Proceedings of Machine Learning Research vol 178:1–39, 2022 35th Annual Conference on Learning Theory

Understanding Riemannian Acceleration via a Proximal Extragradient Framework

Jikai Jin JKJIN@PKU.EDU.CN
School of Mathematical Sciences, Peking University, Beijing, China

Suvrit Sra SUVRIT@MIT.EDU
LIDS, Massachusetts Institute of Technology, Cambridge, MA, USA

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

We contribute to advancing the understanding of Riemannian accelerated gradient methods. In particular, we revisit "*Accelerated Hybrid Proximal Extragradient*" (A-HPE), a powerful framework for obtaining Euclidean accelerated methods (Monteiro and Svaiter, 2013). Building on A-HPE, we then propose and analyze Riemannian A-HPE. The core of our analysis consists of two key components: (i) a set of new insights into Euclidean A-HPE itself; and (ii) a careful control of metric distortion caused by Riemannian geometry. We illustrate our framework by obtaining a few existing and new Riemannian accelerated gradient methods as special cases, while characterizing their acceleration as corollaries of our main results.

Keywords: geodesic convexity; Riemannian accelerated gradient; proximal extragradient

1. Introduction

Convexity admits an elegant generalization beyond vector spaces to geodesic metric spaces. There, through the lens of *geodesic convexity* one obtains a rich class of tractable nonconvex optimization problems, which makes the study of geodesically convex optimization potentially of far-reaching value. Main examples where geodesically convex optimization has been studied include certain Riemannian manifolds (Rapcsák, 1991; Udriste, 2013; Boumal, 2022; Sra and Hosseini, 2015; Wiesel, 2012), Hadamard spaces (Bacák, 2014), and non-commutative groups (Bürgisser et al., 2019).

The interest in geodesic convexity is paralleled by the development of optimization algorithms. Early works prove convergence for Riemannian proximal-point (Ferreira and Oliveira, 2002; de Carvalho Bento et al., 2016) and Riemannian analogs of many other Euclidean methods (Rapcsák, 1991; Smith, 1994; Absil et al., 2009), though these works in general do not exploit geodesic convexity, and limit their analyses to asymptotic results. The work (Zhang and Sra, 2016) is the first to provide non-asymptotic rates (and iteration complexity) of first-order methods for geodesically convex optimization on Hadamard manifolds. Subsequent works establish iteration complexities for other optimization methods on Riemannian manifolds, such as variance-reduced methods (Zhang et al., 2016; Sato et al., 2019), adaptive gradient methods (Kasai et al., 2019), Newton-type methods (Hu et al., 2018; Agarwal et al., 2021), among many others.

A key open question is whether it is possible to develop accelerated gradient methods on Riemannian manifolds. Zhang and Sra (2018) develop the first such method, and they show that the method achieves acceleration in a small neighborhood of the global minimum. Later, Ahn and Sra (2020) show that a Riemannian version of Nesterov's method converges globally, at a rate strictly faster than gradient descent and *eventually attains full acceleration*, which is defined as follows:

Definition 1 *We say that a gradient-based method eventually achieves acceleration if for optimizing an L-smooth, μ-geodesically strongly-convex function f, it outputs a sequence $\{w_k\}_{k\geq 1}$ with computational complexity $\mathcal{O}(k)$ that satisfies*

$$f(w_k) - f^* = \mathcal{O}\left((1-\tau_1)(1-\tau_2)\cdots(1-\tau_k)\right), \tag{1}$$

where $\tau_k \geq c\frac{\mu}{L}$ for some constant $c > 0$, and $\lim_{k\to+\infty} \tau_k = \Omega(\sqrt{\mu/L})$.

Motivation of this work. The abovementioned work leads one to wonder whether we can develop methods that attain full acceleration from the start, without a "burn in" period. Unfortunately, recent work (Hamilton and Moitra, 2021; Criscitiello and Boumal, 2021b) shows that full acceleration is impossible in general, which suggests that the best we can hope for is eventual acceleration.

Despite this recent progress on lower and upper bounds characterizing Riemannian acceleration, there is a considerable gap between the study of acceleration in Euclidean space versus Riemannian manifolds.[1] While numerous Euclidean accelerated methods beyond the canonical one of Nesterov have been studied, it is still unknown whether they also generalize to the Riemannian setting, and as such, a systematic understanding of Riemannian acceleration is still lacking.

This gap motivates us to study Riemannian acceleration more closely. We start by revisiting *Accelerated Hybrid Proximal Extragradient* (A-HPE), a powerful framework for convex optimization (Monteiro and Svaiter, 2013). Indeed, it can be shown that Nesterov's optimal method is a special case of A-HPE; Monteiro and Svaiter (2013) also propose a second-order method A-NPE, which is a specific implementation of their framework and has complexity $\widetilde{\mathcal{O}}\left(\varepsilon^{-2/7}\right)$ for ε-suboptimality. A hitherto unknown property of A-HPE is that it can recover a wide range of accelerated methods that have been independently proposed in past literature, e.g., the accelerated extragradient descent method of Diakonikolas and Orecchia (2018), the algorithm with an extra gradient descent step in (Chen and Luo, 2019, Section 4), the extra-point method of Huang and Zhang (2021), among others.

The A-HPE framework also has implications beyond usual first-order methods. A-NPE is used to design optimal second-order method in (Arjevani et al., 2019), and more generally, a number of works (Bubeck et al., 2019; Jiang et al., 2019) show that A-HPE can also induce optimal higher-order methods for smooth convex functions. Carmon et al. (2020) considers a different setting where one has access to a ball oracle, and they show that combining A-HPE with line search yields an accelerated method that is near-optimal. Moreover, A-HPE was extended to strongly-convex functions in (Barré et al., 2022; Marques Alves, 2022).

Overview and main contributions. In light of the above motivation, we believe that A-HPE can help us uncover 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, the key contributions of our work may be summarized as follows:

– We revisit Euclidean A-HPE in Section 2, and propose to view it as the linear coupling of two approximate proximal point iterates. This viewpoint produces a simple, new analysis of A-HPE.

– We introduce Riemannian A-HPE in Section 3, which we analyze by following our Euclidean approach, while localizing the challenges posed by Riemannian geometry. Specifically, we discover that besides the metric distortion that appears in previous works (Zhang and Sra, 2018; Ahn and Sra, 2020), there is an *additional distortion* that must be controlled.

1. We limit our discussion to the convex case, and refer the reader to the recent work (Criscitiello and Boumal, 2021a) that studies acceleration for non-geodesically-convex problems on manifolds.

Algorithm 1 Accelerated hybrid proximal extragradient (A-HPE) method

Input : Objective function f, initial point x_0, step size $\lambda_k > 0, k = 1, 2, \cdots$, a sequence $\{\sigma_k\}$ in $[0, 1]$, initial weight $A_0 \geq 0$

$z_0 = x_0$

for $k = 1, 2, \cdots$ **do**

$a_{k+1} \leftarrow \frac{(1+2\mu A_k)\lambda_{k+1} + \sqrt{(1+2\mu A_k)^2\lambda_{k+1}^2 + 4(1+\mu A_k)A_k\lambda_{k+1}}}{2}$

$A_{k+1} \leftarrow A_k + a_{k+1}$

$y_k \leftarrow x_k + \frac{a_{k+1}(1+\mu A_k)}{A_{k+1}+\mu(a_{k+1}A_k + A_k A_{k+1})}(z_k - x_k)$

$\varepsilon_k \leftarrow \frac{\sigma_k^2}{2(1+\lambda_k\mu)^2}\|x_{k+1} - y_k\|^2$

choose $(x_{k+1}, v_{k+1}) \in \texttt{iprox}_f(y_k, \lambda_k, \varepsilon_k)$

$z_{k+1} \leftarrow z_k + \frac{a_{k+1}}{1+\mu A_{k+1}}(\mu(x_{k+1} - z_k) - v_{k+1})$

end

– In Section 3.4, we first consider the case without additional distortion, for which we prove global eventual acceleration that not only generalizes (Ahn and Sra, 2020), but also offers global guarantees for some other Riemannian counterparts of Euclidean first-order accelerated methods.

– In Section 3.5, we then tackle the general case with additional distortion, which we handle by leveraging geometric bounds on Riemannian manifolds. For this general case, we obtain new local acceleration results akin to (Zhang and Sra, 2018).

– Finally, in Section 4, we discuss a number of accelerated first-order methods as special cases.

Notation and terminology. Throughout the paper $\langle \cdot, \cdot \rangle$ denotes inner product in a Euclidean space, and $\|\cdot\|$ its induced norm. For a closed convex set $\mathcal{X} \in \mathbb{R}^d$, we define the projection $\mathcal{P}_{\mathcal{X}}(x) := \operatorname{argmin}_{y\in\mathcal{X}} \|x - y\|$. For a convex function $f : \mathbb{R}^d \to \mathbb{R}$, the *proximal mapping* of f is given by $\texttt{prox}_f(x) := \arg\min_{u\in\mathbb{R}^d} f(u) + \frac{1}{2}\|u - x\|^2$. For a μ-strongly convex function $f : \mathbb{R}^d \to \mathbb{R}$, we define the 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. A New Analysis of Euclidean A-HPE

We now revisit the Euclidean A-HPE framework—see Algorithm 1. We propose to analyze A-HPE via the proximal point method, leading to a novel analysis that is simpler and more intuitive (in our opinion) than previous approaches (Barré et al., 2022; Marques Alves, 2022). More importantly, this analysis helps us develop Riemannian A-HPE, our main focus.

Throughout this section we assume that f is μ-strongly-convex. Our description follows (Barré et al., 2022) and relies on the key concept of *inexact proximal operators*. Our definition below is equivalent to the one in (Barré et al., 2022, Definition 2.3) that relies on the primal-dual gap of a proximal function. We use our version for ease of analysis. Appendix A provides additional intuition on this concept by relating it to ε-subgradients.

Definition 2 *(Barré et al., 2022, Lemma 2.4) We write* $(x, v) \in \texttt{iprox}_f(y, \lambda, \varepsilon)$ *if*

$$\frac{1}{2(1+\lambda\mu)^2}\|x - y + \lambda v\|^2 + \frac{\lambda}{1+\lambda\mu}\big(f(x) - f(w) - \langle x - w, v\rangle + \frac{\mu}{2}\|x - w\|^2\big) \leq \varepsilon, \qquad (2)$$

where $w \in \mathbb{R}^d$ *satisfies* $v - \mu x + \mu w \in \partial f(w)$ *and* $\varepsilon \geq 0$.

If $\varepsilon = 0$ and $w = x$, then $v \in \partial f(x)$ and $x + \lambda v = y$, which recovers the *exact* proximal operator.

With Definition 2 in hand, up to the specification of the sequences $\{\lambda_k\}$ and $\{\sigma_k\}$, all steps of Algorithm 1 are implementable. Hence, we are ready to state the main result of this section.

Theorem 3 *For the iterates produced by Algorithm 1, we have the function suboptimality bound*

$$f(x_k) - f(x^*) \leq \frac{A_0(f(x_0) - f(x^*)) + \frac{1+\mu A_0}{2}\|x_0 - x^*\|^2}{A_k} = \mathcal{O}\big(\Pi_{i=1}^{k-1}\big(1+\max\big\{\mu\lambda_i, \sqrt{\mu\lambda_i}\big\}\big)^{-1}\big).$$

Assuming that f is also L-smooth, a number of first-order methods (including Nesterov's method) can be considered as a special cases of Algorithm 1, with the choice $\lambda_i = \mathcal{O}(1/L)$. Theorem 3 then implies that these methods have the optimal convergence rate of $\mathcal{O}\big((1 + \sqrt{\mu/L})^{-k}\big)$. We do not present concrete examples here since this is not our main focus, but we will include detailed discussions about such special cases for the Riemannian setting later in the paper.

2.1. Overview of the proof of Theorem 3

We now overview our proof technique for Theorem 3, which sheds light on the specific parameter choices and updates that comprise Algorithm 1. To aid exposition, we trade simplicity for rigor in our overview below, and defer a fully rigorous proof to Appendix B.

Motivated by (Allen-Zhu and Orecchia, 2017; Ahn, 2020), we view Algorithm 1 as a combination of two approximate PPM (proximal point method) updates, each using a different notion of approximation. The first uses the inexact proximal operator from Theorem 2, while the second arises from minimizing a quadratic lower-approximation of f. When properly combined, these two steps allow us to prove the following theorem that immediately implies Theorem 3.

Theorem 4 *The potential function* $p_k := f(x_k) + \frac{1+\mu A_k}{2}\|z_k - x^*\|^2$ *is decreasing for all* $k \geq 0$.

The potential function in Theorem 4 has two terms: the objective $f(x_k)$ and a distance term involving $\|z_k - x^*\|^2$. We analyze these terms separately; they are associated with the two approximate PPM steps alluded to above, and the amount they change with k must be carefully combined to ensure $p_k \geq p_{k+1}$. We start with Theorem 5 to bound the change in function value.

Lemma 5 *Denote* $\nabla_{k+1} = v_{k+1} + \mu\,(w_{k+1} - x_{k+1}) \in \partial f(w_{k+1})$*; when* ε_k *is small, we have*

$$f(x_{k+1}) \lesssim f(w_{k+1}) + \tfrac{1}{2\mu}\left(\|v_{k+1}\|^2 - \|\nabla_{k+1}\|^2\right) \tag{3a}$$

$$\leq f(x_k) - \tfrac{\mu}{2}\|x_k - x_{k+1} + \mu^{-1}v_{k+1}\|^2 + \tfrac{1}{2\mu}\|v_{k+1}\|^2. \tag{3b}$$

The proof of Theorem 5 is given in Appendix B. Inequality (3b) is not exact; we omit an additional term that depends on ε_k for ease of presentation. Inequality (3) 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 the change in function values.

Next, we bound the change in the distance term. Observe that Line 8 of Algorithm 1 is nothing but $z_{k+1} \leftarrow \frac{1+\mu A_k}{1+\mu A_{k+1}} z_k + \frac{\mu a_{k+1}}{1+\mu A_{k+1}}\left(w_{k+1} - \mu^{-1}\nabla_{k+1}\right) = \operatorname{argmin}_z\{f_{w_{k+1}}(z) + \frac{1+\mu A_k}{2a_{k+1}}\|z - z_k\|^2\}$, where $f_{w_{k+1}}(z) := f(w_{k+1}) + \langle \nabla_{k+1}, z - w_{k+1}\rangle + \frac{\mu}{2}\|z - w_{k+1}\|^2$ is a lower quadratic approximation of f. Theorem 6 then helps us bound this change.

Lemma 6 (Approximate distance change) *When ε_k is small, we have*

$$\begin{aligned}
&\tfrac{1+\mu A_k}{2}\|z_k - x^*\|^2 - \tfrac{1+\mu A_{k+1}}{2}\|z_{k+1} - x^*\|^2 \\
&\geq a_{k+1}(f(w_{k+1}) - f(x^*)) + \tfrac{\mu a_{k+1}(1+\mu A_k)}{2(1+\mu A_{k+1})}\|z_k - w_{k+1} + \mu^{-1}\nabla_{k+1}\|^2 - \tfrac{a_{k+1}}{2\mu}\|\nabla_{k+1}\|^2 \quad \text{(4a)} \\
&\gtrsim a_{k+1}(f(x_{k+1}) - f(x^*)) + \tfrac{\mu a_{k+1}(1+\mu A_k)}{2(1+\mu A_{k+1})}\|z_k - x_{k+1} + \mu^{-1}v_{k+1}\|^2 - \tfrac{a_{k+1}}{2\mu}\|v_{k+1}\|^2 . \quad \text{(4b)}
\end{aligned}$$

Note that (4a) is very similar to the prox-grad inequality (Beck, 2017, Theorem 10.16) and the fundamental inequality of mirror descent (Allen-Zhu and Orecchia, 2017, Section 2.2) that imply *contraction of distance* with a proximal iteration. Inequality (4b) is not exact since it depends on ε_k. Again, the term $\|v_{k+1}\|$ prevents us from directly deducing contraction of the distance to x^*.

The inequalities in Theorem 5 and Theorem 6 reveal a challenge faced when proving descent of the potential function: we must control the magnitude of v_{k+1}. Specifically, consider the situation where the positive terms $\|x_k - x_{k+1} + \mu^{-1}v_{k+1}\|$ in (3b) and $\|z_k - x_{k+1} + \mu^{-1}v_{k+1}\|$ in (4b) are small but $\|v_{k+1}\|$ is large. But when ε_k is small, Theorem 2 also implies that $x_{k+1} - y_k \approx -\lambda_k v_{k+1}$, which further implies that $y_k - x_{k+1} + \mu^{-1}v_{k+1} \approx \left(\mu^{-1} + \lambda_k\right) v_{k+1}$ is large. This observation suggests that a contradiction is arrived at if we choose y_k on the line segment connecting x_k and z_k, i.e., $y_k = \tau x_k + (1-\tau)z_k$. Why? Since in this case, if $y_k - x_{k+1} + \mu^{-1}v_{k+1}$ is large, then we can directly deduce that (a convex combination) of the terms $\|x_k - x_{k+1} + \mu^{-1}v_{k+1}\|$ and $\|z_k - x_{k+1} + \mu^{-1}v_{k+1}\|$ is large using Cauchy-Schwarz. Remarkably, this argument suggests that we should choose τ such that $\tau : 1-\tau$ is equal to ratio of the coefficients of $\|x_k - x_{k+1} + \mu^{-1}v_{k+1}\|^2$ and $\|z_k - x_{k+1} + \mu^{-1}v_{k+1}\|^2$. Therefore we can use these terms to cancel out the error induced by v_{k+1}, and ultimately attain the desired potential function descent, leading to Theorem 4.

3. From Euclidean to Riemannian A-HPE

We are now ready to generalize Euclidean A-HPE to the Riemannian setting (more precisely, to Hadamard manifolds). In Section 3.1 we recall key notation for the Riemannian setting, and Section 3.2 is dedicated to the analysis of our proposed framework, Riemannian A-HPE.

3.1. Riemannian preliminaries and notation

We refer the readers to standard textbooks, e.g., (Lee, 2006; Jost, 2008) for an in-depth introduction; we recall below key notation and concepts.

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 $\texttt{Exp}_x$, 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* $\texttt{Exp}_x^{-1} : \mathcal{M} \to T_x\mathcal{M}$ is well-defined.

We use $d(x,y)$ to denote the *Riemannian distance* between x and y. The definition of exponential map implies that $d(x,y) = \left\|\texttt{Exp}_x^{-1}(y)\right\|$. We will also use the *tangent space distance*: $d_w(x,y) :=$

$\|\mathtt{Exp}_w^{-1}(x) - \mathtt{Exp}_w^{-1}(y)\|$. Note that $d_w(x,y) \leq d(x,y)$ for all $w,x,y \in \mathcal{M}$. We say that a function $f : \mathcal{M} \to \mathbb{R}$ is μ-geodesically-strongly-convex for $\mu \geq 0$, if for any $x \in \mathcal{M}$ there exists a non-empty set $\partial f(x)$, such that for all $y \in \mathcal{M}$ and $v \in \partial f(x)$ we have

$$f(y) \geq f(x) + \langle v, \mathtt{Exp}_x^{-1}(y)\rangle + \tfrac{\mu}{2} d^2(x,y).$$

Thus $f_x(y) := f(x) + \langle v, \mathtt{Exp}_x^{-1}(y)\rangle + \frac{\mu}{2} d^2(x,y)$ is a lower approximation of f.

We use Γ_x^y to denote the *parallel transport* from $T_x\mathcal{M}$ to $T_y\mathcal{M}$ along the geodesic connecting x and y. Using parallel transport, we can define a natural generalization of L-smoothness to the Riemannian setting. We say that $f : \mathcal{M} \to \mathbb{R}$ is *L-smooth* if for all $x, y \in \mathcal{M}$, we have

$$\|\Gamma_x^y \nabla f(x) - \nabla f(y)\| \leq L \cdot d(x,y). \tag{5}$$

3.2. The proposed Riemannian A-HPE framework

In this section, we first present a straightforward generalization of Euclidean A-HPE (Algorithm 1) to the Riemannian setting—see Algorithm 2. Then, we introduce a number results useful in its convergence analysis. Our presentation largely follows the Euclidean setting, except for a number of new challenges posed by Riemannian geometry. Throughout, we assume that f is μ-geodesically strongly convex, and that the sectional curvature of $\mathcal{M}$ lies in $[-K, 0]$.

Algorithm 2 Riemannian accelerated hybrid proximal extragradient method

Input : Objective function f, initial point x_0, 'reference' step size $\lambda > 0$, $\sigma_k \in (0,1)$ and initial weights $A_0, B_0 \geq 0$

$z_0 \leftarrow x_0$

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

choose a valid distortion rate δ_k according to Theorem 10

$\theta_k \leftarrow$ the smaller root of $B_k(1-\theta)^2 = \mu\lambda_k\theta\left((1-\theta)B_k + \frac{\mu}{2}\delta_k A_k\right)$

$B_{k+1} \leftarrow \frac{B_k}{\theta_k\delta_k}$, $a_{k+1} \leftarrow 2\mu^{-1}(1-\theta_k)B_{k+1}$ and $A_{k+1} \leftarrow A_k + a_{k+1}$

$y_k \leftarrow \mathtt{Exp}_{x_k}\left(\frac{\theta_k a_{k+1}}{A_k + \theta a_{k+1}} \mathtt{Exp}_{x_k}^{-1}(z_k)\right)$

choose $(x_{k+1}, v_{k+1}) \in \mathtt{iprox}_f^{w_{k+1}}(y_k, \lambda_k, \varepsilon_k)$ with $\varepsilon_k = \frac{\sigma_k^2}{2(1+\lambda_k\mu)^2} d^2_{w_{k+1}}(x_{k+1}, y_k)$

$z_{k+1} \leftarrow \mathtt{Exp}_{w_{k+1}}\left((1-\theta_k)\mathtt{Exp}_{w_{k+1}}^{-1}(x_{k+1}) + \theta_k \mathtt{Exp}_{w_{k+1}}^{-1}(z_k) - \frac{1-\theta_k}{\mu} v_{k+1}\right)$

end

Beyond the natural replacement of vector space operations with their Riemannian counterparts, there are two key differences between Algorithm 1 and Algorithm 2: (i) the latter uses a Riemannian version of the *iprox* operator; and (ii) it uses additional parameters (B_k and δ_k) in its updates. We define the Riemannian *iprox* operator as follows.

Definition 7 (Riemannian inexact proximal operator) *For $x, y, w \in \mathcal{M}$, $v \in T_w\mathcal{M}$ and $\lambda, \varepsilon \geq 0$, we write $(x,v) \in \mathtt{iprox}_f^w(y, \lambda, \varepsilon)$ if we have the inequality*

$$\frac{\|\mathtt{Exp}_w^{-1}(x) - \mathtt{Exp}_w^{-1}(y) + \lambda v\|^2}{2(1+\lambda\mu)^2} + \frac{\lambda\left(f(x) - f(w) - \left\langle \mathtt{Exp}_w^{-1}(x), v\right\rangle + \frac{\mu}{2} d^2(x,w)\right)}{1+\lambda\mu} \leq \varepsilon, \tag{6}$$

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

Key among the additional parameters is δ_k, the *distortion rate* that is used to model the non-linearity of the exponential map. The concept of distortion rate is not new, and was introduced in (Ahn and Sra, 2020) to analyze potential function decrease. We will formally define it in Theorem 9. By setting $\delta_k = 0$ and $B_k = \frac{1+\mu A_k}{2}$, Algorithm 2 recovers Algorithm 1 in the Euclidean setting. Before discussing technical details, let us give an informal statement of our main result. In subsequent sections, we sketch its proof and provide the formal statements, while full, rigorous proofs are deferred to Appendix C.

Theorem 8 (informal version of Theorem 13 and Theorem 18) *Under mild conditions on the choice of w_{k+1}, for μ-strongly convex and L-smooth function f, the following statements hold:*

*(1). Suppose w_{k+1} lies on the geodesic between x_k and z_k, then the iterates $\{x_k\}$ generated by Algorithm 2 eventually achieve acceleration (*cf. *Theorem 1) with arbitrary initialization.*

(2). In the general case, the iterates $\{x_k\}$ achieve acceleration as long as the initialization is in a $\mathcal{O}\big(K^{-1/2}\big(\mu/L\big)^{3/4}\big)$ neighbourhood of x^*.

As we will see later, the Riemannian analogs of Nesterov's method considered in (Zhang and Sra, 2018) and (Ahn and Sra, 2020) belong to the first case in Theorem 8, and thus, follow as corollaries of our main result. We will also discuss additional instances of each case in Section 4.

3.3. Potential Function Analysis for Riemannian A-HPE

Similar to the Euclidean setting, we define the potential function

$$p_k = A_k \cdot (f(x_k) - f(x^*)) + B_k \cdot d^2_{w_k}(z_k, x^*). \tag{7}$$

Note that in the above definiton, we use the tangent space distance d_{w_k} 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.

Our Euclidean analysis is based on a separate analysis of two approximate PPM schemes, one leading to Theorem 5 for function value, and the other leading to Theorem 6 for distance to x^*. While it is straightforward to generalize Theorem 5 to the Riemannian setting by using strong-convexity and Theorem 7, it is hard to bound the distance term $B_k \cdot d^2_{w_k}(z_k, x^*) - B_{k+1} \cdot d^2_{w_{k+1}}(z_{k+1}, x^*)$ because it involves vectors in two different tangent spaces $T_{w_k}\mathcal{M}$ and $T_{w_{k+1}}\mathcal{M}$. Taking cue from (Ahn and Sra, 2020), we also use the notion of *distortion rates* to overcome this issue.

Definition 9 *We say that $\delta_k > 0$ is a* valid distortion rate *if $d^2_{w_{k+1}}(z_k, x^*) \leq \delta_k d^2_{w_k}(z_k, x^*)$.*

To be able to use valid distortion rates in an actual algorithm, it is crucial to avoid dependence on the unknown optimal point x^*. To that end, the next lemma shows that one can obtain a valid distortion rate in terms of $d(w_k, z_k)$ instead.

Lemma 10 ((Ahn and Sra, 2020, Lemma 4.1)) *For any points $x, y, z \in \mathcal{M}$, we have $d^2(x, y) \leq T_K(d(x, z))d^2_z(x, y)$, where the function $T_K(\cdot)$ is defined as*

$$T_K(r) := \begin{cases} \max\big\{1 + 4\big(\frac{\sqrt{K}r}{\tanh(\sqrt{K}r)} - 1\big), \big(\frac{\sinh(2\sqrt{K}\cdot r)}{2\sqrt{K}\cdot r}\big)^2\big\}, & \text{if } r > 0, \\ 1, & \text{if } r = 0. \end{cases}$$

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

Assuming access to valid distortion rates, we can obtain the Riemannian analog to Theorem 6.

Lemma 11 *Suppose that $\delta_k > 0$ is a valid distortion rate, and $B_{k+1} = \frac{B_k}{\theta_k \delta_k}$, then*

$$B_k d^2_{w_k}(z_k, x^*) - B_{k+1} d^2_{w_{k+1}}(z_{k+1}, x^*) \geq (1-\theta_k) B_{k+1} \big(\tfrac{2}{\mu}(f(w_{k+1}) - f(x^*)) - \tfrac{1}{\mu^2}\|\nabla_{k+1}\|^2 + \theta_k \big\| \mathtt{Exp}^{-1}_{w_{k+1}}(z_k) - \mathtt{Exp}^{-1}_{w_{k+1}}(x_{k+1}) + \mu^{-1} v_{k+1} \big\|^2 \big).$$

Now, it remains to combine the two PPM schemes (based on function value and distance to x^*) to obtain a bound for the potential function. Specifically, we can prove the following key lemma.

Lemma 12 *Let $a_{k+1} = A_{k+1} - A_k = \frac{2}{\mu}(1-\theta_k) B_{k+1}$, and p_k be given by (7). Then,*

$$\begin{aligned} p_k - p_{k+1} \geq{}& \tfrac{\mu}{2}(\theta_k a_{k+1} + A_k) \big\| \mathtt{Exp}^{-1}_{w_{k+1}}(y'_k) - \mathtt{Exp}^{-1}_{w_{k+1}}(x_{k+1}) + \mu^{-1} v_{k+1} \big\|^2 \\ &+ \tfrac{A_{k+1}}{2\lambda_k \sigma_k} \big\| \mathtt{Exp}^{-1}_{w_{k+1}}(y_k) - \mathtt{Exp}^{-1}_{w_{k+1}}(x_{k+1}) - \lambda_k v_{k+1} \big\|^2 \\ &+ \tfrac{\mu \theta_k a_{k+1} A_k}{2(A_k + \theta_k a_{k+1})} d^2_{w_{k+1}}(x_k, z_k) - \tfrac{\sigma_k A_{k+1}}{2\lambda_k} d^2_{w_{k+1}}(x_{k+1}, y_k) - \tfrac{A_{k+1}}{2\mu}\|v_{k+1}\|^2, \end{aligned} \tag{8}$$

where

$$y'_k = \mathtt{Exp}_{w_{k+1}} \big(\tfrac{A_k}{A_k + \theta_k a_{k+1}} \mathtt{Exp}^{-1}_{w_{k+1}}(x_k) + \tfrac{\theta_k a_{k+1}}{A_k + \theta_k a_{k+1}} \mathtt{Exp}^{-1}_{w_{k+1}}(z_k) \big). \tag{9}$$

The main feature of Theorem 12 is the presence of a new point y'_k in (8). In the Euclidean setting, we combined the two approximate PPM schemes by choosing y_k on the line segment between x_k and z_k. Generalizing this update rule to the Riemannian setting, we naturally choose y_k on the geodesic connecting x_k and z_k. However, here a subtle complication arises: since we are working with vectors in the tangent space $\mathcal{T}_{w_{k+1}}$, what we really want is that $\mathtt{Exp}^{-1}_{w_{k+1}}(y_k)$ be a convex combination of $\mathtt{Exp}^{-1}_{w_{k+1}}(x_k)$ and $\mathtt{Exp}^{-1}_{w_{k+1}}(z_k)$. This subtlety explains why y'_k as defined by (9) appears in the bound (8). Further, note that since y'_k depends on w_{k+1}, it *cannot* be used as the update rule of y_k.

In Euclidean A-HPE, y'_k does not complicate matters since we always have $y_k = y'_k$. However, $y_k \neq y'_k$ in general for the Riemannian setting, which prevents us from mimicking the Euclidean analysis. Indeed, Theorem 12 highlights an *additional distortion* that arises for Riemannian A-HPE and is not present in previous works that focus on Nesterov's method (Zhang and Sra, 2018; Ahn and Sra, 2020). Indeed, the algorithms analyzed in these previous works are a special case of the general A-HPE framework, where the particular specialization of the updates bypasses the additional distortion that arises more generally. We expand on these observations below.

3.4. Basic A-HPE: convergence without additional distortion

We first consider the special case where $y_k = y'_k$. This equality holds as long as w_{k+1} is chosen on the geodesic connecting x_k and z_k. In this case, we can derive potential decrease from Theorem 12 by using an analysis similar to the Euclidean setting. Doing so, we obtain the following main result regarding the convergence rate of Riemannian A-HPE.

Theorem 13 *Suppose in Algorithm 2, we choose $\lambda_k = \lambda$ and w_{k+1} lies on the geodesic connecting x_k and z_k such that $d(w_{k+1}, y_k) = \mathcal{O}(1)$, then we have*

$$f(x_k) - f(x^*) \leq p_0 / A_k, \quad \text{and} \quad \lim_{k \to +\infty} A_{k+1}/A_k = 1 + \mu\lambda + \sqrt{\mu\lambda(1+\mu\lambda)}. \tag{10}$$

***Proof sketch*:** The first inequality follows directly from potential decrease. Define $\xi_k = {}^{a_k}/_{A_k}$, then it suffices to show that $\lim_{k\to+\infty} \xi_k = \sqrt{\frac{\mu\lambda}{1+\mu\lambda}}$. The proof relies on the following recursive equation:

$$\delta_k \xi_{k+1}\left(\xi_{k+1} - {}^{\mu\lambda}/_{1+\mu\lambda}\right) = \xi_k^2(1-\xi_{k+1}). \tag{11}$$

If $\delta_k = 1$, then we can show that $\{\xi_k\}$ converges to the fixed point of (11), which is $\sqrt{\frac{\mu\lambda}{1+\mu\lambda}}$. In our setting, δ_k is not constant, but potential decrease implies that $\delta_k - 1$ converges to 0 at a linear rate. Therefore, we can still obtain the desired result. □

The assumption $d(w_{k+1}, y_k) = \mathcal{O}(1)$ ensures that the iterates of Algorithm 2 are uniformly bounded; otherwise the distortion error can become arbitrarily large. This assumption is trivially true when $w_{k+1} = y_k$, which holds for a number first-order methods that we will discuss in Section 4. Theorem 13 also immediately implies the following result, which plays a crucial role when studying first-order methods as special cases of Algorithm 2.

Corollary 14 *Suppose that f is L-smooth and $\lambda_k = \lambda = \Theta(1/L)$. Under the conditions in Theorem 13, Algorithm 2 eventually achieves acceleration.*

Finally, we bound the number of iterations sufficient for achieving full acceleration.

Theorem 15 *Under the assumptions in Theorem 13, if f is L-smooth and $\lambda_k = \lambda = \mathcal{O}({}^1/_L)$, then we have $\xi_k \geq \frac{1}{2}\sqrt{\frac{\mu\lambda}{1+\mu\lambda}}$ after $T = \widetilde{\mathcal{O}}({}^L/_\mu)$ iterations, where $\widetilde{\mathcal{O}}$ hides logarithmic terms. As a result, Algorithm 2 achieves acceleration in at most $\widetilde{\mathcal{O}}({}^L/_\mu)$ iterations.*

3.5. The general case of A-HPE: handling additional distortion

In general, we do not have $y_k = y_k'$, so that an *additional distortion* appears in the analysis of Riemannian A-HPE. To overcome the challenge posed by this distortion, we take an approach that is based on deriving an upper bound for the tangent space distance between y_k and y_k'. We need the following lemma, which is a variant of (Sun et al., 2019, Section B.3).

Lemma 16 *(Sun et al., 2019, Lemma 3) Let $x \in \mathcal{M}$ and $y, a \in T_x\mathcal{M}$. Let $z = \mathtt{Exp}_x(a)$, then*

$$d\left(\mathtt{Exp}_x(y+a), \mathtt{Exp}_z\left(\Gamma_x^z y\right)\right) \leq \min\{\|a\|, \|y\|\} S_K(\|a\| + \|y\|),$$

where $S_K(r) = \cosh(\sqrt{K}r) - {}^{\sinh(\sqrt{K}r)}/_{\sqrt{K}r}$.

Note that a key feature of the function S_K is that $\lim_{r\to 0} S_K(r) = 0$. By using Theorem 16, we can obtain an upper bound on $d_{w_{k+1}}(y_k, y_k')$ in terms of $S_K(\cdot)$ and a distance term.

Lemma 17 *We have for all $k \geq 1$ that*

$$d_{w_{k+1}}(y_k, y_k') \leq 2d^*(w_{k+1}; x_k, z_k) \cdot S_K\left(d(x_k, z_k) + d^*(w_{k+1}; x_k, z_k)\right), \tag{12}$$

where $d^*(w; x, z) := \min\left\{d(w, y) \mid y = \mathtt{Exp}_x(t \cdot \mathtt{Exp}_x^{-1}(z)), t \in [0,1]\right\}$ is the distance from w to the geodesic connecting x and z.

When $d^2_{w_{k+1}}(y_k, y'_k)$ is small, we can imagine that the algorithm still behaves similar to the $y_k = y'_k$ case studied in Section 3.4. From a technical standpoint, to lower-bound the potential difference $p_k - p_{k+1}$, the key difference between the Riemannian setting with the Euclidean setting is the presence of an additional negative term that depends on $d^2_{w_{k+1}}(y_k, y'_k)$. As a result, potential decrease can still be guaranteed if the RHS of (12) is smaller than the positive terms. This yields our main result for the convergence of Algorithm 2 in the general case, as stated below.

Theorem 18 (informal) *Suppose that f is L-smooth, $\sigma_k = \sigma \in (0,1)$ and $\lambda_k = \lambda = \mathcal{O}(1/L)$. Under regularity conditions on the choice of w_{k+1}, if the initialization satisfies $d(x_0, x^*) = \mathcal{O}\big(K^{-1/2}(\mu/L)^{3/4}\big)$ and $B_0 = \frac{\mu}{2}A_0 > 0$, then potential decrease holds, and $\xi_k := \frac{a_k}{A_k} = \Theta\big(\sqrt{\frac{\mu\lambda}{1+\mu\lambda}}\big)$.*

***Proof sketch*:** The proof is by induction on k. When $k = 0$, by using regularity conditions on w_1, we can derive an upper bound for the RHS of (12), which implies potential decrease. Now suppose that potential decrease holds for k. By the definition of the potential function p_k, we can show that $d^2(x_k, x^*)$ and $d^2(z_k, x^*) = \mathcal{O}(K^{-1}(\mu/L)^{1/2})$.

Note that the distortion rate $\delta_k \le 1 + \mathcal{O}\left(Kd^2(w_k, z_k)\right)$, and under regularity conditions on w_k we can bound δ_k by $1 + \mathcal{O}(\sqrt{\mu/L})$; ξ_{k+1} can then be lower-bounded using the recursive equation (35) and the lower bound for ξ_k. Finally, the RHS of (12) can be directly upper-bounded using the bounds for x_k, z_k and regularity conditions on w_{k+1}, which implies potential decrease for $k+1$. □

The regularity conditions on the sequence $\{w_k\}$ are described formally in Theorem 41, and they play a crucial role in Theorem 18. In short, they require that $\{w_k\}$ is not too far away from the sequence $\{x_k\}$ and $\{z_k\}$, since otherwise the algorithm may suffer from large distortion error.

Theorem 18 implies that as long as the initialization is inside a $\mathcal{O}(K^{-1/2}(\mu/L)^{3/4})$ neighbourhood of the global minimum x^*, then it can achieve the accelerated rate.

Corollary 19 *Under the assumptions of Theorem 18, we have*

$$f(x_k) - f(x^*) \le c_1 K^{-1} L(\mu/L)^{\frac{3}{2}} \cdot \left(1 - c_2\sqrt{\mu/L}\right)^k,$$

for some numerical constants $c_1, c_2 > 0$.

4. Special cases of Riemannian A-HPE: acceleration of several first-order methods

Inspired by Nesterov's method, a number of different accelerated methods have been proposed in the Euclidean setting (Diakonikolas and Orecchia, 2018; Chen and Luo, 2019; Huang and Zhang, 2021). These methods are empirically observed to be superior in some aspects (e.g., robustness to noise, possibly smaller constants in convergence bounds, etc.) However, they are derived using a variety of very different techniques, which obscures their common origin. In contrast, we observe that all of them can be deduced from A-HPE quite naturally and straightforwardly.

At the same time, in the Riemannian setting only a generalized version of Nesterov's method is known to achieve acceleration (Zhang and Sra, 2018; Ahn and Sra, 2020). Can we design other accelerated methods, similar to those in the Euclidean setting? The answer is "yes," and we discuss below several special cases obtained from our Riemannian A-HPE framework. We divide these special cases into two categories: (i) those without additional distortion (Section 3.4), and which eventually attain acceleration with arbitrary initialization due to Theorem 13; and (ii) those that can suffer additional distortion studied in Section 3.5, for which local acceleration is ensured by Theorem 18. Detailed derivations of the methods studied in this section are given in Appendix D.

4.1. Accelerated methods without additional distortion

Riemannian Nesterov's method. Nesterov's method has a direct generalization to the Riemannian setting, as proposed and analyzed in (Zhang and Sra, 2018; Ahn and Sra, 2020); it takes the form:

$$\begin{aligned} y_k &= \mathtt{Exp}_{x_k}\big(\tfrac{\theta_k a_{k+1}}{A_k+\theta a_{k+1}}\mathtt{Exp}_{x_k}^{-1}(z_k)\big),\\ x_{k+1} &= \mathtt{Exp}_{y_k}(-\lambda\nabla f(y_k)),\\ z_{k+1} &= \mathtt{Exp}_{y_k}(\theta_k\mathtt{Exp}_{y_k}^{-1}(z_k) - \mu^{-1}(1-\theta_k)\nabla f(y_k)). \end{aligned} \tag{13}$$

We can derive this algorithm from Algorithm 2 by choosing $w_{k+1} = y_k$, $x_{k+1} = \mathtt{Exp}_{y_k}(-\lambda_k\nabla f(y_k))$ and $v_{k+1} = \nabla f(y_k) + \mu\mathtt{Exp}_{y_k}^{-1}(x_{k+1})$. Additional distortion is not present since $w_{k+1} = y_k$. We also recover the result of Ahn and Sra (2020) that (13) can eventually achieve acceleration; the local acceleration result of Zhang and Sra (2018) can also be directly deduced from Theorem 18.

Riemannian Nesterov's method with multiple gradient steps. We can also perform multiple gradient descent (GD) steps from y_k to obtain x_{k+1}. Chen and Luo (2019, Algorithm 3) present a method of this type in the Euclidean setting. Here we consider a Riemannian version of their method:

$$\begin{aligned} y_k &= \mathtt{Exp}_{x_k}\big(\tfrac{\theta_k a_{k+1}}{A_k+\theta a_{k+1}}\mathtt{Exp}_{x_k}^{-1}(z_k)\big),\\ \tilde{x}_{k+1} &= \mathtt{Exp}_{y_k}(-\lambda\nabla f(y_k)),\\ x_{k+1} &= \mathtt{Exp}_{\tilde{x}_{k+1}}(-\lambda\nabla f(\tilde{x}_{k+1})),\\ z_{k+1} &= \mathtt{Exp}_{y_k}\big(\theta_k\mathtt{Exp}_{y_k}^{-1}(z_k) - \mu^{-1}(1-\theta_k)\nabla f(y_k)\big). \end{aligned} \tag{14}$$

Method (14) can be derived from Algorithm 2 by choosing x_{k+1} as the result of two GD steps; the other variables the same as Riemannian Nesterov's method.

4.2. Accelerated methods with additional distortion

Riemannian accelerated extra-gradient descent (RAXGD). We consider a Riemannian version of the accelerated extra-gradient method (AXGD) proposed by Diakonikolas and Orecchia (2018):

$$\begin{aligned} y_k &= \mathtt{Exp}_{x_k}\big(\tfrac{\theta_k a_{k+1}}{A_k+\theta a_{k+1}}\mathtt{Exp}_{x_k}^{-1}(z_k)\big),\\ x_{k+1} &= \mathtt{Exp}_{y_k}(-\lambda\nabla f(y_k)),\\ z_{k+1} &= \mathtt{Exp}_{x_{k+1}}\big(\theta_k\mathtt{Exp}_{x_{k+1}}^{-1}(z_k) - \mu^{-1}(1-\theta_k)\nabla f(x_{k+1})\big). \end{aligned} \tag{15}$$

Method (15) can be recovered from Algorithm 2 by choosing $v = \nabla f(x_{k+1})$, and $w_{k+1} = x_{k+1} = \mathtt{Exp}_{y_k}(-\lambda_k\nabla f(y_k))$. While Diakonikolas and Orecchia (2018) obtain AXGD via a specifically chosen discretization of suitable continuous-time dynamics, we observe that (R)AXGD can be deduced from A-HPE quite straightforwardly.

We can also derive from Algorithm 2 a generalized version of RAXGD; please refer to Appendix D for more details and discussions.

The extra-point framework of Huang and Zhang (2021). Recently, a general framework was proposed by Huang and Zhang (2021) for obtaining accelerated methods in the Euclidean setting (*cf.* eq., (26) therein). We observe that their framework has a natural interpretation via the PPM viewpoint discussed in Section 2, though upon using a less general version of update rules compared with

A-HPE. A detailed comparison between their framework and A-HPE is provided in Appendix D.3, where we also present a Riemannian generalization of their algorithm. Using our approach of analyzing Riemannian A-HPE, local acceleration can be shown for the resulting algorithm, while for a special case (corresponding to the algorithm described in (Huang and Zhang, 2021, eq.(38))), global eventual acceleration can also be achieved.

5. Conclusion and future directions

In this paper, we propose an alternative viewpoint of the Euclidean A-HPE framework of (Monteiro and Svaiter, 2013) via the proximal point method. This viewpoint allows us to derive a simple and novel convergence analysis of A-HPE; it also plays a pivotal role in obtaining Algorithm 2, our proposed generalization of A-HPE to the Riemannian setting. While most of our Euclidean proof generalizes to the Riemannian setting, there is an additional distortion caused by the non-linearity of the exponential map that we must overcome; we model this distortion by leveraging geometric tools to complete the convergence analysis. Our main results include local acceleration of Riemannian A-HPE in its most general form, which we sharpen to global (eventual) acceleration whenever additional distortion is not present. We demonstrate the generality of our framework by discussing several accelerated first-order methods as special cases, recovering the recent results (Zhang and Sra, 2018; Ahn and Sra, 2020) as special cases, obtaining Riemannian counterparts of other accelerated (Euclidean) algorithms, and deriving new algorithms from our framework.

An aspect more basic worth noting is that this work also contributes toward a more thorough understanding of accelerated methods on Riemannian manifolds. Even on Euclidean spaces, our PPM-based approach may be of independent interest, since it provides a unified way for analyzing several accelerated methods that have been proposed in the literature and analyzed using a number of different techniques. Nonetheless, there are some important questions that remain unanswered.

First, we only show local convergence in the general case where additional distortion arises. It is unclear whether Riemannian A-HPE can indeed fail to converge in some cases, or whether the locality restriction is a shortcoming of our analysis. Nevertheless, we believe that some regularization conditions on the specification of the *iprox* operator (e.g., the conditions in Theorem 41) are necessary, since large distortion error would unavoidably impact the rate of convergence.

Second, in this paper we focus on accelerated first-order methods for strongly-convex functions on non-positively curved manifolds. The main challenge of the convex setting is that the effect of metric distortion would not asymptotically vanish as in the strong-convex setting. For manifolds with positive curvature, it is necessary to restrict the iterates inside a convex set, for example by using projection operators, but this may hurt the analysis of acceleration. Also, as discussed in Section 1, the A-HPE framework can also lead to optimal higher-order methods in Euclidean setting. However, to the best of our knowledge, optimal higher-order methods and their convergence rates are not known in the Riemannian setting. It may be useful (and feasible) to design such methods based on the Riemannian A-HPE framework introduced in this paper.

Finally, a broader goal in the study of acceleration is to develop theory and algorithms for non-Euclidean settings beyond those offered by Riemannian geometry.

Acknowledgments

We thank Kwangjun Ahn (MIT) and Xiang Cheng (MIT) for valuable discussions, and thank Shuailing Feng for pointing out a mistake in an initial version of this paper. Jikai Jin is is partially supported by the elite undergraduate training program of School of Mathematical Sciences in Peking University. SS acknowledges support from an NSF CAREER Grant (1846088).

References

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

Naman Agarwal, Nicolas Boumal, Brian Bullins, and Coralia Cartis. Adaptive regularization with cubics on manifolds. *Mathematical Programming*, 188(1):85–134, 2021.

Kwangjun Ahn. From proximal point method to Nesterov's acceleration. *arXiv preprint arXiv:2005.08304*, 2020.

Kwangjun Ahn and Suvrit Sra. From Nesterov's estimate sequence to Riemannian acceleration. In *Conference on Learning Theory*, pages 84–118. PMLR, 2020.

Zeyuan Allen-Zhu and Lorenzo Orecchia. Linear coupling: An ultimate unification of gradient and mirror descent. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2017.

Yossi Arjevani, Ohad Shamir, and Ron Shiff. Oracle complexity of second-order methods for smooth convex optimization. *Mathematical Programming*, 178(1):327–360, 2019.

Miroslav Bacák. *Convex analysis and optimization in Hadamard spaces*. de Gruyter, 2014.

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.

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. *Open Journal of Mathematical Optimization*, 3:1–15, 2022.

Amir Beck. *First-order methods in optimization*. SIAM, 2017.

Nicolas Boumal. An introduction to optimization on smooth manifolds. To appear with Cambridge University Press, Jan 2022. URL http://www.nicolasboumal.net/book.

Arne Brøndsted and Ralph Tyrrell Rockafellar. On the subdifferentiability of convex functions. *Proceedings of the American Mathematical Society*, 16(4):605–611, 1965.

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.

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.

Yair Carmon, Arun Jambulapati, Qijia Jiang, Yujia Jin, Yin Tat Lee, Aaron Sidford, and Kevin Tian. Acceleration with a ball optimization oracle. *Advances in Neural Information Processing Systems*, 33, 2020.

Long Chen and Hao Luo. First order optimization methods based on Hessian-driven Nesterov accelerated gradient flow. *arXiv preprint arXiv:1912.09276*, 2019.

Christopher Criscitiello and Nicolas Boumal. An accelerated first-order method for non-convex optimization on manifolds. *arXiv:2008.02252*, 2021a.

Christopher Criscitiello and Nicolas Boumal. Negative curvature obstructs acceleration for geodesically convex optimization, even with exact first-order oracles. *arXiv preprint arXiv:2111.13263*, 2021b.

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.

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.

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

Linus Hamilton and Ankur Moitra. A no-go theorem for robust acceleration in the hyperbolic plane. *Advances in Neural Information Processing Systems*, 34, 2021.

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.

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

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.

Jürgen Jost. *Riemannian geometry and geometric analysis*. Springer, seventh edition, 2008.

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.

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

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. *Optimization Methods and Software*, pages 1–31, 2022.

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.

Tamás Rapcsák. Geodesic convexity in nonlinear optimization. *Journal of Optimization Theory and Applications*, 69(1):169–183, 1991.

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.

Steven T Smith. Optimization techniques on Riemannian manifolds. *Fields institute communications*, 3(3):113–135, 1994.

Suvrit Sra and Reshad Hosseini. Conic geometric optimization on the manifold of positive definite matrices. *SIAM Journal on Optimization*, 25(1):713–739, 2015.

Yue Sun, Nicolas Flammarion, and Maryam Fazel. Escaping from saddle points on Riemannian manifolds. *Advances in Neural Information Processing Systems*, 32:7276–7286, 2019.

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

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

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

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

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.

Appendix A. Connection between *iprox* and ε-subgradient

In this section, we show the equivalence between the *iprox* operator (cf. Theorem 2) and the notion of ε-subdifferential (Brøndsted and Rockafellar, 1965, Section 3).

Definition 20 *Suppose that $h : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex and $x \in \mathbb{R}^d$. We say that $u \in \mathbb{R}^d$ is an ε-subgradient of f at x if the inequality*

$$f(y) \geq f(x) + \langle u, y - x\rangle + \frac{\mu}{2}\|y - x\|^2 - \varepsilon$$

holds for all $y \in \mathbb{R}^d$.

Note that the condition $v - \mu x + \mu w \in \partial f(w)$ in Theorem 2 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 μ-strongly convex. Hence $w \in \arg\max_z \Phi(z)$, and for any z we have

$$f(z) \geq f(x) + \langle z - x, v\rangle + \frac{\mu}{2}\|x - z\|^2 - \frac{1 + \lambda\mu}{\lambda}\varepsilon.$$

In other words, v is an $\frac{1+\lambda\mu}{\lambda}\varepsilon$-subgradient of f at x. The inequality (2) further implies that $x+\lambda v \approx y$. Thus Theorem 2 indeed defines an approximation to the exact proximal point, for which $x + \lambda v = y$ and $v \in \partial f(x)$.

Appendix B. Details and proofs of Section 2

We define the potential function

$$p_k = A_k(f(x_k) - f(x^*)) + \frac{1 + \mu A_k}{2}\|z_k - x^*\|^2 \tag{16}$$

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 (Barré et al., 2021) 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 21 (Interpolation implies contraction) *For all $p, q \in \mathbb{R}$ such that $p + q > 0$, we have*

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

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

$$z_{k+1} \leftarrow \frac{1+\mu A_k}{1+\mu A_{k+1}} z_k + \frac{\mu a_{k+1}}{1+\mu A_{k+1}} w_{k+1} - \frac{a_{k+1}}{1+\mu A_{k+1}} \nabla_{k+1}. \tag{17}$$

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

Lemma 22 *We have*

$$\begin{aligned} &\frac{1+\mu A_k}{2}\|z_k - x^*\|^2 - \frac{1+\mu A_{k+1}}{2}\|z_{k+1} - x^*\|^2 \geq a_{k+1}(f(w_{k+1}) - f(x^*)) \\ &+ \frac{\mu a_{k+1}(1+\mu A_k)}{2(1+\mu A_{k+1})}\|z_k - w_{k+1} + \mu^{-1}\nabla_{k+1}\|^2 - \frac{a_{k+1}}{2\mu}\|\nabla_{k+1}\|^2 \end{aligned} \tag{18}$$

Proof: First note that

$$\begin{aligned} &\frac{1+\mu A_{k+1}}{2}\|z_{k+1} - x^*\|^2 - \frac{1+\mu A_k}{2}\|z_k - x^*\|^2 \\ &= \frac{\mu a_{k+1}}{2}\left\|\frac{1+\mu A_{k+1}}{\mu a_{k+1}}(z_{k+1} - x^*) - \frac{1+\mu A_k}{\mu a_{k+1}}(z_k - x^*)\right\|^2 \qquad (19a) \\ &\quad - \frac{(1+\mu A_k)(1+\mu A_{k+1})}{2\mu a_{k+1}}\|z_{k+1} - z_k\|^2 \\ &= \frac{\mu a_{k+1}}{2}\|x^* - w_{k+1} + \mu^{-1}\nabla_{k+1}\|^2 - \frac{\mu a_{k+1}(1+\mu A_k)}{2(1+\mu A_{k+1})}\|z_k - x_{k+1} + \mu^{-1}v_{k+1}\|^2 \qquad (19b) \end{aligned}$$

where Theorem 21 is used in (19a), and (19b) follows from (17). Thus, by strong convexity of f and the definition of w_{k+1} (see Theorem 2) we have

$$\begin{aligned} f(x^*) &\geq f(w_{k+1}) + \langle \nabla_{k+1}, x^* - w_{k+1}\rangle + \frac{\mu}{2}\|x^* - w_{k+1}\|^2 \\ &= f(w_{k+1}) + \frac{\mu}{2}\|x^* - w_{k+1} + \mu^{-1}\nabla_{k+1}\|^2 - \frac{1}{2\mu}\|\nabla_{k+1}\|^2 \end{aligned}$$

so that

$$\begin{aligned} a_{k+1}(f(x^*) - f(w_{k+1})) &\geq \frac{1+\mu A_{k+1}}{2}\|z_{k+1} - x^*\|^2 - \frac{1+\mu A_k}{2}\|z_k - x^*\|^2 \\ &+ \frac{\mu a_{k+1}(1+\mu A_k)}{2(1+\mu A_{k+1})}\|z_k - w_{k+1} + \mu^{-1}\nabla_{k+1}\|^2 - \frac{a_{k+1}}{2\mu}\|\nabla_{k+1}\|^2 \end{aligned}$$

as desired. □

Remark 23 *The derivation of (19) reveals the connection between the choice of parameters in the update (17) 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. Theorem 11).*

Now it suffices to deal with the function value terms. Strong convexity implies that

$$f(x_k) \geq f(w_{k+1}) + \frac{\mu}{2}\|x_k - w_{k+1} + \mu^{-1}\nabla_{k+1}\|^2 - \frac{1}{2\mu}\|\nabla_{k+1}\|^2 \tag{20}$$

and

$$f(x_{k+1}) \geq f(w_{k+1}) + \frac{\mu}{2}\|\mu^{-1}v_{k+1}\|^2 - \frac{1}{2\mu}\|\nabla_{k+1}\|^2 \tag{21}$$

while the definition of w_{k+1} implies

$$\begin{aligned}\frac{\sigma_k^2}{2}\|x_{k+1} - y_k\|^2 &\geq \frac{1}{2}\|x_{k+1} - y_k + \lambda_k v_{k+1}\|^2 \\ &+ \lambda_k(1+\lambda_k\mu)\left(f(x_{k+1}) - f(w_{k+1}) + \frac{1}{2\mu}\left(\|\nabla_{k+1}\|^2 - \|v_{k+1}\|^2\right)\right)\end{aligned} \tag{22}$$

We now seek a correct linear combination of the above inequalities to match the coefficient of $p_k - p_{k+1}$. Note that adding (21) and (22) leads to the following simpler inequality

$$\|x_{k+1} - y_k + \lambda_k v_{k+1}\|^2 \leq \sigma_k^2\|x_{k+1} - y_k\|^2 \tag{23}$$

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

Lemma 24 *We have for all $k \geq 0$ that*

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

Proof: By combining the inequalities (18),(20),(22) we have

$$\begin{aligned}&p_k - p_{k+1} \\ &= \underbrace{\left(\frac{1+\mu A_k}{2}\|z_k - x^*\|^2 - \frac{1+A_{k+1}}{2}\|z_{k+1} - x^*\|^2 + a_{k+1}\left(f(x^*) - f(w_{k+1})\right)\right)}_{\text{use } \textit{Theorem 22}} \\ &\quad + \underbrace{A_k(f(x_k) - f(w_{k+1}))}_{\text{use (20)}} + \underbrace{A_{k+1}(f(w_{k+1}) - f(x_{k+1}))}_{\text{use (22)}} \\ &\geq \frac{\mu a_{k+1}(1+\mu A_k)}{2(1+\mu A_{k+1})}\|z_k - x_{k+1} + \mu^{-1}v_{k+1}\|^2 + \frac{\mu A_k}{2}\|x_k - x_{k+1} + \mu^{-1}v_{k+1}\|^2 \\ &\quad - \frac{A_{k+1}}{2\mu}\|\nabla_{k+1}\|^2 + \frac{A_{k+1}}{2\mu}\left(\|\nabla_{k+1}\|^2 - \|v_{k+1}\|^2\right) \\ &\quad + \frac{A_{k+1}}{\lambda_k(1+\lambda_k\mu)}\left(\frac{1}{2}\|x_{k+1} - y_k + \lambda_k v_{k+1}\|^2 - \frac{\sigma_k^2}{2}\|x_{k+1} - y_k\|^2\right)\end{aligned} \tag{24}$$

We now show that the last expression in the above inequality is positive. Recall that in Section 2.1 we made the intuitive argument which shows that the "positive term" of form $\theta_z\|z_k - x_{k+1} +$

$\mu^{-1}v_{k+1}\|^2 + \theta_x\|x_k - x_{k+1} + \mu^{-1}v_{k+1}\|^2$ cannot be small. Formally, the choice of y_k implies that

$$\begin{aligned}
&\frac{\mu a_{k+1}(1+\mu A_k)}{2(1+\mu A_{k+1})}\|z_k - x_{k+1} + \mu^{-1}v_{k+1}\|^2 + \frac{\mu A_k}{2}\|x_k - x_{k+1} + \mu^{-1}v_{k+1}\|^2 \\
&\geq \frac{\mu\left(A_{k+1} + \mu(a_{k+1}A_k + A_kA_{k+1})\right)}{2(1+\mu A_{k+1})}\|y_k - x_{k+1} + \mu^{-1}v_{k+1}\|^2 \\
&\quad + \frac{\mu A_k a_{k+1}(1+\mu A_k)}{2\left(A_{k+1} + \mu(a_{k+1}A_k + A_kA_{k+1})\right)}\|x_k - z_k\|^2 \\
&= \frac{\mu a_{k+1}^2}{2\lambda_k(1+\mu A_{k+1})}\|y_k - x_{k+1} + \mu^{-1}v_{k+1}\|^2 + \frac{\mu\lambda_k A_k(1+\mu A_k)}{2a_{k+1}}\|x_k - z_k\|^2
\end{aligned}$$

where we have used the following equation

$$a_{k+1}^2 = \lambda_k\left(A_{k+1} + \mu(a_{k+1}A_k + A_kA_{k+1})\right) \tag{25}$$

to simplify the expression. We can now deduce from (23) that the right hand side of (24) is lower bounded by

$$\begin{aligned}
&\frac{\mu a_{k+1}^2}{2\lambda_k(1+\mu A_{k+1})}\|y_k - x_{k+1} + \mu^{-1}v_{k+1}\|^2 + \frac{\mu\lambda_k A_k(1+\mu A_k)}{2a_{k+1}}\|x_k - z_k\|^2 \\
&\quad - \frac{A_{k+1}}{2\mu}\|v_{k+1}\|^2 + \frac{A_{k+1}}{\lambda_k}\left(\tfrac{1}{2}\|x_{k+1} - y_k + \lambda_k v_{k+1}\|^2 - \tfrac{\sigma_k^2}{2}\|x_{k+1} - y_k\|^2\right).
\end{aligned}$$

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

$$\alpha\|x_{k+1} - y_k\|^2 + 2\beta\left\langle x_{k+1} - y_k, v_{k+1}\right\rangle + \gamma\|v_{k+1}\|^2 \tag{26}$$

where

$$\begin{aligned}
\alpha &= \frac{\mu a_{k+1}^2}{2\lambda_k(1+\mu A_{k+1})} + \frac{1-\sigma_k^2}{2}\frac{A_{k+1}}{\lambda_k} \\
\beta &= \frac{a_{k+1}^2}{2\lambda_k(1+\mu A_{k+1})} - \frac{1}{2}A_{k+1} = -\frac{\mu a_{k+1}^2}{2(1+\mu A_{k+1})} \\
\gamma &= \frac{a_{k+1}^2}{2\mu\lambda_k(1+\mu A_{k+1})} - \frac{A_{k+1}}{2\mu} + \frac{1}{2}\lambda_k A_{k+1} \\
&= \frac{1}{2}\lambda_k A_{k+1} - \frac{a_{k+1}^2}{2(1+\mu A_{k+1})} = \frac{\mu\lambda_k a_{k+1}^2}{2(1+\mu A_{k+1})}
\end{aligned}$$

where we have used (25) 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 (26) 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 Theorem 21, which suggests the correct way of doing this combination.

Theorem 3 is now a direct corollary of Theorem 24.

Theorem 25 *(Theorem 3 restated) For the iterates produced by Algorithm 1, we have*

$$\begin{aligned} f(x_k) - f(x^*) &\leq \frac{1}{A_k}\left(A_0(f(x_0) - f(x^*)) + \frac{1+\mu A_0}{2}\|x_0 - x^*\|^2\right) \\ &= \mathcal{O}\left(\Pi_{i=1}^k \left(1 + \max\left\{\mu\lambda_i, \sqrt{\mu\lambda_i}\right\}\right)^{-1}\right) \end{aligned}$$

Proof: Since $p_0 \geq p_k \geq A_k(f(x_k) - f(x^*))$, we have

$$f(x_k) - f(x^*) \leq \frac{1}{A_k}p_0 = \frac{1}{A_k}\left(A_0(f(x_0) - f(x^*)) + \frac{1+\mu A_0}{2}\|x_0 - x^*\|^2\right).$$

Note that

$$\begin{aligned} a_{k+1} = A_{k+1} - A_k &= \frac{\left(1+2\mu A_k\right)\lambda_{k+1} + \sqrt{\left(1+2\mu A_k\right)^2\lambda_{k+1}^2 + 4\left(1+\mu A_k\right)A_k\lambda_{k+1}}}{2} \\ &\geq A_k \max\left\{\mu\lambda_{k+1}, \sqrt{\mu\lambda_{k+1}}\right\}, \end{aligned}$$

so that the conclusion follows. □

Appendix C. Details of Section 3

C.1. Some useful properties of Algorithm 2

The following lemma characterize the growth rate of sequence $\{A_k\}$, which is closely related to the convergence rate of Algorithm 2.

Lemma 26 *For all $k \geq 0$, we have $A_{k+1} = (1+\mu\lambda_k)(\theta_k a_{k+1} + A_k)$.*

Proof: Since

$$(1-\theta_k)B_k = (1-\theta_k)\theta_k\delta_k B_{k+1} = \frac{\mu}{2}\theta_k\delta_k a_{k+1},$$

the equation $B_k(1-\theta_k)^2 = \mu\lambda_k\theta_k\left((1-\theta_k)B_k + \frac{\mu}{2}\delta_k A_k\right)$ can be equivalently written as

$$\begin{aligned} &(1-\theta_k)\frac{\mu}{2}\theta_k\delta_k a_{k+1} = \mu\lambda_k\theta_k \cdot \frac{\mu}{2}\delta_k(A_k + \theta_k a_{k+1}) \\ \Leftrightarrow\ &(1-\theta_k)a_{k+1} = \mu\lambda_k(A_k + \theta_k a_{k+1}) \\ \Leftrightarrow\ &A_{k+1} = A_k + a_{k+1} = (1+\mu\lambda_k)(A_k + \theta_k a_{k+1}). \end{aligned}$$

The conclusion follows. □

The next lemma reveals the relationship between the ratio of coefficients A_k and B_k and an important quantity $\xi_k = \frac{a_k}{A_k}$ (defined in the proof of Theorem 13). Recall that in the Euclidean setting, we have the equation $B_k = \frac{1+\mu A_k}{2}$, but the situation is more complex in the Riemannian setting due to the distortion rate δ_k.

Lemma 27 *For any $k \geq 0$, we have*

$$\frac{B_{k+1}}{A_{k+1}} = \frac{1+\mu\lambda_k}{2\lambda_k}\left(\frac{a_{k+1}}{A_{k+1}}\right)^2 = \frac{1+\mu\lambda_k}{2\lambda_k}\xi_{k+1}^2.$$

Proof: Recall that we have $A_{k+1} = (1+\mu\lambda_k)(\theta_k a_{k+1} + A_k) = (1+\mu\lambda_k)(A_{k+1} - (1-\theta_k)a_{k+1})$, so that

$$1 - \theta_k = \frac{\mu\lambda_k A_{k+1}}{(1+\mu\lambda_k)a_{k+1}}.$$

We can then obtain

$$\frac{B_{k+1}}{A_{k+1}} = \frac{\mu}{2}\frac{a_{k+1}}{(1-\theta_k)A_{k+1}} = \frac{1+\mu\lambda_k}{2\lambda_k}\left(\frac{a_{k+1}}{A_{k+1}}\right)^2,$$

as desired. □

C.2. Potential function analysis

Lemma 28 (restatement of Theorem 11) *Suppose that $\delta_k > 0$ is a valid distortion rate and $B_{k+1} = \frac{B_k}{\theta_k\delta_k}$, then*

$$B_k d_{w_k}^2(z_k, x^*) - B_{k+1} d_{w_{k+1}}^2(z_{k+1}, x^*) \geq (1-\theta_k)B_{k+1}\left(\frac{2}{\mu}(f(w_{k+1}) - f(x^*)) - \frac{1}{\mu^2}\|\nabla_{k+1}\|^2\right) + \theta_k(1-\theta_k)B_{k+1}\left\|\mathtt{Exp}_{w_{k+1}}^{-1}(z_k) - \mathtt{Exp}_{w_{k+1}}^{-1}(x_{k+1}) + \mu^{-1}v_{k+1}\right\|^2$$

Proof: Since δ_k is a valid distortion rate, we have

$$B_k d_{w_k}^2(z_k, x^*) \geq \frac{B_k}{\delta_k} d_{w_{k+1}}^2(z_k, x^*) \tag{27}$$

This implies that

$$B_{k+1} d_{w_{k+1}}^2(z_{k+1}, x^*) - B_k d_{w_k}^2(z_k, x^*) \leq B_{k+1} d_{w_{k+1}}^2(z_{k+1}, x^*) - \theta_k B_{k+1} d_{w_{k+1}}^2(z_k, x^*) \tag{28a}$$

$$= (1-\theta_k)B_{k+1}\left(\frac{1}{1-\theta_k} d_{w_{k+1}}(z_{k+1}, x^*) - \frac{\theta_k}{1-\theta_k} d_{w_{k+1}}(z_k, x^*)\right)^2 \tag{28b}$$

$$- \frac{\theta_k}{1-\theta_k}\left(d_{w_{k+1}}(z_{k+1}, x^*) - d_{w_{k+1}}(z_k, x^*)\right)^2$$

$$= (1-\theta_k)B_{k+1}\left\|\mathtt{Exp}_{w_{k+1}}^{-1}(x^*) - \mathtt{Exp}_{w_{k+1}}^{-1}(x_{k+1}) + \mu^{-1}v_{k+1}\right\|^2 \tag{28c}$$

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

where (28a) follows from (27) and $\theta_k B_{k+1} = \frac{B_k}{\delta_k}$, (28b) uses Theorem 21, and (28c) follows from the definition of z_{k+1}. On the other hand, by strong convexity of f, we have

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

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

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

The conclusion follows by plugging this inequality into (28). Note that the steps after (28a) are essentially the same as the Euclidean setting, because all the calculations are done in the tangent space $T_{w_{k+1}}\mathcal{M}$. □

We then proceed to derive a Riemannian analog of Theorem 24, where we proved the potential decrease in the Euclidean setting. By following the same approach as Theorem 24, we can see that the inequality would involve an additional point y'_k.

Lemma 29 (restatement of Theorem 12) *Suppose that* $a_{k+1} = A_{k+1} - A_k = \frac{2}{\mu}(1-\theta_k)B_{k+1}$, *then*

$$\begin{aligned} p_k - p_{k+1} \geq\ & \frac{\mu}{2}(\theta_k a_{k+1} + A_k)\left\|\mathtt{Exp}_{w_{k+1}}^{-1}(y'_k) - \mathtt{Exp}_{w_{k+1}}^{-1}(x_{k+1}) + \mu^{-1}v_{k+1}\right\|^2 \\ & + \frac{A_{k+1}}{2\lambda_k\sigma_k}\left\|\mathtt{Exp}_{w_{k+1}}^{-1}(y_k) - \mathtt{Exp}_{w_{k+1}}^{-1}(x_{k+1}) - \lambda_k v_{k+1}\right\|^2 \\ & + \frac{\mu\theta_k a_{k+1}A_k}{2(A_k+\theta_k a_{k+1})}d^2_{w_{k+1}}(x_k, z_k) - \frac{\sigma_k A_{k+1}}{2\lambda_k}d^2_{w_{k+1}}(x_{k+1}, y_k) - \frac{A_{k+1}}{2\mu}\|v_{k+1}\|^2 \end{aligned} \tag{29}$$

where

$$y'_k = \mathtt{Exp}_{w_{k+1}}\left(\frac{A_k}{A_k+\theta_k a_{k+1}}\mathtt{Exp}_{w_{k+1}}^{-1}(x_k) + \frac{\theta_k a_{k+1}}{A_k+\theta_k a_{k+1}}\mathtt{Exp}_{w_{k+1}}^{-1}(z_k)\right) \tag{30}$$

Proof: Recall the the argument in Theorem 24 basically uses strong convexity and the definition of Euclidean *iprox* to lower bound the potential decrease with a quadratic function, and the choice of parameters ensure that the quadratic is positive definite. In the Riemannian setting, since 'vector' on a manifold is undefined, we need to work with vectors in a tangent space instead. In the following, we work in the tangent space $\mathcal{T}_{w_{k+1}}$. This choice is quite natural, since straightforwardly generalizing of the proof of Theorem 24 would involve exponential maps at w_{k+1}. In $\mathcal{T}_{w_{k+1}}$, our goal is to derive a quadratic function to lower bound $p_k - p_{k+1}$.

Strong convexity implies that

$$f(x_k) \geq f(w_{k+1}) + \frac{\mu}{2}\|\mathtt{Exp}_{w_{k+1}}^{-1}(x_k) + \mu^{-1}\nabla_{k+1}\|^2 - \frac{1}{2\mu}\|\nabla_{k+1}\|^2$$

and

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

and the definition of Riemannian *iprox* operator (6) implies that

$$\begin{aligned} \frac{\sigma_k^2}{2}\|\mathtt{Exp}_{w_{k+1}}^{-1}(x_{k+1}) - \mathtt{Exp}_{w_{k+1}}^{-1}(y_k)\|^2 \geq\ & \frac{1}{2}\|\mathtt{Exp}_{w_{k+1}}^{-1}(x_{k+1}) - \mathtt{Exp}_{w_{k+1}}^{-1}(y_k) + \lambda_k v_{k+1}\|^2 \\ & + \lambda_k(1+\lambda_k\mu)\left(f(x_{k+1}) - f(w_{k+1}) + \frac{1}{2\mu}\left(\|\nabla_{k+1}\|^2 - \|v_{k+1}\|^2\right)\right) \end{aligned}$$

Combining the above inequalities, we have

$$p_k - p_{k+1} = \left(B_k d^2_{w_k}(z_k, x^*) - B_{k+1} d^2_{w_{k+1}}(z_{k+1}, x^*) + \frac{2}{\mu}(1-\theta_k) B_{k+1}(f(x^*) - f(w_{k+1})) \right) \tag{31a}$$

$$\begin{aligned}
&+ A_k(f(x_k) - f(w_{k+1})) + A_{k+1}(f(w_{k+1}) - f(x_{k+1})) \\
&\geq \frac{\mu A_k}{2} \left\| \mathtt{Exp}^{-1}_{w_{k+1}}(x_k) - \mathtt{Exp}^{-1}_{w_{k+1}}(x_{k+1}) + \mu^{-1} v_{k+1} \right\|^2 \\
&+ \frac{\mu}{2} \theta_k a_{k+1} \left\| \mathtt{Exp}^{-1}_{w_{k+1}}(z_k) - \mathtt{Exp}^{-1}_{w_{k+1}}(x_{k+1}) + \mu^{-1} v_{k+1} \right\|^2 \\
&+ \frac{A_{k+1}}{2\lambda_k \sigma_k} \left\| \mathtt{Exp}^{-1}_{w_{k+1}}(y_k) - \mathtt{Exp}^{-1}_{w_{k+1}}(x_{k+1}) - \lambda_k v_{k+1} \right\|^2 \\
&- \frac{\sigma_k A_{k+1}}{2\lambda_k} d^2_{w_{k+1}}(x_{k+1}, y_k) - \frac{A_{k+1}}{2\mu} \|v_{k+1}\|^2
\end{aligned}$$

where we use the condition $a_{k+1} = \frac{2}{\mu}(1-\theta_k) B_{k+1}$ in (31a). Finally, Theorem 21 implies that

$$\begin{aligned}
&\frac{\mu A_k}{2} \left\| \mathtt{Exp}^{-1}_{w_{k+1}}(x_k) - \mathtt{Exp}^{-1}_{w_{k+1}}(x_{k+1}) + \mu^{-1} v_{k+1} \right\|^2 \\
&\quad + \frac{\mu}{2} \theta_k a_{k+1} \left\| \mathtt{Exp}^{-1}_{w_{k+1}}(z_k) - \mathtt{Exp}^{-1}_{w_{k+1}}(x_{k+1}) + \mu^{-1} v_{k+1} \right\|^2 \\
&= \frac{\mu}{2} (\theta_k a_{k+1} + A_k) \left\| \mathtt{Exp}^{-1}_{w_{k+1}}(y'_k) - \mathtt{Exp}^{-1}_{w_{k+1}}(x_{k+1}) + \mu^{-1} v_{k+1} \right\|^2 \\
&\quad + \frac{\mu \theta_k a_{k+1} A_k}{2(A_k + \theta_k a_{k+1})} d^2_{w_{k+1}}(x_k, z_k).
\end{aligned}$$

The conclusion follows. The final equation in the proof explains why y'_k would appear in (29). □

C.3. Convergence without the additional distortion

In the Riemannian setting, it is not *guaranteed* that y_k is the same as y'_k, and this may give rise to the *additional distortion*, as shown in Theorem 12. However, recall that our definition of *iprox* allows flexible choices of x_{k+1}, w_{k+1} and v_{k+1}. We can see that in some special cases, we still have $y_k = y'_k$. The following proposition provides sufficient condition for this to hold. It can be easily derived from the definition of y_k and y'_k.

Proposition 30 *Suppose that w_{k+1} lies on the geodesic connecting x_k and z_k, then $y_k = y'_k$.*

We now move on to theoretical analysis under the condition $y_k = y'_k$. The right hand side of (8) is a quadratic function of $\mathtt{Exp}^{-1}_{w_{k+1}}(y_k) - \mathtt{Exp}^{-1}_{w_{k+1}}(x_{k+1})$ and v_{k+1}, after ignoring the non-negative $d_{w_{k+1}}(x_k, z_k)$ term. We can then prove the following lemma for potential decrease. The equation $A_{k+1} = (1 + \mu\lambda_k)(\theta_k a_{k+1} + A_k)$ plays a crucial role in the proof.

Lemma 31 *Suppose that $\sigma_k < 1$, then*

$$\begin{aligned}\frac{(1-\sigma_k)A_{k+1}}{2\lambda_k} d^2_{w_{k+1}}(x_{k+1}, y_k) \le \frac{\mu}{2}(\theta_k a_{k+1} + A_k) \left\| \mathtt{Exp}^{-1}_{w_{k+1}}(y_k) - \mathtt{Exp}^{-1}_{w_{k+1}}(x_{k+1}) + \mu^{-1} v_{k+1} \right\|^2 \\ + \frac{A_{k+1}}{2\lambda_k} \left\| \mathtt{Exp}^{-1}_{w_{k+1}}(y_k) - \mathtt{Exp}^{-1}_{w_{k+1}}(x_{k+1}) - \lambda_k v_{k+1} \right\|^2 \\ - \frac{\sigma_k A_{k+1}}{2\lambda_k} d^2_{w_{k+1}}(x_{k+1}, y_k) - \frac{A_{k+1}}{2\mu} \|v_{k+1}\|^2\end{aligned}$$

Proof: First note that 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):

$$\mathtt{RHS} - \mathtt{LHS} = \alpha d^2_{w_{k+1}}(x_{k+1}, y_k) + 2\beta \left\langle \mathtt{Exp}^{-1}_{w_{k+1}}(y_k) - \mathtt{Exp}^{-1}_{w_{k+1}}(x_{k+1}), v_{k+1} \right\rangle + \gamma \|v_{k+1}\|^2$$

where

$$\begin{aligned}\alpha &= \frac{\mu}{2}(\theta_k a_{k+1} + A_k) + \frac{(1-\sigma_k)A_{k+1}}{2\lambda_k} = \frac{A_{k+1}}{2}\left(\frac{\mu}{1+\mu\lambda_k} + \frac{1-\sigma_k}{\lambda_k}\right) \\ \beta &= \frac{1}{2}(\theta_k a_{k+1} + A_k) - \frac{A_{k+1}}{2} = -\frac{A_{k+1}}{2} \cdot \frac{\mu\lambda_k}{1+\mu\lambda_k} \\ \gamma &= \frac{1}{2\mu}(\theta_k a_{k+1} + A_k) + \frac{\lambda_k A_{k+1}}{2} - \frac{A_{k+1}}{2\mu} = \frac{A_{k+1}}{2} \cdot \frac{\mu\lambda_k^2}{1+\mu\lambda_k}.\end{aligned}$$

Note that

$$\beta^2 = \left(\alpha - \frac{(1-\sigma_k)A_{k+1}}{2\lambda_k}\right)\gamma,$$

we can thus obtain

$$\mathtt{RHS} - \mathtt{LHS} \ge \frac{(1-\sigma_k)A_{k+1}}{2\lambda_k} d^2_{w_{k+1}}(x_{k+1}, y_k)$$

as desired. □

Combining Theorem 11 and Theorem 31, we can see that the potential sequence $\{p_k\}$ is non-increasing:

Corollary 32 *Suppose that $y_k = y'_k$, then the following inequality holds:*

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

In particular, we have $p_{k+1} \le p_k$, so that $p_k \le p_0$ for all $k \ge 1$.

Finally, we can prove the following theorem, which says that if w_{k+1} is chosen on the geodesic connecting x_k and z_k, then Algorithm 2 provably achieves eventual acceleration with arbitrary initialization.

Theorem 33 (restatement of Theorem 13) *Suppose that in Algorithm 2, we choose $\lambda_k = \lambda$ and w_{k+1} lies on the geodesic connecting x_k and z_k s.t. $d(w_{k+1}, y_k) = \mathcal{O}(1)$, then we have*

$$f(x_k) - f(x^*) \le \frac{p_0}{A_k}, \quad d^2_{w_{k+1}}(z_k, x^*) \le \frac{p_0}{B_k} \le \frac{2p_0}{\mu a_k}. \tag{32}$$

Moreover, we have

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

Proof: The first two inequalities follow from Theorem 32 and

$$a_{k+1} = (1-\theta_k)\frac{B_k}{\delta_k\theta_k} = 2\mu^{-1}(1-\theta_k)B_{k+1} < 2\mu^{-1}B_{k+1}, \quad \forall k \geq 0. \tag{33}$$

We now prove (33). This is equivalent to

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

Define $\xi_k = \frac{a_k}{A_k}$ for $k \geq 1$. Note that the update of θ_k and a_{k+1} in Algorithm 2 implies that

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

Thus

$$\begin{aligned}\delta_k a_{k+1}^2 &= 2\lambda\left(B_k A_{k+1} + \frac{\mu}{2}\delta_k A_k a_{k+1}\right)\\ (1+\mu\lambda)a_{k+1}^2 &= 2\delta_k^{-1}\lambda A_{k+1}\left(B_k + \frac{\mu}{2}\delta_k a_{k+1}\right) = 2\lambda A_{k+1}B_{k+1}\end{aligned} \tag{34}$$

As a result, we have $\frac{B_k}{A_k} = \frac{1+\mu\lambda}{2\lambda}\xi_k^2$. The above derivations only holds for $k \geq 1$, we artificially define $\xi_0 = \sqrt{\frac{2\lambda}{1+\mu\lambda}\frac{B_k}{A_k}}$, so that for all $k \geq 0$, rewrite the equation $B_{k+1} = \frac{B_k}{\delta_k} + \frac{\mu}{2}a_{k+1}$ in terms of ξ as

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

or equivalently,

$$\delta_k\xi_{k+1}^2 = \xi_k^2(1-\xi_{k+1}) + \frac{\mu\lambda}{1+\mu\lambda}\delta_k\xi_{k+1} \tag{35}$$

Before proceeding to analyze the recursive equation (35), we first prove that $\lim_{k\to+\infty}\delta_k = 1$. This is in fact necessary since otherwise $\{\xi_k\}$ would not converge to the fixed point $\sqrt{\frac{\mu\lambda}{1+\mu\lambda}}$.

Since $\lim_{k\to+\infty}A_k = +\infty$, we have $x_k \to x^*$ and $d_{w_{k+1}}(x_{k+1}, y_k) \to 0$, by Theorem 32. By assumption, $d(w_{k+1}, y_k)$ is bounded, so that

$$d(x_{k+1}, y_k) \leq d_{w_{k+1}}(x_{k+1}, y_k) + 2d(w_{k+1}, y_k)$$

is bounded, which implies that the sequence $\{y_k\}$ is bounded. Thus $\{w_k\}$ is also bounded.

Since $A_{k+1} \geq (1+2\mu\lambda)A_k$, we have $a_{k+1} = A_{k+1} - A_k \geq 2\mu\lambda A_k$, so that $\lim_{k\to+\infty}a_k = +\infty$ and $d_{w_{k+1}}^2(z_k, x^*) \leq \frac{p_0}{a_k} \to 0$. Since $w_{k+1} = \mathcal{O}(1)$, the distortion inequality Theorem 10 implies that $d(z_k, x^*) \to 0$. Note that w_{k+1} lies on the geodesic connecting x_k and z_k, and $\mathcal{M}$ has non-positive curvature, we have

$$d(w_{k+1}, x^*) \leq \max\{d(x_k, x^*), d(z_k, x^*)\} \to 0.$$

Hence $\delta_k = T_K(d(w_k, z_k)) \to 1$ as $k \to +\infty$.

We now return to (35). We first show that for any $\varepsilon > 0$, we have

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

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

The recursive relation (35) 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 δ_k becomes larger and ξ_k becomes smaller, then ξ_{k+1} also becomes smaller. Based on this observation, we first choose k_0 such that $\delta_k \le 1 + \varepsilon\sqrt{\frac{\mu\lambda}{1+\mu\lambda}}$ for all $k \ge k_0$, and then construct a *reference sequence* $\{\zeta_k\}_{k\ge k_0}$ defined as

$$\zeta_{k_0} = \xi_{k_0}, \quad \delta\zeta_{k+1} \left(\zeta_{k+1} - \frac{\mu\lambda}{1+\mu\lambda} \right) = \zeta_k^2 (1 - \zeta_{k+1}), \quad \delta = 1 + \varepsilon\sqrt{\frac{\mu\lambda}{1+\mu\lambda}}$$

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

$$\varphi(x) = \frac{1}{2\delta} \left(\frac{\mu\lambda}{1+\mu\lambda}\delta - x^2 + \sqrt{\left(x^2 - \frac{\mu\lambda}{1+\mu\lambda}\delta \right)^2 + 4\delta x^2} \right) \tag{36}$$

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 $\varphi'(x) \ge 0$. On the other hand,

$$\begin{aligned} \varphi'(x) < 1 &\Leftrightarrow \left(x\left(x^2 - \frac{\mu\lambda}{1+\mu\lambda}\delta \right) + 2\delta x \right)^2 < (x+\delta)^2 \left(\left(x^2 - \frac{\mu\lambda}{1+\mu\lambda}\delta \right)^2 + 4\delta x^2 \right) \\ &\Leftarrow 4\delta x^2 \left(x^2 - \frac{\mu\lambda}{1+\mu\lambda}\delta \right)^2 + 4\delta^2 x^2 < (x+\delta)^2 \cdot 4\delta x^2 \end{aligned} \tag{37}$$

which trivially holds, since $\delta > 1$. Since φ is continuously differentiable, we have $\sup_{x\in[0,1]} \varphi'(x) < 1$ i.e. φ is a contraction mapping. Since $\zeta_k \in [0,1]$, $\forall k \ge k_0$, it converges exponentially fast to a fixed point of φ, 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

$$\liminf_{k\to+\infty} \xi_k \ge \liminf_{k\to+\infty} \zeta_k \ge (1-\varepsilon)\sqrt{\frac{\mu\lambda}{1+\mu\lambda}}$$

To prove the desired result, it remains show that

$$\limsup_{k\to+\infty} \xi_k \le \sqrt{\frac{\mu\lambda}{1+\mu\lambda}}$$

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 $\sqrt{\frac{\mu\lambda}{1+\mu\lambda}}$. □

Although Theorem 13 shows that Algorithm 2 eventually achieves acceleration in the sense of Theorem 1, it might also be helpful to know how fast the sequence $\{\tau_k\}$ (cf. Theorem 1) achieves the order of $\mathcal{O}\left(\sqrt{\frac{\mu}{L}}\right)$ i.e. how long the 'burn-in' period takes to achieve full acceleration. Since at this point we are focusing on acceleration for smooth strongly-convex functions, in the following we always assume that f is L-smooth.

Lemma 34 *Suppose that $d(w_{k+1}, y_k) = \mathcal{O}(1)$, and $\lambda_k = \lambda = \frac{c}{L}$, where $c \in (0,1)$ is a numerical constant, then*

$$\delta_k - 1 \le C_0 \left(1 + c\frac{\mu}{L}\right)^{-k}$$

where C_0 is a constant that may depend on L, μ and initialization, but independent of k.

Proof: Let D be a uniform upper bound of $d(w_{k+1}, y_k)$. Since $a_k \ge c\frac{\mu}{L}A_k \ge c\frac{\mu}{L}\left(1 + c\frac{\mu}{L}\right)^k A_0$, we have

$$d^2_{w_{k+1}}(z_k, x^*) \le \frac{2L}{\mu^2 A_0}\left(1 + c\frac{\mu}{L}\right)^{-k} p_0$$

Recall that in the proof of Theorem 13 we have shown that $\{w_k\}$ is bounded, and it's easy to see that the upper bound only depends on initialization and $d(w_{k+1}, y_k)$, by Theorem 10 we have

$$d^2(z_k, x^*) \le \frac{2C_1 L}{c\mu^2 A_0}\left(1 + c\frac{\mu}{L}\right)^{-k} p_0$$

for some $C_1 \ge 1$ that only depends on initialization and D. Since w_{k+1} lies on the geodesic between x_k and z_k, we have

$$\begin{aligned} d^2(w_{k+1}, x^*) &\le \max\left\{d^2(x_k, x^*), d^2(z_k, x^*)\right\} \\ &\le \max\left\{2\mu^{-1}(f(x_k) - f(x^*)), d^2(z_k, x^*)\right\} \le \frac{2C_1 L}{c\mu^2 A_0}\left(1 + c\frac{\mu}{L}\right)^{-k} p_0 \end{aligned}$$

As a result,

$$d^2(w_k, z_k) \le 2\left(d^2(w_k, x^*) + d^2(z_k, x^*)\right) \le \frac{12C_1 L}{c\mu^2 A_0}\left(1 + c\frac{\mu}{L}\right)^{-(k-1)} p_0 \tag{38}$$

Finally since $T_K(r) = 1 + \mathcal{O}(r^2)$ for small r, we have

$$\delta_k - 1 = \mathcal{O}\left(\left(1 + c\frac{\mu}{L}\right)^{-k}\right),$$

as desired. □

Theorem 35 (restatement of Theorem 15) *Suppose that $d(w_{k+1}, y_k) = \mathcal{O}(1)$ and $\lambda_k = \lambda = \mathcal{O}\left(\frac{1}{L}\right)$, then we have $\xi_k \ge \frac{1}{2}\sqrt{\frac{\mu\lambda}{1+\mu\lambda}}$ after $T = \widetilde{\mathcal{O}}\left(\frac{L}{\mu}\right)$ iterations, where $\widetilde{\mathcal{O}}$ hides logarithmic terms which may depend on L, μ and the initialization. As a result, Algorithm 2 achieves acceleration in at most $\widetilde{\mathcal{O}}\left(\frac{L}{\mu}\right)$ iterations.*

Proof: We consider the recursive equation of ξ_k derived in the proof of Theorem 13:

$$\delta_k \xi_{k+1} \left(\xi_{k+1} - \frac{\mu\lambda}{1+\mu\lambda} \right) = \xi_k^2 (1 - \xi_{k+1}) \tag{39}$$

The previous lemma implies that

$$\delta_k \leq 1 + \sqrt{\frac{\mu\lambda}{1+\mu\lambda}} \tag{40}$$

holds after $\widetilde{\mathcal{O}}\left(\frac{L}{\mu}\right)$ iterations, where $\widetilde{\mathcal{O}}$ hides logarithmic terms. In the following, we study how many iterations are needed for $\xi_k \geq \frac{1}{2}\sqrt{\frac{\mu\lambda}{1+\mu\lambda}}$ after (40) is guaranteed to hold.

Indeed, note that smaller δ_k and larger ξ_k implies a larger ξ_{k+1} in (39), it suffices to consider the case $\delta_k = \delta = 1 + \sqrt{\frac{\mu\lambda}{1+\mu\lambda}}$.

Now we study the behavior of $\varphi(x)$ defined in (36) more carefully. Its derivative $\varphi'(x)$ can be written as

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

Hence for all $\delta, x > 0$ we have $\varphi'(x) \leq \frac{1}{\sqrt{2}}$. This implies that with a constant δ, (36) converges to its fixed point in $\widetilde{\mathcal{O}}(1)$ iterations. Since for $\delta = 1 + \sqrt{\frac{\mu\lambda}{1+\mu\lambda}}$, its fixed point is larger than $1 + \frac{1}{2}\sqrt{\frac{\mu\lambda}{1+\mu\lambda}}$, we conclude that a total number of $\widetilde{\mathcal{O}}\left(\frac{L}{\mu}\right)$ iterations are needed for $\xi_k \geq \frac{1}{2}\sqrt{\frac{\mu\lambda}{1+\mu\lambda}}$ to hold. □

C.4. The general case

This subsection provides details and proofs of our main results for the general case, where the additional distortion is present. We begin with the following result, which shows that we need to control the distance between y_k and y'_k.

Lemma 36 *Suppose that $\sigma_k < 1$, then*

$$\begin{aligned} p_k - p_{k+1} \geq{}& \frac{(1-\sigma_k)A_{k+1}}{4\lambda_k} d^2_{w_{k+1}}(x_{k+1}, y_k) + \frac{\mu\theta_k a_{k+1} A_k}{2(A_k + \theta_k a_{k+1})} d^2_{w_{k+1}}(x_k, z_k) \\ &+ \frac{(1-\sigma_k)A_{k+1}}{6}\sqrt{\frac{\mu\lambda_k}{1+\mu\lambda_k}} d_{w_{k+1}}(x_{k+1}, y_k)\|v_{k+1}\| \\ &- \mu(\theta_k a_{k+1} + A_k) d_{w_{k+1}}(y_k, y'_k) \cdot \left\| \mathtt{Exp}^{-1}_{w_{k+1}}(y_k) - \mathtt{Exp}^{-1}_{w_{k+1}}(x_{k+1}) + \mu^{-1} v_{k+1} \right\| \end{aligned} \tag{41}$$

Proof: Note that

$$\begin{aligned}
&\left\|\mathtt{Exp}_{w_{k+1}}^{-1}(y_k') - \mathtt{Exp}_{w_{k+1}}^{-1}(x_{k+1}) + \mu^{-1}v_{k+1}\right\|^2 \\
&\geq \left\|\mathtt{Exp}_{w_{k+1}}^{-1}(y_k) - \mathtt{Exp}_{w_{k+1}}^{-1}(x_{k+1}) + \mu^{-1}v_{k+1}\right\|^2 \\
&\quad + 2\left\langle \mathtt{Exp}_{w_{k+1}}^{-1}(y_k') - \mathtt{Exp}_{w_{k+1}}^{-1}(y_k), \mathtt{Exp}_{w_{k+1}}^{-1}(y_k) - \mathtt{Exp}_{w_{k+1}}^{-1}(x_{k+1}) + \mu^{-1}v_{k+1}\right\rangle \\
&\geq \left\|\mathtt{Exp}_{w_{k+1}}^{-1}(y_k) - \mathtt{Exp}_{w_{k+1}}^{-1}(x_{k+1}) + \mu^{-1}v_{k+1}\right\|^2 \\
&\quad - 2d_{w_{k+1}}(y_k, y_k') \cdot \left\|\mathtt{Exp}_{w_{k+1}}^{-1}(y_k) - \mathtt{Exp}_{w_{k+1}}^{-1}(x_{k+1}) + \mu^{-1}v_{k+1}\right\|
\end{aligned}$$

where the last line follows from Cauchy-Schwarz inequality. The remaining steps of the proof is similar to Theorem 31, except that we also need to incorporate the $\left\langle \mathtt{Exp}_{w_{k+1}}^{-1}(y_k) - \mathtt{Exp}_{w_{k+1}}^{-1}(x_{k+1}), v_{k+1}\right\rangle$ into the bound. Indeed we have

$$\begin{aligned}
&\alpha d_{w_{k+1}}^2(x_{k+1}, y_k) + 2\beta\left\langle \mathtt{Exp}_{w_{k+1}}^{-1}(y_k) - \mathtt{Exp}_{w_{k+1}}^{-1}(x_{k+1}), v_{k+1}\right\rangle + \gamma\|v_{k+1}\|^2 \\
&\geq \frac{(1-\sigma_k)A_{k+1}}{4\lambda_k} d_{w_{k+1}}^2(x_{k+1}, y_k) + A_{k+1}\left(\sqrt{\frac{\mu^2\lambda_k^2}{(1+\mu\lambda_k)^2} + \frac{(1-\sigma_k)\mu\lambda_k}{2(1+\mu\lambda_k)}} - \frac{\mu\lambda_k}{1+\mu\lambda_k}\right) \cdot \\
&\quad \left\|\mathtt{Exp}_{w_{k+1}}^{-1}(y_k) - \mathtt{Exp}_{w_{k+1}}^{-1}(x_{k+1})\right\| \|v_{k+1}\| \\
&\geq \frac{(1-\sigma_k)A_{k+1}}{4\lambda_k} d_{w_{k+1}}^2(x_{k+1}, y_k) + \frac{(1-\sigma_k)A_{k+1}}{6}\sqrt{\frac{\mu\lambda_k}{1+\mu\lambda_k}} d_{w_{k+1}}(x_{k+1}, y_k)\|v_{k+1}\|
\end{aligned}$$

where α, β and γ are the coefficients defined in Theorem 31. Hence, by Theorem 29 we have

$$\begin{aligned}
p_k - p_{k+1} &\geq \frac{\mu}{2}(\theta_k a_{k+1} + A_k)\left\|\mathtt{Exp}_{w_{k+1}}^{-1}(y_k') - \mathtt{Exp}_{w_{k+1}}^{-1}(x_{k+1}) + \mu^{-1}v_{k+1}\right\|^2 \\
&\quad + \frac{A_{k+1}}{2\lambda_k\sigma_k}\left\|\mathtt{Exp}_{w_{k+1}}^{-1}(y_k) - \mathtt{Exp}_{w_{k+1}}^{-1}(x_{k+1}) - \lambda_k v_{k+1}\right\|^2 \\
&\quad + \frac{\mu\theta_k a_{k+1}A_k}{2(A_k + \theta_k a_{k+1})} d_{w_{k+1}}^2(x_k, z_k) - \frac{\sigma_k A_{k+1}}{2\lambda_k} d_{w_{k+1}}^2(x_{k+1}, y_k) - \frac{A_{k+1}}{2\mu}\|v_{k+1}\|^2 \\
&\geq \alpha d_{w_{k+1}}^2(x_{k+1}, y_k) + 2\beta\left\langle \mathtt{Exp}_{w_{k+1}}^{-1}(y_k) - \mathtt{Exp}_{w_{k+1}}^{-1}(x_{k+1}), v_{k+1}\right\rangle + \gamma\|v_{k+1}\|^2 \\
&\quad + \frac{\mu\theta_k a_{k+1}A_k}{2(A_k + \theta_k a_{k+1})} d_{w_{k+1}}^2(x_k, z_k) \\
&\quad - \mu(\theta_k a_{k+1} + A_k) d_{w_{k+1}}(y_k, y_k') \cdot \left\|\mathtt{Exp}_{w_{k+1}}^{-1}(y_k) - \mathtt{Exp}_{w_{k+1}}^{-1}(x_{k+1}) + \mu^{-1}v_{k+1}\right\| \\
&\geq \frac{(1-\sigma_k)A_{k+1}}{4\lambda_k} d_{w_{k+1}}^2(x_{k+1}, y_k) + \frac{\mu\theta_k a_{k+1}A_k}{2(A_k + \theta_k a_{k+1})} d_{w_{k+1}}^2(x_k, z_k) \\
&\quad + \frac{(1-\sigma_k)A_{k+1}}{6}\sqrt{\frac{\mu\lambda_k}{1+\mu\lambda_k}} d_{w_{k+1}}(x_{k+1}, y_k)\|v_{k+1}\| \\
&\quad - \mu(\theta_k a_{k+1} + A_k) d_{w_{k+1}}(y_k, y_k') \cdot \left\|\mathtt{Exp}_{w_{k+1}}^{-1}(y_k) - \mathtt{Exp}_{w_{k+1}}^{-1}(x_{k+1}) + \mu^{-1}v_{k+1}\right\|
\end{aligned}$$

as desired. □

In order to ensure potential decrease, it suffices to control the magnitude of the error term $d_{w_{k+1}}(y_k, y_k')$, as shown in the corollary below:

Corollary 37 *Suppose that*

$$d_{w_{k+1}}(y_k, y_k') \le \frac{1-\sigma_k}{6} \min\left\{\sqrt{\mu\lambda_k}, 1\right\} d_{w_{k+1}}(x_{k+1}, y_k),$$

then we have the following inequality which implies potential decrease:

$$p_k - p_{k+1} \ge \frac{(1-\sigma_k)A_{k+1}}{12\lambda_k} d^2_{w_{k+1}}(x_{k+1}, y_k) + \frac{\mu\theta_k a_{k+1} A_k}{2(A_k + \theta_k a_{k+1})} d^2_{w_{k+1}}(x_k, z_k).$$

Proof: Under the given condition, we can see that

$$\begin{aligned}
&\frac{(1-\sigma_k)A_{k+1}}{6}\sqrt{\frac{\mu\lambda_k}{1+\mu\lambda_k}} d_{w_{k+1}}(x_{k+1}, y_k) \\
&\ge \frac{1-\sigma_k}{6}\sqrt{\mu\lambda_k(1+\mu\lambda_k)}(\theta_k a_{k+1} + A_k) \cdot \frac{6}{1-\sigma_k}\frac{1}{\sqrt{\mu\lambda_k}} d_{w_{k+1}}(y_k, y_k') \\
&\ge (\theta_k a_{k+1} + A_k) d_{w_{k+1}}(y_k, y_k')
\end{aligned}$$

and

$$\frac{(1-\sigma_k)A_{k+1}}{6\lambda_k} d_{w_{k+1}}(x_{k+1}, y_k) \ge \frac{1+\mu\lambda_k}{\lambda_k}(\theta_k a_{k+1} + A_k) d_{w_{k+1}}(y_k, y_k') \ge \mu(\theta_k a_{k+1} + A_k) d_{w_{k+1}}(y_k, y_k')$$

Plugging the above inequalities into (41), we obtain the desired result. □

Lemma 38 (restatement of Theorem 16) *Suppose that $x \in \mathcal{M}$ and $y, a \in T_x\mathcal{M}$. Let $z = \mathtt{Exp}_x(a)$, then we have*

$$d\left(\mathtt{Exp}_x(y+a), \mathtt{Exp}_z\left(\Gamma_x^z y\right)\right) \le \min\{\|a\|, \|y\|\} S_K(\|a\| + \|y\|)$$

where

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

Proof: Define $\gamma(t) = \mathtt{Exp}_x(ta)$ and the curve

$$t \to c(r,t) = \mathtt{Exp}_{\gamma(t)}\left(r\Gamma_x^{\gamma(t)}(y + (1-t)a)\right)$$

for fixed r. Let $J_t^{\text{norm}}(r) = \frac{\mathrm{d}}{\mathrm{d}t} c(r,t)$, then it is shown in (Sun et al., 2019, Section B.3) that

$$d\left(\mathtt{Exp}_x(y+a), \mathtt{Exp}_z\left(\Gamma_x^z y\right)\right) \le \int_0^1 \|J_t^{\text{norm}}(1)\| \,\mathrm{d}t$$

Moreover, for fixed $t \in [0,1]$, let $\tilde{z} = \Gamma_x^{\gamma(t)}(y + (1-t)a)$ and $\rho_t = \|y + (1-t)a\| = \|\tilde{z}\|$, then its easy to see that $\|z\| \le \|y\| + \|a\|$, and the proof in (Sun et al., 2019, Section B.3) implies that

$$\|J_t^{\text{norm}}(1)\| \le \|J_t^{\text{norm}}(0)\| \, S_K(\rho_t) \le \frac{\|a\| \cdot \|y\|}{\rho_t} S_K(\rho_t) \le \frac{\|a\| \cdot \|y\|}{\|a\| + \|y\|} S_K(\|a\| + \|y\|),$$

where the last step follows from the observation that $r^{-1}S_K(r)$ is increasing in r, by Taylor's expansion. Hence the result follows. □

The following lemma gives an upper bound of $d_{w_{k+1}}(y_k, y_k')$ in terms of function S_k and a distance term.

Lemma 39 (restatement of Theorem 17) *We have for all $k \geq 1$ that*

$$d_{w_{k+1}}(y_k, y_k') \leq 2d^*(w_{k+1}; x_k, z_k) \cdot S_K\left(d(x_k, z_k) + d^*(w_{k+1}; x_k, z_k)\right)$$

where $d^*(w; x, z) = \min\left\{d(w, y) : y = \mathtt{Exp}_x(t \cdot \mathtt{Exp}_x^{-1}(z), t \in [0, 1]\right\}$ is the distance from w to the geodesic connecting x and z.

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

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

Suppose that w is a point on the geodesic connecting x_k and z_k, then

$$y_k = \mathtt{Exp}_w\left((1-\tau)\mathtt{Exp}_w^{-1}(x_k) + \tau \mathtt{Exp}_w^{-1}(z_k)\right)$$

We moreover define

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

Applying Theorem 38 with $x = w, z = w_{k+1}$ gives

$$d(y_k, y_k'') \leq d(w_{k+1}, w) \cdot S_K\left(d(x_k, z_k) + 2d(w_{k+1}, w)\right)$$

On the other hand,

$$\begin{aligned}
d_{w_{k+1}}(y_k', y_k'') &\leq (1-\tau) d\left(x_k, \mathtt{Exp}_{w_{k+1}}\left(\Gamma_w^{w_{k+1}}\mathtt{Exp}_w^{-1}(x_k) + \mathtt{Exp}_{w_{k+1}}^{-1}(w)\right)\right) \\
&\quad + \tau d\left(x_k, \mathtt{Exp}_{w_{k+1}}\left(\Gamma_w^{w_{k+1}}\mathtt{Exp}_w^{-1}(z_k) + \mathtt{Exp}_{w_{k+1}}^{-1}(w)\right)\right) \\
&\leq (1-\tau) d(w_{k+1}, w) \cdot S_K\left(d(w, x_k) + 2d(w, w_{k+1})\right) \\
&\quad + \tau d(w_{k+1}, w) \cdot S_K\left(d(w, z_k) + 2d(w, w_{k+1})\right) \\
&\leq d(w_{k+1}, w) \cdot S_K\left(d(x_k, z_k) + 2d(w, w_{k+1})\right)
\end{aligned}$$

Combining the above inequalities, we obtain

$$d_{w_{k+1}}(y_k, y_k') \leq 2d(w_{k+1}, w) \cdot S_K\left(d(x_k, z_k) + 2d(w, w_{k+1})\right)$$

The conclusion now follows from the definition of d^*. □

Corollary 40 *Suppose that*

$$12d^*(w_{k+1}; x_k, z_k) \cdot S_K\left(d(x_k, z_k) + 2d^*(w_{k+1}; x_k, z_k)\right) \leq (1-\sigma_k)\min\left\{\sqrt{\mu\lambda_k}, 1\right\} d_{w_{k+1}}(x_{k+1}, y_k), \tag{42}$$

then we have the following inequality which implies potential decrease:

$$p_k - p_{k+1} \geq \frac{(1-\sigma_k)A_{k+1}}{12\lambda_k} d_{w_{k+1}}^2(x_{k+1}, y_k) + \frac{\mu\theta_k a_{k+1} A_k}{2(A_k + \theta_k a_{k+1})} d_{w_{k+1}}^2(x_k, z_k). \tag{43}$$

The corollary can be seen as a generalized version of the potential-decrease result we obtained in Theorem 32. Indeed, when w_{k+1} lies on the geodesic between x_k and z_k, then the left hand side of (43) equals zero, so that (43) is guaranteed to hold.

We are now ready to prove our main result.

Theorem 41 (formal version of Theorem 18) *Assume f is L-smooth, and suppose that*

- $\sigma_k = \sigma \in (0,1)$.
- *The sequence $\{w_k\}$ satisfies $d^2(w_k, x^*) \leq \omega \max\{d(x_i, x^*), 0 \leq i \leq k; d(z_j, z^*), 0 \leq j \leq k-1\}$.*
- $d^*(w_{k+1}; x_k, z_k) \leq \rho_1 d_{w_{k+1}}(x_{k+1}, y_k)$ *and* $d^*(w_{k+1}; x_k, z_k) \leq \rho_2 \max\{d(x_k, x^*), d(z_k, x^*)\}$.
- *The step size $\lambda_k = \lambda = \frac{c^2}{L}$ where $c \in (0,1)$ is a fixed constant.*
- *The initialization satisfies $d(x_0, x^*) \leq \frac{\tau}{20} K^{-\frac{1}{2}} \left(\frac{\mu}{L}\right)^{\frac{3}{4}}$ and $B_0 = \frac{\mu}{2} A_0 > 0$, where*

$$\tau \leq \min\left\{\sqrt{\frac{c}{2(2\omega+5)}}, \sqrt{\frac{25(1-\sigma)c}{2\rho_1(7+10\rho_2^2)}}\right\}.$$

then for all $k \geq 0$, the following statements hold:

(1). Potential decrease (43) holds.

(2). $d^2(x_k, x^*) \leq \left(\frac{L}{\mu} + 1\right) d^2(x_0, x^*) \leq \frac{\tau^2}{200} K^{-1} \left(\frac{\mu}{L}\right)^{\frac{1}{2}}$.

(3). The distortion rate $\delta_k \leq 1 + \frac{2\omega+5}{10}\tau^2\sqrt{\frac{\mu}{L}}$.

(4). $\xi_k := \frac{a_k}{A_k} \geq \frac{9}{10}\sqrt{\frac{\mu\lambda}{1+\mu\lambda}}$ *and hence* $\frac{B_k}{A_k} \geq \frac{2}{5}\mu$.

(5). $d^2(z_k, x^*) \leq \frac{1}{80} K^{-1} \left(\frac{\mu}{L}\right)^{\frac{1}{2}}$.

Proof: We prove the result by induction on k. Specifically, for $k \geq 0$, we first prove (2),(3) and (5) hold for k, and then use them to derive (1),(4) for $k+1$, completing one round of induction step. When $k = 0$, (2) follows from

$$\delta_0 \leq 1 + 4Kd^2(w_0, z_0) \leq 1 + 8K\left(d^2(w_0, x^*) + d^2(z_0, x^*)\right) \leq 1 + \frac{\omega+1}{50}\tau^2\sqrt{\frac{\mu}{L}},$$

and the rest follows from the assumptions. Now suppose that the statements hold for $1, 2, \cdots, k-1$. Consider the case for k.

The induction hypothesis implies that $d^2(x_k, x^*) \leq \frac{\tau^2}{200} K^{-1}\sqrt{\frac{\mu}{L}}$, and

$$\begin{aligned} d^2_{w_k}(z_k, x^*) &\leq \frac{1}{B_k} p_0 \leq \frac{5}{2\mu}\frac{1}{A_0}\left(A_0(f(x_0) - f(x^*) + B_0 d^2_{w_0}(z_0, x^*)\right) \\ &\leq \frac{5}{2\mu}\left(\frac{L}{2}d^2(x_0, x^*) + \frac{\mu}{2}d^2(x_0, x^*)\right) \\ &= \frac{5}{4}\left(\frac{L}{\mu} + 1\right) d^2(x_0, x^*) \leq \frac{\tau^2}{160} K^{-1}\sqrt{\frac{\mu}{L}} \end{aligned}$$

On the other hand, since

$$d^2(w_k, x^*) \leq \omega d^2(x_k, x^*) \leq \frac{\omega\tau^2}{200} K^{-1} \left(\frac{\mu}{L}\right)^{\frac{1}{2}} < \frac{1}{2K}, \tag{44}$$

the distortion inequality (10) implies that

$$d^2(z_k, x^*) \leq (1 + 4Kd^2(w_k, x^*)) d^2_{w_k}(z_k, x^*) \leq \frac{\tau^2}{80} K^{-1} \sqrt{\frac{\mu}{L}}. \tag{45}$$

The inequalities (44) and (45) together implies that

$$d^2(z_k, w_k) \leq 2\left(d^2(w_k, x^*) + d^2(z_k, x^*)\right) \leq \frac{2\omega + 5}{50} \tau^2 K^{-1} \sqrt{\frac{\mu}{L}}$$

Hence, the distortion rate δ_k can be bounded as follows:

$$\delta_k \leq 1 + 4Kd^2(w_k, z_k) < 1 + \frac{2\omega + 5}{10} \tau^2 \sqrt{\frac{\mu}{L}} \leq 1 + \frac{c}{20} \sqrt{\frac{\mu}{L}} \leq 1 + \frac{1}{10} \sqrt{\frac{\mu\lambda}{1 + \mu\lambda}}.$$

The induction hypothesis implies that $\xi_k \geq \frac{9}{10} \sqrt{\frac{\mu\lambda}{1+\mu\lambda}} =: \xi_*$, and recall the equation

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

To show $\xi_{k+1} \geq \frac{9}{10} \sqrt{\frac{\mu\lambda}{1+\mu\lambda}}$, it suffices to show that

$$\delta_k \xi_* \left(\xi_* - \frac{\mu\lambda}{1 + \mu\lambda}\right) \leq \xi_*^2 (1 - \xi_*) \Leftrightarrow \delta_k \left(1 - \frac{10}{9} \xi_*\right) \leq 1 - \xi_*$$

The final equation holds since $\delta \leq 1 + \frac{1}{9}\xi_*$.

Now it remains to show potential decrease $p_{k+1} \leq p_k$; it suffices to prove that (42) holds. Since $S_K(r) \leq \frac{1}{3} K r^2$ when $Kr^2 \leq 1$, the assumptions imply that

$$\begin{aligned}
&12 d^*(w_{k+1}; x_k, z_k) \cdot S_K\left(d(x_k, z_k) + 2d^*(w_{k+1}; x_k, z_k)\right) \\
&\leq 4\rho_1 d_{w_{k+1}}(x_{k+1}, y_k) \cdot K \left(d(x_k, z_k) + 2d^*(w_{k+1}; x_k, z_k)\right)^2 \\
&\leq 4\rho_1 K \left(\frac{7}{100} \tau^2 K^{-1} + \frac{1}{10} \rho_2^2 \tau^2 K^{-1}\right) \sqrt{\frac{\mu}{L}} d_{w_{k+1}}(x_{k+1}, y_k) \\
&\leq \frac{1 - \sigma}{2} c \sqrt{\frac{\mu}{L}} d_{w_{k+1}}(x_{k+1}, y_k) \leq (1 - \sigma) \sqrt{\mu\lambda} d_{w_{k+1}}(x_{k+1}, y_k)
\end{aligned}$$

so that (42) holds. The proof is completed. □

Finally, we have the following corollary on acceleration for smooth functions.

Corollary 42 *Under the assumptions of Theorem 41, we have*

$$f(x_k) - f(x^*) \leq \frac{1}{A_k} p_0 \leq \frac{\tau^2}{400} K^{-1} L \left(\frac{\mu}{L}\right)^{\frac{3}{2}} \left(1 - \frac{9\sqrt{c}}{10\sqrt{2}} \sqrt{\frac{\mu}{L}}\right)^k$$

Appendix D. Details of Section 4

In this section, we provide detailed description of the algorithms we discussed in Section 4 and verification that they can be recovered from the Riemannian A-HPE framework. Throughout this section, we assume that f is L-smooth.

D.1. Algorithms without the additional distortion

First, we look at the *Riemannian Nesterov's method*, which is proposed and studied in Zhang and Sra (2018); Ahn and Sra (2020) and, to the best of our knowledge, the only provably accelerated method in our setting. The update of this method is given in Algorithm 3.

Algorithm 3 *Riemannian* Nesterov's Method

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

$z_0 \leftarrow x_0$ and $\lambda \leftarrow \frac{\sigma^2}{2L}$

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

choose a valid distortion rate δ_k according to Theorem 10

$\theta_k \leftarrow$ the smaller root of $B_k(1-\theta)^2 = \mu\lambda\theta\left((1-\theta)B_k + \frac{\mu}{2}\delta_k A_k\right)$

$B_{k+1} \leftarrow \frac{B_k}{\theta_k \delta_k}$, $a_{k+1} \leftarrow 2\mu^{-1}(1-\theta_k)B_{k+1}$ and $A_{k+1} \leftarrow A_k + a_{k+1}$

$y_k \leftarrow \mathtt{Exp}_{x_k}\left(\frac{\theta_k a_{k+1}}{A_k + \theta a_{k+1}} \mathtt{Exp}_{x_k}^{-1}(z_k)\right)$

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

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

end

Proposition 43 *Algorithm 3 can be recovered from Algorithm 2 by choosing* $\sigma_k = \sigma$, $\lambda \in \left(0, \frac{\sigma^2}{2L}\right)$, $w_{k+1} = y_k$, $x_{k+1} = \mathtt{Exp}_{y_k}(-\lambda_k \nabla f(y_k))$ *and* $v_{k+1} = \nabla f(y_k) + \mu \mathtt{Exp}_{y_k}^{-1}(x_{k+1})$.

Proof: It remains to check that the specified update rule satisfies the inequality (6) in the definition of *iprox*. Indeed we have

$$\begin{aligned}
\text{LHS} &= \frac{\lambda_k}{2(1+\lambda_k\mu)}\left(f(x_{k+1}) - f(y_k) - \left\langle \mathtt{Exp}_{y_k}^{-1}(x_{k+1}), \nabla f(y_k)\right\rangle\right) \\
&\quad + \left(\frac{\lambda_k^2\mu^2}{2(1+\lambda_k\mu)^2} - \frac{\lambda_k\mu}{2(1+\lambda_k\mu)}\right) d^2(y_k, x_{k+1}) \\
&\leq \frac{\lambda_k L}{2(1+\lambda_k\mu)} d^2(y_k, x_{k+1}) \leq \frac{\sigma_k^2}{2(1+\lambda_k\mu)^2} d^2(y_k, x_{k+1}) = \text{RHS}
\end{aligned}$$

so that the result follows. □

The second example is given in Algorithm 4. It is a direct generalization of the accelerated method (Chen and Luo, 2019, Algorithm 3) to Riemannian setting, and can be viewed as a variant of Nesterov's method with an additional gradient descent step. To the best of our knowledge, the algorithm is new and its convergence property is not known in Riemannian setting.

Algorithm 4 *Riemannian* Nesterov's method with an extra gradient step

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

$z_0 \leftarrow x_0$ and $\lambda \leftarrow \frac{\sigma^2}{2L}$

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

choose a valid distortion rate δ_k according to Theorem 10

$\theta_k \leftarrow$ the smaller root of $B_k(1-\theta)^2 = \mu\lambda\theta\left((1-\theta)B_k + \frac{\mu}{2}\delta_k A_k\right)$

$B_{k+1} \leftarrow \frac{B_k}{\theta_k \delta_k}$, $a_{k+1} \leftarrow 2\mu^{-1}(1-\theta_k)B_{k+1}$ and $A_{k+1} \leftarrow A_k + a_{k+1}$

$x_k \leftarrow \mathtt{Exp}_{\tilde{x}_k}(-\lambda \nabla f(\tilde{x}_k))$

$y_k \leftarrow \mathtt{Exp}_{x_k}\left(\frac{\theta_k a_{k+1}}{A_k + \theta a_{k+1}} \mathtt{Exp}_{x_k}^{-1}(z_k)\right)$

$\tilde{x}_{k+1} \leftarrow \mathtt{Exp}_{y_k}(-\lambda \nabla f(y_k))$

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

end

Proposition 44 *Algorithm 4 can be recovered from Algorithm 2 by choosing* $\sigma_k = \sigma = \frac{3}{4}$, $w_{k+1} = y_k$, $v_{k+1} = \nabla f(y_k) + \mu \mathtt{Exp}_{y_k}^{-1}(x_{k+1})$ *and* x_{k+1} *defined by*

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

Proof: It suffices to check that the *iprox* definition is satisfied. Smoothness implies that

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

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

$$\begin{aligned}
&\|\mathtt{Exp}_{y_k}^{-1}(x_{k+1}) + \lambda_k v_{k+1}\|^2 = \|(1+\mu\lambda_k)\mathtt{Exp}_{y_k}^{-1}(x_{k+1}) + \lambda_k \nabla f(y_k)\|^2 \\
&= (1+\mu\lambda_k)^2 d^2(x_{k+1}, y_k) + 2\lambda_k(1+\mu\lambda_k)\left\langle \mathtt{Exp}_{y_k}^{-1}(x_{k+1}), \nabla f(y_k)\right\rangle + \lambda_k^2 \|\nabla f(y_k)\|^2 \\
&\leq (1+\mu\lambda_k)^2 d^2(x_{k+1}, y_k) + \lambda_k^2 \|\nabla f(y_k)\|^2 \\
&\quad - 2\lambda_k(1+\mu\lambda_k)\left(f(y_k) - f(x_{k+1}) + \frac{\mu}{2} d^2(y_k, x_{k+1})\right) \\
&= (1+\mu\lambda_k) d^2(x_{k+1}, y_k) + d^2(y_k, \tilde{x}_{k+1}) \\
&\quad - 2\lambda_k(1+\mu\lambda_k)\left(\frac{1}{\lambda_k} - \frac{L}{2}\right)\left(d^2(y_k, \tilde{x}_{k+1}) + d^2(\tilde{x}_{k+1}, x_{k+1})\right) \\
&\leq (1+\mu\lambda_k) d^2(x_{k+1}, y_k) \\
&\quad - 2\left((1+\mu\lambda_k)\left(1 - \frac{L}{2}\lambda_k\right) - \frac{1}{2}\right)\left(d^2(y_k, \tilde{x}_{k+1}) + d^2(\tilde{x}_{k+1}, x_{k+1})\right) \\
&\leq \left(\frac{L}{2}\lambda_k(1+\mu\lambda_k) + \frac{1}{2}\right) d^2(x_{k+1}, y_k)
\end{aligned} \tag{46}$$

Finally, the choice of λ satisfies $L\lambda(1+\mu\lambda) \leq \frac{1}{2}$, hence the result follows. □

D.2. Algorithms with the additional distortion

In this section, we discuss specific examples of first-order methods that can be obtained from Algorithm 2 as special cases. The setting considered here is more general than the previous subsection,

in that we do not require that w_{k+1} is chosen on the geodesic connecting x_k and z_k, and we can apply Theorem 19 to obtain local (full) acceleration.

We first present a method, called *Riemannian accelerated extra-gradient descent (RAXGD)*, in Algorithm 5. To see its difference with Riemannian Nesterov's method, note that it uses two gradients each iteration. RAXGD can be seen as a Riemannian and strongly-convex version of the *accelerated extra-gradient method* proposed by Diakonikolas and Orecchia (2018). To the best of our knowledge, this method has not been proposed or studied before.

Algorithm 5 *Riemannian* accelerated extra-gradient descent

Input : Objective function f, initial point x_0, $\sigma_k \in (0,1)$, parameters L, μ, initial weight $A_0, B_0 > 0$

$z_0 \leftarrow x_0$ and $\lambda \leftarrow \frac{\sigma}{L}$

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

choose a valid distortion rate δ_k according to Theorem 10

$\theta_k \leftarrow$ the smaller root of $B_k(1-\theta)^2 = \mu\lambda_k\theta\left((1-\theta)B_k + \frac{\mu}{2}\delta_k A_k\right)$

$B_{k+1} \leftarrow \frac{B_k}{\theta_k\delta_k}$, $a_{k+1} = 2\mu^{-1}(1-\theta_k)B_{k+1}$ and $A_{k+1} = A_k + a_{k+1}$

$y_k \leftarrow \mathtt{Exp}_{x_k}\left(\frac{\theta_k a_{k+1}}{A_k+\theta a_{k+1}}\mathtt{Exp}_{x_k}^{-1}(z_k)\right)$

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

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

end

The following proposition shows that Algorithm 5 can be considered as a special case of Algorithm 2.

Proposition 45 *Algorithm 5 can be recovered from Algorithm 2 by choosing* $\sigma_k = \sigma \in (0,1)$, $\lambda \le \frac{\sigma}{L}$, $v = \nabla f(x_{k+1})$ *and* $w_{k+1} = x_{k+1} = \mathtt{Exp}_{y_k}(-\lambda_k\nabla f(y_k))$. *Moreover, the conditions in Theorem 41 are satisfied with* $\rho_1 = \rho_2 = \omega = 1$.

Proof: We have $(x_{k+1}, v_{k+1}) \in \mathtt{iprox}_f^{w_{k+1}}(y_k, \lambda_k, \varepsilon_k)$, since

$$\begin{aligned}\left\|\mathtt{Exp}_{x_{k+1}}^{-1}(y_k) - \lambda_k\nabla f(x_{k+1})\right\| &= \lambda_k\left\|\Gamma_{y_k}^{x_{k+1}}\nabla f(y_k) - \nabla f(x_{k+1})\right\| \\ &\le L\lambda_k d(y_k, x_{k+1}) \le \sigma_k d(x_{k+1}, y_k).\end{aligned}$$

Finally, note that

$$d^*(w_{k+1}; x_k, z_k) \le d(x_{k+1}, y_k) \le \frac{1}{L}\left\|\nabla f(y_k)\right\| \le d(y_k, x^*),$$

the conclusion follows. □

We can also design new accelerated algorithms by choosing different realizations of the *iprox* operator in Algorithm 2. This can lead to novel algorithms that are previously unknown even in Euclidean setting. In the following we derive from Algorithm 2 a generalized version of RAXGD, given in Algorithm 6.

Rather than obtain x_{k+1} directly from a gradient descent step, it allows arbitrary choices of x_{k+1} as long as a distance inequality

$$d(w_{k+1}, x_{k+1}) \le \frac{1-\sigma}{3}d(y_k, x_{k+1}) \tag{47}$$

Algorithm 6 Generalized Riemannian accelerated extra-gradient descent

Input : Objective function f, initial point x_0, $\sigma_k \in (0,1)$, parameters L, μ, initial weight $A_0, B_0 > 0$
$z_0 \leftarrow x_0$ and $\lambda \leftarrow \frac{\sigma}{2L}$
for $k = 0, 1, \cdots$ **do**
choose a valid distortion rate δ_k according to Theorem 10
$\theta_k \leftarrow$ the smaller root of $B_k(1-\theta)^2 = \mu\lambda_k\theta\left((1-\theta)B_k + \frac{\mu}{2}\delta_k A_k\right)$
$B_{k+1} \leftarrow \frac{B_k}{\theta_k\delta_k}$, $a_{k+1} = 2\mu^{-1}(1-\theta_k)B_{k+1}$ and $A_{k+1} = A_k + a_{k+1}$
$y_k \leftarrow \mathtt{Exp}_{x_k}\left(\frac{\theta_k a_{k+1}}{A_k+\theta a_{k+1}}\mathtt{Exp}_{x_k}^{-1}(z_k)\right)$
$w_{k+1} \leftarrow \mathtt{Exp}_{y_k}(-\lambda\nabla f(y_k))$
choose x_{k+1} such that $d(w_{k+1}, x_{k+1}) \leq \frac{1-\sigma}{3}d(y_k, x_{k+1})$
$z_{k+1} \leftarrow \mathtt{Exp}_{x_{k+1}}\left(\theta_k\mathtt{Exp}_{x_{k+1}}^{-1}(z_k) - \mu^{-1}(1-\theta_k)\nabla f(x_{k+1})\right)$
end

is satisfied. The inequality is obviously satisfied when $x_{k+1} = w_{k+1}$, which reduces to Algorithm 5. Intuitively, (47) implies that x_{k+1} is obtained by starting from y_k and following an 'approximately descent' direction. In Euclidean setting, the solution set of (47) for x_{k+1} is a region enclosed by an Apollonius circle that contains w_{k+1}.

Proposition 46 *Algorithm 6 is a special case of Algorithm 2. Moreover, the conditions in Theorem 41 holds with* $\rho_1 = 4, \rho_2 = 1$ *and* $\omega = \frac{3}{2}$.

Proof: To check that Algorithm 6 can be obtained from Algorithm 2, it suffices to verify that the update of x_{k+1} satisfies (6).

Since f is L-smooth, we have

$$\frac{\lambda}{1+\mu\lambda}\left(f(x_{k+1}) - f_{w_{k+1}}(x_{k+1})\right) \leq \frac{L\lambda}{2}d^2(x_{k+1}, w_{k+1}) \leq \frac{1}{2}d^2(x_{k+1}, y_k).$$

On the other hand

$$\begin{aligned}
&\|\mathtt{Exp}_{w_{k+1}}^{-1}(x_{k+1}) - \mathtt{Exp}_{w_{k+1}}^{-1}(y_k) + \lambda v_{k+1}\|^2 \\
&= \|(1+\mu\lambda)\mathtt{Exp}_{w_{k+1}}^{-1}(x_{k+1}) - \mathtt{Exp}_{w_{k+1}}^{-1}(y_k) + \lambda\nabla f(w_{k+1})\|^2 \\
&\leq 2(1+\mu\lambda)^2 d^2(w_{k+1}, x_{k+1}) + 2\lambda^2\left\|\nabla f(w_{k+1}) - \Gamma_{y_k}^{w_{k+1}}\nabla f(y_k)\right\|^2 \\
&\leq 2(1+\mu\lambda)^2 d^2(w_{k+1}, x_{k+1}) + 2L^2\lambda^2 d^2(w_{k+1}, y_k) \\
&\leq 2\left((1+\mu\lambda)^2 + 2L^2\lambda^2\right)d^2(w_{k+1}, x_{k+1}) + 4L^2\lambda^2 d^2(x_{k+1}, y_k) \\
&\leq 3(1+\mu\lambda)^2 d^2(w_{k+1}, x_{k+1}) + \sigma^2 d^2(x_{k+1}, y_k) \\
&\leq (1+\mu\lambda)^2 d^2(x_{k+1}, y_k)
\end{aligned}$$

where the last step uses (47). Hence (6) holds. □

By Theorem 41, we deduce that Algorithm 6 achieves acceleration when initialized in an $\mathcal{O}\left(K^{-\frac{1}{2}}\left(\frac{\mu}{L}\right)^{\frac{3}{4}}\right)$. To the best of our knowledge, Algorithm 6 has not been studied even for strongly-convex functions in the Euclidean setting.

We emphasize that the purpose of introducing Algorithm 6 is to show that Algorithm 2 can lead to many different types of accelerated first-order methods. There are of course other ways to specify the *iprox* operator, which would lead to many interesting algorithms.

D.3. Discussion of the extra-point framework in Huang and Zhang (2021)

In a recent work Huang and Zhang (2021), the authors propose an extra-point approach motivated by the analysis of classical accelerated methods. Based on this idea, they propose a framework for smooth strongly-convex optimization, which is in a quite general form and contains a total of 9 parameters. For convenience we give the detailed updates of their framework below.

$$p_k \leftarrow t_1 x_k + t_2 z_k \tag{48a}$$
$$y_k \leftarrow \text{a solution of } \langle \nabla f(y_k), p_k - y_k \rangle \geq 0 \tag{48b}$$
$$\widetilde{x}_{k+1} \leftarrow y_k - \frac{t_3}{L} \nabla f(y_k) \tag{48c}$$
$$x_{k+1} \leftarrow y_k - \frac{t_4}{L} \nabla f(\widetilde{x}_{k+1}) - \frac{t_5}{L} \left(\nabla f(\widetilde{x}_{k+1}) - \nabla f(y_k) \right) + t_6 \left(\widetilde{x}_{k+1} - y_k \right) \tag{48d}$$
$$z_{k+1} \leftarrow t_7 z_k + t_8 y_k - t_9 \nabla f(y_k) \tag{48e}$$

The authors derive sufficient conditions on the choice of $t_i, 1 \leq i \leq 9$ so that (48) can achieve acceleration. While their framework looks complicated, in the following we show that it can be interpreted quite naturally from the PPM viewpoint introduced in Section 2.

First, (48e) is very similar to the update of z_{k+1} in A-HPE; one can see this by comparing it with (17), with the choice $w_{k+1} = y_k$ and $v_{k+1} = \nabla f(y_k) + \mu (x_{k+1} - y_k)$. With properly chosen constants t_7, t_8, t_9, (48e) can then be interpreted as an approximate PPM scheme.

Second, (48c) and (48d) together give a gradient-descent-type update formula of x_{k+1}. In particular, (48d) can also be written as

$$x_{k+1} \leftarrow y_k - \frac{t_3 t_6 - t_5}{L} \nabla f(y_k) - \frac{t_4 + t_5}{L} \nabla f(\widetilde{x}_{k+1}),$$

which is very similar to the Riemannian Nesterov's method with multiple gradient steps that we introduced in Algorithm 4. As a result, the update of x_{k+1} can also be interpreted as another approximate PPM scheme.

Recall the arguments in Section 2 that the final step is to combine these two schemes and obtain potential decrease. In A-HPE this is implemented by a simple convex combination of the iterates x_k and z_k. However, in (48) the procedure is more complex: first a convex combination is obtained (i.e. the update of p_k), and then y_k is chosen to be any solution of the inequality (48b).

This procedure, in fact, can be easily justified by one additional step in the analysis: intuitively, y_k is a *refinement* of the convex combination. Specifically, as argued in the proof of Theorem 24, the combination of two PPM approaches is implemented by the following inequality:

$$\theta_z \|z_k - x_{k+1} + \mu^{-1} v_{k+1}\|^2 + \theta_x \|x_k - x_{k+1} + \mu^{-1} v_{k+1}\|^2 \geq (\theta_z + \theta_x) \|p_k - x_{k+1} + \mu^{-1} v_{k+1}\|^2.$$

Since $x_{k+1} - \mu^{-1} v_{k+1} = y_k - \mu^{-1} \nabla f(y_k)$, we have

$$\|p_k - x_{k+1} + \mu^{-1} v_{k+1}\|^2 = \|p_k - y_k + \mu^{-1} \nabla f(y_k)\|^2 \geq \|\nabla f(y_k)\|^2 = \|y_k - x_{k+1} + \mu^{-1} v_{k+1}\|^2,$$

where the inequality exactly follows from (48b)!

Now we have seen that the framework of (Huang and Zhang, 2021) uses the same idea of approximate-PPM as A-HPE, except that the combination step is more general. On the other hand, the framework is limited to the choice of $w_{k+1} = y_k$ in the definition of *iprox*, while A-HPE allows more flexible choices.

Finally, we provide a natural extension of the framework to the Riemannian setting:

$$
\begin{aligned}
p_k &\leftarrow \mathtt{Exp}_{x_k}\left(t_2 \mathtt{Exp}_{x_k}^{-1}(z_k)\right) \\
y_k &\leftarrow \text{a solution of } \left\langle \nabla f(y_k), \mathtt{Exp}_{y_k}^{-1}(p_k) \right\rangle \geq 0 \\
\widetilde{x}_{k+1} &\leftarrow \mathtt{Exp}_{y_k}\left(-\frac{t_3}{L}\nabla f(y_k)\right) \\
x_{k+1} &\leftarrow \mathtt{Exp}_{\widetilde{x}_{k+1}}\left((1-t_6)\mathtt{Exp}_{\widetilde{x}_{k+1}}^{-1}(y_k) - \frac{t_4}{L}\nabla f(\widetilde{x}_{k+1}) - \frac{t_5}{L}\left(\nabla f(\widetilde{x}_{k+1}) - \Gamma_{y_k}^{\widetilde{x}_{k+1}}\nabla f(y_k)\right)\right) \\
z_{k+1} &\leftarrow \mathtt{Exp}_{y_k}\left(t_7 \mathtt{Exp}_{y_k}^{-1}(z_k) - t_9 \nabla f(y_k)\right)
\end{aligned}
$$

Local acceleration of the framework can be shown using our in approach Section 3. For the special case $y_k = p_k$, the additional distortion disappears and the framework attains global eventual acceleration.