Online Zeroth-order Optimisation on Hadamard Manifolds

Alejandro I. Maass\(^1\), Chris Manzie\(^1\), Dragan Neˇsi´c\(^1\), Jonathan H. Manton\(^1\), and Iman Shames\(^2\)

\(^1\)Department of Electrical and Electronic Engineering, The University of Melbourne
\(^2\)College of Engineering & Computer Science, The Australian National University
* e-mails: \{maassa,manziec,dnesic,jmanton\}@unimelb.edu.au; iman.shames@anu.edu.au

Abstract

We study numerical optimisation algorithms that use zeroth-order information to minimise time-varying geodesically-convex cost functions on Riemannian manifolds. In the Euclidean setting, zeroth-order algorithms have received a lot of attention in both the time-varying and time-invariant case. However, the extension to Riemannian manifolds is much less developed. We focus on Hadamard manifolds, which are a special class of Riemannian manifolds with global nonpositive curvature that offer convenient grounds for the generalisation of convexity notions. Specifically, we derive bounds on the expected instantaneous tracking error, and we provide algorithm parameter values that minimise the algorithm’s performance. Our results illustrate how the manifold geometry in terms of the sectional curvature affects these bounds. Additionally, we provide dynamic regret bounds for this online optimisation setting. To the best of our knowledge, these are the first bounds on tracking error and dynamic regret for online zeroth-order Riemannian optimisation. Lastly, via numerical simulations, we demonstrate the applicability of our algorithm on an online Karcher mean problem.

1 Introduction

Time-varying optimisation problems are popular in the machine learning community under the framework known as online convex optimisation (OCO) \([22]\). OCO is a promising methodology for modelling sequential tasks, and the main goal is developing algorithms that can track trajectories of the optimisers of the time-varying optimisation problem (up to asymptotic error bounds). We refer the reader to \([37]\) for a review of available algorithms and applications. In this paper, we study the following sequence of optimisation problems

\[
\min_{x \in \mathcal{X} \subseteq \mathcal{M}} f_k(x),
\]

where \(\mathcal{M}\) is a Hadamard submanifold embedded in \(\mathbb{R}^n\), \(\mathcal{X}\) is a closed geodesically convex set, and \(f_k\) belongs to the class of continuous functions

\[
\mathcal{F} := \{f_k : \mathcal{M} \to \mathbb{R} \mid k \in \mathbb{K}\},
\]

where \(\mathbb{K} := \mathbb{N}_0 \cup \left\{ j + \frac{1}{2} \mid j \in \mathbb{N}_0 \right\} = \{0, 1/2, 1, 3/2, 2, 5/2, \ldots \}\), and every \(\{f_k\}_{k \in \mathbb{K}} \in \mathcal{F}\) is geodesically \(L\)-smooth and geodesically strongly convex (see Definitions 2 and 3 further below). Hadamard
manifolds are a special class of Riemannian manifolds that provide proper grounds for the generalisation of convexity notions from the Euclidean setting to nonlinear spaces [3]. Technically speaking, Hadamard manifolds are Riemannian manifolds that are complete, simply connected, and with global nonpositive curvature [7]—we define these concepts formally in Section 2. Some classical examples include hyperbolic spaces, manifolds of positive definite matrices, and $\mathbb{R}^n$.

We assume that the analytical form of the functions $\{f_k\}_{k \in \mathbb{K}}$ and their gradients are not available, and we can only obtain function evaluations via a zeroth-order oracle. Each cost function is thus seen as a black-box with time-varying input-output map. This scenario is found in many applications where derivatives are either unavailable, or too expensive to compute, see e.g. [29, 30] for machine learning, [23] for online controller tuning, [14] for deep neural networks, and [17] for mobile fog computing.

Consequently, we aim to generate solutions to (1) using random gradient-free iterates of the form

$$x_{k+1} = \mathcal{P}_X \left[ \text{Exp}_{x_k}(-\alpha_k g_{\eta,k^+}(x_k, u_k)) \right],$$

where $\mathcal{P}_X$ denotes the projection that maps a point $x \in \mathcal{M}$ to $\mathcal{P}_X(x) \in \mathcal{X} \subset \mathcal{M}$ such that $\text{dist}(x, \mathcal{P}_X(x)) < \text{dist}(x, y)$, for all $y \in \mathcal{X} \setminus \{\mathcal{P}_X(x)\}$, and $\mathcal{P}_X(x) = x$ for $x \in \mathcal{X}$. The positive constant $\alpha_k$ denotes the step size, $g_{\eta,k^+}$ is the oracle, and $\text{Exp}_{x_k}(\cdot)$ denotes the exponential mapping which we define further below. We consider an extension of the recently proposed zeroth-order oracle for optimisation over Riemannian manifolds in [27] to make it suitable for our online optimisation setting. That is,

$$g_{\eta,k^+}(x, u) := \frac{f_{k^+}(\text{Exp}_x(\eta u)) - f_k(x)}{\eta},$$

where $k^+ := k + 1/2$, $k \in \mathbb{N}_0$, $u = P u_0 \in T_x \mathcal{M}$, $u_0 \sim \mathcal{N}(0, I_n) \in \mathbb{R}^n$, and $P \in \mathbb{R}^{n \times n}$ is the orthogonal projection matrix onto the tangent space $T_x \mathcal{M}$ of $\mathcal{M}$ at $x$. We note that the proposed oracle is based on the one recently introduced in [27] for time-invariant cost functions; however, ours allows for $f$ to be time-varying as per our class of functions $\mathcal{F}$. Note that we are using a particular case of the general retraction $R_x(\eta u)$ in [27] which is the exponential mapping $\text{Exp}_x(\eta u)$.

### 1.1 Related work

Our work lies in the general field of optimisation over Riemannian manifolds which has received a lot of attention recently in the literature given its applications to machine learning [43], signal processing [31], dictionary learning [38], and low-rank matrix completion [39], among others. For the time-invariant—or offline—setting, several results have been proposed in the literature, and we briefly review them below. For instance, [10] provided convergence rates for deterministic Riemannian gradient descent and smooth cost functions. Stochastic algorithms were also considered for smooth Riemannian optimisation in [8, 24, 41, 43, 45]. Particularly, [8] extended the classical stochastic gradient descent algorithms to the Riemannian case, and provided convergence results. The authors in [43] introduced the variance–reduced RSVRG method and considered Riemannian optimisation of finite sums of geodesically smooth functions. The work [24] proposed a Riemannian stochastic recursive gradient algorithm (RSRG) which provides notable computational advantages in comparison to RSVRG. The work [45] introduced the Riemannian SPIDER method for non-convex Riemannian optimisation as a simple and efficient extension of the Euclidean SPIDER counterpart. Lastly, [41] studied stochastic projection-free methods for constrained optimisation of smooth functions on Riemannian manifolds. The stochastic Riemannian Frank-Wolfe methods for
nonconvex and geodesically convex problems are introduced. For non-smooth cost functions, Riemannian subgradient methods have been proposed in [28], manifold ADMM methods in [25], manifold proximal gradient (ManPG) methods in [16], manifold proximal point algorithms (ManPPA) in [15], and stochastic ManPG in [40].

None of the aforementioned works have considered the zeroth-order setting, in which the oracle makes available only cost function values as opposed to first-order or second-order information. To the best of our knowledge, zeroth-order Riemannian optimisation for time-invariant cost functions have been considered in [13,21,27]. Particularly, [21] presented the extended Riemannian stochastic derivative-free optimisation (RSDFO) algorithm, and proved it converges in finitely many steps in compact connected Riemannian manifolds. The authors in [13] extended the derivative-free optimisation method by [33] to Riemannian manifolds, but did not provide any complexity or convergence results. Just recently, [27] provided the first complexity results for both deterministic and stochastic zeroth-order Riemannian optimisation. Their zeroth-order methods rely on an estimator of the Riemannian gradient based on a modification of the Gaussian smoothing technique from the seminal work by Nesterov [32]. In [27], the authors illustrated that the proposed zeroth-order method has a comparable performance to its first-order counterparts in many applications of interest such as matrix approximation, k-PCA, sparse PCA, and the Karcher mean problem.

We note that the existing works on zeroth-order Riemannian optimisation listed above do not consider the online setting (1) in which the cost-function is allowed to be time-varying. This problem, however, has been widely studied for $\mathcal{M} \equiv \mathbb{R}^n$, see e.g. [5,12,18,19,42], in which it is assumed that cost function evaluations are carried out simultaneously when computing two-point estimates of the gradient. Later on, [36] relaxed this assumption and allowed the cost function to change between function evaluations, which added an extra modelling layer that better respects the time-varying nature of the problem.

1.2 Contributions

Our contributions are threefold:

- We extend the OCO framework presented in [36] from the Euclidean setting to the case where the cost function is defined on a Hadamard manifold. Our proposed algorithm uses an extension of the zeroth-order oracle recently presented in [27] that allows for the function to be time-varying.

- We provide asymptotic bounds on the expected instantaneous tracking error, which to the best of our knowledge, are the first error bounds for online zeroth-order optimisation on Riemannian manifolds available in the literature. Our results illustrate how the manifold geometry—in terms of the sectional curvature—influences the performance of the algorithm. In addition, we provide explicit choices for the algorithm parameters—step size and oracle’s precision—that minimise the performance of the algorithm.

- Lastly, we provide dynamic regret bounds for our time-varying setting. These are also the first available regret bounds in the literature for online zeroth-order optimisation on Riemannian manifolds, which had not been previously derived even for the Euclidean case.

2 Preliminaries

In this section, we present a brief introduction on the basics of manifold optimization. For a more in-depth revision we refer the reader to [1]. A smooth manifold is a topological manifold $\mathcal{M}$ with
a globally defined differentiable structure. At any point \( x \) on a smooth manifold, tangent vectors are defined as the tangents of parametrised curves passing through \( x \). The tangent space \( T_x M \) of a manifold \( M \) at \( x \) is defined as the set of all tangent vectors at the point \( x \). Tangent vectors on manifolds generalise the notion of a directional derivative. Formally, we can define the tangent space \( T_x M \) as follows

\[
T_x M := \{ \gamma'(0) : \gamma(0) = x, \; \gamma([-\varepsilon, \varepsilon]) \subset M \text{ for some } \varepsilon > 0, \; \gamma \text{ is differentiable} \}.
\]

The tangent bundle of a differentiable manifold \( M \) is the manifold \( TM \) that assembles all the tangent vectors in \( M \). As a set, it is the disjoint union of all the tangent planes, i.e. \( TM := \sqcup_{x \in M} T_x M \). The dimension of a manifold \( M \), denoted as \( d \), is the dimension of the Euclidean space that the manifold is locally homeomorphic to. In particular, the dimension of the tangent space is always equal to the dimension of the manifold.

A Riemannian manifold is a couple \((M, g)\), where \( M \) is a smooth manifold equipped with a smoothly varying inner product (Riemannian metric) on the tangent space at every point, i.e. \( g(\cdot, \cdot) := \langle \cdot, \cdot \rangle_x : T_x M \times T_x M \to \mathbb{R} \). Without loss of generality, when the Riemannian metric is clear from the context, we simply talk about “the Riemannian manifold \( M \)”. Throughout, we assume the Levi-Civita connection is associated with \((M, g)\).

**Definition 1 (Riemannian gradient)** Suppose \( f \) is a smooth function on the Riemannian manifold \( M \). The Riemannian gradient \( \nabla f(x) \) is defined as the unique element of \( T_x M \) satisfying 

\[
\langle \xi, \nabla f(x) \rangle_x = \frac{d}{dt} f(\gamma(t)) \bigg|_{t=0} \text{ for any } \xi \in T_x M, \text{ where } \gamma(t) \text{ is a curve in } M \text{ such that } \gamma(0) = x \text{ and } \gamma'(0) = \xi.
\]

A Riemannian submanifold \( M \) of a Riemannian manifold \( \mathcal{N} \) is a submanifold of \( \mathcal{N} \) equipped with the Riemannian metric inherited from \( \mathcal{N} \). Since in this paper we assume that \( M \) is a Riemannian submanifold embedded in \( \mathbb{R}^n \), \( M \) is equipped with the Riemannian metric inherited from \( \mathbb{R}^n \). We thus write the inner product on the tangent space \( \langle \cdot, \cdot \rangle_x \) at every point \( x \in M \) as \( \langle \cdot, \cdot \rangle_x = \langle \cdot, \cdot \rangle \), with the right-hand side being the Euclidean inner-product. Consequently, the Riemannian gradient in Definition 1 becomes the projection of its Euclidean gradient onto the tangent space, that is \( \nabla f(x) = P_{T_x M}(\nabla f(x)) \), where \( \nabla f(x) \) denotes the Euclidean gradient of \( f \) at \( x \).

We now introduce a family of local parametrisations often called retractions. In a nutshell, retractions allow us to move on a manifold (i.e. move in the direction of a tangent vector) while staying on the manifold. Formally, we say that a retraction mapping \( R_x \) is a smooth mapping from \( T_x M \) to \( M \) such that \( R_x(0) = x \), and the differential at 0 is an identity mapping, i.e. \( \frac{d}{dt} R_x(t\xi) \bigg|_{t=0} = \xi \), for all \( \xi \in T_x M \). We refer to the latter as the local rigidity condition. In other words, for every tangent vector \( \xi \in T_x M \), the curve \( \gamma_\xi : t \mapsto R_x(t\xi) \) satisfies \( \gamma_\xi(0) = \xi \). Moving along the curve \( \gamma_\xi \) is thought of as moving in the direction \( \xi \) while constrained to the manifold \( M \). In particular, the exponential mapping \( \text{Exp}_x \) is a retraction that generates geodesics. A geodesic is a curve representing in some sense the shortest path between two points in a Riemannian manifold.

Throughout this paper, we assume that \( M \) is a Hadamard manifold, which is a special type of Riemannian manifold. A Hadamard manifold is complete, simply connected, and has nonpositive sectional curvature everywhere \([3, 7]\). Complete refers to the domain of the exponential mapping being the whole tangent bundle \( TM \), and simply connected means there are no circular paths that cannot be shrunk to a point. Hadamard manifolds have strong properties. For instance, there exists a unique geodesic between any two points on \( M \), and the exponential map is globally invertible at any point, \( \text{Exp}_x^{-1} : M \to T_x M \). The geodesic distance is thus given by \( \text{dist}(x, y) = \|\text{Exp}_x^{-1}(y)\| = \|\text{Exp}_x^{-1}(y)\| \), where \( \|\cdot\| \) is the norm associated with the Riemannian metric, which in our setting corresponds to the Euclidean norm as discussed above. On a Hadamard manifold, the
notion of parallel transport provides a way to transport a vector along a geodesic. It is defined as the operator $\Gamma^y_x : T_xM \to T_yM$ which maps $v \in T_xM$ to $\Gamma^y_x(v) \in T_yM$ while preserving the inner product, i.e. $\langle u, v \rangle_x = \langle \Gamma^y_x(u), \Gamma^y_x(v) \rangle_y$.

We introduce some important definitions for our class of cost functions.

**Definition 2 (Geodesically $L$-smoothness)** A differentiable function $f \in F$ is said to be geodesically $L$-smooth if there exists $L \geq 0$ such that the following inequality holds for all $x, y \in M$,

$$\|\text{grad} f(x) - \Gamma^y_x(\text{grad} f(y))\| \leq L \|x - y\|, \quad \text{(4)}$$

where we recall that $\|x - y\|$ denotes the geodesic distance between $x$ and $y$, and $\Gamma^y_x$ is the parallel transport from $y$ to $x$.

It can be shown that if $f$ is geodesically $L$-smooth, then for any $x, y \in M$ we have [44,45]

$$f(y) \leq f(x) + \langle \text{grad} f(x), \text{Exp}_x^{-1}(y) \rangle_x + \frac{L}{2} \|x - y\|^2. \quad \text{(5)}$$

**Definition 3 (Geodesic strong convexity)** A function $f \in F$ is said to be geodesically $\sigma$-strongly convex if there exists $\sigma \in (0, L]$ such that for any $x, y \in M$,

$$f(y) \geq f(x) + \langle \text{grad} f(x), \text{Exp}_x^{-1}(y) \rangle_x + \frac{\sigma}{2} \|x - y\|^2.$$

Throughout this paper, the expectation operator is defined as

$$\mathbb{E}[g(u_0)] := \frac{1}{\nu} \int_{\mathbb{R}^n} g(u_0)e^{-\frac{1}{2}\|u_0\|^2}du_0,$$

where $\nu$ is the normalising constant, see also [27].

### 3 Online optimisation using zeroth-order Riemannian oracles

Before presenting our results we state the underlying assumptions.

**Assumption 1**

(a) Every $f \in F$ in (1) is geodesically $L$-smooth and geodesically strongly convex as per Definitions 2 and 3, respectively.

(b) $\exists V \geq 0$ such that $\text{dist}(x_{k+1}^*, x_{k+1}^*) \leq V$, where $x_k^* := \arg \min_{x \in M} f_k(x)$ for all $k \in \mathbb{K}$.

(c) $\exists \delta > 0$ such that $|f_k(x) - f_{k+1}(x)| \leq \delta$ for all $x \in M$.

(d) The sectional curvature of $M$ is lower-bounded by $\kappa \leq 0$.

(e) $\exists R > 0$ such that $\max_{y, z \in \mathcal{X}} \|y - z\| \leq R$.

Similar assumptions have been used in the literature when $M \equiv \mathbb{R}^n$, see e.g. [36] and [19]. Particularly, (a) generalises assumptions such as Lipschitz gradient smoothness and strong convexity often adopted in convex optimisation over $\mathbb{R}^n$. Item (b) bounds the change in minimiser, and (c) bounds the variation between consecutive cost functions. Note that $\delta \rightarrow 0$ corresponds to the special case where the cost-function does not vary between evaluations and thus $g_{n,k+1}(x,u) = g_{n,k}(x,u)$.
in (3). Lastly, (d) assumes a lower bound in the curvature of $\mathcal{M}$, and (e) an upper bound in the diameter of $\mathcal{X}$, which are common assumptions in Riemannian optimisation see e.g. [8,44], and [27].

We note that since $f \in \mathcal{F}$ is geodesically strongly convex as per Assumption 1(a), it is immediate to show that each $f \in \mathcal{F}$ satisfies

$$ -\langle \text{grad} f(x), \text{Exp}_{x}^{-1}(x^*) \rangle \geq \frac{\sigma}{2} \text{dist}(x, x^*)^2, $$

for all $x \in \mathcal{M}$, where $x^* := \arg \min_{x \in \mathcal{M}} f(x)$. We will use this inequality in our analysis further below. Inequality (6) is the Riemannian counterpart of the restricted secant inequality in $\mathbb{R}^n$.

### 3.1 Error bounds

The main objective of this section is to show that the proposed algorithm (2) can closely follow (or track) the optimisers of the time-varying optimisation problem (1) as time grows. To that end, we define the tracking error as

$$ e_k := \text{dist}(x_k, x_k^+) = \|\text{Exp}_{x_k}^{-1}(x_k^*)\| = \|\text{Exp}_{x_k^+}^{-1}(x_k)\|, $$

and the estimation error as

$$ \bar{e}_k := \text{dist}(x_{k+1}, x_k^+) = \|\text{Exp}_{x_{k+1}}^{-1}(x_k^*)\| = \|\text{Exp}_{x_k^+}^{-1}(x_{k+1})\|. $$

The objects $e_k$ and $\bar{e}_k$ are also known as pre-update optimality gap and post-update optimality gap, respectively, see e.g. [19].

To state our main tracking result, we require some intermediate steps which we present in the following. First, a common technique for analysing optimisation algorithms is to write the estimation error in terms of the tracking error by using the law of cosines in the Euclidean space. Unfortunately, this equality does not exist for general nonlinear spaces, and in fact, there are no corresponding analytical expressions. However, in [44], a trigonometric distance bound for Alexandrov spaces with curvature bounded below was proposed. Alexandrov spaces are length spaces with curvature bounded below and form a generalisation of Riemannian manifolds with sectional curvature bounded below. The result uses the properties of geodesic triangles, and it can be used as an analogue to the law of cosines given its fundamental nature. For our setting, this inequality is formalised in the lemma below.

**Lemma 1** For any Riemannian manifold $\mathcal{M}$ with a sectional curvature lower bounded by $\kappa$, and any points $x_k^+, x_k \in \mathcal{M}$, the update $x_{k+1} = \mathcal{P}_\mathcal{X}[\exp_{x_k}(-\alpha_k g_{\eta,k}(x_k, u_k))]$ satisfies

$$ e_k^2 \leq e_k^2 + 2\alpha_k \langle g_{\eta,k}(x_k, u_k), \exp_{x_k}^{-1}(x_k^+ \rangle + \zeta(\kappa, e_k) \alpha_k^2 \|g_{\eta,k}(x_k, u_k)\|^2, $$

where $\zeta(\kappa, e_k) := e_k \sqrt{|\kappa|}/\tanh(e_k \sqrt{|\kappa|})$.

**Proof** See Appendix B.

The next intermediate step is to derive a relation between the conditional expectations of $e_{k+1}$ and $\bar{e}_k$. By Assumption 1(b) and the triangle inequality for the geodesic distance $\text{dist}(\cdot, \cdot)$, we can write

$$ e_{k+1} = \text{dist}(x_{k+1}, x_{k+1}^+) $$

$$ \leq \text{dist}(x_{k+1}, x_k^+) + \text{dist}(x_k^+, x_{k+1}^+) $$

$$ \leq \bar{e}_k + 2V $$

$$ \mathbb{E}[e_{k+1} \mid x_k] \leq \mathbb{E}[\bar{e}_k \mid x_k] + 2V. $$

(8)
The last required intermediate step is the proposition below, where we present bounds related to the zeroth-order oracle that are essential to show our optimiser tracking result.

**Proposition 1** Under Assumption 1(a),(c), the following holds.

1. \( \|E[g_{\eta,k^+}(x,u)] - \nabla f_{k^+}(x)\| \leq \frac{L\eta}{2}(d+3)^{3/2} + \frac{\delta}{\eta}d^{1/2}. \)

2. \( E\left[\|g_{\eta,k^+}(x,u)\|^2\right] \leq \frac{L^2\eta^2}{2}(d+6)^3 + 2L\delta(d+4)^2 + \frac{2\delta^2}{\eta^2}d + 2(d+4)\|\nabla f_{k^+}(x)\|^2. \)

**Proof** See Appendix C.

Proposition 1 is the extension of Proposition 2.1 by [27] to the time-varying case. Note that our bounds depend on \( \delta \), which is the upper bound on the cost function variation by means of Assumption 1(c). We emphasise that Proposition 1 recovers the oracle bounds by [27] for \( \delta \to 0 \) (time-invariant case).

We are now in a position to state a general result that illustrates how the conditional expectation of the tracking error evolves in time for any given step size \( \alpha_k \). This is the main tool required to show that algorithm (2) can track the optimisers of (1).

**Theorem 1** Consider the iterates \( x_{k+1} = \mathcal{P}_X[\text{Exp}_{x_k}(-\alpha_k g_{\eta,k^+}(x_k,u_k))] \) with \( \alpha_k > 0 \) and \( g_{\eta,k^+} \) as per (3). Then, under Assumption 1(a)--(d) we have that, for all \( k \in \mathbb{N}_0 \),

\[
E[e_{k+1}|x_k] \leq \sqrt{\psi(e_k)} + 2V, \quad (9)
\]

where

\[
\psi(e_k) := (2(d+4)L^2\zeta(\kappa,e_k)\alpha_k^2 - \sigma\alpha_k + 1)e_k^2 + \alpha_k\left(L\eta(d+3)^{3/2} + \frac{2\delta}{\eta}d^{1/2}\right)e_k + \left(\frac{L^2\eta^2}{2}(d+6)^3 + 2L\delta(d+4)^2 + \frac{2\delta^2}{\eta^2}d\right)\zeta(\kappa,e_k)\alpha_k^2.
\]

**Proof** See Appendix D.

We can see that the conditional expectation of the tracking error depends on \( \psi(e_k) \) which in turn depends on the parameters of the problem such as the Lipschitz constant \( L \), manifold curvature \( \kappa \), step size \( \alpha_k \), oracle’s precision \( \eta \), and manifold dimension \( d \). For instance, we can see that \( \psi \) increases with \( L \), however, its dependence on other parameters such as the step size and oracle’s precision is not trivial. However, it turns out that if we pick a constant step size, we can obtain a simpler expression for the expected tracking error. Particularly, if we choose a constant step size in the interval \((0,\sigma/(2L^2(d+4)\zeta(\kappa,R)))\), we can show that the expected tracking error remains bounded for \( k \to \infty \), which is the main objective of this section. This is formalised in the below corollary.

**Corollary 1** Under Assumption 1, if \( \alpha_k = \alpha \in \left(0,\frac{\sigma}{2L^2(d+4)\zeta(\kappa,R)}\right) \) for all \( k \in \mathbb{N}_0 \), then

\[
\limsup_{k \to \infty} E[e_k] \leq \Delta := \frac{D + 2V}{1 - \rho}, \quad (10)
\]

where \( \rho := \sqrt{2(d+4)L^2\zeta(\kappa,R)\alpha^2 - \sigma\alpha + 1}, \quad D := \alpha \max\{\theta_1, \theta_2\} \), \( \theta_1 := \frac{L\eta(d+3)^{3/2} + (2/\eta)\delta d^{1/2}}{2\rho} \), and

\[
\theta_2 := \sqrt{\left(\frac{L^2\eta^2}{2}(d+6)^3 + 2L\delta(d+4)^2 + \frac{2\delta^2}{\eta^2}d\right)\zeta(\kappa,R)}.
\]
Proof See Appendix E.

Corollary 1 shows that the algorithm can closely follow the optimisers of the time-varying optimisation problem in (1) as \( k \) tends to infinity as long as we pick a constant step size \( \alpha_k = \alpha \) in algorithm (2) in the interval \( \alpha \in (0, \sigma/(2L^2(d + 4)\zeta(\kappa, R))) \). This is represented via the upper bound \( \Delta \), which we use as a performance metric of the algorithm.

Remark 1 It is worth noticing that \( \Delta \) depends on the manifold geometry through \( \kappa \), as opposed to the Euclidean counterpart by [36]. Moreover, \( \Delta \) depends on the intrinsic dimension \( d \) of the manifold, and not on the Euclidean ambient space dimension \( n \) which could be considerably larger. This is due to the fact that we work directly on the manifold and perform appropriate extensions of notions such as geodesic convexity, exponential maps, etc. Therefore, it would be more costly to work in the larger ambient Euclidean space, and our Riemannian method should be used preferably, unless a specific structure in the larger space can be considerably exploited to simplify calculations. The reported dependence of \( \Delta \) on the manifold curvature is consistent with recent literature. For instance, in Riemannian SVRG algorithms, the convergence rate of the algorithm depends on the manifold curvature as observed by [43], see also [44] for similar conclusions on subgradient methods.

In brief, we have shown that algorithm (2) can closely follow the optimisers of the time-varying optimisation problem (1) with a performance of \( \Delta \). We emphasise that this holds for any choice of constant step size in the interval \( \alpha \in (0, \sigma/(2L^2(d + 4)\zeta(\kappa, R))) \), and any choice of oracle’s precision \( \eta \). Then, it is important to ask whether we can find specific expressions for the step size \( \alpha \) and oracle’s precision \( \eta \) such that the performance metric \( \Delta \) is minimised. The following theorem provides an answer to this question by means of such expressions.

Theorem 2 Let \( \bar{\eta} := (4\delta^2 d/(L^2(d + 6)^3))^{1/4} \), and \( \bar{\alpha} \) be the root\(^1\) of \( A\alpha^2 + B\alpha + C = 0 \) in the interval \( (0, \frac{\sigma}{2L^2(d + 4)\zeta(\kappa, R)}) \) with

\[
A := (8V L^2 \zeta(\kappa, R)(d + 4) + \sigma \bar{\theta})^2 - 8\bar{\theta}^2 L^2 \zeta(\kappa, R)(d + 4), \\
B := -4V \left( \bar{\theta} \sigma^2 + 8V L^2 \zeta(\kappa, R)(d + 4)\sigma + 8\bar{\theta} L^2 \zeta(\kappa, R)(d + 4) \right), \\
C := (2\sigma V + 2\bar{\theta})^2 - 4\bar{\theta}^2, \\
\bar{\theta} := \sqrt{\left( \frac{L^2 \bar{\eta}^2}{2} (d + 6)^3 + 2L\delta(d + 4)^2 + \frac{2\delta^2}{\eta^2} d \right) \zeta(\kappa, R)}.
\]

Then, \( \bar{\eta} \) and \( \bar{\alpha} \) minimise \( \Delta \) in (10).

Proof See Appendix F.

As a summary of the results of this section, we essentially stated that, if we use algorithm (2) with choices of step size \( \alpha \) and oracle’s precision \( \eta \) as per Theorem 2, then the algorithm will track the optimisers of the time-varying optimisation problem (1) with performance \( \Delta \), which is in fact optimal for this choice of algorithm parameters.

3.2 Regret bounds

In this section, we provide dynamic regret bounds for our setting, for which we impose an extra assumption on the gradient.

\(^1\)We note that the choice of step size \( \bar{\alpha} \) in Theorem 2 always exists since \( \Delta|_{\eta=\bar{\eta}} \) is convex in \( \alpha \) over the interval \( (0, \frac{\sigma}{2L^2(d + 4)\zeta(\kappa, R)}) \).
Assumption 2 For all $f \in \mathcal{F}$ and $x \in \mathcal{X}$, $\exists G > 0$ such that $\|\nabla f(x)\| \leq G$.

Consider the following regret definitions, as counterparts to our tracking and estimation errors respectively,

\[
\text{Reg}_{T}^{\text{Track}} := \sum_{k=0}^{T} \mathbb{E}[f_{k+}(x_{k})] - f_{k+}(x_{k}^*) , \quad (11a)
\]

\[
\text{Reg}_{T}^{\text{Est.}} := \sum_{k=0}^{T} \mathbb{E}[f_{k+}(x_{k+1})] - f_{k+}(x_{k}^*) . \quad (11b)
\]

For these definitions of regret we present the following bounds.

Theorem 3 If the step-size $\alpha_k$ and oracle’s precision $\eta_k$ are chosen as

\[
0 < \alpha_k < \min \left\{ \sqrt{-B_k - (B_k^2 - 4AC_k)^{\frac{1}{2}}} \frac{\sigma}{2L^2 (d+4) \zeta (\kappa, R)} \right\} , \quad (12a)
\]

\[
0 < \sqrt{-B_k - (B_k^2 - 4AC_k)^{\frac{1}{2}}} \leq \eta_k \leq \sqrt{-B_k + (B_k^2 - 4AC_k)^{\frac{1}{2}}} , \quad (12b)
\]

where

\[
\bar{A} := 4L^2 \delta^2 (d+4) \zeta^2 - 4L^2 \sigma^2 d (d+6)^3 \zeta^2 , \quad \bar{B}_k := -\frac{4L^2 \delta (d+4) \zeta^2}{\bar{T}_k} , \quad \bar{C}_k := \frac{\bar{c}^4}{\bar{T}_k^2}
\]

\[
A := \frac{L^2 (d+6)^3 \zeta}{2} , \quad B_k := 2L \delta (d+4)^2 \zeta - \frac{\bar{c}^2}{\alpha_k^2 \bar{T}_k} , \quad C := 2 \delta^2 d \zeta,
\]

and $T_k = 2^m$ for $k \in [2^m - 1, 2^{m+1} - 2]$, $m \in \mathbb{N}_0$. Then, the tracking and estimation regrets satisfy, for any $T \geq 1$,

\[
\text{Reg}_{T}^{\text{Track}} \leq \frac{G}{1 - \max \{ \rho_0, \rho_T \}} \left( \mathbb{E}[\epsilon_0] - \rho_T \mathbb{E}[\epsilon_T] + \frac{\bar{c} \sqrt{T}}{1 - \sqrt{2}} \sqrt{T} + V_T \right) , \quad (13a)
\]

\[
\text{Reg}_{T}^{\text{Est.}} \leq \frac{G}{1 - \max \{ \rho_1, \rho_{T+1} \}} \left( \mathbb{E}[\epsilon_0] - \rho_{T+1} \mathbb{E}[\epsilon_T] + \frac{\bar{c} \sqrt{T}}{1 - \sqrt{2}} \sqrt{T} + \max \{ \rho_1, \rho_{T+1} \} V_T \right) , \quad (13b)
\]

where $\bar{c} > 0$, $V_T := \sum_{k=0}^{T-1} \text{dist}(x_{k+}, x_{k+1}^*)$, and $\rho_k := \sqrt{2(d+4)L^2 \zeta (\kappa, R) \alpha_k^2 - \sigma \alpha_k + 1}$.

Proof See Appendix G. ■

Theorem 3 provides a choice of step size $\alpha_k$ and oracle’s precision $\eta_k$ such that the regret satisfies (13). Note that $\alpha_k$ and $\eta_k$ are piecewise constant in periods of length $2^m$. For example, the periods $T_k$ have the form $T_0 = 1, T_1 = 2, T_2 = 2, T_3 = \cdots = T_6 = 4$, etc., and thus $(\alpha_1, \eta_1) = (\alpha_2, \eta_2), (\alpha_3, \eta_3) = \cdots = (\alpha_6, \eta_6)$, and so on.

We note that the regret bounds in (13) are sub-linear in $T$, disregarding $V_T$. This error term $V_T$ relates to the change in minimisers given the time-varying nature of the problem at hand. Obviously, if $V_T$ is sub-linear in $T$, we can see that the presented regret bounds would be sub-linear in $T$ as well.
4 Numerical example

To validate our results, we apply our zeroth-order algorithm to the problem of computing the Karcher mean of a collection of symmetric positive definite (SPD) matrices, also known as Riemannian centre of mass or Fréchet mean [6]. This problem appears in a number of applications such as medical imaging [20], image segmentation [34], signal estimation [26], and particle filtering [9]. Note that the Karcher mean is guaranteed to exist and be unique on a Hadamard manifold, see e.g. [4].

We consider a time-varying version of the Karcher mean problem which arises in online scenarios. For instance, we may want to find a central representative for a collection of online noisy measurements of a moving object. Formally, we consider that the measurements are $N$ SPD matrices of dimension $m \times m$ that become available at each time $k$, which we denote by $\{A_{k,1}, \ldots, A_{k,N}\}$.

The manifold of SPD matrices is defined as $\mathcal{M} := \{X \in \mathbb{R}^{m \times m} : X = X^T > 0\}$. If we equip $\mathcal{M}$ with the Riemannian metric

$$
\langle M, N \rangle_X := \text{Tr} \left\{ X^{-1}MX^{-1}N \right\}, \quad M, N \in T_X \mathcal{M},
$$

for every $X \in \mathcal{M}$, then the SPD manifold is a Hadamard manifold [3]. The Riemannian distance is given by

$$
dist(X, Y) := \left\| \log \left( X^{-1/2}YX^{-1/2} \right) \right\|_F,
$$

where $\|\cdot\|_F$ corresponds to the Euclidean (or Frobenius) norm, and the exponential mapping is

$$
\text{Exp}_X(M) = X^{1/2}\exp \left( X^{-1/2}MX^{-1/2} \right) X^{1/2}, \quad M \in T_X \mathcal{M},
$$

for every $X \in \mathcal{M}$, where $\exp$ denotes the matrix exponential. The time-varying cost function is defined as

$$
f_k(X) := \frac{1}{2N} \sum_{i=1}^{N} \text{dist}(X, A_{k,i})^2, \quad k \in \mathbb{N}_0.
$$

The Karcher mean for each set of measurements $\{A_{k,i}\}_{i=1}^{N}$ received at time $k$ is the unique minimiser of $f_k(X)$, i.e. $x_k^* := \arg \min_{X \in \mathcal{M}} f_k(X)$, for all $k \in \mathbb{N}_0$. The cost function (14) is known to be geodesically strongly convex with $\sigma = 1$ and geodesically $\zeta$-smooth, see e.g. [44] and [43]. We note that (14) corresponds to the case where $f_{k^+} = f_k$ in Assumption 1(c), i.e. $\delta \to 0$, and we use $\delta = 0.001$. We consider $N = 10$, three problem sizes $m \in \{3, 9, 100\}$, and the manifold dimension is $d = m(m + 1)/2$. For this example, we estimated $V = 0.5$ and $\zeta = 1.5$. The step size and oracle’s precision for each $m \in \{3, 9, 100\}$ are chosen as per Theorem 2, which gives $\bar{\sigma}_3 = 0.0074$, $\bar{\sigma}_9 = 0.0015$, and $\bar{\sigma}_{100} = 1.46 \cdot 10^{-5}$, and $\bar{\eta}_3 = 0.0089$, $\bar{\eta}_9 = 0.005$, and $\bar{\eta}_{100} = 5.13 \cdot 10^{-4}$. The matrices $\{A_{k,i}\}_{i=1}^{N}$ were randomly generated using the MANOPT toolbox in MATLAB, see [11].

Figure 1 depicts the average tracking error $\mathbb{E}[e_k]$ after implementing our zeroth-order algorithm for 100 random runs and different values of problem size. We can see that the algorithm (2) successfully tracks the optimisers up to an asymptotic error bound. That is, the average tracking errors converge to a ball which is upper bounded by some $\Delta$ for $k \to \infty$. The theoretical asymptotic bounds from Corollary 1 are $\Delta_3 = 543.73$, $\Delta_9 = 2666$, and $\Delta_{100} = 2750 \cdot 10^2$, for $m \in \{3, 9, 100\}$. Comparing with the asymptotic values in Figure 1, we can conclude that these bounds can become conservative depending on the application. Obtaining less conservative bounds is an open question for future research. It can also be seen that the larger the problem size, the longer $\mathbb{E}[e_k]$ takes to converge.
Figure 1: Algorithm’s average tracking error—over 100 random scenarios—for different values of problem size $m \in \{3, 9, 100\}$.

**Remark 2** In the context of our example, it is worth mentioning that for applications such as diffusion tensor imaging, the Euclidean averaging of SPD matrices often leads to a “swelling effect”, artificial extra diffusion introduced in computation, see e.g. [2]. Particularly, it means the determinant of the Euclidean mean can be strictly larger than the original determinants. In diffusion tensor imaging, diffusion tensors correspond to covariance matrices of the local Brownian motion of water molecules. Introducing more diffusion is physically unacceptable in this context. Therefore, Riemannian approaches such as the one presented in this paper would be preferred.

5 Conclusions

A gradient-free algorithm for the minimisation of time-varying cost functions on Hadamard manifolds was proposed. Bounds on the expectation of the tracking error and on dynamic regret were derived, and choices for algorithm parameters such that the asymptotic tracking error bound is minimised were provided. Finally, the theoretical results were validated via numerical experiments.

Future work includes the extension to a more general class of Riemannian manifolds by trying to relax the way we sample the random vector $u$ in (3), and also the extension to the non-smooth time-varying case. In addition, looking at the problem in a different set of coordinates at each step is also an interesting future direction. Lastly, exploring different gradient approximations in the oracle for this time-varying context is also of interest.

A Auxiliary lemmas

**Lemma 2** ([3]) Let $\mathcal{M}$ be a Hadamard manifold and $\mathcal{X} \subset \mathcal{M}$ a closed convex set. Then, the mapping $\mathcal{P}_\mathcal{X}(x) := \{ y \in \mathcal{X} : \text{dist}(x, z) = \inf_{z \in \mathcal{X}} \text{dist}(x, z) \}$ is single-valued and nonexpansive, that is, we have $\text{dist}(\mathcal{P}_\mathcal{X}(x), \mathcal{P}_\mathcal{X}(y)) \leq \text{dist}(x, y)$ for every $x, y \in \mathcal{M}$.

**Lemma 3** ([36]) Let $x, y, a, c \geq 0$ and $b \in \mathbb{R}$. Then $x^2 \leq ay^2 + by + c$ implies $x \leq y\sqrt{a} + D$, where $D := \max \left\{ \frac{b}{2\sqrt{a}}, \sqrt{c} \right\}$.

**Lemma 4** ([27]) Suppose $\mathcal{X}$ is a $d$-dimensional subspace of $\mathbb{R}^n$, with orthogonal projection matrix $P \in \mathbb{R}^{n \times n}$, $u_0 \sim \mathcal{N}(0, I_n)$, and $u = Pu_0$ is the orthogonal projection of $u_0$ onto $\mathcal{X}$. Then,
Figure 2: Illustration of the geodesic triangle used in Lemma 5.

(a) \( x = \frac{1}{\nu} \int_{\mathbb{R}^n} (x, u) u e^{-\frac{1}{2} \|u\|^2} du_0, \forall x \in \mathcal{X}. \)

(b) For \( p \in [0, 2] \), \( E[\|u\|^p] \leq d^p/2 \), and if \( p \geq 2 \), then \( E[\|u\|^p] \leq (d + p)^{p/2} \).

(c) \( E[\|\langle \text{grad} f_k(x), u \rangle\|^2] \leq (d + 4) \|\text{grad} f_k(x)\|^2. \)

Lemma 5 (44) If \( a, b, c \) are the sides (i.e. lengths) of a geodesic triangle in an Alexandrov space with curvature lower bounded by \( \kappa \), and \( A \) is the angle between sides \( b \) and \( c \), then

\[
a^2 \leq \frac{c^2}{\sqrt{|\kappa|}} b^2 + c^2 - 2bc \cos(A).
\]

B Proof of Lemma 1.

Let \( \tilde{x}_{k+1} := \text{Exp}_{x_k}( -\alpha_k g_{\eta,k+1}(x_k, u_k) ) \), and consider the geodesic triangle depicted in Figure 2 with vertices \( x^*_k \), \( x_k \), and \( \tilde{x}_{k+1} \), and sides \( a := \text{dist}(\tilde{x}_{k+1}, x^*_k) \), \( b := \text{dist}(x_k, \tilde{x}_{k+1}) \), and \( c := e_k = \text{dist}(x_k, x^*_k) = \|\text{Exp}_x^{-1}(x^*_k)\| \). For this triangle, we have that \( \text{dist}(x_k, \tilde{x}_{k+1}) = \|\text{Exp}_x^{-1}(\tilde{x}_{k+1})\| = \alpha_k \|g_{\eta,k+1}(x_k, u_k)\| \). In addition, we have that

\[
bc \cos(A) = \langle -\alpha_k g_{\eta,k+1}(x_k, u_k), \text{Exp}_x^{-1}(x^*_k) \rangle.
\]

Then, by Lemma 5,

\[
\text{dist}(\tilde{x}_{k+1}, x^*_k)^2 \leq e_k^2 + 2\alpha_k \langle g_{\eta,k+1}(x_k, u_k), \text{Exp}_x^{-1}(x^*_k) \rangle + \zeta(\kappa, e_k) \alpha_k^2 \|g_{\eta,k+1}(x_k, u_k)\|^2. \quad (15)
\]

Lastly, note that by Lemma 2, \( \text{dist}(\tilde{x}_{k+1}, x^*_k)^2 \geq \text{dist}(x_{k+1}, x^*_k)^2 = e_k^2 \), and thus the result follows immediately from (15).
C Proof of Proposition 1.

(a) We complete the proof with the following steps. Essentially we want to quantify how well the expectation of the oracle approximates the real gradient.

\[
\|\mathbb{E} [g_{\eta,k}^+(x,u)] - \text{grad} f_k^+(x)\|
\]

Lemma 4(a)
\[
\leq \frac{1}{\nu} \int_{\mathbb{R}^n} \left( \frac{f_k^+(\text{Exp}_x(\eta u)) - f_k^+(x)}{\eta} \right.
\]
\[+ \langle \text{grad} f_k^+(x), u \rangle \right) u e^{-\frac{1}{2}\|u\|^2} du_0
\]
\[
\leq \frac{1}{\eta \nu} \int_{\mathbb{R}^n} |f_k^+(\text{Exp}_x(\eta u)) - f_k^+(x) - \langle \text{grad} f_k^+(x), \eta u \rangle |
\]
\[+ f_k^+(x) - f_k^+(x) \|u\| e^{-\frac{1}{2}\|u\|^2} du_0
\]
Assumption 1(a),(c)
\[
\leq \frac{L \eta^2}{2} \frac{\|u\|^2 + \delta}{\nu^2} \|u\| e^{-\frac{1}{2}\|u\|^2} du_0
\]
\[= \frac{L \eta^2}{2} \int_{\mathbb{R}^n} \|u\|^2 e^{-\frac{1}{2}\|u\|^2} du_0 + \delta \frac{\eta}{\nu^2} \int_{\mathbb{R}^n} \|u\| e^{-\frac{1}{2}\|u\|^2} du_0
\]
Lemma 4(b)
\[
\leq \frac{L \eta}{2} (d + 3)^{3/2} + \frac{\delta}{\eta} q^{1/2}.
\]

(b) To compute a bound on \( \mathbb{E} \left[ \|g_{\eta,k}^+(x,u)\|^2 \right] \), we proceed by definition of \( g_{\eta,k}^+(x,u) \), and thus first we bound \( f_k^+(\text{Exp}_x(\eta u)) - f_k^+(x) \). We start by adding a convenient zero, that is, we add and subtract both \( f_k^+(x) \) and \( \langle \text{grad} f_k^+(x), \eta u \rangle \), and then we use Assumptions 1(a) and 1(c). That is,
\[
(f_k^+(\text{Exp}_x(\eta u)) - f_k^+(x))^2 = (f_k^+(\text{Exp}_x(\eta u)) - f_k^+(x) - \langle \text{grad} f_k^+(x), \eta u \rangle)
\]
\[+ f_k^+(x) - f_k^+(x) \|u\| e^{-\frac{1}{2}\|u\|^2} du_0
\]
\[
\leq \left( \frac{L \eta^2}{2} \|u\|^2 + \delta + \langle \text{grad} f_k^+(x), \eta u \rangle \right)^2
\]
\[
\leq 2 \left( \frac{L \eta^2}{2} \|u\|^2 + \delta \right)^2 + 2\eta \langle \text{grad} f_k^+(x), u \rangle^2.
\]
Consequently,
\[
\mathbb{E} \left[ \|g_{\eta,k}^+(x,u)\|^2 \right] = \frac{(f_k^+(\text{Exp}_x(\eta u)) - f_k^+(x))^2}{\eta^2} \mathbb{E} \left[ \|u\|^2 \right]
\]
\[
\leq \frac{L^2 \eta^2}{2} \mathbb{E} \left[ \|u\|^6 \right] + 2L \delta \mathbb{E} \left[ \|u\|^4 \right] + \frac{2\delta^2}{\eta^2} \mathbb{E} \left[ \|u\|^2 \right]
\]
\[+ 2 \mathbb{E} \left[ \|\langle \text{grad} f_k^+(x), u \rangle u \| 2 \right].
\]
The proof is thus complete by means of Lemma 4(b),(c).

D Proof of Theorem 1.

In this proof our main goal is to obtain (9). To achieve this, we note that (8) gives us a relation between \( \mathbb{E} [e_{k+1} \mid x_k] \) and \( \mathbb{E} [\bar{e}_k \mid x_k] \). Therefore, we first compute the following. We take conditional
expectations in (7) and obtain

\[
\mathbb{E} \left[ \bar{e}_k^2 \mid x_k \right] \leq e_k^2 + 2\alpha_k \left( \mathbb{E} \left[ g_{\eta,k}(x_k, u_k) \mid x_k \right], \text{Exp}^{-1}_k(x_{k+}^*) \right) \\
+ \zeta(\kappa, e_k)\alpha_k^2 \mathbb{E} \left[ \left\| g_{\eta,k}(x_k, u_k) \right\|^2 \mid x_k \right] \\
\leq e_k^2 + 2\alpha_k \left( \mathbb{E} \left[ g_{\eta,k}(x_k, u_k) \right], \text{Exp}^{-1}_k(x_{k+}^*) \right) \\
+ \zeta(\kappa, e_k)\alpha_k^2 \left( \frac{L^2 \eta^2}{2}(d + 6)^3 + 2L\delta(d + 4)^2 \\
+ \frac{2\delta^2}{\eta^2}d + 2(d + 4)\left\| \text{grad} f_k(x_k) \right\|^2 \right), \\
\tag{16}
\]

where the last inequality follows from Proposition 1(b). Below, we focus on computing the term \( \langle \mathbb{E} \left[ g_{\eta,k}(x_k, u_k) \right], \text{Exp}^{-1}_k(x_{k+}^*) \rangle \) in (16). We add a convenient zero, in this case we add and subtract \( \text{grad} f_k(x_k) \)

\[
\langle \mathbb{E} \left[ g_{\eta,k}(x_k, u_k) \right], \text{Exp}^{-1}_k(x_{k+}^*) \rangle = \langle \mathbb{E} \left[ g_{\eta,k}(x_k, u_k) \right] - \text{grad} f_k(x_k), \text{Exp}^{-1}_k(x_{k+}^*) \rangle \\
+ \langle \text{grad} f_k(x_k), \text{Exp}^{-1}_k(x_{k+}^*) \rangle.
\]

Then, by the Cauchy-Scharwz inequality and (6),

\[
\langle \mathbb{E} \left[ g_{\eta,k}(x_k, u_k) \right], \text{Exp}^{-1}_k(x_{k+}^*) \rangle \\
\leq \left\| \mathbb{E} \left[ g_{\eta,k}(x_k, u_k) \right] - \text{grad} f_k(x_k) \right\| \|\text{Exp}^{-1}_k(x_{k+}^*)\| - \frac{\sigma}{2}\text{dist}(x_k, x_{k+}^+)^2 \\
= \left\| \mathbb{E} \left[ g_{\eta,k}(x_k, u_k) \right] - \text{grad} f_k(x_k) \right\| e_k - \frac{\sigma}{2}e_k^2 \\
\leq \left( \frac{L\eta}{2}(d + 3)^{3/2} + \frac{\delta}{\eta}d^{1/2} \right) e_k - \frac{\sigma}{2}e_k^2,
\]

where the last inequality follows from Proposition 1(a).

There is now only one term in (16) that remains to be bounded, which is \( \| \text{grad} f_k(x_k) \|^2 \). To that end, note that from (4) and the reverse triangle inequality,

\[
\| \| \text{grad} f(x) \| - \| \Gamma^c \| \text{grad} f(y) \| \| \leq L\text{dist}(x, y),
\]

which, in turn, implies that \( \| \text{grad} f_k(x_k) \| \leq Le_k \). Therefore, by using the latter together with (17) into (16), we obtain

\[
\mathbb{E} \left[ \bar{e}_k^2 \mid x_k \right] \leq e_k^2 + 2\alpha_k \left( \left( \frac{L\eta}{2}(d + 3)^{3/2} + \frac{\delta}{\eta}d^{1/2} \right) e_k - \frac{\sigma}{2}e_k^2 \right) \\
+ \zeta(\kappa, e_k)\alpha_k^2 \left( \frac{L^2 \eta^2}{2}(d + 6)^3 + 2L\delta(d + 4)^2 + \frac{2\delta^2}{\eta^2}d + 2(d + 4)L^2e_k^2 \right) \\
= \left( 2(d + 4)L^2 \zeta(\kappa, e_k)\alpha_k^2 - \sigma\alpha_k + 1 \right) e_k^2 \\
+ \alpha_k \left( L\eta(d + 3)^{3/2} + \frac{2\delta}{\eta}d^{1/2} \right) e_k \\
+ \left( \frac{L^2 \eta^2}{2}(d + 6)^3 + 2L\delta(d + 4)^2 + \frac{2\delta^2}{\eta^2}d \right) \zeta(\kappa, e_k)\alpha_k^2.
\]

By Jensen’s inequality we get \( \mathbb{E} \left[ \bar{e}_k \mid x_k \right]^2 \leq \mathbb{E} \left[ \bar{e}_k^2 \mid x_k \right] \), and the proof is thus complete from applying (8).
E  Proof of Corollary 1.

The first part of the proof consists in simplifying the expression for $\psi(e_k)$ that comes from Theorem 1 given the choice of constant step-size and also the bound on $\zeta$. Note that Assumption 1(e) implies that $\zeta(\kappa, e_k) \leq \zeta(\kappa, R)$. Then, for $\alpha_k = \alpha$,

$$
\psi(e_k) = \left(2(d + 4)L^2\zeta(\kappa, R)\alpha^2 - \sigma\alpha + 1\right)e_k^2
+ \alpha \left(L\eta(d + 3)^{3/2} + \frac{2\delta}{\eta}d^{1/2}\right)e_k
+ \left(\frac{L^2\eta^2}{2}(d + 6)^3 + 2L\delta(d + 4)^2 + \frac{2\delta^2}{\eta^2}d\right)\zeta(\kappa, R)\alpha^2.
$$

The second part of the proof boils down to using the above simplified expression for $\psi(e_k)$ to obtain a recursive expression for $\mathbb{E}[e_k]$ so that we can iterate it and compute (10). Recall from the proof of Theorem 1 that $\mathbb{E}[\bar{e}_k | x_k]^2 \leq \psi(e_k) = ae_k^2 + be_k + c$. Therefore, Lemma 3 implies $\mathbb{E}[\bar{e}_k | x_k] \leq \rho e_k + D$. We then use the latter inequality in (8) to get $\mathbb{E}[e_{k+1} | x_k] \leq \rho e_k + D + 2V$. Applying expectation then leads to the following recursive equation for $\mathbb{E}[e_k]$,

$$
\mathbb{E}[e_{k+1}] \leq \rho\mathbb{E}[e_k] + D + 2V. \quad (18)
$$

Since $\rho < 1$ for $0 < \alpha < \frac{\sigma}{2L^2(d + 4)\zeta(\kappa, R)}$, we can iterate (18) and obtain (10) in the limit $k \to \infty$, concluding the proof.

F  Proof of Theorem 2.

The first part of the proof consists in showing that $\theta_2 > \theta_1$, which will allow us to write $D = \alpha \max\{\theta_1, \theta_2\} = \alpha \theta_2$. Note that $\rho^2 = 2(d + 4)L^2\zeta(\kappa, R)\alpha^2 - \sigma\alpha + 1 > \frac{L^2}{2}\alpha^2 - \sigma\alpha + 1 \geq \frac{\alpha^2}{2}\alpha^2 - \sigma\alpha + 1$, where the last inequality follows from Assumption 1(a). Then, $2\rho^2 > (\sigma\alpha)^2 - 2\sigma\alpha + 2 = (\sigma\alpha - 1)^2 + 1 \geq 1$, which implies that $2\rho > \sqrt{2}$. Therefore,

$$
\theta_1 = \frac{L\eta(d + 3)^{3/2} + (2/\eta)\delta d^{1/2}}{2\rho} < \frac{L\eta(d + 3)^{3/2} + (2/\eta)\delta d^{1/2}}{\sqrt{2}},
$$

and thus

$$
\theta_1^2 < \frac{L^2\eta^2}{2}(d + 3)^3 + 2L\delta d^{1/2}(d + 3)^{3/2} + \frac{2\delta^2}{\eta^2}d. \quad (19)
$$

On the other hand,

$$
\theta_2^2 = \frac{L^2\eta^2}{2}\zeta(\kappa, R)(d + 6)^3 + 2L\delta\zeta(\kappa, R)(d + 4)^2 + \frac{2\delta^2}{\eta^2}\zeta(\kappa, R)d,
$$
and note that, by definition, \( \zeta(\kappa, R) \geq 1 \) for all \( \kappa \) and \( R \). We also note that \((d + 4)^2 > d^{1/2}(d + 3)^{3/2}\) for all \( d \geq 0 \). Hence,

\[
\theta_2^2 \geq \frac{L^2 \eta^2}{2} (d + 6)^3 + 2L \delta (d + 4)^2 + \frac{2 \delta^2}{\eta^2} d
\]

\[
> \frac{L^2 \eta^2}{2} (d + 3)^3 + 2L \delta d^{1/2}(d + 3)^{3/2} + \frac{2 \delta^2}{\eta^2} d
\]

\[
(19) \quad \theta_1^2,
\]

which implies that \( D = \alpha \max\{\theta_1, \theta_2\} = \alpha \theta_2 \), and thus \( \Delta = \frac{\alpha \theta_2 + 2V}{1 - \rho} \).

Now that we have an expression for \( \Delta \), the second part of the proof consists in computing \( \alpha \) and \( \eta \) that minimise \( \Delta \). We note that \( \Delta \) depends on \( \eta \) through \( \theta_2 \) only, and it depends on \( \alpha \) through \( \rho \) only. We can see that

\[
\frac{\partial \theta_2}{\partial \eta} = 0 \implies \zeta(\kappa, R) \left( \frac{L^2 (d + 6)^3 \eta - \frac{4 \delta^2 d}{\eta^2}}{1 - \sqrt{2(d + 4) L^2 \zeta(\kappa, R) \alpha^2 - \sigma \alpha + 1}} \right) = 0,
\]

from which we conclude that \( \eta = \left(4 \delta^2 d/(L^2 (d + 6)^3)\right)^{1/4} \) minimises \( \theta_2 \), and since the denominator of \( \Delta \) is independent of \( \eta \), \( \Delta \) attains its minimum at \( \eta \).

Next, we obtain \( \bar{\alpha} \) that minimises \( \Delta \). To that end, we compute the derivative of \( \Delta \) evaluated at \( \bar{\eta} \) with respect to \( \alpha \) and set it to be equal to zero. Note that

\[
\frac{\partial \Delta}{\partial \alpha} \bigg|_{\eta=\bar{\eta}} = \frac{\partial}{\partial \alpha} \left( \frac{\alpha \bar{\theta} + 2V}{1 - \sqrt{2(d + 4) L^2 \zeta(\kappa, R) \alpha^2 - \sigma \alpha + 1}} \right) = 0
\]

implies

\[
2 \bar{\theta} \sqrt{2L^2(d + 4) \zeta(\kappa, R) \alpha^2 - \sigma \alpha + 1} + 8L^2 V \zeta(\kappa, R)(d + 4) \alpha + \bar{\sigma} \bar{\alpha} - 2\sigma V - 2\bar{\theta} = 0,
\]

and then

\[
\left( (8L^2 V \zeta(\kappa, R)(d + 4) + \bar{\sigma} \bar{\alpha}) \alpha - (2\sigma V + 2\bar{\theta}) \right)^2 = 4 \bar{\theta}^2 \left( 2L^2(d + 4) \zeta(\kappa, R) \alpha^2 - \sigma \alpha + 1 \right).
\]

Lastly, we simply group terms in (20) to write \( A\alpha^2 + B\alpha + C = 0 \) with \( A, B \) and \( C \) as per the theorem statement, completing the proof.

**G Proof of Theorem 3.**

To prove Theorem 3 we need the following two intermediate lemmas.

**Lemma 6** Let \( \rho_k := \sqrt{2(d + 4)L^2 \zeta(\kappa, R) \alpha_k^2 - \sigma \alpha_k + 1} \), and suppose \( \alpha_k \in (0, \sigma/(2L^2(d+4) \zeta(\kappa, R))) \) for all \( k \in \mathbb{N}_0 \). Then,

\[
\sum_{k=0}^{T-1} \mathbb{E}[e_k] \leq \frac{1}{1 - \max\{\rho_0, \rho_T\}} \left( \mathbb{E}[e_0] - \rho_T \mathbb{E}[e_T] + \sum_{k=0}^{T-1} D_k + V_T \right),
\]

where \( V_T := \sum_{k=0}^{T-1} \text{dist}(x_{k+1}^*, x_{k+1}^*) \), and \( D_k := \alpha_k \sqrt{\left( \frac{L^2 \eta_k^2}{2} (d + 6)^3 + 2L \delta (d + 4)^2 + \frac{2 \delta^2}{\eta_k^2} \right) \zeta} \).
Proof. We know that $e_{k+1} \leq \bar{e}_k + \text{dist}(x^*_k, x^*_{k+1})$, and thus $\mathbb{E}[e_{k+1} \mid x_k] \leq \mathbb{E}[\bar{e}_k \mid x_k] + \text{dist}(x^*_k, x^*_{k+1})$. Note that this is a relaxed version of (8) since we do not use the change on minimiser bound from Assumption 1(b). Therefore, from the proof of Corollary 1, it is easy to see that for any $\alpha_k > 0$,

$$\mathbb{E}[e_{k+1}] \leq \rho_k \mathbb{E}[e_k] + D_k + \text{dist}(x^*_k, x^*_{k+1}).$$

(21)

Summing both sides of (21) over and adding $\mathbb{E}[\epsilon_0]$ to both sides,

$$\sum_{k=0}^{T} \mathbb{E}[e_k] \leq \mathbb{E}[\epsilon_0] + \sum_{k=0}^{T} \rho_k \mathbb{E}[e_{k-1}] + \sum_{k=1}^{T} D_{k-1} + \sum_{k=1}^{T} \text{dist}(x^*_{k-1}, x^*_k)$$

$$= \mathbb{E}[\epsilon_0] + \sum_{k=0}^{T-1} \rho_k \mathbb{E}[e_k] + \sum_{k=0}^{T-1} D_k + \sum_{k=0}^{T-1} \text{dist}(x^*_{k}, x^*_{k+1})$$

$$= \mathbb{E}[\epsilon_0] - \rho_T \mathbb{E}[\epsilon_T] + \sum_{k=0}^{T-1} \rho_k \mathbb{E}[e_k] + \sum_{k=0}^{T-1} D_k + \sum_{k=0}^{T-1} \text{dist}(x^*_{k+1}, x^*_{k+1})$$

$$\leq \mathbb{E}[\epsilon_0] - \rho_T \mathbb{E}[\epsilon_T] + \max\{\rho_0, \rho_T\} \sum_{k=0}^{T} \mathbb{E}[e_k] + \sum_{k=0}^{T} D_k + \sum_{k=0}^{T} \text{dist}(x^*_{k+1}, x^*_{k+1}),$$

where the last inequality follows from the definition of $\rho_k$, and the proof is complete by noting that $\max\{\rho_0, \rho_T\} < 1$ since $\alpha_k \in (0, \sigma/(2L^2(d+4)\zeta(\kappa, R)))$.

Lemma 7. Let $\rho_k := \sqrt{2(d+4)L^2\zeta(\kappa, R)}\alpha_k^2 - \sigma\alpha_k + 1$, and suppose $\alpha_k \in (0, \sigma/(2L^2(d+4)\zeta(\kappa, R)))$ for all $k \in \mathbb{N}_0$. Then,

$$\sum_{k=0}^{T} \mathbb{E}[\epsilon_k] \leq \frac{1}{1 - \max\{\rho_1, \rho_{T+1}\}} \left( \mathbb{E}[\epsilon_0] - \rho_{T+1} \mathbb{E}[\epsilon_T] + \sum_{k=1}^{T} D_k + \max\{\rho_1, \rho_{T+1}\} V_T \right),$$

where $V_T := \sum_{k=0}^{T-1} \text{dist}(x^*_{k+1}, x^*_{k+1})$.

Proof. From the proof of Corollary 1, we can conclude that $\mathbb{E}[\epsilon_k] \leq \rho_k \mathbb{E}[e_k] + D_k$. Then, by the triangle inequality of the Riemannian distance, we can write $e_k = \text{dist}(x_k, x^*_k) \leq \text{dist}(x_k, x^*_{k-1}) + \text{dist}(x^*_{k-1}, x^*_{k})$. Therefore,

$$\mathbb{E}[\epsilon_k] \leq \rho_k \mathbb{E}[\epsilon_{k-1}] + \rho_k \text{dist}(x^*_{k-1}, x^*_{k}) + D_k.$$
We proceed similarly to the proof of Lemma 6, that is,

\[
\sum_{k=0}^{T} \mathbb{E} [\tilde{e}_k] \leq \mathbb{E} [\tilde{e}_0] + \sum_{k=1}^{T} \rho_k \mathbb{E} [\tilde{e}_{k-1}] + \sum_{k=1}^{T-1} \rho_k \text{dist}(x^*_{k+1}, x^*_{k-1}) + \sum_{k=1}^{T} D_k
\]

\[
\leq \mathbb{E} [\tilde{e}_0] + \sum_{k=0}^{T-1} \rho_{k+1} \mathbb{E} [\tilde{e}_k] + \sum_{k=0}^{T} \rho_{k+1} \text{dist}(x^*_{k+1}, x^*_{k+1}) + \sum_{k=1}^{T} D_k
\]

\[
= \mathbb{E} [\tilde{e}_0] - \rho_{T+1} \mathbb{E} [\tilde{e}_T] + \sum_{k=0}^{T-1} \rho_{k+1} \mathbb{E} [\tilde{e}_k] + \sum_{k=0}^{T} \rho_{k+1} \text{dist}(x^*_{k+1}, x^*_{k+1}) + \sum_{k=1}^{T} D_k
\]

\[
\leq \mathbb{E} [\tilde{e}_0] - \rho_{T+1} \mathbb{E} [\tilde{e}_T] + \max\{\rho_1, \rho_{T+1}\} \sum_{k=0}^{T-1} \mathbb{E} [\tilde{e}_k]
\]

\[
+ \max\{\rho_1, \rho_{T+1}\} \sum_{k=0}^{T-1} \text{dist}(x^*_{k+1}, x^*_{k+1}) + \sum_{k=1}^{T} D_k,
\]

completing the proof. \hfill \blacksquare

Now we can proceed with the proof of Theorem 3. Note that the geodesic strong convexity of \( f_k^+ \), the Cauchy-Schwarz inequality, and Assumption 1 imply

\[
f_k^+(x_k) - f_k^+(x^*_k) \leq - \langle \text{grad} f_k^+(x_k), \text{Exp}^{-1}(x^*_k) \rangle \leq \| \text{grad} f_k^+(x_k) \| e_k \leq G e_k
\]

\[
\mathbb{E} [f_k^+(x_k)] - f_k^+(x^*_k) \leq G \mathbb{E} [e_k].
\]

Similarly, it is immediate to show that \( \mathbb{E} [f_k^+(x^*_k)] - f_k^+(x^*_k) \leq G \mathbb{E} [\tilde{e}_k] \). Therefore, by means of Lemmas 6, 7, and the regret definitions in (11), we can immediately obtain the following upper bounds on \( \text{Reg}^\text{Track}_T \) and \( \text{Reg}^\text{Est}_T \),

\[
\text{Reg}^\text{Track}_T \leq \frac{G}{1 - \max\{\rho_0, \rho_T\}} \left( \mathbb{E} [\tilde{e}_0] - \rho_{T+1} \mathbb{E} [\tilde{e}_T] + \sum_{k=0}^{T-1} D_k + V_T \right),
\]

\[
\text{Reg}^\text{Est}_T \leq \frac{G}{1 - \max\{\rho_1, \rho_{T+1}\}} \left( \mathbb{E} [\tilde{e}_0] - \rho_{T+1} \mathbb{E} [\tilde{e}_T] + \sum_{k=1}^{T} D_k + \max\{\rho_1, \rho_{T+1}\} V_T \right).
\]

The next step of the proof follows by using the so-called doubling-trick from [35, Section 2.3.1], which divides the algorithm rounds into periods of increasing size, specifically, in periods of \( 2^m \) rounds, \( m \in \mathbb{N}_0 \). Formally, this allows us to write, for any \( T \geq 1 \),

\[
\sum_{k=0}^{T-1} D_k \leq \sum_{m=0}^{\lceil \log_2(T) \rceil} \sum_{k=2^m-1}^{2^m+1-2} D_k.
\]

Then, if for each period of \( 2^m \) rounds we can find a step size and oracle’s precision such that
\( D_k \leq \frac{c}{\sqrt{2m}} \) for some \( c > 0 \), \( k \in [2^m - 1, 2^{m+1} - 2] \), \( m \in \mathbb{N}_0 \), then we would have

\[
\sum_{k=0}^{T-1} D_k \leq \sum_{m=0}^{[\log_2(T)]} \sum_{k=2^{m+1}-2}^{2^{m+1}-2} D_k \leq \sum_{m=0}^{[\log_2(T)]} \frac{\bar{c}\sqrt{2m}}{1-\sqrt{2}} = \frac{\bar{c}(1 - (\sqrt{2})^{[\log_2(T)]+1}}{1-\sqrt{2}} \leq \frac{\bar{c}\sqrt{2}}{1-\sqrt{2}} \leq \frac{\bar{c}\sqrt{T}}{1-\sqrt{2}},
\]

which would prove the regret bound for \( \text{Reg}_T^{\text{Track}} \) in (13a). We can also bound \( \sum_{k=1}^{T} D_k \) exactly as above to prove \( \text{Reg}_T^{\text{Est}} \) in (13b). Therefore, to conclude the proof, we have to show that the choices of \( \alpha_k \) and \( \eta_k \) in the theorem statement indeed imply \( D_k \leq \frac{c}{\sqrt{2m}} \) in each period of length \( 2^m \), which is what we do below.

Fix a period of length \( 2^m \), that is, consider \( T_k = 2^m \) for \( k \in [2^m - 1, 2^{m+1} - 2] \), \( m \in \mathbb{N}_0 \). Recall that \( D_k := \alpha_k \sqrt{\left( (L^2\eta_k^2/(2d+6)^3 + 2L\delta(d+4)^2 + \frac{24^2}{\eta_k^2} d \right)} \). Then, in each period, \( D_k \leq \frac{c}{\sqrt{T_k}} \) implies

\[
\frac{L^2(d+6)^3 \zeta^4}{2} \eta_k^4 + \left( 2L\delta(d+4)^2 \zeta - \frac{\bar{c}^2}{\alpha_k^2 T_k} \right) \eta_k^2 + 2\delta^2 d \zeta \leq 0,
\]

which is a quadratic inequality on \( \eta_k^2 \). The solutions to \( \mathcal{A}\eta_k^4 + B_k\eta_k^2 + C = 0 \) are given by

\[
\begin{align*}
x_1 &= -B_k + (B_k^2 - 4AC)^{\frac{1}{2}}, \\
x_2 &= -B_k - (B_k^2 - 4AC)^{\frac{1}{2}}.
\end{align*}
\]

Since \( \mathcal{A} \geq 0 \) and \( C > 0 \), a necessary condition such that \( x_1, x_2 > 0 \) is that \( B_k < 0 \), which implies that

\[
\alpha_k < \sqrt{\frac{\bar{c}^2}{2L\delta(d+4)^2 \zeta (\kappa, R) T_k}}.
\]

We also need that \( \Delta_k := B_k^2 - 4AC \geq 0 \), which imposes an extra condition on the step size \( \alpha_k \), that is,

\[
0 \leq B_k^2 - 4AC = \left( 2L\delta(d+4)^2 \zeta - \frac{\bar{c}^2}{\alpha_k^2 T_k} \right)^2 - 4L^2\delta^2 d(d+6)^3 \zeta^2
\]

\[
= 4L^2\delta^2 (d+4)^4 \zeta^2 - 4L^2\delta^2 d(d+6)^3 \zeta^2 - \frac{4L\delta(d+4)^2 \zeta \bar{c}^2}{\alpha_k^2 T_k} + \frac{\bar{c}^4}{\alpha_k^4 T_k^2} \]

which can be written as a quadratic equation on \( \alpha_k^2 \),

\[
0 \leq \left[ \frac{4L^2\delta^2 (d+4)^4 \zeta^2 - 4L^2\delta^2 d(d+6)^3 \zeta^2}{\mathcal{A}} \right] \alpha_k^4 - \frac{4L\delta(d+4)^2 \zeta \bar{c}^2}{T_k} \alpha_k^2 + \frac{\bar{c}^4}{T_k^2}. \quad (23)
\]

19
The solutions to \( \bar{A}\alpha_k^4 + \bar{B}\alpha_k^2 + \bar{C} = 0 \) are

\[
\begin{align*}
y_1 &= \frac{-\bar{B}_k + (\bar{B}_k^2 - 4\bar{A}\bar{C}_k)^{\frac{1}{2}}}{2\bar{A}}, \\
y_2 &= \frac{-\bar{B}_k - (\bar{B}_k^2 - 4\bar{A}\bar{C}_k)^{\frac{1}{2}}}{2\bar{A}},
\end{align*}
\]

We analyse the determinant \( \Delta_k := \bar{B}_k^2 - 4\bar{A}\bar{C}_k \), that is,

\[
\begin{align*}
\Delta_k &= \frac{16L^2\delta^2(d+4)^4\zeta^2c^4}{T_k^2} - 4 \left[ 4L^2\delta^2(d+4)^4\zeta^2 - 4L^2\delta^2d(d+6)^3\zeta^2 \right] \frac{c^4}{T_k^2} \\
&= \frac{16L^2\delta^2d(d+6)^3\zeta^2c^4}{T_k^2},
\end{align*}
\]

which is always non-negative. Additionally, note that \( \bar{A} \geq 0 \) for \( d < 4 \), and \( \bar{A} \leq 0 \) for \( d \geq 5 \). We will use this fact to write a closed-form choice for the step size \( \alpha_k \) in each period \( T_k \) of length \( 2^m \).

- For \( d < 4 \) (\( \bar{A} \geq 0 \)), we note that \( \bar{B}_k^2 - 4\bar{A}\bar{C}_k \leq \bar{B}_k^2 \) since \( \bar{C}_k > 0 \). This fact together with \( \bar{B}_k \leq 0 \) implies that \( y_1 \geq y_2 \geq 0 \). Then, from (23) \( \alpha_k \) needs to satisfy

\[
\bar{A}(\alpha_k^2 - y_1)(\alpha_k^2 - y_2) \geq 0
\]

which holds for \( \alpha_k \geq \sqrt{y_1} \) or \( 0 < \alpha_k \leq \sqrt{y_2} \).

- For \( d \geq 5 \) (\( \bar{A} \leq 0 \)), we note that \( \bar{B}_k^2 - 4\bar{A}\bar{C}_k \geq \bar{B}_k^2 \) since \( \bar{C}_k > 0 \). Moreover, since \( \bar{B}_k \leq 0 \), we have that \( y_1 \leq 0 \) and \( y_2 \geq 0 \). Therefore, from (23), \( \bar{A}(\alpha_k^2 - y_2)(\alpha_k^2 + |y_1|) \geq 0 \) which satisfied with \( 0 < \alpha_k \leq \sqrt{y_2} \).

In addition, we need that \( \alpha_k < \frac{\sigma}{2L^2(d+4)\zeta(\kappa, R)} \) (to make \( \rho_k < 1 \)). Therefore, we choose

\[
\alpha_k < \min \left\{ \sqrt{y_2}, \frac{\sigma}{2L^2(d+4)\zeta(\kappa, R)} \right\}.
\]

Since this choice of \( \alpha_k \) implies that \( \Delta_k \geq 0 \), then \( \bar{A}\eta_k^4 + \bar{B}\eta_k^2 + \bar{C} = 0 \) has two distinct positive real solutions \( x_1 \) and \( x_2 \). Then, from (22), we recall that \( \eta_k \) needs to be chosen such that

\[
\bar{A}\eta_k^4 + \bar{B}\eta_k^2 + \bar{C} \leq 0 \iff \bar{A}(\eta_k^2 - x_1)(\eta_k^2 - x_2) \leq 0.
\]

Since \( \bar{A} \geq 0 \) and \( \bar{C} > 0 \), then \( \bar{B}_k^2 - 4\bar{A}\bar{C} \leq \bar{B}_k^2 \) and \( 0 \leq x_2 \leq x_1 \). Therefore, the choice of \( \eta_k \) needs to satisfy

\[
\frac{-\bar{B}_k - (\bar{B}_k^2 - 4\bar{A}\bar{C})^{\frac{3}{2}}}{2\bar{A}} \leq \eta_k \leq \frac{-\bar{B}_k + (\bar{B}_k^2 - 4\bar{A}\bar{C})^{\frac{3}{2}}}{2\bar{A}},
\]

concluding the proof.
References

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

[2] Vincent Arsigny, Pierre Fillard, Xavier Pennec, and Nicholas Ayache. Geometric means in a novel vector space structure on symmetric positive-definite matrices. *SIAM journal on matrix analysis and applications*, 29(1):328–347, 2007.

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

[4] Marcel Berger. *A panoramic view of Riemannian geometry*. Springer Science & Business Media, 2012.

[5] Omar Besbes, Yonatan Gur, and Assaf Zeevi. Non-stationary stochastic optimization. *Operations research*, 63(5):1227–1244, 2015.

[6] Dario A Bini and Bruno Iannazzo. Computing the Karcher mean of symmetric positive definite matrices. *Linear Algebra and its Applications*, 438(4):1700–1710, 2013.

[7] Richard L Bishop and Barrett O’Neill. Manifolds of negative curvature. *Transactions of the American Mathematical Society*, 145:1–49, 1969.

[8] Silvere Bonnabel. Stochastic gradient descent on Riemannian manifolds. *IEEE Transactions on Automatic Control*, 58(9):2217–2229, 2013.

[9] Claudio J Bordin Jr, Caio G de Figueredo, and Marcelo GS Bruno. Nonlinear state estimation on unit spheres using manifold particle filtering. *Digital Signal Processing*, 81:50–56, 2018.

[10] Nicolas Boumal, Pierre-Antoine Absil, and Coralia Cartis. Global rates of convergence for nonconvex optimization on manifolds. *IMA Journal of Numerical Analysis*, 39(1):1–33, 2019.

[11] Nicolas Boumal, Bamdev Mishra, P-A Absil, and Rodolphe Sepulchre. Manopt, a Matlab toolbox for optimization on manifolds. *The Journal of Machine Learning Research*, 15(1):1455–1459, 2014.

[12] Sébastien Bubeck and Nicolo Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *arXiv preprint arXiv:1204.5721*, 2012.

[13] Amit Chattopadhyay, Suvishamuthu Easter Selvan, and Umberto Amato. A derivative-free Riemannian Powell’s method, minimizing hartley-entropy-based ICA contrast. *IEEE transactions on neural networks and learning systems*, 27(9):1983–1990, 2015.

[14] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pages 15–26, 2017.

[15] Shixiang Chen, Zengde Deng, Shiqian Ma, and Anthony Man-Cho So. Manifold proximal point algorithms for dual principal component pursuit and orthogonal dictionary learning. In *2019 53rd Asilomar Conference on Signals, Systems, and Computers*, pages 259–263. IEEE, 2019.
[16] Shixiang Chen, Shiqian Ma, Anthony Man-Cho So, and Tong Zhang. Proximal gradient method for nonsmooth optimization over the Stiefel manifold. SIAM Journal on Optimization, 30(1):210–239, 2020.

[17] Tianyi Chen and Georgios B Giannakis. Bandit convex optimization for scalable and dynamic IoT management. IEEE Internet of Things Journal, 6(1):1276–1286, 2018.

[18] Chao-Kai Chiang, Chia-Jung Lee, and Chi-Jen Lu. Beating bandits in gradually evolving worlds. In Conference on Learning Theory, pages 210–227, 2013.

[19] Rishabh Dixit, Amrit Singh Bedi, Ruchi Tripathi, and Ketan Rajawat. Online learning with inexact proximal online gradient descent algorithms. IEEE Transactions on Signal Processing, 67(5):1338–1352, 2019.

[20] P Thomas Fletcher and Sarang Joshi. Riemannian geometry for the statistical analysis of diffusion tensor data. Signal Processing, 87(2):250–262, 2007.

[21] Robert Simon Fong and Peter Tino. Stochastic derivative-free optimization on Riemannian manifolds. arXiv preprint arXiv:1908.06783, 2019.

[22] Elad Hazan. Introduction to online convex optimization. arXiv preprint arXiv:1909.05207, 2019.

[23] Alex S Ira, Chris Manzie, Iman Shames, Robert Chin, Dragan Nešić, Hayato Nakada, and Takeshi Sano. Tuning of multivariable model predictive controllers through expert bandit feedback. International Journal of Control, pages 1–9, 2020.

[24] Hiroyuki Kasai, Hiroyuki Sato, and Bamdev Mishra. Riemannian stochastic recursive gradient algorithm. In International Conference on Machine Learning, pages 2516–2524, 2018.

[25] Artiom Kovnatsky, Klaus Glashoff, and Michael M Bronstein. MADMM: a generic algorithm for non-smooth optimization on manifolds. In European Conference on Computer Vision, pages 680–696. Springer, 2016.

[26] Sebastian A Kurtek, Anuj Srivastava, and Wei Wu. Signal estimation under random time-warpings and nonlinear signal alignment. In Advances in Neural Information Processing Systems, pages 675–683, 2011.

[27] Jiaxiang Li, Krishnakumar Balasubramanian, and Shiqian Ma. Stochastic zeroth-order Riemannian derivative estimation and optimization. arXiv preprint arXiv:2003.11238, 2020.

[28] Xiao Li, Shixiang Chen, Zengde Deng, Qing Qu, Zhihui Zhu, and Anthony Man Cho So. Weakly convex optimization over Stiefel manifold using Riemannian subgradient-type methods. arXiv, pages arXiv–1911, 2019.

[29] Dhruv Malik, Ashwin Pananjady, Kush Bhatia, Koulik Khamaru, Peter Bartlett, and Martin Wainwright. Derivative-free methods for policy optimization: Guarantees for linear quadratic systems. In The 22nd International Conference on Artificial Intelligence and Statistics, pages 2916–2925. PMLR, 2019.

[30] Horia Mania, Aurelia Guy, and Benjamin Recht. Simple random search of static linear policies is competitive for reinforcement learning. In Advances in Neural Information Processing Systems, pages 1800–1809, 2018.
[31] Jonathan H Manton. Optimization algorithms exploiting unitary constraints. *IEEE Transactions on Signal Processing*, 50(3):635–650, 2002.

[32] Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

[33] Michael JD Powell. An efficient method for finding the minimum of a function of several variables without calculating derivatives. *The computer journal*, 7(2):155–162, 1964.

[34] Yogesh Rathi, Allen Tannenbaum, and Oleg Michailovich. Segmenting images on the tensor manifold. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.

[35] Shai Shalev-Shwartz et al. Online learning and online convex optimization. *Foundations and trends in Machine Learning*, 4(2):107–194, 2011.

[36] Iman Shames, Daniel Selvaratnam, and Jonathan H Manton. Online optimization using zeroth order oracles. *IEEE Control Systems Letters*, 4(1):31–36, 2019.

[37] Andrea Simonetto, Emiliano Dall’Anese, Santiago Paternain, Geert Leus, and Georgios B Giannakis. Time-varying convex optimization: Time-structured algorithms and applications. *Proceedings of the IEEE*, 2020.

[38] Ju Sun, Qing Qu, and John Wright. Complete dictionary recovery over the sphere II: Recovery by Riemannian trust-region method. *IEEE Transactions on Information Theory*, 63(2):885–914, 2016.

[39] Bart Vandereycken. Low-rank matrix completion by Riemannian optimization. *SIAM Journal on Optimization*, 23(2):1214–1236, 2013.

[40] Bokun Wang, Shiqian Ma, and Lingzhou Xue. Riemannian stochastic proximal gradient methods for nonsmooth optimization over the Stiefel manifold. *arXiv preprint arXiv:2005.01209*, 2020.

[41] Melanie Weber and Suvrit Sra. Projection-free nonconvex stochastic optimization on Riemannian manifolds. *arXiv preprint arXiv:1910.04194*, 2019.

[42] Tianbao Yang, Lijun Zhang, Rong Jin, and Jinfeng Yi. Tracking slowly moving clairvoyant: Optimal dynamic regret of online learning with true and noisy gradient. In *International Conference on Machine Learning*, pages 449–457, 2016.

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

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

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