A Continuous-time Perspective for Modeling Acceleration in Riemannian Optimization

Foivos Alimisis **Antonio Orvieto** **Gary Bécigneul** **Aurelien Lucchi**
ETH Zürich, Switzerland

Abstract

We propose a novel second-order ODE as the continuous-time limit of a Riemannian accelerated gradient-based method on a manifold with curvature bounded from below. This ODE can be seen as a generalization of the ODE derived for Euclidean spaces, and can also serve as an analysis tool. We study the convergence behavior of this ODE for different classes of functions, such as geodesically convex, strongly-convex and weakly-quasi-convex. We demonstrate how such an ODE can be discretized using a semi-implicit and Nesterov-inspired numerical integrator, that empirically yields stable algorithms which are faithful to the continuous-time analysis and exhibit accelerated convergence.

1 Introduction

A core problem in machine learning is finding a minimum of a function $f : H \to \mathbb{R}$. In the vast majority of machine learning applications, H represents either a Euclidean space or a Riemannian manifold. Among the most popular types of methods to optimize f are first-order methods, such as gradient descent which simply updates a sequence of iterates $\{x_k\}$ by stepping in the opposite direction of the gradient $\nabla f(x_k)$. In the case $H = \mathbb{R}^n$, gradient descent as a first-order method has been shown to achieve a suboptimal convergence rate. In a seminal paper (Nesterov, 1983), Nesterov showed that one can construct an optimal – a.k.a. accelerated – algorithm that achieves faster rates of convergence for both convex and strongly-convex functions. The convergence analysis of this algorithm relies heavily on the linear structure of the space H and it is not until recently that a first adaptation to Riemannian spaces has been derived in (Zhang and Sra, 2018). Their algorithm is shown to obtain an accelerated rate of convergence for *geodesically* strongly-convex functions. These functions are of particular interest as they are non-convex in the Euclidean sense and they occur in some fundamental problems, see e.g. (Zhang and Sra, 2016, 2018).

In this manuscript, we take a different direction from previous works that have focused on analyzing the discrete-time form of Nesterov acceleration. We instead derive a continuous-time model that generalizes the work of Su et al. (2014) to non-Euclidean spaces. The resulting second-order ODE is shown to exhibit an approximate equivalence to Nesterov acceleration, and can therefore be used as an analysis tool. We prove theoretically that the continuous-time process corresponding to the derived differential equation has an accelerated rate of convergence for various types of functions. As in (Su et al., 2014), one can also obtain different discrete-time algorithms from such an ODE. We here focus on a discretization scheme that we show empirically to yield an accelerated rate of convergence.

In summary, our main contributions are:

- We derive a second-order differential equation that can serve as an analysis tool for a Riemannian variant of accelerated gradient descent.
- We analyze the convergence behavior of this ODE for three different types of functions: geodesically convex, strongly-convex and weakly-quasi-convex.
- As a byproduct of our convergence analysis, we establish some new technical results about the Hessian of the Riemannian distance function. These results could be of general interest.
- We prove that in the case of Riemannian gradient descent applied to geodesically strongly convex functions, the discrete and continuous trajectories remain close. The extension of this result to an accelerated method is however non-trivial.
- We provide empirical results on several problems of interest in order to confirm the validity of our theoretical analysis and discretization scheme.

Proceedings of the 23^{rd}International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108. Copyright 2020 by the author(s).

2 Related work

Accelerated Gradient Descent/Flow. The first *practical* accelerated algorithm in a vector space is due to (Nesterov, 1983). Since then, the community has shown a deep interest in understanding the mechanism underlying acceleration. A recent trend has been to look at acceleration from a continuous-time viewpoint. In such a framework, accelerated gradient descent is seen as the discretization of a second-order ODE. In (Su et al., 2014), a second order differential equation to capture the dynamics of the classical algorithm from Nesterov in the convex case is formulated. In (Wibisono et al., 2016), study continuous accelerated dynamics introducing the concept of Bregman Lagrangian. In (Wilson et al., 2016), substitute the classical estimate sequences technique by a family of Lyapunov functions in both discrete and continuous time. In (Shi et al., 2018), Shi et al. show that differential equations are rough approximators of real learning dynamics, i.e. a given algorithm can generate many continuous models. Finally, the same authors showed in (Shi et al., 2019) that symplectic integration (see (Hairer et al., 2006)) has deep links to Nesterov's method.

Riemannian optimization. Research in the field of Riemannian optimization has recently encountered a lot of interest. A seminal book in the field is (Absil et al., 2009) who gives a comprehensive review of many standard optimization methods except accelerated methods. More recently, Zhang and Sra (2016) proved convergence rates for Riemannian gradient descent applied to the class of geodesically convex functions. Acceleration in a Riemannian framework was discussed by Liu et al. (2017) who claimed to have designed Riemannian accelerated methods with guaranteed convergence rates but as discussed in (Zhang and Sra, 2018), their method relies on finding the exact solution to a nonlinear equation and it is not clear how difficult this problem is. Subsequently, Zhang and Sra (2018) developed the first computationally tractable accelerated algorithm on a Riemannian manifold, but their approach only has provable convergence for geodesically strongly-convex objectives. In contrast, we here address the problem of achieving acceleration for the *weaker* class of weakly-quasi-convex objective functions.

3 Background

We review some basic notions from Riemannian geometry that are required in our analysis. For a full review, we refer the reader to a classical textbook, for instance (Spivak, 1979).

Manifolds. A differentiable manifold M is a topological space that is locally Euclidean. This means that for any point $x \in M$, we can find a neighborhood that is diffeomorphic to an open subset of some Euclidean space. This Euclidean space can be proved to have the same dimension, regardless of the chosen point, called the dimension of the manifold. A Riemannian manifold (M, g) is a differentiable manifold equipped with a Riemannian metric g_x, i.e. an inner product for each tangent space T_xM at $x \in M$. We denote the inner product of $u, v \in T_xM$ with $\langle u, v\rangle_x$ or just $\langle u, v\rangle$ when the tangent space is obvious from context. Similarly we consider the norm as the one induced by the inner product at each tangent space.

Geodesics Geodesics are curves $\gamma : [0, 1] \to M$ of constant speed and of (locally) minimum length. They can be thought of as the Riemannian generalization of straight lines in Euclidean spaces. Geodesics are used to construct the exponential map $\exp_x : T_xM \to M$, defined by $\exp_x(v) = \gamma(1)$, where γ is the unique geodesic such that $\gamma(0) = x$ and $\dot{\gamma}(0) = v$. The exponential map is locally a diffeomorphism. Using the notion of geodesics, we can define an intrinsic distance d between two points in the Riemannian manifold M, as the infimum of lengths of geodesics that connect these two points. Geodesics also provide a way to transport vectors from one tangent space to another. This operation called parallel transport is usually denoted by $\Gamma_x^y : T_xM \to T_yM$. Closely linked to geodesics is the notion of injectivity radius. Given a point $x \in M$, we define the injectivity radius at x (denoted $\mathrm{inj}(x)$), the radius of the biggest ball around x, where the exponential map $\exp_x$ is a diffeomorphism. We denote the inverse of the exponential map inside this ball by $\log_x$.

Vector fields and covariant derivative. The correct notion to capture second order changes on a Riemannian manifold is called covariant differentiation and it is induced by the fundamental property of Riemannian manifolds to be equipped with a connection. The fact that a connection can always be defined in a Riemannian manifold is the subject of the fundamental theorem of Riemannian geometry. We are interested in a specific type of connection, called the Levi-Civita connection, which induces a specific type of covariant derivative. For our purpose, it will however be sufficient to define the notion of covariant derivative using the (simpler) notion of parallel transport. First, we state the definition of a vector field on a Riemannian manifold.

Definition 1. *Let M be a Riemannian manifold. A vector field X in M is a smooth map $X : M \to \mathcal{T}M$, where $\mathcal{T}M$ is the tangent bundle, i.e. the collection of all tangent vectors in all tangent spaces of M, such that $p \circ X$ is the identity (p is the projection from $\mathcal{T}M$ to M).*

One can see a vector field as an infinite collection of imaginary curves, the so-called integral curves (formally they are solutions of first-order differential equations on M).

Definition 2. *Given two vector fields X, Y in a Riemannian manifold M, we define the covariant derivative of B along A to be*

$$\nabla_X Y(p) := \lim_{h\to 0} \frac{\Gamma_{\gamma(h)}^{\gamma(0)} Y(\gamma(h)) - Y(p)}{h},$$

with γ the unique integral curve of A passing from p.

Geodesic convexity. We remind the reader of the basic definitions needed in Riemannian optimization.

Definition 3. *A subset $A \subseteq M$ of a Riemannian manifold M is called geodesically uniquely convex, if every two points in A are connected by a unique geodesic.*

Definition 4. *A function $f : M \to \mathbb{R}$ is called geodesically convex, if $f(\gamma(t)) \leq (1-t)f(p) + tf(q)$, for $t \in [0,1]$, where γ is any geodesic connecting $p, q \in M$.*

Given a function $f : M \to \mathbb{R}$, the notions of differential and (Riemannian) inner product allow us to define the Riemannian gradient of f at $x \in M$, which is a tangent vector belonging to the tangent space based at x, $T_x M$.

Definition 5. *The Riemannian gradient* gradf *of a (real-valued) function $f : M \to \mathbb{R}$ at a point $x \in M$, is the tangent vector at x, such that $\langle \mathrm{gradf}(x), u\rangle = df(x)u$ [1], for any $u \in T_x M$.*

Given the notion of Riemannian gradient and covariant derivative we can define the notion of Riemannian Hessian.

Definition 6. *Given vector fields A, B in M, we define the Hessian operator of f to be*

$$\mathrm{Hess}(f)(A, B) := \langle \nabla_A \operatorname{grad} f, B\ \rangle.$$

Using the Riemannian inner product and the Riemannian gradient, we can formulate an equivalent definition for geodesic convexity for a smooth function f defined in a geodesically uniquely convex domain A (the inverse of the exponential map is well-defined).

Proposition 1. *Let a smooth, geodesically convex function $f : A \to \mathbb{R}$. Then, for any $x, y \in A$,*

$$f(x) - f(y) \geq \langle \mathrm{gradf}(y), \log_y(x)\rangle.$$

As in the Euclidean case, any local minimum of a geodesically convex function is a global minimum.
In a similar manner we can define geodesic strong convexity.

Definition 7. *A smooth function $f : A \to \mathbb{R}$ is called geodesically μ-strongly convex, $\mu > 0$, if $\forall x, y \in A$*

$$f(x) - f(y) \geq \langle \mathrm{gradf}(y), \log_y(x)\rangle + \frac{\mu}{2}\|\log_y(x)\|^2.$$

If a function f is geodesically strongly convex with a non-empty set of minima, then there is only one minimum and it is global.
We now generalize the well-known notion of Euclidean weak-quasi-convexity (see (Guminov and Gasnikov, 2017)) to Riemannian manifolds.

Definition 8. *A function $f : A \to \mathbb{R}$ is called geodesically α-weakly-quasi-convex with respect to $c \in M$, if*

$$\alpha(f(x) - f(c)) \leq -\langle \mathrm{gradf}(x), \log_x(c)\rangle$$

for some fixed $\alpha \in (0,1]$ and any $x \in M$.

It is easy to see that weak-quasi-convexity implies that any local minimum of f is also a global minimum.
Using the notion of parallel transport we can define when f is geodesically L-smooth, i.e. has Lipschitz continuous gradient in a suitable differential-geometric way.

Definition 9. *A function $f : M \to \mathbb{R}$ is called L-smooth if $\forall x, y \in M$ and geodesic γ connecting them*

$$\|\mathrm{gradf}(x) - \Gamma_y^x \mathrm{gradf}(y)\| \leq Ll(\gamma),$$

where Γ is the parallel transport along γ and $l(\gamma)$ the length of γ.

Geodesic L-smoothness has similar properties to its Euclidean analogue. Namely, a two times differentiable function is L-smooth, if and only if the norm of its Riemannian Hessian is bounded by L.

Curvature. In this paper, we make the standard assumption that the input space is not "infinitely curved". In order to make this statement rigorous, we need the notion of sectional curvature K, which is a measure of how sharply the manifold is curved (or how "far" from being flat our manifold is), "two-dimensionally".

4 Hessian of the distance function

Before discussing the design and analysis of accelerated flows on manifolds, it is necessary to derive a crucial geometric result. During a first read, the reader may skip this section or return to it later to understand some of the technicalities in Section 5.

In Euclidean spaces, the law of cosines relates the lengths of the sides of a triangle to the cosine of one of its angles. One can also adapt this result to non-linear spaces as we will demonstrate next. We first derive a lemma that provides a bound on the Hessian of a

[1] df denotes the differential of f, i.e. $df(x)[u] = \lim_{t\to 0} \frac{f(c(t)) - f(x)}{t}$, where $c : I \to M$ is a smooth curve such that $c(0) = x$ and $\dot{c}(0) = u$.

variant of the the Riemannian squared distance function $-\frac{1}{2}d(X,p)^2$ for the curve $X : I \to M$ and $p \in M$. Alternatively, the Hessian of $-\frac{1}{2}d(X,p)^2$ can be seen as the covariant derivative of $\log_{X(t)}(p)$.

Lemma 2. *For a Riemannian manifold M with curvature bounded above by $K_{\max}$ and below by $K_{\min}$ and* $\operatorname{diam}(M) \leq D < \begin{cases} \frac{\pi}{\sqrt{K_{\max}}} & , K_{\max} > 0 \\ \infty & , K_{\max} \leq 0 \end{cases}$, *we have that*

$$\delta \|\dot{X}\|^2 \leq \langle \nabla_{\dot{X}} \log_X(p), -\dot{X} \rangle \leq \zeta \|\dot{X}\|^2,$$

where

$$\delta := \begin{cases} 1 & , K_{\max} \leq 0 \\ \sqrt{K_{\max}} d(X,p) \cot(\sqrt{K_{\max}} d(X,p)) & , K_{\max} > 0 \end{cases}$$

and

$$\zeta := \begin{cases} \sqrt{-K_{\min}} d(X,p) \coth(\sqrt{-K_{\min}} d(X,p)) & , K_{\min} < 0 \\ 1 & , K_{\min} \geq 0 \end{cases}.$$

Corollary 2.1. *Let a geodesic triangle Δabc in a Riemannian manifold M of curvature bounded above by $K_{\max}$ and $\operatorname{diam}(M) \leq D$. We denote be B the angle between the edges ab and bc. If $K_{\max} > 0$, we assume in addition that $D < \frac{\pi}{\sqrt{K_{\max}}}$. Then*

$$(ac)^2 \geq \delta (bc)^2 + (ab)^2 - 2(ab)(bc)\cos(B)$$

where δ is defined as

$$\delta = \begin{cases} 1 & , K_{\max} \leq 0 \\ \sqrt{K_{\max}} d(q,a) \cot(\sqrt{K_{\max}} d(q,a)) & , K_{\max} > 0 \end{cases}$$

for some $q \in M$ along the edge bc.

Note that one can also recover Lemma 5 in (Zhang and Sra, 2016) as a corollary of Lemma 2.

Properties of the cost as function of curvature. Given a geodesically uniquely convex subset $A \subset M$ and $p \in A$, we consider two points $x, y \in A$. We are interested in bounding distances in the geodesic triangle Δxyp. Corollary 2.1 states that

$$d(x,p)^2 \geq \delta d(x,y)^2 + d(y,p)^2 - 2\langle \log_y(p), \log_y(x) \rangle$$

Taking into consideration that the gradient of the function $f(x) = d(x,p)^2$ is $\operatorname{gradf}(x) = -2\log_x(p)$, the last inequality is equivalent to

$$f(x) \geq f(y) + \langle \operatorname{gradf}(y), \log_y(x) \rangle + \frac{2\delta}{2} \|\log_x(y)\|^2$$

As shown in the appendix, this inequality is tight in the spherical case. This inequality also means that f is either geodesically 2δ-strongly convex, convex (but not strongly-convex) or not convex, if $\delta > 0$, $\delta = 0$, or $\delta < 0$ respectively. The first case happens, when $d(x,p) < \frac{\pi}{2\sqrt{K_{\max}}}$, the second when $d(x,p) = \frac{\pi}{2\sqrt{K_{\max}}}$ and the third when $\frac{\pi}{2\sqrt{K_{\max}}} < d(x,p) < \frac{\pi}{\sqrt{K_{\max}}}$. However, note that the function f is always 1-weakly-quasi-convex with respect to its global minimizer p. Indeed, from the definition $f(x) = d(x,p)^2$, we have $f(x) - f(p) = \|\log_x(p)\|^2$ and $-\langle \operatorname{gradf}(x), \log_x(p) \rangle = 2\langle \log_x(p), \log_x(p) \rangle = 2\|\log_x(p)\|^2$, which combined gives us $f(x) - f(p) \leq -\langle \operatorname{gradf}(x), \log_x(p) \rangle$.

Example for a sphere. Consider a manifold M as a sphere with constant curvature K. As a geodesically uniquely convex domain A, we take the ball $B_r(p)$ centered at $p \in A$ and with radius r. If $r < \frac{\pi}{2\sqrt{K}}$, then $\delta > 0$, while if $r = \frac{\pi}{2\sqrt{K}}$ (i.e. A is an open hemisphere), then $\delta = 0$. The problem of minimizing $f(x) = d(x,p)^2$ is therefore either geodesically strongly-convex or geodesically convex depending on the value of r. Alternatively, if we choose to construct our geodesically uniquely convex domain A as an open hemisphere with $p \in A$ not at the center, then there are points with distance from p more than $\frac{\pi}{2\sqrt{K}}$. Thus δ is negative and f is not geodesically convex. Given that $f(x) = d(x,p)^2$ is always 1-weakly-quasi-convex, the problem of minimizing f is weakly-quasi-convex but not convex.

Duality smoothness/convexity. Lemma 5 in (Zhang and Sra, 2016) states that the function $f(x) = d(x,p)^2$ is 2ζ-smooth. This shows that there is some sort of duality between convexity and smoothness with respect to the curvature of the manifold. For a given function $d(x,p)^2$, a smaller curvature makes the function more convex while also making it less smooth.

5 Accelerated flows

Recall that the problem that we investigate is minimizing a function $f : M \to \mathbb{R}$. A fundamental algorithm to solve this problem is Riemannian gradient descent (RGD), which takes the form $x_{k+1} = \exp_{x_k}(-\eta \operatorname{gradf}(x_k))$, where $\eta > 0$ is the so-called learning rate. The convergence properties of this method, extensively explored in (Zhang and Sra, 2016), can be successfully studied (see (Munier, 2007) and the appendix) by the means of its continuous-time limit $\dot{X} + \operatorname{gradf}(X) = 0$.

In contrast, we are not aware of any prior work investigating the continuous-time formulation of an accelerated method. Hence, taking inspiration from the seminal work of Su et al. (2014), we consider the following differential equation to model acceleration:

$$\boxed{\nabla \dot{X} + c\dot{X} + \operatorname{gradf}(X) = 0} \qquad \text{(RNAG-ODE)}$$

For the convex and weakly-quasi-convex cases, we choose $c := c(t) = \frac{v}{t}$, where v is a constant to be determined later. From now on, we define ζ as

$$\zeta := \begin{cases} \sqrt{-K_{\min}} D \coth(\sqrt{-K_{\min}} D) & , K_{\min} < 0 \\ 1 & , K_{\min} \geq 0 \end{cases}$$

where D is an upper bound for the working domain. Next, following (Zhang and Sra, 2018), we make the following set of assumptions, which we will keep for the rest of the paper.

Assumptions Given $A \subseteq M$, and $f : M \to \mathbb{R}$,

1. The sectional curvature K inside A is bounded from below, i.e. $K \geq K_{\min}$.
2. M is a complete manifold, such that any two points are connected by some geodesic.
3. A is a geodesically uniquely convex subset of M, such that $\operatorname{diam}(A) \leq D$. The exponential map is globally a diffeomorphism.
4. f is geodesically L-smooth and all its minima are inside A.
5. We have granted access to oracles which compute the exponential and logarithmic maps as well as the Riemannian gradient of f efficiently.
6. All the solutions of our derived differential equations remain inside A.

Note that the first four assumptions are standard in Riemmanian optimization (see, ((Munier, 2007; Zhang and Sra, 2016, 2018))). The fifth assumption is mostly required for computational purpose. The last assumption could potentially be relaxed by relying on a barrier function or a projection step.

5.1 Existence of a solution

For strongly-convex functions, we will choose $c(t)$ to be constant, in which case existence and uniqueness of the solution can be shown to hold globally due to completeness of M.

When $c(t) = \frac{v}{t}$, the proof is not as simple and involves the use of the Arzela-Ascoli theorem for sequences of curves on Riemannian manifolds, in a similar vein as in (Su et al., 2014). However, we cannot guarantee the uniqueness of the solution. The proof is provided in the appendix.

Lemma 3. *The differential equation*

$$\nabla \dot{X} + \frac{v}{t}\dot{X} + \operatorname{gradf}(X) = 0 \tag{1}$$

where v is a positive constant, has a global solution $X : [0, \infty) \to M$ under the initial conditions $X(0) = x_0 \in A$ and $\dot{X}(0) = 0$.

The proof relies on the following result that might be of independent interest and is close to the fundamental theorem of calculus for vector fields on Riemannian manifolds.

Lemma 4. *Consider a vector field A along the smooth curve $X : [a, b] \to M$ in a Riemannian manifold M. Then*

$$\Gamma_{X(b)}^{X(a)} A(b) - A(a) = \int_a^b \Gamma_{X(t)}^{X(a)} \nabla A(t) dt$$

where Γ is the parallel transport along the curve X.

5.2 The convex case

Now we are ready to analyze the convergence rate of the solutions of Eq. 1, starting from a point $X(0) \in A$, to a minimizer x^* of a geodesically convex function f.

Theorem 5. *Let f be a geodesically convex function. Any solution of the differential equation*

$$\nabla \dot{X} + \frac{1 + 2\zeta}{t}\dot{X} + \operatorname{gradf}(X) = 0 \tag{2}$$

converges to a minimizer x^* of f with rate

$$f(X) - f(x^*) \leq \frac{2\zeta \|\log_{x_0}(x^*)\|^2}{t^2} \quad (t > 0).$$

Proof sketch. The proof is done by showing that the following Lyapunov function is decreasing:

$$\epsilon(t) = t^2(f(X) - f(x^*)) + 2\| - \log_X(x^*) + \frac{t}{2}\dot{X}\|^2 \\ +2(\zeta - 1)\|\log_X(x^*)\|^2.$$

The novelty compared to (Su et al., 2014) is the last curvature-dependent summand. Complete proof in the appendix. □

5.3 The weakly-quasi-convex case

For α-weakly-quasi convex functions, we have the following result.

Theorem 6. *Let f be a geodesically α-weakly-quasi-convex function. Any solution of the differential equation*

$$\nabla \dot{X} + \frac{1 + \frac{2}{\alpha}\zeta}{t}\dot{X} + \operatorname{gradf}(X) = 0 \tag{3}$$

converges to a minimizer x^* of f with rate

$$f(X) - f^* \leq \frac{2\zeta \|\log_{x_0}(x^*)\|^2}{\alpha^2 t^2} \quad (t > 0).$$

The proof is similar to the one of the convex case and can be found in the appendix. Note here that α can be larger than 1. An important specific case is the Riemannian squared distance $d(x, p)^2$, where $\alpha = 2$.

5.4 The strongly-convex case

Recall that we have a constant friction term for strongly-convex functions, which yields an ODE similar to Equation 7 in (Wilson et al., 2016) for the Euclidean case.

Theorem 7. *Let f be a geodesically μ-strongly convex function. The solution of the differential equation*

$$\nabla\dot{X} + \left(\frac{1}{\sqrt{\zeta}} + \sqrt{\zeta}\right)\sqrt{\mu}\dot{X} + \mathrm{gradf}(X) = 0 \tag{4}$$

converges to a minimizer x^* of f with rate

$$f(X) - f^* \leq \frac{\frac{\mu}{2}\|\log_{x_0}(x)\|^2 + f(x_0) - f^*}{e^{\sqrt{\frac{\mu}{\zeta}}t}} \quad (t > 0).$$

Proof sketch. The proof (see appendix) shows that the following energy function is monotonically decreasing:

$$\epsilon(t) = e^{\sqrt{\frac{\mu}{\zeta}}t}\Big(\frac{\mu}{2\zeta}\| - \log_X(x^*) + \sqrt{\zeta/\mu}\dot{X}\|^2$$

$$f(X) - f^* + \frac{\mu(\zeta-1)}{2\zeta}\|\log_X(x^*)\|^2\Big).$$

□

Note that the constant $\sqrt{\zeta} + \frac{1}{\sqrt{\zeta}}$ is always greater or equal than 2 and equality holds only when $\zeta = 1$, in which case we recover the Euclidean formulation.

5.5 Comparison to the Euclidean case

Compared to the ODE of (Su et al., 2014), the second derivative of the curve X has been substituted with the covariant derivative of the vector field $\dot{X}$. This is the usual intrinsic way to capture second order changes on manifolds. The Lyapunov functions chosen in the analysis are such that the covariant derivative arises when taking its derivative, which explains why the results derived in Section 4 are needed in our analysis. Also interesting is the effect of the curvature: we note that it is involved in both the friction term of the ODE and in the convergence rates. The positive-curvature case matches the Euclidean one, while the negative-curvature case yields worse constants in terms of theoretical guarantees. This seems to validate the intuition that convergence is easier in spaces with larger curvature, which is also consistent with the results of (Zhang and Sra, 2016).

6 Discretization

We now design and test a Nesterov-inspired semi-implicit integration scheme that translates the ODEs above into implementable accelerated optimization methods. Starting from the ODE $\nabla\dot{X} + \alpha(t)\dot{X} + \mathrm{gradf}(X) = 0$ and following the Euclidean modus operandi of (Shi et al., 2019; Betancourt et al., 2018), our first step is to introduce a velocity variable $V = \dot{X}$. Hence, we can write $\nabla V = -\alpha(t)V - \mathrm{gradf}(X)$.

The semi-implicit Euler method in Euclidean spaces is a numerical integrator tailored to second-order ODEs, which leverages on the velocity/position decomposition and is widely used in physics because of its energy and volume conservation properties, that in turn imply good stability and small integration errors, (Hairer et al., 2006). This scheme consists of a standard forward-Euler update on the velocity variable v_k, followed by an update on the position variable x_k using the just updated value of the velocity, i.e. v_{k+1}. Namely, if $M = \mathbb{R}^d$, we have

$$\begin{cases} v_{k+1} = \beta_k v_k - h\nabla f(x_k) \\ x_{k+1} = x_k + hv_{k+1} \end{cases} \tag{5}$$

where $\beta_k := 1 - h\alpha(kh)$ is the momentum parameter and h is the integration step-size which, if small enough, guarantees [2] $X(kh) \approxeq x_k$. Inspired from the recent success of similar integrators in yielding accelerated algorithms (Shi et al., 2019; Maddison et al., 2018), we next provide a simple adaptation of the semi-implicit method to the Riemannian setting.

Algorithm 1 SIRNAG

```
$x_0 \leftarrow$ random point on $M$;
$v_0 \leftarrow 0 \in T_{x_0}M$;
$h \leftarrow$ some small number $> 0$ (integration step);
**if** geod. strongly-convexity **then**
    $\beta_k \leftarrow 1 - h\frac{(1+\zeta)\sqrt{\mu}}{\sqrt{\zeta}}$;
**else if** geod. weak-quasi-convexity **then**
    $\beta_k \leftarrow \frac{k-1}{k+2\zeta/\alpha}$;
**end if**
**for** $k \geq 0$ **do**
    **Option I**: $a_k \leftarrow \beta_k v_k - h\mathrm{gradf}(x_k)$;
    **Option II**: $a_k \leftarrow \beta_k v_k - h\mathrm{gradf}(\exp_{x_k}(h\beta_k v_k))$
    $x_{k+1} \leftarrow \exp_{x_k}(ha_k)$;
    $v_{k+1} \leftarrow \Gamma_{x_k}^{x_{k+1}} a_k$;
**end for**
```

We start by noting that, since we require $v_k \in T_{x_k}M$ for all k, our method will have to include parallel transport of velocity vectors along the geodesics of the manifold. However, we can postpone this operation to the very end: indeed, if we let $a_k := \beta_k v_k - h\mathrm{gradf}(x_k)$, then $a_k \in T_{x_k}M$ and we can update the position directly using a forward-Euler step: $x_{k+1} = \exp_{x_k}(ha_k)$. To conclude, we need to transport the just used velocity a_k to $T_{x_{k+1}}M$: $v_{k+1} = \Gamma_{x_k}^{x_{k+1}} a_k$.

We summarize the content of the last lines in Algorithm 1 (with Option I) and provide a variant (Option II), inspired by the reformulation of Nesterov's method provided by Sutskever et al. (2013). The latter shows that Equation (5) is exactly Nesterov's method (Nesterov, 2018) once we replace $\nabla f(x_k)$ with

[2] For a *fixed* interval $[0, T]$ with $T = Kh$ $(K \in \mathbb{N})$, we have $\|X(kh) - x_k\| = O(h)$ for all $0 \leq k \leq K$ (Hairer et al., 2006). However, the notation hides an exponential dependency on T: i.e., does not imply shadowing (see Section 7).

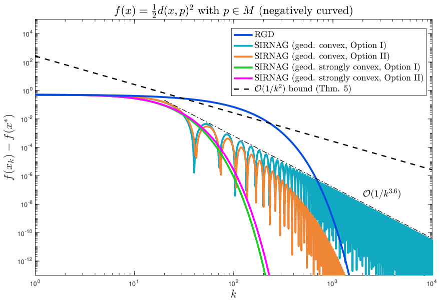

Figure 1: Dynamics of SIRNAG ($h = 0.1$) and RGD ($\eta = h^2$, see footnote 5) on a subset of diameter $D = 1$ of the hyperbolic space $M = \mathbb{H}^2$ ($K = -1$, hence $\zeta = \coth(1) \approxeq 1.313$) equipped with the convex (actually, strongly convex) toy function $f(x) = \frac{1}{2}d(x,p)^2$ for $p \in M$. Plotted is also the bound found in Theorem 5, discretized.

$\nabla f(x_k + h\beta_k v_k)$ (the so-called *corrected gradient*). In our setting, we can similarly use $\text{gradf}(\exp_{x_k}(h\beta_k v_k))$. **As a result, Algorithm 1 with Option II *reduces to Nesterov's method* when $M = \mathbb{R}^d$.**

Experiments. Inspired by the relevance of hyperbolic geometry in machine learning (Zhang et al., 2018; Sra and Hosseini, 2015), we start our empirical study by illustrating some properties of SIRNAG on manifolds with constant *negative curvature*. Fig. 1 shows that our integrator is stable and can achieve, on simple functions, a rate that is actually faster than the prediction of Theorem 5, in perfect agreement with previous observations for similar costs in the Euclidean setting (Zhang et al., 2018; Betancourt et al., 2018). Moreover, as expected, Option II provides a speedup[3] over Option I because it is closer to the original Nesterov's method. Next, to test the tightness of the oracle bound provided by Theorem 5, we use our algorithm to solve a high-dimensional eigenvalue problem. Indeed, the leading unit eigenvector of a symmetric matrix $Q \in \mathbb{R}^{m\times m}$ maximizes $x^T Qx$ over the unit sphere $M = \mathbb{S}^{m-1}$ (constant *positive curvature*). It is well known (Dieuleveut et al., 2017) that such objectives, when $M = \mathbb{R}^m$, are hard to optimize if Q is high-dimensional and ill-conditioned, and are therefore able to truly showcase the acceleration phenomenon[4] for convex but not necessarily strongly convex functions. Fig. 2 shows that this fact translates to the manifold setting: indeed, the suboptimality of SIRNAG decays as $1/k^2$ — as predicted by our continuous-time analysis — in contrast to RGD[5] which behaves like $\mathcal{O}(1/k)$.

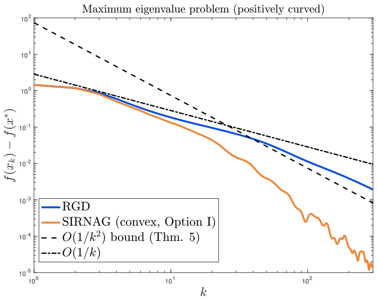

Figure 2: Performance of SIRNAG (convex, i.e. $\beta_k = \frac{k-1}{k+2\zeta}$) against RGD in finding the maximum eigenvalue of a 5-thousand dimensional ill-conditioned matrix. Plotted is also the bound found in Theorem 5, discretized.

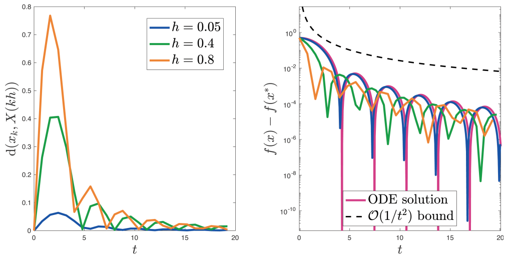

Figure 3: Convergence of SIRNAG (Option I) to the solution of Equation (1), same settings as Fig. 1. Solution to the ODE approximated by SIRNAG (Option I) with an extremely small integration step: $h = 10^{-5}$. The error peak is proportional to the step-size (see next section).

To conclude, as an ultimate test for our discretization procedure, we verify the convergence of SIRNAG to NAG-ODE as $h \to 0$ in Fig. 3. Finally, the code to reproduce the experiments above is available online[6].

7 Shadowing in model spaces

So far, we have shown that the discretization of our second-order ODE empirically exhibits an accelerated rate of convergence and follows the continuous-time limit. The reader might wonder whether any theoretical guarantee can be established to bound the error between the continuous-time and discrete-time process (i.e. predict the results of Fig. 3). In the following, we will show that such guarantees can be obtained for a descent method such as RGD when compared to its limiting ODE (studied in (Munier, 2007)). Further, in the next section, we discuss why the extension to accelerated methods is non-trivial. We will rely on the

[3] Actually Option II in the geodesically strongly-convex case seems a bit slower. This happens because f is of a very particular form, and is well known in the Euclidean literature (see e.g. Proposition 1 in (Lessard et al., 2016)).

[4] Indeed, high dimensional quadratics are used to build lower bounds in convex optimization (Nesterov, 2018).

[5] For RGD we used a stepsize (i.e. a gradient multiplication factor) $\eta \leq 1/\lambda_{\max}(Q)$, where $\lambda_{\max}(Q)$ is the maximum eigenvalue of Q. This is the standard choice in the Euclidean setting, also motivated by the results in (Zhang and Sra, 2016). To get the same gradient multiplication factor and correspondence with the optimal parameters is Nesterov's method, in SIRNAG we choose $h = \sqrt{1/\lambda_{\max}(Q)}$. For further details, we direct the reader to the first few pages of (Su et al., 2014).

[6] `https://github.com/aorvieto/riemann-continuous.git`

shadowing lemma for metric spaces (Ombach, 1993; Brin and Stuck, 2002) and use the contraction property of RGD, as well as common concepts from the theory of dynamical systems (Brin and Stuck, 2002). We briefly review the required definitions and we refer the reader to (Orvieto and Lucchi, 2019) for detailed explanations. We consider a dynamical system on a Riemannian manifold M, i.e. a map $\Psi : M \to M$.

Definition 10. *A sequence $(x_k)_{k=0}^{\infty}$ is an* ***orbit*** *of Ψ if, for all $k \in \mathbb{N}$, $x_{k+1} = \Psi(x_k)$.*

Definition 11. *A sequence $(y_k)_{k=0}^{\infty}$ is a* ***δ−pseudo-orbit*** *of Ψ if, for all $k \in \mathbb{N}$, $d(y_{k+1}, \Psi(y_k)) \leq \delta$.*

Definition 12. *A pseudo-orbit $(y_k)_{k=0}^{\infty}$ of Ψ is* ***ϵ−shadowed*** *if there exists an orbit $(x_k)_{k=0}^{\infty}$ of Ψ such that, for all $k \in \mathbb{N}$, $d(x_k, y_k) \leq \epsilon$.*

In this section, we pick Ψ to be the dynamical system associated with Riemannian gradient descent, which maps x to $\exp_x(-h\mathrm{gradf}(x))$. Its orbit $(x_k)_{k=0}^{\infty}$ is a sequence of iterates returned by RGD. As a candidate pseudo-orbit, we pick $(y_k)_{k=0}^{\infty}$ to the sequence of points derived from the iterative application of φ_h — the time-h flow of the ODE $\dot{y} = -\mathrm{gradf}(y), y(0) = y_0 \in M$), which is itself a dynamical system. The latter sequence represents our ODE approximation of the algorithm Ψ. Our goal in this subsection is to show that, under some conditions, the sequence $(y_k)_{k=0}^{\infty}$ is *close to* an orbit of Ψ, uniformly in k — i.e. that it is shadowed by Ψ. To prove this result, we need a fundamental lemma.

Lemma 8. *(Contraction map shadowing (Ombach, 1993)) Assume that Ψ is uniformly contracting with constant $0 < \rho < 1$. Then, for every $\epsilon > 0$, there exists $\delta > 0$ such that every δ-pseudo-orbit $(y_k)_{k=0}^{\infty}$ of Ψ is ϵ-shadowed by the orbit $(x_k)_{k=0}^{\infty}$ of Ψ starting at $x_0 = y_0$. Moreover, $\delta \leq (1-\rho)\epsilon$.*

To use this result, we first need to prove that the ODE orbit $(y_k)_{k=0}^{\infty}$ is actually a pseudo-orbit of Ψ. This result is standard in numerical analysis, and can be also found (in a less general form) as Proposition 2 in (Absil and Malick, 2012). We assume, in analogy with (Orvieto and Lucchi, 2019), that $f : M \to \mathbb{R}$ is a C^2 function such that[7] for all points on the ODE solution, $\|\mathrm{gradf}(x)\| \leq \ell$ and $\mu \leq \|\mathrm{Hessf}(x)\| \leq L$.

Proposition 9. *There exists a constant C, independent of h but dependent on ℓ, L and the Riemannian structure of M, such that, for any $y_0 \in M$ and $k \in \mathbb{N}$,*

$$d(y_{k+1}, \exp_{y_k}(-h\mathrm{gradf}(y_k))) \leq Ch^2.$$

Last, we need to prove that Ψ is uniformly contracting. We state the result for manifolds of constant curvature K and note that passing to the bounded-curvature case can be done easily by Rauch comparison theorems. We start by defining the following quantities:

$$\zeta := \begin{cases} 1 & , K \geq 0 \\ \sqrt{-K}D\coth(\sqrt{-K}D) & , K < 0 \end{cases}$$

$$\lambda := \begin{cases} 1 & , K \geq 0 \\ \sinh(\sqrt{-K}D)/(\sqrt{-K}D) & , K < 0 \end{cases}$$

Lemma 10. *Let $x_1, x_2 \in M$, where M is a Riemannian manifold of constant curvature K and $\mathrm{diam}(M) \leq D$. If $K > 0$ we further assume that $D < \frac{\pi}{\sqrt{K}}$. Then, for $\xi := \lambda(\zeta - h\mu)$ we have*

$$d(\exp_{x_1}(-h\mathrm{gradf}(x_1)), \exp_{x_2}(-h\mathrm{gradf}(x_2))) \leq \xi d(x_1, x_2),$$

Note that, in the positive curvature case, we recover $\xi = 1 - h\mu$, in analogy with the result of (Orvieto and Lucchi, 2019). Finally we can state our shadowing result, which is now simple application of the contraction map shadowing Lemma.

Theorem 11. *Let $\epsilon > \frac{4C(\lambda\zeta-1)}{\lambda^2\mu^2}$. Any orbit $(y_k)_{k=0}^{\infty}$ of Riemannian gradient flow is ϵ-shadowed by an orbit $(x_k)_{k=0}^{\infty}$ of Riemannian gradient descent, given that $\mu > \frac{\lambda\zeta-1}{\lambda h}$ and*

$$h \leq \min\left\{\left(\frac{\lambda\mu}{2C} + \sqrt{\frac{\lambda^2\mu^2}{4C^2} - \frac{\lambda\zeta-1}{C\epsilon}}\right)\epsilon, \frac{1}{L}\right\}.$$

In the flat and positive-curvature case $\lambda = \zeta = 1$ and we recover Theorem 3 in (Orvieto and Lucchi, 2019).

8 Discussion

We proposed a second-order ODE which gives rise to a family of accelerated methods for weakly-quasi-convex and strongly-convex optimization. Using a modified semi-implicit integration scheme, we derived a cheap iterative Nesterov-inspired algorithm which is numerically stable and empirically achieves an accelerated rate of convergence for optimization problems defined over manifolds, under both positive and negative curvature. As future work, it would be desirable to establish a general shadowing theory for the second-order ODE we studied, in order to guarantee that the discretization error can be provably kept under control. As a first step towards such an ambitious goal, we derived a shadowing result for Riemannian gradient descent. We note that, as also noted by Orvieto and Lucchi (2019), the main difficulty in the construction of such a result for accelerated algorithms is the mysterious lack of contraction of momentum methods, which are notoriously non-descending and heavily oscillating. Finally, the continuous-time representation derived in this manuscript might serve for other applications, such as analyzing the escape speed from saddle points (Criscitiello and Boumal, 2019; Sun et al., 2019) or for speeding-up the optimization of nonconvex functions as in (Carmon et al., 2017).

[7] This easily holds if f is geodesically μ-strongly convex and L-smooth. In this case, ℓ depends on the initial condition y_0 and on L.

Acknowledgements The authors would like to thank professors Urs Lang and Nicolas Boumal for helpful discussions regarding the content of this work. Gary Bécigneul is funded by the Max Planck ETH Center for Learning Systems.

References

P-A Absil and Jérôme Malick. Projection-like retractions on matrix manifolds. *SIAM Journal on Optimization*, 22(1):135–158, 2012.

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

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

Michael Brin and Garrett Stuck. *Introduction to dynamical systems.* Cambridge university press, 2002.

Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Convex until proven guilty: Dimension-free acceleration of gradient descent on non-convex functions. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 654–663. JMLR. org, 2017.

Christopher Criscitiello and Nicolas Boumal. Efficiently escaping saddle points on manifolds. In *Advances in Neural Information Processing Systems*, pages 5985–5995, 2019.

Aymeric Dieuleveut, Nicolas Flammarion, and Francis Bach. Harder, better, faster, stronger convergence rates for least-squares regression. *The Journal of Machine Learning Research*, 18(1):3520–3570, 2017.

Sergey Guminov and Alexander Gasnikov. Accelerated methods for α-weakly-quasi-convex problems. *arXiv preprint arXiv:1710.00797*, 2017.

Ernst Hairer, Christian Lubich, and Gerhard Wanner. *Geometric numerical integration: structure-preserving algorithms for ordinary differential equations*, volume 31. Springer Science & Business Media, 2006.

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

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

Chris J Maddison, Daniel Paulin, Yee Whye Teh, Brendan O'Donoghue, and Arnaud Doucet. Hamiltonian descent methods. *arXiv preprint arXiv:1809.05042*, 2018.

Julien Munier. Steepest descent method on a riemannian manifold: the convex case. *Balkan Journal of Geometry & Its Applications*, 12(2), 2007.

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

Yurii E Nesterov. A method for solving the convex programming problem with convergence rate $o(1/k^2)$. In *Dokl. akad. nauk Sssr*, volume 269, pages 543–547, 1983.

Jerzy Ombach. The simplest shadowing. In *Annales Polonici Mathematici*, volume 58, pages 253–258, 1993.

Antonio Orvieto and Aurelien Lucchi. Shadowing properties of optimization algorithms. In *Advances in Neural Information Processing Systems*, pages 12671–12682, 2019.

Bin Shi, Simon S Du, Michael I Jordan, and Weijie J Su. Understanding the acceleration phenomenon via high-resolution differential equations. *arXiv preprint arXiv:1810.08907*, 2018.

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.

Michael Spivak. *A Comprehensive Introduction to Differential Geometry*, volume 1 of *10*. Publish or perish, 2 edition, 1979. ISBN 0914098837.

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

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

Yue Sun, Nicolas Flammarion, and Maryam Fazel. Escaping from saddle points on riemannian manifolds. *arXiv preprint arXiv:1906.07355*, 2019.

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147, 2013.

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

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

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

Hongyi Zhang and Suvrit Sra. Towards riemannian accelerated gradient methods. *arXiv preprint arXiv:1806.02812*, 2018.

Jingzhao Zhang, Aryan Mokhtari, Suvrit Sra, and Ali Jadbabaie. Direct runge-kutta discretization achieves acceleration. In *Advances in Neural Information Processing Systems*, pages 3900–3909, 2018.