A Riemannian Accelerated Proximal Extragradient Framework and its Implications

Jikai Jin
School of Mathematical Sciences, Peking University, Beijing, China
jkjin@pku.edu.cn

Suvrit Sra
LIDS, Massachusetts Institute of Technology, Cambridge, MA, USA
suvrit@mit.edu

Abstract

The study of accelerated gradient methods in Riemannian optimization has recently witnessed notable progress. However, in contrast with the Euclidean setting, a systematic understanding of acceleration is still lacking in the Riemannian setting. We revisit the Accelerated Hybrid Proximal Extragradient (A-HPE) method of Monteiro and Svaiter [29], a powerful framework for obtaining accelerated Euclidean methods. Subsequently, we propose a Riemannian version of A-HPE. The basis of our analysis of Riemannian A-HPE is a set of insights into Euclidean A-HPE, which we combine with a careful control of distortion caused by Riemannian geometry. We describe a number of Riemannian accelerated gradient methods as concrete instances of our framework.

1 Introduction

The theory of convex optimization is well-studied and efficient algorithms have been developed for different function classes [30]. Convexity plays a crucial role in the polynomial-time complexity of these algorithms for finding approximate minimizers, which is generally intractable for non-convex problems. Notwithstanding the fruitful theoretical results for convex optimization, the requirement of convexity is quite restrictive and does not hold in many real-world problems.

The notion of convexity in Euclidean space has a vast generalization to certain metric spaces, specifically in the guise of geodesic convexity. Some non-convex problems in Euclidean space can be transformed into geodesically convex problems in some non-Euclidean space [21, 36], which reveals their potential tractability.

This promise of geodesic convexity motivates the study of optimization in non-Euclidean spaces, such as Riemannian manifolds [1, 10, 34], Hadamard spaces [7] and non-commutative groups [14]. Some early works provide convergence guarantees for proximal-point methods [17, 19] and other optimization algorithms [32] on Riemannian manifolds, but they do not assume geodesic convexity, so that their analysis are limited to asymptotic results. The work [37] is the first to provide non-asymptotic rates of a number of first-order methods for geodesically convex optimization on Hadamard manifolds. A subsequent line of works establish iteration complexities for other optimization methods in the Riemannian setting, such as variance-reduced methods [31, 39], adaptive gradient methods [25], Frank-Wolfe [35], and Newton-type methods [22].
A key open question is whether it is possible to develop accelerated gradient methods on Riemannian manifolds. In [38] the authors develop the first method of such a kind, but the acceleration therein is attained only in a small neighbourhood of the global minimum. Later Ahn and Sra [3] show that a Riemannian version of Nesterov’s method enjoys global convergence, displaying a rate strictly better than Riemannian gradient descent and eventually attaining full acceleration. On the other hand, the work of Hamilton and Moitra [20] shows that full acceleration is not possible in general, which suggests that eventual acceleration is the best we can hope for.

Despite this progress, we still lack a systematic understanding of acceleration in the Riemannian setting. There are a large number of existing works dedicated to the understanding of acceleration in Euclidean spaces and a variety of different accelerated methods have been proposed. It is still unknown whether these methods can also be generalized to the Riemannian setting.

Among these works on acceleration in the Euclidean setting, the one motivating our paper is [29], which proposes an accelerated hybrid proximal extra-gradient (A-HPE) framework for convex functions. The authors of [29] show that Nesterov’s optimal method can be seen as a special case of A-HPE, and they also propose a second-order method A-NPE which is a specific implementation of their framework and has complexity $\tilde{O}(\varepsilon^{-2/7})$. Recently, this framework has been extended to strongly-convex functions [5, 8].

A hitherto unknown property of the A-HPE framework is that it can recover a wide range of accelerated methods that are independently proposed in past literature, such as the accelerated extra-gradient descent [18], the algorithm with an extra gradient descent step in [16, Section 4], among others. The A-HPE framework also has implications beyond first-order methods. A-NPE is used to design optimal second-order method in [6], and more generally, a number of works [12, 24] show that A-HPE can also induce optimal higher-order methods for smooth convex functions. A recent work [15] considers a different setting where we have access to a ball oracle, and it shows that combining A-HPE with line search yields an accelerated method that is near-optimal.

In view of these findings, we believe that A-HPE can help reveal the fundamental ideas behind the acceleration phenomenon. The main goal of this paper is to propose a Riemannian version of A-HPE and provide global convergence guarantees for this framework. To that end, we first revisit the Euclidean version of A-HPE in Section 2. We propose to view this framework as the linear coupling of two approximate proximal point iterates, and we demonstrate that this viewpoint produces a clean and simple analysis. We then introduce our Riemannian A-HPE framework in Section 3, and by following the Euclidean approach, we localize the main challenges of the analysis in the Riemannian setting. We overcome these challenges by leveraging geometric bounds on Riemannian manifolds, and this allows us to derive convergence guarantees for the Riemannian framework. Finally in Section 4 we consider a number of accelerated methods (including the recent work [3]) as special cases of Riemannian A-HPE, and our result for the framework implies that these methods provably achieve acceleration.

1.1 Notation and terminology

Throughout this paper we use $\langle \cdot, \cdot \rangle$ to denote inner product in an Euclidean space and $\| \cdot \|$ its induced norm. For a convex set $\mathcal{X} \subseteq \mathbb{R}^d$, we define the projection $\mathcal{P}_\mathcal{X}(x) := \arg\min_{y \in \mathcal{X}} \|x - y\|$. 

For a convex function $f : \mathbb{R}^d \to \mathbb{R}$, the proximal mapping of $f$ is given by
$$\text{prox}_f(x) := \arg\min_{u \in \mathbb{R}^d} f(u) + \frac{1}{2}\|u - x\|^2.$$ 
Equivalently, $u = \text{prox}_f(x)$ is the unique solution to $x - u \in \partial f(u)$.

For a $\mu$-strongly convex function $f : \mathbb{R}^d \to \mathbb{R}$, we can define a quadratic function
$$f_w(x) := f(w) + \langle x - w, \nabla \rangle + \frac{\mu}{2}\|x - w\|^2,$$
for $w \in \mathbb{R}^d$ and $\nabla \in \partial f(w)$, so that $f_w(x) \leq f(x)$ for all $x$.

## 2 Analysis of Euclidean A-HPE

The problem of minimizing a convex function $f$ in an Euclidean space is equivalent to finding a solution $x$ to $0 \in \partial f(x)$. This can be seen as a special case of the monotone inclusion problem $0 \in T(x)$, where $T$ is a maximal monotone point-to-set operator. The proximal point method (PPM) proposed by [27] is an iterative scheme for solving this problem. Specifically, it generates a sequence \(\{x_k\}\) according to the update rule $x_k = (\lambda_k T + I)^{-1}(x_{k-1})$.

While the PPM update is generally intractable for convex optimization problems, where $T = \partial f(x)$, it motivates the design and analysis of many implementable algorithms, which usually rely on relaxed notions of the PPM iterate. In [28] the authors propose the Hybrid Proximal Extragradient (HPE) framework and analyze its iteration complexity. This framework can be used to design first-order methods for smooth convex optimization. However, the convergence rate of these methods are similar to vanilla gradient descent and are slower than Nesterov’s accelerated method.

By incorporating ideas of Nesterov’s method, [29] proposed an Accelerated Hybrid Proximal Extragradient (A-HPE) framework. A-HPE converges faster than HPE and contains Nesterov’s method as a special case. Recently, some works [5, 8] extend the A-HPE framework to strongly-convex functions. We note that this framework is quite general and provides us a systematic way of designing accelerated methods.

### 2.1 The A-HPE framework

In this section, we revisit the A-HPE framework in Euclidean setting for strongly-convex functions. By connecting the update of the framework to that of the PPM method, we are able to come up a novel analysis which is much simpler and more intuitive than previous approaches [5, 8], and our analysis also serves as a guideline for generalization to the Riemannian setting.

Throughout this section we assume that $f$ is $\mu$-strongly-convex. Our description of the framework follows [8] and relies on the following notion called inexact proximal operator.

**Definition 2.1.** [8, Lemma 2.4] We write
$$(x, v) \in \text{iprox}_f(y, \lambda, \epsilon)$$
if the following inequality holds:
$$\frac{1}{2(1 + \lambda \mu)}\|x - y + \lambda v\|^2 + \frac{\lambda}{1 + \lambda \mu} (f(x) - f(w) - \langle x - w, v \rangle) + \frac{\mu}{2}\|x - w\|^2 \leq \epsilon \quad (2.1)$$
where $w \in \mathbb{R}^d$ satisfies $v - \mu x + \mu w \in \partial f(w)$. 

3
The inequality Equation (2.1) can be equivalently written as
\[ \frac{1}{2(1+\lambda\mu)^2} \|x - y + \lambda v\|^2 + \frac{\lambda}{1+\lambda\mu} \left( f(x) - f(w) - \langle x - w, \nabla \rangle - \frac{\mu}{2} \|x - w\|^2 \right) \leq \epsilon \] (2.2)
or
\[ \frac{1}{2(1+\lambda\mu)^2} \|x - y + \lambda v\|^2 + \frac{\lambda}{1+\lambda\mu} \left( f(x) - f(w) + \frac{1}{2\mu} \|
abla\|^2 - \frac{1}{2\mu} \|v\|^2 \right) \leq \epsilon \] (2.3)
where \( \nabla = v - \mu x + \mu w \in \partial f(w) \).

For \( \epsilon = 0 \), if we choose \( w = x \) and in the definition, then \( v \in \partial f(x) \) and \( x + \lambda v = y \), recovering the definition of exact proximal operator.

The definition we provide here is equivalent to the one in [8, Definition 2.3] that relies on the primal-dual gap of a proximal function. We use this version in this paper for convenience of analysis. To intuitively understand this definition, we now discuss its connection to the notion of \( \epsilon \)-subdifferential [11, Section 3].

**Definition 2.2.** Suppose that \( h : \mathbb{R}^d \to \mathbb{R} \) is \( \mu \)-strongly convex and \( x \in \mathbb{R}^d \). We say that \( u \in \mathbb{R}^d \) is an \( \epsilon \)-subgradient of \( f \) at \( x \) if the inequality
\[ f(y) \geq f(x) + \langle u, y - x \rangle + \frac{\mu}{2} \|y - x\|^2 - \epsilon \]
holds for all \( y \in \mathbb{R}^d \).

Note that the condition \( v - \mu x + \mu w \in \partial f(w) \) in Definition 2.1 implies that \( 0 \in \partial \Phi(w) \), where
\[ \Phi(z) = f(x) - f(z) - \langle x - z, v \rangle + \frac{\mu}{2} \|x - z\|^2 \]
Moreover, \( \Phi(z) \) is concave since \( f \) is \( \mu \)-strongly convex. Hence \( w \in \arg \max_z \Phi(z) \), and for any \( z \) we have
\[ f(z) \geq f(x) + \langle z - x, v \rangle + \frac{\mu}{2} \|z - x\|^2 - \frac{1 + \lambda\mu}{\lambda} \epsilon. \]
In other words, \( v \) is an \( \frac{1 + \lambda\mu}{\lambda} \epsilon \)-subgradient of \( f \) at \( x \). The inequality Equation (2.1) further implies that \( x + \lambda v \approx y \). Thus Definition 2.1 indeed defines an approximation to the exact proximal point,
for which $x + \lambda v = y$ and $v \in \partial f(x)$.

We now present the A-HPE framework in Euclidean setting in Algorithm 1. In contrast with previous versions of A-HPE [5, 8], here we allow $A_0$ to be arbitrary non-negative real number, and we introduce an additional sequence of variables \{$R_k$\} that differs from \{$A_k$\} only by a constant. The version presented in [5, 8] corresponds to the special case $A_0 = R_0 = 0$. We also define $w_{k+1}$ as the variable $w$ that appears in Definition 2.1, associated with $(x_{k+1}, v_{k+1})$. Note that $w_k$ would only appear in the analysis and is not needed in Algorithm 1.

We define $\nabla_{k+1} := v_{k+1} + \mu (w_{k+1} - x_{k+1}) \in \partial f(w_{k+1})$, so the last line of Algorithm 1 can be re-written as

$$z_{k+1} \leftarrow \frac{1 + \mu R_k}{1 + \mu R_{k+1}} z_k + \frac{\mu a_{k+1}}{1 + \mu R_{k+1}} w_{k+1} - \frac{a_{k+1}}{1 + \mu R_{k+1}} \nabla_{k+1}$$

(2.4)

2.2 Overview of the proof

We first provide an overview of the proof, which will shed light on the main ideas behind Algorithm 1. As a result, the derivations in this subsection may not be fully rigorous; we will present a formal proof in the next subsection.

Motivated by previous works [2, 4], we view Algorithm 1 as a combination of two approximate proximal point iterations, using different notions of approximation. The first uses the inexact proximal iteration in Definition 2.1; the second is implemented in the last line of Algorithm 1, or equivalently Equation (2.4). Indeed, we can rewrite the update rule as

$$z_{k+1} \leftarrow \arg \min_z \left\{ f(w_{k+1}) + \frac{1 + \mu R_k}{2a_{k+1}} \| z - z_k \|^2 \right\}$$

where $f(w_{k+1}) = f(w_{k+1}) + \langle \nabla_{k+1}, z - w_{k+1} \rangle + \frac{\mu}{2} \| z - w_{k+1} \|^2$ is a quadratic lower-approximation of $f$.

We first look at what we can obtain if we perform the above two updates separately. Since $(x_{k+1}, v_{k+1}) \in \text{iprox}(y, \lambda, \varepsilon_k)$, when $\varepsilon_k$ is small, we have

$$f(x_{k+1}) \lesssim f(w_{k+1}) + \frac{1}{2\mu} \left( \| v_{k+1} \|^2 - \| \nabla_{k+1} \|^2 \right)$$

(2.5a)

$$\leq f(x_k) - \frac{\mu}{2} \| x_k - x_{k+1} + \mu^{-1} v_{k+1} \|^2 + \frac{1}{2\mu} \| v_{k+1} \|^2$$

(2.5b)

where Equation (2.5a) follows from Equation (2.3), and Equation (2.5b) uses strong convexity of $f$ and $x_{k+1} - \mu^{-1} v_{k+1} = w_{k+1} - \mu^{-1} \nabla_{k+1}$. Equation (2.5) can be understood as a descent inequality for the function value at \{ $x_k$ \}, albeit with an “error term” $\| v_{k+1} \|$. When this term is large, we may no longer be able to control these function values.
We define the potential function

\[ p_k = A_k(f(x_k) - f(x^*)) + \frac{1 + \mu R_k}{2} \|z_k - z^*\|^2 \]  

(2.7)

our goal is to show that the sequence \( \{p_k\} \) is non-increasing, so that we can obtain a bound for \( f(x_k) - f(x^*) \).

In the work [8] the authors also use a potential function approach to show convergence of A-HPE. Motivated by our linear coupling viewpoint, we present our analysis in a clearer way,
which is helpful for addressing the key challenges that may arise in the Riemannian setting.

We first present a simple lemma which will be used to simplify our analysis. It can be checked using simple algebraic calculations, so we omit its proof here.

**Lemma 2.3 (Interpolation implies contraction).** For all \( p, q \in \mathbb{R} \) such that \( p + q > 0 \), we have

\[
p\|x\|^2 + q\|y\|^2 = (p + q)\left|\frac{p}{p+q}x + \frac{q}{p+q}y\right|^2 + \frac{pq}{p+q}\|x - y\|^2.
\]

The following lemma deals with the squared-distance terms in the potential function.

**Lemma 2.4.** We have

\[
\frac{1 + \mu R_k}{2}\|z_k - x^*\|^2 - \frac{1 + \mu R_{k+1}}{2}\|z_{k+1} - x^*\|^2 \geq a_{k+1}(f(w_{k+1}) - f(x^*))
\]

(2.8)

\[
+ \frac{\mu a_{k+1}(1 + \mu R_k)}{2(1 + \mu R_{k+1})}\|z_k - w_{k+1} + \mu^{-1}\nabla_{k+1}\|_2 - \frac{a_{k+1}}{2\mu}\|\nabla_{k+1}\|^2
\]

(2.9a)

\[
= \frac{\mu a_{k+1}}{2}\|x^* - w_{k+1} + \mu^{-1}\nabla_{k+1}\|_2 - \frac{\mu a_{k+1}(1 + \mu R_k)}{2(1 + \mu R_{k+1})}\|z_k - x_{k+1} + \mu^{-1}v_{k+1}\|^2
\]

(2.9b)

where Lemma 2.3 is used in Equation (2.9a), and Equation (2.9b) follows from Equation (2.4).

Thus, by strong convexity of \( f \) and the definition of \( w_{k+1} \) (see Definition 2.1) we have

\[
f(x^*) \geq f(w_{k+1}) + \langle \nabla_{k+1}, x^* - w_{k+1} \rangle + \frac{\mu}{2}\|x^* - w_{k+1}\|^2
\]

\[
= f(w_{k+1}) + \frac{\mu}{2}\|x^* - w_{k+1} + \mu^{-1}\nabla_{k+1}\|^2 - \frac{1}{2\mu}\|\nabla_{k+1}\|^2
\]

so that

\[
a_{k+1}(f(x^*) - f(w_{k+1})) \geq \frac{1 + \mu R_{k+1}}{2}\|z_{k+1} - x^*\|^2 - \frac{1 + \mu R_k}{2}\|z_k - x^*\|^2
\]

\[
+ \frac{\mu a_{k+1}(1 + \mu R_k)}{2(1 + \mu R_{k+1})}\|z_k - w_{k+1} + \mu^{-1}\nabla_{k+1}\|_2 - \frac{a_{k+1}}{2\mu}\|\nabla_{k+1}\|^2
\]

as desired.

**Remark 2.5.** The derivation of Equation (2.9) reveals the connection between the choice of parameters in the update Equation (2.4) and the growth of coefficient of the distance term in the construction of potential function. This observation will provide guidelines for choosing parameters in the Riemannian setting (cf. Lemma 3.4).
Now it suffices to deal with the function value terms. Strong convexity implies that
\[ f(x_k) \geq f(w_{k+1}) + \frac{\mu}{2} \|x_k - w_{k+1} + \mu^{-1} \nabla_{k+1}\|^2 - \frac{1}{2\mu} \|\nabla_{k+1}\|^2 \] (2.10)
and
\[ f(x_{k+1}) \geq f(w_{k+1}) + \frac{\mu}{2} \|\mu^{-1}v_{k+1}\|^2 - \frac{1}{2\mu} \|\nabla_{k+1}\|^2 \] (2.11)
while the definition of \( w_{k+1} \) implies
\[
\frac{\sigma_k^2}{2} \|x_{k+1} - y_k\|^2 \geq \frac{1}{2} \|x_{k+1} - y_k + \lambda_k v_{k+1}\|^2 \\
+ \lambda_k (1 + \lambda_k \mu) \left( f(x_{k+1}) - f(w_{k+1}) + \frac{1}{1+\lambda_k \mu} \left( \|\nabla_{k+1}\|^2 - \|v_{k+1}\|^2 \right) \right) \] (2.12)

We now seek a correct linear combination of the above inequalities to match the coefficient of \( p_k - p_{k+1} \). Note that adding (2.11) and (2.12) leads to the following simpler inequality
\[ \|x_{k+1} - y_k + \lambda_k v_{k+1}\|^2 \leq \sigma_k^2 \|x_{k+1} - y_k\|^2 \] (2.13)

The following lemma proves non-increasing of the potential function, which is based on the above observations and results.

**Lemma 2.6.** We have for all \( k \geq 0 \) that
\[ p_k - p_{k+1} \geq \frac{\mu \lambda_k A_k (1 + \mu R_k)}{2a_{k+1}} \|x_k - z_k\|^2 + \frac{(1 - \sigma_k^2 A_{k+1})}{2\lambda_k} \|x_{k+1} - y_k\|^2 \]

**Proof.** By combining the inequalities (2.8), (2.10), (2.12) we have
\[
p_k - p_{k+1} \\
\leq \left( \frac{1 + \mu R_k}{2} \|z_k - x^*\|^2 - \frac{1 + R_k}{2} \|z_{k+1} - x^*\|^2 + a_{k+1} (f(x^*) - f(w_{k+1})) \right) \\
+ A_k (f(x_k) - f(w_{k+1})) + A_{k+1} (f(w_{k+1}) - f(x_{k+1})) \tag{2.14}
\]
\[
\geq \frac{\mu a_{k+1} (1 + \mu R_k)}{2(1 + \mu R_{k+1})} \|z_k - x_{k+1} + \mu^{-1} v_{k+1}\|^2 + \frac{\mu A_k}{2} \|x_k - x_{k+1} + \mu^{-1} v_{k+1}\|^2 \\
- \frac{A_{k+1}}{2\mu} \|\nabla_{k+1}\|^2 + \frac{A_{k+1}}{2\mu} \left( \|\nabla_{k+1}\|^2 - \|v_{k+1}\|^2 \right) \\
+ \frac{A_{k+1}}{\lambda_k (1 + \lambda_k \mu)} \left( \frac{1}{2} \|x_{k+1} - y_k + \lambda_k v_{k+1}\|^2 - \sigma_k^2 \|x_{k+1} - y_k\|^2 \right)
\]

We now show that the last expression in the above inequality is positive. Recall that in Section 2.2 we made the intuitive argument which shows that the "positive term" of form \( \theta z \|z_k - x_{k+1} +
\( \mu^{-1}v_{k+1}^2 + \theta_x \| x_k - x_{k+1} + \mu^{-1}v_{k+1} \|_2^2 \) cannot be small. Formally, the choice of \( y_k \) implies that

\[
\frac{\mu a_{k+1} (1 + \mu R_k)}{2(1 + \mu R_{k+1})} \| z_k - x_k + \mu^{-1}v_{k+1} \|_2^2 + \frac{\mu A_k}{2} \| x_k - x_{k+1} + \mu^{-1}v_{k+1} \|_2^2 \\
\geq \frac{\mu (A_{k+1} + \mu (a_{k+1} R_k + A_k R_{k+1}))}{2(1 + \mu R_{k+1})} \| y_k - x_{k+1} + \mu^{-1}v_{k+1} \|_2^2 \\
+ \frac{\mu A_ka_{k+1} (1 + \mu R_k)}{2(A_{k+1} + \mu (a_{k+1} R_k + A_k R_{k+1}))} \| x_k - z_k \|_2^2 \\
= \frac{\mu a_{k+1}^2}{2 \lambda_k (1 + \mu R_{k+1})} \| y_k - x_{k+1} + \mu^{-1}v_{k+1} \|_2^2 + \frac{\mu \lambda_k A_k (1 + \mu R_k)}{2a_{k+1}} \| x_k - z_k \|_2^2
\]

where we have used the following equation

\[ a_{k+1}^2 = \lambda_k (A_{k+1} + \mu (a_{k+1} R_k + A_k R_{k+1})) \]  \hspace{1cm} (2.15)

to simplify the expression. We can now deduce from eq. (2.13) that the right hand side of eq. (2.14) is lower bounded by

\[
\frac{\mu a_{k+1}^2}{2 \lambda_k (1 + \mu R_{k+1})} \| y_k - x_{k+1} + \mu^{-1}v_{k+1} \|_2^2 + \frac{\mu \lambda_k A_k (1 + \mu R_k)}{2a_{k+1}} \| x_k - z_k \|_2^2 \\
- \frac{A_{k+1}}{2 \mu} \| v_{k+1} \|_2^2 + \frac{A_{k+1}}{\lambda_k} \left( \frac{1}{2} \| x_{k+1} - y_k + \lambda_k v_{k+1} \|_2^2 - \frac{\sigma^2}{2} \| x_{k+1} - y_k \|_2^2 \right).
\]

Now except from the \( \| x_k - z_k \|_2^2 \) term which is non-negative, the rest can be written as

\[
\alpha \| x_{k+1} - y_k \|_2^2 + 2 \beta \langle x_{k+1} - y_k, v_{k+1} \rangle + \gamma \| v_{k+1} \|_2^2
\]  \hspace{1cm} (2.16)

where

\[
\alpha = \frac{\mu a_{k+1}^2}{2 \lambda_k (1 + \mu R_{k+1})} + \frac{1 - \sigma^2}{2} \frac{A_{k+1}}{\lambda_k} \\
\beta = \frac{a_{k+1}^2}{2 \lambda_k (1 + \mu R_{k+1})} - \frac{1}{2} A_{k+1} = - \frac{\mu a_{k+1}^2}{2(1 + \mu R_{k+1})} \\
\gamma = \frac{a_{k+1}^2}{2 \mu \lambda_k (1 + \mu R_{k+1})} - \frac{A_{k+1}}{2 \mu} + \frac{1}{2} \lambda_k A_{k+1} \\
= \frac{1}{2} \lambda_k A_{k+1} - \frac{a_{k+1}^2}{2(1 + \mu R_{k+1})} = \frac{\mu \lambda_k a_{k+1}^2}{2(1 + \mu R_{k+1})}
\]

where we have used Equation (2.15) to simplify the expressions. Now it’s easy to see that the desired inequality holds.

We now make some remarks on the previous lemma.

1. Firstly, we can see from the proof that the choice of \( a_{k+1} \) guarantees that the quadratic function Equation (2.16) is non-negative. The correct way of obtaining \( a_{k+1} \) is to first deduce the quadratic function and then determine a proper choice of \( a_{k+1} \) such that the function is always non-negative. This approach will be used to derive the update rule of \( a_{k+1} \) in the Riemannian setting, where additional parameters need to be introduced due to the distortion phenomenon.
2. Secondly, as we have discussed before, \( x_k \) and \( z_k \) can both be regarded as an approximate proximal point iterate, and the point \( y_k \) is chosen on the segment between \( x_k \) and \( z_k \) in order to combine these two approaches. The ratio \( \| x_k - y_k \| : \| y_k - z_k \| \) follows naturally from the analysis and Lemma 2.3, which suggests the correct way of doing this combination.

We conclude this section by providing the convergence rate of A-HPE in Euclidean setting, which is a direct corollary of Lemma 2.6.

**Theorem 2.7.** For the iterates produced by Algorithm 1, we have

\[
f(x_k) - f(x^*) \leq \frac{1}{A_k} \left( A_0(f(x_0) - f(x^*)) + \frac{1}{2} \| x_0 - x^* \|^2 \right)
\]

\[
= \mathcal{O} \left( \Pi_{i=1}^{k-1} \left( 1 + \max \left\{ \mu \lambda_i, \sqrt{\mu \lambda_i} \right\} \right)^{-1} \right)
\]

Assuming that \( f \) is \( L \)-smooth, a number of first-order methods (including Nesterov’s method) can be considered as a special case of Algorithm 1, with the choice \( \lambda_i = \mathcal{O} \left( \frac{1}{L} \right) \). Theorem 2.7 then implies that these methods have optimal convergence rate of \( \mathcal{O} \left( \left( 1 + \sqrt{\frac{\mu}{L}} \right)^{-k} \right) \). We do not present concrete examples here since this is not the main focus of this paper, but we will make detailed discussions about this in the Riemannian setting in Section 4.

3 Riemannian A-HPE and its convergence analysis

The goal of this section is to generalize the results in the previous section to the optimization of function \( f(x) \) on a Riemannian manifold. In Section 3.1 we recall some useful notations and properties of Riemannian manifolds, and Section 3.2 is dedicated to the analysis of our proposed framework, Riemannian A-HPE.

### 3.1 Preliminaries of Riemannian geometry

In this subsection we recall some basic concepts from Riemannian geometry, and we refer the readers to standard textbooks [13, 26] for an in-depth introduction.

A smooth manifold \( \mathcal{M} \) is called a Riemannian manifold if an inner product \( \langle \cdot, \cdot \rangle_x \) is defined in the tangent space \( T_x \mathcal{M} \) for all \( x \in \mathcal{M} \) and the inner product varies smoothly in \( x \). In this section we use the notation \( \langle \cdot, \cdot \rangle \) and omit the dependence on \( x \), since it is clear from the context. We define \( \| \cdot \| \) to be the norm induced by the inner product i.e. \( \| v \| = \sqrt{\langle v, v \rangle} \).

A curve on \( \mathcal{M} \) is called a geodesic if it is locally distance-minimizing. The exponential map, denoted by \( \exp_x \), maps a vector \( v \in T_x \mathcal{M} \) to a point \( y \in \mathcal{M} \) such that there exists a geodesic \( \gamma : [0, 1] \to \mathcal{M} \) such that \( \gamma(0) = x \), \( \gamma(1) = y \) and \( \gamma'(0) = v \). We assume that the sectional curvature of \( \mathcal{M} \) is non-positive and lower bounded by \(-K\), where \( K \) is a positive real number. Under this assumption, any two points on \( \mathcal{M} \) are connected by a unique geodesic, and thus the inverse exponential map \( \exp_x^{-1} : \mathcal{M} \to T_x \mathcal{M} \) is well-defined. In a Euclidean space, \( \exp_x(y) = x + y \) and \( \exp_x^{-1}(y) = y - x \). Hence the exponential map provides a generalization of basic arithmetic operations to the Riemannian setting.

We use \( d(x, y) \) to denote the Riemannian distance between \( x \) and \( y \), that is, the length of the geodesic connecting them. The definition of exponential map implies that \( d(x, y) = \| \exp_x^{-1}(y) \| \).
In this section we will also make use of the tangent space distance, which is defined as 
\[ d_w(x, y) = \| \text{Exp}^{-1}_w(x) - \text{Exp}^{-1}_w(y) \| \]. Since by definition the geodesic is the shortest path between two points, we have 
\[ d_w(x, y) \leq d(x, y) \] for all \( w, x, y \in M \).

For functions on Riemannian manifolds, we can define the notion of geodesic convexity, which is a generalization of convexity in Euclidean spaces. Specifically, we say that a function \( f : M \to \mathbb{R} \) is \( \mu \)-geodesically strongly convex for \( \mu \geq 0 \) if for any \( x \in M \) there exists a non-empty set \( \partial f(x) \), such that for all \( y \in M \) and \( v \in \partial f(x) \) we have

\[
    f(y) \geq f(x) + \langle v, \text{Exp}^{-1}_x(y) \rangle + \frac{\mu}{2} \| \text{Exp}^{-1}_x(y) \|^2
\]

Thus \( f(x) := f(x) + \langle v, \text{Exp}^{-1}_x(y) \rangle + \frac{\mu}{2} \| \text{Exp}^{-1}_x(y) \|^2 \) is a lower approximation of \( f \). An equivalent definition is that

\[
    f(c(t)) \leq (1 - t)f(c(0)) + tf(c(1)) - \frac{t(1 - t)\mu}{2} L(c)^2
\]

for any geodesic \( c : [0, 1] \to M \). Here \( L(c) \) is the length of \( c \).

We use \( \Gamma^y_x \) to denote the parallel transport from \( T_xM \) to \( T_yM \) along the geodesic connecting \( x \) and \( y \). Parallel transport provides a natural way to compare vectors that lie in different tangent spaces. We say that \( f : M \to \mathbb{R} \) is \( L \)-smooth if it is differentiable and has \( L \)-Lipschitz gradient, in the sense that

\[
    \| \Gamma^y_x \nabla f(x) - \nabla f(y) \| \leq L \cdot d(x, y), \quad (3.1)
\]

for all \( x, y \in M \). This provides a natural generalization of smoothness to the Riemannian setting.

### 3.2 Riemannian A-HPE and its convergence

In this section, we first present the Riemannian A-HPE framework, which is a generalization of the (Euclidean) A-HPE framework (cf. Algorithm 1) to the Riemannian setting, and then we establish its convergence properties. We assume that \( f \) is \( \mu \)-geodesically strongly convex and the sectional curvature of \( M \) is bounded in \([ -K, 0 ]\).

#### 3.2.1 Algorithm and setup for its analysis

The update procedure of Riemannian A-HPE framework is given in Algorithm 2. There are some key features of this Riemannian framework, compared with its Euclidean counterpart:

A. **Adaptive step size.** The parameter \( D_k \) is choosing adaptively and serves as a reasonable upper bound for \( d(w_{k+1}, y_k) \). Note that this bound shall be derived at the beginning of iteration \( k \).

The step size depends on \( D_k \) and thus it is also adaptive. Intuitively, this is because we need to overcome some distortion phenomenon in the Riemannian setting, so in the first few iterations when we are still far from \( x^* \), we need to use smaller step size. This is in contrast with the Euclidean setting, where in theoretical analysis, algorithms such as gradient descent and Nesterov’s method use a constant step size.

B. **Riemannian inexact proximal operator.** We need a Riemannian version of the inexact proximal operator which is defined below.
Algorithm 2 Strongly convex Riemannian accelerated hybrid proximal extragradient method

**Input:** Objective function $f$, initial point $x_0$, ‘reference’ step size $\lambda > 0$, $\sigma_k \in (0, 1)$, bounded sequence $\{c_k\}$ and initial weight $A_0, B_0 \geq 0$

1: set $z_0 = x_0$
2: for $k = 0, 1, \cdots$ do
3: choose a valid distortion rate $\delta_k$ according to Lemma 3.3
4: choose $D_k > 0$ based on history information
5: set $\lambda_k = \min \left\{ \lambda, \frac{1-\sigma_k}{\delta_k \epsilon_k + D_k} \right\}$ (S$_K$ is defined in Lemma 3.8)
6: set $\theta_k$ to be the smaller root of $B_k(1 - \theta)^2 = \mu \lambda_k \theta \left( (1 - \theta)B_k + \frac{\mu}{2} \delta_k A_k \right)$
7: set $B_{k+1} = \frac{B_k}{\delta_k \theta_k}$
8: set $a_{k+1} = \frac{1-\theta_k}{\delta_k \theta_k} B_k$
9: set $A_{k+1} = A_k + a_{k+1}$
10: set $y_k = \text{Exp}_{x_k} \left( \frac{\theta_k a_{k+1}}{A_k + \theta_k a_{k+1}} \text{Exp}_{x_k}^{-1}(z_k) \right)$
11: set $\epsilon_k = \frac{\sigma_k^2}{\sqrt{2(1+\lambda_k \mu)^2}} d_{w_k}^2(x_{k+1}, y_k)$
12: choose $(x_{k+1}, v_{k+1}) \in \text{iprox}^{w_k}_{\epsilon_k+1}(y_k, \lambda_k \epsilon_k)$ such that $d(w_{k+1}, y_k) \leq D_k$ and $d(w_{k+1}, y_k) \leq \epsilon_k d_{w_k}(x_{k+1}, y_k)$
13: choose $r_k > 0$ as an upper bound for $d(w_{k+1}, x^*)$ and $B_k = \{ \|x\| \leq r_k \} \subset \mathbb{R}^d$
14: set $z_{k+1} = \text{Exp}_{w_{k+1}} B_k \left( (1 - \theta_k) \text{Exp}_{w_{k+1}}^{-1}(x_{k+1}) + \theta_k \text{Exp}_{w_{k+1}}^{-1}(z_k) - \frac{\lambda_k \theta_k}{2} v_{k+1} \right)$
15: end for

**Definition 3.1** (Riemannian inexact proximal operator). For $x, y \in M, v \in T_{x^*} M$ and $\lambda, \epsilon \geq 0$, we write

$$ (x, v) \in \text{iprox}_f^{w}(y, \lambda, \epsilon) $$

if we have the inequality

$$ \frac{1}{2(1+\lambda \mu)^2} \|\text{Exp}_w^{-1}(x) - \text{Exp}_w^{-1}(y) + \lambda v\|^2 $$

$$ + \frac{\lambda}{1+\lambda \mu} \left( f(x) - f(w) - \left\langle \text{Exp}_w^{-1}(x), v \right\rangle + \frac{\mu}{2} d^2(x, w) \right) \leq \epsilon $$

and $v - \mu \text{Exp}_w^{-1}(x) \in \partial f(w)$. 

A difference with the Euclidean setting is that the update rule of $z_{k+1}$ in Algorithm 2 now require the knowledge of $w_{k+1}$. One may notice that if we first fix $x_{k+1}$ and $v_{k+1}$, then $w_{k+1}$ needs to be solved from a non-linear condition $v - \mu \text{Exp}_w^{-1}(x) \in \partial f(w)$. As we demonstrate in Section 4, for specific implementations of the framework we choose these quantities simultaneously so that we can always ensure that $w_{k+1}$ can be updated in an explicit form, without involving any non-trivial sub-problems.

C. **Projection operator in the update of $z_k$**. We update $z_k$ via an exponential map at $w_{k+1}$ since we will be working in the tangent space $T_{w_{k+1}}$ in subsequent analysis. The update rule involves a projection operator performed in the tangent space, which is absent in the Euclidean version. The main reason for this modification is that we require some boundedness properties of $z_{k+1}$ for convergence of the framework. The radius of ball $B_k$ depends on an upper
estimate of \(d(w_{k+1}, x^*)\), which should be obtained without knowledge of \(x^*\). For example, \(r_k = \mu^{-1}\|\nabla f(w_{k+1})\|\) is a valid choice, due to strong convexity of \(f\). The sequences \(\{r_k\}\) and \(\{D_k\}\) are both introduced to control the step size \(\lambda_k\). They are not needed in some special cases where the distortion phenomenon in Lemma 3.5 is not present; we will discuss these cases in Section 4.1.

### 3.2.2 Analysis of Algorithm 2

We are now ready to begin the analysis of Algorithm 2. Similar to the Euclidean setting, we define the potential function

\[p_k = A_k \cdot (f(x_k) - f(x^*)) + B_k \cdot d_w^2(z_k, x^*).\]

Note that in the above definition, we use the tangent space distance \(d_w\) rather than the Riemannian distance \(d\). Indeed, when generalizing our analysis to the Riemannian setting, we need to work with vectors in tangent spaces, so that it is more convenient to use the tangent space distance here.

In the previous section, a key step for bounding potential decrease is to bound the decrease of the squared-distance term (cf. Lemma 2.4). In Riemannian setting, however, it is hard to bound the term \(B_k \cdot d_w^2(z_k, x^*) - B_{k+1} \cdot d_w^2(z_{k+1}, x^*)\) because it involves the vectors in two different tangent spaces \(T_{w_k}M\) and \(T_{w_{k+1}}M\). Similar to [3], we introduce the notion of distortion rate to overcome this issue.

**Definition 3.2.** We say that \(\delta_k > 0\) is a valid distortion rate if it satisfies

\[d_{w_{k+1}}^2(z_k, x^*) \leq \delta_k d_{w_k}^2(z_k, x^*)\]

The next lemma shows that we can obtain a valid distortion rate in terms of \(d(w_k, z_k)\), and independent of \(x^*\) which we do not know during the optimization process.

**Lemma 3.3.** [3, Lemma 4.1] For any points \(x, y, z \in M\), we have \(d^2(x, y) \leq T_K(d(x, z))d^2(x, y)\), where

\[T_K(r) := \max \left\{ 1 + 4 \left( \frac{\sqrt{K}}{\tanh(\sqrt{K}r)} - 1 \right), \left( \frac{\sinh(2\sqrt{K}r)}{2\sqrt{K}r} \right)^2 \right\}, \quad \text{if } r > 0,
\]

\[1, \quad \text{if } r = 0,\]

In particular, \(\delta_k = T_K(d(w_k, z_k))\) is a valid distortion rate.

Now, we can prove the following lemma which is a Riemannian counterpart of Lemma 2.4.

**Lemma 3.4.** Suppose that \(\delta_k > 0\) is a valid distortion rate and \(B_{k+1} = \frac{B_k}{\delta_k}\), then

\[B_k d_w^2(z_k, x^*) - B_{k+1} d_w^2(z_{k+1}, x^*) \geq (1 - \theta_k) B_{k+1} \left( \frac{2}{\mu} (f(w_{k+1}) - f(x^*)) - \frac{1}{\mu^2} \|\nabla_{k+1}\|^2 \right) + \theta_k (1 - \theta_k) B_{k+1} \left\| \text{Exp}_{w_{k+1}}^{-1}(z_k) - \text{Exp}_{w_{k+1}}^{-1}(x_{k+1}) + \mu^{-1} v_{k+1} \right\|^2\]
Proof. We define an additional point

$$
\tilde{z}_{k+1} = \text{Exp}_{w_{k+1}} \left( (1 - \theta_k)\text{Exp}_{w_{k+1}}^{-1}(x_{k+1}) + \theta_k\text{Exp}_{w_{k+1}}^{-1}(z_k) - \frac{1 - \theta_k}{\mu}v_{k+1} \right)
$$

so that \(\text{Exp}_{w_{k+1}}^{-1}(z_{k+1}) = \mathcal{P}_{B_k} \left( \text{Exp}_{w_{k+1}}^{-1}\tilde{z}_{k+1} \right)\). Since \(\text{Exp}_{w_{k+1}}^{-1}(x^*) \in B_k\) as guaranteed by the algorithm, we have \(d_{w_{k+1}}(z_{k+1}, x^*) \leq d_{w_{k+1}}(\tilde{z}_{k+1}, x^*)\) due to the contraction property of the projection operator.

Since \(\delta_k\) is a valid distortion rate, we have

$$
B_kd_{w_k}^2(z_k, x^*) \geq \frac{B_k}{\delta_k}d_{w_{k+1}}^2(z_k, x^*)
$$

(3.3)

This implies that

$$
B_{k+1}d_{w_{k+1}}^2(z_{k+1}, x^*) - B_kd_{w_k}^2(z_k, x^*) \leq B_{k+1}d_{w_{k+1}}^2(\tilde{z}_{k+1}, x^*) - B_kd_{w_k}^2(z_k, x^*) \leq B_{k+1}d_{w_{k+1}}^2(\tilde{z}_{k+1}, x^*) - \theta_kB_{k+1}d_{w_{k+1}}^2(z_k, x^*)
$$

(3.4a)

$$
\leq (1 - \theta_k)B_{k+1} \left( \frac{1}{1 - \theta_k}d_{w_{k+1}}(\tilde{z}_{k+1}, x^*) - \frac{\theta_k}{1 - \theta_k}d_{w_{k+1}}(z_k, x^*) \right)^2
$$

(3.4b)

$$
= (1 - \theta_k)B_{k+1} \left\| \text{Exp}_{w_{k+1}}^{-1}(x^*) - \text{Exp}_{w_{k+1}}^{-1}(x_{k+1}) + \mu^{-1}v_{k+1} \right\|^2 - \theta_k(1 - \theta_k)B_{k+1} \left\| \text{Exp}_{w_{k+1}}^{-1}(z_k) - \text{Exp}_{w_{k+1}}^{-1}(x_{k+1}) + \mu^{-1}v_{k+1} \right\|^2
$$

(3.4c)

(3.4d)

where Equation (3.4a) uses the property of projection operator, Equation (3.4b) follows from Equation (3.3) and \(\theta_kB_{k+1} = \frac{B_k}{\delta_k}\), Equation (3.4c) uses Lemma 2.3, and Equation (3.4d) follows from the definition of \(\tilde{z}_{k+1}\). On the other hand, by strong convexity of \(f\), we have

$$
f(x^*) - f(w_{k+1}) \geq \left\langle \text{Exp}_{w_{k+1}}^{-1}(x^*), \nabla_{k+1} \right\rangle + \frac{\mu}{2} \left\| \text{Exp}_{w_{k+1}}^{-1} \right\|^2
$$

$$
= \frac{\mu}{2} \left\| \text{Exp}_{w_{k+1}}^{-1}(x^*) + \mu^{-1}\nabla_{k+1} \right\|^2 - \frac{1}{2\mu} \left\| \nabla_{k+1} \right\|^2
$$

$$
= \frac{\mu}{2} \left\| \text{Exp}_{w_{k+1}}^{-1}(x^*) - \text{Exp}_{w_{k+1}}^{-1}(x_{k+1}) + \mu^{-1}v_{k+1} \right\|^2 - \frac{1}{2\mu} \left\| \nabla_{k+1} \right\|^2
$$

The conclusion follows by plugging this inequality into Equation (3.4). Note that the steps after Equation (3.4b) are essentially the same as the Euclidean setting, because all the calculations are done in the tangent space \(T_{w_{k+1}} M\).

We then proceed to derive a Riemannian analog of Lemma 2.6, where we proved the potential decrease in the Euclidean setting. By following the same approach, we can see that the inequality would involve an additional point \(y'_k\).
Lemma 3.5. Suppose that $a_{k+1} = A_{k+1} - A_k = \frac{2}{\mu}(1 - \theta_k)B_{k+1}$, then

\[ p_k - p_{k+1} \geq \frac{\mu}{2}(\theta_k a_{k+1} + A_k) \left\| \text{Exp}_{w_{k+1}}^{-1}(y'_k) - \text{Exp}_{w_{k+1}}^{-1}(x_{k+1}) + \mu^{-1}v_{k+1} \right\|^2 \]

\[ + \frac{A_{k+1}}{2\lambda_k \sigma_k^2} \left\| \text{Exp}_{w_{k+1}}^{-1}(y_k) - \text{Exp}_{w_{k+1}}^{-1}(x_{k+1}) - \lambda_k v_{k+1} \right\|^2 \]

\[ + \frac{\mu \theta_k a_{k+1}}{2(A_k + \theta_k a_{k+1})} d^2_{w_{k+1}}(x_k, y_k) - \frac{\sigma_k A_{k+1}}{2\lambda_k} d^2_{w_{k+1}}(x_{k+1}, y_{k+1}) - \frac{A_{k+1}}{2\mu} \left\| v_{k+1} \right\|^2 \]  

(3.5)

where

\[ y'_k = \text{Exp}_{w_{k+1}}^{-1} \left( \frac{A_k}{A_k + \theta_k a_{k+1}} \text{Exp}_{w_{k+1}}^{-1}(x_k) + \frac{\theta_k a_{k+1}}{A_k + \theta_k a_{k+1}} \text{Exp}_{w_{k+1}}^{-1}(z_k) \right) \]

(3.6)

Proof. Strong convexity implies that

\[ f(x_k) \geq f(w_{k+1}) + \frac{\mu}{2} \left\| \text{Exp}_{w_{k+1}}^{-1}(x_k) + \mu^{-1} \nabla_k \right\|^2 - \frac{1}{2\mu} \left\| \nabla_k \right\|^2 \]

and

\[ f(x_{k+1}) \geq f(w_{k+1}) + \frac{\mu}{2} \left\| \mu^{-1}v_{k+1} \right\|^2 - \frac{1}{2\mu} \left\| \nabla_k \right\|^2 , \]

and the definition of Riemannian iprox operator Equation (3.2) implies that

\[ \frac{\sigma_k^2}{2} \left\| \text{Exp}_{w_{k+1}}^{-1}(x_{k+1}) - \text{Exp}_{w_{k+1}}^{-1}(y_k) \right\|^2 \geq \frac{1}{2} \left\| \text{Exp}_{w_{k+1}}^{-1}(x_{k+1}) - \text{Exp}_{w_{k+1}}^{-1}(y_k) + \lambda_k v_{k+1} \right\|^2 \]

\[ + \lambda_k (1 + \lambda_k \mu) \left( f(x_{k+1}) - f(w_{k+1}) + \frac{1}{2\mu} \left( \left\| \nabla_k \right\|^2 - \left\| v_{k+1} \right\|^2 \right) \right) \]

Combining the above inequalities, we have

\[ p_k - p_{k+1} \]

\[ = \left( B_k d^2_{w_{k+1}}(z_k, x^*) - B_{k+1} d^2_{w_{k+1}}(z_{k+1}, x^*) + \frac{2}{\mu}(1 - \theta_k)B_{k+1}(f(x^*) - f(w_{k+1})) \right) \]

\[ + A_k(f(x_k) - f(w_{k+1})) + A_{k+1}(f(w_{k+1}) - f(x_{k+1})) \]

\[ \geq \frac{\mu A_k}{2} \left\| \text{Exp}_{w_{k+1}}^{-1}(x_k) - \text{Exp}_{w_{k+1}}^{-1}(x_{k+1}) + \mu^{-1}v_{k+1} \right\|^2 \]

\[ + \frac{\mu \theta_k a_{k+1}}{2\lambda_k \sigma_k^2} \left\| \text{Exp}_{w_{k+1}}^{-1}(y_k) - \text{Exp}_{w_{k+1}}^{-1}(x_{k+1}) + \mu^{-1}v_{k+1} \right\|^2 \]

\[ + \frac{A_{k+1}}{2\lambda_k} d^2_{w_{k+1}}(x_{k+1}, y_{k+1}) - \frac{A_{k+1}}{2\mu} \left\| v_{k+1} \right\|^2 \]  

(3.7a)

where we use the condition $a_{k+1} = \frac{2}{\mu}(1 - \theta_k)B_{k+1}$ in (3.7a). Finally, Lemma 2.3 implies that

\[ \frac{\mu A_k}{2} \left\| \text{Exp}_{w_{k+1}}^{-1}(x_k) - \text{Exp}_{w_{k+1}}^{-1}(x_{k+1}) + \mu^{-1}v_{k+1} \right\|^2 \]

\[ + \frac{\mu \theta_k a_{k+1}}{2\lambda_k \sigma_k^2} \left\| \text{Exp}_{w_{k+1}}^{-1}(y_k) - \text{Exp}_{w_{k+1}}^{-1}(x_{k+1}) + \mu^{-1}v_{k+1} \right\|^2 \]

\[ = \frac{\mu}{2}(\theta_k a_{k+1} + A_k) \left\| \text{Exp}_{w_{k+1}}^{-1}(y'_k) - \text{Exp}_{w_{k+1}}^{-1}(x_{k+1}) + \mu^{-1}v_{k+1} \right\|^2 \]

\[ + \frac{\mu \theta_k a_{k+1} A_k}{2(A_k + \theta_k a_{k+1})} d^2_{w_{k+1}}(x_k, z_k) . \]
The conclusion follows.

Remark 3.6. The final equation in the proof explains why \( y'_k \) would appear in Equation (3.5). In the Euclidean setting we always have \( y_k = y'_k \), but this is not necessarily true for the Riemannian setting, which prevents us from directly following the Euclidean approach. Moreover, the definition of \( y'_k \) depends on \( w_{k+1} \), so it cannot be used as the update rule of \( y_k \). As a result, Lemma 3.5 highlights an additional distortion that arises in the analysis of Riemannian A-HPE, which is not present in previous works that focus on the Riemannian version of Nesterov’s method \cite{3, 38}.

Naturally, we hope that \( y_k \) and \( y'_k \) are close, so that replacing \( y_k \) with \( y'_k \) in Equation (3.5) would only induce a small error term. More formally, we will use the following inequality, which is a direct consequence of Cauchy-Schwarz:

\[
\left\| \text{Exp}_{w_{k+1}}^{-1}(y'_k) + u \right\|^2 \geq \frac{3}{4} \left\| \text{Exp}_{w_{k+1}}^{-1}(y_k) + u \right\|^2 - 3 \cdot d^2_{w_{k+1}}(y_k, y'_k)
\]

This inequality suggests that there are two major challenges in the analysis:

- The coefficient of the term \( \left\| \text{Exp}_{w_{k+1}}^{-1}(y_k) + u \right\|^2 \) becomes smaller, compared with the analysis in the Euclidean setting.
- We need to develop an upper bound for the term \( d^2_{w_{k+1}}(y_k, y'_k) \).

The following lemma deals with the first challenge. We use the observation in previous section that the update rule of \( a_{k+1} \) is determined by ensuring the non-negativity of a quadratic function. The update rule of \( \theta_k \) in Algorithm 2 (which also determines \( a_{k+1} \)), while looks complicated, actually follows from the simple equation \( A_{k+1} = (1 + \mu \lambda_k)(\theta_k a_{k+1} + A_k) \) derived in the proof.

Lemma 3.7. Suppose that \( \sigma_k \leq \frac{3}{4} \), then

\[
\frac{(1 - \sigma_k) A_{k+1}}{2\lambda_k} d^2_{w_{k+1}}(x_{k+1}, y_k)
\]

\[
\leq \frac{3\mu}{8} (\theta_k a_{k+1} + A_k) \left\| \text{Exp}_{w_{k+1}}^{-1}(y_k) - \text{Exp}_{w_{k+1}}^{-1}(x_{k+1}) + \mu^{-1}v_{k+1} \right\|^2
\]

\[
+ \frac{A_{k+1}}{2\lambda_k} \left\| \text{Exp}_{w_{k+1}}^{-1}(y_k) - \text{Exp}_{w_{k+1}}^{-1}(x_{k+1}) - \lambda_kv_{k+1} \right\|^2
\]

\[
+ \frac{\mu \theta_k a_{k+1} A_k}{2(A_k + \theta_k a_{k+1})} d^2_{w_{k+1}}(x_k, z_k) - \frac{\sigma_k A_{k+1}}{2\lambda_k} d^2_{w_{k+1}}(x_{k+1}, y_k)
\]

Proof. The difference of the right hand side and left hand side of the inequality can be written in the following form (where we omit the \( d^2_{w_{k+1}}(x_k, z_k) \) term, which is non-negative):

\[
a d^2_{w_{k+1}}(x_{k+1}, y_k) + 2\beta \left\langle \text{Exp}_{w_{k+1}}^{-1}(y_k) - \text{Exp}_{w_{k+1}}^{-1}(x_{k+1}), v_{k+1} \right\rangle + \gamma \|v_{k+1}\|^2
\]

where

\[
\alpha = \frac{3\mu}{8} (\theta_k a_{k+1} + A_k) + \frac{(1 - \sigma_k) A_{k+1}}{2\lambda_k}
\]

\[
\beta = \frac{3}{8} (\theta_k a_{k+1} + A_k) - A_{k+1}
\]

\[
\gamma = \frac{3}{8\mu} (\theta_k a_{k+1} + A_k) + \frac{\lambda_k A_{k+1}}{2} - \frac{A_{k+1}}{2\mu}
\]
Let $a = \theta_k a_{k+1} + A_k$ and $b = A_{k+1}$, then

$$4(\alpha \gamma - \beta^2) = \frac{3}{4} \left( \frac{1 - \sigma_k}{\lambda_k \mu} + \lambda_k \mu - 1 \right) ab - \left( \sigma_k - \frac{1 - \sigma_k}{\lambda_k \mu} \right) b^2.$$ 

It follows from line 5 in Algorithm 2 and the choice of $a_{k+1}$ in Lemma 3.5 that $A_{k+1} = (1 + \mu \lambda_k)(\theta_k a_{k+1} + A_k)$, so $\beta^2 \leq \alpha \gamma$ when $\sigma_k \leq \frac{3}{4}$. This proves the result. \hfill \Box

It follows from Lemma 3.5 and Lemma 3.7 that

$$p_k - p_{k+1} \geq \frac{\mu \theta_k a_{k+1} A_k}{2(1 + \theta_k a_{k+1})} d_{w_{k+1}}^2 (x_k, z_k) + \frac{(1 - \sigma_k) A_{k+1}}{2\lambda_k} d_{w_{k+1}}^2 (x_{k+1}, y_k)$$

$$- \frac{3\mu}{2}(\theta_k a_{k+1} + A_k) d_{w_{k+1}}^2 (y_k, y_k') \tag{3.8}$$

We now turn to bound the distance $d_{w_{k+1}} (y_k, y_k')$. We need the following lemma, which is proved in [33, Section B.3].

**Lemma 3.8.** [33, Lemma 3] Suppose that $x \in \mathcal{M}$ and $y, a \in T_x \mathcal{M}$. Let $z = \text{Exp}_x (a)$, then we have

$$d(\text{Exp}_x (y + a), \text{Exp}_x (\Gamma z y)) \leq \min \{ \|a\|, \|y\| \} S_k(\|a\| + \|y\|)$$

where

$$S_k(r) = \frac{1}{r} \left( \cosh \left( \sqrt{K} r \right) - \frac{\sinh \left( \sqrt{K} r \right)}{\sqrt{K} r} \right)$$

Note that a key feature of the function $S_K$ is that $\lim_{r \to 0} S_K(r) = 0$.

**Lemma 3.9.** We have for all $k \geq 1$ that

$$d_{w_{k+1}} (y_k, y_k') \leq d(w_{k+1}, y) \cdot S_k (d(x_k, z_k) + d(y_k, w_{k+1}))$$

**Proof.** Let $\tau = \frac{\theta_k a_{k+1}}{A_k + \theta_k a_{k+1}}$, then by definition

$$y_k' = \text{Exp}_{w_{k+1}} \left( (1 - \tau) \text{Exp}_{w_{k+1}}^{-1} (x_k) + \tau \text{Exp}_{w_{k+1}}^{-1} (z_k) \right)$$

On the other hand, $y_k$ can alternatively be written as

$$y_k = \text{Exp}_{w_{k+1}} \left( (1 - \tau) \Gamma_{y_k}^{w_{k+1}} \text{Exp}_{y_k}^1 (x_k) + \tau \Gamma_{y_k}^{w_{k+1}} \text{Exp}_{y_k}^{-1} (z_k) + \text{Exp}_{w_{k+1}}^1 (y_k) \right)$$

Hence by Lemma 3.8 we have

$$d_{w_{k+1}} (y_k, y_k') \leq (1 - \tau) d \left( x_k, \text{Exp}_{w_{k+1}} \left( \Gamma_{y_k}^{w_{k+1}} \text{Exp}_{y_k}^{-1} (x_k) + \text{Exp}_{w_{k+1}}^1 (y_k) \right) \right)$$

$$+ \tau d \left( z_k, \text{Exp}_{w_{k+1}} \left( \Gamma_{y_k}^{w_{k+1}} \text{Exp}_{y_k}^{-1} (z_k) + \text{Exp}_{w_{k+1}}^1 (y_k) \right) \right)$$

$$\leq (1 - \tau) d(w_{k+1}, y_k) \cdot S_k (d(y_k, x_k) + d(y_k, w_{k+1}))$$

$$+ \tau d(w_{k+1}, y) \cdot S_k (d(y_k, z_k) + d(y_k, w_{k+1}))$$

$$\leq d(w_{k+1}, y_k) \cdot S_k (d(x_k, z_k) + d(y_k, w_{k+1}))$$

as desired. \hfill \Box
Remark 3.10. The above lemma provides guidance for choosing the step size $\lambda_k$. To ensure sufficient decrease of potential function, we would like to have (cf. Equation (3.8))

$$
\frac{(1 - \sigma_k)A_{k+1}^2}{4\lambda_k} d_{w_{k+1}}^2(x_{k+1}, y_k) \geq \frac{3\mu}{2} (\theta_k a_{k+1} + A_k) d_{w_{k+1}}^2(y_k, y'_k)
$$

so that we can obtain the following descent inequality for the potential function:

$$
p_k - p_{k+1} \geq \frac{\mu \theta_k a_{k+1} A_k}{2(A_k + \theta_k a_{k+1})} d_{w_{k+1}}^2(x_k, z_k) + \frac{(1 - \sigma_k)A_{k+1}^2}{4\lambda_k} d_{w_{k+1}}^2(x_{k+1}, y_k)
$$

The update of Algorithm 2 implies that $d_{w_{k+1}}(x_{k+1}, y_k) \geq c_k^{-1} d(w_{k+1}, y_k)$. Note that $A_{k+1} = (1 + \mu \lambda_k)(\theta_k a_{k+1} + A_k)$, so the previous lemma implies that it suffices to have

$$
\lambda_k \leq \frac{1 - \sigma_k}{6c_k^2 S_k^2} (d(x_k, z_k) + d(w_{k+1}, y_k)) \leq \frac{1 - \sigma_k}{6c_k^2} \left( \frac{d(w_{k+1}, y_k)}{d(w_{k+1}, y'_k)} \right)^2
$$

We now summarize our findings.

Theorem 3.11. Suppose that $\sigma_k \leq \frac{3}{4}$, the iterates produced by Algorithm 2 satisfies Equation (3.9). In particular, we have

$$
\begin{align*}
    f(x_k) - f(x^*) &\leq \frac{p_0}{A_k} \leq \frac{p_0}{A_0} \cdot \prod_{i=0}^{k-1} (1 + 2\mu \lambda_i)^{-1} \\
    d_{w_{k+1}}^2(z_k, x^*) &\leq \frac{p_0}{B_k} \leq \frac{p_0}{A_0} \cdot (\mu \lambda_{k-1})^{-1} \prod_{i=0}^{k-2} (1 + 2\mu \lambda_i)^{-1}
\end{align*}
$$

Proof. We have seen that Algorithm 2 ensures (3.9), which in particular implies that

$$
p_0 \geq p_k = A_k (f(x_k) - f(x^*)) + B_k d_{w_{k+1}}^2(z_k, x^*)
$$

On the other hand, $A_{k+1} = (1 + \mu \lambda_k)(\theta_k a_{k+1} + A_k) \geq (1 + \mu \lambda_k) A_k \geq \prod_{i=0}^{k-1} (1 + \mu \lambda_i) A_0$ and $B_{k+1} = \frac{\mu}{2(1 - \theta_k)} A_{k+1} \geq \frac{\mu \lambda}{2(1 + \lambda \mu)} A_{k+1}$. Thus (3.10) holds.

The following theorem is the main result of this section, which shows linear convergence of the iterates of Algorithm 2. Moreover, we are able to characterize the asymptotic convergence rate.

Theorem 3.12. Suppose that $\sigma_k \leq \frac{3}{4}$ and $\{r_k\}$ is bounded, then the iterates generated by Algorithm 2 are uniformly bounded. Moreover,

1. If $\{D_k\}$ is bounded, then $x_k \to x^*$ and $z_k \to x^*$ as $k \to +\infty$.
2. If $\lim_{k \to +\infty} D_k = 0$, then there exists $N \in \mathbb{Z}_+$ such that $\lambda_k = \lambda$ for $k > N$.
3. Suppose that the conclusion of (2) holds, then

$$
\lim_{k \to +\infty} \frac{a_k}{A_k} = \sqrt{\frac{\mu \lambda}{1 + \mu \lambda}}
$$

Before giving the proof of this result, we first make some remarks on its assumptions and implications.

First, the assumption that $D_k = o(1)$ seems restrictive at first glance; however, it can be satisfied by inductively choosing $D_k$ based on the history of past iterations. Intuitively, suppose
that \( x_k \) and \( z_k \) are near \( x^* \) (which can be deduced from (1)), then \( y_k \) is also near \( x^* \), according to Algorithm 2. Hence, if we choose \( w_{k+1} \) reasonably based on \( x_k, y_k \) and \( z_k \), then it is also near \( x^* \). In this case \( d(w_{k+1}, y_k) \) is small, and most importantly, we are able to develop upper bounds of it which is computable before iteration \( k \), i.e. independent of the parameter choices at iteration \( k \). Similarly, the assumption that \( \{r_k\} \) is bounded can be verified for reasonable choices of \( w_{k+1} \). We will illustrate how to check the assumptions of Theorem 3.12 in Section 4.

Second, the item (3) of the above theorem implies that

\[
\lim_{k \to +\infty} \frac{A_{k+1}}{A_k} = 1 + \mu \lambda + \sqrt{\mu \lambda (1 + \mu \lambda)}
\]

This shows that the convergence rate depends on the choice of \( \lambda \). Note that there are two different phases regarding the convergence rate: when \( \mu \lambda \ll 1 \), we have \( \lim_{k \to +\infty} \frac{A_{k+1}}{A_k} \approx 1 + \sqrt{\mu \lambda} \), while for \( \mu \lambda \gg 1 \), \( \lim_{k \to +\infty} \frac{A_{k+1}}{A_k} \approx 1 + 2 \mu \lambda \). In particular, for \( \lambda = O \left( \frac{1}{\mu} \right) \), the theorem implies that Algorithm 2 eventually achieves the accelerated rate of Nesterov’s method.

**Proof.** We first show that all iterates produced by the algorithm are bounded. Since we have potential decrease as explained in Remark 3.10, we have that \( f(x_k) - f(x^*) \) is bounded, so \( \{x_k\} \) is bounded by strong convexity of \( f \). The boundedness of \( \{r_k\} \) implies that \( \{z_k\} \) is bounded. Thus \( \{y_k\} \) is bounded since \( y_k \) lies on the geodesic between \( x_k \) and \( z_k \). Finally, \( \{w_k\} \) is bounded since \( d(w_{k+1}, y_k) \leq c_k d(x_{k+1}, y_k) \) and \( \{c_k\} \) is bounded.

Now we proceed to show the statements of the theorem.

(1) The boundedness of iterates imply that \( c_k S_k \left( d(x_k, z_k) + 2d(w_{k+1}, y_k) \right) \) is uniformly bounded, so that \( \lambda_k \) has a uniform positive lower bound, which we denote by \( \lambda^* \).

Since \( \lambda_k \geq \lambda^* \) for all \( k \), we deduce that \( f(x_k) - f(x^*) \to 0 \) and \( d_{w_{k+1}}(z_k, x^*) \to 0 \) from Theorem 3.11. Since \( f \) is strongly convex, \( f(x_k) - f(x^*) \geq \frac{\mu}{2} d^2(x_k, x^*) \) and hence \( d(x_k, x^*) \to 0 \). On the other hand, since \( \{w_{k+1}\} \) is bounded, the distortion rate \( T_k(w_{k+1}, z_k) \) is also bounded according to Lemma 3.3. Thus \( d_{w_{k+1}}(z_k, x^*) \to 0 \) implies \( d(z_k, x^*) \to 0 \), as desired.

(2) The result of (1) implies that \( d(x_k, z_k) \to 0 \). Since \( D_k \to 0 \) by assumption, we have

\[
\lim_{k \to +\infty} c_k S_k \left( d(x_k, z_k) + 2D_k \right) = 0
\]

Hence by line 5 of Algorithm 2, \( \lambda_k = \lambda \) for sufficiently large \( k \).

(3) Define \( \xi_k = \frac{a_k}{\lambda_k} \). Note that the update of \( \theta_k \) and \( a_{k+1} \) in Algorithm 2 implies that

\[
\delta_k a_{k+1}^2 = 2 \lambda_k \left( \frac{\mu}{2} \delta_k A_k + B_k \right) a_{k+1} - 2 \lambda_k A_k B_k = 0
\]

Thus

\[
\delta_k^2 a_{k+1}^2 = 2 \lambda_k \left( B_k A_{k+1} + \frac{\mu}{2} \delta_k A_k a_{k+1} \right)
\]

\[
(1 + \mu \lambda_k) a_{k+1}^2 = 2 \delta_k^{-1} \lambda_k A_{k+1} \left( B_k + \frac{\mu}{2} \delta_k a_{k+1} \right) = 2 \lambda_k A_{k+1} B_{k+1}
\]

(3.11)

As a result, we have \( \frac{B_k}{\lambda_k} = \frac{1 - \mu \lambda_k}{2 \lambda_k - 1} \xi_k^2 \). We can now rewrite the equation \( B_{k+1} = \frac{B_k}{\lambda_k} + \frac{\mu}{2} a_{k+1} \) in terms of \( \xi \) as

\[
\delta_k \frac{1 + \mu \lambda_k}{2 \lambda_k} a_{k+1} = \frac{1 + \mu \lambda_k}{2 \lambda_k - 1} \xi_k (1 - \xi_k) + \frac{\mu}{2} \delta_k \xi_{k+1}
\]
For sufficiently large \( k \), we have \( \lambda_k = \lambda \), so that the above equation can be simplified as

\[
\delta_k \xi_{k+1}^2 = \xi_k^2(1 - \xi_{k+1}) + \frac{\mu \lambda}{1 + \mu \lambda} \delta_k \xi_{k+1}
\]  

(3.12)

We first show that for any \( \varepsilon > 0 \), we have

\[
\liminf_{k \to +\infty} \xi_k \geq (1 - \varepsilon) \frac{\sqrt{\mu \lambda}}{1 + \mu \lambda}
\]

Since \( d(w_{k+1}, y_k) \leq D_k \to 0 \) by assumption, and \( y_k \to x^* \), we have \( w_{k+1} \to x^* \). The definition of \( \delta_k \) then implies that \( \lim_{k \to +\infty} \delta_k = 1 \).

The recursive relation (3.12) can be rewritten as

\[
\delta_k \xi_{k+1} \left( \xi_{k+1} - \frac{\mu \lambda}{1 + \mu \lambda} \right) = \xi_k^2(1 - \xi_{k+1})
\]

Note that: if \( \delta_k \) becomes larger and \( \xi_k \) becomes smaller, then \( \xi_{k+1} \) also becomes smaller. Based on this observation, we first choose \( k_0 \) such that \( \delta_k \leq 1 + \varepsilon \sqrt{\frac{\mu \lambda}{1 + \mu \lambda}} \) for all \( k \geq k_0 \), and then construct a reference sequence \( \{\xi_k\}_{k \geq k_0} \) defined as

\[
\xi_{k_0} = \xi_{k_0}, \quad \delta_{k_0} \left( \xi_{k_0} - \frac{\mu \lambda}{1 + \mu \lambda} \right) = \xi_{k_0}^2(1 - \xi_{k_0}), \quad \delta = 1 + \varepsilon \sqrt{\frac{\mu \lambda}{1 + \mu \lambda}}
\]

Then we have \( \xi_k \geq \xi_{k_0} \) for all \( k \geq k_0 \). Alternatively, we can write the recursion above as \( \xi_{k+1} = \varphi(\xi_k) \), where

\[
\varphi(x) = \frac{1}{2\delta} \left( \frac{\mu \lambda}{1 + \mu \lambda} \delta - x^2 + \sqrt{x^2 - \frac{\mu \lambda}{1 + \mu \lambda} \delta} \right)
\]

We have

\[
\varphi'(x) = -\frac{x}{\delta} + \frac{x \left( x^2 - \frac{\mu \lambda}{1 + \mu \lambda} \delta \right) + 2\delta x}{\delta \sqrt{\left( x^2 - \frac{\mu \lambda}{1 + \mu \lambda} \delta \right)^2 + 4\delta x^2}}
\]

The observation made above implies that \( \varphi'(x) \geq 0 \). On the other hand,

\[
\varphi'(x) < 1 \iff \left( x \left( x^2 - \frac{\mu \lambda}{1 + \mu \lambda} \delta \right) + 2\delta x \right)^2 < (x + \delta)^2 \left( x^2 - \frac{\mu \lambda}{1 + \mu \lambda} \delta \right)^2 + 4\delta x^2
\]

\[
\Leftrightarrow 4\delta x^2 \left( x^2 - \frac{\mu \lambda}{1 + \mu \lambda} \delta \right)^2 + 4\delta^2 x^2 < (x + \delta)^2 \cdot 4\delta x^2
\]  

(3.13)

which trivially holds, since \( \delta > 1 \). Since \( \varphi \) is continuously differentiable, we have \( \sup_{x \in [0,1]} \varphi'(x) < 1 \) i.e. \( \varphi \) is a contraction mapping. Since \( \xi_k \in [0,1] \), \( \forall k \geq k_0 \), it converges exponentially fast to a fixed point of \( \varphi \), which is the positive root of the equation \( x^2 + (\delta - 1)x - \frac{\mu \lambda}{1 + \mu \lambda} \delta = 0 \). It’s easy to check that this root is larger than \( (1 - \varepsilon) \sqrt{\frac{\mu \lambda}{1 + \mu \lambda}} \), so that

\[
\liminf_{k \to +\infty} \xi_k \geq \liminf_{k \to +\infty} \xi_k \geq (1 - \varepsilon) \sqrt{\frac{\mu \lambda}{1 + \mu \lambda}}
\]
To prove the desired result, it remains show that
\[
\limsup_{k \to +\infty} \xi_k \leq \sqrt{\frac{\mu \lambda}{1 + \mu \lambda}}
\]
This can be similarly shown by constructing a reference sequence with \( \delta = 1 \) in the recursion, and the reference sequence converges to the fixed point corresponding to \( \delta = 1 \), which is \( \sqrt{\frac{\mu \lambda}{1 + \mu \lambda}} \). □

4 Accelerated first-order methods as special cases

The main objective of this section is to provide some first-order accelerated gradient methods that can be derived as special cases of the Riemannian A-HPE framework in Algorithm 2. Specifically, we show that for \( L \)-smooth \( \mu \)-strongly convex functions, these method eventually achieves acceleration. Our approach is based on verifying the assumptions in Theorem 3.12 with step size \( \lambda = O \left( \frac{1}{L} \right) \). To the best of our knowledge, the only existing work that proves result of this kind is [3] for a Riemannian version of Nesterov’s method [30]. Here, the Riemannian A-HPE framework leads to many other possibilities.

4.1 Acceleration without the additional distortion

In this subsection we discuss accelerated methods which does not incur the additional distortion issue highlighted in Lemma 3.5; that is, we have \( y_k = y_k' \). It then follows from Equation (3.8) that we directly obtain potential decrease, and we can choose constant step size \( \lambda_k = \lambda \) in Algorithm 2. We now discuss specific examples that fall into this category.

(i). Riemannian Nesterov’s method We choose \( \sigma_k = \sigma \), \( \lambda > 0 \) such that \( \lambda < \frac{\sigma^2}{\sqrt{1}} \), \( w_{k+1} = y_k \), \( x_{k+1} = \text{Exp}_{y_k}(-\lambda_k \nabla f(y_k)) \) and \( v_{k+1} = \nabla f(y_k) + \mu \text{Exp}_{y_k}^{-1}(x_{k+1}) \). This choice recovers the Nesterov’s accelerated method in the Euclidean setting, and is the same as its Riemannian counterpart proposed in [3] (although the choice of parameters may be slightly different). We summarize it in Algorithm 3 and we have the following convergence result.

**Corollary 4.1.** Suppose that \( \sigma_k = \sigma \leq \frac{3}{4} \) and \( \lambda_k = \lambda = \frac{\sigma^2}{\sqrt{1}} \), then Algorithm 3 is a special case of Algorithm 2 and eventually achieves acceleration.

**Proof.** It remains to check that the inequality Equation (3.2) holds. Indeed we have
\[
\text{LHS} = \frac{\lambda_k}{2(1 + \lambda_k \mu)} \left( f(x_{k+1}) - f(y_k) - \langle \text{Exp}_{y_k}^{-1}(x_{k+1}), \nabla f(y_k) \rangle \right)
+ \left( \frac{\lambda_k^2 \mu^2}{2(1 + \lambda_k \mu)^2} - \frac{\lambda_k \mu}{2(1 + \lambda_k \mu)} \right) d^2(y_k, x_{k+1})
\leq \frac{\lambda_k L}{2(1 + \lambda_k \mu)} d^2(y_k, x_{k+1}) \leq \frac{\sigma_k^2}{2(1 + \lambda_k \mu)^2} d^2(y_k, x_{k+1}) = \text{RHS}
\]
Hence Theorem 3.12 implies that \( \lim_{k \to +\infty} \frac{\Delta_{k+1}}{\Delta_k} = 1 + O \left( \sqrt{\mu/L} \right) \) i.e. the algorithm eventually achieves full acceleration. □

(ii). Riemannian Nesterov’s method with an additional gradient step Rather than taking one gradient descent step in the update rule of \( x_{k+1} \) in Algorithm 3, we may also consider taking multiple
Algorithm 3 Riemannian Nesterov’s Method

Input: Objective function $f$, initial point $x_0$, $\sigma \in (0, \frac{2}{\theta})$, parameters $L, \mu$, initial weight $A_0 \geq 0$

set $z_0 = x_0$ and $\lambda = \frac{\sigma^2}{\theta^2}$

for $k = 0, 1, \cdots$ do

choose a valid distortion rate $\delta_k$ according to Lemma 3.3

set $\delta_k$ to be the smaller root of $B_k(1 - \theta)^2 = \mu \lambda \theta ((1 - \theta)B_k + \frac{\sqrt{2}}{2} \delta_k A_k)$

set $B_{k+1} = \frac{B_k}{\delta_k}$

set $a_{k+1} = \frac{1 - \theta}{\delta_k} B_k$

set $A_{k+1} = A_k + a_{k+1}$

set $y_k = \text{Exp}_{y_k} \left( \frac{\delta_k A_k}{\delta_k A_k + 1} \text{Exp}_{x_k}^1(z_k) \right)$

set $x_{k+1} \leftarrow \text{Exp}_{y_k}(-\lambda \nabla f(y_k))$

set $z_{k+1} \leftarrow \text{Exp}_{y_k} \left( \theta_k \text{Exp}_{y_k}^{-1}(z_k) - \mu^{-1}(1 - \theta_k) \nabla f(y_k) \right)$

end for

Gradient steps. Specifically, we still choose $\sigma_k = \sigma = \frac{3}{\theta}$, $w_{k+1} = y_k, v_{k+1} = \nabla f(y_k) + \mu \text{Exp}_{y_k}^{-1}(x_{k+1})$, and we update $x_{k+1}$ as follows:

$$\tilde{x}_{k+1} = \text{Exp}_{x_{k+1}}(-\lambda_k \nabla f(x_{k+1})), \quad x_{k+1} = \text{Exp}_{x_{k+1}}(-\lambda_k \nabla f(\tilde{x}_{k+1}))$$

Now we need to determine the step size $\lambda_k$ to guarantee that the inexact proximal inequality Equation (2.1) is satisfied. We still have

$$f(x_{k+1}) - f(y_k) - \left\langle \text{Exp}_{y_k}^{-1}(x_{k+1}), \nabla f(y_k) \right\rangle \leq \frac{L}{2} d^2(x_{k+1}, y_k)$$

On the other hand, we can bound $||\text{Exp}_{y_k}^{-1}(x_{k+1}) + \lambda_k v_{k+1}||^2$ as follows:

$$||\text{Exp}_{y_k}^{-1}(x_{k+1}) + \lambda_k v_{k+1}||^2 = ||(1 + \mu \lambda_k) \text{Exp}_{y_k}^{-1}(x_{k+1}) + \lambda_k \nabla f(y_k)||^2$$

$$= (1 + \mu \lambda_k)^2 d^2(x_{k+1}, y_k) + 2\lambda_k(1 + \mu \lambda_k) \left\langle \text{Exp}_{y_k}^{-1}(x_{k+1}), \nabla f(y_k) \right\rangle + \lambda_k^2 \nabla f(y_k) ||^2$$

$$\leq (1 + \mu \lambda_k)^2 d^2(x_{k+1}, y_k) + \lambda_k^2 \nabla f(y_k) ||^2$$

$$- 2\lambda_k(1 + \mu \lambda_k) \left( f(y_k) - f(x_{k+1}) + \frac{\mu}{2} d^2(y_k, x_{k+1}) \right)$$

$$= (1 + \mu \lambda_k)^2 d^2(x_{k+1}, y_k) + \frac{L}{2} d^2(y_k, \tilde{x}_{k+1})$$

$$- 2\lambda_k(1 + \mu \lambda_k) \left( \frac{1}{\lambda_k} - \frac{L}{2} \right) \left( d^2(y_k, \tilde{x}_{k+1}) + d^2(\tilde{x}_{k+1}, x_{k+1}) \right)$$

$$\leq (1 + \mu \lambda_k)^2 d^2(x_{k+1}, y_k)$$

$$- 2 \left( (1 + \mu \lambda_k) \left( 1 - \frac{L}{2} \lambda_k \right) - \frac{1}{2} \right) \left( d^2(y_k, \tilde{x}_{k+1}) + d^2(\tilde{x}_{k+1}, x_{k+1}) \right)$$

$$\leq \left( \frac{L}{2} \lambda_k(1 + \mu \lambda_k) + \frac{1}{2} \right) d^2(x_{k+1}, y_k)$$

Thus it suffices to have $L \lambda_k(1 + \mu \lambda_k) \leq \frac{1}{4}$. We summarize the method we obtain in Algorithm 4. Similar to Algorithm 3, by choosing $w_{k+1} = y_k$ we avoid the distortion caused by the update rule of $y_k$. The above analysis implies the following result.
Corollary 4.2. Suppose that $\sigma_k = \sigma \leq \frac{3}{4}$ and $\lambda_k = O\left(\frac{1}{k}\right)$ such that $L\lambda_k(1 + \mu \lambda_k) \leq \frac{1}{4}$, then Algorithm 4 eventually achieves full acceleration.

Algorithm 4 can be seen as a Riemannian version of [16, Algorithm 3], in which the authors arrive at the algorithm using a discretization scheme of a continuous-time gradient flow. Similar methods based on this “multiple step” scheme is proposed and analyzed in recent work [23] for the Euclidean setting. In contrast, our approach is based on showing that this “multiple step” procedure is essentially a specific implementation of the inexact proximal operator, thereby providing a more transparent way to understand why such method would converge.

Algorithm 4 Riemannian Nesterov’s method with an extra gradient step

**Input:** Objective function $f$, initial point $x_0$, $\sigma \in (0, \frac{3}{4})$, parameters $L, \mu$, initial weight $A_0 \geq 0$

```
set $z_0 = x_0$ and $\lambda = \frac{\sigma^2}{L}$
for $k = 0, 1, \cdots$ do
    choose a valid distortion rate $\delta_k$ according to Lemma 3.3
    set $\theta_k$ to be the smaller root of $B_k(1-\theta)^2 = \mu \lambda \theta ((1-\theta)B_k + \frac{\mu}{\sigma^2} \delta_k A_k)$
    set $B_{k+1} = \frac{B_k}{\theta_k}$
    set $a_{k+1} = \frac{\mu}{\sigma^2} \delta_k B_k$
    set $A_{k+1} = A_k + a_{k+1}$
    set $x_k = \exp_{\tilde{x}_k} (-\lambda \nabla f(x_k))$
    set $y_k = \exp_{y_k} \left( \frac{\theta_k a_{k+1}}{A_{k+1}} \exp_{\tilde{x}_k}^{-1}(z_k) \right)$
    set $\tilde{x}_{k+1} = \exp_{y_k} (-\lambda \nabla f(y_k))$
    set $z_{k+1} = \exp_{y_k} \left( \theta_k \exp_{\tilde{x}_k}^{-1}(z_k) - \mu^{-1}(1-\theta_k) \nabla f(y_k) \right)$
end for
```

4.2 Acceleration with additional distortion

In this subsection we present a Riemannian accelerated extra-gradient descent method that requires us to deal with the additional distortion caused by $y_k$. This method can be seen as a Riemannian version of the accelerated extra-gradient method proposed by [18].

Specifically, we choose $\sigma_k = \sigma \leq \frac{3}{4}$, $\lambda \leq \frac{3}{4}$, $\nu = \nabla f(x_{k+1})$ and $w_{k+1} = x_{k+1} = \exp_{y_k} (-\lambda_k \nabla f(y_k))$. We can show the following result for these choices.

Corollary 4.3. Algorithm 5, obtained by the above choices of parameters, can be seen as a special case of Algorithm 2 with $c_k = 1$ and

$$D_k = \sqrt{\mu} \max \left\{ \| \nabla f(x_k) \|, \| \nabla f(z_k) \| \right\},$$

and it eventually achieves acceleration.

**Proof.** It is straightforward to check that $c_k = 1$. We also have $(x_{k+1}, v_{k+1}) \in \text{prox}_{\nu_k+1}^y(y_k, \lambda_k, \varepsilon_k)$, since

$$\left\| \exp_{x_k}^{-1}(y_k) - \lambda_k \nabla f(x_{k+1}) \right\| = \lambda_k \left\| \Gamma_{y_k}^{\nu_{k+1}} \nabla f(y_k) - \nabla f(x_{k+1}) \right\| \leq L \lambda_k d(y_k, x_{k+1}) \leq \sigma_k d(x_{k+1}, y_k).$$
Note that \( y'_k \) and \( y_k \) are possibly different in this case.

The choice of \( x_{k+1} \) implies that
\[
d(y_k, x_{k+1}) = \lambda_k \|\nabla f(y_k)\| \leq \frac{3}{4L} \|\nabla f(y_k)\|.
\]

Since \( y_k \) lies on the geodesic connecting \( x_k \) and \( x_k, f(y_k) \leq \max\{f(x_k), f(z_k)\} \). Hence by \( L \)-smoothness and \( \mu \)-strong convexity we have
\[
\|\nabla f(y_k)\|^2 \leq 2L(f(y_k) - f(x^*)) \leq 2L (\max\{f(x_k), f(z_k)\} - f(x^*))
\]
\[
\leq \frac{L}{\mu} \max \{\|\nabla f(x_k)\|^2, \|\nabla f(z_k)\|^2\}
\]

Thus
\[
d(y_k, x_{k+1}) \leq \frac{3}{4L} \|\nabla f(y_k)\| \leq \sqrt{L} \mu \max \{\|\nabla f(x_k)\|, \|\nabla f(z_k)\|\} = D_k,
\]
so that our choice of \( D_k \) serves as a valid upper bound of \( d(w_{k+1}, y_k) \) and hence ensures potential decrease. Since \( p_0 \geq p_k \geq A_k (f(x_k) - f(x^*)) \geq \frac{\mu}{2} A_k \|x_k - x^*\|^2 \), the iterates \( \{x_k\} \) are bounded. Thus \( \{w_k\} \) is also bounded. We choose \( r_k = \mu^{-1} \|\nabla f(x_{k+1})\| \), which is an upper bound for \( d(w_{k+1}, x^*) \). Also, \( r_k \leq \frac{\mu}{2} d(x_{k+1}, x^*) \) so that it is bounded, as required by Theorem 3.12.

By Theorem 3.12, all the iterates are uniformly bounded. It remains to check that \( D_k = o(1) \). But the boundedness of the iterates imply that the step size \( \lambda_k \) has a uniform positive lower bound, hence \( x_k \to x^* \) and \( z_k \to x^* \). Therefore, \( D_k = o(1) \) by its definition.

Hence, all the conditions in Theorem 3.12 hold, and the result follows. \( \square \)

**Algorithm 5 Riemannian accelerated extra-gradient descent**

**Input:** Objective function \( f \), initial point \( x_0 \), \( \sigma_k \in (0, \frac{3}{4}) \), parameters \( L, \mu \), initial weight \( A_0 \geq 0 \)

- set \( z_0 = x_0 \) and \( \lambda = \frac{\mu}{2} \)
- for \( k = 0, 1, \ldots \) do
  - choose a valid distortion rate \( \delta_k \) according to Lemma 3.3
  - set \( D_k = 3 \sqrt{L} \mu \max \{\|\nabla f(x_k)\|, \|\nabla f(z_k)\|\} \)
  - set \( \lambda_k = \min \{\lambda_k, \frac{1 - \sigma_k}{24 \mu \mu \mu_k^2 (d(x_k, z_k) + 2D_k)}\} \)
  - set \( \theta_k \) to be the smaller root of \( B_k (1 - \theta)^2 = \mu \lambda_k \theta ((1 - \theta) B_k + \frac{\mu}{2} \delta_k A_k) \)
  - set \( B_{k+1} = \frac{B_k}{\delta_k} \)
  - set \( a_{k+1} = \frac{1 - \theta_k}{\theta_k} B_k \)
  - set \( A_{k+1} = A_k + a_{k+1} \)
  - set \( y_k = \text{Exp}_{x_k} \left( \frac{\theta_k a_{k+1}}{A_k + \delta_k a_{k+1}} \text{Exp}_{x_k}^{-1}(z_k) \right) \)
  - set \( x_{k+1} = \text{Exp}_{y_k} \left( -\lambda_k \nabla f(y_k) \right) \)
  - choose \( r_k = \mu^{-1} \|\nabla f(x_{k+1})\| \) and let \( B_k = \{\|x\| \leq r_k\} \subset \mathbb{R}^d \)
  - set \( z_{k+1} = \text{Exp}_{x_{k+1}} \left( \mathcal{P}_{B_k} \left( \theta_k \text{Exp}_{x_{k+1}}^{-1}(z_k) - \mu^{-1} (1 - \theta_k) \nabla f(x_{k+1}) \right) \right) \)
- end for
4.3 Other possibilities

In previous subsections, we set \( w_{k+1} \) to be \( y_k \) and \( x_{k+1} \) which are the two most simple choices. This subsection is dedicated to the discussion of other choices that lead to new Riemannian accelerated methods.

For any fixed \( \tau \in [0, 1] \), we choose \( w_{k+1} = \text{Exp}_{x_k} \left( \tau \text{Exp}_{x_k}^{-1}(y_k) \right) \), \( v_{k+1} = -\mu \text{Exp}_{w_{k+1}}^{-1}(x_{k+1}) + \nabla f(w_{k+1}) \) and \( x_{k+1} \) be generated by some descent procedure (e.g. \( x_{k+1} = \text{Exp}_{y_k}(-\lambda_k \nabla f(y_k)) \)) similar to previous subsections. The quantities \( c_k \) and \( D_k \) can be chosen as in Section 4.2. It is not hard to check that these choices satisfy Equation (3.2), and similar to the arguments in Section 4.2, we can check that the conditions in Theorem 3.12 hold with \( \lambda = \mathcal{O} \left( \frac{1}{T} \right) \).

One may also choose \( w_{k+1} \) outside the geodesic between \( x_{k+1} \) and \( y_k \). Since \( f \) is \( L \)-smooth, we always have

\[
 f(x_{k+1}) - f(w_{k+1}) \leq \frac{L}{2} d^2(x_{k+1}, w_{k+1})
\]

so that the second term in Equation (3.2) is under control as long as we choose \( \lambda_k = \mathcal{O} \left( \frac{1}{T} \right) \) and \( d(x_{k+1}, w_{k+1}) = \mathcal{O} \left( d(x_{k+1}, y_k) \right) \). Note that the second requirement is already met since \( \{c_k\} \) (defined in Algorithm 2) is bounded. For the first term of Equation (3.2), we require that

\[
 \|\text{Exp}_{w_{k+1}}^{-1}(x_{k+1}) - \text{Exp}_{w_{k+1}}^{-1}(y_k) + \lambda v_{k+1}\|^2 \leq \rho \cdot d^2(x_{k+1}, y_k)
\]

for some \( \rho < 1 \). This is equivalent to

\[
 \|(1 - \mu \lambda_k)\text{Exp}_{w_{k+1}}^{-1}(x_{k+1}) - \text{Exp}_{w_{k+1}}^{-1}(y_k) + \lambda \nabla f(w_{k+1})\|^2 \leq \rho \cdot d^2(x_{k+1}, y_k)
\]

This allows a broader range for choosing \( x_{k+1} \) and \( w_{k+1} \). For example, consider \( w_{k+1} = \text{Exp}_{y_k}(-\lambda_k \nabla f(y_k)) \), then

\[
 \|\text{Exp}_{w_{k+1}}^{-1}(y_k) + \lambda \nabla f(w_{k+1})\| \leq L \lambda_k \cdot d(y_k, w_{k+1})
\]

Thus, any choices of \( x_{k+1} \) with \( d(w_{k+1}, x_{k+1}) + L \lambda_k d(y_k, w_{k+1}) \leq \sqrt{\rho} \cdot d(x_{k+1}, y_k) \) is valid. This can be further simplified using \( d(y_k, w_{k+1}) \leq d(y_k, x_{k+1}) + d(x_{k+1}, w_{k+1}) \), which yields the condition

\[
 d(w_{k+1}, x_{k+1}) \leq \frac{\sqrt{\rho} - L \lambda_k}{1 + L \lambda_k} d(y_k, x_{k+1})
\]

(4.2)

We briefly sketch how to check the conditions in Theorem 3.12. We assume WLOG that \( \rho < \frac{1}{4} \). The inequality Equation (3.9) implies that \( d_{w_{k+1}}(x_{k+1}, y_k) \) is bounded. By triangle inequality, \( d_{w_{k+1}}(x_{k+1}, y_k) \geq d(x_{k+1}, y_k) - 2d(w_{k+1}, x_{k+1}) \), so it follows from Equation (4.2) that \( d(x_{k+1}, y_k) \) is bounded. Moreover, we can choose \( c_k = \frac{1}{\sqrt{\rho}} \) in Algorithm 2. Since \( x_{k+1} \) is bounded by potential decrease, we have that \( y_k \) is bounded. Thus \( w_{k+1} = \text{Exp}_{y_k}(-\lambda_k \nabla f(y_k)) \) is bounded; so is \( r_k = \mu^{-1} \|\nabla f(w_{k+1})\| \). Finally the condition \( D_k = o(1) \) follows from the arguments in Section 4.2. We summarize our findings below.

**Corollary 4.4.** Suppose that in Algorithm 2 we choose \( \lambda_k = \mathcal{O} \left( \frac{1}{T} \right) \), \( w_{k+1} = \text{Exp}_{y_k}(-\lambda_k \nabla f(y_k)) \), and \( x_{k+1} \) such that \( d(w_{k+1}, x_{k+1}) \leq \frac{\sqrt{\rho} - L \lambda_k}{1 + L \lambda_k} d(y_k, x_{k+1}) \) for some given numerical constant \( 0 < \rho < \frac{1}{4} \), then all the conditions in Theorem 3.12 hold, and the algorithm eventually achieves acceleration.
5 Conclusion

In this paper, we propose a Riemannian version of the A-HPE framework and establish its global convergence. Our analysis is based on a novel viewpoint of A-HPE which may be of independent interest. Moreover, we show that the framework can lead to a wide range of different accelerated methods on Riemannian manifolds. In the Euclidean setting, the A-HPE framework also leads to optimal higher-order methods if combined with line-search, and we are looking forward to the study of these methods in the Riemannian setting.

References

1. P-A Absil, Robert Mahony, and Rodolphe Sepulchre. Optimization algorithms on matrix manifolds. Princeton University Press, 2009.
2. Kwangjun Ahn. From proximal point method to Nesterov’s acceleration. arXiv preprint arXiv:2005.08304, 2020.
3. Kwangjun Ahn and Suvrit Sra. From Nesterov’s estimate sequence to Riemannian acceleration. In Conference on Learning Theory, pages 84–118. PMLR, 2020.
4. Zeyuan Allen-Zhu and Lorenzo Orecchia. Linear coupling: An ultimate unification of gradient and mirror descent. arXiv preprint arXiv:1407.1537, 2014.
5. M Marques Alves. Variants of the A-HPE and large-step A-HPE algorithms for strongly convex problems with applications to accelerated high-order tensor methods. arXiv preprint arXiv:2102.02045, 2021.
6. Yossi Arjevani, Ohad Shamir, and Ron Shiff. Oracle complexity of second-order methods for smooth convex optimization. Mathematical Programming, 178(1):327–360, 2019.
7. Miroslav Bacák. Convex analysis and optimization in Hadamard spaces. de Gruyter, 2014.
8. Mathieu Barré, Adrien Taylor, and Francis Bach. A note on approximate accelerated forward-backward methods with absolute and relative errors, and possibly strongly convex objectives. arXiv preprint arXiv:2106.15536, 2021.
9. Amir Beck. First-order methods in optimization. SIAM, 2017.
10. Nicolas Boumal. An introduction to optimization on smooth manifolds.
11. Arne Brøndsted and Ralph Tyrrell Rockafellar. On the subdifferentiability of convex functions. Proceedings of the American Mathematical Society, 16(4):605–611, 1965.
12. Sébastien Bubeck, Qijia Jiang, Yin Tat Lee, Yuanzhi Li, and Aaron Sidford. Near-optimal method for highly smooth convex optimization. In Conference on Learning Theory, pages 492–507. PMLR, 2019.
13. Dmitri Burago, Iu D Burago, Yuri Burago, Sergei Ivanov, Sergei V Ivanov, and Sergei A Ivanov. A course in metric geometry, volume 33. American Mathematical Soc., 2001.
14. Peter Bürgisser, Cole Franks, Ankit Garg, Rafael Oliveira, Michael Walter, and Avi Wigderson. Towards a theory of non-commutative optimization: Geodesic 1st and 2nd order methods for moment maps and polytopes. In 2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS), pages 845–861. IEEE, 2019.
15. Yair Carmon, Arun Jambulapati, Qijia Jiang, Yujia Jin, Yin Tat Lee, Aaron Sidford, and Kevin
16. Long Chen and Hao Luo. First order optimization methods based on Hessian-driven Nesterov accelerated gradient flow. *arXiv preprint arXiv:1912.09276*, 2019.

17. Glaydston de Carvalho Bento, João Xavier da Cruz Neto, and Paulo Roberto Oliveira. A new approach to the proximal point method: convergence on general Riemannian manifolds. *Journal of Optimization Theory and Applications*, 168(3):743–755, 2016.

18. Jelena Diakonikolas and Lorenzo Orecchia. Accelerated extra-gradient descent: A novel accelerated first-order method. In *9th Innovations in Theoretical Computer Science Conference (ITCS 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.

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

20. Linus Hamilton and Ankur Moitra. No-go theorem for acceleration in the hyperbolic plane. *arXiv preprint arXiv:2101.05657*, 2021.

21. Reshad Hosseini and Suvrit Sra. Matrix manifold optimization for Gaussian mixtures. *Advances in Neural Information Processing Systems*, 28:910–918, 2015.

22. Jiang Hu, Andre Milzarek, Zaiwen Wen, and Yaxiang Yuan. Adaptive quadratically regularized newton method for Riemannian optimization. *SIAM Journal on Matrix Analysis and Applications*, 39(3):1181–1207, 2018.

23. Kevin Huang and Shuzhong Zhang. A unifying framework of accelerated first-order approach to strongly monotone variational inequalities. *arXiv preprint arXiv:2103.15270*, 2021.

24. Bo Jiang, Haoyue Wang, and Shuzhong Zhang. An optimal high-order tensor method for convex optimization. In *Conference on Learning Theory*, pages 1799–1801. PMLR, 2019.

25. Hiroyuki Kasai, Pratik Jawanpuria, and Bamdev Mishra. Riemannian adaptive stochastic gradient algorithms on matrix manifolds. In *International Conference on Machine Learning*, pages 3262–3271. PMLR, 2019.

26. John M Lee. *Riemannian manifolds: an introduction to curvature*, volume 176. Springer Science & Business Media, 2006.

27. Bernard Martinet. Regularisation, d’inéquations variationelles par approximations successives. *Revue Francaise d’informatique et de Recherche operationelle*, 1970.

28. Renato DC Monteiro and Benar Fux Svaiter. On the complexity of the hybrid proximal extragradient method for the iterates and the ergodic mean. *SIAM Journal on Optimization*, 20(6):2755–2787, 2010.

29. Renato DC Monteiro and Benar Fux Svaiter. An accelerated hybrid proximal extragradient method for convex optimization and its implications to second-order methods. *SIAM Journal on Optimization*, 23(2):1092–1125, 2013.

30. Yurii Nesterov et al. *Lectures on convex optimization*, volume 137. Springer, 2018.

31. Hiroyuki Sato, Hiroyuki Kasai, and Bamdev Mishra. Riemannian stochastic variance reduced gradient algorithm with retraction and vector transport. *SIAM Journal on Optimization*, 29(2):1444–1472, 2019.

32. Steven T Smith. Optimization techniques on Riemannian manifolds. *Fields institute communications*, 3(3):113–135, 1994.
33. Yue Sun, Nicolas Flammarion, and Maryam Fazel. Escaping from saddle points on Riemannian manifolds. *Advances in Neural Information Processing Systems*, 32:7276–7286, 2019.

34. Constantin Udriste. *Convex functions and optimization methods on Riemannian manifolds*, volume 297. Springer Science & Business Media, 2013.

35. Melanie Weber and Suvrit Sra. Nonconvex stochastic optimization on manifolds via Riemannian Frank-Wolfe methods. *arXiv preprint arXiv:1910.04194*, 2019.

36. Ami Wiesel. Geodesic convexity and covariance estimation. *IEEE transactions on signal processing*, 60(12):6182–6189, 2012.

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

38. Hongyi Zhang and Suvrit Sra. An estimate sequence for geodesically convex optimization. In *Conference On Learning Theory*, pages 1703–1723. PMLR, 2018.

39. Hongyi Zhang, Sashank J Reddi, and Suvrit Sra. Riemannian SVRG: Fast stochastic optimization on Riemannian manifolds. *Advances in Neural Information Processing Systems*, 29:4592–4600, 2016.