From Nesterov's Estimate Sequence to Riemannian Acceleration

Kwangjun Ahn KJAHN@MIT.EDU

Suvrit Sra SUVRIT@MIT.EDU

Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology

Editors: Jacob Abernethy and Shivani Agarwal

Abstract

We propose the first global accelerated gradient method for Riemannian manifolds. Toward establishing our results, we revisit Nesterov's estimate sequence technique and develop a conceptually simple alternative from first principles. We then extend our analysis to Riemannian acceleration, localizing the key difficulty into *"metric distortion."* We control this distortion via a novel geometric inequality, which enables us to formulate and analyze global Riemannian acceleration.

1. Introduction

Non-convex optimization is in general intractable. But occasionally, special problem structure can enable tractability. An important instance of such structure is that of *geodesic convexity* (*g-convexity*), a generalization of convexity that is defined along geodesics in a metric space (Gromov, 1978; Burago et al., 2001; Bridson and Haefliger, 2013). Tractability through the lens of g-convexity has been fruitful in several applications (e.g., see (Zhang and Sra, 2016, §1.1)) and also some purely theoretical questions (Bürgisser et al., 2019; Goyal and Shetty, 2019) (see also §1.2 of this paper).

Paralleling the theory and applications of g-convexity is the progress on algorithms, primarily set in Riemannian manifolds (Udriste, 1994; Absil et al., 2009) and CAT(0) spaces (Bacák, 2014). Earlier studies focus on *asymptotic* analysis, while Zhang and Sra (2016) obtain the first *non-asymptotic* iteration complexity analysis for Riemannian (stochastic) gradient methods. Subsequent works establish iteration complexity for Riemannian proximal-point methods (Bento et al., 2017), Frank-Wolfe (Weber and Sra, 2019), variance reduced methods (Zhang et al., 2016; Kasai et al., 2016; Zhang et al., 2018; Zhou et al., 2019), trust-region methods (Agarwal et al., 2018), among others.

Despite this progress, a landmark result of Euclidean optimization has eluded the Riemannian setting: namely, a Riemannian analog of Nesterov's accelerated gradient method (Nesterov, 1983). This gap motivates the central question of our paper:

Is it possible to develop accelerated gradient methods for Riemannian manifolds?

This natural question turns out to be highly non-trivial: Nesterov's analysis relies deeply on the linear structure of Euclidean space, and recent efforts could make only *partial* progress–see §1.2 for details.

1.1. Overview of our main results

We take a major step toward answering the above question by developing the *first global accelerated first-order method* for Riemannian manifolds, informally stated as Theorem 1.1; the formal statement is Theorem 4.1. Toward establishing Theorem 1.1 we first revisit Nesterov's (Euclidean) estimate sequence technique (§2) and develop an alternative analysis based on *potential functions* (*Lyapunov functions* (Lyapunov, 1992)). See §5 for precise positioning of our approach within existing work.

Theorem 1.1 (Informal) *Let f be L-smooth and μ-strongly convex in a geodesic sense. Then, there exists a* computationally tractable *optimization algorithm satisfying*

$$f(x_t) - f(x_*) = O\left((1-\xi_1)(1-\xi_2)\cdots(1-\xi_t)\right),$$

*where $\{\xi_t\}$ satisfies (i) $\{\xi_t\}_{t\geq 1} > \mu/L$ (**strictly faster** than gradient descent); and (ii) $\exists \lambda \in (0,1)$ such that $\forall \epsilon > 0$, $|\xi_t - \sqrt{\mu/L}| \leq \epsilon$, for $t \geq \Omega\big(\frac{\log(1/\epsilon)}{\log(1/\lambda)}\big)$ (eventually achieves **full acceleration**).*

Remarkably, the parameters of the algorithm determined—from first principles—by our analysis *exactly* satisfy the complicated recursive relations derived by Nesterov, thereby offering a simple, new alternative to his techniques (§2.3). Moreover, we develop a simple fixed-point iteration that reveals how accelerated convergence rates can be obtained from such complicated recursive relations (§2.4), again providing an elementary alternative to Nesterov's original analysis.

Building on this new viewpoint, we extend our approach to the Riemannian setting (§3 and §4). Here, we introduce a crucial but *a priori* non-obvious modification to the potential function (§3.2). Specifically, we propose using "*projected distances*" instead of Riemannian distances in the potential function, which helps us localize the main difficulty caused by Riemannian geometry into "metric distortion." Already for the simplified setting of constant metric distortion, our analysis implies the local acceleration results of (Zhang and Sra, 2018) (Corollary 3.2). To tackle global acceleration, we establish a novel metric distortion inequality based on comparison theorems in Riemannian geometry (§4.1). We then show how distortion can be estimated at each iteration based, which proves critical to obtain a computationally tractable algorithm (Algorithm 1). We show that distortion decreases over iterations (§4.2), which ultimately leads to Theorem 1.1 (formal result, Theorem 4.1).

1.2. Related work

A few recent works also seek to answer the main question of this paper. The first attempt (Liu et al., 2017) reduces the task to solving nonlinear equations, but it is unclear whether these equations are even feasible or tractably solvable. Alimisis et al. (2020) establish a Riemannian analog of the differential-equation approach to acceleration (Su et al., 2014), and they analyze second-order ODEs on Riemannian manifolds. Then, they employ discretization from the Euclidean case (Betancourt et al., 2018; Shi et al., 2019) to derive first-order methods. But it is unclear whether these methods achieve acceleration, as such discretization *does not* directly yield Nesterov's method even in the Euclidean case. Moreover, as we shall see (Remark 4.2), their global control of metric distortion cannot capture *full* acceleration; one must control metric distortions *locally*. See § 4.1 for details.

The most concrete progress is in (Zhang and Sra, 2018) that proves accelerated convergence, albeit only *locally* in a neighborhood whose radius vanishes as the condition number and the curvature bound grow. They do not characterize how the algorithm behaves outside such a local neighborhood, in stark contrast with our *global* acceleration result. See §3.2 for a detailed comparison.

2. Warm up in the Euclidean case: alternative analysis of Nesterov's optimal method

As a building block for the Riemannian setting, let us revisit the Euclidean setting. In particular, we consider Nesterov's optimal method which is derived based on an ingenious construction called an *estimate sequence* (2018, Ch. 2.2.1): For $t \geq 0$, the iterates are updated as

$$x_{t+1} \leftarrow y_t + \alpha_{t+1}(z_t - y_t) \tag{2.1a}$$

$$y_{t+1} \leftarrow x_{t+1} - \gamma_{t+1}\nabla f(x_{t+1}) \tag{2.1b}$$

$$z_{t+1} \leftarrow x_{t+1} + \beta_{t+1}(z_t - x_{t+1}) - \eta_{t+1}\nabla f(x_{t+1}), \tag{2.1c}$$

for given initial iterates $y_0 = z_0 \in \mathbb{R}^n$. This construction yields optimal first-order methods that achieve the lower bounds under the black-box complexity model (Nemirovski and Yudin, 1983). Note that the updates (2.1) can be also derived without resorting to estimate sequences: for instance, see Appendix A for a derivation based on the *linear coupling* framework due to Allen-Zhu and Orecchia (2017) and see (Ahn, 2020) for a derivation based on the proximal point method.

Despite its fundamental nature, there is a well-known puzzling aspect of Nesterov's construction: To guarantee the standing assumption of the estimate sequence technique (2018, (2.2.3)), Nesterov's original analysis (2018, page 87) finds complicated recursive relations between parameters $\alpha, \beta, \gamma, \eta$ via some non-trivial algebraic "tricks." These tricks are carried out in a fortuitous manner, obscuring the driving principle and the scope of the underlying technique. Notably, Zhang and Sra (2018) favor estimate sequences over other approaches, but still achieve only local acceleration.

Therefore, in our search for global acceleration, we first revisit Euclidean acceleration from first-principles. In particular, we provide an alternative analysis of iteration (2.1) that sheds new light on the scope of Nesterov's original analysis. Our analysis employs a *potential function*[1], a classical tool from control theory (Lyapunov, 1992) that has received a resurgence of interest recently (see §5). Roughly, the potential-function analysis proceeds as follows:

1. **Choose potential:** First, choose an error measure $\mathcal{E}_t$ that "measures" how close the iterates at step t are to the optimal solution; then define the potential function as $\Phi_t := A_t \mathcal{E}_t$.
2. **Ensure potential decrease:** Choose an increasing sequence A_t so that Φ_t is decreasing.

Once Φ_t is chosen as above, it implies that $\mathcal{E}_t \leq \mathcal{E}_0 / A_t$, yielding a convergence rate of $O\left(1/A_t\right)$.

2.1. Choosing the potential function

The key to potential function based analysis is to choose the "correct" performance measure. For an iterate u_t at step t, two prototypical choices are (i) the suboptimality $\mathcal{E}_t = f(u_t) - f(x_*)$; and (ii) the distance to an optimal point $\|u_t - x_*\|$. Indeed, many existing analyses correspond to choosing either one as the performance measure, as explicitly noted in (Bansal and Gupta, 2019).

For iteration (2.1), it turns out that a weighted sum of the suboptimality $f(y_t) - f(x_*)$ and the distance $\|z_t - x_*\|^2$ is the "correct" performance measure, i.e., we choose the potential function as

$$\Phi_t := A_t \cdot (f(y_t) - f(x_*)) + B_t \cdot \|z_t - x_*\|^2, \tag{2.2}$$

for some $A_t > 0$ and $B_t \geq 0$. By taking a weighted sum of the two measures, this performance measure does not require either one to be monotonically decreasing over iterations. This property, also known as *non-relaxational property*, was a key innovation in Nesterov's landmark work (1983). Why we choose y_t for the cost and z_t for the distance will become clearer soon (see Remark 2.3).

We note that the current form of the potential (2.2) is not new; it also appears in prior works (Wilson et al., 2016; Diakonikolas and Orecchia, 2019; Bansal and Gupta, 2019), although with different motivations; see §5 for a detailed perspective; see also Appendix A for additional connections.

2.2. Potential difference calculations

Having chosen the potential function (2.2), the main goal now is to choose the parameters A_{t+1}, B_{t+1}, α_{t+1}, β_{t+1}, γ_{t+1}, η_{t+1} so that the potential decreases, i.e., $\Phi_{t+1} - \Phi_t \leq 0$. To that end, we

[1] Also known as *Lyapunov function* in control theory or *invariant* in theoretical computer science and mathematics.

first express the potential difference $\Phi_{t+1} - \Phi_t$ more simply and derive a manageable upper bound using *first principles*. Using definition (2.2), the difference $\Phi_{t+1} - \Phi_t$ can be split into two parts:

$$A_{t+1} \cdot (f(y_{t+1}) - f(x_*)) - A_t \cdot (f(y_t) - f(x_*)) \tag{2.3}$$

$$+ \; B_{t+1} \cdot \|z_{t+1} - x_*\|^2 - B_t \cdot \|z_t - x_*\|^2 \,. \tag{2.4}$$

Since $\alpha, \beta, \gamma, \eta$ will only appear with index $t+1$, we drop their subscripts for simplicity. We first relate the terms for step $t+1$ with those for step t. To do that, we recast (2.1). Using the notation $\mathsf{Grad}_{s \cdot \nabla}(x) := x - s \cdot \nabla$, we rewrite the updates (2.1b) and (2.1c) as

$$y_{t+1} = \mathsf{Grad}_{\gamma \cdot \nabla f(x_{t+1})}(x_{t+1}) \tag{2.1b$'$}$$

$$z_{t+1} = \mathsf{Grad}_{\eta \cdot \nabla f(x_{t+1})}(x_{t+1} + \beta(z_t - x_{t+1}))\,, \tag{2.1c$'$}$$

respectively. Now the difference between (2.1b$'$) and (2.1c$'$) is clear: the first is an *exact* gradient step in the sense that $\nabla = \nabla f(x)$, while the second step is *inexact*. Hence, in relating the terms for step $t+1$ with those for step t, we need to invoke different analyses for two different gradient steps.

We begin with two folklore results for gradient steps corresponding to exact and inexact steps.

Proposition 2.1 (Descent lemma) *Assume $\nabla = \nabla f(x)$, and let $y = \mathsf{Grad}_{s \cdot \nabla}(x)$. If f is L-smooth, then the gradient step decreases cost: $f(y) - f(x) \leq -s\,(1 - Ls/2)\,\|\nabla\|^2$.*

Proposition 2.2 *Let $z = \mathsf{Grad}_{s \cdot \nabla}(x)$. Then, for any x_*, $\|z - x_*\|^2 - \|x - x_*\|^2 = s^2\,\|\nabla\|^2 + 2s\,\langle \nabla, x_* - x \rangle$, i.e., (inexact) gradient step decreases the distance to x_* as long as direction $-\nabla$ is well aligned with the vector $x_* - x$ and has sufficiently small norm.*

Remark 2.3 *The two steps above reveal why we use y_t for the cost term and z_t for the distance term in (2.2): Proposition 2.1 deals with the cost, while Proposition 2.2 deals with the distance.*

Now we apply Proposition 2.1 to (2.1b$'$) and Proposition 2.2 to (2.1c$'$). For clarity, we denote:

$$\Delta_\gamma := \gamma(1 - L\gamma/2)\,, \quad \nabla := \nabla f(x_{t+1})\,, \quad X := x_{t+1} - x_*\,, \text{ and } W := z_t - x_{t+1}\,. \tag{2.5}$$

With this notation, Propositions 2.1 and 2.2 imply: $f(y_{t+1}) \leq f(x_{t+1}) - \Delta_\gamma\,\|\nabla\|^2$ and $\|z_{t+1} - x_*\|^2 = \|X + \beta W\|^2 + \eta^2\,\|\nabla\|^2 - 2\eta\,\langle \nabla, X + \beta W \rangle$. Plugging these two back into to (2.3) and (2.4), one can derive the following upper bound on $\Phi_{t+1} - \Phi_t$ in terms of the vectors ∇, X, W from *first principles* (i.e., using only smoothness and (strong) convexity; see Appendix E.1):

$$\Phi_{t+1} - \Phi_t \leq C_1 \cdot \|W\|^2 + C_2 \cdot \|X\|^2 + C_3\,\|\nabla\|^2 + C_4 \cdot \langle W, X \rangle + C_5 \cdot \langle W, \nabla \rangle + C_6 \cdot \langle X, \nabla \rangle\,, \tag{2.6}$$

$$\text{where } \begin{cases} C_1 := \beta^2 B_{t+1} - B_t - \frac{\mu}{2}\frac{\alpha^2}{(1-\alpha)^2} A_t\,, & C_2 := B_{t+1} - B_t - \frac{\mu}{2}(A_{t+1} - A_t)\,, \\ C_3 := \eta^2 B_{t+1} - \Delta_\gamma \cdot A_{t+1}\,, & C_4 := 2 \cdot (\beta B_{t+1} - B_t)\,, \\ C_5 := \frac{\alpha}{1-\alpha} A_t - 2\beta\eta B_{t+1}\,, \quad \text{and} & C_6 := (A_{t+1} - A_t) - 2\eta B_{t+1}\,. \end{cases}$$

Notice that the three vectors ∇,X,W are rooted at x_{t+1}. This choice is deliberate; it proves crucial in the Riemannian case where we will need them to lie in the same *tangent space*. See Appendix E.3.

2.3. Ensuring potential decrease

Having established the bound (2.6), our goal is to now choose $A_{t+1}, B_{t+1}, \alpha, \beta, \gamma, \eta$ given A_t, B_t so that (2.6) is non-positive (recall that we have dropped indices of $\alpha_{t+1}, \beta_{t+1}, \gamma_{t+1}, \eta_{t+1}$). In general,

it is difficult to ensure non-positivity of a symbolic expression; but since (2.6) is a quadratic form, one avenue might be to turn it into a *negative sum of squares* ("−SoS"). The simplest strategy to make it "−SoS" would be to try to make the coefficients C_4, C_5, C_6 of the cross terms 0, while making C_1, C_2, C_3 non-positive. It turns out this strategy *fully* determines the parameters, as follows:

- *Coefficients of cross terms characterize α, β, η in terms of A_{t+1}, B_{t+1}:* From $C_6 = 0$, we get $\eta = (A_{t+1} - A_t)/(2B_{t+1})$, and from $C_4 = 0$, we get $\beta = B_t/B_{t+1}$. Plugging these choices into $C_5 = 0$, we obtain the equation $\frac{\alpha}{1-\alpha} A_t = (A_{t+1} - A_t)B_t/B_{t+1}$. To summarize:

$$\eta = \tfrac{A_{t+1}-A_t}{2B_{t+1}}\,, \quad \beta = \tfrac{B_t}{B_{t+1}} \quad \text{and} \quad \tfrac{\alpha}{1-\alpha} = \tfrac{(A_{t+1}-A_t)B_t}{A_t B_{t+1}}\,. \tag{2.7}$$

- *For a fixed γ, coefficients of squared terms determines A_{t+1}, B_{t+1} based on given A_t, B_t:* Beginning with $C_3 \le 0$, we replace η with the one from (2.7) to obtain the inequality

$$(A_{t+1} - A_t)^2/(4\Delta_\gamma \cdot A_{t+1}) \le B_{t+1}\,. \tag{2.8}$$

  Plugging (2.8) into $C_2 \le 0$, we get an inequality only in terms of A_{t+1} (assuming γ is fixed):

$$(A_{t+1} - A_t)^2/(4\Delta_\gamma \cdot A_{t+1}) - (A_{t+1} - A_t)\tfrac{\mu}{2} \le B_t\,. \tag{2.9}$$

  Recall that we need to choose A_{t+1} as large as possible; it turns out that the largest possible A_{t+1} satisfies (2.9) with equality (hence (2.8) as well). To see why, let us follow Nesterov's notation and use the *suboptimality shrinking ratio* $1 - \xi := {}^{A_t}/_{A_{t+1}}$.[2] With this, inequality (2.9) becomes:

$$\xi(\xi - 2\mu\Delta_\gamma)/(1 - \xi) \le 4\Delta_\gamma \cdot B_t/A_t\,. \tag{2.10}$$

  In (2.10) note that the RHS is a nonnegative constant (assuming $\Delta_\gamma > 0$ is already chosen) and the LHS is an increasing function on $[2\mu\Delta_\gamma, 1)$ whose value is 0 at $2\mu\Delta_\gamma$ and approaches $+\infty$ as $\xi \to 1$. Hence, the largest ξ (equivalently, the largest A_{t+1}) satisfies (2.10) (or equivalently, (2.9)) with equality. Consequently, this choice of ξ also satisfies (2.8) with equality. One can then verify that this choice satisfies $\beta^2 B_{t+1} \le B_t$ and hence implies $C_1 \le 0$ (see Appendix E.2).

- *Lastly, choose γ from $(0, 2/L)$:* Now the last variable to determine is γ. The above calculations are valid as long as $\Delta_\gamma > 0$, so we can arbitrarily choose γ in $(0, 2/L)$. Note that most accelerated methods in the literature choose $\gamma = 1/L$ since it is the maximizer of Δ_γ.

Combining the above identities, we can express $A_{t+1}, B_{t+1}, \alpha, \beta, \eta$ in terms of ξ, γ, μ, L; see Appendix E.2. After recovering the indices of $\alpha, \beta, \gamma, \eta$, our findings can be summarized in the main result of this section as follows (see Appendix E.2 for the proof):

Theorem 2.1 (Parameter choice for potential decrease) *Given y_t, z_t and $A_t, B_t > 0$ and $\gamma_{t+1} \in (0, 2/L)$, let $\Delta_\gamma := \gamma_{t+1}(1 - L\gamma_{t+1}/2)$ and $\xi_t := \sqrt{4\Delta_\gamma \cdot {}^{B_t}/_{A_t}}$. Then, choose parameters as per:*

1. *Compute $\xi_{t+1} \in [2\mu\Delta_\gamma, 1)$ satisfying $\frac{\xi_{t+1}(\xi_{t+1} - 2\mu\Delta_\gamma)}{1-\xi_{t+1}} = \xi_t^2$.* (2.11)
2. *Determine parameters based on ξ_{t+1}: $A_{t+1} = \frac{A_t}{1-\xi_{t+1}}$, $B_{t+1} = \frac{\xi_{t+1}^2}{1-\xi_{t+1}} \cdot \frac{A_t}{4\Delta_\gamma}$, $\alpha_{t+1} = \frac{\xi_{t+1} - 2\mu\Delta_\gamma}{1 - 2\mu\Delta_\gamma}$, $\beta_{t+1} = 1 - 2\mu\Delta_\gamma \xi_{t+1}^{-1}$, and $\eta_{t+1} = 2\Delta_\gamma \xi_{t+1}^{-1}$.*

Then, y_{t+1}, z_{t+1} defined as per iteration (2.1) satisfy $\Phi_{t+1} \le \Phi_t$ (see (2.2)), or equivalently,

$$f(y_{t+1}) - f(x_*) + \tfrac{\xi_{t+1}^2}{4\Delta_\gamma} \cdot \|z_{t+1} - x_*\|^2 \le (1 - \xi_{t+1}) \cdot \left[f(y_t) - f(x_*) + \tfrac{\xi_t^2}{4\Delta_\gamma} \cdot \|z_t - x_*\|^2\right].$$

[2] Nesterov's analysis finds $\alpha_i \in (0, 1)$ s.t. $f(y_t) - f(x_*) \le \prod_{i=1}^t (1 - \alpha_i) \cdot \left[f(x_0) - f(x_*) + C\,\|x_0 - x_*\|^2\right]$ for a constant $C > 0$ and iterate x_0 (2018, Thm. 2.2.1)). These α_i's exactly correspond to our suboptimality shrinking ratios.

Remarkably the parameter choices obtained by Theorem 2.1 *exactly match* those of Nesterov's "General Scheme for Optimal Method" (2018, (2.2.1)). Hence, our approach recovers Nesterov's optimal method that encompasses both strongly and non-strongly convex costs, *without* requiring the estimate sequence technique. Another byproduct of our analysis is the convergence of z_t to x_* for $\mu > 0$ (in which case, $\xi > 0$), a property otherwise proved via additional analysis (see e.g., (Gasnikov and Nesterov, 2018, Corollary 1)). This convergence plays a crucial role in the Riemannian setting (see §4.2). Observe that upon applying Theorem 2.1 recursively, we can deduce that

$$f(y_t) - f(x_*) = O\left((1-\xi_1)(1-\xi_2)\cdots(1-\xi_t)\right). \tag{2.12}$$

Thus, to identify the convergence rate of iteration (2.1) with parameters chosen via Theorem 2.1, we only need to study how the sequence $\{\xi_t\}$ evolves. This evolution is the focus of the next subsection.

2.4. Identifying the convergence rate of (2.1): a simple analysis based on fixed-point iteration

We study evolution of ξ_t for the strongly convex case ($\mu > 0$) assuming that γ_t is fixed to a constant $\gamma \in (0, 2/L)$; this assumption is not stringent as most works in the literature choose $\gamma_t \equiv 1/L$.

Our approach offers an alternative to its counterpart in Nesterov's book (2018, Lemma 2.2.4). In contrast to Nesterov's analysis based on clever algebraic manipulations, our approach *directly* analyzes the evolution of the sequence by studying a simple *fixed point iteration*. More importantly, our fixed-point based approach generalizes better to the Riemannian setting. As byproduct of our approach, we can also remove a technical condition on ξ_0 required by Nesterov's analysis. See Remark 2.4.

Now let us examine the recursive relation satisfied by ξ_t. Recall from Theorem 2.1 the following *nonlinear* recursive relation on ξ_t's:

$$\xi_{t+1}(\xi_{t+1} - 2\mu\Delta_\gamma)/(1-\xi_{t+1}) = \xi_t^2\,. \tag{2.13}$$

Our objective is to characterize the evolution of ξ_t. Intuitively, (2.13) can be construed as a recursive relation for computing the root of $\phi(v) = \psi(v)$, where $\phi(v) := \frac{v(v-2\mu\Delta_\gamma)}{(1-v)}$ and $\psi(v) := v^2$. Since the root is equal to $v = \sqrt{2\mu\Delta_\gamma}$, one can guess that $\xi_t \to \sqrt{2\mu\Delta_\gamma}$. See Figure 1 for illustration. The following lemma confirms this guess.

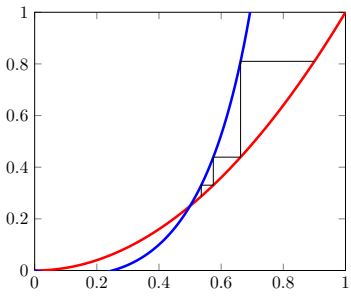

Figure 1: An illustration of the evolution of (2.13) for $2\mu\Delta_\gamma = 0.25$. We plot $\phi = \frac{v(v-2\mu\Delta_\gamma)}{(1-v)}$ in blue and $\psi(v) = v^2$ in red.

Lemma 2.1 (Evolution of (2.13)) *For an arbitrary initial value $\xi_0 \geq 0$, let ξ_t ($t \geq 1$) be the sequence of numbers defined as per (2.13). Then, $\xi_t \in [2\mu\Delta_\gamma, 1)$ for all $t \geq 1$. Furthermore, if*

$$\begin{cases} \xi_0 > \sqrt{2\mu\Delta_\gamma}\,, \\ \xi_0 = \sqrt{2\mu\Delta_\gamma}\,, \\ \xi_0 < \sqrt{2\mu\Delta_\gamma}\,, \end{cases} \quad \textit{then} \quad \begin{cases} \xi_t \searrow \sqrt{2\mu\Delta_\gamma} \text{ as } t \to \infty\,. \\ \xi_t \equiv \sqrt{2\mu\Delta_\gamma}\,. \\ \xi_t \nearrow \sqrt{2\mu\Delta_\gamma} \text{ as } t \to \infty\,. \end{cases}$$

In particular, the convergences are geometric.

Proof The proof and the formal statement (Lemma D.1) are provided in Appendix D. ■

Lemma 2.1 delivers the desired accelerated convergence rate:

Corollary 2.1 *If $\xi_0 \geq \sqrt{2\mu\Delta_\gamma}$, then $f(y_t) - f(x_*) = O(\prod_{i=1}^t (1-\sqrt{2\mu\Delta_\gamma})) = O(\exp(-t\sqrt{2\mu\Delta_\gamma}))$. In particular, setting $\gamma = 1/L$, $f(y_t) - f(x_*) = O(\exp(-t\sqrt{\mu/L}))$.*

Remark 2.4 (Removing technical conditions in Nesterov's analysis) *Nesterov's original analysis requires a technical condition on the initial value* ξ_0*:* $\sqrt{\mu/L} \leq \xi_0 \leq {}^{(2(3+\mu/L))}/_{(3+\sqrt{21+4\mu/L})}$ *(Nesterov, 2018, (2.2.21)). In contrast, our analysis reveals that the upper bound on* ξ_0 *is not needed; the lower bound is also not needed in the sense that* ξ_t *converges to* $\sqrt{\mu/L}$*, the accelerated rate.*

3. Generalization to the non-Euclidean case: Riemannian potential function analysis

This section develops the first key ingredient towards obtaining our main theorem (Theorem 4.1), namely, Theorem 3.1 that is a Riemannian analog of Theorem 2.1.

3.1. Riemannian geometry and Riemannian analog of Nesterov method

We begin by recalling some basic concepts from Riemannian geometry, and defer to textbooks (e.g., (Jost, 2008; Burago et al., 2001)) for more. A *Riemannian manifold* is a smooth manifold M equipped with a smoothly varying inner product $\langle \cdot, \cdot \rangle_x$ (the *Riemannian metric*) defined for each $x \in M$ on the tangent space $T_x M$. With the concept of length of curves, one can introduce a distance d on M, and consequently, view (M, d) as a metric space. Length also allows us to define analogs of straight lines, namely *geodesics*: A curve is a *geodesic* if it is locally distance minimizing. The notion of curvature that we will need is *sectional curvature*, which characterizes curvature by measuring Gaussian curvatures of 2-dimensional submanifolds of M. We make the following key assumption:

Assumption 1 *We assume that the sectional curvature is lower bounded by* $-\kappa$ *for some nonnegative constant* κ*. This is a widely used standard assumption in Riemannian geometry; see e.g., (Burago et al., 2001, Chapter 10) and (Perelman, 1995).*

Operations on manifolds. We can define analogs of vector addition and subtraction on Riemannian manifolds via *exponential maps*. An exponential map $\mathrm{Exp}_x : T_x M \to M$ maps $v \in T_x M$ to $g(1) \in M$ for a geodesic g with $g(0) = x$ and $g'(0) = v$. Notice that $\mathrm{Exp}_x(v) \in M$ is an analog of vector addition "$x + v$." Similarly, the inverse map $\mathrm{Exp}_x^{-1}(y) \in T_x M$ is an analog of vector subtraction "$y - x$." For Exp_x^{-1} to be well-defined for each x, we assume that any two points on M are connected by a unique geodesic. This property is called *uniquely geodesic*, and is valid locally for general Riemannian manifolds and globally for *non-positively curved* manifolds (more precisely, manifolds with globally non-positive sectional curvatures). We assume further that $\mathrm{Exp}_x, \mathrm{Exp}_x^{-1}$ can be computed at each x, as is the case for many widely used matrix manifolds (Absil et al., 2009).[3]

Convexity. The notion of convexity can be extended to Riemannian manifolds using geodesics where convex combinations of two points are defined along geodesics connecting them. This generalized notion of convexity is called *geodesic convexity* (*g-convexity* for short) (Gromov, 1978). One can also define geodesic-smoothness and (strong) g-convexity akin to their Euclidean counterparts.

Assumption 2 *We assume that the cost function* f *is geodesically* L*-smooth and* μ*-strongly convex (formal definitions in Appendix E.3; see also (Zhang and Sra, 2016, Section 2.3)).*

Using the above noted Riemannian analogs of vector operations, Nesterov's method (2.1) turns into:

$$x_{t+1} \leftarrow \mathrm{Exp}_{y_t}\left(\alpha_{t+1} \mathrm{Exp}_{y_t}^{-1}(z_t)\right) \tag{3.1a}$$

$$y_{t+1} \leftarrow \mathrm{Exp}_{x_{t+1}}\left(-\gamma_{t+1} \nabla f(x_{t+1})\right) \tag{3.1b}$$

$$z_{t+1} \leftarrow \mathrm{Exp}_{x_{t+1}}\left(\beta_{t+1} \mathrm{Exp}_{x_{t+1}}^{-1}(z_t) - \eta_{t+1} \nabla f(x_{t+1})\right). \tag{3.1c}$$

[3] For computational reasons, exponential maps are often approximated by cheaper approximations (e.g., retractions). Analyzing the effect of such approximations is not addressed in this paper and is left as an open question.

See Figure 2 for an illustration of (3.1). Note that updates (3.1b) and (3.1c) are well-defined since $\nabla f(x_{t+1})$ lies in the tangent space $T_x M$. We are now ready to analyze the Riemannian iteration (3.1).

3.2. Riemannian potential function analysis and metric distortion

Since (3.1) is a direct analog of its Euclidean counterpart (2.1), one may be tempted to use the potential function $\Psi_t := A_t \cdot (f(y) - f(x_*)) + B_t \cdot d(z_t, x_*)^2$ that is a direct analog of the potential (2.2). However, it turns out that the following less direct choice is much more advantageous:

$$\Psi_t := A_t \cdot (f(y_t) - f(x_*)) + B_t \cdot \left\| \mathrm{Exp}_{x_t}^{-1}(z_t) - \mathrm{Exp}_{x_t}^{-1}(x_*) \right\|_{x_t}^2 . \tag{3.2}$$

The distance term in (3.2) is preferable to $d(z_t, x_*)^2$ because it lets us use Euclidean geometry (since it is defined on the tangent space $T_{x_t} M \cong \mathbb{R}^n$) to control it. To simplify notation, we define:

Definition 3.1 (Projected distance) *For any three points $u, v, w \in M$, the projected distance between v and w with respect to u is defined as $d_u(v, w) := \left\| \mathrm{Exp}_u^{-1}(v) - \mathrm{Exp}_u^{-1}(w) \right\|_u$.*

There is, however, one fundamental hurdle *inherent* to comparing distances in the Riemannian setting: we need to handle the *incompatibility* of metrics between two different points. A key advantage of the potential function analysis is that one only needs to focus on comparing the distances appearing in *adjacent* terms, namely Ψ_t and Ψ_{t+1}, which simplifies the argument considerably. Motivated by the potential (3.2), we define the following quantity for comparing distances:

Definition 3.2 (Valid distortion rate) *We say δ_t is a valid distortion rate at iteration $t \geq 1$ if the following inequality holds: $d_{x_t}(z_{t-1}, x_*)^2 \leq \delta_t \cdot d_{x_{t-1}}(z_{t-1}, x_*)^2$.*

Assuming the existence of valid distortion rates at *each* iteration, we can analyze iteration (3.1) analogously to the analysis in §2.2 and §2.3 to obtain the main result of this section.

Theorem 3.1 (Riemannian analog of Theorem 2.1) *Given y_t, z_t and $A_t, B_t > 0$ and $\gamma_{t+1} \in (0, 2/L)$, let $\Delta_\gamma := \gamma_{t+1}(1 - L\gamma_{t+1}/2)$ and $\xi_t := \sqrt{4\Delta_\gamma \cdot B_t / A_t}$. Assume that $\delta_{t+1} > 1$ is a valid distortion rate at iteration $t+1$. Let us choose parameters as per:*

1. *Compute $\xi_{t+1} \in [2\mu\Delta_\gamma, 1)$ satisfying $\frac{\xi_{t+1}(\xi_{t+1} - 2\mu\Delta_\gamma)}{1 - \xi_{t+1}} = \frac{1}{\boldsymbol{\delta_{t+1}}} \xi_t^2$.* (3.3)
2. *Compute $A_{t+1}, B_{t+1}, \alpha_{t+1}, \beta_{t+1}, \eta_{t+1}$ as in Theorem 2.1.*

Then, y_{t+1}, z_{t+1} generated via iteration (3.1) satisfy $\Psi_{t+1} \leq \Psi_t$ (see (3.2)), or equivalently,

$$f(y_{t+1}) - f(x_*) + \tfrac{\xi_{t+1}^2}{4\Delta_\gamma} \cdot d_{x_{t+1}}(z_{t+1}, x_*)^2 \leq (1 - \xi_{t+1}) \cdot \left[f(y_t) - f(x_*) + \tfrac{\xi_t^2}{4\Delta_\gamma} \cdot d_{x_t}(z_t, x_*)^2 \right] .$$

Proof We sketch the proof here, deferring precise details to Appendix E.3. The proof resembles the arguments in §2.2 and §2.3, except for the appearance of valid distortion rates in (3.3). Using the Riemannian analogs of Propositions 2.1 and 2.2, the following vectors lying in the *same* tangent space $T_{x_{t+1}} M$ constitute counterparts of (2.5):

$$\tilde{W} := \mathrm{Exp}_{x_{t+1}}^{-1}(z_t) , \quad \tilde{X} := -\mathrm{Exp}_{x_{t+1}}^{-1}(x_*) , \quad \text{and} \quad \tilde{\nabla} := \nabla f(x_{t+1}) , \tag{3.4}$$

With these vectors, akin to (2.6), it is again straightforward to derive the following upper bound on $\Psi_{t+1} - \Psi_t$ in terms of the vectors $\tilde{\nabla}, \tilde{X}, \tilde{W}$ (here, $\|\cdot\|$ denotes $\|\cdot\|_{x_{t+1}}$ and $\langle \cdot, \cdot \rangle$ denotes $\langle \cdot, \cdot \rangle_{x_{t+1}}$):

$$\Psi_{t+1}-\Psi_t \le \tilde{C}_1\cdot\|\tilde{W}\|^2+\tilde{C}_2\cdot\|\tilde{X}\|^2+\tilde{C}_3\|\tilde{\nabla}\|^2+\tilde{C}_4\cdot\langle\tilde{W},\tilde{X}\rangle+\tilde{C}_5\cdot\langle\tilde{W},\tilde{\nabla}\rangle+\tilde{C}_6\cdot\langle\tilde{X},\tilde{\nabla}\rangle, \quad (3.5)$$

$$\text{where } \begin{cases} \tilde{C}_1 := \beta_{t+1}^2 B_{t+1} - \frac{B_t}{\boldsymbol{\delta_{t+1}}} - \frac{\mu}{2}\frac{\alpha_{t+1}^2}{(1-\alpha_{t+1})^2}A_t\,, & \tilde{C}_2 := B_{t+1} - \frac{B_t}{\boldsymbol{\delta_{t+1}}} - \frac{\mu}{2}(A_{t+1}-A_t)\,, \\ \tilde{C}_3 := \eta_{t+1}^2 B_{t+1} - \Delta_\gamma\cdot A_{t+1}\,, & \tilde{C}_4 := 2\cdot\left(\beta_{t+1}B_{t+1} - \frac{B_t}{\boldsymbol{\delta_{t+1}}}\right)\,, \\ \tilde{C}_5 := \frac{\alpha_{t+1}}{1-\alpha_{t+1}}A_t - 2\beta_{t+1}\eta_{t+1}B_{t+1}\,, & \text{and} \quad \tilde{C}_6 := (A_{t+1}-A_t) - 2\eta_{t+1}B_{t+1}\,. \end{cases}$$

See Appendix E.3.1 for details. Notice the similarity between (3.5) and (2.6): the only difference is that the B_t's in (2.6) are replaced with B_t/δ_{t+1}'s here. This difference is attributed to the definition of valid distortion rate (Definition 3.2); also, in the derivation of (3.5), we use $-B_t\cdot d_{x_t}(z_t,x_*)^2 \le -\frac{B_t}{\delta_{t+1}}\cdot d_{x_{t+1}}(z_t,x_*)^2$, which precisely accounts for the appearance of B_t/δ_{t+1} instead of B_t.

Having this counterpart (3.5) of (2.6), we follow §2.3 to make (3.5) a negative sum of squares. It turns out that due to similarity between (3.5) and (2.6), the same derivation holds modulo the appearance of δ_{t+1} in the denominator of (3.3). See Appendix E.3.2 for precise details. ■

As before, we can deduce from Theorem 3.1 the suboptimality gap bound (2.12). Hence, to identify the convergence rate we only need to determine the evolution of $\{\xi_t\}$. We provide an illustrative example below, before moving onto the full accelerated algorithm in §4.

Illustrative example: constant distortion rate. Assume that μ is positive[4], and consider the simplified case where $\delta_t \equiv \delta \ge 1$ for all $t \ge 0$. Under this constant distortion condition, similarly to recursion (2.13), one can obtain a recursive relation on $\{\xi_t\}$ by choosing $\gamma_t \equiv \gamma$:

$$\xi_{t+1}(\xi_{t+1} - 2\mu\Delta_\gamma)/(1-\xi_{t+1}) = \xi_t^2/\boldsymbol{\delta}\,. \quad (3.6)$$

Analogously to Lemma 2.1, we can establish geometric convergence of ξ_t to the fixed point $\xi(\delta)$ of (3.6) (see Lemma D.1). Solving for $\xi(\delta)$ explicitly, we obtain the following analog of Corollary 2.1:

Corollary 3.1 *Assume $\mu > 0$. If $\xi_0 \ge \xi(\delta) := \frac{1}{2}\sqrt{(\delta-1)^2+8\delta\mu\Delta_\gamma} - \frac{1}{2}(\delta-1)$, then the following convergence rate holds: $f(y_t) - f(x_*) = O\big(\prod_{i=1}^t(1-\xi(\delta))\big) = O\big(\exp(-t\cdot\xi(\delta))\big)$. In particular, setting $\gamma = 1/L$, $f(y_t) - f(x_*) = O\Big((\exp\big(-\frac{t}{2}\{\sqrt{(\delta-1)^2+4\delta\mu/L} - \frac{t}{2}(\delta-1)\}\big)\Big)$.*

A notable aspect of Corollary 3.1 is that it characterizes a trade-off between the metric distortion and the convergence rate of the resulting algorithm. This point is elaborated by the following remark:

Remark 3.3 (Properties of $\xi(\delta)$) *When there is no distortion, i.e., $\delta = 1$, then $\xi(1) = \sqrt{2\mu\Delta_\gamma}$ since (3.6) becomes (2.13). Moreover, one can verify that $\xi(\delta)$ is (strictly) decreasing in δ, implying that the algorithm's performance gets worse as the distortion gets severer (see Appendix D.1 for verification). Hence, $\xi(\delta) > \lim_{\delta\to\infty}\xi(\delta) = 2\mu\Delta_\gamma$ for all $\delta > 1$, implying that the convergence rate is always* **strictly** *better than gradient descent no matter how severe the distortion is.*

The above example already recovers the local acceleration result of Zhang and Sra (2018). More specifically, they showed that if $d(x_0,x_*)$ is bounded by $1/20\cdot\kappa^{1/2}(L/\mu)^{-3/4}$, then the distortion is bounded by $\delta = 1 + 1/5\cdot(L/\mu)^{-1/2}$; see Appendix F therein. Simplifying $\xi(\delta)$ for this choice of δ, we obtain the following strengthening of their main result (Zhang and Sra, 2018, Theorem 3):

Corollary 3.2 (Local acceleration) *Let $\delta = 1 + \frac{1}{5}\cdot(\mu/L)^{1/2}$, $\gamma = 1/L$ and $\xi_0 \ge \xi(\delta)$. Then, assuming $d(x_0,x_*) \le \frac{1}{20}\cdot\kappa^{1/2}(\mu/L)^{3/4}$, we have $f(y_t) - f(x_*) = O(\exp(-\frac{9}{10}t\sqrt{\mu/L}))$. In particular, $\xi_t = \xi(\delta)$ for all $t \ge 0$, recovers (Zhang and Sra, 2018, Algorithm 2).*

[4] One can also obtain the results for the case $\mu = 0$ from the case $\mu > 0$ through well-known folklore reductions, e.g., (Gasnikov and Nesterov, 2018, Theorem 4); see Appendix H.

4. Riemannian Accelerated Gradient Method

Thus far, the analysis assumed existence of valid distortion rates. But the key question is: *are valid distortion rates available to the method?* We provide a positive answer below and therewith propose a new Riemannian accelerated gradient method. For clarity, we will focus on Riemannian manifolds with globally non-positive sectional curvatures unless stated otherwise; the development for positively-curved manifolds is analogous and is deferred to Appendix I.

4.1. Valid distortion rates and Riemannian accelerated gradient method

We estimate metric distortion by first invoking a classical comparison theorem of Rauch (1951).

Proposition 4.1 *Let $x, y, z \in M$, a Riemannian manifold with curvature lower bounded by $-\kappa < 0$. Let $S_\kappa(r) := \left(\frac{\sinh(\sqrt{\kappa}r)}{\sqrt{\kappa}r}\right)^2$; then, we have $d(y,z)^2 \leq S_\kappa(\max\{d(x,y), d(x,z)\}) \cdot d_x(y,z)^2$.*

Proof A direct consequence of the Rauch comparison theorem; see Appendix C. ■

Applying Proposition 4.1 to the points x_t, z_t, x_*, it is straightforward to conclude:

$$d_{x_{t+1}}(z_t, x_*) \overset{(\clubsuit)}{\leq} d(z_t, x_*)^2 \overset{(\spadesuit)}{\leq} S_\kappa(\max\{d(x_t, z_t),\ d(x_t, x_*)\}) \cdot d_{x_t}(z_t, z_*)^2,$$

where (♣) is due to Topogonov's comparison theorem (see e.g., (Burago et al., 2001, Section 6.5)); and (♠) is due to Proposition 4.1. Hence, $\delta_t = S_\kappa(\max\{d(x_t, z_t),\ d(x_t, x_*)\})$ is a valid distortion rate. Unfortunately, this distortion rate depends on $d(x_t, x_*)$, which is in general unavailable to the algorithm. We overcome this crucial issue by developing a new distortion inequality.

Lemma 4.1 (Improved metric distortion inequality) *Let x, y, z be points on Riemannian manifold M with sectional curvatures lower bounded by $-\kappa < 0$. Then for $T_\kappa : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 1}$ defined as*

$$T_\kappa(r) := \begin{cases} \max\left\{1 + 4\left(\frac{\sqrt{\kappa}r}{\tanh(\sqrt{\kappa}r)} - 1\right),\ \left(\frac{\sinh(2\sqrt{\kappa}\cdot r)}{2\sqrt{\kappa}\cdot r}\right)^2\right\}, & \text{if } r > 0, \\ 1, & \text{if } r = 0, \end{cases} \tag{4.1}$$

the following inequality holds: $d(y,z)^2 \leq T_\kappa(d(x,y)) \cdot d_x(y,z)^2$.

Proof The proof uses Proposition 4.1 and a Riemannian trigonometric inequality due to (Zhang and Sra, 2016, Lemma 6). See Appendix C for a formal statement and the proof. ■

Note that T_κ behaves similarly to S_κ. Most importantly, $\lim_{r \to 0+} T_\kappa(r) = 1$, implying that the effect of distortion diminishes as the distance decreases. Hence, one can essentially regard Lemma 4.1 as a version of Proposition 4.1 in which the term $\max\{d(x,y), d(x,z)\}$ is replaced with $d(x,y)$. Thanks to Lemma 4.1, now we have $T_\kappa(d(x_t, z_t))$ as a valid distortion rate, which is *accessible* to the algorithm at iteration t. Therefore, we propose the following algorithm:

Algorithm 1 (Riemannian accelerated gradient method) **Input:** *$x_0 = y_0 = z_0 \in M$; constant $\xi_0 > 0$; $\gamma \in (0, 2/L)$; $\Delta_\gamma := \gamma(1 - L\gamma/2)$; integer T.*

for $t = 0, 1, 2, \ldots, T$**:**

Compute the distortion rate $\delta_{t+1} := T_\kappa(d(x_t, z_t))$ as per (4.1).

Find $\xi_{t+1} \in [2\mu\Delta_\gamma, 1)$ such that $\xi_{t+1}(\xi_{t+1} - 2\mu\Delta_\gamma)/(1 - \xi_{t+1}) = \xi_t^2/\delta_{t+1}$.

Compute $\alpha_{t+1} := \frac{\xi_{t+1} - 2\mu\Delta_\gamma}{1 - 2\mu\Delta_\gamma}$, $\beta_{t+1} := 1 - 2\mu\Delta_\gamma\xi_{t+1}^{-1}$, and $\eta_{t+1} := 2\Delta_\gamma\xi_{t+1}^{-1}$.

Update the next step iterates as per (3.1) with $\gamma_{t+1} := \gamma$.

end for

Remark 4.2 (Innovations relative to previous methods) *A noticeable innovation in Algorithm 1 lies in its use of the adaptive metric distortion rate $T_\kappa(d(x_t, z_t))$. This is in stark contrast with previous approaches Zhang and Sra (2016, 2018); Alimisis et al. (2020) that use a global metric distortion rate based on the diameter of the domain. As we shall we in § 4.2, our adaptive metric distortion control is a crucial ingredient for achieving* full *acceleration.*

Remark 4.3 *Note that $T_\kappa(d(x_t, z_t))$ is a worst-case upper bound on the valid distortion rate, and hence, if additional information on local geometry is accessible, one can possibly come up with a better estimate and replace $T_\kappa(d(x_t, z_t))$ in Algorithm 1 with the estimate.*

4.2. Convergence rate analysis of the proposed method

Having proposed the algorithm, our final task is to analyze its convergence rate. From Remark 3.3, we know the algorithm achieves a *full* acceleration when δ_t is close to 1. Due to the property $\lim_{r\to 0+} T_\kappa(r) = 1$, one therefore needs to show that $d(x_t, z_t)$ is close to 0. Although $d(x_t, z_t) = 0$ for $t = 0$, one can quickly notice that it is not true for $t \geq 1$.

Now one natural follow-up question is whether $d(x_t, z_t)$ shrinks over iterations. As we have seen in §2.3, the convergence of the iterates to the optimal point is a direct consequence of our potential function analysis. Similarly, one can immediately see that $d_{x_t}(z_t, x_*) \to 0$. It turns out that from this shrinking projected distance, one can also deduce $d(x_t, z_t) \to 0$ under mild conditions:

Lemma 4.2 (Shrinking $d(x_t, z_t)$) *Assume $\mu > 0$ and let $D_0 := f(x_0) - f(x_*) + \xi_0^2/4\Delta_\gamma \cdot d(x_0, x_*)^2$. If $1 < \gamma L < 2 - \xi_t$ and $\xi_t > 2\mu\Delta_\gamma$ hold at iteration $t \geq 1$, then Algorithm 1 satisfies: $d(x_t, z_t) \leq \mathcal{C}_{\mu,L,\gamma}\left[D_0 \prod_{j=1}^{t-1}(1-\xi_j)\right]^{1/2}$, where $\mathcal{C}_{\mu,L,\gamma} > 0$ is a constant depending only on μ, L, γ.*

Proof The proof relies on elementary geometric inequalities (see Appendix F). ■

Note that the assumption $\gamma L \in (1, 2 - \xi_t]$ can be roughly read as "$\gamma L \in (1, 2 - \sqrt{\mu/L}]$" because Remark 3.3 ensures that $\xi(\delta) \leq \sqrt{2\mu\Delta_\gamma} < \sqrt{\mu/L}$ for all $\delta \geq 1$. More precisely, since ξ_t quickly converges to the fixed point, one can easily ensure $\xi_t \leq \sqrt{\mu/L}$ after few iterations. Formalizing this argument, we finally obtain our main theorem (which formalizes Theorem 1.1):

Theorem 4.1 (Global acceleration of Algorithm 1) *Assume $0 < \mu < L$ and $\gamma L \in (1, 2 - \sqrt{\mu/L}]$. Let $\Delta_\gamma := \gamma(1 - L\gamma/2)$ and $\lambda := 1 - 8\mu\Delta_\gamma/(5+\sqrt{5}) \in (0, 1)$. Then for any $\xi_0 > 0$, Algorithm 1 satisfies the following accelerated convergence:*

$$f(y_t) - f(x_*) = O\left((1-\xi_1)(1-\xi_2)\cdots(1-\xi_t)\right), \tag{4.2}$$

where $\{\xi_t\}$ is a sequence such that (i) $\xi_t > 2\mu\Delta_\gamma$ $\forall t \geq 0$ and (ii) for all $\epsilon > 0$, $|\xi_t - \sqrt{2\mu\Delta_\gamma}| \leq \epsilon$ whenever $t = \Omega\left(\frac{\log(1/\epsilon)}{\log(1/\lambda)}\right)$, where the constant involved in $\Omega(\cdot)$ depends only on μ, L, γ, κ.

Proof (4.2) is immediate from Theorem 3.1. For the convergence of $\{\xi_t\}$, see Appendix G. ■

Since $\Delta_\gamma \to 1/(2L)$ as $\gamma \to 1/L$, one can achieve the convergence rate *arbitrarily* close to the full acceleration rate by choosing γ bigger but sufficiently close to $1/L$. This concludes our main results.

5. Comparison with other potential function analyses

In this section, we compare existing potential function analyses with our approach. We discuss here the most directly relevant works; for additional related work, please see Appendix B and also (Taylor and Bach, 2019, Appendix B).

The potential function (2.2) has appeared in prior works on accelerated methods, corroborating its *suitability*. Compared with our analysis, the main difference is that the existing analyses either work for (i) the case $\mu = 0$, or (ii) just the fixed-step case $\xi_t = \sqrt{\mu/L}$. We highlight that our analysis is the first to recover–from first principles–Nesterov's general scheme that smoothly interpolates the cases $\mu = 0$ and $\mu > 0$. Moreover, our analysis allows ξ_t to vary, which is crucial in the Riemannian case where the recursive relation changes over iterations.

Function (2.2) appears in (Wilson et al., 2016, Proposition 4) within the context of a continuous dynamics approach to acceleration. That work studies methods for discretizing *accelerated* ODEs derived in (Su et al., 2014; Wibisono et al., 2016) to transform the continuous dynamics into discrete methods. In that context, they show that (2.2) is a discretization of a *canonical* Lyapunov function. Another appearance is in (Diakonikolas and Orecchia, 2019), where they extend the continuous dynamics view via an *approximate duality gap technique*. Roughly, to analyze a first-order method, they consider an upper bound U_t and a lower bound L_t on the optimal value $f(x_*)$. Their analysis then proceeds by showing the gap $G_t := U_t - L_t$ diminishes with the rate α_t, i.e., $\alpha_t G_t$ is decreasing, which corresponds to showing $A_t \mathcal{E}_t$ is decreasing in our language (§2). Although motivated mostly for continuous dynamics, their techniques cover discrete methods with some modifications. In particular, their choice of G_t for accelerated method corresponds to (2.2) (see §4.2 therein).

Yet another appearance of (2.2) is (Bansal and Gupta, 2019, (5.50)), wherein the motivation was to modify the potential function analyses for gradient descent to design and analyze accelerated methods. They propose the idea of running *two different* gradient steps and linearly combining them to achieve desired accelerated convergence. Following their argument, it turns out (2.2) is the right choice. Indeed, their approach bears resemblance to the *linear coupling* framework (Allen-Zhu and Orecchia, 2017), in which (2.2) has even more canonical interpretations; see Appendix A.

6. Conclusion

In this paper, we establish the first *global* accelerated gradient method for (strongly convex) Riemannian optimization. To that end, we first revisit the Euclidean case and present an alternative approach to Nesterov's estimate sequences, shedding new light on the scope of his technique that has puzzled researchers for many years. We then consider the Riemannian case and propose a method that converges strictly faster than gradient descent, quickly attaining the full accelerated convergence rate within a few iterations. While results for the non-strongly convex setting are also developed via a well-known reduction argument, discovering a direct approach remains open. We believe our results mark fundamental progress toward understanding acceleration in non-Euclidean settings, and hope that our work motivates a richer study of Riemannian acceleration, while contributing to the goal of bringing our understanding of Riemannian optimization at par with the Euclidean setting.

Acknowledgements

SS and KA acknowledge a support from NSF CAREER grant Number 1846088. KA also acknowledges Kwanjeong Educational Foundation for their support.

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. *arXiv:1806.00065*, 2018.

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

Foivos Alimisis, Antonio Orvieto, Gary Bécigneul, and Aurelien Lucchi. A continuous-time perspective for modeling acceleration in Riemannian optimization. *To appear in AISTATS*, 2020.

Zeyuan Allen-Zhu and Lorenzo Orecchia. Linear coupling: An ultimate unification of gradient and mirror descent. In *ITCS*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2017.

Necdet Serhat Aybat, Alireza Fallah, Mert Gurbuzbalaban, and Asuman Ozdaglar. Robust accelerated gradient methods for smooth strongly convex functions. *To appear in SIAM Journal on Optimization*, 2019.

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

Nikhil Bansal and Anupam Gupta. Potential-function proofs for gradient methods. *Theory of Computing*, 15(4):1–32, 2019. doi: 10.4086/toc.2019.v015a004.

Glaydston C Bento, Orizon P Ferreira, and Jefferson G Melo. Iteration-complexity of gradient, subgradient and proximal point methods on Riemannian manifolds. *Journal of Optimization Theory and Applications*, 173(2):548–562, 2017.

Michael Betancourt, Michael Jordan, and Ashia Wilson. On symplectic optimization. *Preprint arXiv:1802.03653*, 2018.

Martin R Bridson and André Haefliger. *Metric spaces of non-positive curvature*, volume 319. Springer Science & Business Media, 2013.

Dmitri Burago, Yurii Burago, and Sergei Ivanov. *A course in metric geometry*, volume 33. American Mathematical Soc., 2001.

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 *FOCS*, pages 845–861. IEEE, 2019.

Isaac Chavel. *Riemannian geometry: a modern introduction*, volume 98. Cambridge university press, 2006.

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

Saman Cyrus, Bin Hu, Bryan Van Scoy, and Laurent Lessard. A robust accelerated optimization algorithm for strongly convex functions. In *ACC*, pages 1376–1381. IEEE, 2018.

Jelena Diakonikolas and Lorenzo Orecchia. The approximate duality gap technique: A unified theory of first-order methods. *SIAM Journal on Optimization*, 29(1):660–689, 2019.

Yoel Drori and Marc Teboulle. Performance of first-order methods for smooth convex minimization: a novel approach. *Mathematical Programming*, 145(1-2):451–482, 2014.

Ramsay Dyer, Gert Vegter, and Mathijs Wintraecken. Riemannian simplices and triangulations. *Geometriae Dedicata*, 179(1):91–138, 2015.

Alexander Vladimirovich Gasnikov and Yu E Nesterov. Universal method for stochastic composite optimization problems. *Computational Mathematics and Mathematical Physics*, 58(1):48–64, 2018.

Navin Goyal and Abhishek Shetty. Sampling and optimization on convex sets in Riemannian manifolds of non-negative curvature. In *COLT*, pages 1519–1561, 2019.

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

Bin Hu and Laurent Lessard. Dissipativity theory for Nesterov's accelerated method. In *ICML*, pages 1549–1557. JMLR, 2017.

Jürgen Jost. *Riemannian geometry and geometric analysis*, volume 42005. Springer, 2008.

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

Donghwan Kim and Jeffrey A Fessler. Optimized first-order methods for smooth convex minimization. *Mathematical programming*, 159(1-2):81–107, 2016.

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

Yuanyuan Liu, Fanhua Shang, James Cheng, Hong Cheng, and Licheng Jiao. Accelerated first-order methods for geodesically convex optimization on Riemannian manifolds. In *NIPS*, pages 4868–4877, 2017.

Aleksandr M. Lyapunov. The general problem of the stability of motion. *International Journal of Control*, 55(3):531–534, 1992.

Arkadi Nemirovski and David Yudin. *Problem complexity and method efficiency in optimization.* Chichester: Wiley, 1983.

Yurii Nesterov. A method for unconstrained convex minimization problem with the rate of convergence $O\left(1/k^2\right)$. In *Doklady AN USSR*, volume 269, pages 543–547, 1983.

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

G. Perelman. Spaces with curvature bounded below. In *Proceedings of the International Congress of Mathematicians*, pages 517–525, Basel, 1995. Birkhäuser Basel. ISBN 978-3-0348-9078-6.

Harry Ernest Rauch. A contribution to differential geometry in the large. *Annals of Mathematics*, pages 38–55, 1951.

Sam Safavi, Bikash Joshi, Guilherme França, and José Bento. An explicit convergence rate for Nesterov's method from SDP. In *ISIT*, pages 1560–1564. IEEE, 2018.

Bin Shi, Simon S Du, Weijie Su, and Michael I Jordan. Acceleration via symplectic discretization of high-resolution differential equations. In *NeurIPS*, pages 5745–5753, 2019.

Weijie Su, Stephen Boyd, and Emmanuel Candes. A differential equation for modeling Nesterov's accelerated gradient method: theory and insights. In *NIPS*, pages 2510–2518, 2014.

Adrien Taylor and Francis Bach. Stochastic first-order methods: non-asymptotic and computer-aided analyses via potential functions. In *COLT*, 2019.

Adrien Taylor, Bryan Van Scoy, and Laurent Lessard. Lyapunov functions for first-order methods: Tight automated convergence guarantees. In *ICML*, pages 4897–4906, 2018.

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

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

Andre Wibisono, Ashia C Wilson, and Michael I Jordan. A variational perspective on accelerated methods in optimization. *PNAS*, 113(47):E7351–E7358, 2016.

Ashia Wilson, Benjamin Recht, and Michael Jordan. A Lyapunov analysis of momentum methods in optimization. *Preprint arXiv:1611.02635*, 2016.

Hongyi Zhang and Suvrit Sra. First-order methods for geodesically convex optimization. In *COLT*, pages 1617–1638, 2016.

Hongyi Zhang and Suvrit Sra. An estimate sequence for geodesically convex optimization. In *COLT*, pages 1703–1723, 2018.

Hongyi Zhang, Sashank Reddi, and Suvrit Sra. Riemannian SVRG: fast stochastic optimization on Riemannian manifolds. In *NIPS*, pages 4592–4600, 2016.

Jingzhao Zhang, Hongyi Zhang, and Suvrit Sra. R-SPIDER: a fast Riemannian stochastic optimization algorithm with curvature independent rate. *Preprint arXiv:1811.04194*, 2018.

Pan Zhou, Xiaotong Yuan, Shuicheng Yan, and Jiashi Feng. Faster first-order methods for stochastic non-convex optimization on Riemannian manifolds. *IEEE TPAMI*, 2019.

Appendix A. Interpretations via linear coupling

Recently, Allen-Zhu and Orecchia (2017) established a framework of designing fast first-order methods called *linear coupling*. The principal observation therein is that the two most fundamental first-order methods, namely *gradient* and *mirror* descent, have *complementary* performances, and one might therefore design faster first-order methods by *linearly coupling* the two methods. In this section, we will discuss how one can derive from linear coupling (i) Nesterov's optimal method iterations (2.1); and (ii) our choice of potential function (2.2) (which offers an alternative way to motivate the potential function; we omit mentioning this connection in the main text because the "coupling" idea does not admit an easy Riemannian analogue).

A.1. Nesterov's iteration from linear coupling

Let us now see how to obtain the main iteration (2.1) via linear coupling. Denote by $\mathsf{Grad}_{s\cdot\nabla}(x)$ and $\mathsf{Mirr}_{s\cdot\nabla}(x)$ a gradient step and a mirror step, respectively. If we choose the Bregman divergence

associated with mirror descent to be $D(u,v) = \frac{1}{2}\|u-v\|_2^2$, then (2.1) can be rewritten as follows:

$$\begin{aligned}
x_{t+1} &\leftarrow \alpha_{t+1} z_t + (1-\alpha_{t+1}) y_t \\
w_{t+1} &\leftarrow \tilde{\beta}_{t+1} z_t + (1-\tilde{\beta}_{t+1}) y_t \\
\nabla_{t+1} &\leftarrow \nabla f(x_{t+1}) \\
y_{t+1} &\leftarrow \mathsf{Grad}_{\gamma_{t+1}\nabla_{t+1}}(x_{t+1}) \\
z_{t+1} &\leftarrow \mathsf{Mirr}_{\eta_{t+1}\nabla_{t+1}}(w_{t+1}) \,,
\end{aligned}$$

where $\tilde{\beta}_{t+1} = \alpha_{t+1} + (1-\alpha_{t+1})\beta_{t+1}$. Note that these steps clearly respect linear coupling: for each step, we compute two different linear combinations of z_t and y_t and run gradient and mirror steps from each combination to obtain the next iterates y_{t+1} and z_{t+1}, respectively. Indeed, the original algorithm considered in (Allen-Zhu and Orecchia, 2017) chooses $\beta' \equiv 1$ and is hence a special case of the above steps. One concrete advantage of viewing iteration (2.1) in the above form is that then it can be naturally generalized to other settings where the smoothness of f is defined with respect to a norm different from ℓ_2.

A.2. Choosing a potential function via linear coupling

Another advantage of the linear coupling view is that one can derive our choice of potential function (2.2) naturally. To see this, first note that the folklore analysis of gradient descent deals with the cost value $f(y)$, while that of mirror descent deals with the distance to an optimal point, or more generally, the Bregman divergence $D(z, x_*)$. (See e.g. (Allen-Zhu and Orecchia, 2017, §2) for details.) Since the algorithm is a *linear combination* of the two methods, it is then natural to consider a *linear combination* of the two performance measures, arriving at (2.2) since our case corresponds to the setting where the Bregman divergence is chosen to be $D(z, x_*) = \frac{1}{2}\|z - x_*\|_2^2$.

Appendix B. Comparison with SDP-based potential function analysis

Another prominent approaches related to potential function analysis are developed based on solving SDPs (Drori and Teboulle, 2014; Lessard et al., 2016; Taylor et al., 2018; Taylor and Bach, 2019). The primary distinction between our approach and most SDP-based approaches is that our analysis is *analytical*, whereas the analyses therein are *numerical*. More specifically, the existing works require *numeric values* of parameters (e.g., α, β, L, μ) because they find suitable potential functions via solving SDPs. Note that one *cannot* solve SDPs unless the numeric coefficients are given. Abstractly, our choice of parameters in Theorem 2.1 can be interpreted as an *analytical* solution to the *symbolic* versions of SDPs formulated in the prior works.

Notable exceptions are (Kim and Fessler, 2016; Hu and Lessard, 2017; Safavi et al., 2018; Cyrus et al., 2018; Aybat et al., 2019), in which small SDPs are solved *analytically*. Specifically, some optimized step sizes for Nesterov's method are derived via solving small SDPs explicitly in (Kim and Fessler, 2016; Safavi et al., 2018); robust versions of gradient methods are derived analytically via classical control-theoretic arguments in (Cyrus et al., 2018; Aybat et al., 2019), and Nesterov's method is reinterpreted using *dissipativity theory* in (Hu and Lessard, 2017). Indeed borrowing the dissipativity interpretation from (Hu and Lessard, 2017), one can interpret our calculations in §2.3 as finding an *analytic* solution to a dissipation inequality (Theorem 2 therein) for our case.

Appendix C. Some inequalities from Riemannian geometry (proof of Lemma 4.1)

This section is devoted to proving Lemma 4.1. The proof requires two ingredients: Proposition 4.1 and a (Riemannian) trigonometric inequality due to (Zhang and Sra, 2016, Lemma 6).

We begin with the first key ingredient: Proposition 4.1. Its proof is based on the following version of the Rauch comparison theorem (Chavel, 2006, Theorem IX.2.3):

Proposition C.1 (Rauch comparison theorem) *Let M be a Riemannian manifold with sectional curvatures lower bounded by $-\kappa < 0$. Then, for any $x \in M$ and $u \in T_x M$, the following upper bound on the operator norm of the differential of the exponential map holds:*

$$\left\| d(\mathrm{Exp}_x)_u \right\|_{op} \leq \frac{\sinh(\sqrt{\kappa}\,\|u\|)}{\sqrt{\kappa}\,\|u\|}\,.$$

Proof Let $u_0 := u/\|u\|$. First, it follows from the definition that the exponential map is radially isometric, i.e., $\|d(\mathrm{Exp}_x)_u(u_0)\| = 1$. Next, due to Rauch comparison theorem (Chavel, 2006, Theorem IX.2.3), for any v orthogonal to u, we have $\|d(\mathrm{Exp}_x)_u(v)\| \leq \frac{\sinh(\sqrt{\kappa}\|u\|)}{\sqrt{\kappa}\|u\|}\|v\|$. Since any vector in $T_u(T_x M)$ can be represented as a linear combination of u_0 and vectors orthogonal to u_0, the proof follows. ■

Now, we are ready to prove Proposition 4.1:

Proposition C.2 (Restatement of Proposition 4.1) *Let x, y, z be points on Riemannian manifold M with sectional curvatures lower bounded by $-\kappa < 0$. Then, the following inequality holds:*

$$d(y,z) \leq \frac{\sinh(\sqrt{\kappa}\max\{d(x,y), d(x,z)\})}{\sqrt{\kappa}\max\{d(x,y), d(x,z)\}} \cdot d_x(y,z)\,.$$

Proof To upper bound the distance $d(y,z)$ in terms of the projected distance $d_x(y,z)$, consider a path $p : [0,1] \to T_x M$ defined as $p(t) = (1-t)\cdot \mathrm{Exp}_x^{-1}(y) + t \cdot \mathrm{Exp}_x^{-1}(z)$. Then, its image $\mathrm{Exp}_x(p)$ is a path on M connecting y to z. By definition of the distance on the manifold, $d(y,z)$ is clearly upper bounded by the length of $\mathrm{Exp}_x(p)$. On the other hand, using Proposition C.1, the length of $\mathrm{Exp}_x(p)$ can be upper bounded as follows (since $\|p'(t)\| = \left\|\mathrm{Exp}_x^{-1}(y) - \mathrm{Exp}_x^{-1}(z)\right\| = d_x(y,z)$):

$$\begin{aligned}\int_0^1 \left\| \frac{d}{dt}\mathrm{Exp}_x(p(t)) \right\| dt &\leq \int_0^1 \left\| d(\mathrm{Exp}_x)_{p(t)} \right\|_{\mathrm{op}} \cdot \left\| p'(t) \right\| dt \\ &\leq \frac{\sinh(\sqrt{\kappa}\max\{d(x,y), d(x,z)\})}{\sqrt{\kappa}\max\{d(x,y), d(x,z)\}} \cdot d_x(y,z)\,,\end{aligned}$$

where the last inequality follows from the fact that $\|p(t)\|$ is upper bounded by $\max\{\|p(0)\|, \|p(1)\|\} = \max\{d(x,y), d(x,z)\}$. ■

We now move on to the second key ingredient, namely a Riemannian trigonometric inequality:

Proposition C.3 (Riemannian trigonometric inequality) *Let M be a Riemannian manifold with sectional curvatures lower bounded by $-\kappa < 0$. Let x, y, z be the vertices of a geodesic triangle with the lengths of the opposite side being a, b, c, respectively, and A be the angle of the triangle at the vertex x, then we have the following inequality:*

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

Proof See (Zhang and Sra, 2016, §3.1) and (Cordero-Erausquin et al., 2001, Lemma 3.12). ■

With these ingredients we now prove Lemma 4.1; we actually prove the following strengthening:

Lemma C.1 *Let x, y, z be points on Riemannian manifold M with sectional curvatures lower bounded by $-\kappa < 0$. Define the function $\widehat{T_\kappa} : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 1}$ as*

$$\widehat{T_\kappa}(r) := \begin{cases} \min_{\epsilon>0} \max \left\{ 1 + \left(1+\epsilon^{-1}\right)^2 \left(\frac{\sqrt{\kappa} r}{\tanh(\sqrt{\kappa} r)} - 1 \right), \left(\frac{\sinh\left((1+\epsilon)\sqrt{\kappa}\cdot r\right)}{(1+\epsilon)\sqrt{\kappa}\cdot r} \right)^2 \right\} & \text{if } r > 0, \\ 1, & \text{if } r = 0. \end{cases}$$

Then, the following inequality holds: $d(y,z)^2 \leq \widehat{T_\kappa}(d(x,y)) \cdot d_x(y,z)^2$.

Note that $\widehat{T_\kappa}(r) \leq T_\kappa(r)$ for all $r \geq 0$ (T_κ is equal to choosing $\epsilon = 1$ in the definition of $\widehat{T_\kappa}$.) Hence, Lemma C.1 immediately implies Lemma 4.1.

Proof *[Proof of Lemma C.1]* Let us fix an arbitrary constant $\epsilon > 0$. We will separately handle two cases: (i) $(1+\epsilon) \cdot d(x,y) < d(x,z)$; and (ii) $(1+\epsilon) \cdot d(x,y) \geq d(x,z)$.

Case (i). Applying Proposition C.3 to $\triangle xyz$, and letting $\zeta := \frac{\sqrt{\kappa} d(x,y)}{\tanh(\sqrt{\kappa} d(x,y))}$, we obtain:

$$\begin{aligned} d(y,z)^2 &\leq d(x,y)^2 + \zeta \cdot d(x,z)^2 - 2\left\langle \mathrm{Exp}_x^{-1}(y), \mathrm{Exp}_x^{-1}(z) \right\rangle \\ &= (\zeta - 1) \cdot d(x,z)^2 + d(x,y)^2 + d(x,z)^2 - 2\left\langle \mathrm{Exp}_x^{-1}(y), \mathrm{Exp}_x^{-1}(z) \right\rangle \\ &= (\zeta - 1) \cdot d(x,z)^2 + d_x(y,z)^2, \end{aligned}$$

where the last line follows from the Euclidean law of cosines. On the other hand, from the Euclidean triangle inequality (consider the triangle $\triangle xyz$ in the tangent space $T_x M$), $d_x(y,z) \geq (d(x,z) - d(x,y)) > \frac{\epsilon}{1+\epsilon} \cdot d(x,z)$. Hence, combining these two, we get

$$\begin{aligned} d(y,z)^2 &\leq (\zeta - 1) \cdot d(x,z)^2 + d_x(y,z)^2 \\ &\leq \left(1+\epsilon^{-1}\right)^2 \cdot (\zeta - 1) \cdot d_x(y,z)^2 + d_x(y,z)^2 \\ &= \left[1 + \left(1+\epsilon^{-1}\right)^2 \cdot (\zeta - 1)\right] \cdot d_x(y,z)^2. \end{aligned} \tag{C.1}$$

Case (ii). For the case $(1+\epsilon) \cdot d(x,y) \geq d(x,z)$, Proposition C.2 implies:

$$d(y,z)^2 \leq \left(\frac{\sinh\left((1+\epsilon)\sqrt{\kappa} \cdot d(x,y)\right)}{(1+\epsilon)\sqrt{\kappa} \cdot d(x,y)} \right)^2 \cdot d_x(y,z)^2. \tag{C.2}$$

Therefore, combining (C.1) and (C.2), the proof is completed. ■

Appendix D. Analysis of the key recursive relations ((2.13) and (3.6))

To ease notation, we replace $2\mu\Delta_\gamma$ with a constant $a \in (0,1)$ and consider:

$$\frac{\xi_{t+1}(\xi_{t+1} - a)}{1 - \xi_{t+1}} = \frac{1}{\delta} \cdot \xi_t^2. \tag{D.1}$$

In particular, when $\delta = 1$ and $a = 2\mu\Delta_\gamma$, equation (D.1) recovers (2.13). The parameter $\delta > 1$ is present to cover the recursion (3.6) for the Riemannian case. Below, we state and prove the following general statement of Lemma 2.1.

Lemma D.1 *For any constants $\delta \geq 1$ and $a \in (0,1)$, and an initial value $\xi_0 \geq 0$, the followings properties are true about the recursive relation (D.1):*

1. $\xi(\delta) := \frac{1}{2}\sqrt{(\delta-1)^2 + 4\delta a} - \frac{1}{2}(\delta - 1)$ *is the unique fixed point of* (D.1).
2. $\lim_{t\to\infty} \xi_t \downarrow \xi(\delta)$ *if* $\xi_0 > \xi(\delta)$; $\xi_t \equiv \xi(\delta)$ *if* $\xi_0 = \xi(\delta)$; *and* $\lim_{t\to\infty} \xi_t \uparrow \xi(\delta)$ *if* $0 \leq \xi_0 < \xi(\delta)$.
3. $|\xi_t - \xi(\delta)| \leq \left(\frac{1}{\sqrt{\delta}}\left(1 - \frac{4}{5+\sqrt{5}} \cdot \frac{a}{\sqrt{\delta}}\right)\right)^{t-1} |\xi_1 - \xi(\delta)|$ *for all* $t \geq 1$.

Proof Define $\phi(v) := \frac{v(v-a)}{1-v}$ and $\psi(v) := \frac{1}{\delta}v^2$. Then, recursion (D.1) can be rewritten as

$$\phi(\xi_{t+1}) = \psi(\xi_t)\,. \tag{D.2}$$

Now, in order to understand (D.2), let us study the properties of the two functions. First, note that ψ is increasing on $\mathbb{R}_{\geq 0}$ and ϕ is increasing on $[a,1)$ with $\phi(a) = 0$ and $\lim_{v\to 1-}\phi(v) = \infty$. Indeed, ϕ is increasing since $\frac{d}{dv}\phi(v) = \frac{1-a}{(1-v)^2} - 1 \geq \frac{1}{1-a} - 1 > 0$.

Hence, one can consider the inverse function of the restriction $\phi|_{[a,1)}$. We will simply denote the inverse function by ϕ^{-1}. Letting $\tau := \phi^{-1} \circ \psi$, (D.2) can be rewritten as:

$$\xi_{t+1} = \tau(\xi_t)\,. \tag{D.3}$$

Note that $\tau : \mathbb{R}_{\geq 0} \to [a,1)$, and hence, $\xi_t \in [a,1)$ for all $t \geq 1$. Since τ is increasing, there is at most one fixed point, i.e., $v \geq 0$ s.t. $\tau(v) = v$. Solving $\tau(v) = v$, or equivalently, $\phi(v) = \psi(v)$ on $v \in [a,1)$ yields $v = \xi(\delta)$. Hence, $\xi(\delta)$ is the unique fixed point of (D.3).

From this observation and the fact that ϕ and ψ are both increasing on the respective domains, we have $\phi < \psi$ for $x \in [a, \xi(\delta))$, and $\phi > \psi$ for $x \in (\xi(\delta), 1)$. Consequently, $\{\xi_t\}$ is increasing if $\xi_0 \in [0, \xi(\delta))$ and decreasing if $\xi_0 > \xi(\delta)$.

Now we prove the geometric convergence of (D.3) to $\xi(\delta)$. To that end, let us first express τ explicitly. One can easily verify that the closed form expression of ϕ^{-1} is equal to

$$\phi^{-1}(v) = \frac{1}{2}\left(\sqrt{(v-a)^2 + 4v} - (v-a)\right)\,.$$

Therefore, we have

$$\tau(v) = \phi^{-1}(\psi(v)) = \phi^{-1}(v^2/\delta) = \frac{1}{2}\left(\sqrt{(v^2/\delta - a)^2 + 4v^2/\delta} - (v^2/\delta - a)\right)\,.$$

Due to mean value theorem, the key ingredient for showing the geometric convergence is to bound the derivative of τ. Indeed, if we can establish that $|\tau'(v)| \leq K < 1$ for $v \in [a,1)$, then we have

$$|\xi_{t+1} - \xi(\delta)| = |\tau(\xi_t) - \tau(\xi(\delta))| \leq K \cdot |\xi_t - \xi(\delta)|\,. \tag{D.4}$$

Letting $\theta(v) := \frac{v(v^2-a)+2v}{\sqrt{(v^2-a)^2+4v^2}} - v$, one can express the derivative τ' in terms of θ:

$$\begin{aligned}\tau'(v) &= \frac{\frac{v}{\delta}(v^2/\delta - a) + 2\frac{v}{\delta}}{\sqrt{(v^2/\delta - a)^2 + 4v^2/\delta}} - \frac{v}{\delta} = \frac{1}{\sqrt{\delta}} \cdot \left(\frac{\frac{v}{\sqrt{\delta}}(v^2/\delta - a) + 2\frac{v}{\sqrt{\delta}}}{\sqrt{(v^2/\delta - a)^2 + 4v^2/\delta}} - \frac{v}{\sqrt{\delta}} \right) \\ &= \frac{1}{\sqrt{\delta}} \cdot \theta(v/\sqrt{\delta})\end{aligned}$$

Hence, it suffices to show that $\theta(v) < 1$ for $v \in (0,1)$. Proposition D.1 below shows this claim.

Proposition D.1 $0 \le \theta(v) < 1 - \frac{4}{5+\sqrt{5}} \cdot v$ *holds for* $v \in (0,1)$.

Proof $\theta(v) \ge 0$ trivially holds since τ is increasing (recall that τ is a composition of increasing functions). Now let us prove the upper bound. We first consider the case $a < v \le \sqrt{a}$. Since $v^2 \le a$,

$$\theta(v) = \frac{-v(a-v^2)+2v}{\sqrt{(v^2-a)^2+4v^2}} - v \le \frac{2v}{\sqrt{(v^2-a)^2+4v^2}} - v \le 1 - v\,.$$

Next, consider the case $v > \sqrt{a}$. Then, $v^2 > a$, and hence

$$\begin{aligned}\theta(v) &= \frac{v(v^2-a)+2v}{\sqrt{(v^2-a)^2+4v^2}} - v = \frac{2v}{\sqrt{(v^2-a)^2+4v^2}} - v \cdot \frac{\sqrt{(v^2-a)^2+4v^2} - (v^2-a)}{\sqrt{(v^2-a)^2+4v^2}} \\ &= \frac{2v}{\sqrt{(v^2-a)^2+4v^2}} - v \cdot \frac{4v^2}{\sqrt{(v^2-a)^2+4v^2}\left(\sqrt{(v^2-a)^2+4v^2} + (v^2-a)\right)} \\ &= \frac{2v}{\sqrt{(v^2-a)^2+4v^2}} - v \cdot \frac{4v^2}{(v^2-a)^2+4v^2+(v^2-a)\sqrt{(v^2-a)^2+4v^2}} \\ &\overset{(\clubsuit)}{\le} 1 - v \cdot \frac{4v^2}{v^2+4v^2+v\sqrt{v^2+4v^2}} = 1 - \frac{4}{5+\sqrt{5}} \cdot v\,.\end{aligned}$$

where ($\clubsuit$) follows since $v \in (\sqrt{a}, 1)$; in particular, we have $0 \le v^2 - a \le v^2 \le v$. Combining the two cases, we complete the proof. ■

From Proposition D.1 and inequality (D.4), the proof of the geometric convergence follows. ■

D.1. Justification of Remark 3.3

In this section, we verify that for any fixed $a \in (0,1)$,

$$\xi(\delta) := \frac{\sqrt{(\delta-1)^2 + 4\delta a} - (\delta - 1))}{2} \quad \text{is decreasing in } \delta \ge 1.$$

Note that for $\delta \geq 1$ we have

$$\begin{aligned}\frac{d}{d\delta}\xi(\delta) &= \frac{2(\delta-1)+4a}{4\sqrt{(\delta-1)^2+4\delta a}} - \frac{1}{2} = \frac{2(\delta-1)+4a-2\sqrt{(\delta-1)^2+4\delta a}}{4\sqrt{(\delta-1)^2+4\delta a}} \\ &= \frac{2\sqrt{((\delta-1)+2a)^2}-2\sqrt{(\delta-1)^2+4\delta a}}{4\sqrt{(\delta-1)^2+4\delta a}} \\ &= \frac{2\sqrt{(\delta-1)^2+4a(\delta-1)+4a^2}-2\sqrt{(\delta-1)^2+4\delta a}}{4\sqrt{(\delta-1)^2+4\delta a}} < 0\,,\end{aligned}$$

where the last inequality is due to the fact that $-4a+4a^2 < 0$ since $a < 1$.

Appendix E. Potential function analyses (Theorems 2.1 and 3.1)

E.1. Derivation of the upper bound on the potential difference (2.6)

We recall the notations (2.5): $\Delta_\gamma := \gamma(1-L\gamma/2)$, $\nabla := \nabla f(x_{t+1})$, $X := x_{t+1} - x_*$, and $W := z_t - x_{t+1}$. Let us first express (2.4) in terms of the vectors ∇, X, W using Proposition 2.2:

$$\begin{aligned}(2.4) &= B_{t+1}\cdot\|x_{t+1}+\beta(z_t-x_{t+1})-x_*\|^2 + B_{t+1}\eta^2\cdot\|\nabla f(x_{t+1})\|^2 \\ &\quad + 2B_{t+1}\eta\cdot\langle\nabla f(x_{t+1}), x_*-x_{t+1}-\beta(z_t-x_{t+1})\rangle - B_t\cdot\|z_t-x_*\|^2 \\ &= B_{t+1}\cdot\|X+\beta W\|^2 + B_{t+1}\eta^2\cdot\|\nabla\|^2 - 2B_{t+1}\eta\cdot\langle\nabla, X+\beta W\rangle - B_t\cdot\|W+X\|^2 \\ &= (B_{t+1}-B_t)\cdot\|X\|^2 + (\beta^2 B_{t+1}-B_t)\cdot\|W\|^2 + \eta^2 B_{t+1}\cdot\|\nabla\|^2 \\ &\quad + 2(\beta B_{t+1}-B_t)\cdot\langle X, W\rangle - 2\beta\eta B_{t+1}\langle W,\nabla\rangle - 2\eta B_{t+1}\cdot\langle X,\nabla\rangle\,.\end{aligned} \tag{E.1}$$

For (2.3), we apply Proposition 2.1 and rearrange terms to obtain:

$$\begin{aligned}(2.3) &= A_{t+1}\cdot(f(y_{t+1})-f(x_*)) - A_t\cdot(f(y_t)-f(x_*)) \\ &\leq A_{t+1}\cdot(f(x_{t+1})-f(x_*)) - A_{t+1}\Delta_\gamma\cdot\|\nabla f(x_{t+1})\|^2 - A_t\cdot(f(y_t)-f(x_*)) \\ &= A_t\cdot(f(x_{t+1})-f(y_t)) + (A_{t+1}-A_t)\cdot(f(x_{t+1})-f(x_*)) - \Delta_\gamma A_{t+1}\cdot\|\nabla f(x_{t+1})\|^2\,. \\ &\overset{(\clubsuit)}{\leq} A_t\cdot\langle\nabla f(x_{t+1}), x_{t+1}-y_t\rangle + (A_{t+1}-A_t)\cdot\langle\nabla f(x_{t+1}), x_{t+1}-x_*\rangle \\ &\quad - A_t\frac{\mu}{2}\cdot\|x_{t+1}-y_t\|^2 - (A_{t+1}-A_t)\frac{\mu}{2}\cdot\|x_{t+1}-x_*\|^2 - \Delta_\gamma A_{t+1}\cdot\|\nabla f(x_{t+1})\|^2\,,\end{aligned} \tag{E.2}$$

where ($\clubsuit$) follows from μ-strong convexity of f (in particular, $f(u)-f(v) \leq \langle\nabla f(u), u-v\rangle - \frac{\mu}{2}\|u-v\|^2$). Now using the identity $x_{t+1}-y_t = \frac{\alpha}{1-\alpha}(z_t-x_{t+1}) = \frac{\alpha}{1-\alpha}W$, one can express (E.2) in terms of ∇, X, W:

$$\begin{aligned}(E.2) &= \frac{\alpha}{1-\alpha}A_t\cdot\langle\nabla, W\rangle + (A_{t+1}-A_t)\cdot\langle\nabla, X\rangle \\ &\quad - \frac{\mu}{2}\left(\frac{\alpha}{1-\alpha}\right)^2 A_t\cdot\|W\|^2 - \frac{\mu}{2}(A_{t+1}-A_t)\cdot\|X\|^2 - \Delta_\gamma A_{t+1}\cdot\|\nabla\|^2\end{aligned} \tag{E.3}$$

Combining (E.1) and (E.3), we obtain the desired upper bound (2.6).

E.2. Proof of Theorem 2.1

We seek to express parameters $A_{t+1}, B_{t+1}, \alpha_{t+1}, \beta_{t+1}, \eta_{t+1}$ in terms of ξ_{t+1} and γ_{t+1}. From the equality version of (2.8), i.e.,

$$B_{t+1} = \frac{(A_{t+1} - A_t)^2}{4\Delta_\gamma \cdot A_{t+1}},$$

B_{t+1} can be easily expressed in terms of ξ_{t+1}: using the relation $1 - \xi_{t+1} := A_t/A_{t+1}$, we have $(\frac{A_{t+1}-A_t}{A_{t+1}})^2 = \xi_{t+1}^2$, and hence

$$B_{t+1} = \left(\frac{A_{t+1} - A_t}{A_{t+1}}\right)^2 \cdot \frac{A_{t+1}}{4\Delta_\gamma} = \xi_{t+1}^2 \cdot \frac{A_t/(1-\xi_{t+1})}{4\Delta_\gamma} = \frac{\xi_{t+1}^2}{1-\xi_{t+1}} \cdot \frac{A_t}{4\Delta_\gamma}. \tag{E.4}$$

From this identity we can also conclude that

$$\frac{A_{t+1}}{B_{t+1}} = \frac{\frac{A_t}{(1-\xi_{t+1})}}{\frac{\xi_{t+1}^2}{1-\xi_{t+1}} \cdot \frac{A_t}{4\Delta_\gamma}} = \frac{4\Delta_\gamma}{\xi_{t+1}^2}. \tag{E.5}$$

Let us recall the expressions (2.7) for the step sizes:

$$\eta_{t+1} = \frac{A_{t+1} - A_t}{2B_{t+1}}, \tag{E.6}$$

$$\beta_{t+1} = \frac{B_t}{B_{t+1}} \quad \text{and} \tag{E.7}$$

$$\frac{\alpha_{t+1}}{1-\alpha_{t+1}} = \frac{(A_{t+1} - A_t)B_t}{A_t B_{t+1}}. \tag{E.8}$$

Let us also recall the recursive relation:

$$\frac{\xi_{t+1}(\xi_{t+1} - 2\mu\Delta_\gamma)}{1-\xi_{t+1}} = \xi_t^2 = 4\Delta_\gamma \cdot \frac{B_t}{A_t} \tag{E.9}$$

Using the relations above, we can now express $\alpha_{t+1}, \beta_{t+1}, \eta_{t+1}$ in terms of ξ_{t+1}:

$$\eta_{t+1} \overset{(E.6)}{=} \frac{A_{t+1} - A_t}{2B_{t+1}} = \frac{A_{t+1} - A_t}{A_{t+1}} \cdot \frac{A_{t+1}}{2B_{t+1}} \overset{(E.5)}{=} \xi_{t+1} \cdot \frac{2\Delta_\gamma}{\xi_{t+1}^2} = 2\Delta_\gamma \xi_{t+1}^{-1},$$

$$\beta_{t+1} \overset{(E.7)}{=} \frac{B_t}{B_{t+1}} \overset{(E.4)}{=} \frac{1-\xi_{t+1}}{\xi_{t+1}^2} \cdot 4\Delta_\gamma \cdot \frac{B_t}{A_t} \overset{(E.9)}{=} \frac{\xi_{t+1} - 2\mu\Delta_\gamma}{\xi_{t+1}} = 1 - 2\mu\Delta_\gamma \xi_{t+1}^{-1}, \quad \text{and}$$

$$\frac{\alpha_{t+1}}{1-\alpha_{t+1}} \overset{(E.8)}{=} \frac{(A_{t+1} - A_t)B_t}{A_t B_{t+1}} = \frac{A_{t+1} - A_t}{A_{t+1}} \cdot \frac{B_t}{A_t} \cdot \frac{A_{t+1}}{B_{t+1}}$$
$$\overset{(E.9)\&(E.5)}{=} \xi_{t+1} \cdot \frac{\xi_{t+1}(\xi_{t+1} - 2\mu\Delta_\gamma)}{4\Delta_\gamma(1-\xi_{t+1})} \cdot \frac{4\Delta_\gamma}{\xi_{t+1}^2} = \frac{\xi_{t+1} - 2\mu\Delta_\gamma}{1-\xi_{t+1}}.$$

With the above choices of parameters, one can easily check that $\alpha_{t+1}, \beta_{t+1}$ both lie in $[0, 1]$ since $\xi_{t+1} \in [2\mu\Delta_\gamma, 1)$.

One last thing we need to check is $\beta^2 B_{t+1} \leq B_t$. Indeed, since $\beta_{t+1} \in [0, 1]$ $\beta_{t+1}^2 B_{t+1} \leq \beta_{t+1} B_{t+1} = B_t$ (due to (E.7)), implying $C_1 \leq 0$.

E.3. Proof of Theorem 3.1

We first introduce the definitions of geodesic (strong) convexity and smoothness. For simplicity, we assume that the function $f : M \to \mathbb{R}$ is differentiable throughout the definitions, and we denote by $\nabla f(x) \in T_x M$ the gradient of f at x.

Definition E.1 (Geodesic (strong) convexity) *f is said to be geodesically μ-strongly convex if*

$$f(y) \geq f(x) + \left\langle \nabla f(x), \operatorname{Exp}_x^{-1}(y) \right\rangle_x + \frac{\mu}{2} \cdot d(x,y)^2 \quad \text{for any } x, y \in M,$$

where $\langle \cdot, \cdot \rangle_x$ denotes the inner product in the tangent space of x induced by the Riemannian metric.

Definition E.2 (Geodesic smoothness) *$f : M \to \mathbb{R}$ is said to be geodesically L-smooth if*

$$f(y) \leq f(x) + \left\langle \nabla f(x), \operatorname{Exp}_x^{-1}(y) \right\rangle_x + \frac{L}{2} \cdot d(x,y)^2 \quad \text{for any } x, y \in M.$$

An equivalent definition is

$$\left\| \nabla f(x) - \Gamma_y^x \nabla f(y) \right\|_x \leq L \cdot d(x,y) \quad \text{for any } x, y \in M,$$

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

With these definitions, we can establish Riemannian analogues of Propositions 2.1 and 2.2:

Proposition E.3 *Let $y = \operatorname{Exp}_x(-s \cdot \nabla f(x))$. If f is geodesically L-smooth, then $f(y) - f(x) \leq -s(1 - Ls/2) \|\nabla f(x)\|_x^2$.*

Proof By the geodesic L-smoothness of f, we have $f(y) \leq f(x) + \left\langle \nabla f(x), \operatorname{Exp}_x^{-1}(y) \right\rangle_x + \frac{L}{2} \cdot d(x,y)^2 = f(x) + \langle \nabla f(x), -s \nabla f(x) \rangle_x + \frac{L}{2} \|-s \nabla f(x)\|_x^2 = f(x) - s\left(1 - \frac{Ls}{2}\right) \|\nabla f(x)\|_x^2$. ∎

Proposition E.4 *Let $z = \operatorname{Exp}_u(v - s \cdot \nabla f(u))$ for some vector $v \in T_x M$. Then, for any x_*, $d_u(z, x_*)^2 - d_u(\operatorname{Exp}_u(v), x_*)^2 = s^2 \|\nabla f(u)\|_u^2 + 2s \left\langle \nabla f(u), \operatorname{Exp}_u^{-1}(x_*) - v \right\rangle_u$.*

Proof The proof follows immediately from the definition of the projected distances (Definition 3.1):

$$\begin{aligned} d_u(z, x_*)^2 &= \left\| \operatorname{Exp}_u^{-1}(z) - \operatorname{Exp}_u^{-1}(x_*) \right\|_u^2 = \left\| v - s \cdot \nabla f(u) - \operatorname{Exp}_u^{-1}(x_*) \right\|_u^2 \\ &= \left\| v - \operatorname{Exp}_u^{-1}(x_*) \right\|_u^2 + \|-s \cdot \nabla f(u)\|_u^2 + 2 \left\langle -s \cdot \nabla f(u), v - \operatorname{Exp}_u^{-1}(x_*) \right\rangle_u \end{aligned}$$

which recovers the conclusion of Proposition E.4. ∎

Now we prove Theorem 3.1. It turns out one can establish an upper bound on the potential difference $\Psi_{t+1} - \Psi_t$ analogously to (2.6). The key difference in the Riemannian case is that instead of W, X, ∇, we now have the following three vectors in the same tangent space $T_{x_{t+1}} M$:

$$\tilde{W} := \operatorname{Exp}_{x_{t+1}}^{-1}(z_t), \quad \tilde{X} := -\operatorname{Exp}_{x_{t+1}}^{-1}(x_*), \text{ and } \tilde{\nabla} := \nabla f(x_{t+1}). \tag{E.10}$$

As pointed out in §2.2, the fact that these three vectors lie in the same tangent space is crucial for the analysis to follow. Using Propositions E.3 and E.4, one can derive the following upper bound similarly to Appendix E.1 (hereafter, $\|\cdot\|$ denotes $\|\cdot\|_{x_{t+1}}$ and $\langle \cdot, \cdot \rangle$ denotes $\langle \cdot, \cdot \rangle_{x_{t+1}}$):

$$\tilde{C}_1 \cdot \|\tilde{W}\|^2 + \tilde{C}_2 \cdot \|\tilde{X}\|^2 + \tilde{C}_3\|\tilde{\nabla}\|^2 + \tilde{C}_4 \cdot \langle \tilde{W}, \tilde{X}\rangle + \tilde{C}_5 \cdot \langle \tilde{W}, \tilde{\nabla}\rangle + \tilde{C}_6 \cdot \langle \tilde{X}, \tilde{\nabla}\rangle, \quad \text{(E.11)}$$

$$\text{where } \begin{cases} \tilde{C}_1 := \beta_{t+1}^2 B_{t+1} - \frac{B_t}{\boldsymbol{\delta_{t+1}}} - \frac{\mu}{2}\frac{\alpha_{t+1}^2}{(1-\alpha_{t+1})^2} A_t\,, & \tilde{C}_2 := B_{t+1} - \frac{B_t}{\boldsymbol{\delta_{t+1}}} - \frac{\mu}{2}(A_{t+1} - A_t)\,, \\ \tilde{C}_3 := \eta_{t+1}^2 B_{t+1} - \Delta_\gamma \cdot A_{t+1}\,, & \tilde{C}_4 := 2 \cdot \left(\beta_{t+1} B_{t+1} - \frac{B_t}{\boldsymbol{\delta_{t+1}}}\right)\,, \\ \tilde{C}_5 := \frac{\alpha_{t+1}}{1-\alpha_{t+1}} A_t - 2\beta_{t+1}\eta_{t+1} B_{t+1}\,, \quad \text{and} & \tilde{C}_6 := (A_{t+1} - A_t) - 2\eta_{t+1} B_{t+1}\,. \end{cases}$$

E.3.1. DERIVATION OF THE UPPER BOUND ON RIEMANNIAN POTENTIAL DIFFERENCE (E.11)

Recall the definition of the Riemannian potential function (3.2):

$$\Psi_t := A_t \cdot (f(y_t) - f(x_*)) + B_t \cdot d_{x_t}(z_t, x_*)^2\,. \quad \text{(E.12)}$$

From the definition, one can write the potential difference $\Psi_{t+1} - \Psi_t$ as follows:

$$A_{t+1} \cdot (f(y_{t+1}) - f(x_*)) - A_t \cdot (f(y_t) - f(x_*)) \quad \text{(E.13)}$$
$$+\ B_{t+1} \cdot d_{x_{t+1}}(z_{t+1}, x_*)^2 - B_t \cdot d_{x_t}(z_t, x_*)^2\,. \quad \text{(E.14)}$$

First, we use the valid distortion rate (Definition 3.2) to upper bound (E.14) in order to express it in terms of projected distances relative to the same reference point x_{t+1}:

$$\text{(E.14)} \le B_{t+1} \cdot d_{x_{t+1}}(z_{t+1}, x_*)^2 - \tfrac{B_t}{\delta_{t+1}} \cdot d_{x_{t+1}}(z_t, x_*)^2\,. \quad \text{(E.15)}$$

Now, similarly to Section E.1, one can use Proposition E.4 to express the right hand side of (E.15) in terms of the vectors $\tilde{\nabla}, \tilde{X}, \tilde{W}$:

$$\begin{aligned}
&B_{t+1} \cdot d_{x_{t+1}}(z_{t+1}, x_*)^2 - \tfrac{B_t}{\delta_{t+1}} \cdot d_{x_{t+1}}(z_t, x_*)^2 \\
&= B_{t+1} \cdot \left\| \beta_{t+1}\mathrm{Exp}_{x_{t+1}}^{-1}(z_t) - \mathrm{Exp}_{x_{t+1}}^{-1}(x_*) \right\|^2 + B_{t+1}\eta_{t+1}^2 \|\nabla f(x_{t+1})\| \\
&\quad + 2B_{t+1}\eta_{t+1} \left\langle \nabla f(x_{t+1}), \mathrm{Exp}_{x_{t+1}}^{-1}(x_*) - \beta_{t+1}\mathrm{Exp}_{x_{t+1}}^{-1}(z_t) \right\rangle - \tfrac{B_t}{\delta_{t+1}} \cdot d_{x_{t+1}}(z_t, x_*)^2 \\
&= B_{t+1} \cdot \|\tilde{X} + \beta_{t+1}\tilde{W}\|^2 + B_{t+1}\eta_{t+1}^2 \cdot \|\tilde{\nabla}\|^2 - 2B_{t+1}\eta_{t+1} \cdot \langle \tilde{\nabla}, \tilde{X} + \beta_{t+1}\tilde{W}\rangle - \tfrac{B_t}{\delta_{t+1}} \cdot \|\tilde{W} + \tilde{X}\|^2 \\
&= (B_{t+1} - \tfrac{B_t}{\delta_{t+1}}) \cdot \|\tilde{X}\|^2 + (\beta_{t+1}^2 B_{t+1} - \tfrac{B_t}{\delta_{t+1}}) \cdot \|\tilde{W}\|^2 + \eta_{t+1}^2 B_{t+1} \cdot \|\tilde{\nabla}\|^2 \\
&\quad + 2(\beta_{t+1}B_{t+1} - \tfrac{B_t}{\delta_{t+1}}) \cdot \langle \tilde{X}, \tilde{W}\rangle - 2\beta_{t+1}\eta_{t+1}B_{t+1}\langle \tilde{W}, \tilde{\nabla}\rangle - 2\eta_{t+1}B_{t+1} \cdot \langle \tilde{X}, \tilde{\nabla}\rangle\,.
\end{aligned} \quad \text{(E.16)}$$

For (E.13), the derivation is identical to that of (E.3), except that now we use Proposition E.3 in place of Proposition 2.1 and μ-geodesic strong convexity in place of μ-strong convexity. In particular,

$$\begin{aligned}
\text{(E.13)} \le\ & \frac{\alpha_{t+1}}{1-\alpha_{t+1}} A_t \cdot \langle \tilde{\nabla}, \tilde{W}\rangle + (A_{t+1} - A_t) \cdot \langle \tilde{\nabla}, \tilde{X}\rangle \\
& - \frac{\mu}{2}\left(\frac{\alpha_{t+1}}{1-\alpha_{t+1}}\right)^2 A_t \cdot \|\tilde{W}\|^2 - \frac{\mu}{2}(A_{t+1} - A_t) \cdot \|\tilde{X}\|^2 - \Delta_\gamma A_{t+1} \cdot \|\tilde{\nabla}\|^2
\end{aligned} \quad \text{(E.17)}$$

Combining (E.16) and (E.17), we obtain the desired upper bound (E.11).

E.3.2. ENSURING RIEMANNIAN POTENTIAL DECREASE

We now follow §2.3 to make (E.11) a negative sum of squares. We recall the expression of the coefficients for reader's convenience:

$$\begin{cases} \tilde{C}_1 := \beta_{t+1}^2 B_{t+1} - \frac{B_t}{\boldsymbol{\delta_{t+1}}} - \frac{\mu}{2}\frac{\alpha_{t+1}^2}{(1-\alpha_{t+1})^2} A_t \,, & \tilde{C}_2 := B_{t+1} - \frac{B_t}{\boldsymbol{\delta_{t+1}}} - \frac{\mu}{2}(A_{t+1} - A_t) \,, \\ \tilde{C}_3 := \eta_{t+1}^2 B_{t+1} - \Delta_\gamma \cdot A_{t+1} \,, & \tilde{C}_4 := 2 \cdot \left(\beta_{t+1} B_{t+1} - \frac{B_t}{\boldsymbol{\delta_{t+1}}}\right) \,, \\ \tilde{C}_5 := \frac{\alpha_{t+1}}{1-\alpha_{t+1}} A_t - 2\beta_{t+1}\eta_{t+1} B_{t+1} \,, \quad \text{and} & \tilde{C}_6 := (A_{t+1} - A_t) - 2\eta_{t+1} B_{t+1} \,. \end{cases}$$

First, from $\tilde{C}_4 = \tilde{C}_5 = \tilde{C}_6 = 0$, we get:

$$\eta_{t+1} = \frac{A_{t+1} - A_t}{2B_{t+1}} \,, \tag{E.18}$$

$$\beta_{t+1} = \frac{B_t}{\boldsymbol{\delta_{t+1}} B_{t+1}} \,, \text{ and} \tag{E.19}$$

$$\frac{\alpha_{t+1}}{1-\alpha_{t+1}} = \frac{2\beta_{t+1}\eta_{t+1}B_{t+1}}{A_t} = \frac{(A_{t+1} - A_t)B_t}{\boldsymbol{\delta_{t+1}} A_t B_{t+1}} \,. \tag{E.20}$$

Next, from $\tilde{C}_3 \leq 0$, we have $\frac{\eta_{t+1}^2 B_{t+1}}{\Delta_\gamma} \leq A_{t+1}$. Substituting (E.18) to this inequality and rearranging, we obtain the following inequality:

$$\frac{(A_{t+1} - A_t)^2}{4\Delta_\gamma \cdot A_{t+1}} \leq B_{t+1} \tag{E.21}$$

From $\tilde{C}_2 \leq 0$, we have $B_{t+1} - \frac{\mu}{2}(A_{t+1} - A_t) \leq \frac{B_t}{\boldsymbol{\delta_{t+1}}}$. Together with (E.21), we obtain:

$$\frac{(A_{t+1} - A_t)^2}{4\Delta_\gamma \cdot A_{t+1}} - (A_{t+1} - A_t)\frac{\mu}{2} \leq \frac{B_t}{\boldsymbol{\delta_{t+1}}} \,. \tag{E.22}$$

Again, using the suboptimality shrinking ratio $1 - \xi_{t+1} := A_t / A_{t+1}$, (E.22) becomes

$$\frac{\xi_{t+1}(\xi_{t+1} - 2\mu\Delta_\gamma)}{1 - \xi_{t+1}} \leq \frac{4\Delta_\gamma}{\boldsymbol{\delta_{t+1}}} \cdot \frac{B_t}{A_t} \,. \tag{E.23}$$

Then, due to the left hand side of (E.23) being increasing (as a function of ξ_{t+1}) on $[2\mu\Delta_\gamma, 1)$, the largest ξ_{t+1} (or equivalently, the largest A_{t+1}) satisfies (E.23) (or equivalently, (E.22)) with equality:

$$\frac{\xi_{t+1}(\xi_{t+1} - 2\mu\Delta_\gamma)}{1 - \xi_{t+1}} = \frac{4\Delta_\gamma}{\boldsymbol{\delta_{t+1}}} \cdot \frac{B_t}{A_t} \,. \tag{E.24}$$

Consequently, such a choice of ξ_{t+1} (or corresponding A_{t+1}) also satisfies (E.21) with equality.

Now, one can follow the calculations in Appendix E.2 to express parameters in terms of ξ_{t+1}. From the equality version of (E.21), i.e. $\frac{(A_{t+1} - A_t)^2}{4\Delta_\gamma \cdot A_{t+1}} = B_{t+1}$, one can derive the following:

$$B_{t+1} = \left(\frac{A_{t+1} - A_t}{A_{t+1}}\right)^2 \cdot \frac{A_{t+1}}{4\Delta_\gamma} = \xi_{t+1}^2 \cdot \frac{A_t/(1-\xi_{t+1})}{4\Delta_\gamma} = \frac{\xi_{t+1}^2}{1-\xi_{t+1}} \cdot \frac{A_t}{4\Delta_\gamma} \,. \tag{E.25}$$

From (E.25), one can also derive:

$$\frac{A_{t+1}}{B_{t+1}} = \frac{\frac{A_t}{(1-\xi_{t+1})}}{\frac{\xi_{t+1}^2}{1-\xi_{t+1}} \cdot \frac{A_t}{4\Delta_\gamma}} = \frac{4\Delta_\gamma}{\xi_{t+1}^2}\,. \tag{E.26}$$

Combining all the relations above, we can express $\alpha_{t+1}, \beta_{t+1}, \eta_{t+1}$ in terms of ξ_{t+1}. It turns out that the final expressions do not depend on δ_{t+1} and are identical to those in Appendix E.2:

$$\eta_{t+1} \overset{\text{(E.18)}}{=} \frac{A_{t+1}-A_t}{2B_{t+1}} = \frac{A_{t+1}-A_t}{A_{t+1}} \cdot \frac{A_{t+1}}{2B_{t+1}} \overset{\text{(E.26)}}{=} \xi_{t+1} \cdot \frac{2\Delta_\gamma}{\xi_{t+1}^2} = 2\Delta_\gamma \xi_{t+1}^{-1}\,,$$

$$\beta_{t+1} \overset{\text{(E.19)}}{=} \frac{B_t}{\boldsymbol{\delta_{t+1}} B_{t+1}} \overset{\text{(E.25)}}{=} \frac{1-\xi_{t+1}}{\xi_{t+1}^2} \cdot \frac{4\Delta_\gamma}{\boldsymbol{\delta_{t+1}}} \cdot \frac{B_t}{A_t} \overset{\text{(E.24)}}{=} \frac{\xi_{t+1}-2\mu\Delta_\gamma}{\xi_{t+1}} = 1 - 2\mu\Delta_\gamma \xi_{t+1}^{-1}\,, \quad \text{and}$$

$$\begin{aligned}\frac{\alpha_{t+1}}{1-\alpha_{t+1}} &\overset{\text{(E.20)}}{=} \frac{(A_{t+1}-A_t)B_t}{\delta_{t+1}A_tB_{t+1}} = \frac{A_{t+1}-A_t}{A_{t+1}} \cdot \frac{B_t}{\delta_{t+1}A_t} \cdot \frac{A_{t+1}}{B_{t+1}} \\ &\overset{\text{(E.24)\&(E.26)}}{=} \xi_{t+1} \cdot \frac{\xi_{t+1}(\xi_{t+1}-2\mu\Delta_\gamma)}{4\Delta_\gamma(1-\xi_{t+1})} \cdot \frac{4\Delta_\gamma}{\xi_{t+1}^2} = \frac{\xi_{t+1}-2\mu\Delta_\gamma}{1-\xi_{t+1}}\,.\end{aligned}$$

With the above choices of parameters, one can again check that $\alpha_{t+1}, \beta_{t+1}$ both lie in $[0,1]$ since $\xi_{t+1} \in [2\mu\Delta_\gamma, 1)$.

One last thing to check is $\tilde{C}_1 \le 0$. Indeed, since $\beta_{t+1} \in [0,1]$, we have $\beta_{t+1}^2 B_{t+1} \le \beta_{t+1} B_{t+1} = B_t/\boldsymbol{\delta_{t+1}}$ (due to (E.19)), implying $\tilde{C}_1 \le 0$. Therefore, the above choices of parameters satisfy $\tilde{C}_1, \tilde{C}_2, \tilde{C}_3 \le 0$ and $\tilde{C}_4, \tilde{C}_5, \tilde{C}_6 = 0$, and consequently, $\Psi_{t+1} \le \Psi_t$ since $\Psi_{t+1} - \Psi_t \le$ (E.11). This completes the proof of Theorem 3.1.

Appendix F. Proofs of distance shrinking lemma (Lemma 4.2)

We first analyze the convergence distances (which is a direct consequence of Theorem 3.1) below.

Proposition F.1 *Let M be a Riemannian manifold with sectional curvatures lower bounded by $-\kappa < 0$ and upper bounded by 0. Assume that $\mu > 0$ and let $D_0 := f(x_0) - f(x_*) + \frac{1}{4\Delta_\gamma}\xi_0^2 \cdot d(x_0, x_*)^2$. Then, for x_t, y_t, z_t ($t \ge 1$) generated by Algorithm 1 the following bounds hold:*

1. $d_{x_t}(z_t, x_*) \le \sqrt{D_0 \prod_{j=1}^t (1-\xi_j)} \cdot \sqrt{\frac{1}{\mu^2\Delta_\gamma}}$.
2. $d(y_t, x_*) \le \sqrt{D_0 \prod_{j=1}^t (1-\xi_j)} \cdot \sqrt{\frac{2}{\mu}}$.
3. $d_{x_t}(y_t, z_t) \le \sqrt{D_0 \prod_{j=1}^t (1-\xi_j)} \cdot \left(\sqrt{\frac{2}{\mu}} + \sqrt{\frac{1}{\mu^2\Delta_\gamma}}\right)$.

Proof By recursively applying Theorem 3.1, we have the following for any $t \ge 1$:

$$\begin{aligned} f(y_t) - f(x_*) + \frac{1}{4\Delta_\gamma}\xi_t^2 \cdot d_{x_t}(z_t, x_*)^2 &\le \prod_{j=1}^t (1-\xi_j) \cdot \left[f(y_0) - f(x_*) + \frac{1}{4\Delta_\gamma}\xi_0^2 \cdot d_{x_0}(z_0, x_*)^2\right] \\ &= \prod_{j=1}^t (1-\xi_j) \cdot D_0\,, \end{aligned}$$

where the equality follows since $x_0 = y_0 = z_0$ (which implies $d_{x_0}(z_0, x_*) = d(x_0, x_*)$).

Hence, the bound on $d_{x_t}(z_t, x_*)$ follows immediately due to $\xi_t \in [2\mu\Delta_\gamma, 1)$, while the bound on $d(y_t, x_*)$ follows from the μ-strong g-convexity of f (Definition E.1), which implies $\frac{\mu}{2} \cdot d(y_t, x_*)^2 \leq f(y_t) - f(x_*)$. Lastly, the bound on $d_{x_t}(y_t, z_t)$ follows upon noting that

$$d_{x_t}(y_t, z_t) \leq d_{x_t}(y_t, x_*) + d_{x_t}(z_t, x_*) \leq d(y_t, x_*) + d_{x_t}(z_t, x_*), \tag{F.1}$$

which is a consequence of the (Euclidean) triangle inequality together with the fact that the projected distances are shorter than the actual distances (a property of non-postively curved manifolds; see e.g. (Burago et al., 2001, §6.5)). ■

Proposition F.1 above establishes that the projected distance $d_{x_t}(y_t, z_t)$ is shrinking over iterations. From this, we can also show that $d(y_t, z_t)$ is shrinking under mild conditions:

Proposition F.2 *Let $D_0 := f(x_0) - f(x_*) + \frac{1}{4\Delta_\gamma}\xi_0^2 \cdot d(x_0, x_*)^2$. If $\gamma L > 1$, $\gamma L \leq 2 - \xi_{t+1}$ and $\xi_{t+1} > 2\mu\Delta_\gamma$ hold for $t \geq 0$, then Algorithm 1 satisfies:*

$$d(y_t, z_t) \leq \frac{1 - 2\mu\Delta_\gamma}{1 - 2\mu\Delta_\gamma\xi_{t+1}^{-1}} \cdot \sqrt{D_0 \prod_{j=1}^{t}(1 - \xi_j)} \cdot \frac{\left(\sqrt{\frac{2}{\mu}} + \sqrt{\frac{1}{\mu^2\Delta_\gamma}} + \frac{L}{\mu}\sqrt{\frac{2}{\mu}}\right)}{(\gamma L - 1)(\gamma L - 1 + 2\mu\Delta_\gamma)}.$$

Remark F.3 *A careful reader might realize that the appearance of the term $1 - 2\mu\Delta_\gamma\xi_{t+1}^{-1}$ in the denominator of the bound could be potentially problematic since this term could be arbitrarily small in general when ξ_{t+1} is very close to $2\mu\Delta_\gamma$. However, as we shall see shortly, this term gets canceled out with the algorithm parameter $\beta_{t+1} = 1 - 2\mu\Delta_\gamma\xi_{t+1}^{-1}$ (see Algorithm 1) when we use Proposition F.2 to bound the distance of our interest $d(x_t, z_t)$.*

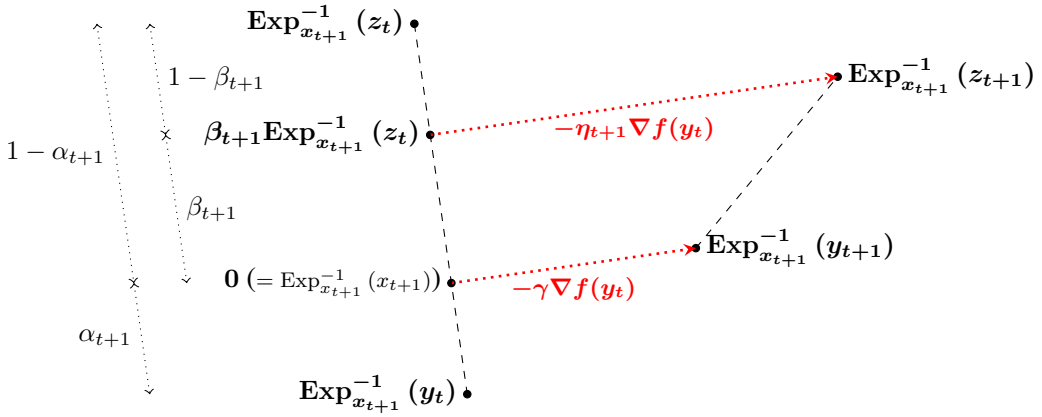

Figure 2: An illustration of the update rule (3.1) on the tangent space $T_{x_{t+1}}M$.

Proof We first recall the following assumption from the proposition statement:

$$1 < \gamma L < 2 - \xi_{t+1} \text{ and } \xi_{t+1} > 2\mu\Delta_\gamma. \tag{F.2}$$

First, from (3.1b) and (3.1c) together with the triangle inequality (see Figure 2),

$$\begin{aligned} d_{x_{t+1}}(y_{t+1}, z_{t+1}) &= \left\| \mathrm{Exp}_{x_{t+1}}^{-1}(y_{t+1}) - \mathrm{Exp}_{x_{t+1}}^{-1}(z_{t+1}) \right\|_{x_{t+1}} \\ &= \left\| -\gamma \nabla f(x_{t+1}) - \beta_{t+1} \mathrm{Exp}_{x_{t+1}}^{-1}(z_t) + \eta_{t+1} \nabla f(x_{t+1}) \right\|_{x_{t+1}} \\ &\geq \beta_{t+1} \left\| \mathrm{Exp}_{x_{t+1}}^{-1}(z_t) \right\|_{x_{t+1}} - |\eta_{t+1} - \gamma| \cdot \|\nabla f(x_{t+1})\|_{x_{t+1}} \\ &= \beta_{t+1} \cdot d(x_{t+1}, z_t) - |\eta_{t+1} - \gamma| \cdot \|\nabla f(x_{t+1})\|_{x_{t+1}} . \end{aligned}$$

Rearranging the above inequality we have

$$\beta_{t+1} \cdot d(x_{t+1}, z_t) \leq d_{x_{t+1}}(y_{t+1}, z_{t+1}) + |\eta_{t+1} - \gamma| \cdot \|\nabla f(x_{t+1})\|_{x_{t+1}} . \tag{F.3}$$

We first simplify the left hand side with the update rules (3.1). First, note that (3.1a) implies that x_{t+1} lies on the geodesic connecting y_t and z_t. Therefore, when representing the iterates on the tangent space $T_{x_{t+1}}M$, the points $\mathrm{Exp}_{x_{t+1}}^{-1}(z_t)$, $\mathrm{Exp}_{x_{t+1}}^{-1}(y_t)$ and 0 $(= \mathrm{Exp}_{x_{t+1}}^{-1}(x_{t+1}))$ on the same line as depicted in Figure 2. Therefore, it is easy to see from Figure 2 that

$$d(x_{t+1}, z_t) = d_{x_{t+1}}(x_{t+1}, y_t) = (1 - \alpha_{t+1}) d_{x_{t+1}}(y_t, z_t) = (1 - \alpha_{t+1}) \cdot d(y_t, z_t) .$$

Substituting this identity to the left hand side of (F.3), (F.3) becomes:

$$\begin{aligned} &\beta_{t+1}(1 - \alpha_{t+1}) \cdot d(y_t, z_t) \\ \leq & d_{x_{t+1}}(y_{t+1}, z_{t+1}) + |\eta_{t+1} - \gamma| \cdot \|\nabla f(x_{t+1})\|_{x_{t+1}} \\ \overset{(\clubsuit)}{\leq} & d_{x_{t+1}}(y_{t+1}, z_{t+1}) + L|\eta_{t+1} - \gamma| \cdot d(x_{t+1}, x_*) \\ \overset{(\spadesuit)}{=} & d_{x_{t+1}}(y_{t+1}, z_{t+1}) + L(\eta_{t+1} - \gamma) \cdot d(x_{t+1}, x_*) \\ \overset{(\heartsuit)}{\leq} & d_{x_{t+1}}(y_{t+1}, z_{t+1}) + L(\eta_{t+1} - \gamma) \cdot d(x_{t+1}, y_t) + L(\eta_{t+1} - \gamma) \cdot d(y_t, x_*) , \\ = & d_{x_{t+1}}(y_{t+1}, z_{t+1}) + L\alpha_{t+1}(\eta_{t+1} - \gamma) \cdot d(y_t, z_t) + L(\eta_{t+1} - \gamma) \cdot d(y_t, x_*) , \end{aligned}$$

where ($\clubsuit$) follows from the geodesic L-smoothness of f: $\|\nabla f(x_{t+1})\|_{x_{t+1}} \leq L \cdot d(x_{t+1}, x_*)$; and ($\spadesuit$) is due to the fact that $\eta_{t+1} - \gamma = 2\Delta_\gamma \xi_{t+1}^{-1} - \gamma = \gamma \xi_{t+1}^{-1}(2 - L\gamma - \xi_{t+1}) > 0$ since $2 - \xi_{t+1} - \gamma L > 0$ from (F.2); ($\heartsuit$) follows from the Riemannian triangle inequality $d(x_{t+1}, x_*) \leq d(x_{t+1}, y_t) + d(y_t, x_*)$; and the last line follows from the identity $d(x_{t+1}, y_t) = \alpha_{t+1} \cdot d(y_t, z_t)$ (see Figure 2).

Moving the term $L\alpha_{t+1}(\eta_{t+1} - \gamma) \cdot d(y_t, z_t)$ to the LHS, we then obtain:

$$[\beta_{t+1}(1 - \alpha_{t+1}) - L\alpha_{t+1}(\eta_{t+1} - \gamma)] \cdot d(y_t, z_t) \leq d_{x_{t+1}}(y_{t+1}, z_{t+1}) + L(\eta_{t+1} - \gamma) \cdot d(y_t, x_*) . \tag{F.4}$$

Since we have seen from Proposition F.1 that the both terms on the right hand side of (F.4) are shrinking, one can prove that $d(y_t, z_t)$ is shrinking as long as one can guarantee that $\beta_{t+1}(1 - \alpha_{t+1}) - L\alpha_{t+1}(\eta_{t+1} - \gamma) > 0$. More formally, Proposition F.2 is a direct consequence the following two statements:

1. The RHS of (F.4) is upper bounded by $\sqrt{D_0\prod_{j=1}^{t}(1-\xi_j)}\cdot\left(\sqrt{\frac{2}{\mu}}+\sqrt{\frac{1}{\mu^2\Delta_\gamma}}+\frac{L}{\mu}\sqrt{\frac{2}{\mu}}\right)$.

2. $\beta_{t+1}(1-\alpha_{t+1})-L\alpha_{t+1}(\eta_{t+1}-\gamma)\geq\frac{1-2\mu\Delta_\gamma\xi_{t+1}^{-1}}{1-2\mu\Delta_\gamma}\cdot(\gamma L-1)(\gamma L-1+2\mu\Delta_\gamma)$. Indeed, with this lower bound one can guarantee that $\beta_{t+1}(1-\alpha_{t+1})-L\alpha_{t+1}(\eta_{t+1}-\gamma)$ is positive due to (F.2): $\gamma L>1$ and $1-2\mu\Delta_\gamma\xi_{t+1}^{-1}>1-2\mu\Delta_\gamma\cdot(2\mu\Delta_\gamma)^{-1}=0$.

Now let us prove the above two statements. From the third conclusion of Proposition F.1, we have $d_{x_{t+1}}(y_{t+1},z_{t+1})\leq\sqrt{D_0\prod_{j=1}^{t+1}(1-\xi_j)}\cdot\left(\sqrt{\frac{2}{\mu}}+\sqrt{\frac{1}{\mu^2\Delta_\gamma}}\right)$. Moreover, from the second conclusion of Proposition F.1, we have:

$$\begin{aligned}L(\eta_{t+1}-\gamma)\cdot d\left(y_t,x_*\right)&\leq L\eta_{t+1}\cdot d\left(y_t,x_*\right)\leq L\eta_{t+1}\cdot\sqrt{D_0\textstyle\prod_{j=1}^{t}(1-\xi_j)}\cdot\sqrt{\frac{2}{\mu}}\\&\leq\sqrt{D_0\textstyle\prod_{j=1}^{t}(1-\xi_j)}\cdot\frac{L}{\mu}\sqrt{\frac{2}{\mu}}\,,\end{aligned}$$

where the last inequality uses $L\eta_{t+1}=2L\Delta_\gamma\xi_{t+1}^{-1}<2L\Delta_\gamma(2\mu\Delta_\gamma)^{-1}\leq\frac{L}{\mu}$. Hence, the first statement follows.

Now, let us prove the second statement. We first recall the parameters in Algorithm 1 for reader's convenience: For $\Delta_\gamma:=\gamma(1-L\gamma/2)$, $\alpha_{t+1}=\frac{\xi_{t+1}-2\mu\Delta_\gamma}{1-2\mu\Delta_\gamma}$, $\beta_{t+1}=1-2\mu\Delta_\gamma\xi_{t+1}^{-1}$, and $\eta_{t+1}=2\Delta_\gamma\xi_{t+1}^{-1}$. Now substituting these parameters to the coefficient, we have:

$$\begin{aligned}&\beta_{t+1}(1-\alpha_{t+1})-L\alpha_{t+1}(\eta_{t+1}-\gamma)\\=&(1-2\mu\Delta_\gamma\xi_{t+1}^{-1})\frac{1-\xi_{t+1}}{1-2\mu\Delta_\gamma}-L\frac{\xi_{t+1}-2\mu\Delta_\gamma}{1-2\mu\Delta_\gamma}\left(2\Delta_\gamma\xi_{t+1}^{-1}-\gamma\right)\\=&\frac{1-2\mu\Delta_\gamma\xi_{t+1}^{-1}}{1-2\mu\Delta_\gamma}\cdot[1-\xi_{t+1}-2L\Delta_\gamma+\gamma L\xi_{t+1}]\end{aligned}$$

Further simplifying the last expression, one obtains the second statement:

$$\begin{aligned}\beta_{t+1}(1-\alpha_{t+1})-L\alpha_{t+1}(\eta_{t+1}-\gamma)=&\frac{1-2\mu\Delta_\gamma\xi_{t+1}^{-1}}{1-2\mu\Delta_\gamma}\cdot\left[(\gamma L-1)^2+(\gamma L-1)\xi_{t+1}\right]\\>&\frac{1-2\mu\Delta_\gamma\xi_{t+1}^{-1}}{1-2\mu\Delta_\gamma}\cdot\left[(\gamma L-1)^2+(\gamma L-1)\cdot 2\mu\Delta_\gamma\right]\,.\end{aligned}$$

where the last line follows from the facts $\xi_{t+1}>2\mu\Delta_\gamma$ and $\gamma L-1>0$. ■

Now, we are finally ready to provide the formal statement and the proof of Lemma 4.2:

Lemma F.1 (Formal statement of Lemma 4.2) *Assume that $\mu>0$. Let $D_0:=f(x_0)-f(x_*)+\frac{1}{4\Delta_\gamma}\xi_0^2\cdot d\left(x_0,x_*\right)^2$. If $\gamma L>1$, $\gamma L\leq 2-\xi_{t+1}$ and $\xi_{t+1}>2\mu\Delta_\gamma$, then Algorithm 1 satisfies:*

$$d\left(x_{t+1},z_{t+1}\right)\leq\mathcal{C}_{\mu,L,\gamma}\cdot\sqrt{D_0\prod_{j=1}^{t}(1-\xi_j)}\,,$$

where $\mathcal{C}_{\mu,L,\gamma}=\frac{\left(\sqrt{\frac{2}{\mu}}+\sqrt{\frac{1}{\mu^2\Delta_\gamma}}+\frac{L}{\mu}\sqrt{\frac{2}{\mu}}\right)(2L\Delta_\gamma+1-2\mu\Delta_\gamma)}{(\gamma L-1)(\gamma L-1+2\mu\Delta_\gamma)}+\frac{L}{\mu}\sqrt{\frac{2}{\mu}}$.

Proof We again recall the parameters in Algorithm 1 for reader's convenience: For $\Delta_\gamma := \gamma(1 - L\gamma/2)$, $\alpha_{t+1} = \frac{\xi_{t+1}-2\mu\Delta_\gamma}{1-2\mu\Delta_\gamma}$, $\beta_{t+1} = 1 - 2\mu\Delta_\gamma\xi_{t+1}^{-1}$, and $\eta_{t+1} = 2\Delta_\gamma\xi_{t+1}^{-1}$. Now, one can use the Euclidean triangle inequality on $T_{x_{t+1}}M$ (see Figure 2) to obtain:

$$\begin{aligned}
d\left(x_{t+1}, z_{t+1}\right) &= d_{x_{t+1}}(x_{t+1}, z_{t+1}) \\
&\leq \beta_{t+1} \cdot d\left(x_{t+1}, z_t\right) + \eta_{t+1} \cdot \|\nabla f(x_{t+1})\|_{x_{t+1}} \\
&\overset{(\clubsuit)}{\leq} \beta_{t+1} \cdot d\left(x_{t+1}, z_t\right) + L\eta_{t+1} \cdot d\left(x_{t+1}, x_*\right) \\
&\overset{(\spadesuit)}{\leq} \beta_{t+1} \cdot d\left(x_{t+1}, z_t\right) + L\eta_{t+1} \cdot d\left(x_{t+1}, y_t\right) + L\eta_{t+1} \cdot d\left(y_t, x_*\right) \\
&\overset{(\heartsuit)}{=} \left(\beta_{t+1}(1-\alpha_{t+1}) + L\eta_{t+1}\alpha_{t+1}\right) \cdot d\left(y_t, z_t\right) + L\eta_{t+1} \cdot d\left(y_t, x_*\right) \\
&\overset{(\diamondsuit)}{=} (1 - \xi_{t+1} + 2L\Delta_\gamma)\frac{1 - 2\mu\Delta_\gamma\xi_{t+1}^{-1}}{1 - 2\mu\Delta_\gamma} \cdot d\left(y_t, z_t\right) + 2L\Delta_\gamma\xi_{t+1}^{-1} \cdot d\left(y_t, x_*\right) ,
\end{aligned}$$

where ($\clubsuit$) is due to the geodesic L-smoothness of f, which implies $\|\nabla f(x_{t+1})\| \leq L \cdot d\left(x_{t+1}, x_*\right)$; ($\spadesuit$) is due to Riemannian triangle inequality; ($\heartsuit$) is due to (3.1a) (see Figure 2); and ($\diamondsuit$) follows from the choice of parameters in Algorithm 1.

Now after we apply Propositions F.1 and F.2 to the last upper bound, and use the fact $\xi_{t+1} \in [2\mu\Delta_\gamma, 1)$ to upper bound ξ_{t+1}'s in the resulting upper bound, Lemma F.1 readily follows. ■

Appendix G. Proof of global acceleration (Theorem 4.1)

We first recall the assumptions in the theorem statement for reader's convenience:

$$0 < \mu < L \text{ and } \gamma L \in (1, 2 - \sqrt{\mu/L}].$$

We first demonstrate that regardless of what initial value $\xi_0 > 0$ we choose, ξ_t becomes less than $\sqrt{\mu/L}$ after a few iterations. Before the demonstration, we denote by $\xi_{t+1} = \tau_{t+1}(\xi_t)$ the recursion $\{\xi_t\}$ in Algorithm 1 follows. In other words, given $\xi_t > 0$, $\xi_{t+1} = \tau_{t+1}(\xi_t)$ is defined as the unique $\xi_{t+1} > 0$ satisfying:

$$\frac{\xi_{t+1}(\xi_{t+1} - 2\mu\Delta_\gamma)}{(1 - \xi_{t+1}} = \frac{\xi_t^2}{\delta_{t+1}} .$$

Proposition G.1 *If $\xi_0 > \sqrt{\mu/L}$, then $\xi_t \leq \sqrt{\mu/L}$ for all t whenever*

$$t \geq \frac{\log\big((\xi_0 - \sqrt{2\mu\Delta_\gamma})/(\sqrt{\mu/L} - \sqrt{2\mu\Delta_\gamma})\big)}{\log\left(1/\big(1 - \frac{8\mu\Delta_\gamma}{5+\sqrt{5}}\big)\right)} . \tag{G.1}$$

If $\xi_0 < \sqrt{\mu/L}$, then $\xi_t \leq \sqrt{\mu/L}$ for all $t \geq 0$.

Proof At some iteration t, we consider two cases depending on whether $\xi_t \leq \sqrt{2\mu\Delta_\gamma}$ or not:

1. First, if $\xi_t \leq \sqrt{2\mu\Delta_\gamma}$, then we evidently have $\xi_{t'} \leq \sqrt{2\mu\Delta_\gamma}$ for all $t' \geq t$. This is due to the fact that the fixed point $\xi(\delta_t)$ is always less than $\sqrt{2\mu\Delta_\gamma}$ together with Lemma D.1.

2. Next, consider the case $\xi_t > \sqrt{2\mu\Delta_\gamma}$. We may assume that $\xi_{t+1} > \sqrt{2\mu\Delta_\gamma}$ (otherwise, $\xi_{t'} \le \sqrt{2\mu\Delta_\gamma}$ for $t' \ge t+1$ due to the first case). Then, the mean value theorem implies:

$$\begin{aligned}\xi_{t+1} - \sqrt{2\mu\Delta_\gamma} &= \tau_{t+1}(\xi_t) - \tau_{t+1}(\tau_{t+1}^{-1}(\sqrt{2\mu\Delta_\gamma})) \\ &\overset{(\clubsuit)}{\le} \frac{1}{\sqrt{\delta_{t+1}}}\left(1 - \frac{4}{5+\sqrt{5}} \cdot \frac{2\mu\Delta_\gamma}{\sqrt{\delta_{t+1}}}\right) \cdot \left(\xi_t - \tau_{t+1}^{-1}(\sqrt{2\mu\Delta_\gamma})\right) \\ &\overset{(\spadesuit)}{<} \left(1 - \tfrac{4}{5+\sqrt{5}} \cdot 2\mu\Delta_\gamma\right) \cdot \left(\xi_t - \sqrt{2\mu\Delta_\gamma}\right),\end{aligned}$$

where (♣) is due to Proposition D.1 together with $\xi_{t+1} > \sqrt{2\mu\Delta_\gamma} \Rightarrow \xi_t > \tau_{t+1}^{-1}(\sqrt{2\mu\Delta_\gamma})$; (♠) follows since $\frac{1}{\sqrt{\delta}}(1 - \frac{4}{(5+\sqrt{5})} \cdot \frac{2\mu\Delta_\gamma}{\sqrt{\delta}})$ for $\delta \ge 1$ is maximized when $\delta = 1$ and $\sqrt{2\mu\Delta_\gamma} < \tau_{t+1}^{-1}(\sqrt{2\mu\Delta_\gamma})$ due to $\sqrt{2\mu\Delta_\gamma} \ge \xi(\delta_{t+1})$ and Lemma D.1. Hence, the distance between ξ_t and $\sqrt{2\mu\Delta_\gamma}$ shrinks geometrically.

Combining the two cases, we conclude the proof. ■

We now study the rate of convergence of $\{\xi_t\}$. To that end, we first study the convergence of $\{\xi(\delta_t)\}$. For simplicity, we assume that $\xi_0 \le \sqrt{\mu/L}$. By Proposition G.1, the arguments below remain true for $\xi_0 > \sqrt{\mu/L}$ after we replace t with $t + (G.1)$. We first characterize $\xi(\delta)$ near $\delta = 1$:

Proposition G.2 *Let* $\xi(\delta) := \frac{1}{2}\left(\sqrt{(\delta-1)^2 + 8\delta\mu\Delta_\gamma} - (\delta - 1)\right)$ *for* $\delta \ge 1$. *Then,* $0 \le \sqrt{2\mu\Delta_\gamma} - \xi(\delta) \le \frac{1}{2}(\delta - 1)$ *for* $1 \le \delta \le 1 + 3/(1 + (4\mu\Delta_\gamma)^{-1})$.

Proof For simplicity, let us write $\delta = 1 + d$. Then, $\xi(1+d) = \frac{1}{2}\left(\sqrt{d^2 + 8\mu\Delta_\gamma(1+d)} - d\right)$. Using the inequality $\sqrt{1+r} \ge 1 + \frac{1}{3}r$ for $0 \le r \le 3$, we get the following as long as $d + \frac{1}{8\mu\Delta_\gamma}d^2 \le 3$:

$$\begin{aligned}\xi(1+d) &\ge \sqrt{2\mu\Delta_\gamma} \cdot \left(1 + \frac{1}{3}d + \frac{1}{24\mu\Delta_\gamma}d^2\right) - \frac{1}{2}d \\ &\ge \sqrt{2\mu\Delta_\gamma} - \left(\frac{1}{2} - \frac{\sqrt{2\mu\Delta_\gamma}}{3}\right)d\,.\end{aligned}$$

Now all we need to check is that $d \le 3/(1 + \frac{1}{4\mu\Delta_\gamma})$ implies $d + \frac{1}{8\mu\Delta_\gamma}d^2 \le 3$. Indeed, if $d \le 3/(1 + \frac{1}{4\mu\Delta_\gamma})$, then we have $d \le 3/(1 + \frac{1}{4\mu\Delta_\gamma}) \le 3/(3/2) = 2$, and hence $d + \frac{1}{8\mu\Delta_\gamma}d^2 = d\big(1 + \frac{d}{8\mu\Delta_\gamma}\big) \le d\big(1 + \frac{1}{4\mu\Delta_\gamma}\big) \le 3$. ■

Next, we characterize the behaviour of the function $T_\kappa(r)$ near $r = 1$.

Proposition G.3 $T_\kappa(r) \le 1 + 2\kappa r^2$ *for* $0 \le r \le \frac{1}{2\sqrt{\kappa}}$.

Proof Using Taylor expansion, one easily easily verify for $0 \le r \le \frac{1}{2\sqrt{\kappa}}$ that

$$\frac{\sqrt{\kappa}r}{\tanh(\sqrt{\kappa}r)} \le 1 + \frac{\kappa}{2}r^2 \quad \text{and} \quad \left(\frac{\sinh(2\sqrt{\kappa}r)}{2\sqrt{\kappa}r}\right)^2 \le 1 + 2\kappa r^2\,.$$

Hence, from the definition of T_κ (see (4.1)), we obtain the desired bound on T_κ. ■

Combining Propositions G.2 and G.3, we obtain the following results:

Proposition G.4 $\sqrt{2\mu\Delta_\gamma} - \xi(T_\kappa(r)) \le \kappa r^2$ *for* $0 \le r \le \sqrt{\frac{3}{1+(4\mu\Delta_\gamma)^{-1}}} \cdot \frac{1}{2\sqrt{\kappa}}$.

Proof Note that $\frac{3}{1+(4\mu\Delta_\gamma)^{-1}} \le \frac{3}{1+2L/\mu} \le 1$, and hence, $\sqrt{\frac{3}{1+(4\mu\Delta_\gamma)^{-1}}} \cdot \frac{1}{2\sqrt{\kappa}} \le \frac{1}{2\sqrt{\kappa}}$. Thus, one can apply Proposition G.3 for $0 \le r \le \sqrt{\frac{3}{1+(4\mu\Delta_\gamma)^{-1}}} \cdot \frac{1}{2\sqrt{\kappa}}$, and obtain $T_\kappa(r) \le 1 + 2\kappa r^2$. Hence, $T_\kappa(r) \le 1 + \frac{1}{2} \cdot \frac{3}{1+(4\mu\Delta_\gamma)^{-1}}$ within the range. Hence, by Proposition G.2, one then obtains $\sqrt{2\mu\Delta_\gamma} - \xi(T_\kappa(r)) \le \kappa r^2$ for $0 \le r \le \sqrt{\frac{3}{1+(4\mu\Delta_\gamma)^{-1}}} \cdot \frac{1}{2\sqrt{\kappa}}$. ■

Let $\mathcal{D}_{\kappa,\mu,\gamma} := \sqrt{\frac{3}{1+(4\mu\Delta_\gamma)^{-1}}} \cdot \frac{1}{2\sqrt{\kappa}}$. Then by Lemma F.1, we can deduce that $d(x_{t+1}, z_{t+1}) \le \mathcal{D}_{\kappa,\mu,\gamma}$ whenever $t \ge 2\frac{\log(\mathcal{C}_{\mu,L,\gamma}\cdot\sqrt{D_0}/\mathcal{D}_{\kappa,\mu,\gamma})}{\log(1/(1-2\mu\Delta_\gamma))}$. Therefore, Proposition G.4 implies that for $t \ge 2\frac{\log(\mathcal{C}_{\mu,L,\gamma}\cdot\sqrt{D_0}/\mathcal{D}_{\kappa,\mu,\gamma})}{\log(1/(1-2\mu\Delta_\gamma)))}$, the following bound holds:

$$\sqrt{2\mu\Delta_\gamma} - \xi\big(T_\kappa\big(d(x_{t+1}, z_{t+1})\big)\big) \le \kappa\mathcal{C}_{\mu,L,\gamma}^2 D_0 (1-2\mu\Delta_\gamma)^t .$$

From this bound, it follows that $\xi\big(T_\kappa\big(d(x_{t+1}, z_{t+1})\big)\big) \in [\sqrt{2\mu\Delta_\gamma} - \epsilon/2, \sqrt{2\mu\Delta_\gamma}]$ whenever

$$t \ge \max\left\{2\frac{\log(\mathcal{C}_{\mu,L,\gamma}\cdot\sqrt{D_0}/\mathcal{D}_{\kappa,\mu,\gamma})}{\log(1/(1-2\mu\Delta_\gamma)))}, \frac{\log(2\kappa\mathcal{C}_{\mu,L,\gamma}^2 D_0/\epsilon)}{\log(1/(1-2\mu\Delta_\gamma)))}\right\}.$$

Now having established the convergence rate of $\{\xi(\delta_t)\}$, we translate it into the convergence rate of $\{\xi_t\}$. Similarly to the proof of Proposition G.1, one can prove that for any $T \ge 0$,

$$|\xi_{T+t} - \xi(\delta_T)| \le \left(1 - \frac{8\mu\Delta_\gamma}{5+\sqrt{5}}\right)^t |\xi_T - \xi(\delta_T)| .$$

From this, one can conclude that $\xi_{t+1} \in [\sqrt{2\mu\Delta_\gamma} - \epsilon, \sqrt{2\mu\Delta_\gamma}]$ whenever

$$t \ge \max\left\{2\frac{\log(\mathcal{C}_{\mu,L,\gamma}\cdot\sqrt{D_0}/\mathcal{D}_{\kappa,\mu,\gamma})}{\log(1/(1-2\mu\Delta_\gamma)))}, \frac{\log(2\kappa\mathcal{C}_{\mu,L,\gamma}^2 D_0/\epsilon)}{\log(1/(1-2\mu\Delta_\gamma)))}\right\} + \frac{\log(2\sqrt{2\mu\Delta_\gamma}/\epsilon)}{\log\left(1/\left(1-\frac{8\mu\Delta_\gamma}{5+\sqrt{5}}\right)\right)},$$

concluding the proof of the the convergence rate of $\{\xi_t\}$ in Theorem 4.1.

Appendix H. Extension to the non-strongly geodesically convex case

In the Euclidean case, it is well-known that one can obtain acceleration guarantees for the non-strongly convex case from the strongly convex case; see e.g., (Gasnikov and Nesterov, 2018, Theorem 4). In this section, we extend such an argument to the Riemannian setting and use it to establish accelerated guarantees for the non-strongly g-convex case under the constant distortion assumption.

To that end, we will need the following properties of the distance function:

Proposition H.1 *Let M be a Riemannian manifold with sectional curvatures lower bounded by $-\kappa < 0$. Then, for a fixed $p \in M$, the distance function $d(x) := \frac{1}{2}d(x,p)^2 : M \to \mathbb{R}$ satisfies:*

1. *d is 1-strongly g-convex in the entire M with $\nabla d(x) = -\mathrm{Exp}_x^{-1}(p)$.*

2. *For $D \geq 0$, d is geodesically $\frac{\sqrt{\kappa}D}{\tanh(\sqrt{\kappa}D)}$-smooth within the domain $\{u \in M \ : \ d(u,p) \leq D\}$.*

Proof Let us first verify the strong g-convexity. Let x, y be arbitrary points on M. Then,

$$d(y,p)^2 \geq d_x(y,p)^2 = d(x,p)^2 + d(x,y)^2 - 2\left\langle \mathrm{Exp}_x^{-1}(p), \mathrm{Exp}_x^{-1}(y)\right\rangle_x .$$

Using the notation $d(\cdot)$ and noting that $\nabla d(x) := -\mathrm{Exp}_x^{-1}(p)$, we get

$$d(y) \geq d(x) + \left\langle \nabla d(x), \mathrm{Exp}_x^{-1}(y)\right\rangle_x + \frac{1}{2}\cdot d(x,y)^2 ,$$

which is precisely the definition of geodesic 1-strong convexity (see Definition E.1). Next, we verify the geodesic smoothness. From the global trigonometry inequality (Proposition C.3),

$$d(y,p)^2 \leq d(x,p)^2 + \frac{\sqrt{\kappa}d(x,p)}{\tanh(\sqrt{\kappa}d(x,p))}\cdot d(x,y)^2 - 2\left\langle \mathrm{Exp}_x^{-1}(p), \mathrm{Exp}_x^{-1}(y)\right\rangle_x ,$$

which can be rewritten as

$$d(y) \leq d(x) + \left\langle \nabla d(x), \mathrm{Exp}_x^{-1}(y)\right\rangle_x + \frac{\sqrt{\kappa}d(x,p)}{2\tanh(\sqrt{\kappa}d(x,p))}\cdot d(x,y)^2 .$$

From this, one can deduce geodesic $\frac{\sqrt{\kappa}D}{\tanh(\sqrt{\kappa}D)}$-smoothness of d (see Definition E.2). ■

The next ingredient is the extension of the folklore reduction argument to the Riemannian case:

Proposition H.2 (Reduction argument) *Given an accuracy $\epsilon > 0$, a Riemannian manifold M, and a point $x_0 \in M$, let $\mu > 0$ be a constant satisfying $\mu \leq \epsilon/d(x_*,x_0)^2$. Suppose that $x_{\mathrm{sol}} \in M$ is an $\epsilon/2$-suboptimal solution to $\underset{x\in M}{\mathrm{minimize}}\left(f(x) + \mu/2\cdot d(x,x_0)^2\right)$. Then, $f(x_{\mathrm{sol}}) - f(x_*) \leq \epsilon$.*

Proof By the definition of x_{sol}, we have $f(x_{\mathrm{sol}}) \leq f(x_*) + \frac{\mu}{2}d(x_*,x_0)^2 + \frac{\epsilon}{2} \leq \epsilon$. ■

Using Propositions H.1 and H.2, Corollary 3.1 can be extended to the non-strongly g-convex case by perturbing the cost function. More specifically, when f is geodesically L-smooth, then $f + \frac{\mu}{2}\cdot d(x,x_0)^2$ is geodesically $L + \mu\frac{\sqrt{\kappa}D}{\tanh(\sqrt{\kappa}D)}$-smooth and μ-strongly convex within $\{u \in M \ : \ d(u,x_0) \leq D\}$. Hence, as long as the algorithm iterates stay within the bounded domain, one can use the reduction argument to obtain accelerated rate for non-strongly convex costs:

Corollary H.1 *Let $\epsilon \in (0,1)$ be an arbitrary accuracy, and f be a geodesically L-smooth function. Assume that there exists $D > 0$ such that*

1. $\epsilon < \frac{L}{2}\cdot d(x_*,x_0)^2 \cdot \frac{\tanh(\sqrt{\kappa}D)}{\sqrt{\kappa}D}$.
2. *All iterates of (3.1) with parameters chosen as per Theorem 3.1 with $\gamma_t \equiv 1/L$, $\mu = \frac{\epsilon}{d(x_*,x_0)^2}$ and $\delta_t \equiv S_\kappa(2D) = \left(\frac{\sinh(\sqrt{\kappa}2D)}{\sqrt{\kappa}2D}\right)^2$ stay within $\{u \in M \ : \ d(u,x_0) \leq D\}$.*

Then, one can find an ϵ-suboptimal solution to $\underset{x\in M}{\mathrm{minimize}}\, f(x)$, within $O\left(\epsilon^{-1/2}\log(1/\epsilon)\right)$ iterations, where the constant involved in $O(\cdot)$ depends only on κ, D, L.

Remark H.3 *It is important to note that Corollary H.1 is not a complete result but rather a proof of concept as it assumes that all iterates with a certain parameter choices stay within a bounded domain. In particular, it would be interesting to see if such an assumption can be guaranteed following the arguments in Appendix F. Moreover, compared to the acceleration result from the Euclidean case (Nesterov, 1983), Corollary H.1 is not fully satisfactory: the target accuracy $\epsilon > 0$ needs to be chosen beforehand, and an extra $\log(1/\epsilon)$ term appears in the iteration complexity. It would be interesting to see if one can overcome such shortcomings of the reduction argument, and we leave it as an open question.*

Proof Let us take $\mu = \epsilon/d(x_*,x_0)^2$. Then, Proposition H.2 implies that arbitrary $\epsilon/2$-suboptimal solution $x_{\mathrm{sol}} \in M$ to $\underset{x\in M}{\mathrm{minimize}} \left(f(x) + \mu/2 \cdot d\left(x, x_0\right)^2\right)$ satisfies $f(x_{\mathrm{sol}}) - f(x_*) \leq \epsilon$.

On the other hand, note that $f + \frac{\mu}{2} \cdot d\left(x, x_0\right)^2$ is geodesically $L + \mu\frac{\sqrt{\kappa}D}{\tanh(\sqrt{\kappa}D)}$-smooth and μ-strongly convex within $\{u \in M \ : \ d\left(u, x_0\right) \leq D\}$. Hence, by choosing $\gamma_t \equiv 1/L$, we have

$$\Delta_\gamma = \frac{1}{L}\left(1 - \frac{L + \frac{\epsilon}{d(x_*,x_0)^2} \cdot \frac{\sqrt{\kappa}D}{\tanh(\sqrt{\kappa}D)}}{2L}\right) \geq \frac{1}{L}\left(1 - \frac{L + \frac{L}{2}}{2L}\right) = \frac{1}{4L},$$

where the inequality follows due to the assumption $\epsilon < \frac{L}{2} \cdot d\left(x_*, x_0\right)^2 \cdot \frac{\tanh(\sqrt{\kappa}D)}{\sqrt{\kappa}D}$.

Since all the iterates stay within a subset of diameter D, Rauch comparison theorem (Proposition 4.1) implies that the constant distortion condition holds with $\delta = S_\kappa(2D)$. Hence, Corollary 3.1 implies that (3.1) with finds an $\epsilon/2$-suboptimal solution within iterations bounded by

$$O\left(\left(\sqrt{(\delta-1)^2 + \epsilon \cdot \frac{\delta}{Ld\left(x_*, x_0\right)^2}} - (\delta - 1)\right)^{-1} \log(2/\epsilon)\right),$$

which is of $O\left(\epsilon^{-1/2}\log(1/\epsilon)\right)$. ■

Appendix I. Extension to positively-curved manifolds

Let us now assume that the sectional curvatures of M is upper bounded by $\sigma \geq 0$. In particular, the case with $\sigma = 0$ corresponds to the non-positively curved case. We first pinpoint the main differences: unlike the the non-positively curved case, M now may not be uniquely geodesic. Instead, one can only guarantee the property within a local neighborhood of M. Consequently, the notion of geodesic convexity can be guaranteed only within a local neighborhood of M. For instance, manifolds with positive sectional curvatures (e.g. spheres) are compact, and hence, they do not admit globally geodesically convex functions other than the constant functions. Following the prior arts (Dyer et al., 2015; Zhang and Sra, 2018), we make the following assumptions to avoid any further complications:

Assumption 3 *The domain $N \subset M$ of f is uniquely geodesic with the diameter bounded by $\frac{\pi}{2\sqrt{\sigma}}$.*

Assumption 4 (Bounded iterates assumption) *All the iterates of Algorithm 1 (whose parameters will be chosen later) remain in N.*

The analysis for the positively curved case is identical to that for the non-positively curved case, modulo an additional geometric inequality due to (Zhang and Sra, 2018):

Proposition I.1 ((Zhang and Sra, 2018, Lemma 7)) *Let x, y, z be points on Riemannian manifold M with sectional curvatures upper bounded by $\sigma \geq 0$. If $d(x,z) \leq \frac{\pi}{2\sqrt{\sigma}}$, then*

$$d_x(y,z)^2 \leq (1 + 2\sigma \cdot d(x,y)^2) \cdot d(y,z)^2 \ .$$

Applying Proposition I.1 to Lemma C.1, we obtain the following metric distortion inequality:

Lemma I.1 (Modification of Lemma C.1) *Let x, x', y, z be points on Riemannian manifold M with sectional curvatures upper and lower bounded by σ and $-\kappa < 0$, respectively. If $d(x',z) \leq \frac{\pi}{2\sqrt{\sigma}}$, then for $\widehat{T_\kappa} : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 1}$ defined as in Lemma C.1, we have*

$$d_{x'}(y,z)^2 \leq \widehat{T_\kappa}(d(x,y)) \cdot (1 + 2\sigma \cdot d\left(x', y\right)^2) \cdot d_x(y,z)^2 \, .$$

From Lemma I.1, one can conclude that at iteration $t \geq 1$,

$$T_\kappa(d(x_t, z_t)) \cdot (1 + 2\sigma \cdot d(y_t, z_t)^2) \tag{I.1}$$

is a valid distortion rate. Thus, one can use (I.1) in lieu of $T_\kappa(d(x_t, z_t))$ for the valid distortion rate in Algorithm 1. Then, one can invoke Theorem 3.1 with the chosen valid distortion rate (I.1) to guarantee the potential decrease. To show that Algorithm 1 with (I.1) eventually achieves full acceleration, the last ingredient is to show that the distances $d(x_t, z_t)$ and $d(y_t, z_t)$ shrink over iterations. Indeed, one can prove that the distances shrink following the arguments in Appendix F. The only difference is that in proving Proposition F.1, one now has the following in place of (F.1):

$$d_{x_t}(y_t, z_t) \leq d_{x_t}(y_t, x_*) + d_{x_t}(z_t, x_*) \leq (1 + \pi^2/2) \cdot d(y_t, x_*) + d_{x_t}(z_t, x_*) \, , \tag{I.2}$$

where the last inequality is due to Proposition I.1 together with the bounded iterates assumption (Assumption 4). Hence the third statement of Proposition F.1 now holds with an additional multiplication constant of $1 + \pi^2/2$. With this modification, the rest follows in exactly the same manner. We skip the details as they significantly overlap with the non-positively curved case.