

---

# Acceleration via silver step-size on Riemannian manifolds with applications to Wasserstein space

---

Anonymous Author(s)

Affiliation

Address

email

## Abstract

1        There is extensive literature on accelerating first-order optimization methods in  
2        a Euclidean setting. Under which conditions such acceleration is feasible in  
3        Riemannian optimization problems is an active area of research. Motivated by the  
4        recent success of varying step-size methods in the Euclidean setting, we undertake  
5        a study of such algorithms in the Riemannian setting. We show that varying step-  
6        size acceleration can be achieved in non-negatively curved Riemannian manifolds  
7        under geodesic smoothness and generalized geodesic convexity, a new notion of  
8        convexity that we introduce to aid our analysis. As a core application, we show  
9        that our method provides the first theoretically guaranteed accelerated optimization  
10       method in Wasserstein spaces. In addition, we numerically validate our method's  
11       applicability to other problems, such as optimization problems on the sphere.

## 12    1 Introduction

13    Consider the Riemannian optimization problem

$$\min_{x \in N} f(x), \tag{1.1}$$

14    where  $N \subseteq M$  is a geodesically convex subset of a Riemannian manifold  $M$ , and  $f : N \rightarrow \mathbb{R}$  is a  
15    continuously differentiable geodesically convex functional. A popular approach to solve (1.1) is via  
16    Riemannian gradient descent (RGD) [ZS16] given by,

$$x_{n+1} = \exp_{x_n}(-\eta_n \text{Grad } f(x_n)), \tag{1.2}$$

17    where  $\exp_x(\cdot)$  is the exponential map at  $x$ ,  $\eta_n$  is the step-size at iteration  $n$ , and  $\text{Grad}$  denotes the  
18    Riemannian gradient. It is known that for geodesically convex and smooth functionals  $f$ , constant  
19    step-size RGD has an  $O(1/n)$  convergence rate as in Euclidean spaces [KY22][Theorem D.2].

20    A natural follow-up question is whether one can find first-order algorithms that achieve an *accelerated*  
21    convergence rate. This is motivated by the success of accelerated first-order methods in Euclidean  
22    settings, most notably Nesterov's method [Nes83], which uses momentum to achieve an  $O(1/n^2)$   
23    rate for convex and smooth objectives. Extensive efforts have been made to achieve the same  
24    accelerated rate using similar acceleration in Riemannian optimization problems under various  
25    settings [LSC<sup>+</sup>17, ZS18, AS20, Sie21, AOBL21, CB22, MR22, KY22, HMJG23]). However, these  
26    works typically rely on additional constraints, stronger assumptions, or modifications to the basic  
27    gradient descent update (1.2). For example, [LSC<sup>+</sup>17] involves an intractable nonlinear operator.  
28    The analysis in [HMJG23] relies on a submanifold structure and establishes acceleration only in  
29    the asymptotic regime. All the other algorithms require both upper and lower sectional curvature

30 bounds. We refer to [dST21] for a general survey of momentum-based acceleration methods, and to  
 31 [KY22][Sections 1, 2] for Riemannian variants.

32 On the other hand, there is a line of work showing that, in the Euclidean case, an accelerated  
 33 convergence rate is possible by using a carefully designed *varying step-size schedule* without any  
 34 modification to vanilla gradient descent. This idea goes back to [You53]; for quadratic functions,  
 35 choosing  $\eta_n$  to be Chebyshev step-sizes in gradient descent achieves the  $O(1/n^2)$  rate. Generalizing  
 36 this idea to general convex and smooth functions, [Alt18, AP24b, AP24c, BA24] introduced the  
 37 *silver step-size* schedule—a carefully designed step-size sequence that guarantees an improved  
 38 convergence rate of  $O(1/n^{\log_2 \rho})$ , where  $\rho = 1 + \sqrt{2}$ . While slower than the  $O(1/n^2)$  rate of  
 39 Nesterov’s acceleration, this method significantly outperforms constant step-size gradient descent  
 40 and shows that standard gradient descent, with a carefully designed step-size schedule, can achieve  
 41 meaningful acceleration. Whether full Nesterov-style acceleration can be attained purely through  
 42 step-size adaptation remains an open question [AP24a]. Motivated by the success of the silver  
 43 step-size schedule in the Euclidean case, in this work, we ask the following question,

44 Is it possible to accelerate Riemannian gradient descent by only using a varying  
 45 step-size schedule without any other modification?

46 **Main contribution** Towards addressing the above question, we make the following contributions.

- 47 1. We introduce a new notion of convexity, which we call *generalized geodesic convexity*.  
 48 Intuitively, for any three points  $x, y, z \in N$ , generalized geodesic convexity requires  $f$  to be  
 49 convex along some curve from  $x$  to  $y$  where the initial velocity is taken in the direction from  $x$   
 50 to  $y$ , but measured in the tangent space at  $z$  instead of  $x$  (see Definition 3.4, Lemma D.6, and  
 51 Figure 1 for details). While well-studied in the optimal transport literature [AGS08, SKL20],  
 52 this form of convexity has not been explored in the context of Riemannian optimization.
- 53 2. For non-negatively curved manifold  $M$  and geodesically  $L$ -smooth, generalized geodesically  
 54 convex function  $f$ , under some technical assumptions we show that RGD with the silver  
 55 step-size schedule achieves the accelerated convergence rate of  $O(1/n^{\log_2 \rho})$ , and the rate  
 56 of  $\exp(-O(n/\kappa^{\log_2 \rho}))$  when  $f$  is in addition geodesically strongly convex with condition  
 57 number  $\kappa$ . These rates match the corresponding rates in the Euclidean case.
- 58 3. One of our main technical contributions is to avoid relying on an equality that is essential  
 59 to the analysis in the Euclidean setting ([AP24c][Equation (8)]), but fails to hold on a  
 60 Riemannian manifold due to metric distortion. Instead, our proof is based on an inequality  
 61 that discards terms affected by uncontrolled metric distortion while preserving the curvature-  
 62 controlled terms (Lemma 5.1, 5.2). Furthermore, compared to Euclidean space, we need to  
 63 handle the intrinsic challenges of Riemannian optimization stemming from metric distortion  
 64 and curvature of the space.
- 65 4. By assuming non-negative curvature and generalized geodesic convexity, our analysis  
 66 achieves acceleration without requiring the curvature upper bound or diameter bound on  $N$ ,  
 67 typically imposed in existing analyses of momentum-based methods.
- 68 5. We show the applicability of our method to Wasserstein space, which has a Riemannian  
 69 structure but lacks a curvature upper bound and diameter bound, and therefore existing  
 70 Riemannian accelerated methods do not apply. In addition, we numerically demonstrate  
 71 the algorithm’s performance on a particular optimization problem defined on the sphere, a  
 72 well-studied positively curved Riemannian manifold.

## 73 2 Background

74 **Riemannian manifolds** In this section, we review the basic concepts of Riemannian manifold while  
 75 deferring the rigorous description to Appendix A.1. At a point  $x$  on a manifold  $M$ , tangent vectors are  
 76 the velocity vectors of smooth curves on  $M$  that pass through  $x$ . The tangent space  $T_x M$  is the vector

space consisting of all such tangent vectors at  $x$ . A Riemannian manifold is a manifold equipped with an inner product  $\langle \cdot, \cdot \rangle_x$  for each tangent space  $T_x M$ , called a Riemannian metric. For  $x, y \in M$ , the distance  $d(x, y)$  is the infimum of the length of all piecewise continuously differentiable curves from  $x$  to  $y$ . A Riemannian gradient of the differentiable function  $f : M \rightarrow \mathbb{R}$  at  $x$  is a tangent vector  $\text{Grad } f(x) \in T_x M$  satisfying  $d_v f(x) = \langle \text{Grad } f(x), v \rangle_x$  for all  $v \in T_x M$ . Here,  $d_v f(x)$  is a directional derivative of  $f$  at  $x$  along the direction  $v$ . For  $(x, v) \in TM$ , where  $TM := \coprod_{x \in M} T_x M$  denotes the tangent bundle, a smooth curve  $\gamma_v : [0, 1] \rightarrow M$  with  $\gamma_v(0) = x$  and  $\gamma'_v(0) = v$  is called a (constant speed) geodesic if it has the locally minimum length with zero acceleration. The exponential map  $\exp_x : T_x M \rightarrow M$  is a map defined by  $\exp_x(v) = \gamma_v(1)$ .  $\exp_x(v)$  transports the point  $x$  in the direction of the tangent vector  $v$ , following the geodesic  $\gamma_v$ . It is known that  $\exp_x$  is a local diffeomorphism in some neighborhood  $U$  of  $0 \in T_x M$ . Hence,  $\exp_x$  allows the inverse on  $U$ , which is called the logarithmic map  $\log_x : \exp_x(U) \rightarrow T_x M$ . While the exponential and logarithmic maps are always locally well-defined, they may not be globally well-defined. A parallel transport  $\Gamma(\gamma)_{t_0}^{t_1} : T_{\gamma(t_0)} M \rightarrow T_{\gamma(t_1)} M$  is a way to transport a tangent vector along the curve  $\gamma$  parallelly. If  $\gamma$  is a geodesic curve such that  $\gamma(0) = x, \gamma(1) = y$ , then we simply denote  $\Gamma(\gamma)_0^1$  as  $\Gamma_x^y$ , a (geodesic) parallel transport from  $T_x M$  to  $T_y M$ .

**Definition 2.1** (Geodesic convexity). *We say  $N \subseteq M$  is a geodesically convex subset of  $M$  if for all  $x, y \in N$  there exists a geodesic  $\gamma$  such that  $\gamma(0) = x, \gamma(1) = y$ , and  $\gamma(t) \in N$  for all  $t \in [0, 1]$ . We say a differentiable function  $f : N \rightarrow \mathbb{R}$  is geodesically  $\alpha$ -strongly convex if for all  $x, y \in N$*

$$f(y) \geq f(x) + \langle \text{Grad } f(x), \log_x y \rangle_x + \frac{\alpha}{2} d^2(x, y).$$

*If the above inequality holds with  $\alpha = 0$ , then  $f$  is said to be geodesically convex.*

**Definition 2.2** (Geodesic smoothness). *We say  $f$  is geodesically  $L$ -smooth if for all  $x, y \in N$*

$$\|\Gamma_y^x \text{Grad } f(y) - \text{Grad } f(x)\|_x \leq Ld(x, y).$$

**Silver Step-size in Euclidean space** In this section, we present the silver step-size schedule [AP24c] for Euclidean optimization problem. Consider the problem (1.1) where  $N \equiv \mathbb{R}^d$ , and  $f$  is convex and  $L$ -smooth. A standard approach is gradient descent, which updates via  $x_{n+1} = x_n - \eta \nabla f(x_n)$  for a fixed step-size  $\eta$ . In contrast, the silver step-size schedule is a sequence of varying step-sizes  $\{\eta_n\}_{n \in \mathbb{N}}$ . For  $n = 2^k - 1$  where  $k \in \mathbb{N}$ ,  $\{\eta_n\}_{n \in \mathbb{N}}$  is given by the following inductively constructed sequence:

$$\eta^{(k+1)} = [\eta^{(k)}, 1 + \rho^{k-1}, \eta^{(k)}], \quad (2.1)$$

where  $\rho = 1 + \sqrt{2}$ . We set  $\eta_0 = \rho - 1$ . For example, for  $k = 1, 2, 3$ ,  $\eta^{(k)}$  has the following form:

$$\eta^{(1)} = [\sqrt{2}], \quad \eta^{(2)} = [\sqrt{2}, 2, \sqrt{2}], \quad \eta^{(3)} = [\sqrt{2}, 2, \sqrt{2}, 2 + \sqrt{2}, \sqrt{2}, 2, \sqrt{2}]$$

In Euclidean optimization, the silver step-size was recently shown to improve the convergence rate of the gradient descent from  $O(1/n)$  to  $O(1/n^{\log_2 \rho})$  [AP24c].

### 3 Silver Step-size RGD: Assumptions and Preliminaries

In this section, we state the assumptions on the manifold and objective function required to solve problem (1.1) using silver step-size RGD (1.2).

**Assumption 3.1** (Assumptions for Riemannian manifold).

1.  $M$  is a complete Riemannian manifold, i.e., any two points are connected by some geodesic.
2.  $N \subseteq M$  is open, geodesically convex subset with non-negative sectional curvature.
3. Exponential maps, logarithmic maps, and parallel transports are all well-defined and computationally tractable on  $N$ .

**Assumption 3.2** (Assumptions on the objective). *We make the following assumptions on  $f : N \rightarrow \mathbb{R}$ .*

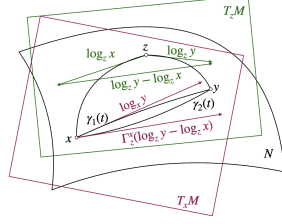


Figure 1: Geometric illustration of generalized geodesic convexity. Usual geodesic convexity means for any  $x, y \in N$ , the function is convex along the geodesic curve  $\gamma_1(t)$ . On the other hand, generalized geodesic convexity with base  $z$  implies the function is convex along a curve  $\gamma_2(t)$ .

1.  $f$  is geodesically convex and has a global minimizer  $x_* \in N$ .
  2. All the iterates of our algorithms are well defined and remain inside  $N$ .
  3. There exists a constant  $L > 0$  such that for all  $x_i, x_j$  in the RGD trajectory,  $i, j = 0, 1, 2, \dots, *$ ,
- $$Q_{ij} := 2L(f(x_i) - f(x_j)) - 2L \left\langle \text{Grad } f(x_j), \log_{x_j} x_i \right\rangle_{x_j} - \left\| \Gamma_{x_i}^{x_j} \text{Grad } f(x_i) - \text{Grad } f(x_j) \right\|_{x_j}^2 \geq 0. \quad (3.1)$$

**Remark 3.3.** Assumptions 3.1 and 3.2, excluding the non-negative curvature and (3.1), are standard in Riemannian optimization literature [AOBL21, KY22, HMJG23] and ensure well-behaved RGD iterates. Whereas we additionally assume non-negative curvature and (3.1), we do not require the curvature upper bound or diameter bound on  $N$  typically assumed in momentum-based algorithms.

Some comments on (3.1) are in order. It is well known that in Euclidean space, (3.1) holds for any  $x_i, x_j \in \mathbb{R}^d$  when  $f$  is convex and  $L$ -smooth [Nes14][Theorem 2.1.5]. However, on Riemannian manifolds, (3.1) can be established under geodesic  $L$ -smoothness together with a stronger form of convexity, which we dub *generalized geodesic convexity*.

**Definition 3.4** (Generalized geodesic convexity). A functional  $f : N \rightarrow \mathbb{R}$  is called *generalized geodesically convex with base  $z \in N$*  if for all  $x, y \in N$ , we have,

$$f(y) \geq f(x) - \langle \Gamma_x^z \text{Grad } f(x), \log_z y - \log_z x \rangle_z. \quad (3.2)$$

$f$  is called *generalized geodesically convex* if (3.2) holds for all  $z \in N$ .

**Remark 3.5** (Geometric interpretation of generalized geodesic convexity). Intuitively, for any three points  $x, y, z \in N$ , generalized geodesic convexity requires  $f$  to be convex along a curve from  $x$  to  $y$ , where the initial velocity is measured in the tangent space at a third point  $z \in N$  (see Definition 3.4, Lemma D.6, and Figure 1 for details). This generalizes standard geodesic convexity, which corresponds to the special case  $z = x$ .

**Remark 3.6.** Albeit new to Riemannian optimization literature, the notion of generalized geodesic convexity is well-established in optimal transport and has found numerous applications in Wasserstein geometry, for example, in the theoretical analysis of the proximal operator in the Wasserstein space [AGS08, SKL20, DBCS23], as well as in the context of  $\Gamma$ -convergence [AGS08][Lemma 9.2.9]. Definition 3.4 provides a Riemannian analogue of this concept.

To give readers more concrete idea, we show one example of generalized geodesically convex functional. We provide the proof and more examples of such functionals in Appendix D.1.1.

**Example 3.7** (Entropy of Gaussian). The Bures-Wasserstein space  $BW(\mathbb{R}^d)$  is the space of Gaussian distributions on  $\mathbb{R}^d$  equipped with Wasserstein geometry. Restricting to zero-mean Gaussians, it becomes a non-negatively curved Riemannian manifold identified with  $SPD(d)$ , the space of symmetric positive definite matrices. The Riemannian metric is defined by  $\langle S, R \rangle_A = \text{tr}(SAR)$  for  $S, R \in \text{Sym}(d)$ . The functional  $\mathcal{H} : SPD(d) \rightarrow \mathbb{R}$  defined by  $\mathcal{H}(A) = -\frac{1}{2} \log \det A$  is generalized geodesically convex under this geometry.

We now establish the relationship between (3.1) and convexity and smoothness, as in Euclidean space. Proposition 3.8 provides a sufficient condition for (3.1).

**Proposition 3.8.** *Let  $f : N \rightarrow \mathbb{R}$  be a geodesically  $L$ -smooth, and generalized geodesically convex function, and for all  $x, y \in N$ ,  $z := \exp_y(-\frac{1}{L}(\text{Grad } f(y) - \Gamma_x^y \text{Grad } f(x))) \in N$ . Then  $f$  satisfies (3.1) for all  $x_i, x_j \in N$ .*

The condition  $z \in N$  is technical and generally requires case-specific verification. In our key application on Wasserstein space, the condition  $z \in N$  is readily satisfied (see Corollary 6.1).

## 4 Main Results

In this section, we present our main convergence results for silver step-size RGD.

**Theorem 4.1.** *Let Assumption 3.1. 3.2 be true and  $n = 2^k - 1$ . Then, for RGD (1.2) with silver step-sizes  $\eta_n/L$  (2.1), we have,*

$$f(x_n) - f(x_*) \leq r_k L d^2(x_0, x_*), \quad r_k = \left(1 + \sqrt{4\rho^{2^k} - 3}\right)^{-1}.$$

Since  $r_k \asymp n^{-\log_2 \rho} \approx n^{-1.2716}$ , Theorem 4.1 shows a better convergence rate than constant step-size RGD on non-negatively curved manifolds, which is  $O(n^{-1})$  [KY22][Appendix D]. Although the theorem is stated for  $n = 2^k - 1$ , our numerical results indicate that the improvement extends to arbitrary  $n \neq 2^k - 1$  as well; see Appendix E.

Our analysis of silver step-size RGD can be extended to geodesically strongly convex functionals. In Euclidean space, a common technique for upgrading convergence guarantees from convex to strongly convex settings is the restarting method [OC15]. The method proceeds as follows:

1. Perform  $m$  steps of gradient descent starting from an initial point  $x_0$  to obtain  $x_m$ .
2. Restart from  $x_m$  with the step-size reset to  $\eta_0$ , and run  $m$  additional steps to obtain  $x_{2m}$ .
3. Repeat this process  $\ell$  times, each time restarting from the most recent iterate with the step-size reset to  $\eta_0$ . After  $\ell$  restarts, the final output is  $x_{\ell m}$ .

Note the total iteration is  $n := \ell m$ . For fixed  $n$ , choosing  $m$  and  $\ell$  appropriately yields the optimal convergence rate for strongly convex objectives. Notably, this approach remains valid in the Riemannian setting with silver step-size RGD.

**Theorem 4.2.** *Consider the same setting of Theorem 4.1. In addition, let  $f$  be geodesically  $\alpha$ -strongly convex with the condition number  $\kappa := L/\alpha$ . Set  $k^* = \lceil \log_\rho \kappa \rceil + 1$ . For any  $\ell \in \mathbb{N}$ , run silver step-size RGD for  $2^{k^*} - 1$  iterations and repeat this process  $\ell$  times, so that the total number of iteration is  $n = \ell(2^{k^*} - 1)$ . Then,*

$$d^2(x_n, x_*) \leq \exp(-\log(\rho/2)n/\kappa^{\log_\rho 2}) d^2(x_0, x_*).$$

*In particular, the algorithm finds an  $\epsilon$ -approximate solution, i.e.,  $d(x_n, x_*)^2 \leq \epsilon$ , in  $O(\kappa^{\log_\rho 2} \log(1/\epsilon))$  number of iterations.*

We provide the proof in Appendix D.2. Since constant step-size RGD finds an  $\epsilon$ -approximate solution for strongly convex objectives in  $O(\kappa \log(1/\epsilon))$  iterations [KY22, Appendix D], our algorithm achieves an improved rate in the strongly convex setting as well. While our theoretical analysis assumes inner iterates of the form  $m = 2^k - 1$ , the algorithm, as previously noted, also performs well numerically for  $m \neq 2^k - 1$ . Thus, in practice, one may use an arbitrary total iteration number  $n$  and select  $m$  and  $\ell$  accordingly to optimize performance.

## 5 Proof Sketch

In this section, we present an outline of the proof of Theorem 4.1 while deferring the main proof to the Appendix B.1. The following two lemmas are the main components of the proof. Without loss of generality, set  $L = 1$ . We also set  $n = 2^k - 1$  for some  $k \in \mathbb{N}$ .

190 **Lemma 5.1.** *Let the conditions of Theorem 4.1 be true. Then,*

$$\begin{aligned} \mathcal{A}_n &:= (4r_k^2)^{-1} \|\text{Grad } f(x_n)\|_{x_n}^2 + r_k^{-1} \langle \text{Grad } f(x_n), \log_{x_n} x_* \rangle_{x_n} + \sum_{i=0}^{n-1} \eta_i^2 \|\text{Grad } f(x_i)\|_{x_i}^2 \\ &\quad + 2 \sum_{i=0}^{n-1} \eta_i \langle \text{Grad } f(x_i), \log_{x_i} x_* \rangle_{x_i} \geq \|\log_{x_n} x_* + (2r_k)^{-1} \text{Grad } f(x_n)\|_{x_n}^2 - d^2(x_0, x_*). \end{aligned}$$

191 While Lemma 5.1 holds with equality in the Euclidean space, metric distortion in the Riemannian  
192 setting prevents this. To address this, we use the non-negative curvature assumption to control the  
193 distortion. The proof is provided in Appendix B.1. Next, we show the following inequality.

194 **Lemma 5.2.** *Let the conditions of Theorem 4.1 be true. Then, for suitably chosen  $\lambda_{ij} \geq 0$ ,*

$$\sum_{i,j=0,\dots,n,*} \lambda_{ij} Q_{ij} \leq r_k^{-1} (f(x_*) - f(x_n)) - \mathcal{A}_n. \quad (5.1)$$

195 Since  $Q_{ij} \geq 0$  all gradient iterates  $x_i, x_j$  by (3.1), Lemma 5.1 and 5.2 together imply

$$f(x_n) - f(x_*) \leq r_k d^2(x_0, x_*).$$

196 We outline the proof of Lemma 5.2, with details deferred to Appendix B.1.

197 *Proof outline for Lemma 5.2.* We begin with the base step of the induction.

198 **Base Step** First, we show the desired inequality (5.1) is valid for  $n = 1$  ( $k = 1$ ).

199 **Lemma 5.3.** *For any arbitrary initialization  $x_0 \in N$ , consider the following RGD update (1.2).*

$$x_1 = \exp_{x_0}(-\eta_0 \text{Grad } f(x_0)),$$

200 where  $\eta_0 = \rho - 1$ . Choose  $\lambda_{ij}$  the same as in [AP24c][Example 2], i.e.,

$$\begin{pmatrix} \lambda_{00} & \lambda_{01} & \lambda_{0*} \\ \lambda_{10} & \lambda_{11} & \lambda_{1*} \\ \lambda_{*0} & \lambda_{*1} & \lambda_{**} \end{pmatrix} = \begin{pmatrix} 0 & \rho & 0 \\ 1 & \frac{\rho}{2r_1} & \rho - 1 \\ \rho - 1 & \frac{1}{2r_1} & 0 \end{pmatrix}. \quad (5.2)$$

201 Then, inequality (5.1) holds.

202 The proof of Lemma 5.3 is deferred to Appendix B.1.

203 **Induction step** Lemma 5.3 validates that the inequality (5.1) holds for the base case  $n = 1$ . In this  
204 section, given we have the inequality (5.1) for  $n = 2^k - 1$  number of iterates, we show by merging  
205 two silver step-sizes, one can get the inequality (5.1) for  $2n + 1 = 2^{k+1} - 1$  number of iterates.

206 **Lemma 5.4.** *Fix  $n = 2^k - 1$ . Take  $\{x_i\}_{i=0,\dots,n} \subset N$  a sequence induced from the RGD with silver  
207 step-size. Suppose there exist  $\lambda_{ij}^{(k)} \geq 0$  such that (5.1) holds. Write*

$$\sigma_{ij} = \lambda_{ij}^{(k)} \mathbb{1}_{\{i,j=0,\dots,n,*\}} + (1 + 2\rho) \lambda_{i-n-1,j-n-1}^{(k)} \mathbb{1}_{\{i,j=n+1,\dots,2n+1,*\}}$$

208 where  $* - n - 1$  is understood to mean  $*$ . Define

$$\begin{aligned} \lambda_{ij}^{(k+1)} &:= \sigma_{ij} - 2\rho \eta_j \mathbb{1}_{\{i=*,j=n+1,\dots,2n\}} + \left(1 + \rho^{k-1} - \frac{1}{2r_k}\right) \mathbb{1}_{\{i=*,j=n\}} \\ &\quad + \left(\frac{1}{2r_{k+1}} - \frac{1+2\rho}{2r_k}\right) \mathbb{1}_{\{i=*,j=2n+1\}}. \end{aligned}$$

209 Then,  $\lambda_{ij}^{(k+1)} \geq 0$ , and satisfies

$$\sum_{i,j=0,\dots,2n+1,*} \lambda_{ij}^{(k+1)} Q_{ij} \leq \frac{f(x_*) - f(x_{2n+1})}{r_{k+1}} - \mathcal{A}_{2n+1}.$$

210 In particular, if  $\lambda_{ij}^{(1)}$  is chosen as in Lemma 5.3, then  $\lambda_{ij}^{(k)} \geq 0$  for all  $k \in \mathbb{N}$  and  $i, j = 0, \dots, 2^k - 1, *$ .



211 The proof is deferred to Appendix B.1. ■

212 **Remark 5.5** (Comparison with the Euclidean case). *In the Euclidean setting, [AP24c] derived*  
 213 *coefficients that satisfy the equality exactly. However, in our case, since we work with an inequality,*  
 214 *it turns out that certain coefficients can be dropped. Specifically, we can discard the coefficients of*  
 215  *$Q_{n,i}$  and  $Q_{2n+1,i}$  for  $i = n, n+1, \dots, 2n+1, *$ . This selective dropout eliminates terms whose*  
 216 *metrics are difficult to control, thereby making the analysis tractable on Riemannian manifolds.*

## 217 6 Applications

218 In this section, we present applications and representative experiments that demonstrate the practicality  
 219 of our algorithm. Implementation detail and additional experiments are provided in Appendix E.

### 220 6.1 Optimization on the 2-Wasserstein Space

221 As noted earlier, the key advantage of our algorithm over existing methods is that its theoretical  
 222 guarantees remain valid even on manifolds without an upper curvature bound. This makes our analysis  
 223 particularly well-suited for the 2-Wasserstein space, which possesses a Riemannian structure but lacks  
 224 a curvature upper bound (see Lemma A.33 and A.43). Furthermore, since our notion of generalized  
 225 geodesic convexity originates from Wasserstein geometry, many functionals defined on it are known  
 226 to be both generalized geodesically convex and geodesically  $L$ -smooth. However, while acceleration  
 227 has been studied in the continuous-time setting [CCT18, WL22], no discrete-time algorithm with  
 228 provable acceleration guarantees was previously available. To the best of our knowledge, our method  
 229 provides the first theoretically guaranteed accelerated algorithm in the 2-Wasserstein space.

230 We briefly introduce the 2-Wasserstein geometry (see Appendix A.2 for details). Let  $\mathcal{P}_{2,ac}(\mathbb{R}^d)$  denote  
 231 the set of probability measures on  $\mathbb{R}^d$  with finite second moments and absolutely continuous with  
 232 respect to the Lebesgue measure,  $\mathcal{L}^2(\mu)$  be the space of square-integrable functions from  $\mathbb{R}^d \rightarrow \mathbb{R}^d$   
 233 under  $\mu \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$ , and  $T_{\#}\mu$  denotes a pushforward of  $\mu$  by  $T$ . For any  $\mu, \nu \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$ , the  
 234 2-Wasserstein metric is defined as:

$$W_2^2(\mu, \nu) := \min_{T \in \mathcal{L}^2(\mu) \text{ s.t. } T_{\#}\mu = \nu} \mathbb{E}_{x \sim \mu} [\|T(x) - x\|^2]. \quad (6.1)$$

235 The metric space  $(\mathcal{P}_{2,ac}(\mathbb{R}^d), W_2)$ , called the 2-Wasserstein space, admits a Riemannian structure  
 236 with tangent space  $T_{\mu}\mathcal{P}_{2,ac}(\mathbb{R}^d) \subset \mathcal{L}^2(\mu)$  and the Riemannian metric given by the  $\mathcal{L}^2(\mu)$  inner  
 237 product. The exponential map is defined by  $\exp_{\mu}(v) = (id + v)_{\#}\mu$ . The map  $T_{\mu,\nu}$  achieving the  
 238 minimum in (6.1) is called an *optimal transport map* from  $\mu$  to  $\nu$ . Then, for a given functional  
 239  $\mathcal{F} : \mathcal{P}_{2,ac}(\mathbb{R}^d) \rightarrow \mathbb{R}$ , denoting Wasserstein gradient by  $\text{Grad}_{W_2} \mathcal{F}(\mu_n)$  (see Definition A.34), silver  
 240 step-size RGD is given by:

$$\mu_{n+1} = \exp_{\mu_n}(-\eta_n \text{Grad}_{W_2} \mathcal{F}(\mu_n)) = (id - \eta_n \text{Grad}_{W_2} \mathcal{F}(\mu_n))_{\#}\mu_n. \quad (6.2)$$

241 Then, we have the following result analogous to Theorem 4.1, and 4.2.

242 **Corollary 6.1** (Accelerated Wasserstein gradient descent by silver step-size). *Suppose a functional*  
 243  *$\mathcal{F} : \mathcal{P}_{2,ac}(\mathbb{R}^d) \rightarrow \mathbb{R}$  is generalized geodesically convex and geodesically  $L$ -smooth with respect to*  
 244 *Wasserstein geometry<sup>1</sup>. Let  $\mu_n$  be a Wasserstein gradient update (6.2). Suppose  $n = 2^k - 1$ . If we*  
 245 *set  $\eta_n$  to be a silver step-size, then we get*

$$\mathcal{F}(\mu_n) - \mathcal{F}(\mu_*) \leq r_k L W_2^2(\mu_0, \mu_*).$$

246 *Suppose  $\mathcal{F}$  is, in addition, geodesically  $\alpha$ -strongly convex with the condition number  $\kappa = L/\alpha$ . Let*  
 247  *$k^* = \lceil \log_{\rho} \kappa \rceil + 1$ . Then by restarting silver step-size RGD every  $(2^{k^*} - 1)$  steps for  $\ell$  times, so that*  
 248  *$n = \ell(2^{k^*} - 1)$ , one obtains*

$$W_2^2(\mu_n, \mu_*) \leq \exp(-\log(\rho/2)n/\kappa^{\log_{\rho} 2}) W_2^2(\mu_0, \mu_*).$$

249 *Again, the algorithm finds an  $\epsilon$ -approximate solution in  $O(\kappa^{\log_{\rho} 2} \log(1/\epsilon))$  number of iterations.*

<sup>1</sup>The notion of generalized geodesic convexity and geodesic smoothness in Wasserstein space is introduced in Definition A.35.

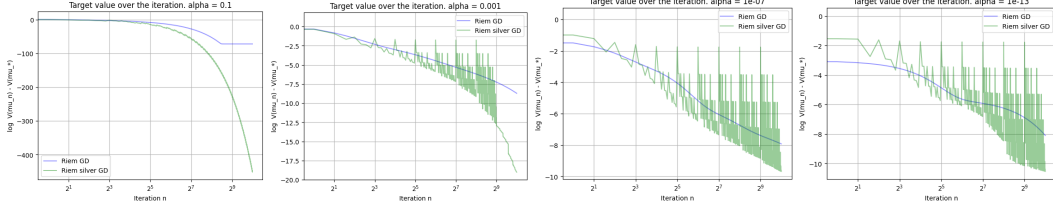


Figure 2: Comparison between silver step-size method and RGD for potential functional optimization in  $BW(\mathbb{R}^d)$ , with different convexity parameters. We set  $\ell = 2^{\lfloor \log_2(\frac{2^{10}-1}{2^{k^*}-1}) \rfloor}$  and  $n = \ell(2^{k^*} - 1)$ , where  $k^*$  being the optimal sub-iterate derived in Theorem 4.2. **Columns:** From left to right, each column corresponds to  $\kappa = 10^1, 10^3, 10^7, 10^{13}$ .

The proof is a direct application of our main theorems and is deferred to Appendix B.2.

For experiments, we set  $N$  to be the Bures-Wasserstein space  $BW(\mathbb{R}^d)$ , the space of non-singular Gaussian distributions in  $\mathbb{R}^d$  equipped with Wasserstein geometry. This set is a geodesically convex subset of  $\mathcal{P}_{2,ac}(\mathbb{R}^d)$ . Moreover,  $BW(\mathbb{R}^d)$  can be identified with a product Riemannian manifold of mean vectors and covariance matrices. Then, (6.2) becomes

$$(m_{n+1}, \Sigma_{n+1}) = \exp_{(m_n, \Sigma_n)}(-\eta_n \text{Grad}_{BW} \mathcal{F}(m_n, \Sigma_n)) \quad (6.3)$$

with  $\exp_{(m_n, \Sigma_n)}(\cdot)$  and Bures-Wasserstein gradient  $\text{Grad}_{BW} \mathcal{F}(m_n, \Sigma_n)$  defined in Definition A.38, A.39. We introduce more detail of  $BW(\mathbb{R}^d)$  geometry in Appendix A.2.1. As our objective functional, we consider an important functional in this space, the *potential functional*:

$$\mathcal{V}(\mu) := \mathbb{E}_{x \sim \mu}[V(x)]$$

where  $V : \mathbb{R}^d \rightarrow \mathbb{R}$ . The following proposition indicates that the potential functional satisfies the conditions required for Corollary 6.1 whenever  $V$  is convex and smooth.

**Proposition 6.2.** *If  $V$  is  $\alpha$ -strongly convex ( $L$ -smooth) in  $\mathbb{R}^d$ , then  $\mathcal{V}$  is generalized geodesically  $\alpha$ -strongly convex (resp.  $L$ -smooth) under both the Wasserstein and Bures-Wasserstein geometries.*

Although Proposition 6.2 is well known [DBCS23][Lemma B.1], we include the proof in Appendix A.2 for completeness. Using the explicit formula of  $\text{Grad}_{BW} \mathcal{V}(m, \Sigma)$  [DBCS23] in (6.3), we obtain the following silver step-size RGD in  $BW(\mathbb{R}^d)$  for  $\mathcal{V}(\mu)$  (6.4):

$$\begin{aligned} m_{n+1} &= m_n - \eta_n \mathbb{E}_{X \sim N(m_n, \Sigma_n)}[\nabla V(X)], \\ \Sigma_{n+1} &= (I - \eta_n \mathbb{E}_{X \sim N(m_n, \Sigma_n)}[\nabla^2 V(X)]) \Sigma_n (I - \eta_n \mathbb{E}_{X \sim N(m_n, \Sigma_n)}[\nabla^2 V(X)]). \end{aligned} \quad (6.4)$$

For our experiment, we choose  $V(x) = \frac{1}{2}(x - m_*)^T \Sigma_*^{-1}(x - m_*)$  defined on  $\mathbb{R}^{10}$ , with  $m_*, \Sigma_*$  being a randomly generated vector and symmetric positive definite matrix respectively. Since  $V$  is a strongly-convex quadratic function, by Proposition 6.2  $\mathcal{V}$  is generalized geodesically  $\alpha$ -strongly convex and geodesically  $L$ -smooth with  $L = 1/\lambda_{\min}(\Sigma_*)$  and  $\alpha = 1/\lambda_{\max}(\Sigma_*)$ . To study the effect of the condition number  $\kappa = L/\alpha$ , we fix  $L = 1$ , and vary  $\alpha$ . Small  $\alpha$  corresponds to convex case, and larger  $\alpha$  stands for the strongly convex case. We choose  $1/L$  as the step-size for constant step-size RGD [KY22]. Figure 2 shows that the silver step-size RGD outperforms constant step-size RGD in both convex and strongly convex case. We provide further implementation detail (e.g., the specific distributions of  $m_*$  and  $\Sigma_*$ ) and additional experiments under various settings (e.g., different random seeds, number of iterations, and comparisons with various constant step-sizes) in Appendix E.

## 6.2 Optimization on the Sphere: Rayleigh Quotient Maximization

While certain functionals are known to be geodesically convex in Wasserstein space, identifying such structure in other Riemannian manifolds is more subtle. Still, Riemannian optimization algorithms have shown strong empirical performance even in the absence of geodesic convexity or smoothness guarantees [AOBL21, KY22, HMJG23]. In this spirit, even though the objective functions are not



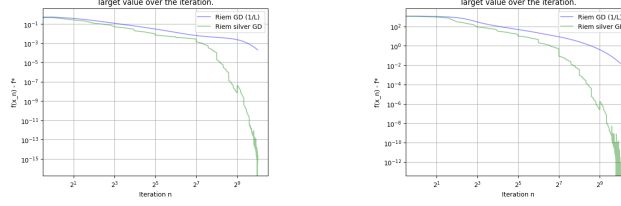


Figure 3: Comparison between silver step-size method and RGD for Rayleigh quotient maximization problem on  $\mathbb{S}^{2500}$ . We set  $n = 2^{10} - 1$  as the total iteration number. **Left:**  $H$  with small eigenvalue gaps. **Right:**  $H$  with large eigenvalue gaps.

generalized geodesically convex, we evaluate our method on the benchmark problem of Rayleigh quotient maximization on the  $d$ -dimensional unit sphere  $\mathbb{S}^{d-1}$ , a standard Riemannian manifold with constant positive curvature  $K_{\min} = K_{\max} = 1$ . Let  $H \in \mathbb{R}^{d \times d}$  be a symmetric matrix, with largest and smallest eigenvalues denoted by  $\lambda_{\max}$  and  $\lambda_{\min}$  respectively. Consider the Rayleigh quotient maximization problem:

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

$f$  is geodesically  $(\lambda_{\max} - \lambda_{\min})$ -smooth [KY22][Proposition 7.1], while not geodesically convex. While the problem is not generalized geodesically convex, Figure 3 illustrates the effectiveness of our method on this problem. The figure is based on experiments conducted on  $\mathbb{S}^{2500}$ . We consider two cases of  $H$ : (1)  $H = \frac{1}{2}(A + A^T)$  where the entries of  $A$  are randomly generated from  $N(0, 1/d)$  as in [KY22] (corresponding to small eigenvalue gaps); and (2) a randomly generated symmetric matrix with  $\lambda_{\max} = d$  and  $\lambda_{\min} = -d$  (corresponding to large eigenvalue gaps). Again, we compared the performance with constant step-size RGD using a step-size of  $1/L$ .

## 7 Conclusion

In this work, we show that for generalized geodesically convex and geodesically  $L$ -smooth functionals on Riemannian manifolds with non-negative curvature, RGD with a silver step-size schedule achieves an accelerated convergence rate—matching that of the Euclidean case. Albeit under a stronger notion of convexity, our algorithm is the first tractable accelerated algorithm for Riemannian manifolds without the curvature upper bound, in particular for the Wasserstein space. A key theoretical novelty of our analysis is that it avoids relying on the crucial equality in [AP24c][Equation (8)] that does not hold on Riemannian manifolds. Instead, our proof is based on the inequalities (Lemma 5.1, 5.2) which accounts for the metric distortion on Riemannian manifolds. Furthermore, we extended the silver step-size analysis to geodesically strongly convex case without modifying the step-size itself using the restarting method. We illustrate our theoretical results on practical problems.

We conclude the paper with some open questions: 1. Proposition 3.8 states a sufficient condition for (3.1), but whether the inequality can be derived from standard geodesic convexity and  $L$ -smoothness alone remains an open question. 2. Many manifolds of interest have negative curvature, where exponential and logarithmic maps are globally defined [ZS18]. As a result, prior work has often focused on non-positively curved settings [AS20, CB23]. However, our analysis does not readily extend to such manifolds (see Appendix D.3 for a heuristic explanation). Extending it to these settings remains an open challenge. 3. It will be quite interesting to extend silver step-size acceleration to other variants of RGD such as stochastic RGD and proximal RGD. 4. Recently, for specific classes of functions in the Euclidean setting, [AP24a] proposed a step-size schedule for gradient descent that achieves the fully accelerated rate  $O(1/n^2)$ , matching that of momentum methods. Extending these ideas to the Riemannian setting would be an intriguing direction for future work.

## References

- [ACGS21] Jason Altschuler, Sinho Chewi, Patrik Robert Gerber, and Austin J Stromme. Averaging on the bures-wasserstein manifold: dimension-free convergence of gradient descent. In *Advances in Neural Information Processing Systems*, 2021.
- [AG08] Luigi Ambrosio and Nicola Gigli. Construction of the parallel transport in the wasserstein space. *Methods Appl. Anal.*, 15(1):1–30, 2008.
- [AGS08] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows in Metric Spaces and in the Space of Probability Measures*. Lectures in Mathematics ETH Zürich. Springer Science & Business Media, 2ed edition, 2008.
- [Alt18] Jason M. Altschuler. Greed, hedging, and acceleration in convex optimization. *Master’s thesis*, 2018.
- [AOBL20] Foivos Alimisis, Antonio Orvieto, Gary Becigneul, and Aurelien Lucchi. A continuous-time perspective for modeling acceleration in riemannian optimization. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, 2020.
- [AOBL21] Foivos Alimisis, Antonio Orvieto, Gary Becigneul, and Aurelien Lucchi. Momentum improves optimization on riemannian manifolds. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, 2021.
- [AP24a] Jason M. Altschuler and Pablo A. Parrilo. Acceleration by random stepsizes: Hedging, equalization, and the arcsine stepsize schedule, 2024.
- [AP24b] Jason M. Altschuler and Pablo A. Parrilo. Acceleration by stepsize hedging: Multi-step descent and the silver stepsize schedule. *J. ACM*, 2024.
- [AP24c] Jason M. Altschuler and Pablo A. Parrilo. Acceleration by stepsize hedging: Silver stepsize schedule for smooth convex optimization. *Mathematical Programming*, 2024.
- [AS20] Kwangjun Ahn and Suvrit Sra. From nesterov’s estimate sequence to riemannian acceleration. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, 2020.
- [BA24] Jinho Bok and Jason M. Altschuler. Accelerating proximal gradient descent via silver stepsizes, 2024.
- [BH06] Rajendra Bhatia and John Holbrook. Riemannian geometry and matrix geometric means. *Linear Algebra and its Applications*, 413, 2006.
- [BJL19] Rajendra Bhatia, Tanvi Jain, and Yongdo Lim. On the bures–wasserstein distance between positive definite matrices. *Expositiones Mathematicae*, 37(2):165–191, 2019.
- [Bou23] Nicolas Boumal. *An introduction to optimization on smooth manifolds*. Cambridge University Press, 2023.
- [Bre91] Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on Pure and Applied Mathematics*, 44, 1991.
- [CB18] Lénaïc Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems, NIPS’18*, page 3040–3050. Curran Associates Inc., 2018.

- [CB22] Christopher Criscitiello and Nicolas Boumal. An accelerated first-order method for non-convex optimization on manifolds. *Foundations of Computational Mathematics*, 2022.
- [CB23] Christopher Criscitiello and Nicolas Boumal. Curvature and complexity: Better lower bounds for geodesically convex optimization. In *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, 2023.
- [CCT18] José A. Carrillo, Young-Pil Choi, and Oliver Tse. Convergence to equilibrium in wasserstein distance for damped euler equations with interaction forces. *Communications in Mathematical Physics*, 365(1):329–361, oct 2018.
- [Che24] Sinho Chewi. Log-concave sampling, 2024.
- [DBCS23] Michael Ziyang Diao, Krishna Balasubramanian, Sinho Chewi, and Adil Salim. Forward-backward gaussian variational inference via jko in the bures-wasserstein space. In *International Conference on Machine Learning*, pages 7960–7991. PMLR, 2023.
- [dST21] Alexandre d’Aspremont, Damien Scieur, and Adrien Taylor. Acceleration methods. *Foundations and Trends® in Optimization*, 5(1-2):1–245, 2021.
- [FAP<sup>+</sup>05] Pierre Fillard, Vincent Arsigny, Xavier Pennec, Paul M. Thompson, and Nicholas Ayache. Extrapolation of sparse tensor fields: Application to the modeling of brain variability. *Lecture Notes in Computer Science*, 3565, 2005.
- [FRF22] Xavier Fernández-Real and Alessio Figalli. *The Continuous Formulation of Shallow Neural Networks as Wasserstein-Type Gradient Flows*, pages 29–57. Springer International Publishing, 05 2022.
- [HMJG21] Andi Han, Bamdev Mishra, Pratik Kumar Jawanpuria, and Junbin Gao. On riemannian optimization over positive definite matrices with the bures-wasserstein geometry. In *Advances in Neural Information Processing Systems*, volume 34. Curran Associates, Inc., 2021.
- [HMJG23] Andi Han, Bamdev Mishra, Pratik Jawanpuria, and Junbin Gao. Riemannian accelerated gradient methods via extrapolation. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, 2023.
- [JKO98] Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the fokker-planck equation. *SIAM J. Math. Anal.*, 29(1):1–17, jan 1998.
- [KPB25] Jakwang Kim, Jiyoung Park, and Anirban Bhattacharya. Robust estimation in metric spaces: Achieving exponential concentration with a fr’echet median. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025.
- [KY22] Jungbin Kim and Insoon Yang. Accelerated gradient methods for geodesically convex optimization: Tractable algorithms and convergence analysis. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*. PMLR, 2022.
- [LCB<sup>+</sup>22] Marc Lambert, Sinho Chewi, Francis Bach, Silvère Bonnabel, and Philippe Rigollet. Variational inference via wasserstein gradient flows. *Advances in Neural Information Processing Systems*, 35:14434–14447, 2022.
- [Lee12] John M. Lee. *Introduction to Smooth Manifolds*, volume 218 of *Graduate Texts in Mathematics*. Springer, 2nd edition, 2012.
- [Lee18] John M. Lee. *Introduction to Riemannian Manifolds*. Springer International Publishing, 2nd edition, 2018.

- [Lot07] John Lott. Some geometric calculations on wasserstein space, 2007.
- [LSC<sup>+</sup>17] Yuanyuan Liu, Fanhua Shang, James Cheng, Hong Cheng, and Licheng Jiao. Accelerated first-order methods for geodesically convex optimization on riemannian manifolds. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.
- [MGB<sup>+</sup>20] Nina Miolane, Nicolas Guigui, Alice Le Brigant, Johan Mathe, Benjamin Hou, Yann Thanwerdas, Stefan Heyder, Olivier Peltre, Niklas Koep, Hadi Zaatiti, Hatem Hajri, Yann Cabanes, Thomas Gerald, Paul Chauchat, Christian Shewmake, Daniel Brooks, Bernhard Kainz, Claire Donnat, Susan Holmes, and Xavier Pennec. Geomstats: A python package for riemannian geometry in machine learning. *Journal of Machine Learning Research*, 21(223):1–9, 2020.
- [MMN18] Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences of the United States of America*, 115:E7665 – E7671, 2018.
- [MPU<sup>+</sup>24] Nina Miolane, Luís F. Pereira, Saiteja Utpala, Nicolas Guigui, Alice Le Brigant, Hzaatiti, Yann Cabanes, Johan Mathe, Niklas Koep, elodiemaignant, ythanwerdas, xpennec, tgeral68, Christian, Tra My Nguyen, Olivier Peltre, John Harvey, pchauchat, julesdeschamps, Quentin Barthélemy, mortenapedersen, Maya95assal, Abdellaoui-Souhail, Adele Myers, Felix Ambellan, Florent-Michel, Stefan Heyder, Shubham Talbar, Yann de Mont-Marin, and Marius. geomstats/geomstats: Geomstats v2.8.0, September 2024.
- [MR22] David Martínez-Rubio. Global riemannian acceleration in hyperbolic and spherical spaces. In *Proceedings of The 33rd International Conference on Algorithmic Learning Theory*, volume 167 of *Proceedings of Machine Learning Research*, 2022.
- [Nes83] Yurii Nesterov. A method for solving the convex programming problem with convergence rate  $O(1/k^2)$ . *Proceedings of the USSR Academy of Sciences*, 269, 1983.
- [Nes14] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer Publishing Company, Incorporated, 1 edition, 2014.
- [Ngu22] Xuan Son Nguyen. A gyrovector space approach for symmetric positive semi-definite matrix learning. In *Computer Vision – ECCV 2022*. Springer Nature Switzerland, 2022.
- [Nie23] Frank Nielsen. A simple approximation method for the fisher–rao distance between multivariate normal distributions. *Entropy*, 25, 2023.
- [OC15] Brendan O’Donoghue and Emmanuel Candès. Adaptive restart for accelerated gradient schemes. *Foundations of Computational Mathematics*, 15:715–732, 2015.
- [Ott01] Felix Otto. The geometry of dissipative evolution equations: The porous medium equation. *Communications in Partial Differential Equations*, 26, 2001.
- [PFA05] Xavier Pennec, Pierre Fillard, and Nicholas Ayache. A riemannian framework for tensor computing. *International Journal of Computer Vision*, 66, 2005.
- [San14] Filippo Santambrogio. Introduction to optimal transport theory, 2014.
- [Sie21] Jonathan W. Siegel. Accelerated optimization with orthogonality constraints. *Journal of Computational Mathematics*, 39, 2021.
- [SKL20] Adil Salim, Anna Korba, and Giulia Luise. The wasserstein proximal gradient algorithm. In *Advances in Neural Information Processing Systems*, 2020.
- [Tak09] Asuka Takatsu. On wasserstein geometry of the space of gaussian measures, 2009.

444 [TKW16] James Townsend, Niklas Koep, and Sebastian Weichwald. Pymanopt: A python toolbox  
445 for optimization on manifolds using automatic differentiation. *Journal of Machine*  
446 *Learning Research*, 17(137):1–5, 2016.

447 [TP22] Yann Thanwerdas and Xavier Pennec.  $O(n)$ -invariant riemannian metrics on spd matrices.  
448 *Linear Algebra and its Applications*, 661, 2022.

449 [Vil08] Cédric Villani. Optimal transport: Old and new. 2008.

450 [WL22] Yifei Wang and Wuchen Li. Accelerated information gradient flow, 2022.

451 [Woj20] Stephan Wojtowytsch. On the convergence of gradient descent training for two-layer  
452 relu-networks in the mean field regime, 2020.

453 [You53] David Young. On richardson’s method for solving linear systems with positive definite  
454 matrices. *Journal of Mathematics and Physics*, 1953.

455 [ZS16] Hongyi Zhang and Suvrit Sra. First-order methods for geodesically convex optimization.  
456 In *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine*  
457 *Learning Research*, 2016.

458 [ZS18] Hongyi Zhang and Suvrit Sra. An estimate sequence for geodesically convex opti-  
459 mization. In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of  
460 *Proceedings of Machine Learning Research*, 2018.

## NeurIPS Paper Checklist

### 1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: Abstract and introduction clearly states our main theorems 4.1: achieving the acceleration using varying step-size under non-negatively curved Riemannian manifolds. In addition, our main application and experiments are well-stated as well.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

### 2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: In Section 3, we clarify which of the assumptions are less standard, and in Section 5, we explain why these assumptions are required for the theoretical analysis. That said, Section 6 presents numerical results demonstrating that our method remains effective even when some of these assumptions are relaxed. Finally, the Conclusion mentions directions that may refine or extend the current approach.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.



- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

### 3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification: We explicitly stated the assumptions in our main body (Assumption [3.1](#), [3.2](#)). In Section [5](#), we provide the sketch of the proof, and the full proof in the Appendix [B.1](#), [B.2](#). We also provided the full reference for the results we cited.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

### 4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: We provide the sufficient detail of the experiment we conducted in Section [6](#) and Appendix [E](#). We also provided the code in our supplementary material.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

## 5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide the full code in our supplementary material, which can reproduce the results of our experiments.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

## 6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We provided the data generating process of the simulation data and explicit algorithm in Section 6 and Appendix E.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

## 7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [\[Yes\]](#)

Justification: We provide the result of multiple experiments and statistical significance in Appendix E.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

## 8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [\[Yes\]](#)

Justification: We stated our computational resource in Appendix E.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

## 9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: Our research is conducted in compliance with the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

## 10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: There is no societal impact of the work performed. Our focus is purely on the algorithmic aspects, rather than on societal applications.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

## 11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Our research does not involve the high risk for misuse.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

## 12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We mentioned and cited the packages we used in Appendix E.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, [paperswithcode.com/datasets](https://paperswithcode.com/datasets) has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

## 13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: The paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

## 14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

#### 15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

#### 16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.