From Proximal Point Method
To Accelerated Methods on Riemannian Manifolds

by

Kwangjun Ahn

B.S., Korea Advanced Institute of Science and Technology (2017)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2021

Author . . . .
Department of Electrical Engineering and Computer Science
May 11, 2021

Certified by . . . .
Suvrit Sra
Associate Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by . . . .
Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students

From Proximal Point Method
To Accelerated Methods on Riemannian Manifolds

by

Kwangjun Ahn

Submitted to the Department of Electrical Engineering and Computer Science
on May 11, 2021, in partial fulfillment of the
requirements for the degree of
Master of Science in Electrical Engineering and Computer Science

Abstract

Recently, there has been significant effort to generalize successful ideas in Euclidean optimization to Riemannian optimization. However, one landmark result of Euclidean optimization has eluded the Riemannian setting: namely, a Riemannian analog of Nesterov's accelerated gradient method (AGM). In this thesis, we establish the first globally accelerated gradient method for Riemannian manifolds.

Toward establishing our result, the first part of the thesis revisits Nesterov's AGM and develops a conceptually simple understanding of it based on the proximal point method (PPM). The main observation is that AGM is in fact an approximation of PPM, which results in simple derivations and analyses of different versions of AGM.

The second part of the thesis then extends our simple approach to the Riemannian case. In our extension, we handle a technical hurdle inherent to the Riemannian case by introducing an appropriate notion of "*metric distortion.*" We control this distortion via a novel geometric inequality, which enables us to formulate and analyze global Riemannian acceleration.

Thesis Supervisor: Suvrit Sra
Title: Associate Professor of Electrical Engineering and Computer Science

Acknowledgments

First, I would like to thank my adviser Prof. Suvrit Sra for helping me perform the research constituting this thesis. This thesis would have not been completed without his help and guidance. I have grown a lot both academically and personally through all the discussions I have had with him. I always appreciate him teaching me how to ask critical questions and think outside the box. With his guidance, I constantly learn, grow and stay inspired. Next, I would like to thank all my collaborators over the last two years. Collaborations have given me a broader perspective on research.

Outside research, I would like to thank all members of MIT Gymnastics Team. Training gymnastics and working out with them have been a great part of my graduate school life. I would have not been able to stay motivated without them.

My last acknowledgement goes to my family. Thank you for your constant support and love. Over the years, I have run into several difficulties and got carried away. It has been always through their love that I was able to regroup myself and bounce back.

Contents

List of Figures

Chapter I

Introduction and outline

I.1 Main question

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 [Gro78, BBB$^+$01, BH13]. Tractability through the lens of g-convexity has been fruitful in several applications (e.g., see [ZS16, §1.1]) and also some purely theoretical questions [BFG$^+$19, GS19].

Paralleling the theory and applications of g-convexity is the progress on algorithms, primarily set in Riemannian manifolds [Udr94, AMS09] and CAT(0) spaces [Bac14]. Earlier studies focus on *asymptotic* analysis, while [ZS16] obtains the first *non-asymptotic* iteration complexity analysis for Riemannian (stochastic) gradient methods. Subsequent works establish iteration complexity for Riemannian proximal-point methods [BFM17], Frank-Wolfe [WS19], variance reduced methods [ZRS16, KSM16, ZZS18, ZYYF19], trust-region methods [ABBC20], 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* (AGM) [Nes83] (see the beginning of Chapter II for background). This gap motivates the main question:

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

In this thesis, we tackle this question based on the following two steps:

① Revisit Nesterov's acceleration in the Euclidean setting and develop a simple derivation and analysis for it.

② Extend the derivation and analysis to the Riemannian setting.

Consequently, this thesis consists of two parts. The first part (consisting of Chapters II and III) addresses ①, and the second part (Chapter IV) addresses ②.

I.2 Outline

To ease presentation, we provide background at the beginning of each chapter and discuss related work at the end of each chapter.

In Chapter II, we develop a simple derivation and analysis for AGM in the Euclidean setting. Our approach is based on connecting AGM to another well-known optimization method called the proximal point method (PPM). In §II.1, we first present a brief background on PPM including its analysis. In §II.2, we then consider two simplest ways to approximate PPM, and discuss their limitations. In §II.3, we discuss how a combination of the two simplest approximations of PPM in fact recovers a version of AGM, and give a simple analysis based on PPM. In §II.4, we demonstrate how our framework recovers other well known versions of AGM, including the momentum version and the similar triangle version.

In Chapter III, we extend our simple derivation and analysis to the strongly convex setting. In §III.1, based on our simple derivation, we derive the most general version of AGM due to Nesterov called "*General Scheme for Optimal Method*" [Nes18, (2.2.7)]. We extend our PPM-based analysis to strongly convex costs. As a warm-up, in §III.2, we first consider the simple case of constant step sizes and recover the famous parameter choice due to Nesterov (e.g., [Nes18, (2.2.22)]). In §III.3, we then consider the general case, and demonstrate that one can recover the elaborate step sizes choice [Nes18, (2.2.22)] from first principles. In §III.4, we provide the proofs of technical statements presented in the chapter.

In Chapter IV, we develop the first global accelerated gradient method for Riemannian manifolds, building on the results form Chapter III. In §IV.1, we provide a short preliminaries on Riemannian geometry required for our development. In §IV.2, we consider a Riemannian analog of "General Scheme for Optimal Method" and analyze it by modifying the techniques from §III.3 using the notion of metric distortion. In §IV.3, we discuss how one can estimate metric distortion rates in terms of the known quantity and develop a globally accelerated gradient method for Riemannian manifolds. In §IV.4, we validate our estimation scheme for metric distortion rates by developing new geometric inequalities. In §IV.5, we analyze how the estimated distortion rate changes over iterations. In §IV.6, we combine all the ingredients together and prove an accelerated convergence rate.

In Chapter V, we conclude this thesis with future directions.

Chapter II

From proximal point method to accelerated gradient methods

This chapter discusses accelerated methods in the Euclidean domain and serves as a stepping stone for answering the main question of this thesis. We remark that readers who are mainly interested in our result for the Riemannian case can skip ahead to Chapter III.

In 1983, Nesterov introduced the *accelerated gradient method* (AGM) for convex optimization. AGM is a gradient method that achieves strictly faster convergence rates than gradient descent. On top of its accelerated rates, AGM is easy to implement, and it has been applied to a myriad of applications. The list applications includes sparse linear regression [BT09], compressed sensing [BBC11], the maximum flow problem [LRS13], and deep neural networks [SMDH13]. Paralleling its success both in theory and practice, there have been a flurry of works trying to understand the scope and principles of AGM [SBC16, KBB15, WWJ16, LRP16, WRJ16, AZO17, DO19].

Despite numerous attempts to understand AGM, one aspect of AGM not well understood in the literature is the fact that it appears in many different forms. In fact, ever since the original version due to Nesterov, there have been several different versions of it; below, we list the most representative ones:

$$\begin{aligned} z_{t+1} &= y_t - \alpha_t^{(1)} \nabla f(y_t)\,, \\ y_{t+1} &= z_{t+1} + \beta_t^{(1)} (z_{t+1} - z_t)\,. \end{aligned}$$

Form I [Nes83, BT09].

$$\begin{aligned} y_t &= \alpha_t^{(2)} x_t + (1 - \alpha_t^{(2)}) z_t\,, \\ z_{t+1} &= y_t - \beta_t^{(2)} \nabla f(y_t)\,, \\ x_{t+1} &= x_t - \gamma_t^{(2)} \cdot \nabla f(y_t)\,. \end{aligned}$$

Form II [Nes18, AZO17].

$$\begin{aligned} y_t &= \alpha_t^{(3)} x_t + (1 - \alpha_t^{(3)}) z_t\,, \\ x_{t+1} &= x_t - \beta_t^{(3)} \nabla f(y_t)\,, \\ z_{t+1} &= \gamma_t^{(3)} x_{t+1} + (1 - \gamma_t^{(3)}) z_t\,. \end{aligned}$$

Form III [AT06, Tse08, GN18].

$$\begin{aligned} y_t &= \alpha_t^{(4)} x_t + (1 - \alpha_t^{(4)}) z_t\,, \\ x_{t+1} &= \beta_t^{(4)} x_t + (1 - \beta_t^{(4)}) y_t - \gamma_t^{(4)} \nabla f(y_t)\,, \\ z_{t+1} &= y_t - \delta_t^{(4)} \nabla f(y_t)\,. \end{aligned}$$

Form IV [Nes18].

Here $\alpha_t^{(\cdot)}, \beta_t^{(\cdot)}, \gamma_t^{(\cdot)}, \delta_t^{(\cdot)}$ are some carefully chosen step sizes.

In this chapter, we present a way to understand the details of AGM from the proximal point method (PPM). Our approach is inspired by that of [Def19]. The main observation is that different versions of AGM can be derived by viewing them as approximations of PPM. On top of simple derivations, the PPM view of AGM also offers simple analyses of different versions of AGM based on the standard analysis of PPM [Gül91]. Moreover, our view gives rise to the key idea of the *method of similar triangles*, a version of AGM shown to have simple extensions to practically relevant settings [Tse08, GN18]. Our approach also readily extends to the strongly convex case as we discuss in Chatper III.

We first provide a brief background on the proximal point method.

II.1 Brief background on the proximal point method

The *proximal point method (PPM)* [Mor65, Mar70, Roc76] is a fundamental method in optimization which solves the minimization of the cost function $f : \mathbb{R}^d \to \mathbb{R}$ by iteratively solving the subproblem

$$x_{t+1} \leftarrow \operatorname*{argmin}_{x \in \mathbb{R}^d} \left\{ f(x) + \frac{1}{2\eta_{t+1}} \|x - x_t\|^2 \right\} \tag{II.1}$$

for a step size $\eta_{t+1} > 0$, where the norm is chosen as the ℓ_2 norm. The motivation of the method is clear: we add a quadratic regularization to make the cost function well conditioned for faster optimization. Nevertheless, solving (II.1) is in general as difficult as solving the original optimization problem, and PPM is largely regarded as a "conceptual" guiding principle for accelerating optimization algorithms [Dru17].

The baseline of our discussion is the following convergence rate of PPM for convex costs proved in a seminal paper by Güler [Gül91] (here x_* denotes a global optimum point, i.e., $x_* \in \operatorname{argmin}_x f(x)$):

$$\boxed{f(x_T) - f(x_*) \leq O\left(\left(\sum_{t=1}^{T} \eta_t \right)^{-1} \right) \quad \text{for any } T \geq 1.} \tag{II.2}$$

In words, one can achieve an arbitrarily fast convergence rate by choosing step sizes η_t's large. Below, we review a short Lyapunov function proof of (II.2), which will serve as a backbone to other analyses.

Proof of (II.2). It turns out that the following Lyapunov function is suitable:

$$\boxed{\Phi_t := \left(\sum_{i=1}^{t} \eta_i \right) \cdot \left(f(x_t) - f(x_*) \right) + \tfrac{1}{2} \|x_* - x_t\|^2 ,} \tag{II.3}$$

where $\Phi_0 := \frac{1}{2} \|x_* - x_0\|^2$ and here and below, $\|\cdot\|$ is the ℓ_2 norm unless stated otherwise. Now, it suffices to show that Φ_t is decreasing, i.e., $\Phi_{t+1} \leq \Phi_t$ for all $t \geq 0$. Indeed, if Φ_t is decreasing, we have $\Phi_T \leq \Phi_0$ for any $T \geq 1$, which precisely recovers (II.2). To that end, we use a standard result:

Proposition II.1 (Proximal inequality (see e.g. [BC11, Proposition 12.26])). *For a convex function $\phi : \mathbb{R}^d \to \mathbb{R}$, let x_{t+1} be the unique minimizer of the following proximal* *step:* $x_{t+1} \leftarrow \operatorname{argmin}_{x\in\mathbb{R}^d} \left\{\phi(x) + \frac{1}{2}\|x - x_t\|^2\right\}$. *Then, for any* $u \in \mathbb{R}^d$,

$$\phi(x_{t+1}) - \phi(u) + \frac{1}{2}\|u - x_{t+1}\|^2 + \frac{1}{2}\|x_{t+1} - x_t\|^2 - \frac{1}{2}\|u - x_t\|^2 \le 0\,.$$

Now Proposition II.1 completes the proof as follows: First, we apply Proposition II.1 with $\phi = \eta_{t+1} f$ and $u = x_*$ and drop the term $\frac{1}{2}\|x_{t+1} - x_t\|^2$ to obtain:

$$\eta_{t+1}\left[f(x_{t+1}) - f(x_*)\right] + \frac{1}{2}\|x_* - x_{t+1}\|^2 - \frac{1}{2}\|x_* - x_t\|^2 \le 0\,. \tag{Ineq_1}$$

Next, from the optimality of x_{t+1}, it readily follows that

$$f(x_{t+1}) - f(x_t) \le 0\,. \tag{Ineq_2}$$

Now, computing (IV.8) + $(\sum_{i=1}^{t}\eta_i)\times$(IV.9) yields $\Phi_{t+1} \le \Phi_t$, which finishes the proof. □

II.1.1 Our conceptual question

Although the convergence rate (II.2) seems powerful, it does not have any practical values as PPM is in general not implementable. Nevertheless, one can ask the following conceptual question:

"Can we efficiently approximate PPM for a large step size η_t?"

Perhaps, the most straightforward approximation would be to replace the cost function f in (II.1) with its lower-order approximations. We implement this idea in the next section.

II.2 Two simple approximations of the proximal point method

To analyze approximation errors, let us assume that the cost function f is L-smooth.

Definition II.1 (Smoothness)**.** For $L > 0$, we say a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth if $f(x) \le f(y) + \langle \nabla f(y), x - y\rangle + \frac{L}{2}\|x-y\|^2$ for any $x, y \in \mathbb{R}^d$.

From the convexity and the L-smoothness of f, we have the following lower and upper bounds: for any $x, y \in \mathbb{R}^d$,

$$\boxed{\underbrace{f(y) + \langle \nabla f(y), x - y\rangle}_{=:\ \mathsf{LOWER}(x;y)} \le f(x) \le \underbrace{f(y) + \langle \nabla f(y), x - y\rangle + \frac{L}{2}\|x-y\|^2}_{=:\ \mathsf{UPPER}(x;y)} .}$$

In this section, we use these bounds to approximate PPM.

II.2.1 First approach: using first-order approximation

Let us first replace f in the objective (II.1) with its lower approximation:

$$x_{t+1} \leftarrow \operatorname*{argmin}_x \left\{ \mathsf{LOWER}(x;x_t) + \frac{1}{2\eta_{t+1}}\|x - x_t\|^2 \right\} . \tag{II.4}$$

Writing the optimality condition, one quickly notices that (II.4) actually leads to gradient descent:

$$x_{t+1} = x_t - \eta_{t+1}\nabla f(x_t) \, . \tag{II.5}$$

Let us see how well (II.4) approximates PPM:

Analysis of the first approach. We first establish counterparts of (IV.8) and (IV.9). First, we apply Proposition II.1 with $\phi(x) = \eta_{t+1}\mathsf{LOWER}(x;x_t)$ and $u = x_*$:

$$\phi(x_{t+1}) - \phi(x_*) + \frac{1}{2}\|x_* - x_{t+1}\|^2 + \frac{1}{2}\|x_{t+1} - x_t\|^2 - \frac{1}{2}\|x_* - x_t\|^2 \le 0 \, .$$

Now using convexity and L-smoothness, we have

$$\phi(x) \le \eta_{t+1} f(x) \le \phi(x) + \frac{L\eta_{t+1}}{2}\|x - x_t\|^2 \, ,$$

and hence the above inequality implies the following analogue of (IV.8):

$$\eta_{t+1}\left[f(x_{t+1}) - f(x_*)\right] + \frac{1}{2}\left\|x_* - x_{t+1}\right\|^2 - \frac{1}{2}\left\|x_* - x_t\right\|^2 \le (\mathcal{E}_1^{\mathsf{GD}}), \qquad (\mathsf{Ineq}_1^{\mathsf{GD}})$$

where $(\mathcal{E}_1^{\mathsf{GD}}) := (\frac{L\eta_{t+1}}{2} - \frac{1}{2})\left\|x_{t+1} - x_t\right\|^2$. Next, we use the L-smoothness of f and the fact $\nabla f(x_t) = -1/\eta_{t+1}(x_{t+1} - x_t)$ (due to (II.5)), to obtain the following analogue of (IV.9):

$$f(x_{t+1}) - f(x_t) \le \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \frac{L}{2}\left\|x_{t+1} - x_t\right\|^2 = (\mathcal{E}_2^{\mathsf{GD}}), \qquad (\mathsf{Ineq}_2^{\mathsf{GD}})$$

where $(\mathcal{E}_2^{\mathsf{GD}}) := (\frac{L}{2} - \frac{1}{\eta_{t+1}})\left\|x_{t+1} - x_t\right\|^2$.

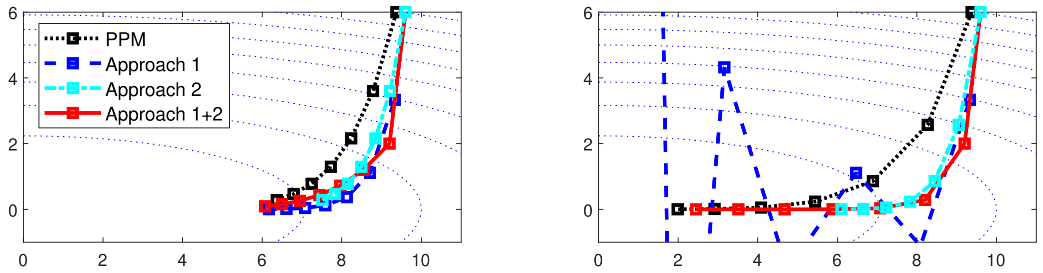
Now paralleling the proof of (II.2), to show that Φ_t (II.3) is a valid Lyapunov function, we need to find the step sizes η_t's that satisfy the following relation: $(\mathcal{E}_1^{\mathsf{GD}}) + (\sum_{i=1}^{t}\eta_i) \times (\mathcal{E}_2^{\mathsf{GD}}) \le 0$. On the other hand, note that both $(\mathcal{E}_1^{\mathsf{GD}})$ and $(\mathcal{E}_2^{\mathsf{GD}})$ become positive numbers when $\eta_{t+1} > 2/L$. Hence, the admissible choices for η_t at each iteration are upper bounded by $2/L$, which together with the PPM convergence rate (II.2) implies that $O(1/\sum_{t=1}^{T}\eta_t) = O(1/T)$ is the best convergence rate one can prove. Indeed, choosing $\eta_t \equiv 1/L$, then we have $(\mathcal{E}_1^{\mathsf{GD}}) = 0$ and $(\mathcal{E}_2^{\mathsf{GD}}) < 0$, obtaining the well-known bound of $f(x_T) - f(x_*) \le \frac{L\|x_0 - x_*\|^2}{2T} = O(1/T)$. □

To summarize, the first approach only leads to a disappointing result: the approximation is valid only for the small step size regime of $\eta_t = O(1)$. We empirically verify this fact for a quadratic cost in Figure II-1. As one can see from Figure II-1, the lower approximation approach (II.4) overshoots for large step sizes like $\eta_t = \Theta(t)$ and quickly steers away from PPM iterates.

II.2.2 Second approach: using smoothness

After seeing the disappointing outcome of the first approach, our second approach is to replace f with its upper approximation due to the L-smoothness:

$$x_{t+1} \leftarrow \operatorname*{argmin}_{x}\left\{\mathsf{UPPER}(x; x_t) + \frac{1}{2\eta_{t+1}}\left\|x - x_t\right\|^2\right\}. \qquad \text{(II.6)}$$



(a) $\eta_t \equiv 1/3$.

(b) $\eta_t = t/3$.

Figure II-1: Iterates comparison between PPM (II.1), the first approach (II.4), the second approach (II.6), and the combined approach (II.8). For the setting, we choose $f(x, y) = 0.1x^2 + y^2$ and $x_0 = (10, 10)$.

Writing the optimality condition, (II.6) actually leads to a conservative update of gradient descent:

$$x_{t+1} = x_t - \frac{1}{L + \eta_{t+1}^{-1}} \nabla f(x_t)\,. \tag{II.7}$$

Note that regardless of how large η_{t+1} we choose, the actual update step size in (II.7) is always upper bounded by $^1/_L$. Although this conservative update prevents the overshooting phenomenon of the first approach, as we increase η_t, this conservative update becomes too tardy to be a good approximation of PPM; see Figure II-1.

II.3 Nesterov's acceleration via alternating two approaches

In the previous section, we have seen that the two simple approximations of PPM both have limitations. Nonetheless, observe that their limitations are opposite to each other: while the first approach is too "reckless," the second approach is too "conservative." This observation motivates us to consider a *combination* of the two approaches which could mitigate each other's limitation.

Remark II.1. A similar interpretation of Nesterov's acceleration as a combination of a reckless step and a conservative step also appeared in [AZO17, BG19]. However, as we shall see, the interpretation based on PPM will bring about a more refined understanding of Nesterov's acceleration.

Let us implement this idea by alternating between the two approximations (II.4) and (II.6) of PPM. The key modification is that for both approximations, we introduce an additional sequence of points $\{y_t\}$ for cost function approximation; i.e., we use the following approximations for the t-th iteration:

$$f(y_t) + \langle \nabla f(y_t), x - y_t \rangle \leq f(x) \leq f(y_t) + \langle \nabla f(y_t), x - y_t \rangle + \frac{L}{2} \left\| x - y_t \right\|^2 .$$

Indeed, this modification is crucial: if we just use approximations at x_t, the resulting alternation merely concatenates (II.4) and (II.6) during each iteration, and the two limitations we discussed in §II.2 will remain in the combined approach.

Having introduced a separate sequence $\{y_t\}$ for cost approximations, we consider the following alternation where during each iteration, we update x_t with (II.4) and y_t with (II.6):

Approximate PPM with alternating two approaches. Given $x_0 \in \mathbb{R}^d$, let $y_0 = x_0$ and run:

$$x_{t+1} \leftarrow \operatorname{argmin}_x \left\{ \mathsf{LOWER}(x; y_t) + \tfrac{1}{2\eta_{t+1}} \left\| x - x_t \right\|^2 \right\}, \qquad \text{(II.8a)}$$

$$y_{t+1} \leftarrow \operatorname{argmin}_x \left\{ \mathsf{UPPER}(x; y_t) + \tfrac{1}{2\eta_{t+1}} \left\| x - x_{t+1} \right\|^2 \right\}. \qquad \text{(II.8b)}$$

In Figure II-1, we empirically verify that (II.8) indeed gets the best of both worlds: this combined approach successfully approximates PPM even for the regime $\eta_t = \Theta(t)$. More remarkably, (II.8) is exactly equal to one version of AGM ("Form II" in the introduction). Turning (II.8) into the equational form by writing the optimality conditions, and introducing an auxiliary iterate $z_{t+1} := y_t - {}^1\!/\!_L \nabla f(y_t)$ (only for simplicity), we obtain the following ($x_0 = y_0 = z_0$):

Equivalent representation of (II.8):

$$y_t = \frac{1/L}{1/L+\eta_t} x_t + \frac{\eta_t}{1/L+\eta_t} z_t \,, \tag{II.9a}$$

$$x_{t+1} = x_t - \eta_{t+1} \nabla f(y_t)\,, \tag{II.9b}$$

$$z_{t+1} = y_t - \tfrac{1}{L} \cdot \nabla f(y_t)\,. \tag{II.9c}$$

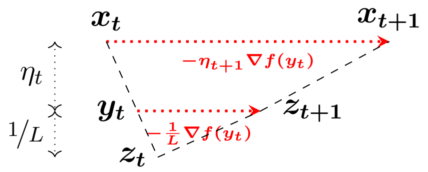


Figure II-2: Illustration of (II.9).

Hence, we arrive at AGM without relying on any non-trivial derivations in the literature such as estimate sequence [Nes18] or linear coupling [AZO17]. To summarize, we have demonstrated:

Nesterov's AGM is a simple approximation of the proximal point method!

Remark II.2. Our derivation is inspired by the work of Defazio [Def19, §5,6]. However, unlike the approach in [Def19], our derivation does not rely on duality, which could be advantageous in the settings where duality fails.

II.3.1 Understanding mysterious parameters of AGM

It is often the case in the literature that the interpolation step (II.9a) is written as an abstract form $y_t = \tau_t x_t + (1-\tau_t) z_t$ with a weight parameter $\tau_t > 0$ to be chosen [AZO17, LRP16, WRJ16, BG19]. That said, in the previous works, τ_t is carefully chosen according to the analysis without conveying much intuition. One important aspect of our PPM view is that it reveals a close relation between the weight parameter τ_t and the step size η_t. More specifically, τ_t is chosen so that the ratio of the distances $\|y_t - x_t\| : \|y_t - z_t\|$ is equal to $\eta_t : 1/L$ (see Figure II-2).

II.3.2 Analysis based on PPM perspective

In order to determine η_t's in (II.9), we revisit the analysis of PPM from §II.2. In turns out that following §II.2.1, one can derive the following analogues of (IV.8) and (IV.9) using Proposition II.1 (we defer the derivations to §II.5.1):

$$\eta_{t+1}(f(z_{t+1}) - f(x_*)) + \frac{1}{2}\|x_* - x_{t+1}\|^2 - \frac{1}{2}\|x_* - x_t\|^2 \le (\mathcal{E}_1^{\mathsf{AGM}}), \tag{$\mathsf{Ineq}_1^{\mathsf{AGM}}$}$$

$$f(z_{t+1}) - f(z_t) \le (\mathcal{E}_2^{\mathsf{AGM}}), \tag{$\mathsf{Ineq}_2^{\mathsf{AGM}}$}$$

where $(\mathcal{E}_1^{\mathsf{AGM}}) := (\frac{\eta_{t+1}^2}{2} - \frac{\eta_{t+1}}{2L})\|\nabla f(y_t)\|^2 + L\eta_t\eta_{t+1}\langle \nabla f(y_t), z_t - y_t\rangle$ and $(\mathcal{E}_2^{\mathsf{AGM}}) := -\frac{1}{2L}\|\nabla f(y_t)\|^2 - \langle \nabla f(y_t), z_t - y_t\rangle$. Hence, we modify the Lyapunov function (II.3) by replacing the first x_t with z_t:

$$\Phi_t := (\textstyle\sum_{i=1}^t \eta_i) \cdot (f(z_t) - f(x_*)) + \frac{1}{2}\|x_* - x_t\|^2 . \tag{II.10}$$

We note that (II.10) is not new; it also appears in prior works [WRJ16, DO19, BG19], although with different motivations.

Then as before, to prove the validity of the chosen Lyapunov function, it suffices to verify $(\mathcal{E}_1^{\mathsf{AGM}}) + (\sum_{i=1}^t \eta_i) \cdot (\mathcal{E}_2^{\mathsf{AGM}}) \le 0$, which is equivalent to

$$\tfrac{1}{2L}\left(L\eta_{t+1}^2 - \textstyle\sum_{i=1}^{t+1}\eta_i\right)\|\nabla f(y_t)\|^2 + \left(L\eta_t\eta_{t+1} - \textstyle\sum_{i=1}^t \eta_i\right)\langle \nabla f(y_t), z_t - y_t\rangle \le 0 \tag{II.11}$$

From (II.11), it suffices to choose $\{\eta_t\}$ so that $L\eta_t\eta_{t+1} = \sum_{i=1}^t \eta_i$. Indeed, with such a choice, the coefficient of the inner product term in (II.11) becomes zero and the coefficient of the squared norm term becomes ${}^1\!/_{2L}(L\eta_{t+1}^2 - L\eta_{t+1}\eta_{t+2}) \le 0$ (if $\{\eta_t\}$ is increasing). Indeed, one can quickly notice that choosing $\eta_t = {}^t\!/_{2L}$ satisfies the desired relation. Therefore, we obtain the well known accelerated convergence rate of $f(z_T) - f(x_*) \le \frac{2L\|x_0 - x_*\|^2}{T(T+1)} = O({}^1\!/_{T^2})$ [Nes83].

II.4 Simple generalizations with similar triangles

In §II.3, we have demonstrated that Nesterov's method is nothing but an approximation of PPM. This view point has not only provided simple derivations of versions of AGM, but also offered clear explanations of the step sizes. In this section, we demonstrate that these interpretations offered by PPM actually lead to a great simplification of

Nesterov's AGM in the form of the *method of similar triangles* [Nes18, GN18] which admits simple generalizations to practically relevant settings (constrained composite costs). To that end, let us first consider the unconstrained case.

Our starting point is the observations made in the previous section: (i) from §II.3.1, we have seen $\|y_t - x_t\| : \|y_t - z_t\| = \eta_t : {}^1\!/\!_L$; (ii) from §II.3.2, we have seen that we need to choose $\eta_t = \Theta(t)$, and hence, $\eta_{t+1} \approx \eta_t \gg 1$. From these observations, one can readily see that the triangle $\triangle x_t x_{t+1} z_t$ is approximately similar to $\triangle y_t z_{t+1} z_t$. Therefore, one can simplify AGM by further exploiting this fact: we modify the updates so that the two triangles are indeed *similar*. There are two different ways one can keep the two triangles similar:

1. Update z_{t+1} as before and but now we update x_{t+1} so that the two triangles are similar.

2. Update x_{t+1} as before and but now we update z_{t+1} so that the two triangles are similar.

We discuss the above two ways one by one.

II.4.1 First similar triangles approximation: momentum form of AGM

We first adopt the first way to keep the two triangles similar. We have the following update.

First similar triangle approximation:

$$y_t = \frac{{}^1\!/\!_L}{{}^1\!/\!_L+\eta_t} x_t + \frac{\eta_t}{{}^1\!/\!_L+\eta_t} z_t \,, \tag{II.12a}$$

$$z_{t+1} = y_t - \tfrac{1}{L}\nabla f(y_t)\,, \tag{II.12b}$$

$$x_{t+1} = z_{t+1} + L\eta_t(z_{t+1} - z_t)\,. \tag{II.12c}$$

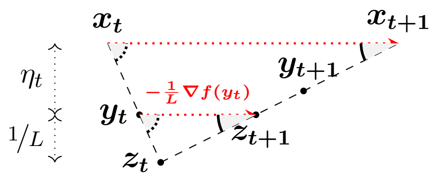


Figure II-3: The updates of (II.12).

In fact, (II.12) can be equivalently expressed without $\{x_t\}$, as illustrated with dots in Figure II-3. More specifically, during the t-th iteration, once we compute (II.12b), one can directly update y_{t+1} via $y_{t+1} = z_{t+1} + \frac{L\eta_t}{L\eta_{t+1}+1}(z_{t+1} - z_t)$. In other words,

$$
\text{(II.12)} \quad \Longleftrightarrow \quad \begin{cases} z_{t+1} = y_t - \frac{1}{L}\nabla f(y_t)\,, \\ y_{t+1} = z_{t+1} + \frac{L\eta_t}{L\eta_{t+1}+1}(z_{t+1} - z_t)\,. \end{cases}
$$

Hence, (II.12) is equal to the well-known momentum form of AGM ("Form I" in the introduction).

Notably, it turns out that our PPM-based analysis suggests the choice of $\{\eta_t\}$ as per the recursive relation $(L\eta_{t+1} + \frac{1}{2})^2 = (L\eta_t + 1)^2 + \frac{1}{4}$, which after substitution $L\eta_t + 1 \leftarrow a_t$ exactly recovers the popular recursive relation $a_{t+1} = \frac{1}{2}(1 + \sqrt{1 + 4a_t^2})$ in [Nes83, BT09]. Below we share the details. The analysis of (II.12) is analogous to the analysis presented in §II.3.2. It turns out that one can derive the following counterparts of (IV.8) and (IV.9); see §II.5.2 for derivations.

$$
\widetilde{\eta}_{t+1}[f(z_{t+1}) - f(x_*)] + \frac{1}{2}\|x_* - x_{t+1}\|^2 - \frac{1}{2}\|x_* - x_t\|^2 \leq (\mathcal{E}_1^{\mathsf{SIM}})\,, \qquad (\mathsf{Ineq}_1^{\mathsf{SIM}})
$$

$$
f(z_{t+1}) - f(z_t) \leq (\mathcal{E}_2^{\mathsf{SIM}})\,, \qquad (\mathsf{Ineq}_2^{\mathsf{SIM}})
$$

where $(\mathcal{E}_1^{\mathsf{SIM}}) := \frac{1}{2}\left(-(L\eta_t + 1)^2 + L\widetilde{\eta}_{t+1}\right) \cdot \|z_{t+1} - y_t\|^2 + \widetilde{\eta}_{t+1} \cdot \langle \nabla f(y_t), z_{t+1} - x_{t+1}\rangle$ and $(\mathcal{E}_2^{\mathsf{SIM}}) := \frac{L}{2}\|z_{t+1} - y_t\|^2 + \langle \nabla f(y_t), z_{t+1} - z_t\rangle$

Having established counterparts of (IV.8) and (IV.9), following §II.3.2, we choose

$$
\Phi_t := \left(\textstyle\sum_{i=1}^t \widetilde{\eta}_i\right) \cdot (f(z_t) - f(x_*)) + \frac{1}{2}\|x_* - x_t\|^2\,. \tag{II.13}
$$

To prove the validity of the chosen Lyapunov function, it suffices to verify

$$
(\mathcal{E}_1^{\mathsf{SIM}}) + \left(\textstyle\sum_{i=1}^t \widetilde{\eta}_i\right) \cdot (\mathcal{E}_2^{\mathsf{SIM}}) \leq 0 \tag{II.14}
$$

which is equivalent to showing (because $z_{t+1} - x_{t+1} = -L\eta_t(z_{t+1} - z_t)$):

$$
\begin{array}{l} \frac{1}{2}\left(-(L\eta_t + 1)^2 + \sum_{i=1}^{t+1} L\widetilde{\eta}_i\right) \cdot \|z_{t+1} - y_t\|^2 \\ + \left(L\eta_t\widetilde{\eta}_{t+1} - \sum_{i=1}^t \widetilde{\eta}_i\right) \langle \nabla f(y_t), z_{t+1} - z_t\rangle \end{array} \leq 0\,. \tag{II.15}
$$

From (II.15), it suffices to choose $\{\eta_t\}$ so that $L\eta_t\widetilde{\eta}_{t+1} = \sum_{i=1}^{t}\widetilde{\eta}_i$. Indeed, with such a choice, the coefficient of the inner product term in (II.11) becomes zero and the coefficient of the squared norm term becomes

$$\begin{aligned}\tfrac{1}{2}\left(-(L\eta_t+1)^2 + \textstyle\sum_{i=1}^{t+1} L\widetilde{\eta}_i\right) &= \frac{1}{2}\left(-(L\eta_t+1)^2 + L\widetilde{\eta}_{t+1} + L\widetilde{\eta}_{t+1}\cdot L\eta_t\right)\\ &= \tfrac{1}{2}\left(-(L\eta_t+1)^2 + L\widetilde{\eta}_{t+1}(L\eta_t+1)\right) = 0\end{aligned}$$

since $L\widetilde{\eta}_{t+1} = L\eta_t + 1$. Indeed, one can actually simplify the relation $L\eta_t\widetilde{\eta}_{t+1} = \sum_{i=1}^{t}\widetilde{\eta}_i$:

$$L\eta_{t+1}\cdot(L\eta_{t+1}+1) = L\eta_{t+1}\cdot L\widetilde{\eta}_{t+2} = \textstyle\sum_{i=1}^{t+1} L\widetilde{\eta}_i = L\widetilde{\eta}_{t+1} + L\eta_t\cdot L\widetilde{\eta}_{t+1} = (L\eta_t+1)^2\,.$$

After rearranging, we obtain the recursive relation: $(L\eta_{t+1}+\frac{1}{2})^2 = (L\eta_t+1)^2 + \frac{1}{4}$, which after the substitution $L\eta_t + 1 = a_t$ exactly recovers the popular recursive relation $a_{t+1} = \frac{1+\sqrt{1+4a_t^2}}{2}$ in [Nes83, BT09].

II.4.2 Second similar triangles approximation: acceleration for composite costs

We now adopt the second way to keep the two triangles similar. We have the following update.

Second similar triangle approximation:

$$y_t = \tfrac{1/L}{1/L+\eta_t}x_t + \tfrac{\eta_t}{1/L+\eta_t}z_t\,, \qquad \text{(II.16a)}$$

$$x_{t+1} = x_t - \eta_{t+1}\nabla f(y_t)\,, \qquad \text{(II.16b)}$$

$$z_{t+1} = \tfrac{1/L}{1/L+\eta_t}x_{t+1} + \tfrac{\eta_t}{1/L+\eta_t}z_t\,. \qquad \text{(II.16c)}$$

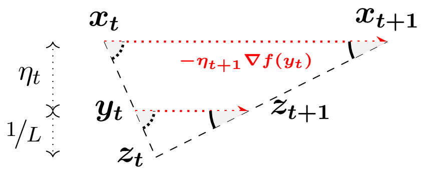


Figure II-4: Illustration of (II.16).

This is "Form III" in the introduction. Below, we provide a PPM-based analysis for a more general setting.

The main advantage of this similar triangles approximation (II.16) becomes clearer in the constraint optimization case: when there is a constraint set, the steps (II.9b) and (II.9c) both become projections steps which could be costly when the constraint set does not admit simple projections. On the other hand, since (II.16) only requires a single projection in each iteration, it minimizes such costly computations.

It turns out (II.16) also admits a simple extension to the practically relevant setting of constrained optimization on composite costs (see e.g. [Nes18, §6.1.3]). More specifically, for a closed convex set $Q \subseteq \mathbb{R}^d$ and a closed[1] convex function $\Psi : Q \to \mathbb{R}$, consider

$$\min_{x \in Q} f^{\Psi}(x) := f(x) + \Psi(x)\,,$$

where $f : Q \to \mathbb{R}$ is a differentiable convex function which is L-smooth with respect to a norm $\|\cdot\|$ that is not necessarily the ℓ_2 norm (i.e., we regard the norm in Definition II.1 to be our chosen norm). For the general norm case, we use the Bregman divergence for the regularization:

Definition II.2. Given a 1-strongly convex (w.r.t the chosen norm $\|\cdot\|$) function $h : Q \to \mathbb{R} \cup \{\infty\}$ that is differentiable on the interior of Q, $D_h(u, v) := h(u) - h(v) - \langle \nabla h(v), u - v\rangle$ for all $u, v \in Q$.

Under the above setting and assumption, (II.16) admits a simple generalization:

Generalization of (II.16) to composite costs:

$$y_t = \tfrac{1/L}{1/L+\eta_t} x_t + \tfrac{\eta_t}{1/L+\eta_t} z_t\,, \tag{II.17a}$$

$$x_{t+1} \leftarrow \operatorname{argmin}_{x \in Q} \left\{ \mathsf{LOWER}(x; y_t) + \tfrac{1}{\eta_{t+1}} D_h(x, x_t) + \Psi(x) \right\}\,, \tag{II.17b}$$

$$z_{t+1} = \tfrac{1/L}{1/L+\eta_t} x_{t+1} + \tfrac{\eta_t}{1/L+\eta_t} z_t\,. \tag{II.17c}$$

Again, the similar triangle approximation (II.17) is computationally advantageous in that it only requires a single projection in each iteration. Now we provide a simple PPM-based analysis of (II.17):

PPM-based analysis of (II.17). To obtain counterparts of (IV.8) and (IV.9), we now use a generalization of Proposition II.1 to the Bregman divergence ([Teb18,

[1] This means that the epigraph of the function is closed. See [Nes18, Definition 3.1.2].

Lemma 3.1]). With such a generalization, we obtain the following inequality for $\phi^{\Psi}(x) := \eta_{t+1}[f(y_t) + \langle \nabla f(y_t), x - y_t\rangle + \Psi(x)]$:

$$\phi^{\Psi}(x_{t+1}) - \phi^{\Psi}(x_*) + D_h(x_*, x_{t+1}) + D_h(x_{t+1}, x_t) - D_h(x_*, x_t) \le 0\,, \tag{II.18}$$

where $x_* \in \operatorname{argmin}_{x\in Q} f^{\Psi}(x)$. Now using (II.18), one can derive from first principles the following inequalities (we defer the derivations to §II.5.3):

$$\eta_{t+1}(f^{\Psi}(z_{t+1}) - f^{\Psi}(x_*)) + D_h(x_*, x_{t+1}) - D_h(x_*, x_t) \le (\mathcal{E}_1^{\mathsf{SIM}'})\,, \tag{$\mathsf{Ineq}_1^{\mathsf{SIM}'}$}$$

$$f^{\Psi}(z_{t+1}) - f^{\Psi}(z_t) \le (\mathcal{E}_2^{\mathsf{SIM}'})\,. \tag{$\mathsf{Ineq}_2^{\mathsf{SIM}'}$}$$

where $(\mathcal{E}_1^{\mathsf{SIM}'}) := -\frac{1}{2}\|x_{t+1} - x_t\|^2 + \eta_{t+1}[\frac{L}{2}\|z_{t+1} - y_t\|^2 + \langle \nabla f(y_t), z_{t+1} - x_{t+1}\rangle + \Psi(z_{t+1}) - \Psi(x_{t+1})]$ and $(\mathcal{E}_2^{\mathsf{SIM}'}) := \frac{L}{2}\|z_{t+1} - y_t\|^2 + \langle \nabla f(y_t), z_{t+1} - z_t\rangle + \Psi(z_{t+1}) - \Psi(z_t)$. Similarly to §II.3.2, yet replacing the norm squared term with the Bregman divergence, we choose

$$\Phi_t := (\textstyle\sum_{i=1}^t \eta_i) \cdot (f^{\Psi}(z_t) - f^{\Psi}(x_*)) + D_h(x_*, x_t)\,.$$

Then, it suffices to show $(\mathcal{E}_1^{\mathsf{SIM}'}) + (\sum_{i=1}^t \eta_i) \cdot (\mathcal{E}_2^{\mathsf{SIM}'}) \le 0$. Using the facts (i) $z_{t+1} - x_{t+1} = L\eta_t(z_t - z_{t+1})$ and (ii) $\|x_{t+1} - x_t\| = (L\eta_t + 1)\|z_{t+1} - y_t\|$ (both are immediate consequences of the similar triangles) and rearranging, one can easily check that $(\mathcal{E}_1^{\mathsf{SIM}'}) + (\sum_{i=1}^t \eta_i) \cdot (\mathcal{E}_2^{\mathsf{SIM}'})$ is equal to

$$\tfrac{1}{2}\left(-(L\eta_t + 1)^2 + L\eta_{t+1} + L\textstyle\sum_{i=1}^t \eta_i\right)\|z_{t+1} - y_t\|^2 \tag{II.19}$$

$$+\left(L\eta_t\eta_{t+1} - \textstyle\sum_{i=1}^t \eta_i\right)\langle \nabla f(y_t), z_t - z_{t+1}\rangle \tag{II.20}$$

$$+\eta_{t+1}[\Psi(z_{t+1}) - \Psi(x_{t+1})] + \left(\textstyle\sum_{i=1}^t \eta_i\right) \cdot [\Psi(z_{t+1}) - \Psi(z_t)]. \tag{II.21}$$

Now choosing $\eta_t = t/2L$ analogously to §II.3.2, one can easily verify (II.19) + (II.20) + (II.21) ≤ 0. Indeed, for (II.19), since $L\eta_t\eta_{t+1} = \sum_{i=1}^t \eta_i$, the coefficient becomes $1/2(L\eta_t + 1)(L\eta_{t+1} - L\eta_t - 1)$ which is a negative number since $L\eta_{t+1} - L\eta_t - 1 = -1/2$; for (II.20), the coefficient becomes zero due to the relation $L\eta_t\eta_{t+1} = \sum_{i=1}^t \eta_i$; lastly, for (II.21), we have

$$(\text{II.21}) = \eta_{t+1}\left[(1+L\eta_t)\Psi(z_{t+1}) - \Psi(x_{t+1}) - L\eta_t\Psi(z_t)\right] \leq 0\,, \tag{II.22}$$

where the equality is due to the relation $L\eta_t\eta_{t+1} = \sum_{i=1}^{t}\eta_i$, and the inequality is due to the update (II.17c) (which can be equivalently written as $(1+L\eta_t)z_{t+1} = x_{t+1} + L\eta_t z_t$) and the convexity of Ψ. Hence, we obtain the accelerated rate of $f^\Psi(z_T) - f^\Psi(x_*) \leq \frac{4LD_h(x_*,x_0)}{T(T+1)} = O(1/T^2)$. □

II.5 Deferred derivations

II.5.1 Deferred derivations from §II.3.2

Let us first derive ($\mathsf{Ineq}_1^{\mathsf{AGM}}$). Applying Proposition II.1 with $\phi(x) = \eta_{t+1}[f(y_t) + \langle \nabla f(y_t), x - y_t\rangle]$ to (II.8a), we obtain:

$$\phi(x_{t+1}) - \phi(x_*) + \frac{1}{2}\left\|x_* - x_{t+1}\right\|^2 + \frac{1}{2}\left\|x_{t+1} - x_t\right\|^2 - \frac{1}{2}\left\|x_* - x_t\right\|^2 \leq 0\,. \tag{II.23}$$

Now from the convexity of f, it holds that $\phi(x_*) \leq \eta_{t+1} f(x_*)$. This together with the L-smoothness of f, it follows that $\phi(x_{t+1}) = \eta_{t+1}[f(y_t) + \langle \nabla f(y_t), z_{t+1} - y_t\rangle + \langle \nabla f(y_t), x_{t+1} - z_{t+1}\rangle] \geq \eta_{t+1}\left[f(z_{t+1}) - \frac{L}{2}\left\|z_{t+1} - y_t\right\|^2 + \langle \nabla f(y_t), x_{t+1} - z_{t+1}\rangle\right]$. Plugging these inequalities back to (IV.12) and rearranging, we obtain the following inequality:

$$\begin{aligned}
&\eta_{t+1}[f(z_{t+1}) - f(x_*)] + \frac{1}{2}\left\|x_* - x_{t+1}\right\|^2 - \frac{1}{2}\left\|x_* - x_t\right\|^2 \\
&\leq -\frac{1}{2}\left\|x_{t+1} - x_t\right\|^2 + \eta_{t+1}\left[\frac{L}{2}\left\|z_{t+1} - y_t\right\|^2 + \langle \nabla f(y_t), z_{t+1} - x_{t+1}\rangle\right]\,.
\end{aligned} \tag{II.24}$$

Now decomposing the inner product term in (II.24) into

$$\eta_{t+1}\left\langle \nabla f(y_t), z_{t+1} - y_t\right\rangle + \eta_{t+1}\left\langle \nabla f(y_t), y_t - x_t\right\rangle + \eta_{t+1}\left\langle \nabla f(y_t), x_t - x_{t+1}\right\rangle,$$

and using $x_{t+1} - x_t = -\eta_{t+1}\nabla f(y_t)$ and $z_{t+1} - y_t = -1/L\nabla f(y_t)$ (which are (II.9b) and (II.9c), respectively), (II.24) becomes $\left(\frac{\eta_{t+1}^2}{2} - \frac{\eta_{t+1}}{2L}\right)\left\|\nabla f(y_t)\right\|^2 + \eta_{t+1}\left\langle \nabla f(y_t), y_t - x_t\right\rangle$.

Now, using the relation $y_t - x_t = L\eta_t(z_t - y_t)$ (which is (II.9a)), we obtain ($\mathcal{E}_1^{\mathsf{AGM}}$). Thus, ($\mathsf{Ineq}_1^{\mathsf{AGM}}$) follows.

Next, ($\mathsf{Ineq}_2^{\mathsf{AGM}}$) readily follows from the L-smoothness and the convexity of f:

$$\begin{aligned} f(z_{t+1}) - f(z_t) &= f(z_{t+1}) - f(y_t) + f(y_t) - f(z_t) \\ &\leq \langle \nabla f(y_t), z_{t+1} - y_t \rangle + \frac{L}{2}\left\| z_{t+1} - y_t \right\|^2 + \langle \nabla f(y_t), y_t - z_t \rangle \\ &\overset{(a)}{=} -\frac{1}{2L}\left\| \nabla f(y_t) \right\|^2 + \langle \nabla f(y_t), y_t - z_t \rangle = (\mathcal{E}_2^{\mathsf{AGM}}), \end{aligned}$$

where (a) is due to $z_{t+1} - y_t = -{}^1\!/_L \nabla f(y_t)$.

II.5.2 Deferred derivations from §II.4.1

We first derive ($\mathsf{Ineq}_1^{\mathsf{SIM}}$). By the updates (II.12), we have $x_{t+1} = x_t - (\eta_t + \frac{1}{L})\nabla f(y_t)$. Letting $\widetilde{\eta}_{t+1} := \eta_t + \frac{1}{L}$, this relation can be equivalently written as:

$$x_{t+1} \leftarrow \operatorname{argmin}_x \left\{ f(y_t) + \langle \nabla f(y_t), x - y_t \rangle + \tfrac{1}{2\widetilde{\eta}_{t+1}} \left\| x - x_t \right\|^2 \right\} \qquad \text{(II.25)}$$

The rest is similar to §II.5.1: we apply Proposition II.1 with $\phi(x) = \widetilde{\eta}_{t+1}[f(y_t) + \langle \nabla f(y_t), x - y_t \rangle]$:

$$\phi(x_{t+1}) - \phi(x_*) + \frac{1}{2}\left\| x_* - x_{t+1} \right\|^2 + \frac{1}{2}\left\| x_{t+1} - x_t \right\|^2 - \frac{1}{2}\left\| x_* - x_t \right\|^2 \leq 0\,. \qquad \text{(II.26)}$$

Now from the convexity, we have $\phi(x_*) \leq \widetilde{\eta}_{t+1} f(x_*)$, and from the L-smoothness, we have

$$\begin{aligned} \phi(x_{t+1}) &= \widetilde{\eta}_{t+1}[f(y_t) + \langle \nabla f(y_t), z_{t+1} - y_t \rangle + \langle \nabla f(y_t), x_{t+1} - z_{t+1} \rangle] \\ &\geq \widetilde{\eta}_{t+1}\left[f(z_{t+1}) - \frac{L}{2}\left\| z_{t+1} - y_t \right\|^2 + \langle \nabla f(y_t), x_{t+1} - z_{t+1} \rangle \right]. \end{aligned}$$

Plugging these inequalities back to (II.26) and rearranging, we obtain the following inequality:

$$\begin{aligned}
&\widetilde{\eta}_{t+1}[f(z_{t+1}) - f(x_*)] + \frac{1}{2}\left\|x_* - x_{t+1}\right\|^2 - \frac{1}{2}\left\|x_* - x_t\right\|^2 \\
&\leq -\frac{1}{2}\left\|x_{t+1} - x_t\right\|^2 + \widetilde{\eta}_{t+1}\left[\frac{L}{2}\left\|z_{t+1} - y_t\right\|^2 + \langle \nabla f(y_t), z_{t+1} - x_{t+1}\rangle\right] \\
&= \frac{1}{2}\left(-(L\eta_t + 1)^2 + L\widetilde{\eta}_{t+1}\right) \cdot \left\|z_{t+1} - y_t\right\|^2 + \widetilde{\eta}_{t+1} \cdot \langle \nabla f(y_t), z_{t+1} - x_{t+1}\rangle = (\mathcal{E}_1^{\mathsf{SIM}}),
\end{aligned}$$

where the last line follows since $\|x_{t+1} - x_t\| = (L\eta_t + 1) \cdot \|z_{t+1} - z_t\|$ (see Figure II-3).

Next we derive ($\mathsf{Ineq}_1^{\mathsf{SIM}}$). From the L-smoothness and the convexity of f:

$$\begin{aligned}
f(z_{t+1}) - f(z_t) &= f(z_{t+1}) - f(y_t) + f(y_t) - f(z_t) \\
&\leq \langle \nabla f(y_t), z_{t+1} - y_t\rangle + \frac{L}{2}\left\|z_{t+1} - y_t\right\|^2 + \langle \nabla f(y_t), y_t - z_t\rangle \\
&= \frac{L}{2}\left\|z_{t+1} - y_t\right\|^2 + \langle \nabla f(y_t), z_{t+1} - z_t\rangle = (\mathcal{E}_2^{\mathsf{SIM}}).
\end{aligned}$$

II.5.3 Deferred derviations from §II.4.2

Let us first derive ($\mathsf{Ineq}_1^{\mathsf{SIM}'}$). From convexity, we have $\phi^\Psi(x_*) \leq \eta_{t+1} f^\Psi(x_*)$, and from the L-smoothness, we have the following lower bound:

$$\begin{aligned}
&\phi^\Psi(x_{t+1}) = \eta_{t+1}[f(y_t) + \langle \nabla f(y_t), z_{t+1} - y_t\rangle + \langle \nabla f(y_t), x_{t+1} - z_{t+1}\rangle + \Psi(x_{t+1})] \\
&\geq \eta_{t+1}\left[f^\Psi(z_{t+1}) - \frac{L}{2}\left\|z_{t+1} - y_t\right\|^2 + \langle \nabla f(y_t), x_{t+1} - z_{t+1}\rangle + \Psi(x_{t+1}) - \Psi(z_{t+1})\right].
\end{aligned}$$

Plugging these back to (II.18), and using the bound $-D_h(x_{t+1}, x_t) \leq -\frac{1}{2}\left\|x_{t+1} - x_t\right\|^2$, ($\mathsf{Ineq}_1^{\mathsf{SIM}'}$) follows.

Next, to derive ($\mathsf{Ineq}_2^{\mathsf{SIM}'}$), we use L-smoothness and the convexity of f to obtain the following:

$$\begin{aligned}
&f^\Psi(z_{t+1}) - f^\Psi(z_t) \leq f(z_{t+1}) - f(y_t) + f(y_t) - f(z_t) + \Psi(z_{t+1}) - \Psi(z_t) \\
&\leq \frac{L}{2}\left\|z_{t+1} - y_t\right\|^2 + \langle \nabla f(y_t), z_{t+1} - z_t\rangle + \Psi(z_{t+1}) - \Psi(z_t),
\end{aligned}$$

which is precisely equal to ($\mathcal{E}_2^{\mathsf{SIM}'}$).

II.6 Related work for Chapter II

Our approach is inspired by that of Defazio [Def19] that establishes an inspiring connection between AGM and PPM. The main observation in that paper is that for strongly convex costs, one can derive a version of AGM from the primal-dual form of PPM with a tweak of geometry. Compared with [Def19], our approach strengthens the connection between AGM and PPM by considering more versions of AGM and their analyses. Another advantage of our approach is that it does not require duality.

We now summarize previous works on developing alternative approaches to Nesterov's acceleration. Most works have studied the continuous limit dynamics of Nesterov's AGM [SBC16, KBB15, WWJ16]. These continuous dynamics approaches have brought about new intuitions about Nesterov's acceleration, and follow-up works have developed analytical techniques for such dynamics [WRJ16, DO19]. Another notable contribution is made based on the linear coupling framework [AZO17]. The main observation is that the two most popular first-order methods, namely gradient descent and mirror descent, have complementary performances, and hence, one can come up with a faster method by linearly coupling the two methods.

PPM has been used to design or interpret other optimization methods [Dru17]. To list few instances, PPM has given rise to fast methods for weakly convex problems [DG19], the prox-linear methods for composite optimizations [BF95, Nes07, LW16], accelerated methods for stochastic optimizations [LMH15], and methods for saddle-point problems [MOP20].

Chapter III

Extension to strongly convex costs

In this chapter, we extend our PPM framework from Chapter II to the case of strongly convex costs. As we shall see, our framework gives rise to a simple derivation of the most general version of AGM due to Nesterov called "*General Scheme for Optimal Method*" [Nes18, (2.2.7)]. We will also extend our PPM-based analysis to strongly convex costs. The derivation and analysis in this chapter will be used as a building block for obtaining accelerated methods over Riemannian manifolds.

III.1 Derivation based on proximal point method

We first make the approximate PPM (II.8) more flexible by considering two separate step sizes.

Approximate PPM with two separate step sizes $\{\eta_t\}$ and $\{\widetilde{\eta}_t\}$. Given $x_0 = y_0 \in \mathbb{R}^d$,

$$x_{t+1} \leftarrow \mathrm{argmin}_x \left\{\mathsf{LOWER}(x; y_t) + \tfrac{1}{2\eta_{t+1}} \left\|x - x_t\right\|^2\right\}, \tag{III.1a}$$

$$y_{t+1} \leftarrow \mathrm{argmin}_x \left\{\mathsf{UPPER}(x; y_t) + \tfrac{1}{2\widetilde{\eta}_{t+1}} \left\|x - x_{t+1}\right\|^2\right\}. \tag{III.1b}$$

Now let us apply our PPM view to the strongly convex cost case.

Definition III.1 (Strong convexity)**.** For $\mu > 0$, we say a differentiable function

$f : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex if $f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle + \frac{\mu}{2} \|x - y\|^2$ for any $x, y \in \mathbb{R}^d$.

Since f is additionally assumed to be strongly convex, one can now strengthen the lower approximation $\mathsf{LOWER}(x; y_t)$ in (III.1a) to $\mathsf{LOWER}(x; y_t) + \frac{\mu}{2} \|x - y_t\|^2$. In other words, we obtain

Approximate PPM for strongly-convex costs. Given $x_0 = y_0 \in \mathbb{R}^d$,

$$x_{t+1} \leftarrow \operatorname*{argmin}_{x \in \mathbb{R}^d} \left\{ \mathsf{LOWER}(x; y_t) + \underbrace{\frac{\mu}{2} \|x - y_t\|^2}_{\substack{\text{additional term due to} \\ \text{strong convexity}}} + \frac{1}{2\eta_{t+1}} \|x - x_t\|^2 \right\}, \quad \text{(III.2a)}$$

$$y_{t+1} \leftarrow \operatorname{argmin}_x \left\{ \mathsf{UPPER}(x; y_t) + \tfrac{1}{2\widetilde{\eta}_{t+1}} \|x - x_{t+1}\|^2 \right\}. \quad \text{(III.2b)}$$

Writing the optimality condition of (III.2), it is straightforward to check that the approximate PPM (III.1) is equivalent to the following updates ($x_0 = y_0 = z_0$):

Equivalent representation of (III.2):

$$y_t = \tfrac{1/L}{1/L + \widetilde{\eta}_t} x_t + \tfrac{\widetilde{\eta}_t}{1/L + \widetilde{\eta}_t} z_t \,, \quad \text{(III.3a)}$$

$$\begin{aligned} x_{t+1} = {} & \tfrac{1/\mu}{1/\mu + \eta_{t+1}} x_t + \tfrac{\eta_{t+1}}{1/\mu + \eta_{t+1}} y_t \\ & - \tfrac{1/\mu \cdot \eta_{t+1}}{1/\mu + \eta_{t+1}} \nabla f(y_t) \end{aligned} \,, \quad \text{(III.3b)}$$

$$z_{t+1} = y_t - \tfrac{1}{L} \nabla f(y_t) \,. \quad \text{(III.3c)}$$

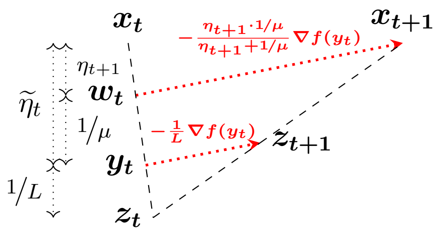

x_t
$-\frac{\eta_{t+1} \cdot 1/\mu}{\eta_{t+1} + 1/\mu} \nabla f(y_t)$
x_{t+1}
η_{t+1}
$\widetilde{\eta}_t$
w_t
$1/\mu$
$-\frac{1}{L} \nabla f(y_t)$
z_{t+1}
y_t
$1/L$
z_t

Figure III-1: Illustration of (III.3).

Note that (III.3) is nothing but "Form IV" in the introduction. Again, our derivation provides new insights into the choices of the AGM step sizes by expressing them in terms of the PPM step sizes η_t's and $\widetilde{\eta}_t$'s.

III.1.1 Relation to well known momentum version

Perhaps, the most well known version of AGM for strongly convex costs is the momentum version due to Nesterov (see, e.g., [Nes18, (2.2.22)])

$$\begin{aligned} z_{t+1} &= y_t - \tfrac{1}{L}\nabla f(y_t)\,, \\ y_{t+1} &= z_{t+1} + \tfrac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}(z_{t+1} - z_t)\,. \end{aligned} \tag{III.4}$$

One might wonder whether one can better understand the step sizes in (III.4) from (III.3).

Let us first recall the well known convergence rate of PPM for strongly convex costs due to Rockafellar [Roc76, (1.14)]:

$$f(x_T) - f(x_*) \le O\left(\textstyle\prod_{t=1}^{T}(1+\mu\eta_t)^{-1}\right) \quad \text{for any } T \ge 1. \tag{III.5}$$

From (III.5), one can see that in order to achieve the accelerated convergence rate $O(\exp(^{-T}/_{\sqrt{\kappa}}))$ where κ is the condition number $^{L}/_{\mu}$, the step sizes η_t must be chosen so that $\eta_t \approx \mu^{-1}(\sqrt{\kappa})^{-1}$. Having said that, the well known version (III.4) corresponds to choosing the following step sizes for (III.3):

$$\eta_t \equiv \eta := \mu^{-1}(\sqrt{\kappa}-1)^{-1} \quad \text{and} \quad \widetilde{\eta}_t \equiv \widetilde{\eta} := \mu^{-1}(\sqrt{\kappa})^{-1}. \tag{III.6}$$

In the next section, we will recover the choice based on PPM-based analysis. Now with such choice of η and $\widetilde{\eta}$, (III.3) becomes:

(III.3) with step sizes chosen as (III.6):

$$y_t = \tfrac{1}{1+\sqrt{\kappa}}x_t + \tfrac{\kappa}{1+\sqrt{\kappa}}z_t\,, \tag{III.7a}$$

$$x_{t+1} = \tfrac{\sqrt{\kappa}-1}{\sqrt{\kappa}}x_t + \tfrac{1}{\sqrt{\kappa}}y_t - \tfrac{\sqrt{\kappa}}{L}\nabla f(y_t)\,, \tag{III.7b}$$

$$z_{t+1} = y_t - \tfrac{1}{L}\nabla f(y_t)\,. \tag{III.7c}$$

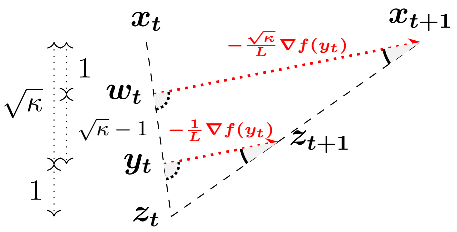


Figure III-2: Illustration of (III.7).

In fact, with the step size choices (III.6), $\triangle w_t x_{t+1} z_t$ is similar to $\triangle y_t z_{t+1} z_t$, and hence the updates (III.7) can be equivalently written without $\{x_t\}$ and $\{w_t\}$:

$$\text{(III.7)} \quad \Longleftrightarrow \quad \text{(III.4)} = \begin{cases} z_{t+1} = y_t - \frac{1}{L}\nabla f(y_t)\,, \\ y_{t+1} = z_{t+1} + \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}(z_{t+1} - z_t)\,. \end{cases}$$

III.2 Analysis based on PPM perspective: simplified version

In Chapter II, we have used the Lyapunov function $\Phi_t := (\sum_{i=1}^t \eta_i) \cdot (f(z_t) - f(x_*)) + \frac{1}{2}\left\|x_* - x_t\right\|^2$ to analyze the non-strongly convex case. For the strongly convex setting, we are shooting for a linear convergence rate (III.5), so we consider the following Lyapunov function:

$$\Phi_t := \prod_{i=1}^{t}(1-\xi_t)^{-1} \cdot \left[f(z_t) - f(x_*) + b_t \cdot \left\|x_t - x_*\right\|^2\right]\,, \tag{III.8}$$

for some $\xi_t \in (0,1)$ and $b_t \geq 0$. Note that the previous Lyapunov function is a special case above (III.8). If one can show Φ_t is decreasing, then we get a convergence rate of $O(\prod_{i=1}^t (1-\xi_t))$.

In this section, as a warm-up, we consider the simplified case where $\xi_t \equiv \xi$ and $b_t \equiv b$, i.e.,

$$\boxed{\Phi_t := (1-\xi)^{-t} \cdot \left[f(z_t) - f(x_*) + b \cdot \left\|x_t - x_*\right\|^2\right]\,.} \tag{III.9}$$

Using this simplified Lyapunov function, we analyze the following constant step sizes version of (III.3). In particular, we rewrite the weights in the updates using simpler notations as below.

(III.3) **with constant step size:** (**$\eta_t \equiv \eta$ and $\widetilde{\eta}_t \equiv \widetilde{\eta}$**):

$$y_t = \tfrac{1/L}{1/L+\widetilde{\eta}} x_t + \tfrac{\widetilde{\eta}}{1/L+\widetilde{\eta}} z_t \qquad =: z_t + \alpha(x_t - z_t)\,, \tag{III.10a}$$

$$x_{t+1} = \tfrac{1/\mu}{1/\mu+\eta} x_t + \tfrac{\eta}{1/\mu+\eta} y_t - \tfrac{1/\mu\cdot\eta}{1/\mu+\eta} \nabla f(y_t) \quad =: y_t + \beta(x_t - y_t) - \tfrac{(1-\beta)}{\mu} \nabla f(y_t)\,, \tag{III.10b}$$

$$z_{t+1} = y_t - \tfrac{1}{L} \nabla f(y_t) \qquad =: y_t - c \nabla f(y_t)\,. \tag{III.10c}$$

Here $\alpha := \frac{1/L}{1/L+\widetilde{\eta}}$, $\beta = \frac{1/\mu}{1/\mu+\eta}$. For later application, we will in general consider $c \in (0, 2/L)$.

Hence, the main goal now is to choose the parameters ξ, b, η, $\widetilde{\eta}$ so that $\Phi_{t+1} - \Phi_t \leq 0$, or equivalently,

$$f(z_{t+1}) - f(x_*) + b \cdot \|x_{t+1} - x_*\|^2 - (1 - \xi) \cdot \left[f(z_t) - f(x_*) + b \cdot \|x_t - x_*\|^2 \right] \tag{III.11}$$

is less than equal to zero. To that end, we first express (III.11) more simply.

III.2.1 Upper bounding the Lyapunov difference

We first express the terms $f(z_{t+1}) - f(x_*)$ and $\|x_{t+1} - x_*\|^2$ in terms of the previous iterates.

- We know from L-smoothness of f that $f(z_{t+1}) - f(y_t) \leq \langle \nabla f(y_t), z_{t+1} - y_t \rangle + \frac{L}{2} \|z_{t+1} - y_t\|^2 = -c(1 - Lc/2) \|\nabla f(y_t)\|^2$.
- By (III.10b), we have $\|x_{t+1} - x_*\|^2 = \left\| y_t - x_* + \beta(x_t - y_t) - \frac{(1-\beta)}{\mu} \nabla f(y_t) - x_* \right\|^2$.

Given these expressions, we streamline our notations as follows.

Notations: we denote the recurring vectors as follows:

$$\Delta_c := c(1 - Lc/2)\,, \quad \nabla := \nabla f(y_t)\,, \quad V := y_t - x_*\,, \text{ and } W := x_t - y_t.$$

Since $c \in (0, 2/L)$, Δ_c is a positive constant.

With these notations, one can rewrite the previous (in)equalities succintly as follows:

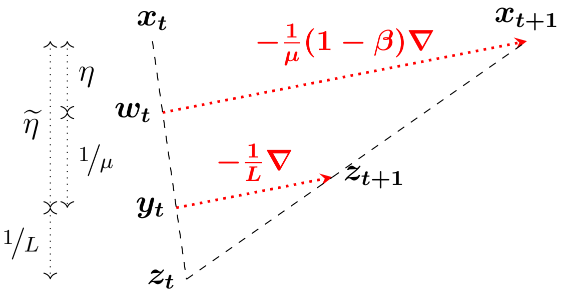
$$\begin{aligned} f(z_{t+1}) - f(y_t) &\le -\Delta_c \left\|\nabla\right\|^2 , \\ \left\|x_{t+1} - x_*\right\|^2 &= \left\|V + \beta W - \frac{1-\beta}{\mu}\nabla\right\|^2 . \end{aligned}$$

Now let us plug these two back to (III.11). Let us first express $b \cdot \left\|x_{t+1} - x_*\right\|^2 - (1-\xi)b \cdot \left\|x_t - x_*\right\|^2$:

$$b \cdot \left\|x_{t+1} - x_*\right\|^2 - (1-\xi)b \cdot \left\|x_t - x_*\right\|^2 = b \cdot \left\|V + \beta W - \frac{1-\beta}{\mu}\nabla\right\|^2 - (1-\xi)b \cdot \left\|V + W\right\|^2$$

$$= \left\{ \begin{aligned} &(b - (1-\xi)b) \cdot \left\|V\right\|^2 + (\beta^2 b - (1-\xi)b) \cdot \left\|W\right\|^2 + \tfrac{(1-\beta)^2}{\mu^2} b \cdot \left\|\nabla\right\|^2 \\ &+2(\beta b - (1-\xi)b) \cdot \langle V, W\rangle - 2\beta\tfrac{1-\beta}{\mu} b \left\langle W, \nabla\right\rangle - 2\tfrac{1-\beta}{\mu} b \cdot \left\langle V, \nabla\right\rangle . \end{aligned} \right. \tag{III.12}$$

Next we express $f(z_{t+1}) - f(x_*) - (1-\xi) \cdot (f(z_t) - f(x_*))$ in terms of ∇, V, W. Below, we recall Figure III-1 with the simplified notations for readers' convenience:



$$\alpha := \frac{1/L}{1/L + \widetilde{\eta}}$$
$$\beta := \frac{1/\mu}{1/\mu + \eta}$$

Using the inequality $f(z_{t+1}) - f(y_t) \le -\Delta_c \left\|\nabla\right\|^2$, we obtain

$$\begin{aligned} &f(z_{t+1}) - f(x_*) - (1-\xi) \cdot (f(z_t) - f(x_*)) \le f(y_t) - f(x_*) - \Delta_c \cdot \left\|\nabla\right\|^2 - (1-\xi) \cdot (f(z_t) - f(x_*)) \\ &\quad = (1-\xi) \cdot (f(y_t) - f(z_t)) + \xi \cdot (f(y_t) - f(x_*)) - \Delta_c \cdot \left\|\nabla\right\|^2 . \\ &\quad \le (1-\xi) \cdot \langle \nabla, y_t - z_t \rangle + \xi \cdot \langle \nabla, V\rangle - \frac{\mu}{2}(1-\xi) \cdot \left\|y_t - z_t\right\|^2 - \frac{\mu}{2}\xi \cdot \left\|V\right\|^2 - \Delta_c \xi \cdot \left\|\nabla\right\|^2 , \end{aligned} \tag{III.13}$$

where the last inequality follows from μ-strong convexity of f: $f(u) - f(v) \le$

$\langle \nabla f(u), u-v\rangle - \frac{\mu}{2}\|u-v\|^2$. Now from (III.10a) (or from the above figure), we have $y_t - z_t = \frac{\alpha}{1-\alpha}(x_t - y_t) = \frac{\alpha}{1-\alpha}W$, and hence, one can express (III.13) fully in terms of ∇, V, W:

$$\begin{aligned}(\text{III.13}) = &\frac{\alpha}{1-\alpha}(1-\xi)\cdot\langle\nabla, W\rangle + \xi\cdot\langle\nabla, V\rangle \\ &-\frac{\mu}{2}\left(\frac{\alpha}{1-\alpha}\right)^2(1-\xi)\cdot\|W\|^2 - \frac{\mu}{2}\xi\cdot\|V\|^2 - \Delta_c\cdot\|\nabla\|^2\end{aligned} \tag{III.14}$$

Combining (III.12) and (III.14), we obtain the following upper bound on $\Phi_{t+1} - \Phi_t$ fully in terms of ∇, V, W:

$$\Phi_{t+1} - \Phi_t \leq C_1\cdot\|W\|^2 + C_2\cdot\|V\|^2 + C_3\|\nabla\|^2 + C_4\cdot\langle W, V\rangle + C_5\cdot\langle W,\nabla\rangle + C_6\cdot\langle V,\nabla\rangle\,, \tag{III.15}$$

$$\text{where}\begin{cases} C_1 := \beta^2 b - (1-\xi)b - \frac{\mu}{2}\frac{\alpha^2}{(1-\alpha)^2}(1-\xi)\,, & C_2 := b - (1-\xi)b - \frac{\mu}{2}\xi\,,\\ C_3 := \frac{(1-\beta)^2}{\mu^2}b - \Delta_c\,, & C_4 := 2(\beta b - (1-\xi)b)\,,\\ C_5 := \frac{\alpha}{1-\alpha}(1-\xi) - 2\beta\frac{1-\beta}{\mu}b\,, & C_6 := \xi - 2\frac{1-\beta}{\mu}b\,. \end{cases}$$

III.2.2 Ensuring decrease in Lyapunov function

Having established the bound (III.15), our goal is to now choose the step sizes α and β (equivalently, η and $\widetilde{\eta}$) and the Lyapunov function parameters ξ, b so that (III.15) is non-positive. Following the approach from Chapter II, one avenue is 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, b in terms of ξ:*
  - *$C_4 = 0$ and $C_6 = 0$ characterize β, b in terms of ξ:* From $C_4 = 0$, we get $\beta = 1-\xi$, and from $C_6 = 0$, we get $(1-\beta)b = \frac{\mu}{2}\xi$, which implies $b = \frac{\mu}{2}$. Hence,

$$\beta = 1-\xi \quad (\Leftrightarrow \mu\eta = \frac{\xi}{1-\xi}) \quad\text{and}\quad b = \frac{\mu}{2}.$$

- $C_5 = 0$ *characterizes* α *in terms of* ξ*:* From $C_5 = 0$, we get $\frac{\alpha}{1-\alpha}(1-\xi) = 2\beta\frac{1-\beta}{\mu}b$. Hence,

$$\frac{\alpha}{1-\alpha} = \xi \ \ (\Leftrightarrow L\widetilde{\eta} = \frac{1}{\xi}).$$

▪ *Coefficients of squared terms determines* ξ*:*

- From $C_3 \leq 0$, we get $\frac{(1-\beta)^2}{\mu^2}b \leq \Delta_c$. Now, Using the characterizations of β and b above, one can rewrite the inequality as

$$\xi^2 \leq 2\mu\Delta_c\,.$$

  Since we want to maximize ξ, we conclude that $\xi = \sqrt{2\mu\Delta_c}$.

- With the choices of α, β, b as above, one can easily check that $C_2 = 0$ and $C_1 = \beta^2 b - (1-\xi)b - \frac{\mu}{2}\frac{\alpha^2}{(1-\alpha)^2}(1-\xi) \leq -\frac{\mu}{2}\frac{\alpha^2}{(1-\alpha)^2} \leq 0$.

Summarizing the calculations thus far, we obtain the following result.

Theorem III.1. For $c \in (0, 2/L)$, let $\Delta_c := c(1 - Lc/2)$. Let us choose parameters as follows:

1. Choose $\xi := \sqrt{2\mu\Delta_c}$.
2. Choose step sizes based on ξ: $\eta = \frac{1}{\mu}\frac{\xi}{1-\xi}$ and $\widetilde{\eta} = \frac{1}{L}\frac{1}{\xi}$.

Then, given x_t, z_t, the next iterates x_{t+1}, x_{t+1} computed as per the constant step version (III.10) satisfy

$$f(y_{t+1}) - f(x_*) + \frac{\mu}{2}\cdot\|z_{t+1} - x_*\|^2 \leq (1-\xi)\cdot\left[f(y_t) - f(x_*) + \frac{\mu}{2}\cdot\|z_t - x_*\|^2\right].$$

To conclude, this shows the convergence rate of $O((1-\sqrt{2\mu\Delta_c})^T)$ of the constant step size version (III.10). In particular, if we choose $\alpha = \frac{1}{L}$, we have $\Delta_c = \frac{1}{2L}$ and $\xi = 1/\sqrt{\kappa}$, thereby achieving the well known convergence rate of $O((1 - 1/\sqrt{\kappa})^T)$. Also, we get $\eta = \frac{1}{\mu}\frac{\xi}{1-\xi} = \mu^{-1}(\sqrt{\kappa}-1)^{-1}$ and $\widetilde{\eta} = \frac{1}{L}\frac{1}{\xi} = \mu^{-1}(\sqrt{\kappa})^{-1}$, which is precisely the choice (III.6).

III.3 Analysis based on PPM perspective: general version

Now we move onto analyzing the general version (III.3). For the general version, we use the general Lyapunov function (III.8) as we recall below:

$$\boxed{\Phi_t := \prod_{i=1}^{t}(1-\xi_t)^{-1}\cdot\left[f(z_t)-f(x_*)+b_t\cdot\|x_t-x_*\|^2\right],} \tag{III.8}$$

for some $\xi_t \in (0,1)$ and $b_t \geq 0$. Also, following the previous analysis for the constant step version, we rewrite (III.3) using simpler notations:

Simper representation of (III.3):

$$y_t = z_t + \alpha_t(x_t - z_t)\,, \tag{III.16a}$$

$$x_{t+1} = y_t + \beta_{t+1}(x_t - y_t) - \tfrac{(1-\beta_{t+1})}{\mu}\nabla f(y_t)\,, \tag{III.16b}$$

$$z_{t+1} = y_t - c\nabla f(y_t)\,. \tag{III.16c}$$

Here $\alpha_t := \frac{1/L}{1/L+\widetilde{\eta}_t}$, $\beta_{t+1} = \frac{1/\mu}{1/\mu+\eta_{t+1}}$. For later application, we will in general consider $c \in (0, 2/L)$.

Hence, the main goal now is to choose the parameters ξ_{t+1}, b_{t+1}, α_t, β_{t+1} so that $\Phi_{t+1} - \Phi_t \leq 0$, or equivalently,

$$f(z_{t+1}) - f(x_*) + b_{t+1}\cdot\|x_{t+1}-x_*\|^2 - (1-\xi_{t+1})\cdot\left[f(z_t)-f(x_*)+b_t\cdot\|x_t-x_*\|^2\right] \tag{III.17}$$

is less than equal to zero. To that end, we first upper bound (III.17) following §III.2.1.

III.3.1 Upper bounding the Lyapunov difference

We use the same notation Δ_c, ∇, V, W as in §III.2.1. Moreover, since we only consider a single iteration, we **drop the subscripts** of $\alpha_t, \beta_{t+1}, \xi_{t+1}$ and simply write them

α, β, ξ. Then, the same derivation as before, except for the fact that now b_{t+1} and b_t could be different, we obtain the following upper bound on $\Phi_{t+1} - \Phi_t$:

$$\Phi_{t+1} - \Phi_t \leq C_1 \cdot \|W\|^2 + C_2 \cdot \|V\|^2 + C_3 \|\nabla\|^2 + C_4 \cdot \langle W, V\rangle + C_5 \cdot \langle W, \nabla\rangle + C_6 \cdot \langle V, \nabla\rangle\,, \tag{III.18}$$

$$\text{where} \begin{cases} C_1 := \beta^2 b_{t+1} - (1-\xi) b_t - \frac{\mu}{2}\frac{\alpha^2}{(1-\alpha)^2}(1-\xi)\,, & C_2 := b_{t+1} - (1-\xi)b_t - \frac{\mu}{2}\xi\,, \\ C_3 := \frac{(1-\beta)^2}{\mu^2} b_{t+1} - \Delta_c\,, & C_4 := 2(\beta b_{t+1} - (1-\xi)b_t)\,, \\ C_5 := \frac{\alpha}{1-\alpha}(1-\xi) - 2\beta\frac{1-\beta}{\mu} b_{t+1}\,, & C_6 := \xi - 2\frac{1-\beta}{\mu} b_{t+1}\,. \end{cases}$$

III.3.2 Ensuring decrease in Lyapunov function

In order to ensure that (III.18) is non-positive. we again make the coefficients C_4, C_5, C_6 of the cross terms 0, while making C_1, C_2, C_3 non-positive. Similarly to §III.2.2, this strategy *fully* determines the parameters.

- *Coefficients of cross terms characterize α, β, b_{t+1} in terms of ξ, b_t:*

  - *$C_4 = 0$ and $C_6 = 0$ characterize β, b_{t+1} in terms of ξ, b_t:* From $C_4 = 0$, we get $\beta b_{t+1} = (1-\xi)b_t$, and from $C_6 = 0$, we get $(1-\beta)b_{t+1} = \frac{\mu}{2}\xi$. Adding them up we obtain

    $$b_{t+1} = \frac{\mu}{2}\xi + (1-\xi)b_t. \tag{III.19}$$

    Plugging this back to $\beta b_{t+1} = (1-\xi)b_t$, we also obtain

    $$\beta = \frac{(1-\xi)b_t}{\frac{\mu}{2}\xi + (1-\xi)b_t}. \tag{III.20}$$

  - *$C_5 = 0$ characterizes α in terms of ξ, b_t:* From $C_5 = 0$, we get $\frac{\alpha}{1-\alpha}(1-\xi) = 2\beta\frac{1-\beta}{\mu}b_{t+1} = \beta\xi$. Hence,

    $$\frac{\alpha}{1-\alpha} = \frac{\xi b_t}{\frac{\mu}{2}\xi + (1-\xi)b_t}. \tag{III.21}$$

- *Coefficients of squared terms determines ξ based on given b_t:*

  - From $C_3 \le 0$, we get $\frac{(1-\beta)^2}{\mu^2} b_{t+1} \le \Delta_c$. Now, since we already have characterized β and b_{t+1} in terms of ξ and b_t, this inequality can be rewritten completely in terms of ξ and b_t. Concretely, using $(1-\beta)b_{t+1} = \frac{\xi\mu}{2}$, one can rewrite the inequality as

$$\xi^2 \le 4\Delta_c b_{t+1}\,. \tag{III.22}$$

    Now, using (III.19), we obtain

$$\frac{\xi^2 - 2\mu\Delta_c\xi}{1-\xi} \le 4\Delta_c b_t\,. \tag{III.23}$$

    In (III.23), note that the RHS is a nonnegative constant and the LHS is an increasing function on $[2\mu\Delta_c, 1)$ whose value is 0 at $2\mu\Delta_c$ and approaches $+\infty$ as $\xi \to 1$. Hence, the largest ξ satisfies (III.23) with equality.

  - One can then easily verify that with these choices, we get $C_2 = 0$ and $C_1 \le 0$.

III.3.3 Summary of analysis

Although the overall derivations and calculations are similar to those in §III.2, the resulting parameter choices are quite complicated as α, β, ξ, b now depend on t. Here we parse the expressions into more interpretable forms. For clarity, we recover the subscripts for the parameters α, β, ξ, b .

To summarize, we have concluded that given b_t from the previous iteration, $\xi_{t+1} \in [2\mu\Delta_c, 1)$ is chosen so that

$$\frac{\xi_{t+1}^2 - 2\mu\Delta_c\xi_{t+1}}{1-\xi_{t+1}} = 4\Delta_c b_t. \tag{III.24}$$

Using this relation, or equivalently $(1-\xi_{t+1})b_t = \frac{\xi_{t+1}^2}{4\Delta_c} - \frac{\mu}{2}\xi_{t+1}$, one can in fact eliminate the appearances of b_t in the expressions (III.19), (III.20) and (III.21) and express them solely in terms of ξ_{t+1} as follows:

$$\begin{aligned}
b_{t+1} &= \frac{\mu}{2}\xi_{t+1} + (1-\xi_{t+1})b_t = \frac{\xi_{t+1}^2}{4\Delta_c},\\
\beta_{t+1} &= \frac{(1-\xi_{t+1})b_t}{\frac{\mu}{2}\xi_{t+1} + (1-\xi_{t+1})b_t} = \frac{{\xi_{t+1}^2}/{4\Delta_c} - {\mu\xi_{t+1}}/{2}}{{\xi_{t+1}^2}/{4\Delta_c}} = 1 - 2\mu\Delta_c\xi_{t+1}^{-1},\\
\frac{\alpha_t}{1-\alpha_t} &= \frac{\xi_{t+1}b_t}{\frac{\mu}{2}\xi_{t+1} + (1-\xi_{t+1})b_t} = \frac{\xi_{t+1}\cdot {\xi_{t+1}^2-2\mu\Delta_c\xi_{t+1}}/{4\Delta_c(1-\xi_{t+1})}}{{\xi_{t+1}^2}/{4\Delta_c}} = \frac{\xi_{t+1}-2\mu\Delta_c}{1-\xi_{t+1}}.
\end{aligned}$$

For α_t, one can solve the expression to $\alpha_t = \frac{\xi_{t+1}-2\mu\Delta_c}{1-2\mu\Delta_c}$. From these expressions, one can easily check that α_t, β_{t+1} both lie in $[0,1]$ since $\xi_{t+1} \in [2\mu\Delta_c, 1)$. For coherency, let us write $b_t = \frac{\xi_t^2}{4\Delta_c}$ for some $\xi_t \geq 0$. Then our findings can be succinctly written as follows.

Theorem III.2 (Parameter choice for decrease in Lyapunov function)**.** For $c \in (0, 2/L)$, let $\Delta_c := c(1-Lc/2)$. Given $\xi_t \geq 0$, let us choose parameters as follows:

1. Compute $\xi_{t+1} \in [2\mu\Delta_c, 1)$ satisfying

$$\frac{\xi_{t+1}(\xi_{t+1}-2\mu\Delta_c)}{1-\xi_{t+1}} = \xi_t^2. \tag{III.25}$$

2. Choose parameters as $\alpha_t = \frac{\xi_{t+1}-2\mu\Delta_c}{1-2\mu\Delta_c}$ and $\beta_{t+1} = 1-2\mu\Delta_c\xi_{t+1}^{-1}$.

Then, given x_t, z_t, the next iterates x_{t+1}, z_{t+1} computed as per (III.16) satisfy

$$f(z_{t+1}) - f(x_*) + \tfrac{\xi_{t+1}^2}{4\Delta_c}\cdot\|x_{t+1}-x_*\|^2 \leq (1-\xi_{t+1})\cdot\left[f(z_t)-f(x_*)+\tfrac{\xi_t^2}{4\Delta_c}\cdot\|x_t-x_*\|^2\right].$$

Remarkably the parameter choices obtained by Theorem III.2 *exactly match* those of Nesterov's "General Scheme for Optimal Method" [Nes18, (2.2.1)]. Hence, our approach recovers Nesterov's optimal method that encompasses both strongly and non-strongly convex costs. Another byproduct of our analysis is the convergence of x_t to x_* for $\mu > 0$ (in which case, $\xi > 0$), a property otherwise proved via additional analysis (see, e.g., [GN18, Corollary 1]). This convergence will play a crucial role in the Riemannian setting. Observe that upon applying Theorem III.2 recursively, we can deduce that

$$f(z_T) - f(x_*) = O\left((1-\xi_1)(1-\xi_2)\cdots(1-\xi_T)\right). \tag{III.26}$$

Thus, to identify the convergence rate of iteration (III.3) with parameters chosen via Theorem III.2, we only need to study how the sequence $\{\xi_t\}$ evolves. This evolution is the focus of the next subsection.

III.3.4 Identifying the convergence rate via fixed-point iteration

Let us study the evolution of $\{\xi_t\}$. Our approach offers an alternative to its counterpart in Nesterov's book [Nes18, 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.

Now let us examine the recursive relation satisfied by ξ_t. Recall from Theorem III.2 the following *nonlinear* recursive relation on ξ_t's:

$$\frac{\xi_{t+1}^2 - 2\mu\Delta_c\xi_{t+1}}{1-\xi_{t+1}} = 4\Delta_c b_t.. \tag{III.24}$$

Our objective is to characterize the evolution of ξ_t. Intuitively, (III.24) can be construed as a recursive relation for computing the root of $\phi(v) = \psi(v)$, where $\phi(v) := \frac{v(v-2\mu\Delta_c)}{(1-v)}$ and $\psi(v) := v^2$. Since the root is equal to $v = \sqrt{2\mu\Delta_c}$, one can guess that $\xi_t \to \sqrt{2\mu\Delta_c}$. See Figure III-3 for illustration. The following lemma confirms this guess.

Lemma III.1 (Evolution of (III.24))**.** For an arbitrary initial value $\xi_0 \geq 0$, let ξ_t ($t \geq 1$) be the sequence of numbers defined as per (III.24). Then, $\xi_t \in [2\mu\Delta_c, 1)$ for all $t \geq 1$. Furthermore, if

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

In particular, the convergences are *geometric.*

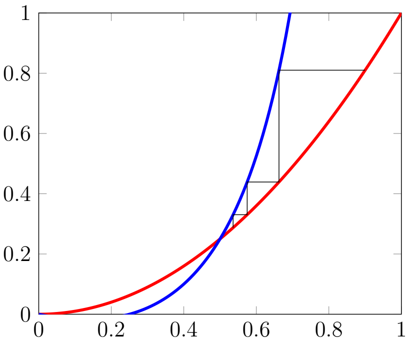


Figure III-3: An illustration of the evolution of (III.24) for $2\mu\Delta_c = 0.25$. We plot $\phi = \frac{v(v-2\mu\Delta_c)}{(1-v)}$ in blue and $\psi(v) = v^2$ in red.

Proof. The proof and the formal statement (Lemma III.2) are provided in §III.4. $\square$

Lemma III.1 delivers the desired accelerated convergence rate:

Corollary III.1. If $\xi_0 \geq \sqrt{2\mu\Delta_c}$, then $f(z_t) - f(x_*) = O(\prod_{i=1}^{t}(1 - \sqrt{2\mu\Delta_c})) = O(\exp(-t\sqrt{2\mu\Delta_c}))$. In particular, setting $c = 1/L$, $f(z_T) - f(x_*) = O(\exp(-T\sqrt{\mu/L}))$.

Remark III.1 (Removing technical conditions in Nesterov's analysis)**.** Nesterov's original analysis requires a technical condition on the initial value ξ_0: $\sqrt{\mu/L} \leq \xi_0 \leq (2(3+\mu/L))/(3+\sqrt{21+4\mu/L})$ [Nes18, (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.

III.4 Analysis of the key recursive relations

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

$$\frac{\xi_{t+1}(\xi_{t+1} - a)}{1 - \xi_{t+1}} = \frac{1}{\delta} \cdot \xi_t^2 \,. \tag{III.27}$$

In particular, when $\delta = 1$ and $a = 2\mu\Delta_c$, equation (III.27) recovers (III.24). The parameter $\delta > 1$ is present to cover the recursion for the Riemannian case (see (IV.5)). Below, we state and prove the following general statement of Lemma III.1.

Lemma III.2. 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 (III.27):

1. $\xi(\delta) := \frac{1}{2}\sqrt{(\delta-1)^2 + 4\delta a} - \frac{1}{2}(\delta - 1)$ is the unique fixed point of (III.27).
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 (III.27) can be rewritten as

$$\phi(\xi_{t+1}) = \psi(\xi_t)\,. \tag{III.28}$$

Now, in order to understand (III.28), 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 $\tau := \phi^{-1} \circ \psi$, (III.28) can be rewritten as:

$$\xi_{t+1} = \tau(\xi_t)\,. \tag{III.29}$$

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 (III.29).

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 (III.29) 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 $|\tau'(v)| \le K < 1$ for $v \in [a,1)$, then we have

$$|\xi_{t+1} - \xi(\delta)| = |\tau(\xi_t) - \tau(\xi(\delta))| \le K \cdot |\xi_t - \xi(\delta)| . \tag{III.30}$$

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 III.1 below shows this claim.

Proposition III.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

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

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

From Proposition III.1 and inequality (III.30), the proof of the geometric convergence follows. □

III.4.1 Justification of Remark IV.1

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

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

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 $a < 1$.

III.5 Related work for Chapter III

Another prominent approaches related to our Lyapunov function analysis are developed based on solving semidefinite programmings (SDP) [DT14, LRP16, TVSL18, TB19]. 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 Lyapunov functions via solving SDPs. Note that one *cannot* solve SDPs unless the numeric coefficients are given. Abstractly, our choice of parameters in Theorem III.2 can be interpreted as an *analytical* solution to the *symbolic* versions of SDPs formulated in the prior works.

Notable exceptions are [KF16, HL17, SJFB18, CHVSL18, AFGO20], in which small SDPs are solved *analytically*. Specifically, some optimized step sizes for Nesterov's method are derived via solving small SDPs explicitly in [KF16, SJFB18]; robust versions of gradient methods are derived analytically via classical control-theoretic arguments in [CHVSL18, AFGO20], and Nesterov's method is reinterpreted using *dissipativity theory* in [HL17].

Chapter IV

Globally accelerated gradient method on Riemmanian manifolds

The goal of this chapter is to generalize the result in Chapter III to the case where the cost is define on a Riemannian manifold. Without going into the details, we first informally state the main result of this chapter as follows; the formal statement is Theorem IV.3.

Theorem IV.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**).

Paralleling Chapter III, our approach is based on Lyapunov function analysis. There is, however, one fundamental hurdle inherent to Lyapunov function analysis in the Riemannian setting: we need to handle the incompatibility of metrics between two different points.

To overcome such a difficulty, we consider two important concepts: "*projected distances*" (Definition IV.3) and "*metric distortion*" (Definition IV.4). First, we introduce a crucial but *a priori* non-obvious modification to the Euclidean Lyapunov function using projected distances. With this modification, the main difficulty caused by Riemannian geometry is localized into metric distortion. Already for the simplified setting of constant metric distortion, our analysis implies the local acceleration results of [ZS18] (Corollary IV.2). To tackle global acceleration, we establish a novel metric distortion inequality based on comparison theorems in Riemannian geometry (§IV.3.1). We then show how distortion can be estimated at each iteration, which proves critical to obtain a computationally tractable algorithm (Algorithm IV.1). We show that distortion decreases over iterations (§IV.3.2), which ultimately leads to Theorem IV.1 (formal result, Theorem IV.3). One notable aspect of our analysis is that for negatively curved spaces, we *do not* assume that the iterates of the algorithm lies in some bounded domain, unlike previous works [ZS18, AOBL20, AOBL21].

We begin by recalling some basic concepts from Riemannian geometry, and defer to textbooks (e.g., [Car92, Jos08, BBB⁺01]) for more.

IV.1 Brief background on Riemannian geometry

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_xM. 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 IV.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., [BBB⁺01, Chapter 10] and [Per95].

Operations on manifolds. We can define analogs of vector addition and subtraction on Riemannian manifolds via *exponential maps*. An exponential map $\mathrm{Exp}_x : T_xM \to$

M maps $v \in T_xM$ 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 $\mathrm{Exp}_x^{-1}(y) \in T_xM$ 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 [AMS09].[1]

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) [Gro78]. One can also define geodesic-smoothness and (strong) g-convexity akin to their Euclidean counterparts. For simplicity, we assume that the function $f : M \to \mathbb{R}$ is differentiable throughout the definitions, and we denote by $\nabla f(x) \in T_xM$ the gradient of f at x.

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

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

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

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

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

[1] 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.

Assumption IV.2. We assume that the cost function f is geodesically L-smooth and μ-strongly convex.

IV.2 Riemannian Lyapunov function analysis

Using the exponential maps, one can write a Riemannian analog of (III.16).

Riemannian analog of (III.16):

$$y_t = \mathrm{Exp}_{z_t}\left(\alpha_t \mathrm{Exp}_{z_t}^{-1}(x_t)\right)\ , \tag{IV.1a}$$

$$x_{t+1} = \mathrm{Exp}_{y_t}\left(\beta_{t+1}\mathrm{Exp}_{y_t}^{-1}(x_t) - \tfrac{(1-\beta_{t+1})}{\mu}\nabla f(y_t)\right), \tag{IV.1b}$$

$$y_{t+1} = \mathrm{Exp}_{y_t}\left(-c\nabla f(y_t)\right)\ . \tag{IV.1c}$$

Here $\alpha_t := \frac{1/L}{1/L+\widetilde{\eta}_t}$, $\beta_{t+1} = \frac{1/\mu}{1/\mu+\eta_{t+1}}$, and $c \in (0, 2/L)$.

Note that updates (IV.1b) and (IV.1c) are well-defined since $\nabla f(y_t)$ lies in the tangent space $T_{y_t}M$. See Figure IV-1 for an illustration of (IV.1).

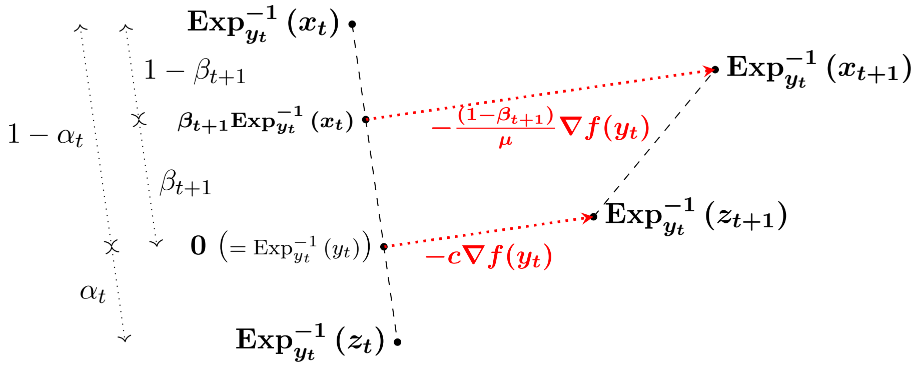

$\mathbf{Exp}_{y_t}^{-1}(x_t)$
$1-\beta_{t+1}$
$1-\alpha_t$
$\beta_{t+1}\mathbf{Exp}_{y_t}^{-1}(x_t)$
$\mathbf{Exp}_{y_t}^{-1}(x_{t+1})$
$-\frac{(1-\beta_{t+1})}{\mu}\nabla f(y_t)$
β_{t+1}
$\mathbf{Exp}_{y_t}^{-1}(z_{t+1})$
$\mathbf{0}\ \left(=\mathrm{Exp}_{y_t}^{-1}(y_t)\right)$
$-c\nabla f(y_t)$
α_t
$\mathbf{Exp}_{y_t}^{-1}(z_t)$

Figure IV-1: An illustration of the update rule (IV.1) on the tangent space $T_{y_t}M$.

We are now ready to analyze the Riemannian iteration (IV.1).

IV.2.1 Riemannian Lyapunov function analysis and metric distortion

Since (IV.1) is a direct analog of its Euclidean counterpart (III.16), one may be tempted to use the Lyapunov function $\Psi_t := \prod_{i=1}^t (1-\xi_t)^{-1} \cdot [f(z_t) - f(x_*) + b_t \cdot d(x_t, x_*)^2]$ that is a direct analog of the Lyapunov (III.8). However, it turns out that the following less direct choice is much more advantageous: for $t \geq 1$,

$$\boxed{\Psi_t := \prod_{i=1}^{t}(1-\xi_t)^{-1} \cdot \left[f(z_t) - f(x_*) + b_t \cdot \left\|\mathrm{Exp}_{y_{t-1}}^{-1}(x_t) - \mathrm{Exp}_{y_{t-1}}^{-1}(x_*)\right\|_{y_{t-1}}^2\right] .} \tag{IV.2}$$

Here, $\Psi_0 = f(x_0) - f(x_*) + b_0 \cdot d\left(x_0, x_*\right)^2$. The distance term in (IV.2) is preferable to $d(x_t, x_*)^2$ 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 IV.3 (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 Lyapunov 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 Lyapunov (IV.2), we define the following quantity for comparing distances:

Definition IV.4 (Valid distortion rate)**.** We say δ_t is a *valid distortion rate* for iteration $t \geq 2$ if the following inequality holds: $d_{y_{t-1}}(x_{t-1}, x_*)^2 \leq \delta_t \cdot d_{y_{t-2}}(x_{t-1}, x_*)^2$.

Assuming the existence of valid distortion rates at *each* iteration, we can analyze iteration (IV.1) analogously to the analysis in Chapter III to obtain the main result of this section.

Theorem IV.2 (Riemannian analog of Theorem III.2)**.** For $c \in (0, 2/L)$, let $\Delta_c := c(1 - Lc/2)$. Assume that $\delta_{t+1} > 1$ is a valid distortion rate for iteration $t + 1$. Given $\xi_t \geq 0$, let us choose parameters as follows:

1. Compute $\xi_{t+1} \in [2\mu\Delta_c, 1)$ satisfying

$$\frac{\xi_{t+1}(\xi_{t+1} - 2\mu\Delta_c)}{1 - \xi_{t+1}} = \frac{1}{\boldsymbol{\delta_{t+1}}}\xi_t^2. \tag{IV.3}$$

2. Choose parameters as $\alpha_t = \frac{\xi_{t+1} - 2\mu\Delta_c}{1 - 2\mu\Delta_c}$ and $\beta_{t+1} = 1 - 2\mu\Delta_c\xi_{t+1}^{-1}$.

Then, given x_t, z_t, the next iterates x_{t+1}, z_{t+1} computed as per (IV.1) satisfy

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

The proof is essentially identical to that of Theorem III.2, except for the appearance of valid distortion rates in (IV.3). We consider analogous notations to Chapter III.

Notations: we denote the recurring vectors as follows:

$$\tilde{\nabla} := \nabla f(y_t), \quad \tilde{W} := \mathrm{Exp}_{y_t}^{-1}(x_t)\ , \quad \tilde{V} := -\mathrm{Exp}_{y_t}^{-1}(x_*)\ .$$

Note that the above three vectors lie in the *same* tangent space $T_{y_t}M$.

With these vectors, akin to (III.18), it is again straightforward to derive the following upper bound on $\Psi_{t+1} - \Psi_t$ in terms of the vectors $\tilde{\nabla}, \tilde{V}, \tilde{W}$ (here, $\|\cdot\|$ denotes $\|\cdot\|_{x_{t+1}}$ and $\langle\cdot,\cdot\rangle$ denotes $\langle\cdot,\cdot\rangle_{x_{t+1}}$):

$$\Psi_{t+1} - \Psi_t \leq \tilde{C}_1 \cdot \|\tilde{W}\|^2 + \tilde{C}_2 \cdot \|\tilde{V}\|^2 + \tilde{C}_3\|\tilde{\nabla}\|^2 + \tilde{C}_4 \cdot \langle\tilde{W}, \tilde{V}\rangle + \tilde{C}_5 \cdot \langle\tilde{W}, \tilde{\nabla}\rangle + \tilde{C}_6 \cdot \langle\tilde{V}, \tilde{\nabla}\rangle, \tag{IV.4}$$

$$\text{where} \begin{cases} C_1 := \beta^2 b_{t+1} - (1-\xi)\frac{b_t}{\boldsymbol{\delta_{t+1}}} - \frac{\mu}{2}\frac{\alpha^2}{(1-\alpha)^2}(1-\xi)\,, & C_2 := b_{t+1} - (1-\xi)\frac{b_t}{\boldsymbol{\delta_{t+1}}} - \frac{\mu}{2}\xi\,, \\ C_3 := \frac{(1-\beta)^2}{\mu^2}b_{t+1} - \Delta_c\,, & C_4 := 2(\beta b_{t+1} - (1-\xi)\frac{b_t}{\boldsymbol{\delta_{t+1}}})\,, \\ C_5 := \frac{\alpha}{1-\alpha}(1-\xi) - 2\beta\frac{1-\beta}{\mu}b_{t+1}\,, & C_6 := \xi - 2\frac{1-\beta}{\mu}b_{t+1}\,. \end{cases}$$

Notice the similarity between (IV.4) and (III.18): the only difference is that the B_t's in (III.18) are replaced with b_t/δ_{t+1}'s here. This difference is precisely attributed to the definition of valid distortion rate (Definition IV.4). In the derivation of (IV.4), we use $-b_t \cdot d_{y_{t-1}}(x_t, x_*)^2 \leq -\frac{b_t}{\delta_{t+1}} \cdot d_{y_t}(x_t, x_*)^2$, which precisely accounts for the appearance of b_t/δ_{t+1} instead of b_t.

Having this counterpart (IV.4) of (III.18), the rest of the proof follows identically. It turns out that due to similarity between (IV.4) and (III.18), the same derivation holds modulo the appearance of δ_{t+1} in the denominator of (IV.3).

As before, we can deduce from Theorem IV.2 the suboptimality gap bound (III.26). 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 §IV.3.

Illustrative example: constant distortion rate. Assume that μ is positive, and consider the simplified case where $\delta_t \equiv \delta \geq 1$ for all $t \geq 0$. Under this constant distortion condition, similarly to recursion (III.24), one can obtain a recursive relation on $\{\xi_t\}$:

$$\frac{\xi_{t+1}(\xi_{t+1} - 2\mu\Delta_c)}{1 - \xi_{t+1}} = \frac{1}{\boldsymbol{\delta}}\xi_t^2. \tag{IV.5}$$

Analogously to Lemma III.1, we can establish geometric convergence of ξ_t to the fixed point $\xi(\delta)$ of (IV.5) (see Lemma III.2). Solving for $\xi(\delta)$ explicitly, we obtain the following analog of Corollary III.1:

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

A notable aspect of Corollary IV.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 IV.1 (Properties of $\xi(\delta)$)**.** When there is no distortion, i.e., $\delta = 1$, then $\xi(1) = \sqrt{2\mu\Delta_c}$ since (IV.5) becomes (III.24). 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 §III.4.1 for verification). Hence, $\xi(\delta) > \lim_{\delta\to\infty} \xi(\delta) = 2\mu\Delta_c$ 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 [ZS18]. 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 [ZS18, Theorem 3]:

Corollary IV.2 (Local acceleration)**.** Let $\delta = 1 + \frac{1}{5} \cdot ({}^\mu/_L)^{1/2}$, $c = 1/L$ and $\xi_0 \geq \xi(\delta)$. Then, assuming $d\,(x_0, x_*) \leq \frac{1}{20} \cdot \kappa^{1/2}({}^\mu/_L)^{3/4}$, we have $f(z_t) - f(x_*) = O(\exp(-\frac{9}{10} t \sqrt{{}^\mu/_L}))$. In particular, $\xi_t = \xi(\delta)$ for all $t \geq 0$, recovers [ZS18, Algorithm 2].

IV.3 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. In Appendix A.1, we discuss how one can extend our result to positively-curved manifolds under an appropriate assumption.

IV.3.1 Valid distortion rates and Riemannian accelerated gradient method

We estimate metric distortion by first invoking a classical comparison theorem of [Rau51].

Proposition IV.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*

$$d\left(y, z\right)^2 \leq S_\kappa(\max\{d\left(x, y\right), d\left(x, z\right)\}) \cdot d_x(y, z)^2.$$

Proof. A direct consequence of the Rauch comparison theorem; see §IV.4. □

Applying Proposition IV.1 to the points y_{t-1}, x_t, x_*, it is straightforward to conclude:

$$d_{y_t}(x_t, x_*) \overset{(\clubsuit)}{\leq} d\left(x_t, x_*\right)^2 \overset{(\spadesuit)}{\leq} S_\kappa(\max\{d\left(y_{t-1}, x_t\right),\ d\left(y_{t-1}, x_*\right)\}) \cdot d_{y_{t-1}}(x_t, x_*)^2,$$

where ($\clubsuit$) is due to Topogonov's comparison theorem (see e.g., [BBB+01, Section 6.5]); and ($\spadesuit$) is due to Proposition IV.1. Hence, $\delta_t = S_\kappa(\max\{d\left(y_{t-1}, z_t\right),\ d\left(y_{t-1}, x_*\right)\})$ 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 IV.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

$$T_\kappa(r) := \begin{cases} \max\left\{1 + 4\left(\frac{\sqrt{\kappa} r}{\tanh(\sqrt{\kappa} r)} - 1\right),\ \left(\frac{\sinh(2\sqrt{\kappa} \cdot r)}{2\sqrt{\kappa} \cdot r}\right)^2\right\}, & \text{if } r > 0, \\ 1, & \text{if } r = 0, \end{cases} \tag{IV.6}$$

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

Proof. The proof uses Proposition IV.1 and a Riemannian trigonometric inequality due to [ZS16, Lemma 6]. See §IV.4 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 IV.1 as a version of Proposition IV.1 in which the term $\max\{d\left(x, y\right), d\left(x, z\right)\}$ is replaced with $d\left(x, y\right)$. Thanks to Lemma IV.1, now we have $T_\kappa(d\,(y_{t-1}, x_t))$ as a valid distortion rate, which is *accessible* to the algorithm at iteration t. Therefore, we propose the following algorithm:

Algorithm IV.1 (Riemannian accelerated gradient method)**. Input:** $x_0 = y_0 = z_0 \in M$; constant $\xi_0 > 0$; $c \in (0, {}^{2}\!/_{L})$; $\Delta_c := c(1 - Lc/2)$; integer T.

```
    for $t = 0, 1, 2, \dots, T$:
        Compute the distortion rate $\delta_{t+1} := T_\kappa(d\,(y_{t-1}, x_t))$ as per (IV.6).
        Find $\xi_{t+1} \in [2\mu\Delta_c, 1)$ such that $\xi_{t+1}(\xi_{t+1} - 2\mu\Delta_c)/(1 - \xi_{t+1}) = \xi_t^2/\delta_{t+1}$.
        Compute $\alpha_t := \frac{\xi_{t+1} - 2\mu\Delta_c}{1 - 2\mu\Delta_c}$ and $\beta_{t+1} := 1 - 2\mu\Delta_c\xi_{t+1}^{-1}$.
        Update the next step iterates as per (IV.1).
    end for
```

Remark IV.2 (Innovations relative to previous methods)**.** A noticeable innovation in Algorithm IV.1 lies in its use of the adaptive metric distortion rate $T_\kappa(d\,(y_{t-1}, x_t))$. This is in stark contrast with previous approaches [ZS16, ZS18, AOBL20] that use a global metric distortion rate based on the diameter of the domain. As we shall we in §IV.3.2, our adaptive metric distortion control is a crucial ingredient for achieving *full* acceleration.

Remark IV.3. Note that $T_\kappa(d\,(y_{t-1}, x_t))$ 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 $T_\kappa(d\,(y_{t-1}, x_t))$ in Algorithm IV.1 with the estimate.

IV.3.2 Convergence rate analysis of the proposed method

Having proposed the algorithm, our final task is to analyze its convergence rate. From Remark IV.1, 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 $d\,(y_{t-1}, x_t)$ is close to 0. Although $d\,(y_{t-1}, x_t) = 0$ for $t = 0$, one can quickly notice that it is not true for $t \geq 1$.

Now one natural follow-up question is whether $d\,(y_{t-1}, x_t)$ shrinks over iterations. As we have seen in §III.2.2, the convergence of the iterates to the optimal point is a direct consequence of our Lyapunov function analysis. Similarly, one can immediately see that $d_{y_{t-1}}(x_t, x_*) \to 0$. It turns out that from this shrinking projected distance, one can also deduce $d\left(y_{t-1}, x_t\right) \to 0$ under mild conditions:

Lemma IV.2 (Shrinking $d\left(y_{t-1}, x_t\right)$)**.** Assume $\mu > 0$ and let $\Psi_0 := f(x_0) - f(x_*) + \frac{\xi_0^2}{4\Delta_c} \cdot d\left(x_0, x_*\right)^2$. If $1 < cL < 2 - \xi_t$ and $\xi_t > 2\mu\Delta_c$ hold at iteration $t \geq 1$, then Algorithm IV.1 satisfies:

$$d\left(y_{t-1}, x_t\right) \leq \mathcal{C}_{\mu,L,c}\big[\Psi_0 \prod_{j=1}^{t-1}(1-\xi_j)\big]^{1/2},$$

where $\mathcal{C}_{\mu,L,c} > 0$ is a constant depending only on μ, L, c.

Proof. See §IV.5. □

Note that the assumption $cL \in (1, 2 - \xi_t]$ can be roughly read as "$cL \in (1, 2 - \sqrt{\mu/L}]$" because Remark IV.1 ensures that $\xi(\delta) \leq \sqrt{2\mu\Delta_c} < \sqrt{\mu/L}$ 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 IV.1):

Theorem IV.3 (Global acceleration of Algorithm IV.1)**.** Assume $0 < \mu < L$ and $cL \in (1, 2 - \sqrt{\mu/L}]$. Let $\Delta_c := c(1 - Lc/2)$ and $\lambda := 1 - {}^{8\mu\Delta_c}/_{(5+\sqrt{5})} \in (0,1)$. Then for any $\xi_0 > 0$, Algorithm IV.1 satisfies the following accelerated convergence:

$$f(z_t) - f(x_*) = O\left((1-\xi_1)(1-\xi_2)\cdots(1-\xi_t)\right), \tag{IV.7}$$

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

Proof. (IV.7) is immediate from Theorem IV.2. For the convergence of $\{\xi_t\}$, see §IV.6. □

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

IV.4 Proof of geometric inequalities

This section is devoted to proving Lemma IV.1. The proof requires two ingredients: Proposition IV.1 and a (Riemannian) trigonometric inequality due to [ZS16, Lemma 6]. We begin with the first key ingredient: Proposition IV.1. Its proof is based on the following version of the Rauch comparison theorem [Cha06, Theorem IX.2.3]:

Proposition IV.2 (Rauch comparison theorem)**.** *Let M be a Riemannian manifold with sectional curvatures lower bounded by $-\kappa < 0$. Then, for any $x \in M$ and $u \in T_xM$, the following upper bound on the operator norm of the differential of the exponential map holds:*

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

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 [Cha06, 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 $T_u(T_xM)$ 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 IV.1:

Proposition IV.3 (Restatement of Proposition IV.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 $p : [0,1] \to T_xM$ 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 IV.2, the length of $\mathrm{Exp}_x(p)$ can be upper bounded as follows (since $\|p'(t)\| = \left\|\mathrm{Exp}_x^{-1}(y) - \mathrm{Exp}_x^{-1}(z)\right\| = d_x(y,z)$):

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

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 IV.4 (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 [ZS16, §3.1] and [CEMS01, Lemma 3.12]. □

With these ingredients we now prove Lemma IV.1; we actually prove the following stronger version.

Lemma IV.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*

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

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 IV.1 immediately implies Lemma IV.1.

Proof. *[Proof of Lemma IV.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 IV.4 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_xM), $d_x(y,z) \geq (d(x,z) - d(x,y)) > \frac{\epsilon}{1+\epsilon}\cdot d(x,z)$. Hence, combining these two, we get

$$\begin{aligned} d(y,z)^2 &\leq (\zeta-1)\cdot d(x,z)^2 + d_x(y,z)^2 \\ &\leq \left(1+\epsilon^{-1}\right)^2\cdot(\zeta-1)\cdot d_x(y,z)^2 + d_x(y,z)^2 \\ &= \left[1+\left(1+\epsilon^{-1}\right)^2\cdot(\zeta-1)\right]\cdot d_x(y,z)^2\,. \end{aligned} \tag{IV.8}$$

Case (ii). For the case $(1+\epsilon)\cdot d(x,y) \geq d(x,z)$, Proposition IV.3 implies:

$$d(y,z)^2 \leq \left(\frac{\sinh\left((1+\epsilon)\sqrt{\kappa}\cdot d(x,y)\right)}{(1+\epsilon)\sqrt{\kappa}\cdot d(x,y)}\right)^2\cdot d_x(y,z)^2\,. \tag{IV.9}$$

Therefore, combining (IV.8) and (IV.9), the proof is completed. □

IV.5 Proof of shrinking distance lemma

In this section, we prove the distance shrinking lemma (Lemma IV.2). We first analyze the convergence distances (which is a direct consequence of Theorem IV.2) below.

Proposition IV.5. *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 $\Psi_0 := f(x_0) - f(x_*) + \frac{\xi_0^2}{4\Delta_c}\cdot d(x_0,x_*)^2$. Then, for x_t, y_t, z_t ($t \geq 1$) generated by Algorithm IV.1* *the following bounds hold:*

1. $d_{y_{t-1}}(x_t, x_*) \leq \sqrt{\Psi_0 \prod_{j=1}^{t}(1-\xi_j)} \cdot \sqrt{\frac{1}{\mu^2\Delta_c}}$.
2. $d\left(z_t, x_*\right) \leq \sqrt{\Psi_0 \prod_{j=1}^{t}(1-\xi_j)} \cdot \sqrt{\frac{2}{\mu}}$.
3. $d_{y_{t-1}}(z_t, x_t) \leq \sqrt{\Psi_0 \prod_{j=1}^{t}(1-\xi_j)} \cdot \left(\sqrt{\frac{2}{\mu}} + \sqrt{\frac{1}{\mu^2\Delta_c}}\right)$.

Proof. By recursively applying Theorem IV.2, we have the following for any $t \geq 1$:

$$f(z_t) - f(x_*) + \frac{\xi_t^2}{4\Delta_c} \cdot d_{y_{t-1}}(x_t, x_*)^2 \leq \prod_{j=1}^{t}(1-\xi_j) \cdot \Psi_0\,.$$

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

$$d_{y_{t-1}}(z_t, x_t) \leq d_{y_{t-1}}(z_t, x_*) + d_{y_{t-1}}(x_t, x_*) \leq d\left(z_t, x_*\right) + d_{y_{t-1}}(x_t, x_*)\,, \tag{IV.10}$$

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. [BBB+01, §6.5]). □

Proposition IV.5 above establishes that the projected distance $d_{y_{t-1}}(z_t, x_t)$ is shrinking over iterations. From this, we can also show that $d\left(z_t, x_t\right)$ is shrinking under mild conditions:

Proposition IV.6. *Let* $\Psi_0 := f(x_0) - f(x_*) + \frac{\xi_0^2}{4\Delta_c} \cdot d\left(x_0, x_*\right)^2$. *If* $cL > 1$, $cL \leq 2 - \xi_{t+1}$ *and* $\xi_{t+1} > 2\mu\Delta_c$ *hold for* $t \geq 0$, *then Algorithm IV.1 satisfies:*

$$d\left(z_t, x_t\right) \leq \frac{1-2\mu\Delta_c}{1-2\mu\Delta_c\xi_{t+1}^{-1}} \cdot \sqrt{\Psi_0 \prod_{j=1}^{t}(1-\xi_j) \cdot \frac{\left(\sqrt{\frac{2}{\mu}} + \sqrt{\frac{1}{\mu^2\Delta_c}} + \frac{L}{\mu}\sqrt{\frac{2}{\mu}}\right)}{(cL-1)(cL-1+2\mu\Delta_c)}}\,.$$

Remark IV.4. A careful reader might realize that the appearance of the term $1 - 2\mu\Delta_c\xi_{t+1}^{-1}$ 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 $2\mu\Delta_c$. However, as we shall see shortly, this term gets canceled out with the algorithm parameter $\beta_{t+1} = 1 - 2\mu\Delta_c\xi_{t+1}^{-1}$ (see Algorithm IV.1) when we use Proposition IV.6 to bound the distance of our interest $d\left(y_{t-1}, x_t\right)$.

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

$$1 < cL < 2 - \xi_{t+1} \text{ and } \xi_{t+1} > 2\mu\Delta_c. \tag{IV.11}$$

Let $\gamma_{t+1} := \frac{(1-\beta_{t+1})}{\mu} = 2\Delta_c\xi_{t+1}^{-1}$. First, from (IV.1b) and (IV.1c) together with the triangle inequality (also see Figure IV-1),

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

Rearranging the above inequality we have

$$\beta_{t+1} \cdot d\left(y_t, x_t\right) \leq d_{y_t}(z_{t+1}, x_{t+1}) + \left|\gamma_{t+1} - c\right| \cdot \left\|\nabla f(y_t)\right\|_{y_t} . \tag{IV.12}$$

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

$$d\left(y_t, x_t\right) = d_{y_t}(y_t, z_t) = (1-\alpha_t)d_{y_t}(z_t, x_t) = (1-\alpha_t) \cdot d\left(z_t, x_t\right) .$$

Substituting this identity to the left hand side of (IV.12), (IV.12) becomes:

$$\begin{aligned}
&\beta_{t+1}(1-\alpha_t)\cdot d\left(z_t, x_t\right) \\
&\quad \le d_{y_t}(z_{t+1}, x_{t+1}) + |\gamma_{t+1}-c|\cdot \left\|\nabla f(y_t)\right\|_{x_{t+1}} \\
&\quad \overset{(\clubsuit)}{\le} d_{y_t}(z_{t+1}, x_{t+1}) + L|\gamma_{t+1}-c|\cdot d\left(y_t, x_*\right) \\
&\quad \overset{(\spadesuit)}{=} d_{y_t}(z_{t+1}, x_{t+1}) + L(\gamma_{t+1}-c)\cdot d\left(y_t, x_*\right) \\
&\quad \overset{(\heartsuit)}{\le} d_{y_t}(z_{t+1}, x_{t+1}) + L(\gamma_{t+1}-c)\cdot d\left(y_t, z_t\right) + L(\gamma_{t+1}-c)\cdot d\left(z_t, x_*\right), \\
&\quad = d_{y_t}(z_{t+1}, x_{t+1}) + L\alpha_t(\gamma_{t+1}-c)\cdot d\left(z_t, x_t\right) + L(\gamma_{t+1}-c)\cdot d\left(z_t, x_*\right)\ ,
\end{aligned}$$

where ($\clubsuit$) follows from the geodesic L-smoothness of f: $\left\|\nabla f(y_t)\right\|_{y_t} \le L\cdot d\left(y_t, x_*\right)$; and ($\spadesuit$) is due to the fact that $\gamma_{t+1}-\gamma = 2\Delta_c\xi_{t+1}^{-1} - c = c\xi_{t+1}^{-1}(2-Lc-\xi_{t+1}) > 0$ since $2-\xi_{t+1}-cL > 0$ from (IV.11); ($\heartsuit$) follows from the Riemannian triangle inequality $d\left(y_t, x_*\right) \le d\left(y_t, z_t\right) + d\left(z_t, x_*\right)$; and the last line follows from the identity $d\left(y_t, z_t\right) = \alpha_t \cdot d\left(z_t, x_t\right)$ (see Figure IV-1).

Moving the term $L\alpha_t(\gamma_{t+1}-c)\cdot d\left(z_t, x_t\right)$ to the LHS, we then obtain:

$$\begin{aligned}
&\left[\beta_{t+1}(1-\alpha_t) - L\alpha_t(\gamma_{t+1}-c)\right]\cdot d\left(z_t, x_t\right) \\
&\quad \le d_{y_t}(z_{t+1}, x_{t+1}) + L(\gamma_{t+1}-c)\cdot d\left(z_t, x_*\right)\ .
\end{aligned} \tag{IV.13}$$

Since we have seen from Proposition IV.5 that the both terms on the right hand side of (IV.13) are shrinking, one can prove that $d\left(z_t, x_t\right)$ is shrinking as long as one can guarantee that $\beta_{t+1}(1-\alpha_t) - L\alpha_{t+1}(\gamma_{t+1}-c) > 0$. More formally, Proposition IV.6 is a direct consequence the following two statements:

1. The RHS of (IV.13) is upper bounded by $\sqrt{\Psi_0\prod_{j=1}^{t}(1-\xi_j)}\cdot\left(\sqrt{\frac{2}{\mu}}+\sqrt{\frac{1}{\mu^2\Delta_c}}+\frac{L}{\mu}\sqrt{\frac{2}{\mu}}\right)$.
2. $\beta_{t+1}(1-\alpha_t) - L\alpha_t(\gamma_{t+1}-c) \ge \frac{1-2\mu\Delta_c\xi_{t+1}^{-1}}{1-2\mu\Delta_c}\cdot(cL-1)(cL-1+2\mu\Delta_c)$. Indeed, with this lower bound one can guarantee that $\beta_{t+1}(1-\alpha_t) - L\alpha_t(\gamma_{t+1}-c)$ is positive due to (IV.11): $cL > 1$ and $1-2\mu\Delta_c\xi_{t+1}^{-1} > 1-2\mu\Delta_c\cdot(2\mu\Delta_c)^{-1} = 0$.

Now let us prove the above two statements. From the third conclusion of Proposition IV.5, we have $d_{y_t}(z_{t+1}, x_{t+1}) \le \sqrt{\Psi_0\prod_{j=1}^{t+1}(1-\xi_j)}\cdot\left(\sqrt{\frac{2}{\mu}}+\sqrt{\frac{1}{\mu^2\Delta_c}}\right)$. Moreover, from the second conclusion of Proposition IV.5, we have:

$$\begin{aligned} L(\gamma_{t+1}-c)\cdot d\left(z_t,x_*\right) \leq L\gamma_{t+1}\cdot d\left(z_t,x_*\right) \leq L\gamma_{t+1}\cdot \sqrt{\Psi_0\prod_{j=1}^{t}(1-\xi_j)}\cdot\sqrt{\tfrac{2}{\mu}} \\ \leq \sqrt{\Psi_0\prod_{j=1}^{t}(1-\xi_j)}\cdot\tfrac{L}{\mu}\sqrt{\tfrac{2}{\mu}}\,, \end{aligned}$$

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

Now, let us prove the second statement. We first recall the parameters in Algorithm IV.1 for reader's convenience: For $\Delta_c := c(1-Lc/2)$, $\alpha_t = \frac{\xi_{t+1}-2\mu\Delta_c}{1-2\mu\Delta_c}$, $\beta_{t+1} = 1-2\mu\Delta_c\xi_{t+1}^{-1}$, and we also defined $\gamma_{t+1} = 2\Delta_c\xi_{t+1}^{-1}$ for simplicity. Now substituting these parameters to the coefficient, we have:

$$\begin{aligned} &\beta_{t+1}(1-\alpha_t) - L\alpha_t(\gamma_{t+1}-c) \\ =&(1-2\mu\Delta_c\xi_{t+1}^{-1})\frac{1-\xi_{t+1}}{1-2\mu\Delta_c} - L\frac{\xi_{t+1}-2\mu\Delta_c}{1-2\mu\Delta_c}\left(2\Delta_c\xi_{t+1}^{-1}-c\right) \\ =&\frac{1-2\mu\Delta_c\xi_{t+1}^{-1}}{1-2\mu\Delta_c}\cdot\left[1-\xi_{t+1}-2L\Delta_c+cL\xi_{t+1}\right] \end{aligned}$$

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

$$\begin{aligned} \beta_{t+1}(1-\alpha_t) - L\alpha_t(\gamma_{t+1}-c) =&\frac{1-2\mu\Delta_c\xi_{t+1}^{-1}}{1-2\mu\Delta_c}\cdot\left[(cL-1)^2+(cL-1)\xi_{t+1}\right] \\ >&\frac{1-2\mu\Delta_c\xi_{t+1}^{-1}}{1-2\mu\Delta_c}\cdot\left[(cL-1)^2+(cL-1)\cdot 2\mu\Delta_c\right]\,. \end{aligned}$$

where the last line follows from the facts $\xi_{t+1} > 2\mu\Delta_c$ and $cL-1>0$. □

Now, we are finally ready to provide the formal statement and the proof of Lemma IV.2:

Lemma IV.3 (Formal statement of Lemma IV.2)**.** Assume that $\mu > 0$. Let $\Psi_0 := f(x_0) - f(x_*) + \frac{1}{4\Delta_c}\xi_0^2 \cdot d\left(x_0, x_*\right)^2$. If $cL > 1$, $cL \leq 2 - \xi_{t+1}$ and $\xi_{t+1} > 2\mu\Delta_c$, then Algorithm IV.1 satisfies:

$$d\left(y_t, x_{t+1}\right) \leq \mathcal{C}_{\mu,L,c} \cdot \sqrt{\Psi_0 \prod_{j=1}^{t}(1-\xi_j)}\,,$$

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

Proof. We again recall the parameters in Algorithm IV.1 for reader's convenience: For $\Delta_c := c(1 - Lc/2)$, $\alpha_t = \frac{\xi_{t+1}-2\mu\Delta_c}{1-2\mu\Delta_c}$, $\beta_{t+1} = 1 - 2\mu\Delta_c\xi_{t+1}^{-1}$, and $\gamma_{t+1} = 2\Delta_c\xi_{t+1}^{-1}$. Now, one can use the Euclidean triangle inequality on $T_{y_t}M$ (see Figure IV-1) to obtain:

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

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

Now after we apply Propositions IV.5 and IV.6 to the last upper bound, and use the fact $\xi_{t+1} \in [2\mu\Delta_c, 1)$ to upper bound ξ_{t+1}'s in the resulting upper bound, Lemma IV.3 readily follows. □

IV.6 Proof of main theorem

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

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

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 IV.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:

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

Proposition IV.7. *If $\xi_0 > \sqrt{\mu/L}$, then $\xi_t \leq \sqrt{\mu/L}$ for all t whenever*

$$t \geq \frac{\log\big((\xi_0 - \sqrt{2\mu\Delta_c})/(\sqrt{\mu/L} - \sqrt{2\mu\Delta_c})\big)}{\log\Big(1/\big(1 - \frac{8\mu\Delta_c}{5+\sqrt{5}}\big)\Big)}\,. \tag{IV.14}$$

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 $\xi_t \leq \sqrt{2\mu\Delta_c}$ or not:

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

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

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

where ($\clubsuit$) is due to Proposition III.1 together with $\xi_{t+1} > \sqrt{2\mu\Delta_c} \Rightarrow \xi_t > \tau_{t+1}^{-1}(\sqrt{2\mu\Delta_c})$; ($\spadesuit$) follows since $\frac{1}{\sqrt{\delta}}(1 - \frac{4}{(5+\sqrt{5})} \cdot \frac{2\mu\Delta_c}{\sqrt{\delta}})$ for $\delta \geq 1$ is maximized when $\delta = 1$ and $\sqrt{2\mu\Delta_c} < \tau_{t+1}^{-1}(\sqrt{2\mu\Delta_c})$ due to $\sqrt{2\mu\Delta_c} \geq \xi(\delta_{t+1})$ and Lemma III.2. Hence, the distance between ξ_t and $\sqrt{2\mu\Delta_c}$ 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 $\xi_0 \leq \sqrt{\mu/L}$. By Proposition IV.7, the arguments below remain true for $\xi_0 > \sqrt{\mu/L}$ after we replace t with $t +$ (IV.14). We first characterize $\xi(\delta)$ near $\delta = 1$:

Proposition IV.8. *Let* $\xi(\delta) := \frac{1}{2}\left(\sqrt{(\delta-1)^2 + 8\delta\mu\Delta_c} - (\delta - 1)\right)$ *for* $\delta \geq 1$. *Then,* $0 \leq \sqrt{2\mu\Delta_c} - \xi(\delta) \leq \frac{1}{2}(\delta - 1)$ *for* $1 \leq \delta \leq 1 + 3/(1 + (4\mu\Delta_c)^{-1})$.

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

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

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

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

Proposition IV.9. $T_\kappa(r) \leq 1 + 2\kappa r^2$ *for* $0 \leq r \leq \frac{1}{2\sqrt{\kappa}}$.

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

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

Hence, from the definition of T_κ (see (IV.6)), we obtain the desired bound on T_κ. □

Combining Propositions IV.8 and IV.9, we obtain the following results:

Proposition IV.10. *$\sqrt{2\mu\Delta_c} - \xi(T_\kappa(r)) \le \kappa r^2$ for $0 \le r \le \sqrt{\frac{3}{1+(4\mu\Delta_c)^{-1}}} \cdot \frac{1}{2\sqrt{\kappa}}$.*

Proof. Note that $\frac{3}{1+(4\mu\Delta_c)^{-1}} \le \frac{3}{1+2L/\mu} \le 1$, and hence, $\sqrt{\frac{3}{1+(4\mu\Delta_c)^{-1}}} \cdot \frac{1}{2\sqrt{\kappa}} \le \frac{1}{2\sqrt{\kappa}}$. Thus, one can apply Proposition IV.9 for $0 \le r \le \sqrt{\frac{3}{1+(4\mu\Delta_c)^{-1}}} \cdot \frac{1}{2\sqrt{\kappa}}$, and obtain $T_\kappa(r) \le 1+2\kappa r^2$. Hence, $T_\kappa(r) \le 1+\frac{1}{2}\cdot\frac{3}{1+(4\mu\Delta_c)^{-1}}$ within the range. Hence, by Proposition IV.8, one then obtains $\sqrt{2\mu\Delta_c} - \xi(T_\kappa(r)) \le \kappa r^2$ for $0 \le r \le \sqrt{\frac{3}{1+(4\mu\Delta_c)^{-1}}} \cdot \frac{1}{2\sqrt{\kappa}}$. □

Let $\mathcal{D}_{\kappa,\mu,c} := \sqrt{\frac{3}{1+(4\mu\Delta_c)^{-1}}} \cdot \frac{1}{2\sqrt{\kappa}}$. Then by Lemma IV.3, we can deduce that $d\left(x_{t+1}, z_{t+1}\right) \le \mathcal{D}_{\kappa,\mu,c}$ whenever $t \ge 2\frac{\log(\mathcal{C}_{\mu,L,c}\cdot\sqrt{D_0}/\mathcal{D}_{\kappa,\mu,c})}{\log(1/(1-2\mu\Delta_c))}$. Therefore, Proposition IV.10 implies that for $t \ge 2\frac{\log(\mathcal{C}_{\mu,L,c}\cdot\sqrt{D_0}/\mathcal{D}_{\kappa,\mu,c})}{\log(1/(1-2\mu\Delta_c)))}$, the following bound holds:

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

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

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

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 IV.7, one can prove that for any $T \ge 0$,

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

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

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

concluding the proof of the the convergence rate of $\{\xi_t\}$ in Theorem IV.3.

IV.7 Related work for Chapter IV

A few works also seek to develop an accelerated methods on Riemannian manifolds. The first attempt by Liu et al. [LSC$^+$17] reduces the task to solving nonlinear equations [LSC$^+$17, (4),(5)]; unfortunately, it is unclear whether these equations are even feasible or tractably solvable, and hence this attempt has been regarded incomplete. Subsequently, Zhang and Sra [ZS18] made a concrete progress by proving an 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. Alimisis et al. [AOBL20] establish a Riemannian analog of the differential-equation approach to acceleration [SBC16], and they analyze second-order ODEs on Riemannian manifolds. Then, they employ discretization from the Euclidean case [BJW18, SDSJ19] 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 discussed in Remark IV.2, their global control of metric distortion cannot capture *full* acceleration; one must control metric distortions *locally*. Another work by Alimisis et al. [AOBL21] considers relaxed settings such as geodesically convex costs and weakly-quasiconvex costs and demonstrates the advantage of their momentum-based algorithm over (plain) gradient descent.

We also summarize noticeable follow up works to our result after our result was published [AS20]. Hamilton and Moitra [HM21] investigate a lower bound for Riemannian acceleration. More specifically, they show that in negatively curved spaces any first-order methods based on a noisy gradient oracle cannot achieve full acceleration, even when the noise of oracle is exponentially small. Their finding suggests that an *eventual* acceleration like our result might be the best one can hope for in negatively curved space. Lastly, Martínez-Rubio [MR20] considers the special case of constant sectional curvature manifolds and develops accelerated methods based on a nontrivial reduction to an Euclidean optimization problem.

Chapter V

Conclusion and future directions

In this thesis, we establish the first *global* Riemannian accelerated gradient method for geodesically strong convex costs. Our approach is based on the two steps: ① Revisit Nesterov's acceleration in the Euclidean setting and develop a simple derivation and analysis for it; ② Extend our approach to the Riemannian setting.

The first part of this thesis (Chapters II and III) addresses ① by developing a simple derivation and analysis for Nesterov's accelerated method based on the proximal point method. We demonstrate that our approach derives several different forms of accelerated methods appeared in the literature and provides simple analyses for them.

The second part of this thesis (Chapter IV) then extends our approach to the Riemannian setting. Two important components of the extension are the use of projected distance for the Lyapunov function and the utilization of metric distortion rate in the analysis. With these components together with novel geometric inequalities, we prove that our proposed algorithm is always faster than Riemmanian gradient descent and quickly achieves the full accelerated convergence rate within a few iterations.

We believe our results mark fundamental progress toward understanding acceleration in non-Euclidean settings. We hope that our work motivates a richer study of Riemannian acceleration, which will eventually bring our understanding of Riemannian optimization at par with the Euclidean setting.

We conclude this thesis with future directions. First open question is a "direct" development of an accelerated method for non-strongly geodesically convex costs. We consider strongly geodesically convex costs, i.e., $\mu > 0$, and our current technique does not directly apply to the case of $\mu = 0$. In the Euclidean case, there is a standard reduction argument from acceleration for strongly convex costs to that for non-strongly convex costs (see e.g., [GN18, Theorem 4]). However, it turns out this standard reduction requires additional assumptions, as we discuss in Appendix A.2. Hence, discovering a direct approach to acceleration for non-strongly convex costs remains an interesting open question. Note that for the special case of constant sectional curvature manifolds, a recent work by Martínez-Rubio [MR20] develops an accelerated method for the non-strongly convex case with the help of line search. However, even in the constant sectional curvature case, characterizing a tight convergence rate seems to be open.

Another direction is to consider a different notion of distance than the projected distance. For our version of Riemannian AGM (IV.1), a suitable notion of distance for the Lyapunov function is the projected distance (IV.2). This motivates the question of whether considering a different notion of distance results in a better version of Riemannian accelerated method than the one considered in this work. More generally, it is relevant to investigate the proximal point method with different choices of distance in the Riemannian setting. Based on our findings in the Euclidean case, a better understanding of Riemannian proximal point method might lead to a better version of Riemannian accelerated method.

Appendix A

Appendix

A.1 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 [DVW15, ZS18], we make the following assumptions to avoid any further complications:

Assumption A.1. The domain $N \subset M$ of f is uniquely geodesic with the diameter bounded by $\frac{\pi}{2\sqrt{\sigma}}$.

Assumption A.2 (Bounded iterates assumption)**.** All the iterates of Algorithm IV.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 [ZS18]:

Proposition A.1 ([ZS18, 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*

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

Applying Proposition A.1 to Lemma IV.1, we obtain the following metric distortion inequality:

Lemma A.1 (Modification of Lemma IV.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 IV.1, we have

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

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

$$T_\kappa(d\left(y_{t-1},x_t\right)) \cdot (1 + 2\sigma \cdot d\left(z_t,x_t\right)^2) \tag{A.1}$$

is a valid distortion rate. Thus, one can use (A.1) in lieu of $T_\kappa(d\left(y_{t-1},x_t\right))$ for the valid distortion rate in Algorithm IV.1. Then, one can invoke Theorem IV.2 with the chosen valid distortion rate (A.1) to guarantee the Lyapunov decrease. To show that Algorithm IV.1 with (A.1) eventually achieves full acceleration, the last ingredient is to show that the distances $d\left(y_{t-1},x_t\right)$ and $d\left(z_t,x_t\right)$ shrink over iterations. Indeed, one can prove that the distances shrink following the arguments in §IV.5. The only difference is that in proving Proposition IV.5, one now has the following in place of (IV.10):

$$d_{y_{t-1}}(z_t,x_t) \leq d_{y_{t-1}}(z_t,x_*) + d_{y_{t-1}}(x_t,x_*) \leq (1+\pi^2/2) \cdot d\left(z_t,x_*\right) + d_{y_{t-1}}(x_t,x_*)\,, \tag{A.2}$$

where the last inequality is due to Proposition A.1 together with the bounded iterates assumption (Assumption A.2). Hence the third statement of Proposition IV.5 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.

A.2 Proof-of-concept result for 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., [GN18, Theorem 4]. In this section, we extend such an argument to the Riemannian setting and use it to discuss accelerated guarantees for the non-strongly g-convex case. Importantly, the reduction argument requires an additional assumption that all iterates stay in a bounded domain and hence does not yield a complete result.

Remark A.1. A recent work by Martínez-Rubio [MR20] also extends the reduction argument to the Riemannian setting. We refer readers to §3 therein for the details.

We first invoke the following properties of the distance function:

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

1. *d is 1-strongly g-convex in the entire M with $\nabla d(x) = -\mathrm{Exp}_x^{-1}\left(p\right)$.*
2. *For $D \geq 0$, d is geodesically $\frac{\sqrt{\kappa}D}{\tanh(\sqrt{\kappa}D)}$-smooth within the domain $\{u \in M \;:\; d\left(u,p\right) \leq D\}$.*

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

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

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 IV.1). Next, we verify the geodesic smoothness. From the global trigonometry inequality (Proposition IV.4),

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

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 IV.2). □

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

Proposition A.3 (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 A.2 and A.3, Corollary IV.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 A.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. $\epsilon < \frac{L}{2} \cdot d\left(x_*, x_0\right)^2 \cdot \frac{\tanh(\sqrt{\kappa}D)}{\sqrt{\kappa}D}$.

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

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

Remark A.2. It is important to note that Corollary A.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 §IV.5.

Proof. Let us take $\mu = {}^{\epsilon}\!/_{d(x_*,x_0)^2}$. Then, Proposition A.3 implies that arbitrary $\epsilon/2$-suboptimal solution $x_{\text{sol}} \in M$ to $\underset{x \in M}{\text{minimize}} \left(f(x) + {}^{\mu}\!/_{2} \cdot d\left(x, x_0\right)^2\right)$ satisfies $f(x_{\text{sol}}) - f(x_*) \leq \epsilon$.

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 $c_t \equiv 1/L$, we have

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

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 IV.1) implies that the constant distortion condition holds with $\delta = S_\kappa(2D)$. Hence, Corollary IV.1 implies that (IV.1) 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 $O\left(\epsilon^{-1/2}\log(1/\epsilon)\right)$. □

Bibliography

[ABBC20] Naman Agarwal, Nicolas Boumal, Brian Bullins, and Coralia Cartis. Adaptive regularization with cubics on manifolds. *Mathematical Programming*, pages 1–50, 2020.

[AFGO20] Necdet Serhat Aybat, Alireza Fallah, Mert Gurbuzbalaban, and Asuman Ozdaglar. Robust accelerated gradient methods for smooth strongly convex functions. *SIAM Journal on Optimization*, 30(1):717–751, 2020.

[AMS09] P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009.

[AOBL20] Foivos Alimisis, Antonio Orvieto, Gary Bécigneul, and Aurelien Lucchi. A continuous-time perspective for modeling acceleration in Riemannian optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 1297–1307. PMLR, 2020.

[AOBL21] Foivos Alimisis, Antonio Orvieto, Gary Becigneul, and Aurelien Lucchi. Momentum improves optimization on Riemannian manifolds. In *International Conference on Artificial Intelligence and Statistics*, pages 1351–1359. PMLR, 2021.

[AS20] Kwangjun Ahn and Suvrit Sra. From Nesterov's estimate sequence to Riemannian acceleration. volume 125 of *Conference on Learning Theory (COLT)*, pages 84–118. PMLR, 09–12 Jul 2020.

[AT06] Alfred Auslender and Marc Teboulle. Interior gradient and proximal methods for convex and conic optimization. *SIAM Journal on Optimization*, 16(3):697–725, 2006.

[AZO17] Zeyuan Allen-Zhu and Lorenzo Orecchia. Linear coupling: An ultimate unification of gradient and mirror descent. In *ITCS 2017*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2017.

[Bac14] Miroslav Bacák. *Convex analysis and optimization in Hadamard spaces*, volume 22. Walter de Gruyter GmbH & Co KG, 2014.

[BBB$^+$01] Dmitri Burago, IU D Burago, Yuri Burago, Sergei A Ivanov, and Sergei Ivanov. *A course in metric geometry*, volume 33. American Mathematical Soc., 2001.

[BBC11] Stephen Becker, Jérôme Bobin, and Emmanuel J Candès. NESTA: A fast and accurate first-order method for sparse recovery. *SIAM Journal on Imaging Sciences*, 4(1):1–39, 2011.

[BC11] Heinz H Bauschke and Patrick L Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*, volume 408. Springer, 2011.

[BF95] James V Burke and Michael C Ferris. A Gauss-Newton method for convex composite optimization. *Mathematical Programming*, 71(2):179–194, 1995.

[BFG$^+$19] 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.

[BFM17] 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.

[BG19] Nikhil Bansal and Anupam Gupta. Potential-function proofs for gradient methods. *Theory of Computing*, 15(4):1–32, 2019.

[BH13] Martin R Bridson and André Haefliger. *Metric spaces of non-positive curvature*, volume 319. Springer Science & Business Media, 2013.

[BJW18] Michael Betancourt, Michael I Jordan, and Ashia C Wilson. On symplectic optimization. *arXiv preprint arXiv:1802.03653*, 2018.

[BT09] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

[Car92] Manfredo Perdigao do Carmo. *Riemannian geometry*. Birkhäuser, 1992.

[CEMS01] 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.

[Cha06] Isaac Chavel. *Riemannian geometry: a modern introduction*, volume 98. Cambridge university press, 2006.

[CHVSL18] Saman Cyrus, Bin Hu, Bryan Van Scoy, and Laurent Lessard. A robust accelerated optimization algorithm for strongly convex functions. In *2018 Annual American Control Conference (ACC)*, pages 1376–1381. IEEE, 2018.

[Def19] Aaron Defazio. On the curved geometry of accelerated optimization. In *Advances in Neural Information Processing Systems*, pages 1764–1773, 2019.

[DG19] Damek Davis and Benjamin Grimmer. Proximally guided stochastic subgradient method for nonsmooth, nonconvex problems. *SIAM Journal on Optimization*, 29(3):1908–1930, 2019.

[DO19] 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.

[Dru17] Dmitriy Drusvyatskiy. The proximal point method revisited. *arXiv preprint: 1712.06038*, 2017.

[DT14] 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.

[DVW15] Ramsay Dyer, Gert Vegter, and Mathijs Wintraecken. Riemannian simplices and triangulations. *Geometriae Dedicata*, 179(1):91–138, 2015.

[GN18] Alexander Vladimirovich Gasnikov and Yu E Nesterov. Universal method for stochastic composite optimization problems. *Computational Mathematics and Mathematical Physics*, 58(1):48–64, 2018.

[Gro78] Mikhail Gromov. Manifolds of negative curvature. *Journals of Differential Geometry*, 13(2):223–230, 1978.

[GS19] Navin Goyal and Abhishek Shetty. Sampling and optimization on convex sets in Riemannian manifolds of non-negative curvature. In *COLT*, pages 1519–1561, 2019.

[Gül91] Osman Güler. On the convergence of the proximal point algorithm for convex minimization. *SIAM Journal on Control and Optimization*, 29(2):403–419, 1991.

[HL17] Bin Hu and Laurent Lessard. Dissipativity theory for Nesterov's accelerated method. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1549–1557. JMLR. org, 2017.

[HM21] Linus Hamilton and Ankur Moitra. No-go theorem for acceleration in the hyperbolic plane. *arXiv preprint arXiv:2101.05657*, 2021.

[Jos08] Jürgen Jost. *Riemannian geometry and geometric analysis*, volume 42005. Springer, 2008.

[KBB15] Walid Krichene, Alexandre Bayen, and Peter L Bartlett. Accelerated mirror descent in continuous and discrete time. In *Advances in Neural Information Processing Systems*, pages 2845–2853, 2015.

[KF16] Donghwan Kim and Jeffrey A Fessler. Optimized first-order methods for smooth convex minimization. *Mathematical programming*, 159(1-2):81–107, 2016.

[KSM16] Hiroyuki Kasai, Hiroyuki Sato, and Bamdev Mishra. Riemannian stochastic variance reduced gradient on Grassmann manifold. *arXiv preprint arXiv:1605.07367*, 2016.

[LMH15] Hongzhou Lin, Julien Mairal, and Zaid Harchaoui. A universal catalyst for first-order optimization. In *Advances in Neural Information Processing Systems*, pages 3384–3392, 2015.

[LRP16] 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.

[LRS13] Yin Tat Lee, Satish Rao, and Nikhil Srivastava. A new approach to computing maximum flows using electrical flows. In *Proceedings of ACM STOC*, pages 755–764, 2013.

[LSC$^+$17] Yuanyuan Liu, Fanhua Shang, James Cheng, Hong Cheng, and Licheng Jiao. Accelerated first-order methods for geodesically convex optimization on Riemannian manifolds. In *Advances in Neural Information Processing Systems*, pages 4868–4877, 2017.

[LW16] Adrian S Lewis and Stephen J Wright. A proximal method for composite minimization. *Mathematical Programming*, 158(1-2):501–546, 2016.

[Mar70] Bernard Martinet. Régularisation d'inéquations variationnelles par approximations successives. rev. française informat. *Recherche Opérationnelle*, 4:154–158, 1970.

[MOP20] Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. In *International Conference on Artificial Intelligence and Statistics*, pages 1497–1507. PMLR, 2020.

[Mor65] Jean-Jacques Moreau. Proximité et dualité dans un espace Hilbertien. *Bulletin de la Société mathématique de France*, 93:273–299, 1965.

[MR20] David Martínez-Rubio. Global Riemannian acceleration in hyperbolic and spherical spaces. *arXiv preprint arXiv:2012.03618*, 2020.

[Nes83] Yurii Nesterov. A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$. In *Doklady AN USSR*, volume 269, pages 543–547, 1983.

[Nes07] Yu Nesterov. Modified Gauss-Newton scheme with worst case guarantees for global performance. *Optimisation methods and software*, 22(3):469–483, 2007.

[Nes18] Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

[Per95] G. Perelman. Spaces with curvature bounded below. In *Proceedings of the International Congress of Mathematicians*, pages 517–525, Basel, 1995. Birkhäuser Basel.

[Rau51] Harry Ernest Rauch. A contribution to differential geometry in the large. *Annals of Mathematics*, pages 38–55, 1951.

[Roc76] R Tyrrell Rockafellar. Monotone operators and the proximal point algorithm. *SIAM journal on control and optimization*, 14(5):877–898, 1976.

[SBC16] Weijie Su, Stephen Boyd, and Emmanuel J Candes. A differential equation for modeling Nesterov's accelerated gradient method: theory and insights. *JMLR*, 17(1):5312–5354, 2016.

[SDSJ19] Bin Shi, Simon S Du, Weijie J Su, and Michael I Jordan. Acceleration via symplectic discretization of high-resolution differential equations. *arXiv preprint arXiv:1902.03694*, 2019.

[SJFB18] 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.

[SMDH13] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *ICML*, pages 1139–1147, 2013.

[TB19] Adrien Taylor and Francis Bach. Stochastic first-order methods: non-asymptotic and computer-aided analyses via potential functions. In *Conference on Learning Theory*, pages 2934–2992. PMLR, 2019.

[Teb18] Marc Teboulle. A simplified view of first order methods for optimization. *Mathematical Programming*, 170(1):67–96, 2018.

[Tse08] Paul Tseng. On accelerated proximal gradient methods for convex-concave optimization. *submitted to SIAM Journal on Optimization*, 2008.

[TVSL18] Adrien Taylor, Bryan Van Scoy, and Laurent Lessard. Lyapunov functions for first-order methods: tight automated convergence guarantees. In *International Conference on Machine Learning*, pages 4897–4906. PMLR, 2018.

[Udr94] Constantin Udriste. *Convex functions and optimization methods on Riemannian manifolds*, volume 297. Springer Science & Business Media, 1994.

[WRJ16] Ashia C Wilson, Benjamin Recht, and Michael I Jordan. A Lyapunov analysis of momentum methods in optimization. *arXiv preprint: 1611.02635*, 2016.

[WS19] Melanie Weber and Suvrit Sra. Nonconvex stochastic optimization on manifolds via Riemannian Frank-Wolfe methods. *arXiv preprint arXiv:1910.04194*, 2019.

[WWJ16] Andre Wibisono, Ashia C Wilson, and Michael I Jordan. A variational perspective on accelerated methods in optimization. *PNAS*, 113(47):E7351–E7358, 2016.

[ZRS16] Hongyi Zhang, Sashank J Reddi, and Suvrit Sra. Riemannian SVRG: Fast stochastic optimization on Riemannian manifolds. In *Advances in Neural Information Processing Systems*, pages 4592–4600, 2016.

[ZS16] Hongyi Zhang and Suvrit Sra. First-order methods for geodesically convex optimization. In *Conference on Learning Theory*, pages 1617–1638, 2016.

[ZS18] Hongyi Zhang and Suvrit Sra. An estimate sequence for geodesically convex optimization. In *Conference On Learning Theory*, pages 1703–1723, 2018.

[ZYYF19] Pan Zhou, Xiaotong Yuan, Shuicheng Yan, and Jiashi Feng. Faster first-order methods for stochastic non-convex optimization on Riemannian manifolds. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[ZZS18] Jingzhao Zhang, Hongyi Zhang, and Suvrit Sra. R-SPIDER: A fast Riemannian stochastic optimization algorithm with curvature independent rate. *arXiv preprint arXiv:1811.04194*, 2018.