
Accelerated Gradient Methods for Geodesically Convex Optimization: Tractable Algorithms and Convergence Analysis

Jungbin Kim¹ Insoon Yang¹

Abstract

We propose computationally tractable accelerated first-order methods for Riemannian optimization, extending the Nesterov accelerated gradient (NAG) method. For both geodesically convex and geodesically strongly convex objective functions, our algorithms are shown to have the same iteration complexities as those for the NAG method on Euclidean spaces, under only standard assumptions. To the best of our knowledge, the proposed scheme is the first fully accelerated method for geodesically convex optimization problems. Our convergence analysis makes use of novel metric distortion lemmas as well as carefully designed potential functions. A connection with the continuous-time dynamics for modeling Riemannian acceleration in (Alimisis et al., 2020) is also identified by letting the stepsize tend to zero. We validate our theoretical results through numerical experiments.

1. Introduction

We consider Riemannian optimization problems of the form

$$\min_{x \in N \subseteq M} f(x), \quad (1)$$

where M is a Riemannian manifold, N is an open geodesically uniquely convex subset of M , and $f : N \rightarrow \mathbb{R}$ is a continuously differentiable *geodesically convex* function. Geodesically convex optimization is the Riemannian version of convex optimization and has salient features such as every local minimum being a global minimum. More interestingly, some (constrained) nonconvex optimization problems defined in the Euclidean space can be considered geodesically convex optimization problems on appropriate

Riemannian manifolds (Vishnoi, 2018, Section 1). Geodesically convex optimization has a wide range of applications, including covariance estimation (Wiesel, 2012), Gaussian mixture models (Hosseini & Sra, 2015; 2020), matrix square root computation (Sra, 2015), metric learning (Zadeh et al., 2016), and optimistic likelihood calculation (Nguyen et al., 2019). See (Zhang & Sra, 2016, Section 1.1) for more examples.

The iteration complexity theory for first-order algorithms is well known when $M = \mathbb{R}^n$. Given an initial point x_0 , gradient descent (GD) updates the iterates as

$$x_{k+1} = x_k - \gamma_k \text{grad } f(x_k). \quad (\text{GD})$$

For a convex and L -smooth objective function f , GD with $\gamma_k = \frac{1}{L}$ finds an ϵ -approximate solution, i.e., $f(x_k) - f(x^*) \leq \epsilon$, in $O\left(\frac{L}{\epsilon}\right)$ iterations. For a μ -strongly convex and L -smooth objective function f , GD with $\gamma_k = \frac{1}{L}$ finds an ϵ -approximate solution in $O\left(\frac{L}{\mu} \log \frac{L}{\epsilon}\right)$ iterations. A major breakthrough in first-order algorithms is the Nesterov accelerated gradient (NAG) method that achieves a faster convergence rate than GD (Nesterov, 1983). Given an initial point $x_0 = z_0$, the NAG scheme updates the iterates as

$$\begin{aligned} y_k &= x_k + \tau_k (z_k - x_k) \\ x_{k+1} &= y_k - \alpha_k \text{grad } f(y_k) \\ z_{k+1} &= y_k + \beta_k (z_k - y_k) - \gamma_k \text{grad } f(y_k). \end{aligned} \quad (\text{NAG})$$

For a convex and L -smooth function f , NAG with $\tau_k = \frac{2}{k+2}$, $\alpha_k = \frac{1}{L}$, $\beta_k = 1$, $\gamma_k = \frac{k+2}{2L}$ (NAG-C) finds an ϵ -approximate solution in $O\left(\sqrt{\frac{L}{\epsilon}}\right)$ iterations (Tseng, 2008). For a μ -strongly convex and L -smooth objective function f , NAG with $\tau_k = \frac{\sqrt{\mu/L}}{1 + \sqrt{\mu/L}}$, $\alpha_k = \frac{1}{L}$, $\beta_k = 1 - \sqrt{\frac{\mu}{L}}$, $\gamma_k = \sqrt{\frac{\mu}{L}} \frac{1}{\mu}$ (NAG-SC) finds an ϵ -approximate solution in $O\left(\sqrt{\frac{L}{\mu}} \log \frac{L}{\epsilon}\right)$ iterations (Nesterov, 2018).

Considering the problem (1) for any Riemannian manifold M , (Zhang & Sra, 2016) successfully generalizes the complexity analysis of GD to Riemannian gradient descent (RGD),

$$x_{k+1} = \exp_{x_k}(-\gamma_k \text{grad } f(x_k)), \quad (\text{RGD})$$

¹Department of Electrical and Computer Engineering, Seoul National University, Seoul, South Korea. Correspondence to: Insoon Yang <insoonyang@snu.ac.kr>.

Table 1. Iteration complexities (required number of iterations to obtain an ϵ -approximate solution) for various accelerated methods on Riemannian manifolds. The notation $\tilde{O}(\cdot)$ and $O^*(\cdot)$ omits $\log(L/\epsilon)$ and $\log(L/\mu)$ factors, respectively (Martínez-Rubio, 2022). For our algorithms, the constant ξ is defined as $\xi = \zeta + 3(\zeta - \delta)$, where ζ and δ are defined in Section 3.2. For the iteration complexity of RAGD (Zhang & Sra, 2018), $\frac{10}{9}$ is not regarded as a constant because this constant arises from their nonstandard assumption

$$d(x_0, x^*) \leq \frac{1}{20\sqrt{\max\{K_{\max}, -K_{\min}\}}} \left(\frac{\mu}{L}\right)^{\frac{3}{4}}.$$

Algorithm	Objective function	Iteration complexity	Remark
Algorithm 1 (Liu et al., 2017)	g-strongly convex	$O\left(\sqrt{L/\mu} \log(L/\epsilon)\right)$	computationally intractable
Algorithm 2 (Liu et al., 2017)	g-convex	$O\left(\sqrt{L/\epsilon}\right)$	computationally intractable
RAGD (Zhang & Sra, 2018)	g-strongly convex	$O\left((10/9)\sqrt{L/\mu} \log(L/\epsilon)\right)$	nonstandard assumption
Algorithm 1 (Ahn & Sra, 2020)	g-strongly convex	$O^*\left(L/\mu + \sqrt{L/\mu} \log(\mu/\epsilon)\right)$	eventually accelerated
RAGDsDR (Alimisis et al., 2021)	g-convex	$O\left(\sqrt{\zeta L/\epsilon}\right)$	only in early stages
(Martínez-Rubio, 2022)	g-convex	$\tilde{O}\left(\sqrt{L/\epsilon}\right)$	only for constant curvature
(Martínez-Rubio, 2022)	g-strongly convex	$O^*\left(\sqrt{L/\mu} \log(\mu/\epsilon)\right)$	only for constant curvature
RNAG-C (ours)	g-convex	$O\left(\xi\sqrt{L/\epsilon}\right)$	
RNAG-SC (ours)	g-strongly convex	$O\left(\xi\sqrt{L/\mu} \log(L/\epsilon)\right)$	

using a lower bound K_{\min} of the sectional curvature and an upper bound D of $\text{diam}(N)$. For completeness, we provide a potential-function analysis in Appendix D to show that RGD with a fixed stepsize has the same iteration complexity as GD.

However, it is still unclear whether a reasonable generalization of NAG to the Riemannian setting is possible with strong theoretical guarantees. When studying the global complexity of Riemannian optimization algorithms, it is common to assume that the sectional curvature of M is bounded below by K_{\min} and bounded above by K_{\max} to prevent the manifold from being overly curved. Unfortunately, (Criscitello & Boumal, 2021; Hamilton & Moitra, 2021) show that even when sectional curvature is bounded, achieving global acceleration is impossible in general. Thus, one might need another common assumption, an upper bound D of $\text{diam}(N)$. This motivates our central question:

Can we design computationally tractable accelerated first-order methods on Riemannian manifolds when the sectional curvature and the diameter of the domain are bounded?

In the literature, there are some partial answers but no full answer to this question (see Table 1 and Section 2). In this paper, we provide a complete answer via new first-order algorithms, which we call the Riemannian Nesterov accelerated gradient (RNAG) method. We show that acceleration is possible on Riemannian manifolds for both geodesically convex (g-convex) and geodesically strongly convex (g-

strongly convex) cases whenever the bounds K_{\min} , K_{\max} , and D are available. The main contributions of this work can be summarized as follows:

- Generalizing Nesterov’s scheme, we propose RNAG, a first-order method for Riemannian optimization. We provide two specific algorithms: RNAG-C (Algorithm 1) for minimizing g-convex functions and RNAG-SC (Algorithm 2) for minimizing g-strongly convex functions. Both algorithms call one gradient oracle per iteration. Our algorithms are computationally tractable in the sense that they only involve exponential maps, logarithm maps, parallel transport, and operations in tangent spaces. In particular, RNAG-C can be interpreted as a variant of NAG-C with high friction in (Su et al., 2014, Section 4.1) (see Appendix B).
- Given the bounds K_{\min} , K_{\max} , and D , we prove that RNAG-C has an $O\left(\sqrt{\frac{L}{\epsilon}}\right)$ iteration complexity (Corollary 5.5), and that RNAG-SC has an $O\left(\sqrt{\frac{L}{\mu}} \log \frac{L}{\epsilon}\right)$ iteration complexity (Corollary 5.7). The crucial steps of the proofs are constructing potential functions as (4) and handling metric distortion using Lemma 5.2 and Lemma 5.3. To the best of our knowledge, this is the first proof for full acceleration in the g-convex case.
- We identify a connection between our algorithms and the ODEs for modeling Riemannian acceleration in (Alimisis et al., 2020) by letting the stepsize tend to zero. This analysis confirms the accelerated convergence of

our algorithms through the lens of continuous-time flows.

2. Related Work

Given a bound D for $\text{diam}(N)$, (Liu et al., 2017) proposed accelerated methods for both g-convex and g-strongly convex cases. Their algorithms have the same iteration complexities as NAG but require a solution to a nonlinear equation at every iteration, which could be as difficult as solving the original problem in general. Given K_{\min} , K_{\max} , and $d(x_0, x^*)$, (Zhang & Sra, 2018) proposed a computationally tractable algorithm for the g-strongly convex case and showed that their algorithm achieves the iteration complexity $O\left(\frac{10}{9}\sqrt{\frac{L}{\mu}}\log\frac{L}{\epsilon}\right)$ when $d(x_0, x^*) \leq \frac{1}{20\sqrt{\max\{K_{\max}, -K_{\min}\}}}\left(\frac{\mu}{L}\right)^{\frac{3}{4}}$. Given only K_{\min} and K_{\max} , (Ahn & Sra, 2020) considered the g-strongly convex case. Although full acceleration is not guaranteed, the authors proved that their algorithm eventually achieves acceleration in later stages. Given K_{\min} , K_{\max} , and D , (Alimisis et al., 2021) proposed a momentum method for the g-convex case. They showed that their algorithm achieves acceleration in early stages. Although this result is not as strong as full acceleration, their theoretical guarantee is meaningful in practical situations. (Martínez-Rubio, 2022) focused on manifolds with constant sectional curvatures, namely a subset of the hyperbolic space or sphere. Their algorithm is accelerated, but it is not straightforward to generalize their argument to any manifolds. Beyond the g-convex setting, (Criscitiello & Boumal, 2020) studied accelerated methods for nonconvex problems. (Lezcano-Casado, 2020) studied adaptive and momentum-based methods using the trivialization framework in (Lezcano-Casado, 2019). Further works on accelerated Riemannian optimization can be found in (Criscitiello & Boumal, 2021, Section 1.6).

Another line of research takes the perspective of continuous-time dynamics as in the Euclidean counterpart (Su et al., 2014; Wibisono et al., 2016; Wilson et al., 2021). For both g-convex and g-strongly convex cases, (Alimisis et al., 2020) proposed ODEs that can model accelerated methods on Riemannian manifolds given K_{\min} and D . (Duruiseaux & Leok, 2021b) extended this result and developed a variational framework. Time-discretization methods for such ODEs on Riemannian manifolds have recently been of considerable interest as well (Duruiseaux & Leok, 2021a; França et al., 2021; Duruiseaux & Leok, 2022).

While many positive results have been obtained for accelerated Riemannian optimization, there are also a few negative results (Hamilton & Moitra, 2021) and (Criscitiello & Boumal, 2021), showing that achieving full acceleration for Riemannian optimization is impossible in general. Because their results involve a growing diameter of domain and most

of the positive results assume that the diameter of domain is bounded by a constant D , the negative result is not contradictory but complementary to the positive results. This indicates that the assumption of bounding the domain by a constant is necessary for achieving full acceleration. See Section 8 for a detailed discussion.

3. Preliminaries

3.1. Background

A Riemannian manifold (M, g) is a real smooth manifold equipped with a Riemannian metric g which assigns to each $p \in M$ a positive-definite inner product $g_p(v, w) = \langle v, w \rangle_p = \langle v, w \rangle$ on the tangent space T_pM . The inner product g_p induces the norm $\|v\|_p = \|v\|$ defined as $\sqrt{\langle v, v \rangle_p}$ on T_pM . The tangent bundle TM of M is defined as $TM = \sqcup_{p \in M} T_pM$. For $p, q \in M$, the geodesic distance $d(p, q)$ between p and q is the infimum of the length of all piecewise continuously differentiable curves from p to q . For nonempty set $N \subseteq M$, the diameter $\text{diam}(N)$ of N is defined as $\text{diam}(N) = \sup_{p, q \in N} d(p, q)$.

For a smooth function $f : M \rightarrow \mathbb{R}$, the Riemannian gradient $\text{grad } f(x)$ of f at x is defined as the tangent vector in T_xM satisfying

$$\langle \text{grad } f(x), v \rangle = df(x)[v],$$

where $df(x) : T_xM \rightarrow \mathbb{R}$ is the differential of f at x . Let $I := [0, 1]$. A geodesic $\gamma : I \rightarrow M$ is a smooth curve of locally minimum length with zero acceleration.¹ In particular, straight lines in \mathbb{R}^n are geodesics. The exponential map at p is defined as, for $v \in T_pM$,

$$\exp_p(v) = \gamma_v(1),$$

where $\gamma_v : I \rightarrow M$ is the geodesic satisfying $\gamma_v(0) = p$ and $\gamma'_v(0) = v$. In general, \exp_p is only defined on a neighborhood of 0 in T_pM . It is known that \exp_p is a diffeomorphism in some neighborhood U of 0. Thus, its inverse is well defined and is called the logarithm map $\log_x : \exp_p(U) \rightarrow T_pM$. For a smooth curve $\gamma : I \rightarrow M$ and $t_0, t_1 \in I$, the parallel transport $\Gamma(\gamma)_{t_0}^{t_1} : T_{\gamma(t_0)}M \rightarrow T_{\gamma(t_1)}M$ is a way of transporting vectors from $T_{\gamma(t_0)}M$ to $T_{\gamma(t_1)}M$ along γ .² When γ is a geodesic, we let $\Gamma_p^q : T_pM \rightarrow T_qM$ denote the parallel transport from T_pM to T_qM .

A subset N of M is said to be geodesically uniquely convex if for every $x, y \in N$, there exists a unique geodesic $\gamma : [0, 1] \rightarrow M$ such that $\gamma(0) = x$, $\gamma(1) = y$, and $\gamma(t) \in N$ for all $t \in [0, 1]$. Let N be a geodesically uniquely

¹The mathematical definition of acceleration is provided in Appendix A.

²The definition using covariant derivatives is contained in Appendix A.

convex subset of M . A function $f : N \rightarrow \mathbb{R}$ is said to be geodesically convex if $f \circ \gamma : [0, 1] \rightarrow \mathbb{R}$ is convex for each geodesic $\gamma : [0, 1] \rightarrow M$ whose image is in N . When f is geodesically convex, we have

$$f(y) \geq f(x) + \langle \text{grad } f(x), \log_x(y) \rangle.$$

Let N be an open geodesically uniquely convex subset of M , and $f : N \rightarrow \mathbb{R}$ be a continuously differentiable function. We say that f is geodesically μ -strongly convex for $\mu > 0$ if

$$f(y) \geq f(x) + \langle \text{grad } f(x), \log_x(y) \rangle + \frac{\mu}{2} \|\log_x(y)\|^2$$

for all $x, y \in N$. We say that f is geodesically L -smooth if

$$f(y) \leq f(x) + \langle \text{grad } f(x), \log_x(y) \rangle + \frac{L}{2} \|\log_x(y)\|^2$$

for all $x, y \in N$. For additional notions from Riemannian geometry that are used in our analysis, we refer the reader to Appendix A as well as the textbooks (Lee, 2018; Petersen, 2016; Boumal, 2020).

3.2. Assumptions

In this subsection, we present the assumptions that are imposed throughout the paper.

Assumption 3.1. The domain N is an open geodesically uniquely convex subset of M . The diameter of the domain is bounded as $\text{diam}(N) \leq D < \infty$. The sectional curvature inside N is bounded below by K_{\min} and bounded above by K_{\max} . If $K_{\max} > 0$, we further assume that $D < \frac{\pi}{\sqrt{K_{\max}}}$.

Assumption 3.1 implies that the exponential map \exp_x is a diffeomorphism for any $x \in N$ (Alimisis et al., 2021).

Assumption 3.2. The objective function $f : N \rightarrow \mathbb{R}$ is continuously differentiable and geodesically L -smooth. Moreover, f is bounded below, and has minimizers, all of which lie in N . A global minimizer is denoted by x^* .

Assumption 3.3. All the iterates x_k and y_k are well-defined on the manifold M remain in N .

Although Assumption 3.3 is common in the literature (Zhang & Sra, 2018; Ahn & Sra, 2020; Alimisis et al., 2021), it is desirable to relax or remove it. We leave the extension as a future research topic.

To implement our algorithms, we also assume that we can compute (or approximate) exponential maps, logarithmic maps, and parallel transport. For many manifolds in practical applications, these maps are implemented in libraries such as (Townsend et al., 2016).

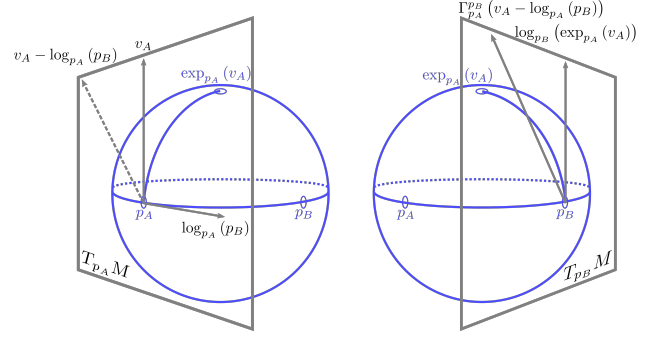


Figure 1. Illustration of the maps $v_A \mapsto \Gamma_{p_A}^{p_B}(v_A - \log_{p_A}(p_B))$ and $v_A \mapsto \log_{p_B}(\exp_{p_A}(v_A))$.

We define the constants $\zeta \geq 1$ and $\delta \leq 1$ as

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

$$\delta = \begin{cases} 1, & \text{if } K_{\max} \leq 0 \\ \sqrt{K_{\max}D} \cot(\sqrt{K_{\max}D}), & \text{if } K_{\max} > 0. \end{cases}$$

These constants naturally arise from the Rauch comparison theorem (Lee, 2018, Theorem 11.7) (Petersen, 2016, Theorem 6.4.3), and many known methods on Riemannian manifolds have a convergence rate depending on some of these constants (Alimisis et al., 2020; 2021; Zhang & Sra, 2016). Note that we can set $\zeta = \delta = 1$ when $M = \mathbb{R}^n$.

4. Algorithms

In this section, we first generalize Nesterov's scheme to the Riemannian setting and then design specific algorithms for both \mathfrak{g} -convex and \mathfrak{g} -strongly convex cases. In (Ahn & Sra, 2020; Zhang & Sra, 2018) NAG is generalized to a three-step algorithm on a Riemannian manifold as

$$\begin{aligned} y_k &= \exp_{x_k}(\tau_k \log_{x_k}(z_k)) \\ x_{k+1} &= \exp_{y_k}(-\alpha_k \text{grad } f(y_k)) \\ z_{k+1} &= \exp_{y_k}(\beta_k \log_{y_k}(z_k) - \gamma_k \text{grad } f(y_k)). \end{aligned} \quad (2)$$

However, it is more natural to define the iterates z_k in the tangent bundle TM , instead of in M .³ Thus, we propose another scheme that involves iterates in TM without using z_k . To associate tangent vectors in different tangent spaces, we use parallel transport, which is a way to transport vectors from one tangent space to another.

³The scheme (2) always uses z_k after mapping it to TM via logarithmic maps. The proof of convergence needs the value of f at x_k and y_k . Thus, these iterates (but not z_k) should be defined in M . Considering continuous-time interpretation, the role of z_k is similar to the role of velocity vector \dot{X} , which is defined in TM (see Section 6).

Algorithm 1 RNAG-C

Input: initial point x_0 , parameters ξ and $T > 0$, step size $s \leq \frac{1}{L}$
 Initialize $\bar{v}_0 = 0 \in T_{x_0}M$.
 Set $\lambda_k = \frac{k+2\xi+T}{2}$.
for $k = 0$ **to** $K - 1$ **do**
 $y_k = \exp_{x_k} \left(\frac{\xi}{\lambda_k + (\xi-1)} \bar{v}_k \right)$
 $x_{k+1} = \exp_{y_k} (-s \text{grad} f(y_k))$
 $v_k = \Gamma_{x_k}^{y_k} (\bar{v}_k - \log_{x_k}(y_k))$
 $\bar{\bar{v}}_{k+1} = v_k - \frac{s\lambda_k}{\xi} \text{grad} f(y_k)$
 $\bar{v}_{k+1} = \Gamma_{y_k}^{x_{k+1}} (\bar{\bar{v}}_{k+1} - \log_{y_k}(x_{k+1}))$
end for
Output: x_K

Given $z_k \in M$ in (2), we define the iterates $v_k = \log_{y_k}(z_k)$, $\bar{v}_k = \log_{x_k}(z_k)$, and $\bar{\bar{v}}_k = \log_{y_{k-1}}(z_k)$ in the tangent bundle TM . It is straightforward to check that the following scheme is equivalent to (2):

$$\begin{aligned}
 y_k &= \exp_{x_k}(\tau_k \bar{v}_k) \\
 x_{k+1} &= \exp_{y_k}(-\alpha_k \text{grad} f(y_k)) \\
 v_k &= \log_{y_k}(\exp_{x_k}(\bar{v}_k)) \\
 \bar{\bar{v}}_{k+1} &= \beta_k v_k - \gamma_k \text{grad} f(y_k) \\
 \bar{v}_{k+1} &= \log_{x_{k+1}}(\exp_{y_k}(\bar{\bar{v}}_{k+1})).
 \end{aligned} \tag{3}$$

In (3), the third and last steps associate tangent vectors in different tangent spaces using the map $T_{p_A}M \rightarrow T_{p_B}M$; $v_A \mapsto \log_{p_B}(\exp_{p_A}(v_A))$. We change these steps by using the map $v_A \mapsto \Gamma_{p_A}^{p_B}(v_A - \log_{p_A}(p_B))$ instead. Technically, this modification allows us to use Lemma 5.3 when handling metric distortion in our convergence analysis. With the change, we obtain the following scheme, which we call RNAG:

$$\begin{aligned}
 y_k &= \exp_{x_k}(\tau_k \bar{v}_k) \\
 x_{k+1} &= \exp_{y_k}(-\alpha_k \text{grad} f(y_k)) \\
 v_k &= \Gamma_{x_k}^{y_k}(\bar{v}_k - \log_{x_k}(y_k)) \\
 \bar{\bar{v}}_{k+1} &= \beta_k v_k - \gamma_k \text{grad} f(y_k) \\
 \bar{v}_{k+1} &= \Gamma_{y_k}^{x_{k+1}}(\bar{\bar{v}}_{k+1} - \log_{y_k}(x_{k+1})).
 \end{aligned} \tag{RNAG}$$

Because RNAG only involves exponential maps, logarithm maps, parallel transport, and operations in tangent spaces, this scheme is computationally tractable, unlike the scheme in (Liu et al., 2017), which involves a non-linear operator. Note that RNAG is different from the scheme (2) because the maps $v_A \mapsto \log_{p_B}(\exp_{p_A}(v_A))$ and $v_A \mapsto \Gamma_{p_A}^{p_B}(v_A - \log_{p_A}(p_B))$ are not equivalent in general (see Figure 1).

By carefully choosing the parameters τ_k , α_k , β_k and γ_k , we finally obtain two algorithms, RNAG-C (Algorithm 1) for

Algorithm 2 RNAG-SC

Input: initial point x_0 , parameter ξ , step size $s \leq \frac{1}{L}$
 Initialize $\bar{v}_0 = 0 \in T_{x_0}M$.
 Set $q = \mu s$.
for $k = 0$ **to** $K - 1$ **do**
 $y_k = \exp_{x_k} \left(\frac{\sqrt{\xi q}}{1 + \sqrt{\xi q}} \bar{v}_k \right)$
 $x_{k+1} = \exp_{y_k}(-s \text{grad} f(y_k))$
 $v_k = \Gamma_{x_k}^{y_k} (\bar{v}_k - \log_{x_k}(y_k))$
 $\bar{\bar{v}}_{k+1} = \left(1 - \sqrt{\frac{q}{\xi}}\right) v_k + \sqrt{\frac{q}{\xi}} \left(-\frac{1}{\mu} \text{grad} f(y_k)\right)$
 $\bar{v}_{k+1} = \Gamma_{y_k}^{x_{k+1}} (\bar{\bar{v}}_{k+1} - \log_{y_k}(x_{k+1}))$
end for
Output: x_K

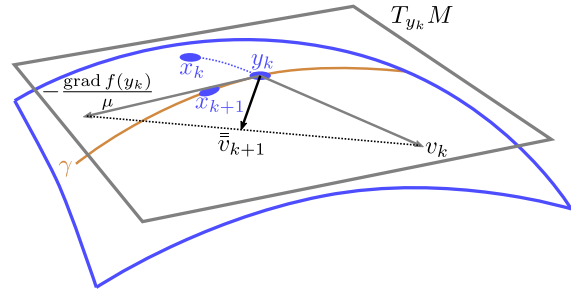


Figure 2. Illustration of RNAG-SC.

the g-convex case, and RNAG-SC (Algorithm 2) for the g-strongly convex case. In particular, we can interpret RNAG-C as a slight variation of NAG-C with high friction (Su et al., 2014, Section 4.1) with the friction parameter $r = 1 + 2\xi$. See Appendix B for a detailed interpretation. Note that we recover NAG-C and NAG-SC from these algorithms when $M = \mathbb{R}^n$ and $\xi = 1$. Figure 2 is an illustration of some steps of RNAG-SC, where the curve γ is a geodesic with $\gamma(0) = y_k$ and $\gamma'(0) = \text{grad} f(y_k)$.

5. Convergence Analysis

5.1. Metric distortion lemma

To handle a potential function involving squared norms in tangent spaces, we need to compare distances in different tangent spaces.

Proposition 5.1. (Alimisis et al., 2020, Lemma 2) *Let γ be a smooth curve whose image is in N . Then, we have*

$$\delta \|\gamma'(t)\|^2 \leq \left\langle D_t \log_{\gamma(t)}(x), -\gamma'(t) \right\rangle \leq \zeta \|\gamma'(t)\|^2.$$

In the proposition above, D_t is a covariant derivative along the curve (see Appendix A). Using this proposition, we obtain the following lemma.

Lemma 5.2. *Let $p_A, p_B, x \in N$ and $v_A \in T_{p_A}M$. If there*

is $r \in [0, 1]$ such that $\log_{p_A}(p_B) = rv_A$, then we have

$$\begin{aligned} & \|v_B - \log_{p_B}(x)\|_{p_B}^2 + (\zeta - 1) \|v_B\|_{p_B}^2 \\ & \leq \|v_A - \log_{p_A}(x)\|_{p_A}^2 + (\zeta - 1) \|v_A\|_{p_A}^2, \end{aligned}$$

where $v_B = \Gamma_{p_A}^{p_B}(v_A - \log_{p_A}(p_B)) \in T_{p_B}M$

In particular, when $r = 1$, Lemma 5.2 recovers a weaker version of (Zhang & Sra, 2016, Lemma 5). We can further generalize this lemma as follows:

Lemma 5.3. *Let $p_A, p_B, x \in N$ and $v_A \in T_{p_A}M$. Define $v_B = \Gamma_{p_A}^{p_B}(v_A - \log_{p_A}(p_B)) \in T_{p_B}M$. If there are $a, b \in T_{p_A}M$, and $r \in (0, 1)$ such that $v_A = a + b$ and $\log_{p_A}(p_B) = rb$, then we have*

$$\begin{aligned} & \|v_B - \log_{p_B}(x)\|_{p_B}^2 + (\xi - 1) \|v_B\|_{p_B}^2 \\ & \leq \|v_A - \log_{p_A}(x)\|_{p_A}^2 + (\xi - 1) \|v_A\|_{p_A}^2 \\ & \quad + \frac{\xi - \delta}{2} \left(\frac{1}{1-r} - 1 \right) \|a\|_{p_A}^2 \end{aligned}$$

for $\xi \geq \zeta$.

As $\exp_{p_A}(v_A) \neq \exp_{p_B}(v_B)$ (see Figure 1), our lemma does not compare the projected distance⁴ between points on the manifold, unlike (Zhang & Sra, 2018, Theorem 10) and (Ahn & Sra, 2020, Lemma 4.1). The proofs of Lemma 5.2 and Lemma 5.3 can be found in Appendix C.

5.2. Main results

We now prove the iteration complexities of RNAG-C and RNAG-SC using potential functions of the form

$$\begin{aligned} \phi_k &= A_k (f(x_k) - f(x^*)) \\ & \quad + B_k \left(\|\bar{v}_k - \log_{x_k}(x^*)\|_{x_k}^2 + (\xi - 1) \|\bar{v}_k\|_{x_k}^2 \right). \end{aligned} \quad (4)$$

The term $(\xi - 1) \|\bar{v}_k\|_{x_k}^2$ is novel compared with the potential function in (Ahn & Sra, 2020), and it measures the kinetic energy (Wibisono et al., 2016). Intuitively, this potential makes sense because a large ξ means high friction (see Appendix B and Section 6). This term is useful when handling metric distortion.

5.2.1. THE GEODESICALLY CONVEX CASE

For the g -convex case, we use a potential function defined as

$$\begin{aligned} \phi_k &= s\lambda_{k-1}^2 (f(x_k) - f(x^*)) \\ & \quad + \frac{\xi}{2} \|\bar{v}_k - \log_{x_k}(x^*)\|^2 + \frac{\xi(\xi - 1)}{2} \|\bar{v}_k\|^2. \end{aligned} \quad (5)$$

⁴For $u, v, w \in M$, the projected distance between v and w with respect to u is defined as $\|\log_u(v) - \log_u(w)\|^2$ (Ahn & Sra, 2020, Definition 3.1).

The following theorem shows that this potential function is decreasing when the parameters ξ and T are chosen appropriately.

Theorem 5.4. *Let f be a g -convex and geodesically L -smooth function. If the parameters ξ and T of RNAG-C satisfy $\xi \geq \zeta$ and*

$$\begin{aligned} & \frac{\xi - \delta}{2} \left(\frac{1}{1 - \xi/\lambda_k} - 1 \right) \\ & \leq (\xi - \zeta) \left(\frac{1}{(1 - \xi/(\lambda_k + \xi - 1))^2} - 1 \right) \end{aligned}$$

for all $k \geq 0$, then the iterates of RNAG-C satisfy $\phi_{k+1} \leq \phi_k$ for all $k \geq 0$, where ϕ_k is defined as (5).

In particular, we can show that the parameters $\xi = \zeta + 3(\zeta - \delta)$ and $T = 4\xi$ satisfy the condition in Theorem 5.4. In this case, the monotonicity of the potential function yields

$$f(x_k) - f(x^*) \leq \frac{1}{s\lambda_{k-1}^2} \phi_k \leq \frac{1}{s\lambda_{k-1}^2} \phi_0.$$

Thus, RNAG-C achieves acceleration. The result is summarized in the following corollary.

Corollary 5.5. *Let f be a g -convex and geodesically L -smooth function. Then, RNAG-C with parameters $\xi = \zeta + 3(\zeta - \delta)$, $T = 4\xi$ and step size $s = \frac{1}{L}$ finds an ϵ -approximate solution in $O\left(\xi\sqrt{\frac{L}{\epsilon}}\right)$ iterations.*

This result implies that the iteration complexity of RNAG-C is the same as that of NAG-C because ξ is a constant. The proofs of Theorem 5.4 and Corollary 5.5 are contained in Appendix E.

5.2.2. THE GEODESICALLY STRONGLY CONVEX CASE

For the g -strongly convex case, we consider a potential function defined as

$$\begin{aligned} \phi_k &= \left(1 - \sqrt{\frac{q}{\xi}}\right)^{-k} \left(f(x_k) - f(x^*) \right. \\ & \quad \left. + \frac{\mu}{2} \|v_k - \log_{y_k}(x^*)\|^2 + \frac{\mu(\xi - 1)}{2} \|v_k\|^2 \right). \end{aligned} \quad (6)$$

This potential function is also shown to be decreasing under appropriate conditions on ξ and s .

Theorem 5.6. *Let f be a geodesically μ -strongly convex and geodesically L -smooth function. If the step size s and the parameter ξ of RNAG-SC satisfy $\xi \geq \zeta$, $\sqrt{\xi q} < 1$, and*

$$\begin{aligned} & \frac{\xi - \delta}{2} \left(\frac{1}{1 - \sqrt{\xi q}} - 1 \right) \left(1 - \sqrt{\frac{q}{\xi}} \right)^2 - \sqrt{\xi q} \left(1 - \sqrt{\frac{q}{\xi}} \right) \\ & \leq (\xi - \zeta) \left(\frac{1}{(1 - \sqrt{\xi q}/(1 + \sqrt{\xi q}))^2} - 1 \right), \end{aligned}$$

then the iterates of RNAG-SC satisfy $\phi_{k+1} \leq \phi_k$ for all $k \geq 0$, where ϕ_k is defined as (6).

In particular, the parameters $\xi = \zeta + 3(\zeta - \delta)$ and $s = \frac{1}{9\xi L}$ satisfy the condition in Theorem 5.6. In this case, by monotonicity of the potential function, we have

$$f(x_k) - f(x^*) \leq \left(1 - \sqrt{\frac{q}{\xi}}\right)^k \phi_0 \leq \left(1 - \sqrt{\frac{q}{\xi}}\right)^k \phi_0,$$

which implies that RNAG-SC achieves acceleration. The following corollary summarizes the result.

Corollary 5.7. *Let f be a geodesically μ -strongly convex and geodesically L -smooth function. Then, RNAG-SC with parameter $\xi = \zeta + 3(\zeta - \delta)$ and step size $s = \frac{1}{9\xi L}$ finds an ϵ -approximate solution in $O\left(\xi \sqrt{\frac{L}{\mu}} \log\left(\frac{L}{\epsilon}\right)\right)$ iterations.*

Because ξ is a constant, the iteration complexity of RNAG-SC is the same as that of NAG-SC. The proofs of Theorem 5.6 and Corollary 5.7 can be found in Appendix F.

6. Continuous-Time Interpretation

In this section, we identify a connection to the ODEs for modeling Riemannian acceleration in (Alimisis et al., 2020, Equations 2 and 4). Specifically, following the informal arguments in (Su et al., 2014, Section 2) and (d'Aspremont et al., 2021, Section 4.8), we obtain ODEs by taking the limit $s \rightarrow 0$ in our schemes. The detailed analysis is contained in Appendix G. For a sufficiently small s , the Euclidean geometry is valid as only a sufficiently small subset of M is considered. Thus, we informally assume $M = \mathbb{R}^n$ for simplicity. We can show that the iterations of RNAG-C satisfy

$$\begin{aligned} & \frac{y_{k+1} - y_k}{\sqrt{s}} \\ &= \frac{\lambda_k - 1}{\lambda_{k+1} + (\xi - 1)} \frac{y_k - y_{k-1}}{\sqrt{s}} \\ & \quad - \frac{\lambda_{k+1}}{\lambda_{k+1} + (\xi - 1)} \sqrt{s} \operatorname{grad} f(y_k) \\ & \quad + \frac{\lambda_k - 1}{\lambda_{k+1} + (\xi - 1)} \sqrt{s} (\operatorname{grad} f(y_{k-1}) - \operatorname{grad} f(y_k)). \end{aligned}$$

We introduce a smooth curve $y(t)$ that is approximated by the iterates of RNAG-C as $y(t) \approx y_{t/\sqrt{s}} = y_k$ with $k = \frac{t}{\sqrt{s}}$. Using the Taylor expansion, we have

$$\begin{aligned} \frac{y_{k+1} - y_k}{\sqrt{s}} &= \dot{y}(t) + \frac{\sqrt{s}}{2} \ddot{y}(t) + o(\sqrt{s}), \\ \frac{y_k - y_{k-1}}{\sqrt{s}} &= \dot{y}(t) - \frac{\sqrt{s}}{2} \ddot{y}(t) + o(\sqrt{s}), \\ \sqrt{s} \operatorname{grad} f(y_{k-1}) &= \sqrt{s} \operatorname{grad} f(y_k) + o(\sqrt{s}). \end{aligned}$$

Letting $s \rightarrow 0$ yields the ODE⁵

$$\nabla_{\dot{y}} \dot{y} + \frac{1 + 2\xi}{t} \dot{y} + \operatorname{grad} f(y) = 0, \quad (7)$$

where the covariant derivative $\nabla_{\dot{y}} \dot{y} = D_t \dot{y}$ is a natural extension of the second derivative \ddot{y} (see Appendix A).

In the g -strongly convex case, we can show that the iterations of RNAG-SC satisfy

$$\begin{aligned} & \frac{y_{k+1} - y_k}{\sqrt{s}} \\ &= \frac{1 - \sqrt{q/\xi}}{1 + \sqrt{\xi q}} \frac{y_k - y_{k-1}}{\sqrt{s}} - \frac{1 + \sqrt{q/\xi}}{1 + \sqrt{\xi q}} \sqrt{s} \operatorname{grad} f(y_k) \\ & \quad + \frac{1 - \sqrt{q/\xi}}{1 + \sqrt{\xi q}} \sqrt{s} (\operatorname{grad} f(y_{k-1}) - \operatorname{grad} f(y_k)). \end{aligned}$$

Through a similar limiting process, we obtain the following ODE:

$$\nabla_{\dot{y}} \dot{y} + \left(\frac{1}{\sqrt{\xi}} + \sqrt{\xi}\right) \sqrt{\mu} \dot{y} + \operatorname{grad} f(y) = 0. \quad (8)$$

Replacing the parameter ξ in the coefficients of our ODEs with ζ , we recover (Alimisis et al., 2020, Equations 2 and 4). Because $\xi \geq \zeta$, the continuous-time acceleration results (Alimisis et al., 2020, Theorems 5 and 7) are valid for our ODEs as well. Thus, this analysis confirms the accelerated convergence of our algorithms through the lens of continuous-time flows.

In both ODEs, the parameter $\xi \geq \zeta$ appears in the coefficient of the friction term \dot{X} , increasing with ξ . Intuitively, this makes sense because ζ is large for an ill-conditioned domain, where $-K_{\min}$ and D are large and thus metric distortion is more severe (where one might want to decrease the effect of momentum).

7. Experiments

In this section, we examine the performance of our algorithms on the Rayleigh quotient maximization problem and the Karcher mean problem. To implement the geometry of manifolds, we used the Python libraries Pymanopt (Townsend et al., 2016) and Geomstats (Miolane et al., 2020). For comparison, we use the known accelerated algorithms RAGD (Zhang & Sra, 2018) for the g -strongly convex case and RNAGsDR with no line search (Alimisis et al., 2021) for the g -convex case. The source code of our RNAG implementation is available online.⁶

⁵When $M \neq \mathbb{R}^n$, we replace \ddot{y} with the acceleration $\nabla_{\dot{y}} \dot{y} = D_t \dot{y}$, where D_t is a covariant derivative along the curve y (see Appendix A).

⁶<https://github.com/jungbinkim1/RNAG>

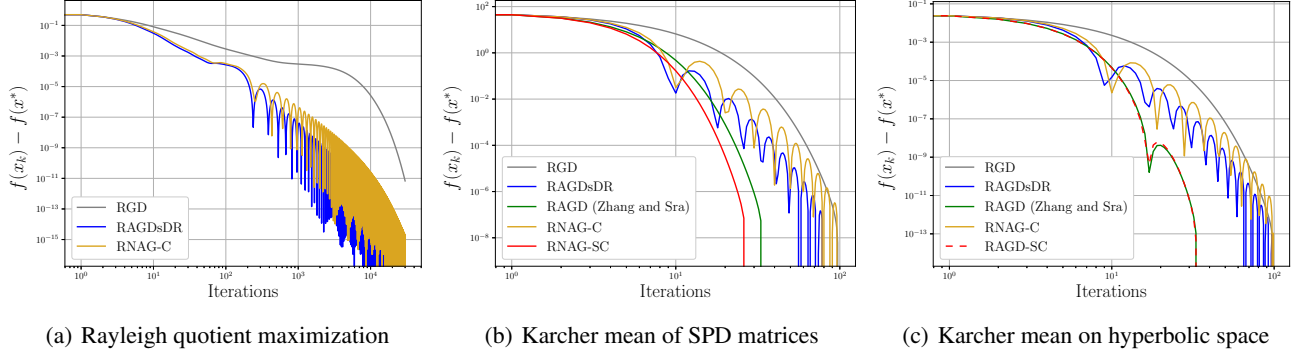


Figure 3. Performances of various Riemannian optimization algorithms on the Rayleigh quotient maximization problem and the Karcher mean problem.

We set the input parameters as $\zeta = 1$ for implementing RAGDsDR, and $\xi = 1$ for implementing our algorithms. The stepsize was chosen as $s = \frac{1}{L}$ in our algorithms.

Rayleigh quotient maximization. Given a real $d \times d$ symmetric matrix A , we consider the problem

$$\min_{x \in \mathbb{S}^{d-1}} f(x) = -\frac{1}{2} x^\top A x.$$

on the unit $(d-1)$ -sphere on \mathbb{S}^{d-1} . For this manifold, we set $K_{\min} = K_{\max} = 1$. We let $d = 1000$ and $A = \frac{1}{2}(B + B^\top)$, where the entries of $B \in \mathbb{R}^{d \times d}$ were randomly generated by the Gaussian distribution $N(0, 1/d)$. We have the smoothness parameter $L = \lambda_{\max} - \lambda_{\min}$ by the following proposition.

Proposition 7.1. *The function f is geodesically $(\lambda_{\max} - \lambda_{\min})$ -smooth, where λ_{\max} and λ_{\min} are the largest and smallest eigenvalues of A , respectively.*

The proof can be found in Appendix H. The result is shown in Figure 3(a). We observe that RNAG-C outperforms RGD and is comparable to RAGDsDR, a known accelerated method for the g-convex case.

Karcher mean of SPD matrices. When $K_{\max} \leq 0$, the Karcher mean (Karcher, 1977) of the points $p_i \in M$ for $i = 1, \dots, n$, is defined as the solution of

$$\min_{x \in M} f(x) = \frac{1}{2n} \sum_{i=1}^n d(x, p_i)^2. \quad (9)$$

The following proposition shows that one can set the strong convexity parameter as $\mu = 1$.

Proposition 7.2. *The function f is geodesically 1-strongly convex.*

The proof can be found in Appendix H. We consider this problem on the manifold $\mathcal{P}(d) \subseteq \mathbb{R}^{d \times d}$ of symmetric positive definite matrices endowed with the Riemannian metric

$\langle X, Y \rangle_P = \text{Tr}(P^{-1} X P^{-1} Y)$. It is known that one can set $K_{\min} = -\frac{1}{2}$ and $K_{\max} = 0$ (Criscitello & Boumal, 2020, Appendix I). We set the dimension and the number of matrices as $d = 100$ and $n = 50$. The matrices p_i were randomly generated using Matrix Mean Toolbox (Bini & Iannazzo, 2013) with condition number 10^6 . We set the smoothness parameter as $L = 10$. The result is shown in Figure 3(b). We observe that RNAG-SC and RAGD (Zhang & Sra, 2018) perform significantly better than RGD. The performances of RNAG-C and RAGDsDR are only slightly better than that of RGD in early stages. This result makes sense because f is g-strongly convex and well-conditioned.

Karcher mean on hyperbolic space. We consider the problem (9) on the hyperbolic space \mathbb{H}^d with the hyperboloid model $\mathbb{H}^d = \{x \in \mathbb{R}^{d+1} : -x_{d+1}^2 + \sum_{k=1}^d x_k^2 = -1\}$. For this manifold, we can set $K_{\min} = K_{\max} = -1$. We set the dimension and the number of points as $d = 1000$ and $n = 10$. First d entries of each point p_i are randomly generated by the Gaussian distribution $N(0, 1/d)$. We set the smoothness parameter as $L = 10$. The result is similar to that of the previous example, and is shown in Figure 3(c).

8. Discussion

In this paper, we have proposed novel computationally tractable first-order methods that achieve Riemannian acceleration for both g-convex and g-strongly convex objective functions whenever the constants K_{\min} , K_{\max} , and D are available. The iteration complexities of RNAG-C and RNAG-SC match those of their Euclidean counterparts. The continuous-time analysis of our algorithms provides an intuitive interpretation of the parameter ξ as a measurement of friction, which is higher when the domain manifold is more ill-conditioned. In fact, the iteration complexities of our algorithms depend on the parameter $\xi \geq \zeta$, which is affected by the values of the constants K_{\min} , K_{\max} , and D . When

ζ is large (i.e., $-K_{\min}$ and D are large), we have a worse guarantee. A possible future direction is to study the effect of the constants K_{\min} , K_{\max} , and D on the complexities of Riemannian optimization algorithms tightly.

Comparison with (Liu et al., 2017). The algorithms in (Liu et al., 2017) achieve acceleration with only standard assumption. However, to implement the operator $\mathbb{S} : (y_{k-1}, x_k, x_{k-1}) \mapsto y_k$ in (Liu et al., 2017, Algorithm 1), one needs to solve the following nonlinear equation at each iteration:

$$\begin{aligned} & (1 - \sqrt{\mu/L})\Gamma_{y_k}^{y_{k-1}} \log_{y_k}(x_k) - \beta\Gamma_{y_k}^{y_{k-1}} \text{grad } f(y_k) \\ & = (1 - \sqrt{\mu/L})^{3/2} \log_{y_{k-1}}(x_{k-1}). \end{aligned}$$

It is unclear whether this equation is solvable in a tractable way or even feasible as noted in (Ahn & Sra, 2020). On the other hand, our algorithms involve only operations in tangent spaces and the exponential map, logarithm map, and parallel transport. Thus, our algorithms are computationally tractable for various manifolds in practice, where the operations above are implementable.

Comparison with (Criscitiello & Boumal, 2021). It is natural to ask how our positive result is not contradictory to the negative result in (Criscitiello & Boumal, 2021). To clarify this, we provide the following two reasons:

(i) We assume that the diameter $\text{diam}(N)$ of the domain N is bounded, which is a more restrictive condition than their assumption that the distance $d(x_0, x^*)$ is bounded.

(ii) We assume that the diameter $\text{diam}(N)$ is bounded by a fixed constant D . Thus, in Corollary 5.5 and Corollary 5.7, ξ does not depend on other parameters such as μ and L . In contrast, (Criscitiello & Boumal, 2021, Theorem 1.3) introduces a bound $\frac{3}{4}r$ of $d(x_0, x^*)$ by letting r be the solution of $\kappa = 12r\sqrt{-K_{\min}} + 9$, thus $r\sqrt{-K_{\min}}$ grows with $\kappa = L/\mu$. A similar discussion can be found in (Martínez-Rubio, 2022, Remark 29).

We believe that the second one is the main reason for our positive results coexist with their negative results. As mentioned in Section 2, their result is not contradictory but complementary to our results.

Acknowledgements

We thank the anonymous reviewers for their insightful suggestions and Dr. Antonio Orvieto and the co-authors of (Alimisis et al., 2021) for allowing us to use a part of their code. This work was supported in part by Samsung Electronics, the National Research Foundation of Korea funded by MSIT(2020R1C1C1009766), and the Information and Communications Technology Planning and Evaluation (IITP) grant funded by MSIT(2022-0-00124, 2022-0-00480).

References

- Ahn, K. and Sra, S. From nesterov’s estimate sequence to riemannian acceleration. In *Proceedings of Thirty Third Conference on Learning Theory*, pp. 84–118, 2020.
- Alimisis, F., Orvieto, A., Becigneul, G., and Lucchi, A. A continuous-time perspective for modeling acceleration in riemannian optimization. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, pp. 1297–1307, 2020.
- Alimisis, F., Orvieto, A., Becigneul, G., and Lucchi, A. Momentum improves optimization on riemannian manifolds. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, pp. 1351–1359, 2021.
- Bini, D. A. and Iannazzo, B. Computing the karcher mean of symmetric positive definite matrices. *Linear Algebra and its Applications*, 438(4):1700–1710, 2013.
- Boumal, N. An introduction to optimization on smooth manifolds. Available online, Aug 2020. URL <http://www.nicolasboumal.net/book>.
- Criscitiello, C. and Boumal, N. An accelerated first-order method for non-convex optimization on manifolds. *arXiv preprint arXiv:2008.02252*, 2020.
- Criscitiello, C. and Boumal, N. Negative curvature obstructs acceleration for geodesically convex optimization, even with exact first-order oracles. *arXiv preprint arXiv:2111.13263*, 2021.
- d’Aspremont, A., Scieur, D., and Taylor, A. Acceleration methods. *arXiv preprint arXiv:2101.09545*, 2021.
- Duruisseaux, V. and Leok, M. Accelerated optimization on riemannian manifolds via discrete constrained variational integrators. *arXiv preprint arXiv:2104.07176*, 2021a.
- Duruisseaux, V. and Leok, M. A variational formulation of accelerated optimization on riemannian manifolds. *arXiv preprint arXiv:2101.06552*, 2021b.
- Duruisseaux, V. and Leok, M. Accelerated optimization on riemannian manifolds via projected variational integrators. *arXiv preprint arXiv:2201.02904*, 2022.
- França, G., Barp, A., Girolami, M., and Jordan, M. I. Optimization on manifolds: A symplectic approach. *arXiv preprint arXiv:2107.11231*, 2021.
- Hamilton, L. and Moitra, A. No-go theorem for acceleration in the hyperbolic plane. *arXiv preprint arXiv:2101.05657*, 2021.

- Hosseini, R. and Sra, S. Matrix manifold optimization for gaussian mixtures. In *Advances in Neural Information Processing Systems*, pp. 910–918, 2015.
- Hosseini, R. and Sra, S. An alternative to em for gaussian mixture models: batch and stochastic riemannian optimization. *Mathematical Programming*, 181(1):187–223, 2020.
- Karcher, H. Riemannian center of mass and mollifier smoothing. *Communications on Pure and Applied Mathematics*, 30(5):509–541, 1977.
- Lee, J. M. *Introduction to Riemannian Manifolds*, volume 176. Springer, 2018.
- Lezcano-Casado, M. Trivializations for gradient-based optimization on manifolds. In *Advances in Neural Information Processing Systems*, pp. 9157–9168, 2019.
- Lezcano-Casado, M. Adaptive and momentum methods on manifolds through trivializations. *arXiv preprint arXiv:2010.04617*, 2020.
- Liu, Y., Shang, F., Cheng, J., Cheng, H., and Jiao, L. Accelerated first-order methods for geodesically convex optimization on riemannian manifolds. In *Advances in Neural Information Processing Systems*, pp. 4868–4877, 2017.
- Martínez-Rubio, D. Global riemannian acceleration in hyperbolic and spherical spaces. In *International Conference on Algorithmic Learning Theory*, pp. 768–826, 2022.
- Miolane, N., Guigui, N., Brigant, A. L., Mathe, J., Hou, B., Thanwerdas, Y., Heyder, S., Peltre, O., Koep, N., Zaatiti, H., Hajri, H., Cabanes, Y., Gerald, T., Chauchat, P., Shewmake, C., Brooks, D., Kainz, B., Donnat, C., Holmes, S., and Pennec, X. Geomstats: A python package for riemannian geometry in machine learning. *Journal of Machine Learning Research*, 21(223):1–9, 2020.
- Nesterov, Y. *Lectures on Convex Optimization*, volume 137. Springer, 2018.
- Nesterov, Y. E. A method for solving the convex programming problem with convergence rate $o(1/k^2)$. In *Dokl. akad. nauk Sssr*, volume 269, pp. 543–547, 1983.
- Nguyen, V. A., Shafieezadeh-Abadeh, S., Yue, M.-C., Kuhn, D., and Wiesemann, W. Calculating optimistic likelihoods using (geodesically) convex optimization. *arXiv preprint arXiv:1910.07817*, 2019.
- Petersen, P. *Riemannian Geometry*, volume 1 of 171. Springer, Cham, 3 edition, 2016. ISBN 978-3-319-26652-7.
- Sra, S. On the matrix square root via geometric optimization. *arXiv preprint arXiv:1507.08366*, 2015.
- Su, W., Boyd, S., and Candes, E. A differential equation for modeling nesterov’s accelerated gradient method: Theory and insights. In *Advances in Neural Information Processing Systems*, pp. 2510–2518, 2014.
- Townsend, J., Koep, N., and Weichwald, S. Pymanopt: A python toolbox for optimization on manifolds using automatic differentiation. *The Journal of Machine Learning Research*, 17(1):4755–4759, 2016.
- Tseng, P. On accelerated proximal gradient methods for convex-concave optimization. *submitted to SIAM Journal on Optimization*, 2(3), 2008.
- Vishnoi, N. K. Geodesic convex optimization: Differentiation on manifolds, geodesics, and convexity. *arXiv preprint arXiv:1806.06373*, 2018.
- Wibisono, A., Wilson, A. C., and Jordan, M. I. A variational perspective on accelerated methods in optimization. *proceedings of the National Academy of Sciences*, 113(47): E7351–E7358, 2016.
- Wiesel, A. Geodesic convexity and covariance estimation. *IEEE Transactions on Signal Processing*, 60(12):6182–6189, 2012.
- Wilson, A. C., Recht, B., and Jordan, M. I. A lyapunov analysis of accelerated methods in optimization. *Journal of Machine Learning Research*, 22(113):1–34, 2021.
- Zadeh, P., Hosseini, R., and Sra, S. Geometric mean metric learning. In *International Conference on Machine Learning*, pp. 2464–2471, 2016.
- Zhang, H. and Sra, S. First-order methods for geodesically convex optimization. In *Conference on Learning Theory*, pp. 1617–1638, 2016.
- Zhang, H. and Sra, S. An estimate sequence for geodesically convex optimization. In *Proceedings of the 31st Conference on Learning Theory*, pp. 1703–1723, 2018.

A. Background

Definition A.1. A smooth vector field V is a smooth map from M to TM such that $p \circ V$ is the identity map, where $p : TM \rightarrow M$ is the projection. The collection of all smooth vector fields on M is denoted by $\mathfrak{X}(M)$.

Definition A.2. Let $\gamma : I \rightarrow M$ be a smooth curve. A smooth vector field V along γ is a smooth map from I to TM such that $V(t) \in T_{\gamma(t)}M$ for all $t \in I$. The collection of all smooth vector fields along γ is denoted by $\mathfrak{X}(\gamma)$.

Proposition A.3 (Fundamental theorem of Riemannian geometry). *There exists a unique operator*

$$\nabla : \mathfrak{X}(M) \times \mathfrak{X}(M) \rightarrow \mathfrak{X}(M) : (U, V) \mapsto \nabla_U V$$

satisfying the following properties for any $U, V, W \in \mathfrak{X}(M)$, smooth functions f, g on M , and $a, b \in \mathbb{R}$:

1. $\nabla_{fU+gW}V = f\nabla_U V + g\nabla_W V$
2. $\nabla_U(aV + bW) = a\nabla_U V + b\nabla_U W$
3. $\nabla_U(fV) = (Uf)V + f\nabla_U V$
4. $[U, V] = \nabla_U V - \nabla_V U$
5. $U\langle V, W \rangle = \langle \nabla_U V, W \rangle + \langle V, \nabla_U W \rangle$,

where $[\cdot, \cdot]$ denotes the Lie bracket. The operator ∇ is called the Levi-Civita connection or the Riemannian connection. The field $\nabla_U V$ is called the covariant derivative of V along U .

From now on, we always assume that M is equipped with the Riemannian connection ∇ .

Proposition A.4. (Boumal, 2020, Section 8.11) *For any smooth vector fields U, V on M , the vector field $\nabla_U V$ at x depends on U only through $U(x)$. Thus, we can write $\nabla_u V$ to mean $(\nabla_U V)(x)$ for any $U \in \mathfrak{X}(M)$ such that $U(x) = u$, without ambiguity.*

For a smooth function $f : M \rightarrow \mathbb{R}$, $\text{grad } f$ is a smooth vector field.

Definition A.5. (Boumal, 2020, Section 8.11) The Riemannian Hessian of a smooth function f on M at $x \in M$ is a self-adjoint linear operator $\text{Hess } f(x) : T_x M \rightarrow T_x M$ defined as

$$\text{Hess } f(x)[u] = \nabla_u \text{grad } f.$$

Proposition A.6. (Boumal, 2020, Section 8.12) *Let $c : I \rightarrow M$ be a smooth curve. There exists a unique operator $D_t : \mathfrak{X}(c) \rightarrow \mathfrak{X}(c)$ satisfying the following properties for all $Y, Z \in \mathfrak{X}(c)$, $U \in \mathfrak{X}(M)$, a smooth function g on I , and $a, b \in \mathbb{R}$:*

1. $D_t(aY + bZ) = aD_t Y + bD_t Z$
2. $D_t(gZ) = g'Z + gD_t Z$
3. $(D_t(U \circ c))(t) = \nabla_{c'(t)} U$ for all $t \in I$
4. $\frac{d}{dt} \langle Y, Z \rangle = \langle D_t Y, Z \rangle + \langle Y, D_t Z \rangle$.

This operator is called the (induced) covariant derivative along the curve c .

We define the acceleration of a smooth curve γ as the vector field $D_t \gamma'$ along γ . Now, we can define the parallel transport using covariant derivatives.

Definition A.7. (Boumal, 2020, Section 10.3) A vector field $Z \in \mathfrak{X}(c)$ is parallel if $D_t Z = 0$.

Proposition A.8. (Boumal, 2020, Section 10.3) *For any smooth curve $c : I \rightarrow M$, $t_0 \in I$ and $u \in T_{c(t_0)}M$, there exists a unique parallel vector field $Z \in \mathfrak{X}(c)$ such that $Z(t_0) = u$.*

Definition A.9. (Boumal, 2020, Section 10.3) Given a smooth curve c on M , the parallel transport of tangent vectors at $c(t_0)$ to the tangent space at $c(t_1)$ along c ,

$$\Gamma(c)_{t_0}^{t_1} : T_{c(t_0)}M \rightarrow T_{c(t_1)}M,$$

is defined by $\Gamma(c)_{t_0}^{t_1}(u) = Z(t_1)$, where $Z \in \mathfrak{X}(c)$ is the unique parallel vector field such that $Z(t_0) = u$.

Proposition A.10. (Boumal, 2020, Section 10.3) The parallel transport operator $\Gamma(c)_{t_0}^{t_1}$ is linear. Also, $\Gamma(c)_{t_1}^{t_2} \circ \Gamma(c)_{t_0}^{t_1} = \Gamma(c)_{t_0}^{t_2}$ and $\Gamma(c)_t^t$ is the identity. In particular, the inverse of $\Gamma(c)_{t_0}^{t_1}$ is $\Gamma(c)_{t_1}^{t_0}$. The parallel transport is an isometry, that is,

$$\langle u, v \rangle_{c(t_0)} = \langle \Gamma(c)_{t_0}^{t_1}(u), \Gamma(c)_{t_0}^{t_1}(v) \rangle_{c(t_1)}.$$

Proposition A.11. (Boumal, 2020, Section 10.3) Consider a smooth curve $c : I \rightarrow M$. Given a vector field $Z \in \mathfrak{X}(c)$, we have

$$D_t Z(t) = \lim_{h \rightarrow 0} \frac{\Gamma(c)_{t+h}^t Z(t+h) - Z(t)}{h}.$$

B. Comparison between RNAG-C and High-Friction NAG-C

In this section, we review high-friction NAG-C in (Su et al., 2014, Section 4.1), and compare it to RNAG-C. For $r \geq 3$, they designed the generalized NAG-C with high friction as

$$\begin{aligned} x_k &= y_{k-1} - s \operatorname{grad} f(y_{k-1}) \\ y_k &= x_k + \frac{k-1}{k+r-1} (x_k - x_{k-1}). \end{aligned}$$

Introducing the third sequence as $z_k = y_k + \frac{k}{r-1} (y_k - x_k)$, we can rewrite this method as

$$\begin{aligned} y_k &= x_k + \frac{r-1}{k+r-1} (z_k - x_k) \\ x_{k+1} &= y_k - s \operatorname{grad} f(y_k) \\ z_{k+1} &= z_k - \frac{k+r-1}{r-1} s \operatorname{grad} f(y_k). \end{aligned} \tag{NAG-C-HF}$$

Note that we can recover NAG-C by letting $r = 3$. The iterates of **NAG-C-HF** satisfy

$$f(x_k) - f(x^*) \leq \frac{(r-1)^2 \|x_0 - x^*\|^2}{2s(k+r-2)^2} \leq \frac{(r-1)^2 \|x_0 - x^*\|^2}{2s(k-2)^2}$$

for $s \leq \frac{1}{L}$ (Su et al., 2014, Theorem 6). Thus, we have $f(x_k) - f(x^*) \leq \epsilon$ whenever

$$(k-2)^2 \geq \frac{(r-1)^2 \|x_0 - x^*\|^2}{2s\epsilon}.$$

In particular, when $s = \frac{1}{L}$ and $r = 1 + 2\xi$, we have the Iteration complexity $O\left(\xi \sqrt{\frac{L}{\epsilon}}\right)$.

For comparison, we write RNAG-C in Euclidean space as

$$\begin{aligned} y_k &= x_k + \frac{2\xi}{k+2\xi+(T+2\xi-2)} (z_k - x_k) \\ x_{k+1} &= y_k - s \operatorname{grad} f(y_k) \\ z_{k+1} &= z_k - \frac{k+2\xi+T}{2\xi} s \operatorname{grad} f(y_k). \end{aligned} \tag{10}$$

One can see that the algorithm (10) is similar to that of **NAG-C-HF** with $r = 1 + 2\xi$, where the only difference occurs in constants that can be ignored as k grows. Note that both algorithms have the same iteration complexity $O\left(\xi \sqrt{\frac{L}{\epsilon}}\right)$ even when we do not ignore the effect of ξ , and lead to the same ODE (Su et al., 2014, Section 4.1)

$$\ddot{y} + \frac{1+2\xi}{t} \dot{y} + \operatorname{grad} f(y) = 0.$$

C. Proofs of Lemma 5.2 and Lemma 5.3

Proposition C.1. (Alimisis et al., 2020, Lemma 12) Let γ be a smooth curve whose image is in N , then

$$\frac{d}{dt} \left\| \log_{\gamma(t)}(x) \right\|^2 = 2 \left\langle D_t \log_{\gamma(t)}(x), \log_{\gamma(t)}(x) \right\rangle = 2 \left\langle \log_{\gamma(t)}(x), -\gamma'(t) \right\rangle.$$

Lemma C.2. Let $p_A, p_B, x \in N$ and $v_A \in T_{p_A}M$. If there is $r \in [0, 1]$ such that $\log_{p_A}(p_B) = rv_A$, then we have

$$\begin{aligned} & \left\| v_B - \log_{p_B}(x) \right\|_{p_B}^2 + (\zeta - 1) \|v_B\|_{p_B}^2 \\ & \leq \left\| v_A - \log_{p_A}(x) \right\|_{p_A}^2 + (\zeta - 1) \|v_A\|_{p_A}^2, \end{aligned}$$

where $v_B = \Gamma_{p_A}^{p_B}(v_A - \log_{p_A}(p_B)) \in T_{p_B}M$

Proof. By geodesic unique convexity of N , there is a unique geodesic γ such that $\gamma(0) = p_A$ and $\gamma(r) = p_B$ whose image lies in N . We can check that $\gamma'(0) = v_A$.⁷ Define the vector field $V(t)$ along γ as $V(t) = \Gamma(\gamma)_0^t(v_A - t\gamma'(0))$. Then, we can check that $V(t) = (1-t)\gamma'(t)$ and $V'(t) = -\gamma'(t)$.⁸ Define the function $w : [0, r] \rightarrow \mathbb{R}$ as $w(t) = \left\| \log_{\gamma(t)}(x) - V(t) \right\|^2$, It follows from Proposition 5.1 and Proposition C.1 that

$$\begin{aligned} \frac{d}{dt} w(t) &= 2 \left\langle D_t \left(\log_{\gamma(t)}(x) - V(t) \right), \log_{\gamma(t)}(x) - V(t) \right\rangle \\ &= 2 \left\langle D_t \log_{\gamma(t)}(x), \log_{\gamma(t)}(x) \right\rangle - 2 \left\langle D_t \log_{\gamma(t)}(x), V(t) \right\rangle - 2 \left\langle D_t V(t), \log_{\gamma(t)}(x) \right\rangle + 2 \left\langle D_t V(t), V(t) \right\rangle \\ &= 2 \left\langle D_t \log_{\gamma(t)}(x), \log_{\gamma(t)}(x) \right\rangle - 2(1-t) \left\langle D_t \log_{\gamma(t)}(x), \gamma'(t) \right\rangle + 2 \left\langle \gamma'(t), \log_{\gamma(t)}(x) \right\rangle + 2 \left\langle D_t V(t), V(t) \right\rangle \\ &= 2(1-t) \left\langle D_t \log_{\gamma(t)}(x), -\gamma'(t) \right\rangle + 2 \left\langle D_t V(t), V(t) \right\rangle \\ &\leq 2(1-t)\zeta \|\gamma'(t)\|^2 + 2 \left\langle D_t V(t), V(t) \right\rangle \\ &= -2\zeta \left\langle -\gamma'(t), (1-t)\gamma'(t) \right\rangle + 2 \left\langle D_t V(t), V(t) \right\rangle \\ &= -2(\zeta - 1) \left\langle D_t V(t), V(t) \right\rangle \\ &= -(\zeta - 1) \left(\frac{d}{dt} \|V(t)\|^2 \right). \end{aligned}$$

Integrating both sides from 0 to r gives

$$w(r) - w(0) \leq \int_0^r -(\zeta - 1) \left(\frac{d}{dt} \|V(t)\|^2 \right) dt = -(\zeta - 1) \left(\|V(r)\|^2 - \|V(0)\|^2 \right).$$

This completes the proof. \square

Lemma C.3. Let $p_A, p_B, x \in N$ and $v_A \in T_{p_A}M$. Define $v_B = \Gamma_{p_A}^{p_B}(v_A - \log_{p_A}(p_B)) \in T_{p_B}M$. If there are $a, b \in T_{p_A}M$, and $r \in (0, 1)$ such that $v_A = a + b$ and $\log_{p_A}(p_B) = rb$, then we have

$$\begin{aligned} & \left\| v_B - \log_{p_B}(x) \right\|_{p_B}^2 + (\xi - 1) \|v_B\|_{p_B}^2 \\ & \leq \left\| v_A - \log_{p_A}(x) \right\|_{p_A}^2 + (\xi - 1) \|v_A\|_{p_A}^2 \\ & \quad + \frac{\xi - \delta}{2} \left(\frac{1}{1-r} - 1 \right) \|a\|_{p_A}^2 \end{aligned}$$

for $\xi \geq \zeta$.

⁷Consider the geodesic $c : t \mapsto \gamma(rt)$. Then $c(0) = p_A$ and $c(1) = p_B$. By definition of the exponential map, $c'(0) = \log_{p_A}(p_B) = rv_A$. Combining this equality with $c'(0) = r\gamma'(0)$ gives the desired result.

⁸A similar argument as in the previous footnote shows the first equality. The second equality follows from Proposition A.11 and the fact that $\gamma'(t)$ is parallel along γ .

Proof. Define γ, V, w as in the proof of Lemma 5.2. As in the proof of Lemma 5.2, we can check that $\gamma'(0) = b$ and $V'(t) = -\gamma'(t)$, and that we have

$$\frac{d}{dt}w(t) = -2 \left\langle D_t \log_{\gamma(t)}(x), V(t) \right\rangle + 2 \langle D_t V(t), V(t) \rangle.$$

Consider the smooth function $f_0 : p \mapsto \frac{1}{2} \|\log_p(x)\|^2$. Because $\text{grad } f_0(p) = -\log_p(x)$, we have $\text{Hess } f_0(\gamma(t))[w] = \nabla_w X$, where $X : p \mapsto -\log_p(x)$ (Alimisis et al., 2020, Section 4). By Proposition 5.1, we have $\delta \|w\|^2 \leq \langle \text{Hess } f_0(\gamma(t))[w], w \rangle \leq \zeta \|w\|^2 \leq \xi \|w\|^2$ (Alimisis et al., 2021, Appendix D). Thus,

$$-\frac{\xi - \delta}{2} \|w\|^2 = \delta \|w\|^2 - \frac{\xi + \delta}{2} \|w\|^2 \leq \left\langle \text{Hess } f_0(\gamma(t))[w] - \frac{\xi + \delta}{2} w, w \right\rangle \leq \xi \|w\|^2 - \frac{\xi + \delta}{2} \|w\|^2 = \frac{\xi - \delta}{2} \|w\|^2.$$

Because $\text{Hess } f_0(\gamma(t))$ is self-adjoint, it is diagonalizable. Thus, the norm of the operator $\text{Hess } f_0(\gamma(t)) - \frac{\xi + \delta}{2} I$ on $T_{\gamma(t)}M$ can be bounded as

$$\left\| \text{Hess } f_0(\gamma(t)) - \frac{\xi + \delta}{2} I \right\| \leq \frac{\xi - \delta}{2}.$$

Now, we have

$$\begin{aligned} -2 \left\langle D_t \log_{\gamma(t)}(x), V(t) \right\rangle &= 2 \langle \nabla_{\gamma'(t)} X, V(t) \rangle \\ &= 2 \langle \text{Hess } f_0(\gamma(t))[\gamma'(t)], V(t) \rangle \\ &= 2 \left\langle \left(\text{Hess } f_0(\gamma(t)) - \frac{\xi + \delta}{2} I \right) (\gamma'(t)), V(t) \right\rangle + 2 \left\langle \frac{\xi + \delta}{2} \gamma'(t), V(t) \right\rangle \\ &\leq 2 \left\| \left(\text{Hess } f_0(\gamma(t)) - \frac{\xi + \delta}{2} I \right) (\gamma'(t)) \right\| \|V(t)\| + 2 \left\langle \frac{\xi + \delta}{2} \gamma'(t), V(t) \right\rangle \\ &\leq 2 \left\| \text{Hess } f_0(\gamma(t)) - \frac{\xi + \delta}{2} I \right\| \|\gamma'(t)\| \|V(t)\| + 2 \left\langle \frac{\xi + \delta}{2} \gamma'(t), V(t) \right\rangle \\ &\leq 2 \frac{\xi - \delta}{2} \|\gamma'(t)\| \|V(t)\| + 2 \left\langle \frac{\xi + \delta}{2} \gamma'(t), V(t) \right\rangle. \end{aligned}$$

Because the parallel transport preserves inner product and norm, we obtain

$$\begin{aligned} -2 \left\langle D_t \log_{\gamma(t)}(x), V(t) \right\rangle &\leq 2 \frac{\xi - \delta}{2} \|b\| \|a + (1-t)b\| + (\xi + \delta) \langle b, a + (1-t)b \rangle \\ &= \frac{\xi - \delta}{2} \frac{1}{1-t} 2 \|(1-t)b\| \|a + (1-t)b\| + (\xi + \delta) \langle b, a + (1-t)b \rangle \\ &\leq \frac{\xi - \delta}{2} \frac{1}{1-t} \left(\|(1-t)b\|^2 + \|a + (1-t)b\|^2 \right) + (\xi + \delta) \langle b, a + (1-t)b \rangle \\ &= \frac{\xi - \delta}{2} \frac{1}{1-t} \|a\|^2 - 2\xi \langle -b, a + (1-t)b \rangle \\ &= \frac{\xi - \delta}{2} \frac{1}{1-t} \|a\|^2 - 2\xi \langle D_t V(t), V(t) \rangle. \end{aligned}$$

Thus, for $t \in (0, r)$

$$\frac{d}{dt}w(t) \leq \frac{\xi - \delta}{2} \frac{1}{1-r} \|a\|^2 - 2(\xi - 1) \langle D_t V(t), V(t) \rangle.$$

Integrating both sides from 0 to r , the result follows. \square

D. Convergence Analysis for RGD

In this section, we review the iteration complexity of RGD with the fixed step size $\gamma_k = s$ under the assumptions in Section 3.2. The results in this section correspond to (Zhang & Sra, 2016, Theorems 13 and 15).

D.1. Geodesically convex case

We define the potential function as

$$\phi_k = s(k + \zeta - 1) (f(x_k) - f(x^*)) + \frac{1}{2} \|\log_{x_k}(x^*)\|^2.$$

The following theorem says that ϕ_k is decreasing.

Theorem D.1. *Let f be a geodesically convex and geodesically L -smooth function. If $s \leq \frac{1}{L}$, then the iterates of RGD satisfy*

$$s(k + \zeta) (f(x_{k+1}) - f(x^*)) + \frac{1}{2} \|\log_{x_{k+1}}(x^*)\|^2 \leq s(k + \zeta - 1) (f(x_k) - f(x^*)) + \frac{1}{2} \|\log_{x_k}(x^*)\|^2$$

for all $k \geq 0$.

Proof. (Step 1). In this step, $\langle \cdot, \cdot \rangle$ and $\|\cdot\|$ always denote the inner product and the norm on $T_{x_k}M$. It follows from the geodesic convexity of f that

$$\begin{aligned} f(x^*) &\geq f(x_k) + \langle \text{grad } f(x_k), \log_{x_k}(x^*) \rangle \\ &= f(x_k) - \frac{1}{s} \langle \log_{x_k}(x_{k+1}), \log_{x_k}(x^*) \rangle. \end{aligned}$$

By the geodesic $\frac{1}{s}$ -smoothness of f , we have

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) + \langle \text{grad } f(x_k), \log_{x_k}(x_{k+1}) \rangle + \frac{1}{2s} \|\log_{x_k}(x_{k+1})\|^2 \\ &= f(x_k) - \frac{1}{2s} \|\log_{x_k}(x_{k+1})\|^2. \end{aligned}$$

Taking a weighted sum of these inequalities yields

$$\begin{aligned} 0 &\geq \left[f(x_k) - f(x^*) - \frac{1}{s} \langle \log_{x_k}(x_{k+1}), \log_{x_k}(x^*) \rangle \right] \\ &\quad + (k + \zeta) \left[f(x_{k+1}) - f(x_k) + \frac{1}{2s} \|\log_{x_k}(x_{k+1})\|^2 \right] \\ &= (k + \zeta) (f(x_{k+1}) - f(x^*)) - (k + \zeta - 1) (f(x_k) - f(x^*)) \\ &\quad - \frac{1}{s} \langle \log_{x_k}(x_{k+1}), \log_{x_k}(x^*) \rangle + \frac{k + \zeta}{2s} \|\log_{x_k}(x_{k+1})\|^2 \\ &\geq (k + \zeta) (f(x_{k+1}) - f(x^*)) - (k + \zeta - 1) (f(x_k) - f(x^*)) \\ &\quad - \frac{1}{s} \langle \log_{x_k}(x_{k+1}), \log_{x_k}(x^*) \rangle + \frac{\zeta}{2s} \|\log_{x_k}(x_{k+1})\|^2 \\ &= (k + \zeta) (f(x_{k+1}) - f(x^*)) - (k + \zeta - 1) (f(x_k) - f(x^*)) \\ &\quad + \frac{1}{2s} \left(\zeta \|\log_{x_k}(x_{k+1})\|^2 - 2 \langle \log_{x_k}(x_{k+1}), \log_{x_k}(x^*) \rangle \right) \\ &= (k + \zeta) (f(x_{k+1}) - f(x^*)) - (k + \zeta - 1) (f(x_k) - f(x^*)) \\ &\quad + \frac{1}{2s} \left(\|\log_{x_k}(x_{k+1}) - \log_{x_k}(x^*)\|^2 + (\zeta - 1) \|\log_{x_k}(x_{k+1})\|^2 - \|\log_{x_k}(x^*)\|^2 \right). \end{aligned}$$

(Step 2: Handling metric distortion). By Lemma 5.2 with $p_A = x_k$, $p_B = x_{k+1}$, $x = x^*$, $v_A = \log_{x_k}(x_{k+1})$, $v_B = 0$, $r = 1$, we have

$$\left\| \log_{x_{k+1}}(x^*) \right\|_{x_{k+1}}^2 \leq \left\| \log_{x_k}(x_{k+1}) - \log_{x_k}(x^*) \right\|_{x_k}^2 + (\zeta - 1) \|\log_{x_k}(x_{k+1})\|_{x_k}^2.$$

Combining this inequality with the result in Step 1 gives

$$\begin{aligned}
 0 &\geq (k + \zeta) (f(x_{k+1}) - f(x^*)) - (k + \zeta - 1) (f(x_k) - f(x^*)) \\
 &\quad + \frac{1}{2s} \left(\|\log_{x_k}(x_{k+1}) - \log_{x_k}(x^*)\|_{x_k}^2 + (\zeta - 1) \|\log_{x_k}(x_{k+1})\|_{x_k}^2 - \|\log_{x_k}(x^*)\|_{x_k}^2 \right) \\
 &\quad + \frac{1}{2s} \left(\|\log_{x_{k+1}}(x^*)\|_{x_{k+1}}^2 - \|\log_{x_k}(x_{k+1}) - \log_{x_k}(x^*)\|_{x_k}^2 - (\zeta - 1) \|\log_{x_k}(x_{k+1})\|_{x_k}^2 \right) \\
 &= (k + \zeta) (f(x_{k+1}) - f(x^*)) - (k + \zeta - 1) (f(x_k) - f(x^*)) + \frac{1}{2s} \|\log_{x_{k+1}}(x^*)\|_{x_{k+1}}^2 - \frac{1}{2s} \|\log_{x_k}(x^*)\|_{x_k}^2 \\
 &= \frac{\phi_{k+1} - \phi_k}{s}.
 \end{aligned}$$

□

Corollary D.2. *Let f be a geodesically convex and geodesically L -smooth function. Then, **RGD** with the step size $s = \frac{1}{L}$ finds an ϵ -approximate solution in $O\left(\frac{\zeta L}{\epsilon}\right)$ iterations.*

Proof. It follows from Theorem D.1 that

$$f(x_k) - f(x^*) \leq \frac{\phi_k}{s(k + \zeta - 1)} \leq \frac{\phi_0}{s(k + \zeta - 1)} = \frac{1}{s(k + \zeta - 1)} \left(s(\zeta - 1) (f(x_0) - f(x^*)) + \frac{1}{2} \|\log_{x_0}(x^*)\|_{x_0}^2 \right).$$

By geodesic $\frac{1}{s}$ -smoothness of f , we have

$$f(x_k) - f(x^*) \leq \frac{1}{s(k + \zeta - 1)} \left(s(\zeta - 1) \frac{1}{2s} \|\log_{x_0}(x^*)\|_{x_0}^2 + \frac{1}{2} \|\log_{x_0}(x^*)\|_{x_0}^2 \right) = \frac{\zeta L}{2(k + \zeta - 1)} \|\log_{x_0}(x^*)\|_{x_0}^2.$$

Thus, we have $f(x_k) - f(x^*) \leq \epsilon$ whenever $k \geq \frac{\zeta L}{2\epsilon} \|\log_{x_0}(x^*)\|_{x_0}^2 - (\zeta - 1)$. Thus we obtain an $O\left(\frac{\zeta L}{\epsilon}\right)$ iteration complexity. □

This result implies that the iteration complexity of **RGD** for geodesically convex case is the same as that of **GD**, since ζ is a constant.

D.2. Geodesically strongly convex case

We define the potential function as

$$\phi_k = (1 - \mu s)^{-k} \left(f(x_k) - f(x^*) + \frac{\mu}{2} \|\log_{x_k}(x^*)\|_{x_k}^2 \right).$$

The following theorem states that ϕ_k is decreasing.

Theorem D.3. *Let f be a geodesically μ -strongly convex and geodesically L -smooth function. If $s \leq \min\left\{\frac{1}{L}, \frac{1}{\zeta\mu}\right\}$, then the iterates of **RGD** satisfy*

$$(1 - \mu s)^{-(k+1)} \left(f(x_{k+1}) - f(x^*) + \frac{\mu}{2} \|\log_{x_{k+1}}(x^*)\|_{x_{k+1}}^2 \right) \leq (1 - \mu s)^{-k} \left(f(x_k) - f(x^*) + \frac{\mu}{2} \|\log_{x_k}(x^*)\|_{x_k}^2 \right)$$

for all $k \geq 0$.

Proof. (Step 1). In this step, $\langle \cdot, \cdot \rangle$ and $\|\cdot\|$ always denote the inner product and the norm on $T_{x_k}M$. Set $q = \mu s$. By geodesic μ -strong convexity of f , we have

$$\begin{aligned}
 f(x^*) &\geq f(x_k) + \langle \text{grad } f(x_k), \log_{x_k}(x^*) \rangle + \frac{\mu}{2} \|\log_{x_k}(x^*)\|_{x_k}^2 \\
 &= f(x_k) - \frac{1}{s} \langle \log_{x_k}(x_{k+1}), \log_{x_k}(x^*) \rangle + \frac{q}{2s} \|\log_{x_k}(x^*)\|_{x_k}^2.
 \end{aligned}$$

By geodesic $\frac{1}{s}$ -smoothness of f , we have

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) + \langle \text{grad } f(x_k), \log_{x_k}(x_{k+1}) \rangle + \frac{1}{2s} \|\log_{x_k}(x_{k+1})\|^2 \\ &= f(x_k) - \frac{1}{2s} \|\log_{x_k}(x_{k+1})\|^2. \end{aligned}$$

Note that $\zeta q \leq 1$. Taking weighted sum of these inequalities, we arrive to the valid inequality

$$\begin{aligned} 0 &\geq q \left[f(x_k) - f(x^*) - \frac{1}{s} \langle \log_{x_k}(x_{k+1}), \log_{x_k}(x^*) \rangle + \frac{q}{2s} \|\log_{x_k}(x^*)\|^2 \right] \\ &\quad + \left[f(x_{k+1}) - f(x_k) + \frac{1}{2s} \|\log_{x_k}(x_{k+1})\|^2 \right] \\ &= f(x_{k+1}) - f(x^*) - (1-q)(f(x_k) - f(x^*)) \\ &\quad - \frac{q}{s} \langle \log_{x_k}(x_{k+1}), \log_{x_k}(x^*) \rangle + \frac{q^2}{2s} \|\log_{x_k}(x^*)\|^2 + \frac{1}{2s} \|\log_{x_k}(x_{k+1})\|^2 \\ &\geq f(x_{k+1}) - f(x^*) - (1-q)(f(x_k) - f(x^*)) \\ &\quad + \frac{q}{2s} \left(-2 \langle \log_{x_k}(x_{k+1}), \log_{x_k}(x^*) \rangle + q \|\log_{x_k}(x^*)\|^2 + \zeta \|\log_{x_k}(x_{k+1})\|^2 \right) \\ &= f(x_{k+1}) - f(x^*) - (1-q)(f(x_k) - f(x^*)) \\ &\quad + \frac{q}{2s} \left(\|\log_{x_k}(x_{k+1}) - \log_{x_k}(x^*)\|^2 + (\zeta - 1) \|\log_{x_k}(x_{k+1})\|^2 - (1-q) \|\log_{x_k}(x^*)\|^2 \right). \end{aligned}$$

(Step 2: Handle metric distortion). By Lemma 5.2 with $p_A = x_k, p_B = x_{k+1}, x = x^*, v_A = \log_{x_k}(x_{k+1}), v_B = 0, r = 1$, we have

$$\left\| \log_{x_{k+1}}(x^*) \right\|_{x_{k+1}}^2 \leq \left\| \log_{x_k}(x_{k+1}) - \log_{x_k}(x^*) \right\|_{x_k}^2 + (\zeta - 1) \left\| \log_{x_k}(x_{k+1}) \right\|_{x_k}^2.$$

Combining this inequality with the result in Step 1 gives

$$\begin{aligned} 0 &\geq f(x_{k+1}) - f(x^*) - (1-q)(f(x_k) - f(x^*)) \\ &\quad + \frac{q}{2s} \left(\left\| \log_{x_k}(x_{k+1}) - \log_{x_k}(x^*) \right\|_{x_k}^2 + (\zeta - 1) \left\| \log_{x_k}(x_{k+1}) \right\|_{x_k}^2 - (1-q) \left\| \log_{x_k}(x^*) \right\|_{x_k}^2 \right) \\ &\quad + \frac{q}{2s} \left(\left\| \log_{x_{k+1}}(x^*) \right\|_{x_{k+1}}^2 - \left\| \log_{x_k}(x_{k+1}) - \log_{x_k}(x^*) \right\|_{x_k}^2 - (\zeta - 1) \left\| \log_{x_k}(x_{k+1}) \right\|_{x_k}^2 \right) \\ 0 &= f(x_{k+1}) - f(x^*) - (1-q)(f(x_k) - f(x^*)) + \frac{q}{2s} \left\| \log_{x_{k+1}}(x^*) \right\|_{x_{k+1}}^2 - \frac{q}{2s} (1-q) \left\| \log_{x_k}(x^*) \right\|_{x_k}^2 \\ &= \left(f(x_{k+1}) - f(x^*) + \frac{\mu}{2} \left\| \log_{x_{k+1}}(x^*) \right\|_{x_{k+1}}^2 \right) - (1-q) \left(f(x_k) - f(x^*) + \frac{\mu}{2} \left\| \log_{x_k}(x^*) \right\|_{x_k}^2 \right) \\ &= (1-q)^{(k+1)} (\phi_{k+1} - \phi_k). \end{aligned}$$

□

Corollary D.4. *Let f be a geodesically μ -strongly convex and geodesically L -smooth function. Then, RGD with step size $s = \frac{1}{\zeta L}$ finds an ϵ -approximate solution in $O\left(\frac{\zeta L}{\mu} \log \frac{L}{\epsilon}\right)$ iterations.*

Proof. By Theorem D.3, we have

$$f(x_k) - f(x^*) \leq (1 - \mu s)^k \phi_k \leq (1 - \mu s)^k \phi_0 = (1 - \mu s)^k \left(f(x_0) - f(x^*) + \frac{\mu}{2} \|\log_{x_0}(x^*)\|^2 \right).$$

It follows from the geodesic L -smoothness of f and the inequality $\left(1 - \frac{\mu}{\zeta L}\right)^k \leq e^{-\frac{\mu}{\zeta L} k}$ that

$$f(x_k) - f(x^*) \leq \left(1 - \frac{\mu}{\zeta L}\right)^k \left(\frac{L}{2} \|\log_{x_0}(x^*)\|^2 + \frac{\mu}{2} \|\log_{x_0}(x^*)\|^2 \right) \leq e^{-\frac{\mu}{\zeta L} k} L \|\log_{x_0}(x^*)\|^2.$$

Thus, we have $f(x_k) - f(x^*) \leq \epsilon$ whenever $k \geq \frac{\zeta L}{\mu} \log\left(\frac{L}{\epsilon} \|\log_{x_0}(x^*)\|^2\right)$. Accordingly, we obtain an $O\left(\frac{\zeta L}{\mu} \log \frac{L}{\epsilon}\right)$ iteration complexity. \square

This result implies that the iteration complexity of **RGD** for g -strongly convex case is the same as that of **GD**, since ζ is a constant. Another proof of the iteration complexity of **RGD** for g -strongly convex functions can be found in (Criscitiello & Boumal, 2021, Proposition 1.8).

E. Convergence Analysis for RNAG-C

Theorem 5.4. *Let f be a g -convex and geodesically L -smooth function. If the parameters ξ and T of RNAG-C satisfy $\xi \geq \zeta$ and*

$$\begin{aligned} & \frac{\xi - \delta}{2} \left(\frac{1}{1 - \xi/\lambda_k} - 1 \right) \\ & \leq (\xi - \zeta) \left(\frac{1}{(1 - \xi/(\lambda_k + \xi - 1))^2} - 1 \right) \end{aligned}$$

for all $k \geq 0$, then the iterates of RNAG-C satisfy $\phi_{k+1} \leq \phi_k$ for all $k \geq 0$, where ϕ_k is defined as (5).

Proof. (Step 1). In this step, $\langle \cdot, \cdot \rangle$ and $\|\cdot\|$ always denote the inner product and the norm on $T_{y_k}M$. It is easy to check that $\text{grad } f(y_k) = -\frac{\xi}{s\lambda_k}(\bar{v}_{k+1} - v_k)$, $\log_{y_k}(x_k) = -\frac{\xi}{\lambda_k - 1}v_k$,⁹ and $\lambda_k^2 - \lambda_k \leq \lambda_{k-1}^2$. By the geodesic convexity of f , we have

$$\begin{aligned} f(x^*) & \geq f(y_k) + \langle \text{grad } f(y_k), \log_{y_k}(x^*) \rangle \\ & = f(y_k) - \frac{\xi}{s\lambda_k} \langle \bar{v}_{k+1} - v_k, \log_{y_k}(x^*) \rangle, \\ f(x_k) & \geq f(y_k) + \langle \text{grad } f(y_k), \log_{y_k}(x_k) \rangle \\ & = f(y_k) + \frac{\xi^2}{s(\lambda_k^2 - \lambda_k)} \langle \bar{v}_{k+1} - v_k, v_k \rangle. \end{aligned}$$

It follows from the geodesic $\frac{1}{s}$ -smoothness of f that

$$\begin{aligned} f(x_{k+1}) & \leq f(y_k) + \langle \text{grad } f(y_k), \log_{y_k}(x_{k+1}) \rangle + \frac{1}{2s} \|\log_{y_k}(x_{k+1})\|^2 \\ & = f(y_k) - \frac{s}{2} \|\text{grad } f(y_k)\|^2 \\ & = f(y_k) - \frac{\xi^2}{2s\lambda_k^2} \|\bar{v}_{k+1} - v_k\|^2. \end{aligned}$$

⁹Note that $y_k = \exp_{x_k}\left(\frac{\xi}{\lambda_k + (\xi - 1)}\bar{v}_k\right)$ and $v_k = \Gamma_{x_k}^{y_k}(\bar{v}_k - \log_{x_k}(y_k)) = \Gamma_{x_k}^{y_k}\left(\left(1 - \frac{\xi}{\lambda_k + (\xi - 1)}\right)\bar{v}_k\right)$. Let γ_1 be the geodesic such that $\gamma_1(0) = x_k$ and $\gamma_1(1) = y_k$, then $\gamma_1'(0) = \log_{x_k}(y_k)$. Let γ_2 be the geodesic defined as $\gamma_2(t) = \gamma_1(1 - t)$. Then, $\log_{y_k}(x_k) = \gamma_2'(0) = -\gamma_1'(1) = -\Gamma_{x_k}^{y_k}(\gamma_1'(0)) = -\Gamma_{x_k}^{y_k}(\log_{x_k}(y_k))$. Now, we have $\log_{y_k}(x_k) = -\Gamma_{x_k}^{y_k}(\log_{x_k}(y_k)) = -\frac{\xi}{\lambda_k + (\xi - 1)}\Gamma_{x_k}^{y_k}(\bar{v}_k) = -\frac{\frac{\xi}{\lambda_k + (\xi - 1)}}{1 - \frac{\xi}{\lambda_k + (\xi - 1)}}v_k = -\frac{\xi}{\lambda_k - 1}v_k$.

Taking a weighted sum of these inequalities yields

$$\begin{aligned}
 0 &\geq \lambda_k \left[f(y_k) - f(x^*) - \frac{\xi}{s\lambda_k} \langle \bar{v}_{k+1} - v_k, \log_{y_k}(x^*) \rangle \right] \\
 &\quad + (\lambda_k^2 - \lambda_k) \left[f(y_k) - f(x_k) + \frac{\xi^2}{s(\lambda_k^2 - \lambda_k)} \langle \bar{v}_{k+1} - v_k, v_k \rangle \right] \\
 &\quad + \lambda_k^2 \left[f(x_{k+1}) - f(y_k) + \frac{\xi^2}{2s\lambda_k^2} \|\bar{v}_{k+1} - v_k\|^2 \right] \\
 &= \lambda_k^2 (f(x_{k+1}) - f(x^*)) - (\lambda_k^2 - \lambda_k) (f(x_k) - f(x^*)) \\
 &\quad - \frac{\xi}{s} \langle \bar{v}_{k+1} - v_k, \log_{y_k}(x^*) \rangle + \frac{\xi^2}{s} \langle \bar{v}_{k+1} - v_k, v_k \rangle + \frac{\xi^2}{2s} \|\bar{v}_{k+1} - v_k\|^2 \\
 &\geq \lambda_k^2 (f(x_{k+1}) - f(x^*)) - \lambda_{k-1}^2 (f(x_k) - f(x^*)) \\
 &\quad + \frac{\xi}{2s} \left(-2 \langle \bar{v}_{k+1} - v_k, \log_{y_k}(x^*) \rangle + 2\xi \langle \bar{v}_{k+1} - v_k, v_k \rangle + \xi \|\bar{v}_{k+1} - v_k\|^2 \right) \\
 &= \lambda_k^2 (f(x_{k+1}) - f(x^*)) - \lambda_{k-1}^2 (f(x_k) - f(x^*)) \\
 &\quad + \frac{\xi}{2s} \left(\|\bar{v}_{k+1} - v_k\|^2 - 2 \langle \bar{v}_{k+1} - v_k, \log_{y_k}(x^*) - v_k \rangle + 2(\xi - 1) \langle \bar{v}_{k+1} - v_k, v_k \rangle + (\xi - 1) \|\bar{v}_{k+1} - v_k\|^2 \right) \\
 &= \lambda_k^2 (f(x_{k+1}) - f(x^*)) - \lambda_{k-1}^2 (f(x_k) - f(x^*)) \\
 &\quad + \frac{\xi}{2s} \left(\|\bar{v}_{k+1} - v_k\|^2 - 2 \langle \bar{v}_{k+1} - v_k, \log_{y_k}(x^*) - v_k \rangle + (\xi - 1) \|\bar{v}_{k+1}\|^2 - (\xi - 1) \|v_k\|^2 \right).
 \end{aligned}$$

Note that

$$\|\bar{v}_{k+1} - \log_{y_k}(x^*)\|^2 - \|v_k - \log_{y_k}(x^*)\|^2 = \|\bar{v}_{k+1} - v_k\|^2 - 2 \langle \bar{v}_{k+1} - v_k, \log_{y_k}(x^*) - v_k \rangle.$$

Thus, we obtain

$$\begin{aligned}
 0 &\geq \lambda_k^2 (f(x_{k+1}) - f(x^*)) - \lambda_{k-1}^2 (f(x_k) - f(x^*)) \\
 &\quad + \frac{\xi}{2s} \left(\|\bar{v}_{k+1} - \log_{y_k}(x^*)\|^2 - \|v_k - \log_{y_k}(x^*)\|^2 + (\xi - 1) \|\bar{v}_{k+1}\|^2 - (\xi - 1) \|v_k\|^2 \right).
 \end{aligned}$$

(Step 2: Handle metric distortion). By Lemma 5.3 with $p_A = y_k$, $p_B = x_{k+1}$, $x = x^*$, $v_A = \bar{v}_{k+1}$, $v_B = \bar{v}_{k+1}$, $a = v_k$, $b = -\gamma_k \text{grad} f(y_k) = -\frac{s\lambda_k}{\xi} \text{grad} f(y_k)$, $r = \frac{s}{\gamma_k} = \frac{\xi}{\lambda_k} \in (0, 1)$, we have

$$\begin{aligned}
 &\left\| \log_{x_{k+1}}(x^*) - \bar{v}_{k+1} \right\|_{x_{k+1}}^2 + (\xi - 1) \|\bar{v}_{k+1}\|_{x_{k+1}}^2 \\
 &\leq \left\| \log_{y_k}(x^*) - \bar{v}_{k+1} \right\|_{y_k}^2 + (\xi - 1) \|\bar{v}_{k+1}\|_{y_k}^2 + \frac{\xi - \delta}{2} \left(\frac{1}{1 - \xi/\lambda_k} - 1 \right) \|v_k\|_{y_k}^2.
 \end{aligned}$$

It follows from Lemma 5.2 with $p_A = x_k$, $p_B = y_k$, $x = x^*$, $v_A = \bar{v}_k$, $v_B = v_k$, $r = \tau_k = \frac{\xi}{\lambda_k + \xi - 1}$ that

$$\begin{aligned}
 \left\| \log_{x_k}(x^*) - \bar{v}_k \right\|_{x_k}^2 + (\xi - 1) \|\bar{v}_k\|_{x_k}^2 &= \left(\left\| \log_{x_k}(x^*) - \bar{v}_k \right\|_{x_k}^2 + (\zeta - 1) \|\bar{v}_k\|_{x_k}^2 \right) + (\xi - \zeta) \|\bar{v}_k\|_{x_k}^2 \\
 &\geq \left(\left\| \log_{y_k}(x^*) - v_k \right\|_{y_k}^2 + (\zeta - 1) \|v_k\|_{y_k}^2 \right) + (\xi - \zeta) \|\bar{v}_k\|_{x_k}^2 \\
 &= \left\| \log_{y_k}(x^*) - v_k \right\|_{y_k}^2 + (\zeta - 1) \|v_k\|_{y_k}^2 + (\xi - \zeta) \frac{1}{(1 - \tau_k)^2} \|v_k\|_{y_k}^2 \\
 &= \left\| \log_{y_k}(x^*) - v_k \right\|_{y_k}^2 + (\xi - 1) \|v_k\|_{y_k}^2 + (\xi - \zeta) \left(\frac{1}{(1 - \tau_k)^2} - 1 \right) \|v_k\|_{y_k}^2,
 \end{aligned}$$

Combining these inequalities with the result in Step 1 gives

$$\begin{aligned}
 0 &\geq s\lambda_k^2 (f(x_{k+1}) - f(x^*)) - \lambda_{k-1}^2 (f(x_k) - f(x^*)) \\
 &\quad + \frac{\xi}{2} \left(\|\bar{v}_{k+1} - \log_{y_k}(x^*)\|^2 + (\xi - 1) \|\bar{v}_{k+1}\|^2 - \|v_k - \log_{y_k}(x^*)\|^2 - (\xi - 1) \|v_k\|^2 \right) \\
 &\quad + \frac{\xi}{2} \left[\left\| \log_{x_{k+1}}(x^*) - \bar{v}_{k+1} \right\|_{x_{k+1}}^2 + (\xi - 1) \|\bar{v}_{k+1}\|_{x_{k+1}}^2 \right. \\
 &\quad \left. - \left\| \log_{y_k}(x^*) - \bar{v}_{k+1} \right\|_{y_k}^2 - (\xi - 1) \|\bar{v}_{k+1}\|_{y_k}^2 - \frac{\xi - \delta}{2} \left(\frac{1}{1 - \xi/\lambda_k} - 1 \right) \|v_k\|_{y_k}^2 \right] \\
 &\quad + \frac{\xi}{2} \left[\left\| \log_{y_k}(x^*) - v_k \right\|_{y_k}^2 + (\xi - 1) \|v_k\|_{y_k}^2 + (\xi - \zeta) \left(\frac{1}{(1 - \tau_k)^2} - 1 \right) \|v_k\|_{y_k}^2 \right. \\
 &\quad \left. - \left\| \log_{x_k}(x^*) - \bar{v}_k \right\|_{x_k}^2 - (\xi - 1) \|\bar{v}_k\|_{x_k}^2 \right] \\
 &= \phi_{k+1} - \phi_k + \frac{\xi}{2} \left((\xi - \zeta) \left(\frac{1}{(1 - \tau_k)^2} - 1 \right) - \frac{\xi - \delta}{2} \left(\frac{1}{1 - \xi/\lambda_k} - 1 \right) \right) \|v_k\|_{y_k}^2 \\
 &\geq \phi_{k+1} - \phi_k.
 \end{aligned}$$

□

Corollary E.1. *Let f be a g -convex and geodesically L -smooth function. Then, RNAG-C with parameters $\xi = \zeta + 3(\zeta - \delta)$, $T = 4\xi$ and step size $s = \frac{1}{L}$ finds an ϵ -approximate solution in $O\left(\xi\sqrt{\frac{L}{\epsilon}}\right)$ iterations.*

Proof. (Step 1: Checking the condition for Theorem 5.4). A straightforward calculation shows that

$$2\left(\frac{1}{1-t} - 1\right) \leq 3\left(\frac{1}{1-(3/4)t} - 1\right).$$

for all $t \in (0, 1/3]$. For convenience, let $r = \frac{s}{\gamma_k} = \frac{\xi}{\lambda_k} = \frac{2\xi}{k+6\xi} \in (0, 1/3]$. Then, $\tau_k = \frac{\xi}{\lambda_k + (\xi - 1)} = \frac{2\xi}{k+6\xi+2(\xi-1)} \geq \frac{2\xi}{k+8\xi} \geq \frac{2\xi}{\frac{4}{3}(k+6\xi)} = \frac{3}{4}r$. Now, we have

$$\begin{aligned}
 (\xi - \zeta) \left(\frac{1}{(1 - \tau_k)^2} - 1 \right) &\geq (\xi - \zeta) \left(\frac{1}{1 - \tau_k} - 1 \right) \\
 &\geq (\xi - \zeta) \left(\frac{1}{1 - \frac{3}{4}r} - 1 \right) \\
 &\geq \frac{\xi - \delta}{2} \left(\frac{1}{1 - r} - 1 \right).
 \end{aligned}$$

(Step 2: Computing iteration complexity). By Theorem 5.4, we have

$$f(x_k) - f(x^*) \leq \frac{\phi_k}{s\lambda_{k-1}^2} \leq \frac{\phi_0}{s\lambda_{k-1}^2} = \frac{1}{s\lambda_{k-1}^2} \left(s\lambda_{-1}^2 (f(x_0) - f(x^*)) + \frac{\xi}{2} \|\log_{x_0}(x^*)\|^2 \right).$$

It follows from the geodesic $\frac{1}{s}$ -smoothness of f that

$$\begin{aligned}
 f(x_k) - f(x^*) &\leq \frac{1}{s\lambda_{k-1}^2} \left(s\lambda_{-1}^2 \frac{1}{2s} \|\log_{x_0}(x^*)\|^2 + \frac{\xi}{2} \|\log_{x_0}(x^*)\|^2 \right) \\
 &= \frac{1}{s\lambda_{k-1}^2} \left(\frac{\lambda_{-1}^2}{2} + \frac{\xi}{2} \right) \|\log_{x_0}(x^*)\|^2 \\
 &= \frac{4L}{(k-1+6\xi)^2} \left(\frac{(6\xi-1)^2}{8} + \frac{\xi}{2} \right) \|\log_{x_0}(x^*)\|^2 \\
 &\leq \frac{4L}{(k-1)^2} \left(\frac{(6\xi-1)^2}{8} + \frac{\xi}{2} \right) \|\log_{x_0}(x^*)\|^2.
 \end{aligned}$$

Thus, we have $f(x_k) - f(x^*) \leq \epsilon$ whenever

$$(k-1)^2 \geq \frac{4L}{\epsilon} \left(\frac{(6\xi-1)^2}{8} + \frac{\xi}{2} \right) \|\log_{x_0}(x^*)\|^2.$$

This implies that RNAG-C has an $O\left(\xi\sqrt{\frac{L}{\epsilon}}\right)$ iteration complexity. \square

F. Convergence Analysis for RNAG-SC

Theorem 5.6. *Let f be a geodesically μ -strongly convex and geodesically L -smooth function. If the step size s and the parameter ξ of RNAG-SC satisfy $\xi \geq \zeta$, $\sqrt{\xi q} < 1$, and*

$$\begin{aligned} & \frac{\xi - \delta}{2} \left(\frac{1}{1 - \sqrt{\xi q}} - 1 \right) \left(1 - \sqrt{\frac{q}{\xi}} \right)^2 - \sqrt{\xi q} \left(1 - \sqrt{\frac{q}{\xi}} \right) \\ & \leq (\xi - \zeta) \left(\frac{1}{\left(1 - \sqrt{\xi q} / (1 + \sqrt{\xi q}) \right)^2} - 1 \right), \end{aligned}$$

then the iterates of RNAG-SC satisfy $\phi_{k+1} \leq \phi_k$ for all $k \geq 0$, where ϕ_k is defined as (6).

Proof. (Step 1). In this step, $\langle \cdot, \cdot \rangle$ and $\|\cdot\|$ always denote the inner product and the norm on $T_{y_k}M$. Set $q = \mu s$. It is straightforward to check that $\text{grad } f(y_k) = \mu \frac{1 - \sqrt{q/\xi}}{\sqrt{q/\xi}} v_k - \mu \frac{1}{\sqrt{q/\xi}} \bar{v}_{k+1}$ and $\log_{y_k}(x_k) = -\sqrt{\xi q} v_k$.¹⁰ By geodesic μ -strong convexity of f , we have

$$\begin{aligned} f(x^*) & \geq f(y_k) + \langle \text{grad } f(y_k), \log_{y_k}(x^*) \rangle + \frac{\mu}{2} \|\log_{y_k}(x^*)\|^2 \\ & = f(y_k) + \mu \frac{1 - \sqrt{q/\xi}}{\sqrt{q/\xi}} \langle v_k, \log_{y_k}(x^*) \rangle - \mu \frac{1}{\sqrt{q/\xi}} \langle \bar{v}_{k+1}, \log_{y_k}(x^*) \rangle + \frac{\mu}{2} \|\log_{y_k}(x^*)\|^2. \end{aligned}$$

It follows from the geodesic convexity of f that

$$\begin{aligned} f(x_k) & \geq f(y_k) + \langle \text{grad } f(y_k), \log_{y_k}(x_k) \rangle \\ & = f(y_k) - \xi \mu \left(1 - \sqrt{\frac{q}{\xi}} \right) \|v_k\|^2 + \xi \mu \langle v_k, \bar{v}_{k+1} \rangle. \end{aligned}$$

By the geodesic $\frac{1}{s}$ -smoothness of f , we have

$$\begin{aligned} f(x_{k+1}) & \leq f(y_k) + \langle \text{grad } f(y_k), \log_{y_k}(x_{k+1}) \rangle + \frac{1}{2s} \|\log_{y_k}(x_{k+1})\|^2 \\ & = f(y_k) - \frac{s}{2} \|\text{grad } f(y_k)\|^2 \\ & = f(y_k) - \frac{s}{2} \left\| \mu \frac{1 - \sqrt{q/\xi}}{\sqrt{q/\xi}} v_k - \mu \frac{1}{\sqrt{q/\xi}} \bar{v}_{k+1} \right\|^2 \\ & = f(y_k) - \frac{\xi \mu}{2} \left(1 - \sqrt{\frac{q}{\xi}} \right)^2 \|v_k\|^2 + \xi \mu \left(1 - \sqrt{\frac{q}{\xi}} \right) \langle v_k, \bar{v}_{k+1} \rangle - \frac{\xi \mu}{2} \|\bar{v}_{k+1}\|^2. \end{aligned}$$

¹⁰Note that $y_k = \exp_{x_k} \left(\frac{\sqrt{\xi q}}{1 + \sqrt{\xi q}} \bar{v}_k \right)$ and $v_k = \Gamma_{x_k}^{y_k}(\bar{v}_k - \log_{x_k}(y_k)) = \Gamma_{x_k}^{y_k} \left(\left(1 - \frac{\sqrt{\xi q}}{1 + \sqrt{\xi q}} \right) \bar{v}_k \right)$. Let γ_1 be the geodesic such that $\gamma_1(0) = x_k$ and $\gamma_1(1) = y_k$, then $\gamma_1'(0) = \log_{x_k}(y_k)$. Let γ_2 be the geodesic defined as $\gamma_2(t) = \gamma_1(1-t)$. Then $\log_{y_k}(x_k) = \gamma_2'(0) = -\gamma_1'(1) = -\Gamma_{x_k}^{y_k}(\gamma_1'(0)) = -\Gamma_{x_k}^{y_k}(\log_{x_k}(y_k))$. Now, we have $\log_{y_k}(x_k) = -\Gamma_{x_k}^{y_k}(\log_{x_k}(y_k)) = -\frac{\sqrt{\xi q}}{1 + \sqrt{\xi q}} \Gamma_{x_k}^{y_k}(\bar{v}_k) = -\frac{\frac{\sqrt{\xi q}}{1 + \sqrt{\xi q}}}{1 - \frac{\sqrt{\xi q}}{1 + \sqrt{\xi q}}} = \sqrt{\xi q} v_k$.

Taking a weighted sum of these inequalities yields

$$\begin{aligned}
 0 &\geq \sqrt{\frac{q}{\xi}} \left[f(y_k) - f(x^*) + \mu \frac{1 - \sqrt{q/\xi}}{\sqrt{q/\xi}} \langle v_k, \log_{y_k}(x^*) \rangle - \mu \frac{1}{\sqrt{q/\xi}} \langle \bar{v}_{k+1}, \log_{y_k}(x^*) \rangle + \frac{\mu}{2} \|\log_{y_k}(x^*)\|^2 \right] \\
 &\quad + \left(1 - \sqrt{\frac{q}{\xi}} \right) \left[f(y_k) - f(x_k) - \xi\mu \left(1 - \sqrt{\frac{q}{\xi}} \right) \|v_k\|^2 + \xi\mu \langle v_k, \bar{v}_{k+1} \rangle \right] \\
 &\quad + \left[f(x_{k+1}) - f(y_k) + \frac{\xi\mu}{2} \left(1 - \sqrt{\frac{q}{\xi}} \right)^2 \|v_k\|^2 - \xi\mu \left(1 - \sqrt{\frac{q}{\xi}} \right) \langle v_k, \bar{v}_{k+1} \rangle + \frac{\xi\mu}{2} \|\bar{v}_{k+1}\|^2 \right] \\
 &= (f(x_{k+1}) - f(x^*)) - \left(1 - \sqrt{\frac{q}{\xi}} \right) (f(x_k) - f(x^*)) \\
 &\quad + \mu \left(1 - \sqrt{\frac{q}{\xi}} \right) \langle v_k, \log_{y_k}(x^*) \rangle - \mu \langle \bar{v}_{k+1}, \log_{y_k}(x^*) \rangle + \frac{\mu}{2} \sqrt{\frac{q}{\xi}} \|\log_{y_k}(x^*)\|^2 \\
 &\quad - \frac{\xi\mu}{2} \left(1 - \sqrt{\frac{q}{\xi}} \right)^2 \|v_k\|^2 + \frac{\xi\mu}{2} \|\bar{v}_{k+1}\|^2.
 \end{aligned}$$

We further notice that

$$\begin{aligned}
 &\|\bar{v}_{k+1} - \log_{y_k}(x^*)\|^2 - \left(1 - \sqrt{\frac{q}{\xi}} \right) \|v_k - \log_{y_k}(x^*)\|^2 \\
 &= \|\bar{v}_{k+1}\|^2 - 2 \langle \bar{v}_{k+1}, \log_{y_k}(x^*) \rangle - \left(1 - \sqrt{\frac{q}{\xi}} \right) \|v_k\|^2 + 2 \left(1 - \sqrt{\frac{q}{\xi}} \right) \langle v_k, \log_{y_k}(x^*) \rangle + \sqrt{\frac{q}{\xi}} \|\log_{y_k}(x^*)\|^2.
 \end{aligned}$$

Therefore, we obtain

$$\begin{aligned}
 0 &\geq \left(f(x_{k+1}) - f(x^*) + \frac{\mu}{2} \|\bar{v}_{k+1} - \log_{y_k}(x^*)\|^2 \right) - \left(1 - \sqrt{\frac{q}{\xi}} \right) \left(f(x_k) - f(x^*) + \frac{\mu}{2} \|v_k - \log_{y_k}(x^*)\|^2 \right) \\
 &\quad + (\xi - 1) \frac{\mu}{2} \|\bar{v}_{k+1}\|^2 - \frac{\xi\mu}{2} \left(1 - \sqrt{\frac{q}{\xi}} \right)^2 \|v_k\|^2 + \frac{\mu}{2} \left(1 - \sqrt{\frac{q}{\xi}} \right) \|v_k\|^2 \\
 &= \left(f(x_{k+1}) - f(x^*) + \frac{\mu}{2} \|\bar{v}_{k+1} - \log_{y_k}(x^*)\|^2 \right) - \left(1 - \sqrt{\frac{q}{\xi}} \right) \left(f(x_k) - f(x^*) + \frac{\mu}{2} \|v_k - \log_{y_k}(x^*)\|^2 \right) \\
 &\quad + (\xi - 1) \frac{\mu}{2} \|\bar{v}_{k+1}\|^2 - (\xi - 1) \frac{\mu}{2} \left(1 - \sqrt{\frac{q}{\xi}} \right) \|v_k\|^2 + \frac{\xi\mu}{2} \sqrt{\frac{q}{\xi}} \left(1 - \sqrt{\frac{q}{\xi}} \right) \|v_k\|^2 \\
 &= \left(f(x_{k+1}) - f(x^*) + \frac{\mu}{2} \|\bar{v}_{k+1} - \log_{y_k}(x^*)\|^2 + (\xi - 1) \frac{\mu}{2} \|\bar{v}_{k+1}\|^2 \right) \\
 &\quad - \left(1 - \sqrt{\frac{q}{\xi}} \right) \left(f(x_k) - f(x^*) + \frac{\mu}{2} \|v_k - \log_{y_k}(x^*)\|^2 + (\xi - 1) \frac{\mu}{2} \|v_k\|^2 \right) \\
 &\quad + \frac{\xi\mu}{2} \sqrt{\frac{q}{\xi}} \left(1 - \sqrt{\frac{q}{\xi}} \right) \|v_k\|^2.
 \end{aligned}$$

(Step 2: Handle metric distortion). It follows from Lemma 5.3 with $p_A = y_k, p_B = x_{k+1}, x = x^*, v_A = \bar{v}_{k+1}, v_B = \bar{v}_{k+1}, a = \left(1 - \sqrt{\frac{q}{\xi}} \right) v_k, b = \sqrt{\frac{q}{\xi}} \left(-\frac{1}{\mu} \right) \text{grad } f(y_k), r = \sqrt{\xi q}$ that

$$\begin{aligned}
 &\left\| \log_{x_{k+1}}(x^*) - \bar{v}_{k+1} \right\|_{x_{k+1}}^2 + (\xi - 1) \|\bar{v}_{k+1}\|_{x_{k+1}}^2 \\
 &\leq \left\| \log_{y_k}(x^*) - \bar{v}_{k+1} \right\|_{y_k}^2 + (\xi - 1) \|\bar{v}_{k+1}\|_{y_k}^2 + \frac{\xi - \delta}{2} \left(\frac{1}{1 - \sqrt{\xi q}} - 1 \right) \left\| \left(1 - \sqrt{\frac{q}{\xi}} \right) v_k \right\|_{y_k}^2.
 \end{aligned}$$

Applying Lemma 5.2 with $p_A = x_k, p_B = y_k, x = x^*, v_A = \bar{v}_k, v_B = v_k, r = \frac{\sqrt{\xi q}}{1 + \sqrt{\xi q}}$ gives

$$\begin{aligned}
 & \left\| \log_{x_k}(x^*) - \bar{v}_k \right\|_{x_k}^2 + (\xi - 1) \|\bar{v}_k\|_{x_k}^2 \\
 &= \left(\left\| \log_{x_k}(x^*) - \bar{v}_k \right\|_{x_k}^2 + (\zeta - 1) \|\bar{v}_k\|_{x_k}^2 \right) + (\xi - \zeta) \|\bar{v}_k\|_{x_k}^2 \\
 &\geq \left(\left\| \log_{y_k}(x^*) - v_k \right\|_{y_k}^2 + (\zeta - 1) \|v_k\|_{y_k}^2 \right) + (\xi - \zeta) \|\bar{v}_k\|_{x_k}^2 \\
 &= \left\| \log_{y_k}(x^*) - v_k \right\|_{y_k}^2 + (\zeta - 1) \|v_k\|_{y_k}^2 + (\xi - \zeta) \frac{1}{\left(1 - \frac{\sqrt{\xi q}}{1 + \sqrt{\xi q}}\right)^2} \|v_k\|_{y_k}^2 \\
 &= \left\| \log_{y_k}(x^*) - v_k \right\|_{y_k}^2 + (\xi - 1) \|v_k\|_{y_k}^2 + (\xi - \zeta) \left(\frac{1}{\left(1 - \frac{\sqrt{\xi q}}{1 + \sqrt{\xi q}}\right)^2} - 1 \right) \|v_k\|_{y_k}^2
 \end{aligned}$$

Combining these inequalities with the result in Step 1 gives

$$\begin{aligned}
 0 &\geq \left(f(x_{k+1}) - f(x^*) + \frac{\mu}{2} \|\bar{v}_{k+1} - \log_{y_k}(x^*)\|_{y_k}^2 + (\xi - 1) \frac{\mu}{2} \|\bar{v}_{k+1}\|_{y_k}^2 \right) \\
 &\quad - \left(1 - \sqrt{\frac{q}{\xi}} \right) \left(f(x_k) - f(x^*) + \frac{\mu}{2} \|v_k - \log_{y_k}(x^*)\|_{y_k}^2 + (\xi - 1) \frac{\mu}{2} \|v_k\|_{y_k}^2 \right) \\
 &\quad + \frac{\mu}{2} \sqrt{\xi q} \left(1 - \sqrt{\frac{q}{\xi}} \right) \|v_k\|_{y_k}^2 \\
 &\quad + \frac{\mu}{2} \left[\left\| \log_{x_{k+1}}(x^*) - \bar{v}_{k+1} \right\|_{x_{k+1}}^2 + (\xi - 1) \|\bar{v}_{k+1}\|_{x_{k+1}}^2 \right. \\
 &\quad \left. - \left\| \log_{y_k}(x^*) - \bar{v}_{k+1} \right\|_{y_k}^2 - (\xi - 1) \|\bar{v}_{k+1}\|_{y_k}^2 - \frac{\xi - \delta}{2} \left(\frac{1}{1 - \sqrt{\xi q}} - 1 \right) \left\| \left(1 - \sqrt{\frac{q}{\xi}} \right) v_k \right\|_{y_k}^2 \right] \\
 &\quad + \frac{\mu}{2} \left[\left\| \log_{y_k}(x^*) - v_k \right\|_{y_k}^2 + (\xi - 1) \|v_k\|_{y_k}^2 + (\xi - \zeta) \left(\frac{1}{\left(1 - \frac{\sqrt{\xi q}}{1 + \sqrt{\xi q}}\right)^2} - 1 \right) \|v_k\|_{y_k}^2 \right. \\
 &\quad \left. - \left\| \log_{x_k}(x^*) - \bar{v}_k \right\|_{x_k}^2 - (\xi - 1) \|\bar{v}_k\|_{x_k}^2 \right] \\
 &= \left(1 - \sqrt{\frac{q}{\xi}} \right)^{k+1} (\phi_{k+1} - \phi_k) \\
 &\quad + \frac{\mu}{2} \left(\sqrt{\xi q} \left(1 - \sqrt{\frac{q}{\xi}} \right) - \frac{\xi - \delta}{2} \left(\frac{1}{1 - \sqrt{\xi q}} - 1 \right) \left(1 - \sqrt{\frac{q}{\xi}} \right)^2 + (\xi - \zeta) \left(\frac{1}{\left(1 - \frac{\sqrt{\xi q}}{1 + \sqrt{\xi q}}\right)^2} - 1 \right) \right) \|v_k\|^2 \\
 &\geq \left(1 - \sqrt{\frac{q}{\xi}} \right)^{k+1} (\phi_{k+1} - \phi_k).
 \end{aligned}$$

□

Corollary F.1. *Let f be a geodesically μ -strongly convex and geodesically L -smooth function. Then, RNAG-SC with parameter $\xi = \zeta + 3(\zeta - \delta)$ and step size $s = \frac{1}{9\xi L}$ finds an ϵ -approximate solution in $O\left(\xi \sqrt{\frac{L}{\mu}} \log\left(\frac{L}{\epsilon}\right)\right)$ iterations.*

Proof. (Step 1: Checking the condition for Theorem 5.6). It is straightforward to check that

$$\frac{\xi - \delta}{2} \left(\frac{1}{1 - t} - 1 \right) \leq (\xi - \zeta) \left(\frac{1}{1 - \frac{t}{1+t}} - 1 \right)$$

for all $t \in (0, 1/3]$. Because $\sqrt{\xi q} = \sqrt{\xi \mu \frac{1}{9\xi^2 L}} = \frac{1}{3} \sqrt{\mu/L} \in (0, 1/3]$, we have

$$\begin{aligned} (\xi - \zeta) \left(\frac{1}{\left(1 - \frac{\sqrt{\xi q}}{1 + \sqrt{\xi q}}\right)^2} - 1 \right) &\geq (\xi - \zeta) \left(\frac{1}{\left(1 - \frac{\sqrt{\xi q}}{1 + \sqrt{\xi q}}\right)} - 1 \right) \\ &\geq \frac{\xi - \delta}{2} \left(\frac{1}{1 - \sqrt{\xi q}} - 1 \right). \end{aligned}$$

Because $\sqrt{\frac{q}{\xi}} \in (0, 1)$, we have

$$\begin{aligned} \frac{\xi - \delta}{2} \left(\frac{1}{1 - \sqrt{\xi q}} - 1 \right) \left(1 - \sqrt{\frac{q}{\xi}}\right)^2 - \sqrt{\xi q} \left(1 - \sqrt{\frac{q}{\xi}}\right) &\leq \frac{\xi - \delta}{2} \left(\frac{1}{1 - \sqrt{\xi q}} - 1 \right) \left(1 - \sqrt{\frac{q}{\xi}}\right)^2 \\ &\leq \frac{\xi - \delta}{2} \left(\frac{1}{1 - \sqrt{\xi q}} - 1 \right). \end{aligned}$$

Combining these inequalities gives the desired condition.

(Step 2: Computing iteration complexity). It follows from Theorem 5.4 that

$$f(x_k) - f(x^*) \leq \left(1 - \sqrt{\frac{q}{\xi}}\right)^k \phi_k \leq \left(1 - \sqrt{\frac{q}{\xi}}\right)^k \phi_0 = \left(1 - \sqrt{\frac{q}{\xi}}\right)^k \left(f(x_0) - f(x^*) + \frac{\mu}{2} \|\log_{x_0}(x^*)\|^2\right).$$

By the geodesic L -smoothness of f , we have

$$\begin{aligned} f(x_k) - f(x^*) &\leq \left(1 - \sqrt{\frac{q}{\xi}}\right)^k \left(\frac{L}{2} \|\log_{x_0}(x^*)\|^2 + \frac{\mu}{2} \|\log_{x_0}(x^*)\|^2\right) \\ &\leq \left(1 - \sqrt{\frac{q}{\xi}}\right)^k L \|\log_{x_0}(x^*)\|^2 \\ &= \left(1 - \sqrt{\frac{\mu}{9\xi^2 L}}\right)^k L \|\log_{x_0}(x^*)\|^2 \\ &\leq e^{-\sqrt{\frac{\mu}{9\xi^2 L}} k} L \|\log_{x_0}(x^*)\|^2 \\ &\leq e^{-\sqrt{\frac{\mu}{9\xi^2 L}} k} L \|\log_{x_0}(x^*)\|^2. \end{aligned}$$

Thus, we have $f(x_k) - f(x^*) \leq \epsilon$ whenever

$$k \geq \sqrt{\frac{9\xi^2 L}{\mu}} \log \left(\frac{L}{\epsilon} \|\log_{x_0}(x^*)\|^2 \right),$$

which implies the $O\left(\xi \sqrt{\frac{L}{\mu}} \log \frac{L}{\epsilon}\right)$ iteration complexity of RNAG-SC. □

G. Continuous-Time Interpretation

G.1. The g-convex case

Because we approximate the curve $y(t)$ by the iterates y_k , we first rewrite RNAG-C in the form using only the iterates y_k as follows:

$$\begin{aligned}
 y_{k+1} - y_k &= x_{k+1} - y_k + \frac{\xi}{\lambda_{k+1} + \xi - 1} \bar{v}_{k+1} \\
 &= -s \operatorname{grad} f(y_k) + \frac{\xi}{\lambda_{k+1} + \xi - 1} (\bar{v}_{k+1} + s \operatorname{grad} f(y_k)) \\
 &= -s \operatorname{grad} f(y_k) + \frac{\xi}{\lambda_{k+1} + \xi - 1} \left(v_k - \frac{s\lambda_k}{\xi} \operatorname{grad} f(y_k) + s \operatorname{grad} f(y_k) \right) \\
 &= \left(-1 + \frac{-\lambda_k + \xi}{\lambda_{k+1} + \xi - 1} \right) s \operatorname{grad} f(y_k) + \frac{\xi}{\lambda_{k+1} + (\xi - 1)} \frac{\lambda_k - 1}{\xi} (y_k - x_k) \\
 &= \frac{1 - \lambda_k - \lambda_{k+1}}{\lambda_{k+1} + (\xi - 1)} s \operatorname{grad} f(y_k) + \frac{\lambda_k - 1}{\lambda_{k+1} + (\xi - 1)} (y_k - x_k) \\
 &= \frac{1 - \lambda_k - \lambda_{k+1}}{\lambda_{k+1} + (\xi - 1)} s \operatorname{grad} f(y_k) + \frac{\lambda_k - 1}{\lambda_{k+1} + (\xi - 1)} (y_k - y_{k-1} + s \operatorname{grad} f(y_{k-1})) \\
 &= \frac{\lambda_k - 1}{\lambda_{k+1} + (\xi - 1)} (y_k - y_{k-1}) - \frac{\lambda_{k+1}}{\lambda_{k+1} + (\xi - 1)} s \operatorname{grad} f(y_k) \\
 &\quad + \frac{\lambda_k - 1}{\lambda_{k+1} + (\xi - 1)} s (\operatorname{grad} f(y_{k-1}) - \operatorname{grad} f(y_k))
 \end{aligned}$$

We introduce a smooth curve $y(t)$ as mentioned in Section 6. Now, dividing both sides of the above equality by \sqrt{s} and substituting

$$\begin{aligned}
 \frac{y_{k+1} - y_k}{\sqrt{s}} &= \dot{y} + \frac{\sqrt{s}}{2} \ddot{y} + o(\sqrt{s}) \\
 \frac{y_k - y_{k-1}}{\sqrt{s}} &= \dot{y} - \frac{\sqrt{s}}{2} \ddot{y} + o(\sqrt{s}) \\
 \sqrt{s} \operatorname{grad} f(y_{k-1}) &= \sqrt{s} \operatorname{grad} f(y_k) + o(\sqrt{s}),
 \end{aligned}$$

we obtain

$$\dot{y} + \frac{\sqrt{s}}{2} \ddot{y} + o(\sqrt{s}) = \frac{\lambda_k - 1}{\lambda_{k+1} + (\xi - 1)} \left(\dot{y} - \frac{\sqrt{s}}{2} \ddot{y} + o(\sqrt{s}) \right) - \frac{\lambda_{k+1}}{\lambda_{k+1} + (\xi - 1)} \sqrt{s} \operatorname{grad} f(y).$$

Dividing both sides by \sqrt{s} and rearranging terms, we have

$$\frac{1}{2} \left(1 + \frac{\lambda_k - 1}{\lambda_{k+1} + (\xi - 1)} \right) \ddot{y} + \frac{1}{\sqrt{s}} \left(1 - \frac{\lambda_k - 1}{\lambda_{k+1} + (\xi - 1)} \right) \dot{y} + \frac{\lambda_{k+1}}{\lambda_{k+1} + (\xi - 1)} \operatorname{grad} f(y) + \frac{o(\sqrt{s})}{\sqrt{s}} = 0.$$

Substituting $k = \frac{t}{\sqrt{s}}$, we can check that $\frac{\lambda_k - 1}{\lambda_{k+1} + (\xi - 1)} \rightarrow 1$, $\frac{\lambda_{k+1}}{\lambda_{k+1} + (\xi - 1)} \rightarrow 1$, and $\frac{1}{\sqrt{s}} \left(1 - \frac{\lambda_k - 1}{\lambda_{k+1} + (\xi - 1)} \right) = \frac{1}{\sqrt{s}} \frac{\lambda_{k+1} - \lambda_k + \xi}{\lambda_{k+1} + (\xi - 1)} = \frac{1}{\sqrt{s}} \frac{1 + 2\xi}{k + T + 4\xi - 2} = \frac{1 + 2\xi}{t + (T + 4\xi - 2)\sqrt{s}} \rightarrow \frac{1 + 2\xi}{t}$ as $s \rightarrow 0$. Therefore, we obtain

$$\ddot{y} + \frac{1 + 2\xi}{t} \dot{y} + \operatorname{grad} f(y) = 0.$$

G.2. The g -strongly convex case

As we approximate the curve $y(t)$ by the iterates y_k , we first rewrite RNAG-C in the form using only the iterates y_k as follows:

$$\begin{aligned}
 y_{k+1} - y_k &= x_{k+1} - y_k + \frac{\sqrt{\xi\mu s}}{1 + \sqrt{\xi\mu s}} \bar{v}_{k+1} \\
 &= -s \operatorname{grad} f(y_k) + \frac{\sqrt{\xi\mu s}}{1 + \sqrt{\xi\mu s}} (\bar{v}_{k+1} + s \operatorname{grad} f(y_k)) \\
 &= -\frac{s}{1 + \sqrt{\xi\mu s}} \operatorname{grad} f(y_k) + \frac{\sqrt{\xi\mu s}}{1 + \sqrt{\xi\mu s}} \left(\left(1 - \sqrt{\frac{\mu s}{\xi}}\right) v_k + \sqrt{\frac{\mu s}{\xi}} \left(-\frac{\operatorname{grad} f(y_k)}{\mu}\right) \right) \\
 &= -\frac{2s}{1 + \sqrt{\xi\mu s}} \operatorname{grad} f(y_k) + \frac{\sqrt{\xi\mu s}}{1 + \sqrt{\xi\mu s}} \left(1 - \sqrt{\frac{\mu s}{\xi}}\right) \frac{1}{\sqrt{\xi\mu s}} (y_k - x_k) \\
 &= -\frac{2s}{1 + \sqrt{\xi\mu s}} \operatorname{grad} f(y_k) + \frac{1 - \sqrt{\mu s/\xi}}{1 + \sqrt{\xi\mu s}} (y_k - y_{k-1} + s \operatorname{grad} f(y_{k-1})) \\
 &= \frac{1 - \sqrt{\mu s/\xi}}{1 + \sqrt{\xi\mu s}} (y_k - y_{k-1}) - \frac{1 + \sqrt{\mu s/\xi}}{1 + \sqrt{\xi\mu s}} s \operatorname{grad} f(y_k) + \frac{1 - \sqrt{\mu s/\xi}}{1 + \sqrt{\xi\mu s}} s (\operatorname{grad} f(y_{k-1}) - \operatorname{grad} f(y_k))
 \end{aligned}$$

Dividing both sides by \sqrt{s} and substituting

$$\begin{aligned}
 \frac{y_{k+1} - y_k}{\sqrt{s}} &= \dot{y} + \frac{\sqrt{s}}{2} \ddot{y} + o(\sqrt{s}) \\
 \frac{y_k - y_{k-1}}{\sqrt{s}} &= \dot{y} - \frac{\sqrt{s}}{2} \ddot{y} + o(\sqrt{s}) \\
 \sqrt{s} \operatorname{grad} f(y_{k-1}) &= \sqrt{s} \operatorname{grad} f(y_k) + o(\sqrt{s})
 \end{aligned}$$

yield

$$\dot{y} + \frac{\sqrt{s}}{2} \ddot{y} + o(\sqrt{s}) = \frac{1 - \sqrt{\mu s/\xi}}{1 + \sqrt{\xi\mu s}} \left(\dot{y} - \frac{\sqrt{s}}{2} \ddot{y} + o(\sqrt{s}) \right) - \frac{1 + \sqrt{\mu s/\xi}}{1 + \sqrt{\xi\mu s}} \sqrt{s} \operatorname{grad} f(y_k).$$

Dividing both sides by \sqrt{s} and rearranging terms, we obtain

$$\frac{1}{2} \left(1 + \frac{1 - \sqrt{\mu s/\xi}}{1 + \sqrt{\xi\mu s}} \right) \ddot{y} + \frac{\left(\sqrt{1/\xi} + \sqrt{\xi} \right) \sqrt{\mu}}{1 + \sqrt{\xi\mu s}} \dot{y} + \frac{1 + \sqrt{\mu s/\xi}}{1 + \sqrt{\xi\mu s}} \operatorname{grad} f(y_k) + \frac{o(\sqrt{s})}{\sqrt{s}} = 0.$$

Taking the limit $s \rightarrow 0$ gives

$$\ddot{y} + \left(\frac{1}{\sqrt{\xi}} + \sqrt{\xi} \right) \sqrt{\mu} \dot{y} + \operatorname{grad} f(y) = 0$$

as desired.

G.3. Experiments

In this section, we empirically show that the iterates of our methods converge to the solution of the corresponding ODEs, as taking the limit $s \rightarrow 0$. We use the Rayleigh quotient maximization problem in Section 7 with $d = 10$ and $\xi = 2$. For RNAG-SC, we set $\mu = 0.1$ (note that the limiting argument above does not use geodesic μ -strong convexity of f). To compute the solution of ODEs (7) and (8), we implement SIRNAG (Option I) (Alimisis et al., 2020) with very small integration step size. The results are shown in Figure 4 and Figure 5.

H. Proofs for Section 7

Proposition H.1. *The function f is geodesically $(\lambda_{\max} - \lambda_{\min})$ -smooth, where λ_{\max} and λ_{\min} are the largest and smallest eigenvalues of A , respectively.*

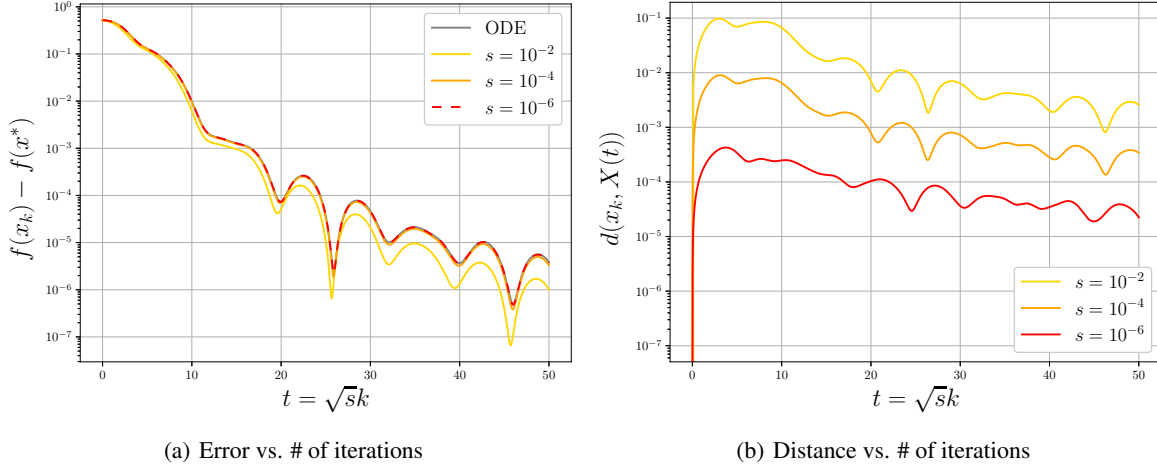


Figure 4. Convergence of RNAG-C to the solution of ODE (7).

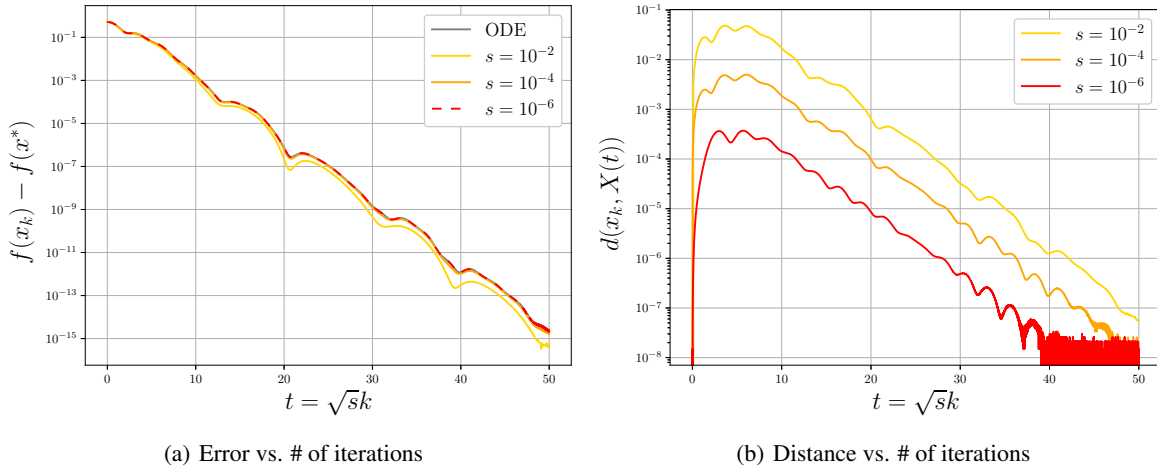


Figure 5. Convergence of RNAG-SC to the solution of ODE (8).

Proof. For $x \in \mathbb{S}^{d-1} \subseteq \mathbb{R}^d$ and a unit tangent vector $v \in T_x M$, we have

$$\exp_x(tv) = \frac{x + \tan(t)v}{\|x + \tan(t)v\|} = \frac{x + \tan(t)v}{\sec(t)}$$

for $t \in I$, where I is a small interval containing 0. We consider the function $h : I \rightarrow M$ defined as

$$\begin{aligned} h(t) &= f(\exp_x(tv)) \\ &= -\frac{1}{2} \cos^2(t) (x + \tan(t)v)^\top A (x + \tan(t)v) \\ &= -\frac{1}{2} h_1(t) h_2(t), \end{aligned}$$

where $h_1(t) = \cos^2(t)$ and $h_2(t) = (x + \tan(t)v)^\top A (x + \tan(t)v)$. Note that $h_1(0) = 1$, $h_1'(0) = 0$, $h_1''(0) = -2$, $h_2(0) = x^\top A x$, $h_2'(0) = 2v^\top A x$, and $h_2''(0) = 2v^\top A v$. Now, by the product rule, we have

$$h''(0) = -\frac{1}{2} h_1''(0) h_2(0) - h_1'(0) h_2'(0) - \frac{1}{2} h_1(0) h_2''(0) = x^\top A x - v^\top A v.$$

Because Rayleigh quotient is always in $[\lambda_{\min}, \lambda_{\max}]$, we have $|h''(0)| \leq (\lambda_{\max} - \lambda_{\min})$. This shows that f is geodesically $(\lambda_{\max} - \lambda_{\min})$ -smooth. \square

Proposition H.2. *The function f is geodesically 1-strongly convex.*

Proof. It is enough to show that the function $x \mapsto d(x, p_i)^2$ is convex. When $K_{\max} \leq 0$, we have $\delta = 1$. Let $\gamma : I \rightarrow M$ be a geodesic whose image is in N . It follows from Proposition C.1 that

$$\begin{aligned} \frac{d^2}{dt^2} \frac{1}{2} d(\gamma(t), p_i)^2 &= \frac{d}{dt} \left\langle \log_{\gamma(t)}(p_i), -\gamma'(t) \right\rangle \\ &= \left\langle D_t \log_{\gamma(t)}(p_i), -\gamma'(t) \right\rangle + \left\langle \log_{\gamma(t)}(p_i), -\gamma''(t) \right\rangle. \end{aligned}$$

Note that $\gamma''(t) = 0$ because γ is a geodesic. Now, Proposition 5.1 gives $\frac{d^2}{dt^2} \frac{1}{2} d(\gamma(t), p_i)^2 \geq 1$. \square