure (Boumal et al., 2016b; Ge et al., 2017; Sun et al., 2017; Kawaguchi, 2016); local convergence rates (Ghadimi and Lan, 2013; Reddi et al., 2016; Agarwal et al., 2016; Carmon et al., 2016); growth conditions that enable fast convergence (Polyak, 1963; Zhang et al., 2016; Attouch et al., 2013; Shamir, 2015); and nonlinear constraints based on Riemannian manifolds (Boumal et al., 2016a; Zhang and Sra, 2016; Zhang et al., 2016; Mishra and Sepulchre, 2016), or more general metric spaces (Ambrosio et al., 2014; Bacák, 2014).

1. as long as Riemannian gradient, exponential map and its inverse are computationally tractable, which is the case for many matrix manifolds (Absil et al., 2009).

In this paper, we focus on nonconvexity from a Riemannian viewpoint and consider gradient based optimization. In particular, we are motivated by Nesterov's accelerated gradient method (Nesterov, 1983), a landmark result in the theory of first-order optimization. By introducing an ingenious "estimate sequence" technique, Nesterov (1983) devised a first-order algorithm that provably outperforms gradient descent, and is *optimal* (in a first-order oracle model) up to constant factors. This result bridges the gap between the lower and upper complexity bounds in smooth first-order convex optimization (Nemirovsky and Yudin, 1983; Nesterov, 2004).

Following this seminal work, other researchers also developed different analyses to explain the phenomenon of acceleration. However, both the original proof of Nesterov and all other existing analyses rely heavily on the linear structure of vector spaces. Therefore, our central question is:

Is linear space structure necessary to achieve acceleration?

Given that the iteration complexity theory of gradient descent generalizes to Riemannian manifolds (Zhang and Sra, 2016), it is tempting to hypothesize that a Riemannian generalization of accelerated gradient methods also works. However, the nonlinear nature of Riemannian geometry poses significant obstructions to either verify or refute such a hypothesis. The aim of this paper is to study existence of accelerated gradient methods on Riemannian manifolds, while identifying and tackling key obstructions and obtaining new tools for global analysis of optimization on Riemannian manifolds as a byproduct.

It is important to note that in a recent work (Liu et al., 2017), the authors claimed to have developed Nesterov-style methods on Riemannian manifolds and analyzed their convergence rates. Unfortunately, this is *not* the case, since their algorithm requires the *exact* solution to a nonlinear equation (Liu et al., 2017, (4) and (5)) on the manifold at every iteration. In fact, solving this nonlinear equation itself can be as difficult as solving the original optimization problem.

1.1. Related work

The first accelerated gradient method in vector space along with the concept of estimate sequence is proposed by Nesterov (1983); (Nesterov, 2004, Chapter 2.2.1) contains an expository introduction. In recent years, there has been a surging interest to either develop new analysis for Nesterov's algorithm or invent new accelerated gradient methods. In particular, Su et al. (2014); Flammarion and Bach (2015); Wibisono et al. (2016) take a dynamical system viewpoint, modeling the continuous time limit of Nesterov's algorithm as a second-order ordinary differential equation. Allen-Zhu and Orecchia (2014) reinterpret Nesterov's algorithm as the linear coupling of a gradient step and a mirror descent step, which also leads to accelerated gradient methods for smoothness defined with non-Euclidean norms. Arjevani et al. (2015) reinvent Nesterov's algorithm by considering optimal methods for optimizing polynomials. Bubeck et al. (2015) develop an alternative accelerated method with a geometric explanation. Lessard et al. (2016) use theory from robust control to derive convergence rates for Nesterov's algorithm.

The design and analysis of Riemannian optimization algorithms as well as some historical perspectives were covered in details in (Absil et al., 2009), although the analysis only focused on local convergence. The first global convergence result was derived in (Udriste, 1994) under the assumption that the Riemannian Hessian is positive definite. Zhang and Sra (2016) established the globally convergence rate of Riemannian gradient descent algorithm for optimizing geodesically convex functions on Riemannian manifolds. Other nonlocal analyses of Riemannian optimization algorithms include stochastic gradient algorithm (Zhang and Sra, 2016), fast incremental algorithm (Zhang et al., 2016; Kasai et al., 2016), proximal point algorithm (Ferreira and Oliveira, 2002) and trust-region algorithm (Boumal et al., 2016a). Absil et al. (2009, Chapter 2) also surveyed some important applications of Riemannian optimization.

1.2. Summary of results

In this paper, we make the following contributions:

1. We propose the first *computationally tractable* accelerated gradient algorithm that, within a radius from the minimizer that depends on the condition number and sectional curvature bounds, is provably faster than gradient descent methods on Riemannian manifolds with bounded sectional curvatures. (Algorithm 2, Theorem 11)

2. We analyze the convergence of this algorithm using a new estimate sequence, which relaxes Nesterov's original assumption and also generalizes to Riemannian optimization. (Lemma 4)

3. We develop a novel bound related to the bi-Lipschitz property of exponential maps on Riemannian manifolds. This fundamental geometric result is essential for our convergence analysis, but should also have other interesting applications. (Theorem 10)

2. Background

We briefly review concepts in Riemannian geometry that are related to our analysis; for a thorough introduction one standard text is (e.g. Jost, 2011). A *Riemannian manifold* $(\mathcal{M}, \mathfrak{g})$ is a real smooth manifold $\mathcal{M}$ equipped with a Riemannain metric $\mathfrak{g}$. The metric $\mathfrak{g}$ induces an inner product structure on each tangent space $T_x\mathcal{M}$ associated with every $x \in \mathcal{M}$. We denote the inner product of $u, v \in T_x\mathcal{M}$ as $\langle u, v\rangle \triangleq \mathfrak{g}_x(u, v)$; and the norm of $u \in T_x\mathcal{M}$ is defined as $\|u\|_x \triangleq \sqrt{\mathfrak{g}_x(u, u)}$; we omit the index x for brevity wherever it is obvious from the context. The angle between u, v is defined as $\arccos \frac{\langle u,v\rangle}{\|u\|\|v\|}$. A geodesic is a constant speed curve $\gamma : [0, 1] \to \mathcal{M}$ that is locally distance minimizing. An exponential map $\mathrm{Exp}_x : T_x\mathcal{M} \to \mathcal{M}$ maps v in $T_x\mathcal{M}$ to y on $\mathcal{M}$, such that there is a geodesic γ with $\gamma(0) = x, \gamma(1) = y$ and $\dot\gamma(0) \triangleq \frac{d}{dt}\gamma(0) = v$. If between any two points in $\mathcal{X} \subset \mathcal{M}$ there is a unique geodesic, the exponential map has an inverse $\mathrm{Exp}_x^{-1} : \mathcal{X} \to T_x\mathcal{M}$ and the geodesic is the unique shortest path with $\|\mathrm{Exp}_x^{-1}(y)\| = \|\mathrm{Exp}_y^{-1}(x)\|$ the geodesic distance between $x, y \in \mathcal{X}$. Parallel transport is the Riemannian analogy of vector translation, induced by the Riemannian metric.

Let $u, v \in T_x\mathcal{M}$ be linearly independent, so that they span a two dimensional subspace of $T_x\mathcal{M}$. Under the exponential map, this subspace is mapped to a two dimensional submanifold of $\mathcal{U} \subset \mathcal{M}$. The sectional curvature $\kappa(x,\mathcal{U})$ is defined as the Gauss curvature of $\mathcal{U}$ at x, and is a critical concept in the comparison theorems involving geodesic triangles (Burago et al., 2001).

The notion of geodesically convex sets, geodesically (strongly) convex functions and geodesically smooth functions are defined as straightforward generalizations of the corresponding vector space objects to Riemannian manifolds. In particular,

- A set $\mathcal{X}$ is called *geodesically convex* if for any $x, y \in \mathcal{X}$, there is a geodesic γ with $\gamma(0) = x, \gamma(1) = y$ and $\gamma(t) \in \mathcal{X}$ for $t \in [0, 1]$.

- We call a function $f : \mathcal{X} \to \mathbb{R}$ *geodesically convex* (g-convex) if for any $x, y \in \mathcal{X}$ and any geodesic γ such that $\gamma(0) = x$, $\gamma(1) = y$ and $\gamma(t) \in \mathcal{X}$ for all $t \in [0, 1]$, it holds that

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

It can be shown that if the inverse exponential map is well-defined, an equivalent definition is that for any $x, y \in \mathcal{X}$, $f(y) \ge f(x) + \langle g_x, \mathrm{Exp}_x^{-1}(y)\rangle$, where g_x is the gradient of f at x (in this work we assume f is differentiable). A function $f : \mathcal{X} \to \mathbb{R}$ is called *geodesically μ-strongly convex* (μ-strongly g-convex) if for any $x, y \in \mathcal{X}$ and gradient g_x, it holds that

$$f(y) \ge f(x) + \langle g_x, \mathrm{Exp}_x^{-1}(y)\rangle + \tfrac{\mu}{2}\|\mathrm{Exp}_x^{-1}(y)\|^2.$$

- We call a vector field $g : \mathcal{X} \to \mathbb{R}^d$ *geodesically L-Lipschitz* (L-g-Lipschitz) if for any $x, y \in \mathcal{X}$,

$$\|g(x) - \Gamma_y^x g(y)\| \le L\|\mathrm{Exp}_x^{-1}(y)\|,$$

where Γ_y^x is the parallel transport from y to x. We call a differentiable function $f : \mathcal{X} \to \mathbb{R}$ *geodesically L-smooth* (L-g-smooth) if its gradient is L-g-Lipschitz, in which case we have

$$f(y) \le f(x) + \langle g_x, \mathrm{Exp}_x^{-1}(y)\rangle + \tfrac{L}{2}\|\mathrm{Exp}_x^{-1}(y)\|^2.$$

Throughout our analysis, for simplicity, we make the following standing assumptions:

Assumption 1 *$\mathcal{X} \subset \mathcal{M}$ is a geodesically convex set where the exponential map* Exp *and its inverse* Exp^{-1} *are well defined.*

Assumption 2 *The sectional curvature in $\mathcal{X}$ is bounded, i.e. $|\kappa(x, \cdot)| \le K, \forall x \in \mathcal{X}$.*

Assumption 3 *f is geodesically L-smooth, μ-strongly convex, and assumes its minimum inside $\mathcal{X}$.*

Assumption 4 *All the iterates remain in $\mathcal{X}$.*

With these assumptions, the problem being solved can be stated formally as $\min_{x\in\mathcal{X}\subset\mathcal{M}} \ f(x)$.

3. Proposed algorithm: RAGD

Our proposed optimization procedure is shown in Algorithm 1. We assume the algorithm is granted access to oracles that can efficiently compute the exponential map and its inverse, as well as the Riemannian gradient of function f. In comparison with Nesterov's accelerated gradient method in vector space (Nesterov, 2004, p.76), we note two important differences: first, instead of linearly combining vectors, the update for iterates is computed via exponential maps; second, we introduce a paired sequence of parameters $\{(\gamma_k, \overline{\gamma}_k)\}_{k=0}^{T-1}$, for reasons that will become clear when we analyze the convergence of the algorithm.

Algorithm 1 provides a general scheme for Nesterov-style algorithms on Riemannian manifolds, leaving the choice of many parameters to users' preference. To further simplify the parameter choice as well as the analysis, we note that the following specific choice of parameters

$$\gamma_0 \equiv \gamma = \frac{\sqrt{\beta^2 + 4(1+\beta)\mu h} - \beta}{\sqrt{\beta^2 + 4(1+\beta)\mu h} + \beta} \cdot \mu, \qquad h_k \equiv h, \forall k \ge 0, \qquad \beta_k \equiv \beta > 0, \forall k \ge 0,$$

Algorithm 1: Riemannian-Nesterov$(x_0, \gamma_0, \{h_k\}_{k=0}^{T-1}, \{\beta_k\}_{k=0}^{T-1})$

Parameters: initial point $x_0 \in \mathcal{X}$, $\gamma_0 > 0$, step sizes $\{h_k \leq \frac{1}{L}\}$, shrinkage parameters $\{\beta_k > 0\}$

initialize $v_0 = x_0$

for $k = 0, 1, \ldots, T-1$ **do**

  Compute $\alpha_k \in (0,1)$ from the equation $\alpha_k^2 = h_k \cdot ((1-\alpha_k)\gamma_k + \alpha_k \mu)$

  Set $\overline{\gamma}_{k+1} = (1-\alpha_k)\gamma_k + \alpha_k \mu$

  Choose $y_k = \mathrm{Exp}_{x_k}\left(\frac{\alpha_k \gamma_k}{\gamma_k + \alpha_k \mu} \mathrm{Exp}_{x_k}^{-1}(v_k)\right)$

  Compute $f(y_k)$ and $\mathrm{grad} f(y_k)$

  Set $x_{k+1} = \mathrm{Exp}_{y_k}\left(-h_k \mathrm{grad} f(y_k)\right)$

  Set $v_{k+1} = \mathrm{Exp}_{y_k}\left(\frac{(1-\alpha_k)\gamma_k}{\overline{\gamma}_{k+1}} \mathrm{Exp}_{y_k}^{-1}(v_k) - \frac{\alpha_k}{\overline{\gamma}_{k+1}} \mathrm{grad} f(y_k)\right)$

  Set $\gamma_{k+1} = \frac{1}{1+\beta_k} \overline{\gamma}_{k+1}$

end

Output: x_T

Algorithm 2: Constant Step Riemannian-Nesterov(x_0, h, β)

Parameters: initial point $x_0 \in \mathcal{X}$, step size $h \leq \frac{1}{L}$, shrinkage parameter $\beta > 0$

initialize $v_0 = x_0$

set $\alpha = \frac{\sqrt{\beta^2 + 4(1+\beta)\mu h} - \beta}{2}, \ \gamma = \frac{\sqrt{\beta^2 + 4(1+\beta)\mu h} - \beta}{\sqrt{\beta^2 + 4(1+\beta)\mu h} + \beta} \cdot \mu, \ \overline{\gamma} = (1+\beta)\gamma$

for $k = 0, 1, \ldots, T-1$ **do**

  Choose $y_k = \mathrm{Exp}_{x_k}\left(\frac{\alpha\gamma}{\gamma + \alpha\mu} \mathrm{Exp}_{x_k}^{-1}(v_k)\right)$

  Set $x_{k+1} = \mathrm{Exp}_{y_k}\left(-h \mathrm{grad} f(y_k)\right)$

  Set $v_{k+1} = \mathrm{Exp}_{y_k}\left(\frac{(1-\alpha)\gamma}{\overline{\gamma}} \mathrm{Exp}_{y_k}^{-1}(v_k) - \frac{\alpha}{\overline{\gamma}} \mathrm{grad} f(y_k)\right)$

end

Output: x_T

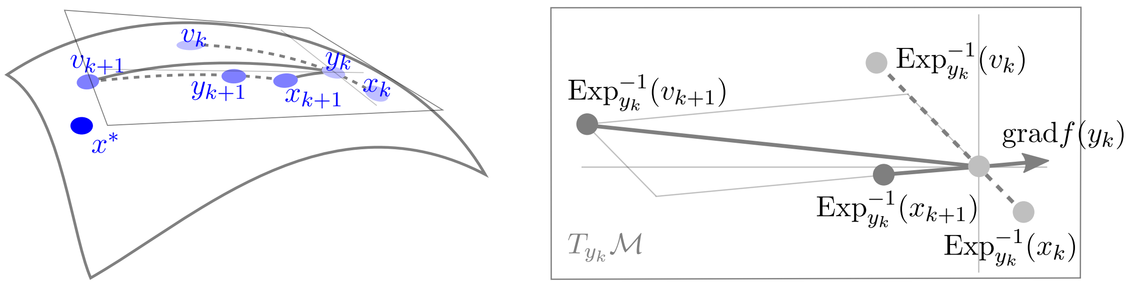

Figure 1: Illustration of the geometric quantities in Algorithm 1. **Left:** iterates and minimizer x^* with y_k's tangent space shown schematically. **Right:** the inverse exponential maps of relevant iterates in y_k's tangent space. Note that y_k is on the geodesic from x_k to v_k (Algorithm 1, Line 1); $\mathrm{Exp}_{y_k}^{-1}(x_{k+1})$ is in the opposite direction of $\mathrm{grad} f(y_k)$ (Algorithm 1, Line 2); also note how $\mathrm{Exp}_{y_k}^{-1}(v_{k+1})$ is constructed (Algorithm 1, Line 3).

which leads to Algorithm 2, a constant step instantiation of the general scheme. We leave the proof of this claim as a lemma in the Appendix.

We move forward to analyzing the convergence properties of these two algorithms in the following two sections. In Section 4, we first provide a novel generalization of Nesterov's estimate sequence to Riemannian manifolds, then show that if a specific tangent space distance comparison inequality (8) always holds, then Algorithm 1 converges similarly as its vector space counterpart. In Section 5, we establish sufficient conditions for this tangent space distance comparison inequality to hold, specifically for Algorithm 2, and show that under these conditions Algorithm 2 converges in $O\left(\sqrt{\frac{L}{\mu}}\log(1/\epsilon)\right)$ iterations, a faster rate than the $O\left(\frac{L}{\mu}\log(1/\epsilon)\right)$ complexity of Riemannian gradient descent.

4. Analysis of a new estimate sequence

First introduced in (Nesterov, 1983), estimate sequences are central tools in establishing the acceleration of Nesterov's method. We first note a weaker notion of estimate sequences for functions whose domain is not necessarily a vector space.

Definition 1 *A pair of sequences $\{\Phi_k(x) : \mathcal{X} \to \mathbb{R}\}_{k=0}^{\infty}$ and $\{\lambda_k\}_{k=0}^{\infty}$ is called a (weak) estimate sequence of a function $f(x) : \mathcal{X} \to \mathbb{R}$, if $\lambda_k \to 0$ and for all $k \geq 0$ we have:*

$$\Phi_k(x^*) \leq (1 - \lambda_k) f(x^*) + \lambda_k \Phi_0(x^*). \tag{1}$$

This definition relaxes the original definition proposed by Nesterov (2004, def. 2.2.1), in that the latter requires $\Phi_k(x) \leq (1 - \lambda_k) f(x) + \lambda_k \Phi_0(x)$ to hold for all $x \in \mathcal{X}$, whereas our definition only assumes it holds at the minimizer x^*. We note that similar observations have been made, e.g., in (Carmon et al., 2017). This relaxation is essential for sparing us from fiddling with the global geometry of Riemannian manifolds.

However, there is one major obstacle in the analysis – Nesterov's construction of quadratic function sequence critically relies on the linear metric and does not generalize to nonlinear space.

An example is given in Figure 2, where we illustrate the distortion of distance (hence quadratic functions) in tangent spaces. The key novelty in our construction is inequality (4) which allows a broader family of estimate sequences, as well as inequality (8) which handles nonlinear metric distortion and fulfills inequality (4). Before delving into the analysis of our specific construction, we recall how to construct estimate sequences and note their use in the following two lemmas.

Lemma 2 *Let us assume that:*

1. *f is geodesically L-smooth and μ-strongly geodesically convex on domain $\mathcal{X}$.*
2. *$\Phi_0(x)$ is an arbitrary function on $\mathcal{X}$.*
3. *$\{y_k\}_{k=0}^{\infty}$ is an arbitrary sequence in $\mathcal{X}$.*
4. *$\{\alpha_k\}_{k=0}^{\infty}$: $\alpha_k \in (0,1)$, $\sum_{k=0}^{\infty} \alpha_k = \infty$.*
5. *$\lambda_0 = 1$.*

Then the pair of sequences $\{\Phi_k(x)\}_{k=0}^{\infty}$, $\{\lambda_k\}_{k=0}^{\infty}$ which satisfy the following recursive rules:

$$\lambda_{k+1} = (1-\alpha_k)\lambda_k, \tag{2}$$

$$\overline{\Phi}_{k+1}(x) = (1-\alpha_k)\Phi_k(x) + \alpha_k \left[f(y_k) + \langle \mathrm{grad} f(y_k), \mathrm{Exp}_{y_k}^{-1}(x) \rangle + \frac{\mu}{2}\|\mathrm{Exp}_{y_k}^{-1}(x)\|^2 \right], \tag{3}$$

$$\Phi_{k+1}(x^*) \leq \overline{\Phi}_{k+1}(x^*), \tag{4}$$

is a (weak) estimate sequence.

The proof is similar to (Nesterov, 2004, Lemma 2.2.2) which we include in Appendix B.

Lemma 3 *If for a (weak) estimate sequence $\{\Phi_k(x) : \mathcal{X} \to \mathbb{R}\}_{k=0}^{\infty}$ and $\{\lambda_k\}_{k=0}^{\infty}$ we can find a sequence of iterates $\{x_k\}$, such that*

$$f(x_k) \leq \Phi_k^* \equiv \min_{x \in \mathcal{X}} \Phi_k(x),$$

then $f(x_k) - f(x^*) \leq \lambda_k(\Phi_0(x^*) - f(x^*)) \to 0$.

Proof By Definition 1 we have $f(x_k) \leq \Phi_k^* \leq \Phi_k(x^*) \leq (1-\lambda_k)f(x^*) + \lambda_k \Phi_0(x^*)$. Hence $f(x_k) - f(x^*) \leq \lambda_k(\Phi_0(x^*) - f(x^*)) \to 0$. ■

Lemma 3 immediately suggest the use of (weak) estimate sequences in establishing the convergence and analyzing the convergence rate of certain iterative algorithms. The following lemma shows that a weak estimate sequence exists for Algorithm 1. Later in Lemma 6, we prove that the sequence $\{x_k\}$ in Algorithm 1 satisfies the requirements in Lemma 3 for our estimate sequence.

Lemma 4 *Let $\Phi_0(x) = \Phi_0^* + \frac{\gamma_0}{2}\|\mathrm{Exp}_{y_0}^{-1}(x)\|^2$. Assume for all $k \geq 0$, the sequences $\{\gamma_k\}$, $\{\overline{\gamma}_k\}$, $\{v_k\}$, $\{\Phi_k^*\}$ and $\{\alpha_k\}$ satisfy*

$$\overline{\gamma}_{k+1} = (1-\alpha_k)\gamma_k + \alpha_k\mu, \tag{5}$$

$$v_{k+1} = \mathrm{Exp}_{y_k}\left(\frac{(1-\alpha_k)\gamma_k}{\overline{\gamma}_{k+1}}\mathrm{Exp}_{y_k}^{-1}(v_k) - \frac{\alpha_k}{\overline{\gamma}_{k+1}}\mathrm{grad}f(y_k)\right) \tag{6}$$

$$\begin{aligned}\Phi_{k+1}^* = &\ (1-\alpha_k)\,\Phi_k^* + \alpha_k f(y_k) - \frac{\alpha_k^2}{2\overline{\gamma}_{k+1}}\|\mathrm{grad}f(y_k)\|^2 \\ &+ \frac{\alpha_k(1-\alpha_k)\gamma_k}{\overline{\gamma}_{k+1}}\left(\frac{\mu}{2}\|\mathrm{Exp}_{y_k}^{-1}(v_k)\|^2 + \langle \mathrm{grad}f(y_k), \mathrm{Exp}_{y_k}^{-1}(v_k)\rangle\right),\end{aligned} \tag{7}$$

$$\gamma_{k+1}\|\mathrm{Exp}_{y_{k+1}}^{-1}(x^*) - \mathrm{Exp}_{y_{k+1}}^{-1}(v_{k+1})\|^2 \leq \overline{\gamma}_{k+1}\|\mathrm{Exp}_{y_k}^{-1}(x^*) - \mathrm{Exp}_{y_k}^{-1}(v_{k+1})\|^2, \tag{8}$$

$$\alpha_k \in (0,1), \quad \sum_{k=0}^{\infty}\alpha_k = \infty, \tag{9}$$

then the pair of sequence $\{\Phi_k(x)\}_{k=0}^{\infty}$ and $\{\lambda_k\}_{k=0}^{\infty}$, defined by

$$\Phi_{k+1}(x) = \Phi_{k+1}^* + \frac{\gamma_{k+1}}{2}\|\mathrm{Exp}_{y_{k+1}}^{-1}(x) - \mathrm{Exp}_{y_{k+1}}^{-1}(v_{k+1})\|^2, \tag{10}$$

$$\lambda_0 = 1, \quad \lambda_{k+1} = (1-\alpha_k)\lambda_k. \tag{11}$$

is a (weak) estimate sequence.

Proof Recall the definition of $\overline{\Phi}_{k+1}(x)$ in Equation (3). We claim that if $\Phi_k(x) = \Phi_k^* + \frac{\gamma_k}{2}\|\mathrm{Exp}_{y_k}^{-1}(x) - \mathrm{Exp}_{y_k}^{-1}(v_k)\|^2$, then we have $\overline{\Phi}_{k+1}(x) \equiv \Phi_{k+1}^* + \frac{\overline{\gamma}_{k+1}}{2}\|\mathrm{Exp}_{y_k}^{-1}(x) - \mathrm{Exp}_{y_k}^{-1}(v_{k+1})\|^2$. The proof of this claim requires a simple algebraic manipulation as is noted as Lemma 5. Now using the assumption (8) we immediately get $\Phi_{k+1}(x^*) \leq \overline{\Phi}_{k+1}(x^*)$. By Lemma 2 the proof is complete. ■

We verify the specific form of $\overline{\Phi}_{k+1}(x)$ in Lemma 5, whose proof can be found in the Appendix C.

Lemma 5 *For all $k \geq 0$, if $\Phi_k(x) = \Phi_k^* + \frac{\gamma_k}{2}\|\mathrm{Exp}_{y_k}^{-1}(x) - \mathrm{Exp}_{y_k}^{-1}(v_k)\|^2$, then with $\overline{\Phi}_{k+1}$ defined as in (3), $\overline{\gamma}_{k+1}$ as in (5), v_{k+1} as in Algorithm 1 and Φ_{k+1}^* as in (7) we have $\overline{\Phi}_{k+1}(x) \equiv \Phi_{k+1}^* + \frac{\overline{\gamma}_{k+1}}{2}\|\mathrm{Exp}_{y_k}^{-1}(x) - \mathrm{Exp}_{y_k}^{-1}(v_{k+1})\|^2$.*

The next lemma asserts that the iterates $\{x_k\}$ of Algorithm 1 satisfy the requirement that the function values $f(x_k)$ are upper bounded by Φ_k^* defined in our estimate sequence.

Lemma 6 *Assume $\Phi_0^* = f(x_0)$, and $\{\Phi_k^*\}$ be defined as in (7) with $\{x_k\}$ and other terms defined as in Algorithm 1. Then we have $\Phi_k^* \geq f(x_k)$ for all $k \geq 0$.*

The proof is standard. We include it in Appendix D for completeness. Finally, we are ready to state the following theorem on the convergence rate of Algorithm 1.

Theorem 7 (Convergence of Algorithm 1) *For any given $T \geq 0$, assume (8) is satisfied for all $0 \leq k \leq T$, then Algorithm 1 generates a sequence $\{x_k\}_{k=0}^{\infty}$ such that*

$$f(x_T) - f(x^*) \leq \lambda_T\left(f(x_0) - f(x^*) + \frac{\gamma_0}{2}\|\mathrm{Exp}_{x_0}^{-1}(x^*)\|^2\right) \tag{12}$$

where $\lambda_0 = 1$ and $\lambda_k = \prod_{i=0}^{k-1}(1-\alpha_i)$.

Proof The proof is similar to (Nesterov, 2004, Theorem 2.2.1). We choose $\Phi_0(x) = f(x_0) + \frac{\gamma_0}{2}\|\mathrm{Exp}_{y_0}^{-1}(x)\|^2$, hence $\Phi_0^* = f(x_0)$. By Lemma 4 and Lemma 6, the assumptions in Lemma 3 hold. It remains to use Lemma 3. ■

5. Local fast rate with a constant step scheme

By now we see that almost all the analysis of Nesterov's generalizes, except that the assumption in (8) is not necessarily satisfied. In vector space, the two expressions both reduce to $x^* - v_{k+1}$ and hence (8) trivially holds with $\gamma = \overline{\gamma}$. On Riemannian manifolds, however, due to the nonlinear Riemannian metric and the associated exponential maps, $\|\mathrm{Exp}_{y_{k+1}}^{-1}(x^*) - \mathrm{Exp}_{y_{k+1}}^{-1}(v_{k+1})\|$ and $\|\mathrm{Exp}_{y_k}^{-1}(x^*) - \mathrm{Exp}_{y_k}^{-1}(v_{k+1})\|$ in general do not equal (illustrated in Figure 2). Bounding the difference between these two quantities points the way forward for our analysis, which is also our main contribution in this section. We start with two lemmas comparing a geodesic triangle and the triangle formed by the preimage of its vertices in the tangent space, in two constant curvature spaces: hyperbolic space and the hypersphere.

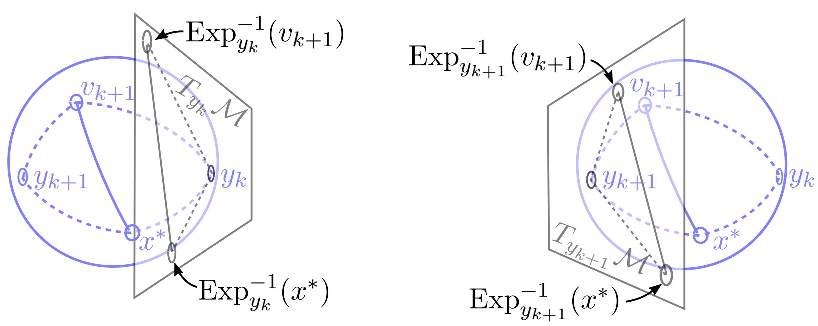

Figure 2: A schematic illustration of the geometric quantities in Theorem 10. Tangent spaces of y_k and y_{k+1} are shown in separate figures to reduce cluttering. Note that even on a sphere (which has constant positive sectional curvature), $d(x^*, v_{k+1})$, $\|\mathrm{Exp}_{y_k}^{-1}(x^*) - \mathrm{Exp}_{y_k}^{-1}(v_{k+1})\|$ and $\|\mathrm{Exp}_{y_{k+1}}^{-1}(x^*) - \mathrm{Exp}_{y_{k+1}}^{-1}(v_{k+1})\|$ generally do not equal.

Lemma 8 (bi-Lipschitzness of the exponential map in hyperbolic space) *Let a, b, c be the side lengths of a geodesic triangle in a hyperbolic space with constant sectional curvature -1, and A is the angle between sides b and c. Furthermore, assume $b \le \frac{1}{4}, c \ge 0$. Let $\triangle\bar{a}\bar{b}\bar{c}$ be the comparison triangle in Euclidean space, with $\bar{b} = b, \bar{c} = c, \bar{A} = A$, then*

$$\bar{a}^2 \le a^2 \le (1 + 2b^2)\bar{a}^2. \tag{13}$$

Proof The proof of this lemma contains technical details that deviate from our main focus; so we defer them to the appendix. The first inequality is well known. To show the second inequality, we have Lemma 13 and Lemma 14 (in Appendix) which in combination complete the proof. ■

We also state without proof that by the same techniques one can show the following result holds.

Lemma 9 (bi-Lipschitzness of the exponential map on hypersphere) *Let a, b, c be the side lengths of a geodesic triangle in a hypersphere with constant sectional curvature 1, and A is the angle between sides b and c. Furthermore, assume $b \leq \frac{1}{4}, c \in [0, \frac{\pi}{2}]$. Let $\triangle \bar{a}\bar{b}\bar{c}$ be the comparison triangle in Euclidean space, with $\bar{b} = b, \bar{c} = c, \bar{A} = A$, then*

$$a^2 \leq \bar{a}^2 \leq (1 + 2b^2)a^2. \tag{14}$$

Albeit very much simplified, spaces of constant curvature are important objects to study, because often their properties can be generalized to general Riemannian manifolds with bounded curvature, specifically via the use of powerful comparison theorems in metric geometry (Burago et al., 2001). In our case, we use these two lemmas to derive a tangent space distance comparison theorem for Riemannian manifolds with bounded sectional curvature.

Theorem 10 (Multiplicative distortion of squared distance on Riemannian manifold) *Let x^*, v_{k+1}, y_k, $y_{k+1} \in \mathcal{X}$ be four points in a g-convex, uniquely geodesic set $\mathcal{X}$ where the sectional curvature is bounded within $[-K, K]$, for some nonnegative number K. Define $b_{k+1} = \max\left\{\|\mathrm{Exp}_{y_k}^{-1}(x^*)\|, \|\mathrm{Exp}_{y_{k+1}}^{-1}(x^*)\|\right\}$. Assume $b_{k+1} \leq \frac{1}{4\sqrt{K}}$ for $K > 0$ (otherwise $b_{k+1} < \infty$), then we have*

$$\|\mathrm{Exp}_{y_{k+1}}^{-1}(x^*) - \mathrm{Exp}_{y_{k+1}}^{-1}(v_{k+1})\|^2 \leq (1 + 5Kb_{k+1}^2)\|\mathrm{Exp}_{y_k}^{-1}(x^*) - \mathrm{Exp}_{y_k}^{-1}(v_{k+1})\|^2. \tag{15}$$

Proof The high level idea is to think of the tangent space distance distortion on Riemannian manifolds of bounded curvature as a consequence of bi-Lipschitzness of the exponential map. Specifically, note that $\triangle y_k x^* v_{k+1}$ and $\triangle y_{k+1} x^* v_{k+1}$ are two geodesic triangles in $\mathcal{X}$, whereas $\|\mathrm{Exp}_{y_k}^{-1}(x^*) - \mathrm{Exp}_{y_k}^{-1}(v_{k+1})\|$ and $\|\mathrm{Exp}_{y_{k+1}}^{-1}(x^*) - \mathrm{Exp}_{y_{k+1}}^{-1}(v_{k+1})\|$ are side lengths of two comparison triangles in vector space. Since $\mathcal{X}$ is of bounded sectional curvature, we can apply comparison theorems.

First, we consider bound on the distortion of squared distance in a Riemannian manifold with constant curvature $-K$. Note that in this case, the hyperbolic law of cosines becomes

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

which corresponds to the geodesic triangle in hyperbolic space with side lengths $\sqrt{K}a$, $\sqrt{K}b$, $\sqrt{K}c$, with the corresponding comparison triangle in Euclidean space having lengths $\sqrt{K}\bar{a}$, $\sqrt{K}\bar{b}$, $\sqrt{K}\bar{c}$. Apply Lemma 8 we have $(\sqrt{K}a)^2 \leq (1+2(\sqrt{K}b)^2)(\sqrt{K}\bar{a})^2$, i.e. $a^2 \leq (1+2Kb^2)\bar{a}^2$. Now consider the geodesic triangle $\triangle y_k x^* v_{k+1}$. Let $\tilde{a} = \|\mathrm{Exp}_{v_{k+1}}^{-1}(x^*)\|, b = \|\mathrm{Exp}_{y_k}^{-1}(v_{k+1})\| \leq b_{k+1}, c = \|\mathrm{Exp}_{y_k}^{-1}(x^*)\|, A = \angle x^* y_k v_{k+1}$, so that $\|\mathrm{Exp}_{y_k}^{-1}(x^*) - \mathrm{Exp}_{y_k}^{-1}(v_{k+1})\|^2 = b^2 + c^2 - 2bc\cos(A)$. By Toponogov's comparison theorem (Burago et al., 2001), we have $\tilde{a} \leq a$ hence

$$\|\mathrm{Exp}_{v_{k+1}}^{-1}(x^*)\|^2 \leq \left(1 + 2Kb_{k+1}^2\right)\|\mathrm{Exp}_{y_k}^{-1}(x^*) - \mathrm{Exp}_{y_k}^{-1}(v_{k+1})\|^2. \tag{16}$$

Similarly, using the spherical law of cosines for a space of constant curvature K

$$\cos(\sqrt{K}a) = \cos(\sqrt{K}b)\cos(\sqrt{K}c) + \sin(\sqrt{K}b)\sin(\sqrt{K}c)\cos(A)$$

and Lemma 9 we can show $\bar{a}^2 \leq (1 + 2Kb^2)a^2$, where $\bar{a}$ is the side length in Euclidean space corresponding to a. Hence by our uniquely geodesic assumption and (Meyer, 1989, Theorem

2.2, Remark 7), with similar reasoning for the geodesic triangle $\triangle y_{k+1}x^*v_{k+1}$, we have $a \leq \|\mathrm{Exp}_{v_{k+1}}^{-1}(x^*)\|$, so that

$$\|\mathrm{Exp}_{y_{k+1}}^{-1}(x^*) - \mathrm{Exp}_{y_{k+1}}^{-1}(v_{k+1})\|^2 \leq \left(1 + 2Kb_{k+1}^2\right) a^2 \leq \left(1 + 2Kb_{k+1}^2\right) \|\mathrm{Exp}_{v_{k+1}}^{-1}(x^*)\|^2. \tag{17}$$

Finally, combining inequalities (16) and (17), and noting that $(1 + 2Kb_{k+1}^2)^2 = 1 + 4Kb_{k+1}^2 + (4Kb_{k+1}^2)Kb^2 \leq 1 + 5Kb_{k+1}^2$, the proof is complete. ∎

Theorem 10 suggests that if $b_{k+1} \leq \frac{1}{4\sqrt{K}}$, we could choose $\beta \geq 5Kb_{k+1}^2$ and $\gamma \leq \frac{1}{1+\beta}\overline{\gamma}$ to guarantee $\Phi_{k+1}(x^*) \leq \overline{\Phi}_{k+1}(x^*)$. It then follows that the analysis holds for k-th step. Still, it is unknown that under what conditions can we guarantee $\Phi_{k+1}(x^*) \leq \overline{\Phi}_{k+1}(x^*)$ hold for all $k \geq 0$, which would lead to a convergence proof. We resolve this question in the next theorem.

Theorem 11 (Local fast convergence) *With Assumptions 1, 2, 3, 4, denote $D = \frac{1}{20\sqrt{K}}\left(\frac{\mu}{L}\right)^{\frac{3}{4}}$ and assume $\mathcal{B}_{x^*,D} := \{x \in \mathcal{M} : d(x, x^*) \leq D\} \subseteq \mathcal{X}$. If we set $h = \frac{1}{L}, \beta = \frac{1}{5}\sqrt{\frac{\mu}{L}}$ and $x_0 \in \mathcal{B}_{x^*,D}$, then Algorithm 2 converges; moreover, we have*

$$f(x_k) - f(x^*) \leq \left(1 - \frac{9}{10}\sqrt{\frac{\mu}{L}}\right)^k \left(f(x_0) - f(x^*) + \frac{\mu}{2}\|\mathrm{Exp}_{x_0}^{-1}(x^*)\|^2\right). \tag{18}$$

Proof sketch. Recall that in Theorem 7 we already establish that if the tangent space distance comparison inequality (8) holds, then the general Riemannian Nesterov iteration (Algorithm 1) and hence its constant step size special case (Algorithm 2) converge with a guaranteed rate. By the tangent space distance comparison theorem (Theorem 10), the comparison inequality should hold if y_k and x^* are close enough. Indeed, we use induction to assert that with a good initialization, (8) holds for each step. Specifically, for every $k > 0$, if y_k is close to x^* and the comparison inequality holds until the $(k-1)$-th step, then y_{k+1} is also close to x^* and the comparison inequality holds until the k-th step. We postpone the complete proof until Appendix F. ∎

6. Discussion

In this work, we proposed a Riemannian generalization of the accelerated gradient algorithm and developed its convergence and complexity analysis. For the first time (to the best of our knowledge), we show gradient based algorithms on Riemannian manifolds can be accelerated, at least in a neighborhood of the minimizer. Central to our analysis are the two main technical contributions of our work: a new estimate sequence (Lemma 4), which relaxes the assumption of Nesterov's original construction and handles metric distortion on Riemannian manifolds; a tangent space distance comparison theorem (Theorem 10), which provides sufficient conditions for bounding the metric distortion and could be of interest for a broader range of problems on Riemannian manifolds.

Despite not matching the standard convex results, our result exposes the key difficulty of analyzing Nesterov-style algorithms on Riemannian manifolds, an aspect missing in previous work. Critically, the convergence analysis relies on bounding a new distortion term per each step. Furthermore, we observe that the side length sequence $d(y_k, v_{k+1})$ can grow much greater than $d(y_k, x^*)$, even if we reduce the "step size" h_k in Algorithm 1, defeating any attempt to control the distortion globally by modifying the algorithm parameters. This is a benign feature in vector space analysis, since (8) trivially holds nonetheless; however it poses a great difficulty for analysis in nonlinear space. Note the stark contrast to (stochastic) gradient descent, where the step length can be effectively controlled by reducing the step size, hence bounding the distortion terms globally (Zhang and Sra, 2016).

A topic of future interest is to study whether assumption (8) can be further relaxed, while maintaining that overall the algorithm still converges. By bounding the squared distance distortion in every step, our analysis provides guarantee for the worst-case scenario, which seems unlikely to happen in practice. It would be interesting to conduct experiments to see how often (8) is violated versus how often it is loose. It would also be interesting to construct some adversarial problem case (if any) and study the complexity lower bound of gradient based Riemannian optimization, to see if geodesically convex optimization is strictly more difficult than convex optimization. Generalizing the current analysis to non-strongly g-convex functions is another interesting direction.

Acknowledgments

The authors thank the anonymous reviewers for helpful feedback. This work was supported in part by NSF-IIS-1409802 and the DARPA Lagrange grant.

References

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

Naman Agarwal, Zeyuan Allen Zhu, Brian Bullins, Elad Hazan, and Tengyu Ma. Finding approximate local minima for nonconvex optimization in linear time. *CoRR*, abs/1611.01146, 2016.

Zeyuan Allen-Zhu and Lorenzo Orecchia. Linear coupling: An ultimate unification of gradient and mirror descent. *arXiv:1407.1537*, 2014.

Luigi Ambrosio, Nicola Gigli, Giuseppe Savaré, et al. Metric measure spaces with Riemannian Ricci curvature bounded from below. *Duke Mathematical Journal*, 163(7):1405–1490, 2014.

Yossi Arjevani, Shai Shalev-Shwartz, and Ohad Shamir. On lower and upper bounds for smooth and strongly convex optimization problems. *arXiv:1503.06833*, 2015.

Hedy Attouch, Jérôme Bolte, and Benar Fux Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized Gauss–Seidel methods. *Mathematical Programming*, 137(1-2):91–129, 2013.

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

Nicolas Boumal, P-A Absil, and Coralia Cartis. Global rates of convergence for nonconvex optimization on manifolds. *arXiv:1605.08101*, 2016a.

Nicolas Boumal, Vlad Voroninski, and Afonso Bandeira. The non-convex Burer-Monteiro approach works on smooth semidefinite programs. In *Advances in Neural Information Processing Systems*, pages 2757–2765, 2016b.

Sébastien Bubeck, Yin Tat Lee, and Mohit Singh. A geometric alternative to Nesterov's accelerated gradient descent. *arXiv:1506.08187*, 2015.

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

Yair Carmon, John C. Duchi, Oliver Hinder, and Aaron Sidford. Accelerated methods for non-convex optimization. *CoRR*, abs/1611.00756, 2016.

Yair Carmon, Oliver Hinder, John C Duchi, and Aaron Sidford. " convex until proven guilty": Dimension-free acceleration of gradient descent on non-convex functions. *arXiv preprint arXiv:1705.02766*, 2017.

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems*, pages 1646–1654, 2014.

OP Ferreira and PR Oliveira. Proximal point algorithm on Riemannian manifolds. *Optimization*, 51(2): 257–270, 2002.

Nicolas Flammarion and Francis Bach. From averaging to acceleration, there is only a step-size. In *Conference on Learning Theory*, pages 658–695, 2015.

Rong Ge, Chi Jin, and Yi Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. *arXiv:1704.00708*, 2017.

Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, pages 315–323, 2013.

Jürgen Jost. *Riemannian Geometry and Geometric Analysis*. Springer Science & Business Media, 2011.

Narendra Karmarkar. A new polynomial-time algorithm for linear programming. In *Proceedings of the sixteenth annual ACM symposium on Theory of computing*, pages 302–311. ACM, 1984.

Hiroyuki Kasai, Hiroyuki Sato, and Bamdev Mishra. Riemannian stochastic variance reduced gradient on Grassmann manifold. *arXiv:1605.07367*, 2016.

Kenji Kawaguchi. Deep learning without poor local minima. In *Advances in Neural Information Processing Systems*, pages 586–594, 2016.

Leonid G Khachiyan. Polynomial algorithms in linear programming. *USSR Computational Mathematics and Mathematical Physics*, 20(1):53–72, 1980.

Laurent Lessard, Benjamin Recht, and Andrew Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26(1):57–95, 2016.

Yuanyuan Liu, Fanhua Shang, James Cheng, Hong Cheng, and Licheng Jiao. Accelerated first-order methods for geodesically convex optimization on Riemannian manifolds. In *Advances in Neural Information Processing Systems*, pages 4875–4884, 2017.

Wolfgang Meyer. Toponogov's theorem and applications. *SMR*, 404:9, 1989.

Bamdev Mishra and Rodolphe Sepulchre. Riemannian preconditioning. *SIAM Journal on Optimization*, 26(1): 635–660, 2016.

Arkadiĭ Semenovich Nemirovsky and David Borisovich Yudin. *Problem complexity and method efficiency in optimization*. Wiley, 1983.

Yurii Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. In *Soviet Mathematics Doklady*, volume 27(2), pages 372–376, 1983.

Yurii Nesterov. *Introductory lectures on convex optimization*, volume 87. Springer Science & Business Media, 2004.

B.T. Polyak. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878, January 1963.

Sashank J. Reddi, Ahmed Hefny, Suvrit Sra, Barnabás Póczos, and Alexander J. Smola. Stochastic variance reduction for nonconvex optimization. In *Proceedings of the 33nd International Conference on Machine Learning, ICML*, pages 314–323, 2016.

Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *arXiv:1309.2388*, 2013.

Ohad Shamir. A Stochastic PCA and SVD Algorithm with an Exponential Convergence Rate. In *International Conference on Machine Learning (ICML-15)*, pages 144–152, 2015.

Weijie Su, Stephen Boyd, and Emmanuel Candes. A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights. In *Advances in Neural Information Processing Systems*, pages 2510–2518, 2014.

Ju Sun, Qing Qu, and John Wright. Complete dictionary recovery over the sphere I: Overview and the geometric picture. *IEEE Transactions on Information Theory*, 63(2):853–884, 2017.

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

Andre Wibisono, Ashia C Wilson, and Michael I Jordan. A variational perspective on accelerated methods in optimization. *Proceedings of the National Academy of Sciences*, page 201614734, 2016.

Hongyi Zhang and Suvrit Sra. First-order methods for geodesically convex optimization. In *29th Annual Conference on Learning Theory (COLT)*, pages 1617–1638, 2016.

Hongyi Zhang, Sashank J. Reddi, and Suvrit Sra. Riemannian SVRG: Fast stochastic optimization on Riemannian manifolds. In *Advances in Neural Information Processing Systems 29*, 2016.

Appendix A. Constant step scheme

Lemma 12 *Pick $\beta_k \equiv \beta > 0$. If in Algorithm 1 we set*

$$h_k \equiv h, \forall k \geq 0, \qquad \gamma_0 \equiv \gamma = \frac{\sqrt{\beta^2 + 4(1+\beta)\mu h} - \beta}{\sqrt{\beta^2 + 4(1+\beta)\mu h} + \beta} \cdot \mu,$$

then we have

$$\alpha_k \equiv \alpha = \frac{\sqrt{\beta^2 + 4(1+\beta)\mu h} - \beta}{2}, \qquad \overline{\gamma}_{k+1} \equiv (1+\beta)\gamma, \qquad \gamma_{k+1} \equiv \gamma, \qquad \forall k \geq 0. \quad (19)$$

Proof Suppose that $\gamma_k = \gamma$, then from Algorithm 1 we have α_k is the positive root of

$$\alpha_k^2 - (\mu - \gamma)h\alpha_k - \gamma h = 0.$$

Also note

$$\mu - \gamma = \frac{\beta\alpha}{(1+\beta)h}, \quad \text{and} \quad \gamma = \frac{\alpha^2}{(1+\beta)h}, \tag{20}$$

hence

$$\begin{aligned}
\alpha_k &= \frac{(\mu - \gamma)h + \sqrt{(\mu-\gamma)^2 h^2 + 4\gamma h}}{2} \\
&= \frac{\beta\alpha}{2(1+\beta)} + \frac{1}{2}\sqrt{\frac{\beta^2\alpha^2}{(1+\beta)^2} + \frac{4\alpha^2}{1+\beta}} \\
&= \alpha
\end{aligned}$$

Furthermore, we have

$$\begin{aligned}
\overline{\gamma}_{k+1} &= (1-\alpha_k)\gamma_k + \alpha_k\mu = (1-\alpha)\gamma + \alpha\mu \\
&= \gamma + (\mu - \gamma)\alpha = \gamma + \beta\frac{\alpha^2}{(1+\beta)h} \\
&= (1+\beta)\gamma
\end{aligned}$$

and $\gamma_{k+1} = \frac{1}{1+\beta}\overline{\gamma}_{k+1} = \gamma$. Since $\gamma_k = \gamma$ holds for $k = 0$, by induction the proof is complete. ∎

Appendix B. Proof of Lemma 2

Proof The proof is similar to (Nesterov, 2004, Lemma 2.2.2) except that we introduce $\overline{\Phi}_{k+1}$ as an intermediate step in constructing $\Phi_{k+1}(x)$. In fact, to start we have $\Phi_0(x) \leq (1-\lambda_0)f(x) + \lambda_0\Phi_0(x) \equiv \Phi_0(x)$. Moreover, assume (1) holds for some $k \geq 0$, i.e. $\Phi_k(x^*) - f(x^*) \leq \lambda_k(\Phi_0(x^*) - f(x^*))$, then

$$\begin{aligned}
\Phi_{k+1}(x^*) - f(x^*) &\leq \overline{\Phi}_{k+1}(x^*) - f(x^*) \\
&\leq (1-\alpha_k)\Phi_k(x^*) + \alpha_k f(x^*) - f(x^*) \\
&= (1-\alpha_k)(\Phi_k(x^*) - f(x^*)) \\
&\leq (1-\alpha_k)\lambda_k(\Phi_0(x^*) - f(x^*)) \\
&= \lambda_{k+1}(\Phi_0(x^*) - f(x^*)),
\end{aligned}$$

where the first inequality is due to our construction of $\Phi_{k+1}(x)$ in (4), the second inequality due to strong convexity of f. By induction we have $\Phi_k(x^*) \le (1-\lambda_k)f(x^*) + \lambda_k \Phi_0(x^*)$ for all $k \ge 0$. It remains to note that condition 4 ensures $\lambda_k \to 0$. ∎

Appendix C. Proof of Lemma 5

Proof We prove this lemma by completing the square:

$$\begin{aligned}
\overline{\Phi}_{k+1}(x) =\ & (1-\alpha_k)\left(\Phi_k^* + \frac{\gamma_k}{2}\|\mathrm{Exp}_{y_k}^{-1}(x) - \mathrm{Exp}_{y_k}^{-1}(v_k)\|^2\right) \\
& + \alpha_k \left(f(y_k) + \langle \mathrm{grad} f(y_k), \mathrm{Exp}_{y_k}^{-1}(x)\rangle + \frac{\mu}{2}\|\mathrm{Exp}_{y_k}^{-1}(x)\|^2\right) \\
=\ & \frac{\overline{\gamma}_{k+1}}{2}\|\mathrm{Exp}_{y_k}^{-1}(x)\|^2 + \left\langle \alpha_k \mathrm{grad} f(y_k) - (1-\alpha_k)\,\gamma_k \mathrm{Exp}_{y_k}^{-1}(v_k), \mathrm{Exp}_{y_k}^{-1}(x)\right\rangle \\
& + (1-\alpha_k)\left(\Phi_k^* + \frac{\gamma_k}{2}\|\mathrm{Exp}_{y_k}^{-1}(v_k)\|^2\right) + \alpha_k f(y_k) \\
=\ & \frac{\overline{\gamma}_{k+1}}{2}\left\|\mathrm{Exp}_{y_k}^{-1}(x) - \left(\frac{(1-\alpha_k)\gamma_k}{\overline{\gamma}_{k+1}}\mathrm{Exp}_{y_k}^{-1}(v_k) - \frac{\alpha_k}{\overline{\gamma}_{k+1}}\mathrm{grad} f(y_k)\right)\right\|^2 + \Phi_{k+1}^* \\
=\ & \Phi_{k+1}^* + \frac{\overline{\gamma}_{k+1}}{2}\left\|\mathrm{Exp}_{y_k}^{-1}(x) - \mathrm{Exp}_{y_k}^{-1}(v_{k+1})\right\|^2
\end{aligned}$$

where the third equality is by completing the square with respect to $\mathrm{Exp}_{y_k}^{-1}(x)$ and use the definition of Φ_{k+1}^* in (7), the last equality is by the definition of y_k in Algorithm 1, and $\overline{\Phi}_{k+1}(x)$ is minimized if and only if $x = \mathrm{Exp}_{y_k}\left(\frac{(1-\alpha_k)\gamma_k}{\overline{\gamma}_{k+1}}\mathrm{Exp}_{y_k}^{-1}(v_k) - \frac{\alpha_k}{\overline{\gamma}_{k+1}}\mathrm{grad} f(y_k)\right) = v_{k+1}$. ∎

Appendix D. Proof of Lemma 6

Proof For $k=0$, $\Phi_k^* \ge f(x_k)$ trivially holds. Assume for iteration k we have $\Phi_k^* \ge f(x_k)$, then from definition (7) we have

$$\begin{aligned}
\Phi_{k+1}^* \ge\ & (1-\alpha_k)\, f(x_k) + \alpha_k f(y_k) - \frac{\alpha_k^2}{2\overline{\gamma}_{k+1}}\|\mathrm{grad} f(y_k)\|^2 + \frac{\alpha_k(1-\alpha_k)\gamma_k}{\overline{\gamma}_{k+1}}\langle \mathrm{grad} f(y_k), \mathrm{Exp}_{y_k}^{-1}(v_k)\rangle \\
\ge\ & f(y_k) - \frac{\alpha_k^2}{2\overline{\gamma}_{k+1}}\|\mathrm{grad} f(y_k)\|^2 + (1-\alpha_k)\left\langle \mathrm{grad} f(y_k), \frac{\alpha_k\gamma_k}{\overline{\gamma}_{k+1}}\mathrm{Exp}_{y_k}^{-1}(v_k) + \mathrm{Exp}_{y_k}^{-1}(x_k)\right\rangle \\
=\ & f(y_k) - \frac{\alpha_k^2}{2\overline{\gamma}_{k+1}}\|\mathrm{grad} f(y_k)\|^2 \\
=\ & f(y_k) - \frac{h_k}{2}\|\mathrm{grad} f(y_k)\|^2,
\end{aligned}$$

where the first inequality is due to $\Phi_k^* \geq f(x_k)$, the second due to $f(x_k) \geq f(y_k) + \langle \mathrm{grad} f(y_k), \mathrm{Exp}_{y_k}^{-1}(x_k) \rangle$ by g-convexity, and the equalities follow from Algorithm 1. On the other hand, we have the bound

$$
\begin{aligned}
f(x_{k+1}) &\leq f(y_k) + \langle \mathrm{grad} f(y_k), \mathrm{Exp}_{y_k}^{-1}(x_{k+1}) \rangle + \frac{L}{2} \| \mathrm{Exp}_{y_k}^{-1}(x_{k+1}) \|^2 \\
&= f(y_k) - h_k \left(1 - \frac{L h_k}{2} \right) \| \mathrm{grad} f(y_k) \|^2 \\
&\leq f(y_k) - \frac{h_k}{2} \| \mathrm{grad} f(y_k) \|^2 \leq \Phi_{k+1}^*,
\end{aligned}
$$

where the first inequality is by the L-smoothness assumption, the equality from the definition of x_{k+1} in Algorithm 1 Line 2, and the second inequality from the assumption that $h_k \leq \frac{1}{L}$. Hence by induction, $\Phi_k^* \geq f(x_k)$ for all $k \geq 0$. ■

Appendix E. Proof of Lemma 8

Lemma 13 *Let a, b, c be the side lengths of a geodesic triangle in a hyperbolic space with constant sectional curvature -1, and A is the angle between sides b and c. Furthermore, assume $b \leq \frac{1}{4}, c \geq \frac{1}{2}$. Let $\triangle \bar{a}\bar{b}\bar{c}$ be the comparison triangle in Euclidean space, with $\bar{b} = b, \bar{c} = c, \bar{A} = A$, then*

$$a^2 \leq (1 + 2b^2)\bar{a}^2 \tag{21}$$

Proof We first apply (Zhang and Sra, 2016, Lemma 5) with $\kappa = -1$ to get

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

We also have

$$\bar{a}^2 = b^2 + c^2 - 2bc\cos(A).$$

Hence we get

$$a^2 - \bar{a}^2 \leq \left(\frac{c}{\tanh(c)} - 1 \right) b^2.$$

It remains to note that for $b \leq \frac{1}{4}, c \geq \frac{1}{2}$,

$$2a^2 \geq 2(c-b)^2 \geq 2\left(c - \frac{1}{4}\right) \geq \frac{c}{\tanh(1/2)} - 1 \geq \frac{c}{\tanh(c)} - 1,$$

which implies $a^2 \leq (1 + 2b^2)\bar{a}^2$. ■

Lemma 14 *Let a, b, c be the side lengths of a geodesic triangle in a hyperbolic space with constant sectional curvature -1, and A is the angle between sides b and c. Furthermore, assume $b \leq \frac{1}{4}, c \leq \frac{1}{2}$. Let $\triangle \bar{a}\bar{b}\bar{c}$ be the comparison triangle in Euclidean space, with $\bar{b} = b, \bar{c} = c, \bar{A} = A$, then*

$$a^2 \leq (1 + b^2)\bar{a}^2 \tag{22}$$

Proof Recall the law of cosines in Euclidean space and hyperbolic space:

$$\bar{a}^2 = \bar{b}^2 + \bar{c}^2 - 2\bar{b}\bar{c}\cos\bar{A}, \tag{23}$$

$$\cosh a = \cosh b \cosh c - \sinh b \sinh c \cos A, \tag{24}$$

and the Taylor series expansion:

$$\cosh x = \sum_{n=0}^{\infty} \frac{1}{(2n)!} x^{2n}, \qquad \sinh x = \sum_{n=0}^{\infty} \frac{1}{(2n+1)!} x^{2n+1}. \tag{25}$$

We let $\bar{b} = b, \bar{c} = c, \bar{A} = A$, from Eq. (23) we have

$$\cosh \bar{a} = \cosh\left(\sqrt{b^2 + c^2 - 2bc\cos A}\right) \tag{26}$$

It is widely known that $\bar{a} \leq a$. Now we use Eq. (25) to expand the RHS of Eq. (24) and Eq. (26), and compare the coefficients for each corresponding term $b^i c^j$ in the two series. Without loss of generality, we assume $i \geq j$; the results for condition $i < j$ can be easily obtained by the symmetry of b, c. We expand Eq. (24) as

$$\begin{aligned}\cosh a = & \left(\sum_{n=0}^{\infty} \frac{1}{(2n)!} b^{2n}\right)\left(\sum_{n=0}^{\infty} \frac{1}{(2n)!} c^{2n}\right) \\ & - \left(\sum_{n=0}^{\infty} \frac{1}{(2n+1)!} b^{2n+1}\right)\left(\sum_{n=0}^{\infty} \frac{1}{(2n+1)!} c^{2n+1}\right)\cos A\end{aligned}$$

where the coefficient $\alpha(i, j)$ of $b^i c^j$ is

$$\alpha(i,j) = \begin{cases} \frac{1}{(2p)!(2q)!}, & \text{if } p, q \in \mathbb{N} \text{ and } i = 2p, j = 2q, \\ \frac{\cos A}{(2p+1)!(2q+1)!}, & \text{if } p, q \in \mathbb{N} \text{ and } i = 2p+1, j = 2q+1, \\ 0, & \text{otherwise.} \end{cases} \tag{27}$$

Similarly, we expand Eq. (26) as

$$\cosh \bar{a} = \sum_{n=0}^{\infty} \frac{1}{(2n)!} \left(b^2 + c^2 - 2bc\cos A\right)^n$$

where the coefficient $\bar{\alpha}(i, j)$ of $b^i c^j$ is

$$\bar{\alpha}(i,j) = \begin{cases} \frac{\sum_{k=0}^{q} \binom{p+q}{p-k,q-k,2k} (2\cos A)^{2k}}{(2p+2q)!}, & \text{if } p, q \in \mathbb{N} \text{ and } i = 2p, j = 2q, \\ \frac{\sum_{k=0}^{q} \binom{p+q+1}{p-k,q-k,2k+1} (2\cos A)^{2k+1}}{(2p+2q+2)!}, & \text{if } p, q \in \mathbb{N} \text{ and } i = 2p+1, j = 2q+1, \\ 0, & \text{otherwise.} \end{cases} \tag{28}$$

We hence calculate their absolute difference

$$\begin{aligned}
&|\alpha(i,j)-\bar{\alpha}(i,j)| \\
&= \begin{cases} \frac{\sum_{k=0}^{q}\binom{p+q}{p-k,q-k,2k}2^{2k}\left(1-(\cos A)^{2k}\right)}{(2p+2q)!}, & \text{if } p,q\in\mathbb{N} \text{ and } i=2p, j=2q, \\ \frac{\sum_{k=0}^{q}\binom{p+q+1}{p-k,q-k,2k+1}2^{2k+1}\left(1-(\cos A)^{2k}\right)|\cos A|}{(2p+2q+2)!}, & \text{if } p,q\in\mathbb{N} \text{ and } i=2p+1, j=2q+1, \\ 0, & \text{otherwise.} \end{cases} \\
&\le \begin{cases} \frac{\sum_{k=0}^{q}\binom{p+q}{p-k,q-k,2k}2^{2k}k}{(2p+2q)!}\sin^2 A, & \text{if } p,q\in\mathbb{N} \text{ and } i=2p, j=2q, \\ \frac{\sum_{k=0}^{q}\binom{p+q+1}{p-k,q-k,2k+1}2^{2k+1}k}{(2p+2q+2)!}\sin^2 A, & \text{if } p,q\in\mathbb{N} \text{ and } i=2p+1, j=2q+1, \\ 0, & \text{otherwise.} \end{cases} \\
&\le \begin{cases} \frac{q\sum_{k=0}^{q}\binom{p+q}{p-k,q-k,2k}2^{2k}}{(2p+2q)!}\sin^2 A, & \text{if } p,q\in\mathbb{N} \text{ and } i=2p, j=2q, \\ \frac{q\sum_{k=0}^{q}\binom{p+q+1}{p-k,q-k,2k+1}2^{2k+1}}{(2p+2q+2)!}\sin^2 A, & \text{if } p,q\in\mathbb{N} \text{ and } i=2p+1, j=2q+1, \\ 0, & \text{otherwise.} \end{cases} \\
&= \begin{cases} \frac{q}{(2p)!(2q)!}\sin^2 A, & \text{if } p,q\in\mathbb{N} \text{ and } i=2p, j=2q, \\ \frac{q}{(2p+1)!(2q+1)!}\sin^2 A, & \text{if } p,q\in\mathbb{N} \text{ and } i=2p+1, j=2q+1, \\ 0, & \text{otherwise.} \end{cases} \qquad (29)
\end{aligned}$$

where the two equalities are due to Lemma 15, the first inequality due to the following fact

$$\begin{aligned}
1-(\cos A)^{2m} &= \left(1-(\cos A)^2\right)\left(1+(\cos A)^2+(\cos A)^4+\cdots+(\cos A)^{2(m-1)}\right) \\
&= \sin^2 A\left(1+(\cos A)^2+(\cos A)^4+\cdots+(\cos A)^{2(m-1)}\right)\le m\sin^2 A
\end{aligned}$$

By setting $q=0$, we see that in the Taylor series of $\cosh a-\cosh\bar{a}$, any term that does not include a factor of c^2 cancels out. By the symmetry of b,c, any term that does not include a factor of b^2 also cancels out. The term with the lowest order of power is thus $\frac{1}{4}b^2c^2\sin^2 A$. Since we have $c\le\frac{1}{2}, b\le\frac{1}{4}$, the terms $|\alpha(i,j)-\bar{\alpha}(i,j)|b^ic^j$ must satisfy

$$\begin{aligned}
\sum_{i,j}|\alpha(i,j)-\bar{\alpha}(i,j)|b^ic^j &\le \left(\frac{1}{4}+\sum_{\substack{i+j=2k,\\ i,j\ge 2,k\ge 3}}\frac{i+j}{2(i!)(j!)}\frac{1}{2^{2k-4}}\right)b^2c^2\sin^2 A \\
&\le \left(\frac{1}{4}+\sum_{k\ge 3}\frac{1}{2^{2k-3}}\right)b^2c^2\sin^2 A\le\frac{1}{2}b^2c^2\sin^2 A \\
&= \frac{1}{2}b^2\bar{a}^2\sin^2 C\le\frac{1}{2}\bar{a}^2b^2
\end{aligned}$$

where the first inequality follows from Eq. (29) and is due to $\min(p,q)\le\frac{i+j}{2}$, the second inequality is due to $\sum_{\substack{i+j=2k\\ i\ge 2,j\ge 2}}\frac{i+j}{(i!)(j!)}\le\frac{(2k)^2}{(k!)^2}\le 1$ for $k\ge 3$ and the last equality is due to Euclidean law of sines. We thus get

$$\cosh a-\cosh\bar{a}\le\sum_{i,j}|\alpha(i,j)-\bar{\alpha}(i,j)|b^ic^j\sin^2 A\le\frac{1}{2}b^2\bar{a}^2 \qquad (30)$$

On the other hand, from the Taylor series of $\cosh$ we have

$$\cosh a - \cosh \bar{a} = \sum_{n=0}^{\infty} \frac{a^{2n} - \bar{a}^{2n}}{(2n)!} \geq \frac{1}{2}(a^2 - \bar{a}^2),$$

hence $a^2 \leq (1 + b^2)\bar{a}^2$.

■

Lemma 15 (Two multinomial identities) *For $p, q \in \mathbb{N}, p \geq q$, we have*

$$\frac{(2p+2q)!}{(2p)!(2q)!} = \sum_{k=0}^{q} \binom{p+q}{p-k, q-k, 2k} 2^{2k} \tag{31}$$

$$\frac{(2p+2q+2)!}{(2p+1)!(2q+1)!} = \sum_{k=0}^{q} \binom{p+q+1}{p-k, q-k, 2k+1} 2^{2k+1} \tag{32}$$

Proof We prove the identities by showing that the LHS and RHS correspond to two equivalent ways of counting the same quantity. For the first identity, consider a set of $2p+2q$ balls b_i each with a unique index $i = 1, \ldots, 2p+2q$, we count how many ways we can put them into boxes B_1 and B_2, such that B_1 has $2p$ balls and B_2 has $2q$ balls. The LHS is obviously a correct count. To get the RHS, note that we can first put balls in pairs, then decide what to do with each pair. Specifically, there are $p+q$ pairs $\{b_{2i-1}, b_{2i}\}$, and we can partition the counts by the number of pairs of which we put one of the two balls in B_2. Note that this number must be even. If there are $2k$ such pairs, which gives us $2k$ balls in B_2, we still need to choose $2(q-k)$ pairs of which both balls are put in B_2, and the left are $p-k$ pairs of which both balls are put in B_1. The total number of counts given k is thus

$$\binom{p+q}{p-k, q-k, 2k} 2^{2k}$$

because we can choose either ball in each of the $2k$ pairs leading to 2^{2k} possible choices. Summing over k we get the RHS. Hence the LHS and the RHS equal. The second identity can be proved with essentially the same argument. ■

Appendix F. Proof of Theorem 11

Proof *The base case.* First we verify that y_0, y_1 is sufficiently close to x^* so that the comparison inequality (8) holds at step $k = 0$. In fact, since $y_0 = x_0$ by construction, we have

$$\|\mathrm{Exp}_{y_0}^{-1}(x^*)\| = \|\mathrm{Exp}_{x_0}^{-1}(x^*)\| \leq \frac{1}{4\sqrt{K}}, \qquad 5K\|\mathrm{Exp}_{y_0}^{-1}(x^*)\|^2 \leq \frac{1}{80}\left(\frac{\mu}{L}\right)^{\frac{3}{2}} \leq \beta. \tag{33}$$

To bound $\|\mathrm{Exp}_{y_1}^{-1}(x^*)\|$, observe that y_1 is on the geodesic between x_1 and v_1. So first we bound $\|\mathrm{Exp}_{x_1}^{-1}(x^*)\|$ and $\|\mathrm{Exp}_{v_1}^{-1}(x^*)\|$. Bound on $\|\mathrm{Exp}_{x_1}^{-1}(x^*)\|$ comes from strong g-convexity:

$$\begin{aligned}\|\mathrm{Exp}_{x_1}^{-1}(x^*)\|^2 &\leq \frac{2}{\mu}(f(x_1) - f(x^*)) \leq \frac{2}{\mu}(f(x_0) - f(x^*)) + \frac{\gamma}{\mu}\|\mathrm{Exp}_{x_0}^{-1}(x^*)\|^2 \\ &\leq \frac{L+\gamma}{\mu}\|\mathrm{Exp}_{x_0}^{-1}(x^*)\|^2,\end{aligned}$$

whereas bound on $\|\mathrm{Exp}_{v_1}^{-1}(x^*)\|$ utilizes the tangent space distance comparison theorem. First, from the definition of $\overline{\Phi}_1$ we have

$$\|\mathrm{Exp}_{y_0}^{-1}(x^*) - \mathrm{Exp}_{y_0}^{-1}(v_1)\|^2 = \frac{2}{\gamma}(\overline{\Phi}_1(x^*) - \Phi_1^*) \leq \frac{2}{\gamma}(\Phi_0(x^*) - f(x^*)) \leq \frac{L+\gamma}{\gamma}\|\mathrm{Exp}_{x_0}^{-1}(x^*)\|^2$$

Then note that (33) implies that the assumption in Theorem 10 is satisfied when $k = 0$, thus we have

$$\|\mathrm{Exp}_{v_1}^{-1}(x^*)\|^2 \leq (1+\beta)\|\mathrm{Exp}_{y_0}^{-1}(x^*) - \mathrm{Exp}_{y_0}^{-1}(v_1)\|^2 \leq \frac{2(L+\gamma)}{\gamma}\|\mathrm{Exp}_{x_0}^{-1}(x^*)\|^2.$$

Together we have

$$\begin{aligned}
\|\mathrm{Exp}_{y_1}^{-1}(x^*)\| &\leq \|\mathrm{Exp}_{x_1}^{-1}(x^*)\| + \frac{\alpha\gamma}{\gamma+\alpha\mu}\|\mathrm{Exp}_{x_1}^{-1}(v_1)\| \\
&\leq \|\mathrm{Exp}_{x_1}^{-1}(x^*)\| + \frac{\alpha\gamma}{\gamma+\alpha\mu}\left(\|\mathrm{Exp}_{x_1}^{-1}(x^*)\| + \|\mathrm{Exp}_{v_1}^{-1}(x^*)\|\right) \\
&\leq \sqrt{\frac{L+\gamma}{\mu}}\|\mathrm{Exp}_{x_0}^{-1}(x^*)\| + \frac{\alpha\gamma}{\gamma+\alpha\mu}\left(\sqrt{\frac{L+\gamma}{\mu}} + \sqrt{\frac{2(L+\gamma)}{\mu}}\right)\|\mathrm{Exp}_{x_0}^{-1}(x^*)\| \\
&\leq \left(1 + \frac{1+\sqrt{2}}{2}\right)\sqrt{\frac{L+\gamma}{\mu}}\|\mathrm{Exp}_{x_0}^{-1}(x^*)\| \\
&\leq \frac{1}{10\sqrt{K}}\left(\frac{\mu}{L}\right)^{\frac{1}{4}} \leq \frac{1}{4\sqrt{K}} \qquad (34)
\end{aligned}$$

which also implies

$$5K\|\mathrm{Exp}_{y_1}^{-1}(x^*)\|^2 \leq \frac{1}{20}\sqrt{\frac{\mu}{L}} \leq \beta \qquad (35)$$

By (34), (35) and Theorem 10 it is hence guaranteed that

$$\gamma\|\mathrm{Exp}_{y_1}^{-1}(x^*) - \mathrm{Exp}_{y_1}^{-1}(v_1)\|^2 \leq \overline{\gamma}\|\mathrm{Exp}_{y_0}^{-1}(x^*) - \mathrm{Exp}_{y_0}^{-1}(v_1)\|^2.$$

The inductive step. Assume that for $i = 0, \dots, k-1$, (8) hold simultaneously, i.e.:

$$\gamma\|\mathrm{Exp}_{y_{i+1}}^{-1}(x^*) - \mathrm{Exp}_{y_{i+1}}^{-1}(v_{i+1})\|^2 \leq \overline{\gamma}\|\mathrm{Exp}_{y_i}^{-1}(x^*) - \mathrm{Exp}_{y_i}^{-1}(v_{i+1})\|^2, \forall i = 0, \dots, k-1$$

and also that $\|\mathrm{Exp}_{y_k}^{-1}(x^*)\| \leq \frac{1}{10\sqrt{K}}\left(\frac{\mu}{L}\right)^{\frac{1}{4}}$. To bound $\|\mathrm{Exp}_{y_{k+1}}^{-1}(x^*)\|$, observe that y_{k+1} is on the geodesic between x_{k+1} and v_{k+1}. So first we bound $\|\mathrm{Exp}_{x_{k+1}}^{-1}(x^*)\|$ and $\|\mathrm{Exp}_{v_{k+1}}^{-1}(x^*)\|$. Note that due to the sequential nature of the algorithm, statements about any step only depend on its previous steps, but not any step afterwards. Since (8) hold for steps $i = 0, \dots, k-1$, the analysis in the previous section already applies for steps $i = 0, \dots, k-1$. Therefore by Theorem 7 and the proof of Lemma 6 we know

$$\begin{aligned}
f(x^*) &\leq f(x_{k+1}) \leq \Phi_{k+1}^* \leq \Phi_{k+1}(x^*) \leq f(x^*) + (1-\alpha)^{k+1}(\Phi_0(x^*) - f(x^*)) \\
&\leq \Phi_0(x^*) = f(x_0) + \frac{\gamma}{2}\|\mathrm{Exp}_{x_0}^{-1}(x^*)\|^2
\end{aligned}$$

Hence we get $f(x_{k+1}) - f(x^*) \leq \Phi_0(x^*) - f(x^*)$ and $\frac{\gamma}{2}\|\mathrm{Exp}_{y_k}^{-1}(x^*) - \mathrm{Exp}_{y_k}^{-1}(v_{k+1})\|^2 \equiv \overline{\Phi}_{k+1}(x^*) - \Phi_{k+1}^* \leq \Phi_0(x^*) - f(x^*)$. Bound on $\|\mathrm{Exp}_{x_{k+1}}^{-1}(x^*)\|$ comes from strong g-convexity:

$$\begin{aligned}\|\mathrm{Exp}_{x_{k+1}}^{-1}(x^*)\|^2 &\leq \frac{2}{\mu}(f(x_{k+1}) - f(x^*)) \leq \frac{2}{\mu}(f(x_0) - f(x^*)) + \frac{\gamma}{\mu}\|\mathrm{Exp}_{x_0}^{-1}(x^*)\|^2 \\ &\leq \frac{L+\gamma}{\mu}\|\mathrm{Exp}_{x_0}^{-1}(x^*)\|^2,\end{aligned}$$

whereas bound on $\|\mathrm{Exp}_{v_{k+1}}^{-1}(x^*)\|$ utilizes the tangent space distance comparison theorem. First, from the definition of $\overline{\Phi}_{k+1}$ we have

$$\|\mathrm{Exp}_{y_k}^{-1}(x^*) - \mathrm{Exp}_{y_k}^{-1}(v_{k+1})\|^2 = \frac{2}{\gamma}(\overline{\Phi}_{k+1}(x^*) - \Phi_{k+1}^*) \leq \frac{2}{\gamma}(\Phi_0(x^*) - f(x^*)) \leq \frac{L+\gamma}{\gamma}\|\mathrm{Exp}_{x_0}^{-1}(x^*)\|^2$$

Then note that the inductive hypothesis implies that

$$\|\mathrm{Exp}_{v_{k+1}}^{-1}(x^*)\|^2 \leq (1+\beta)\|\mathrm{Exp}_{y_k}^{-1}(x^*) - \mathrm{Exp}_{y_k}^{-1}(v_{k+1})\|^2 \leq \frac{2(L+\gamma)}{\gamma}\|\mathrm{Exp}_{x_0}^{-1}(x^*)\|^2$$

Together we have

$$\begin{aligned}\|\mathrm{Exp}_{y_{k+1}}^{-1}(x^*)\| \leq\; & \|\mathrm{Exp}_{x_{k+1}}^{-1}(x^*)\| + \frac{\alpha\gamma}{\gamma+\alpha\mu}\|\mathrm{Exp}_{x_{k+1}}^{-1}(v_{k+1})\| \\ \leq\; & \|\mathrm{Exp}_{x_{k+1}}^{-1}(x^*)\| + \frac{\alpha\gamma}{\gamma+\alpha\mu}\left(\|\mathrm{Exp}_{x_{k+1}}^{-1}(x^*)\| + \|\mathrm{Exp}_{v_{k+1}}^{-1}(x^*)\|\right) \\ \leq\; & \sqrt{\frac{L+\gamma}{\mu}}\|\mathrm{Exp}_{x_0}^{-1}(x^*)\| + \frac{\alpha\gamma}{\gamma+\alpha\mu}\left(\sqrt{\frac{L+\gamma}{\mu}} + \sqrt{\frac{2(L+\gamma)}{\mu}}\right)\|\mathrm{Exp}_{x_0}^{-1}(x^*)\| \\ \leq\; & \left(1 + \frac{1+\sqrt{2}}{2}\right)\sqrt{\frac{L+\gamma}{\mu}}\|\mathrm{Exp}_{x_0}^{-1}(x^*)\| \\ \leq\; & \frac{1}{10\sqrt{K}}\left(\frac{\mu}{L}\right)^{\frac{1}{4}} \leq \frac{1}{4\sqrt{K}}\end{aligned}$$

which also implies that

$$5K\|\mathrm{Exp}_{y_{k+1}}^{-1}(x^*)\|^2 \leq \frac{1}{20}\sqrt{\frac{\mu}{L}} \leq \beta$$

By the two lines of equations above and Theorem 10 it is guaranteed that $\|\mathrm{Exp}_{y_{k+1}}^{-1}(x^*)\| \leq \frac{1}{10\sqrt{K}}\left(\frac{\mu}{L}\right)^{\frac{1}{4}}$ and also

$$\gamma\|\mathrm{Exp}_{y_{k+1}}^{-1}(x^*) - \mathrm{Exp}_{y_{k+1}}^{-1}(v_{k+1})\|^2 \leq \overline{\gamma}\|\mathrm{Exp}_{y_k}^{-1}(x^*) - \mathrm{Exp}_{y_k}^{-1}(v_{k+1})\|^2.$$

i.e. (8) hold for $i = 0, \ldots, k$. This concludes the inductive step.
By induction, (8) hold for all $k \geq 0$, hence by Theorem 7, Algorithm 2 converges, with

$$\alpha_i \equiv \alpha = \frac{\sqrt{\beta^2 + 4(1+\beta)\mu h} - \beta}{2} = \frac{\sqrt{\mu h}}{2}\left(\sqrt{\frac{1}{25} + 4\left(1 + \frac{\sqrt{\mu h}}{5}\right)} - \frac{1}{5}\right) \geq \frac{9}{10}\sqrt{\frac{\mu}{L}}.$$

■