

Gradient Method for Optimization on Riemannian Manifolds with Lower Bounded Curvature

O. P. Ferreira ^{*} M. S. Louzeiro ^{*} L. F. Prudente ^{*}

June 8, 2018

Abstract

The gradient method for minimize a differentiable convex function on Riemannian manifolds with lower bounded sectional curvature is analyzed in this paper. The analysis of the method is presented with three different finite procedures for determining the stepsize, namely, Lipschitz stepsize, adaptive stepsize and Armijo's stepsize. The first procedure requires that the objective function has Lipschitz continuous gradient, which is not necessary for the other approaches. Convergence of the whole sequence to a minimizer, without any level set boundedness assumption, is proved. Iteration-complexity bound for functions with Lipschitz continuous gradient is also presented. Numerical experiments are provided to illustrate the effectiveness of the method in this new setting and certify the obtained theoretical results. In particular, we consider the problem of finding the Riemannian center of mass and the so-called Karcher's mean. Our numerical experiences indicate that the adaptive stepsize is a promising scheme that is worth considering.

Keywords: Gradient method, convex programming, Riemannian manifold, lower bounded curvature, iteration-complexity bound.

AMS subject classification: 90C33 · 49K05 · 47J25

1 Introduction

We consider the gradient method to solve the optimization problem defined by:

$$\min\{f(p) : p \in \mathcal{M}\}, \quad (1)$$

where the constraint set \mathcal{M} is endowed with a structure of a *complete Riemannian manifold with lower bounded curvature* and $f : \mathcal{M} \rightarrow \mathbb{R}$ is a *continuously differentiable convex function*. It is well known that, in several cases, by endowing \mathcal{M} with a suitable Riemannian metric, an Euclidean non-convex constrained problem can be seen as a Riemannian convex unconstrained problem. In addition to this property, we will present

^{*}IME/UFG, Avenida Esperança, s/n, Campus Samambaia, Goiânia, GO, 74690-900, Brazil (e-mails: orizon@ufg.br, mauriciosilvalouzeiro@gmail.com, lfprudente@ufg.br).

some examples showing that endowing the set of constraints with a suitable Riemannian metric the objective function can be also *Riemannian Lipschitz gradient*. Consequently, the geometric and algebraic structure that comes from the Riemannian metric makes possible to greatly reduce the computational cost for solving such problems. Indeed, it is also widely known that, in several contexts, the iteration complexity of the gradient method for convex optimization problems with Lipschitz gradient is much lower than for general nonconvex problems; see for example [6, 18, 28, 33, 38] and references therein. Furthermore, many Euclidean optimization problems are naturally posed on the Riemannian context; see [15, 18, 32, 33]. Then, to take advantage of the Riemannian geometric structure, it is preferable to treat these problems as the ones of finding singularities of gradient vector fields on Riemannian manifolds rather than using Lagrange multipliers or projection methods; see [23, 32, 34]. Accordingly, constrained optimization problems can be viewed as unconstrained ones from a Riemannian geometry point of view. Moreover, Riemannian structures can also opens up new research directions that aid in developing competitive algorithms; see [1, 15, 18, 27, 32, 33]. For this purpose, extensions of concepts and techniques of optimization from Euclidean space to Riemannian context have been quite frequently in recent years. Papers dealing with this subject include, but are not limited to [21, 22, 24, 35, 36, 38, 39].

The gradient method is one of the oldest methods for the minimization of a differentiable function in Euclidean space. Despite having slow convergence rate, the simplicity of implementation, the low memory requirements and cost per iteration, make the gradient method quite attractive to solve large-scale optimization problems. Indeed, the computational cost per iteration is mildly dependent on the dimension of the problem, yielding computational efficiency for this method; see [18, 26, 29]. In addition, the gradient method is the starting point for designing many more sophisticated and efficient algorithms, including fast gradient method, accelerated gradient method and Barzilai-Borwein method; see [25, 37] for a comprehensive study on this subject. To the best of our knowledge the gradient method was the first optimization method to be considered in a Riemannian setting. In order to deal with contained optimization problems in the Euclidean space, Luenberger [23] proposed and established important convergence properties of gradient method by using the Riemannian structure of the constraint set induced by the Euclidean structure. Since then, the gradient method has been studied in general Riemannian manifold. Some early works dealing with this method include [17, 28, 32, 34]. However, the obtained convergence results in these previous works demand that the initial points of the sequence belong to a bounded level set of the objective function establishing only that all its cluster points are stationary. By assuming convexity of the objective function and that the manifolds has *non-negative curvature*, it has been proven in [11] that, for a suitable choice of the stepsize and without any level set boundedness assumption, the whole sequence converges to a solution. Recently new important properties of the gradient method in Riemannian settings have been obtained. For instance, in [39] the authors provided iteration-complexity bounds for convex optimization problems on Hadamard manifolds. In [8], the authors established iteration-complexity bounds without any assumption on the convexity of the

problem and curvature of the manifold. In [7] the gradient method is considered to compute the Karcher mean, which is a strong convex function in the cone of symmetric positive definite matrices endowed with a suitable Riemannian metric. In [2] is studied properties of the gradient method for the problem of finding the global Riemannian center of mass of a set of data points on a Riemannian manifold. In [5] is extended the convergence analysis of the gradient method to the Hadamard setting for continuously differentiable functions which satisfy the Kurdyka-Lojasiewicz inequality.

By the aforementioned we see that the gradient method remains a subject of considerable interest. In spite of its long history, the full convergence of the sequence generated by the gradient method in a general Riemannian manifolds has not yet been established. However, as far as we know, the full convergence of the sequence generated by the gradient method under convexity of the objective function and *lower boundedness of the curvature* of Riemannian manifolds is a new contribution of this paper, which adds important results in the available convergence theory of this method. The analysis of the method is presented with three different finite procedures for determining the stepsize, namely, Lipschitz stepsize, adaptive stepsize and Armijo's stepsize. It should be noted that we use a recent inequality established in [35,36]. Numerical experiments are provided to illustrate the effectiveness of the method in this new setting and certify the obtained theoretical results. In particular, we consider the problem of finding the Riemannian mass center and the so-called Karcher's mean. Our experiments indicate that adaptive size is a promising scheme that is worth considering.

This paper is organized as follows. Section 2 presents some definitions and preliminary results related to the Riemannian geometry that are important throughout our study. In Section 3, we state the gradient algorithm and the three different finite procedures for determining the stepsize. Section 3.1 is devoted to the asymptotic convergence analysis of the method, and in Section 3.2 the iteration-complexity bound is presented. Section 4 provides some examples of functions satisfying the assumptions of our results in the previous sections. In Section 5, we present some numerical experiments to illustrate the behavior of the method. The last section contains some conclusions.

2 Notations and basic concepts

In this section, we recall some concepts, notations, and basics results about Riemannian manifolds. For more details we refer the reader to [13, 28, 31, 34].

We denote by $T_p\mathcal{M}$ the *tangent space* of a finite dimensional Riemannian manifold \mathcal{M} at p . The corresponding norm associated to the Riemannian metric $\langle \cdot, \cdot \rangle$ is denoted by $\| \cdot \|$. We use $\ell(\alpha)$ to denote the length of a piecewise smooth curve $\alpha : [a, b] \rightarrow \mathcal{M}$. The Riemannian distance between p and q in \mathcal{M} is denoted by $d(p, q)$, which induces the original topology on \mathcal{M} , namely, (\mathcal{M}, d) , which is a complete metric space where bounded and closed subsets are compact. The *closed metric ball* in \mathcal{M} centered at the point $p \in \mathcal{M}$ with radius $r > 0$ is denoted by $B[p, r]$. Denote by $\mathcal{X}(\mathcal{M})$, the space of smooth vector fields on \mathcal{M} . Let ∇ be the Levi-Civita connection associated to $(\mathcal{M}, \langle \cdot, \cdot \rangle)$.

For each $t \in [a, b]$ and a piecewise smooth curve $\alpha : [a, b] \rightarrow \mathcal{M}$, ∇ induces an isometry relative to $\langle \cdot, \cdot \rangle$, $P_{\alpha, a, t} : T_{\alpha(a)}\mathcal{M} \rightarrow T_{\alpha(t)}\mathcal{M}$ defined by $P_{\alpha, a, t}v = V(t)$, where V is the unique vector field on the curve α such that $\nabla_{\alpha'(t)}V(t) = 0$ and $V(a) = v$. The isometry $P_{\alpha, a, t}$ is called *parallel transport* along of α joining $\alpha(a)$ to $\alpha(t)$ and, when there is no confusion, it will be denoted by $P_{\alpha, p, q}$. A vector field V along a smooth curve γ is said to be *parallel* iff $\nabla_{\gamma'}V = 0$. If γ' itself is parallel, we say that γ is a *geodesic*. Given that the geodesic equation $\nabla_{\gamma'}\gamma' = 0$ is a second order nonlinear ordinary differential equation, then the geodesic $\gamma = \gamma_v(\cdot, p)$ is determined by its position p and velocity v at p . It is easy to check that $\|\gamma'\|$ is constant. The restriction of a geodesic to a closed bounded interval is called a *geodesic segment*. A geodesic segment joining p to q in \mathcal{M} is said to be *minimal* if its length is equal to $d(p, q)$. A Riemannian manifold is *complete* if the geodesics are defined for any values of $t \in \mathbb{R}$. Hopf-Rinow's theorem asserts that any pair of points in a complete Riemannian manifold \mathcal{M} can be joined by a (not necessarily unique) minimal geodesic segment. Owing to the completeness of the Riemannian manifold \mathcal{M} , the *exponential map* $\exp_p : T_p\mathcal{M} \rightarrow \mathcal{M}$ is given by $\exp_p v = \gamma_v(1, p)$, for each $p \in \mathcal{M}$. In this paper, all manifolds are assumed to be connected, finite dimensional, and complete. For $f : \mathcal{D} \rightarrow \mathbb{R}$ a differentiable function on the open set $\mathcal{D} \subset \mathcal{M}$, the Riemannian metric induces the mapping $f \mapsto \text{grad}f$ associates its *gradient* via the following rule $\langle \text{grad}f(p), X(p) \rangle := df(p)X$, for all $p \in \mathcal{D}$. For a twice-differentiable function, the mapping $f \mapsto \text{hess}f$ associates its *hessian* via the rule $\langle \text{hess}f X, X \rangle := d^2f(X, X)$, for all $X \in \mathcal{X}(\mathcal{D})$, where the last equalities imply that $\text{hess}f X = \nabla_X \text{grad}f$, for all $X \in \mathcal{X}(\mathcal{D})$. We proceeded to recall some concepts and basic properties about convexity in the Riemannian context. For more details see, for example, [28, 34, 35]. For any two points $p, q \in \mathcal{M}$, Γ_{pq} denotes the set of all geodesic segments $\gamma : [0, 1] \rightarrow \mathcal{M}$ with $\gamma(0) = p$ and $\gamma(1) = q$. We use Γ_{pq}^Ω to denote the set of all $\gamma \in \Gamma_{pq}$ such that $\gamma(t) \in \Omega$, for all $t \in [0, 1]$. A nonempty subset $\Omega \subset \mathcal{M}$ is said to be *weakly convex* if, for any $p, q \in \Omega$, there is a minimal geodesic segment joining p to q belonging Ω . A function $f : \mathcal{D} \rightarrow \mathbb{R}$ is said to be *convex* on the set $\Omega \subset \mathcal{D}$ if Ω is weakly convex and for any $p, q \in \Omega$ and $\gamma \in \Gamma_{pq}^\Omega$ the composition $f \circ \gamma : [0, 1] \rightarrow \mathbb{R}$ is a convex function on $[0, 1]$, i.e., $f \circ \gamma(t) \leq (1-t)f(p) + tf(q)$, for all $t \in [0, 1]$; see [35]. For f a differentiable function on \mathcal{D} and a weakly convex set $\Omega \subset \mathcal{D}$, we have the following characterization: f is convex on Ω iff there holds $f(\gamma(t)) \geq f(p) + \langle \text{grad}f(p), \gamma'(0) \rangle$, for all $p, q \in \Omega$ and $\gamma \in \Gamma_{pq}^\Omega$.

The following lemma plays an important role in next sections and its proof, with some minor technical adjustments, can be found in [35, Lemma 3.2]; see also [36]. For simplifying the notations, let

$$\kappa < 0, \quad \hat{\kappa} := \sqrt{|\kappa|}. \quad (2)$$

Lemma 1. *Let \mathcal{M} be a Riemannian manifold with sectional curvature $K \geq \kappa$, and $\hat{\kappa}$ be defined in (2). Assume that f is differentiable and convex on the set $\Omega \subset \mathcal{M}$, $p \in \Omega$ and $\gamma : [0, \infty) \rightarrow \mathcal{M}$ is defined by $\gamma(t) = \exp_p(-t \text{ grad}f(p))$. Then, for any $t \in [0, \infty)$*

and $q \in \Omega$ there holds

$$\cosh(\hat{\kappa}d(\gamma(t), q)) \leq \cosh(\hat{\kappa}d(p, q)) + \hat{\kappa} \cosh(\hat{\kappa}d(p, q)) \sinh(t\hat{\kappa} \|\text{grad } f(p)\|) \left[\frac{t \|\text{grad } f(p)\|}{2} - \frac{\tanh(\hat{\kappa}d(p, q))}{\hat{\kappa}d(p, q)} \frac{f(p) - f(q)}{\|\text{grad } f(p)\|} \right]$$

and, consequently, the following inequality holds

$$d^2(\gamma(t), q) \leq d^2(p, q) + \frac{\sinh(\hat{\kappa}t \|\text{grad } f(p)\|)}{\hat{\kappa}} \left[t \|\text{grad } f(p)\| \frac{\hat{\kappa}d(p, q)}{\tanh(\hat{\kappa}d(p, q))} - \frac{2}{\|\text{grad } f(p)\|} (f(p) - f(q)) \right].$$

Next we present the definition of Lipschitz continuous gradient vector field; see [10].

Definition 1. Let f be a differentiable function on the set \mathcal{D} . The gradient vector field of f is said to be Lipschitz continuous on \mathcal{D} with constant $L \geq 0$ if, for any $p, q \in \mathcal{D}$ and $\gamma \in \Gamma_{pq}^{\mathcal{D}}$, it holds that $\|P_{\gamma, p, q} \text{grad } f(p) - \text{grad } f(q)\| \leq L\ell(\gamma)$.

The norm of the hessian $\text{hess } f$ at $p \in \mathcal{M}$ is given by

$$\|\text{hess } f(p)\| := \sup \{ \|\text{hess } f(p)v\| : v \in T_p\mathcal{M}, \|v\| = 1 \}. \quad (3)$$

In the following result we present a characterization for twice continuously differentiable functions with Lipschitz continuous gradient vector field, which has similar proof of its Euclidean counterpart and will be omitted here.

Lemma 2. Let $f : \mathcal{D} \rightarrow \mathbb{R}$ be a twice continuously differentiable function. The gradient vector field of f is Lipschitz continuous with constant $L \geq 0$ if, and only if, there exists $L \geq 0$ such that $\|\text{hess } f(p)\| \leq L$, for all $p \in \mathcal{D}$.

The next lemma can be found in [6, Corollary 2.1] with minor adjustment. Its proof follows from the definition of convexity of functions and the fundamental theorem of calculus.

Lemma 3. Let f be a differentiable function on the set \mathcal{D} and $a > 0$. Assume that $\text{grad } f$ is Lipschitz continuous on \mathcal{D} with constant $L \geq 0$ and $p \in \Omega$. If $\exp_p(-t \text{grad } f(p)) \in \mathcal{D}$, for all $t \in [0, a]$, then there holds

$$f(\exp_p(-t \text{grad } f(p))) \leq f(p) - \left(1 - \frac{L}{2}t\right) t \|\text{grad } f(p)\|^2, \quad \forall t \in [0, a].$$

Note that if $\mathcal{D} = \mathcal{M}$, then condition $\exp_p(-t \text{grad } f(p)) \in \mathcal{D}$, for all $t \in [0, a]$, in Lemma 3 plays no role. In the following example we present a functions satisfying all the assumptions of Lemma 3 for the case $\mathcal{D} \neq \mathcal{M}$.

Example 1. Let $\mathcal{M} = \{p \in \mathbb{R}^n : \|p\| = 1\}$ the Euclidean sphere and $q \in \mathcal{M}$. Define $\varphi_q(p) := d^2(p, q)/2$, for all $p \in \mathcal{M}$. The function φ_q is differentiable in $\mathcal{D} := \{p \in \mathcal{M} : d(p, q) < 5\pi/6\}$ and convex in $\Omega := \{p \in \mathcal{M} : d(p, q) \leq \pi/2\}$. Furthermore, $\text{grad } \varphi_q$ is Lipschitz continuous on \mathcal{D} , because \mathcal{D} is compact and $\text{hess } \varphi_q$ is continuous in $\mathcal{M} \setminus \{-q\} \supset \mathcal{D}$. Indeed, combining [16, Lemma 3] with Lemma 2 we conclude that

$$L = \sup_{p \in \mathcal{D}} \frac{|\langle p, q \rangle \arccos \langle p, q \rangle|}{\sqrt{1 - \langle p, q \rangle^2}} = \frac{5\pi}{6} \sqrt{3}.$$

Since $\text{grad } \varphi_q(p) = -\exp_p^{-1} q$ for all $p \in \mathcal{M} \setminus \{-q\}$, after some calculations, we conclude that $d(\exp_p(-t \text{grad } \varphi_q(p)), p) \leq td(p, q)$, for all $p \in \mathcal{D}$. Hence, letting $p \in \Omega$ we have

$$d((\exp_p(-t \text{grad } \varphi_q(p))), q) \leq d(\exp_p(-t \text{grad } \varphi_q(p)), p) + d(p, q) \leq (t+1) \frac{\pi}{2},$$

and then $\exp_p(-t \text{grad } \varphi_q(p)) \in \mathcal{D}$, for all $t \in [0, 1/L]$. For more details about the function φ_q ; see [16].

The following concept will be useful in the analysis of the sequence generated by the gradient method. In fact, as we shall prove, the sequence generated by this method satisfies the following definition.

Definition 2. A sequence $\{y_k\}$ in the complete metric space (\mathcal{M}, d) is quasi-Fejér convergent to a set $W \subset \mathcal{M}$ if, for every $w \in W$, there exist a sequence $\{\epsilon_k\} \subset \mathbb{R}$ such that $\epsilon_k \geq 0$, $\sum_{k=1}^{\infty} \epsilon_k < +\infty$, and $d^2(y_{k+1}, w) \leq d^2(y_k, w) + \epsilon_k$, for all $k = 0, 1, \dots$

The main property of a quasi-Fejér sequence is stated in the next result, and its proof is similar to the one proved in [9], by replacing the Euclidean distance by the Riemannian.

Theorem 1. Let $\{y_k\}$ be a sequence in the complete metric space (\mathcal{M}, d) . If $\{y_k\}$ is quasi-Fejér convergent to a nonempty set $W \subset \mathcal{M}$, then $\{y_k\}$ is bounded. If furthermore, a cluster point \bar{y} of $\{y_k\}$ belongs to W , then $\lim_{k \rightarrow \infty} y_k = \bar{y}$.

The study of the gradient method for convex functions is well understood for Riemannian manifold with nonnegative sectional curvature and Hadamard manifolds; see [10, 38, 39]. In order to increase the domain of applications of the method, hereafter, we assume that \mathcal{M} is a complete Riemannian manifolds with sectional curvature $K \geq \kappa$, where $\kappa < 0$, unless the contrary is explicitly stated.

3 The Riemannian gradient method

In this section we state the Riemannian gradient method to solve (1) and the strategies for choosing the stepsize that will be used in our analysis.

Let $f : \mathcal{D} \rightarrow \mathbb{R}$ be differentiable, $\mathcal{D} \subset \mathcal{M}$ be an open set, Ω^* be the *solution set* of the problem (1), $f^* := \inf_{x \in \mathcal{D}} f(x)$ be the *optimum value* of f , and $c \in \mathbb{R}$. From now on, we assume that Ω^* is non-empty and f is convex on the sub-level set $\mathcal{L}_c f$, where

$$\mathcal{L}_c f := \{p \in \mathcal{M} : f(p) \leq c\} \subset \mathcal{D}.$$

The statement of *Riemannian gradient algorithm* to solve the problem (1) is as follows.

Algorithm 1. Gradient algorithm in a Riemannian manifold \mathcal{M}

Step 0. Let $p_0 \in \mathcal{L}_c f$. Set $k = 0$.

Step 1. If $\text{grad} f(p_k) = 0$, then **stop**; otherwise, choose a stepsize $t_k > 0$ and compute

$$p_{k+1} := \exp_{p_k}(-t_k \text{grad} f(p_k)). \quad (4)$$

Step 2. Set $k \leftarrow k + 1$ and proceed to **Step 1**.

In the following we present three different strategies for choosing the stepsize $t_k > 0$ in Algorithm 1. In the first strategy we assume that $\text{grad} f$ is Lipschitz continuous.

Strategy 1 (Lipschitz stepsize). Assume that $\text{grad} f$ is Lipschitz continuous on \mathcal{D} with constant $L \geq 0$ and that $\exp_p(-t \text{grad} f(p)) \in \mathcal{D}$, for all $p \in \mathcal{L}_c f$ and $t \in [0, 1/L]$. Let $\varepsilon > 0$ and take

$$\varepsilon < t_k \leq \frac{1}{L}. \quad (5)$$

Remark 1. If $\mathcal{D} = \mathcal{M}$, then condition $\exp_p(-t \text{grad} f(p)) \in \mathcal{D}$, for all $t \in [0, a]$, in Strategy 1 plays no role. Recall that the function in Example 1 satisfies this condition for $\mathcal{D} \neq \mathcal{M}$.

Despite knowing that $\text{grad} f$ is Lipschitz continuous, in general, the Lipschitz constant is not computable. Next strategy can be used to compute the stepsize without any Lipschitz condition. However, as we shall show, if $\text{grad} f$ is Lipschitz with constant $L > 0$ the stepsize computed is an approximation to the stepsize $1/L$; see [4].

Strategy 2 (adaptive stepsize). Take $\beta \in (0, 1)$, $L_0 > 0$, and $\eta > 1$. Set $t_k := L_k^{-1}$, where $L_k := \eta^{i_k} L_{k-1}$ and

$$i_k := \min \left\{ i : f(\gamma_k(\tau_i)) \leq f(p_k) - \beta \tau_i \|\text{grad} f(p_k)\|^2, \ i = 0, 1, \dots \right\}, \quad (6)$$

where $\tau_i := (\eta^i L_{k-1})^{-1}$ and $\gamma_k(\tau_i) := \exp_{p_k}(-\tau_i \text{grad} f(p_k))$.

Strategy 3 (Armijo's stepsize). Choose $\beta \in (0, 1)$ and take

$$t_k := \max \left\{ 2^{-i} : f(\gamma_k(2^{-i})) \leq f(p_k) - \beta 2^{-i} \|\text{grad} f(p_k)\|^2, \ i = 0, 1, \dots \right\}, \quad (7)$$

where $\gamma_k(2^{-i}) := \exp_{p_k}(-2^{-i} \text{grad} f(p_k))$.

Remark 2. *Strategy 2 can be seen as an Armijo-type line search where the first trial stepsize at iteration k is set to be equal to t_{k-1} . Indeed, taking $L_0 = 1$, and $\eta = 2$ the inequality in (6) can be equivalently rewritten as*

$$f(\gamma_k(2^{-i}t_{k-1})) \leq f(p_k) - \beta 2^{-i}t_{k-1} \|\text{grad } f(p_k)\|^2.$$

The proof of the well-definedness of Strategies 2 and 3 follows the usual arguments and will be omitted. On the other hand, (5) and Lemma 3 imply that, for each $p \in \mathcal{L}_c f$ there holds $\exp_p(-t \text{grad } f(p)) \in \mathcal{L}_c f$, for all $t \in [0, 1/L]$. Hence, the sequence $\{p_k\}$ generated by Algorithm 1 with Strategies 1, 2 or 3 is well-defined. Finally we remark that, due to f be convex, $\text{grad } f(p) = 0$ if and only if $p \in \Omega^*$. Therefore, *from now on we assume that $\text{grad } f(p_k) \neq 0$, or equivalently, $p_k \notin \Omega^*$, for all $k = 0, 1, \dots$*

3.1 Asymptotic convergence Analysis

In this section our goal is to prove that the sequence $\{p_k\}$, generated by the gradient method with Strategies 1, 2 or 3, converges to a solution of problem (1).

Lemma 4. *Let $\{p_k\}$ be generated by Algorithm 1 with Strategies 1, 2 or 3. Then,*

$$f(p_{k+1}) \leq f(p_k) - \nu t_k \|\text{grad } f(p_k)\|^2, \quad k = 0, 1, \dots, \quad (8)$$

where $\nu = 1/2$ for Strategy 1, and $\nu = \beta$ for Strategies 2 and 3. Consequently, $\{f(p_k)\}$ is non-increasing sequence and $\lim_{k \rightarrow +\infty} t_k \|\text{grad } f(p_k)\|^2 = 0$.

Proof. For Strategies 2 and 3, inequality (8) follows directly from (6) and (7), respectively. Now, we assume that $\{p_k\}$ is generated by using Strategy 1. In this case, Lemma 3 implies that

$$f(p_{k+1}) = f(\exp_{p_k}(-t_k \text{grad } f(p_k))) \leq f(p_k) - \left(1 - \frac{L}{2}t_k\right) t_k \|\text{grad } f(p_k)\|^2,$$

for all $k = 0, 1, \dots$. Hence, taking into account (5) we have $1/2 \leq (1 - Lt_k/2)$ and then, (8) follows. Therefore, (8) holds for $\{p_k\}$ generated by using the three strategies. It is immediate from (8) that $\{f(p_k)\}$ is non-increasing. Moreover, (8) implies that

$$\sum_{k=0}^{\ell} t_k \|\text{grad } f(p_k)\|^2 \leq \frac{1}{\nu} \sum_{k=0}^{\ell} f(p_k) - f(p_{k+1}) \leq \frac{1}{\nu} (f(p_0) - f^*),$$

for each nonnegative integer ℓ , which implies that $t_k \|\text{grad } f(p_k)\|^2$ goes to zero, as k goes to infinity, completing the proof. \square

Remark 3. *Whenever f is Lipschitz continuous on \mathcal{D} with constant $L \geq 0$, the stepsize in Strategy 2 can be seen as an approximation for the Lipschitz constant. Indeed, since $L_0 > 0$ and $\eta > 1$ in Strategy 2, we conclude that $t_k := L_k^{-1} \leq L_{k-1}^{-1} = t_{k-1}$, for all*

$k = 0, 1, \dots$. Thus $t_k \leq 1/L_0$, for all $k = 0, 1, \dots$. If $L_0 \geq L$, then it follows from (8) that $t_k \leq 1/L_0$, for all $k = 0, 1, \dots$. Now assume that $L_0 \leq L$. In this case, (8) holds for $t_k = 1/L$ and then (6) implies that $1/(\eta L) \leq t_k$. Therefore,

$$\frac{1}{\eta L} \leq t_k \leq \frac{1}{L_0}, \quad k = 0, 1, \dots \quad (9)$$

Let $p_0 \in \mathcal{M}$. By Lemma 4, we define constant $\rho > 0$ as follows

$$\sum_{k=0}^{\infty} t_k^2 \|\text{grad } f(p_k)\|^2 \leq \rho := \begin{cases} 2[f(p_0) - f^*]/L, & \text{for Strategy 1;} \\ [f(p_0) - f^*]/(\beta L_0), & \text{for Strategy 2;} \\ [f(p_0) - f^*]/\beta, & \text{for Strategy 3.} \end{cases} \quad (10)$$

In the following result, in particular, we bound the sequence $\{p_k\}$ generated by Algorithm 1 with Strategies 1, 2 or 3.

Lemma 5. *Let $q \in \Omega^*$ and $\{p_k\}$ the sequence generated by Algorithm 1 with Strategies 1, 2 or 3. Then there holds*

$$d(p_{k+1}, q) \leq \frac{1}{\sqrt{\kappa}} \cosh^{-1} \left(\cosh(\sqrt{\kappa} d(p_0, q)) e^{\frac{1}{2}(\sqrt{\kappa}\rho) \sinh(\sqrt{\kappa}\rho)} \right), \quad k = 0, 1, \dots \quad (11)$$

Proof. Applying the first inequality of Lemma 1, with $t = t_k$ and $p = p_k$, we have $p_{k+1} = \gamma(t_k)$, and taking into account that $q \in \Omega^*$, we conclude that

$$\cosh(\hat{\kappa} d(p_{k+1}, q)) \leq \cosh(\hat{\kappa} d(p_k, q)) \left[1 + (\hat{\kappa} t_k \|\text{grad } f(p_k)\|)^2 \frac{\sinh(\hat{\kappa} t_k \|\text{grad } f(p_k)\|)}{2\hat{\kappa} t_k \|\text{grad } f(p_k)\|} \right],$$

for all $k = 0, 1, \dots$, where $\hat{\kappa}$ is defined in (2). Since (10) implies $t_k \|\text{grad } f(p_k)\| \leq \sqrt{\rho}$, for all $k = 0, 1, \dots$, and the map $(0, +\infty) \ni t \mapsto \sinh(t)/t$ is increasing, we conclude that

$$\cosh(\hat{\kappa} d(p_{k+1}, q)) \leq \cosh(\hat{\kappa} d(p_k, q)) \left[1 + a (t_k \|\text{grad } f(p_k)\|)^2 \right], \quad k = 0, 1, \dots,$$

where $a := \hat{\kappa}(\sinh(\hat{\kappa}\sqrt{\rho}))/ (2\sqrt{\rho})$. Now note that the last inequality implies that

$$\cosh(\hat{\kappa} d(p_{k+1}, q)) \leq \cosh(\hat{\kappa} d(p_k, q)) e^{a(t_k \|\text{grad } f(p_k)\|)^2}, \quad k = 0, 1, \dots,$$

Therefore, by using (10), it follows that $\cosh(\hat{\kappa} d(p_{k+1}, q)) \leq \cosh(\hat{\kappa} d(p_0, q)) e^{a\rho}$, which is equivalent to (11) by considering the definition of $\hat{\kappa}$ in (2). \square

Let us define the following auxiliary constant

$$C_{\rho, \kappa}^q := \frac{\sinh(\sqrt{\kappa}\rho)}{\sqrt{\kappa}\rho} \left[1 + \cosh^{-1} \left(\cosh(\sqrt{\kappa} d(p_0, q)) e^{\frac{1}{2}(\sqrt{\kappa}\rho) \sinh(\sqrt{\kappa}\rho)} \right) \right], \quad (12)$$

where ρ is defined in (10).

Lemma 6. *Let $\{p_k\}$ be generated by by Algorithm 1 with Strategies 1, 2 or 3. Then, for each $q \in \Omega^*$, there holds*

$$d^2(p_{k+1}, q) \leq d^2(p_k, q) + \frac{t_k}{\nu} \mathcal{C}_{\rho, \kappa}^q [f(p_k) - f(p_{k+1})] + 2t_k [f(q) - f(p_k)], \quad (13)$$

for all $k = 0, 1, \dots$, where $\nu = 1/2$ for Strategy 1 and $\nu = \beta$ for Strategies 2 and 3.

Proof. Define $\gamma_k(t) = \exp_{p_k}(-t \operatorname{grad} f(p_k))$, for all $t \in [0, +\infty)$. Then, $\gamma_k(0) = p_k$ and, from (4), we obtain $\gamma_k(t_k) = p_{k+1}$. Applying second inequality of Lemma 1 with $\gamma = \gamma_k$, after some manipulations, we conclude that

$$d^2(p_{k+1}, q) \leq d^2(p_k, q) + \frac{\sinh(\hat{\kappa} t_k \|\operatorname{grad} f(p_k)\|)}{\hat{\kappa} t_k \|\operatorname{grad} f(p_k)\|} \left[t_k^2 \|\operatorname{grad} f(p_k)\|^2 \frac{\hat{\kappa} d(p_k, q)}{\tanh(\hat{\kappa} d(p_k, q))} + 2t_k [f(q) - f(p_k)] \right], \quad (14)$$

for all $k = 0, 1, \dots$. On the other hand, $t/\tanh(t) \leq 1 + t$, for all $t \geq 0$, and the map $(0, +\infty) \ni t \mapsto \sinh(t)/t$ is increasing and bounded below by 1. Thus, taking into account that (10) implies $t_k \|\operatorname{grad} f(p_k)\| \leq \sqrt{\rho}$ for all $k = 0, 1, \dots$, and considering $f(q) - f(p_k) \leq 0$ for all $k = 0, 1, \dots$, we conclude from (14) that

$$d^2(p_{k+1}, q) \leq d^2(p_k, q) + \frac{\sinh(\hat{\kappa} \sqrt{\rho})}{\hat{\kappa} \sqrt{\rho}} t_k^2 \|\operatorname{grad} f(p_k)\|^2 [1 + \hat{\kappa} d(p_k, q)] + 2t_k [f(q) - f(p_k)],$$

for all $k = 0, 1, \dots$, where ρ is defined in (10). Thus, by Lemma 4, we obtain

$$d^2(p_{k+1}, q) \leq d^2(p_k, q) + \frac{t_k}{\nu} \frac{\sinh(\hat{\kappa} \sqrt{\rho})}{\hat{\kappa} \sqrt{\rho}} [1 + \hat{\kappa} d(p_k, q)] [f(p_k) - f(p_{k+1})] + 2t_k [f(q) - f(p_k)],$$

for all $k = 0, 1, \dots$. Therefore, by Lemma 5 and (12), we have (13), which concludes the proof. \square

Finally we are ready to prove the full convergence of $\{p_k\}$ to a minimizer of f .

Theorem 2. *Let $\{p_k\}$ be generated by by Algorithm 1 with Strategies 1, 2 or 3. Then $\{p_k\}$ converges to a solution of the problem in (1).*

Proof. First note that $f(q) - f(p_k) \leq 0$, for all $k = 0, 1, \dots$ and $q \in \Omega^*$. Hence, (5), (7) and (9) imply $0 < t_k \leq 1/L$ or $0 < t_k \leq 1/L_0$ or $0 < t_k \leq 1$, for all $k = 0, 1, \dots$, for Strategies 1, 2 or 3, respectively. Let $\Gamma := \max\{1, 1/L, 1/L_0\}$. Thus, for Strategies 1, 2 or 3 we conclude from Lemma 6 that

$$d^2(p_{k+1}, q) \leq d^2(p_k, q) + \frac{1}{\nu} \Gamma \mathcal{C}_{\rho, \kappa}^q [f(p_k) - f(p_{k+1})], \quad k = 0, 1, \dots,$$

for all $q \in \Omega^*$. Considering that $\sum_{i=0}^{\infty} [f(p_i) - f(p_{i+1})] \leq [f(p_0) - f^*]$, we conclude that $\{p_k\}$ is quasi-Fejér convergent to Ω^* . Therefore, since Ω^* is non-empty the sequence

$\{p_k\}$ is bounded. Let \bar{p} be a cluster point of $\{p_k\}$ and $\{p_{k_j}\}$ be a subsequence $\{p_k\}$ such that $\lim_{j \rightarrow \infty} p_{k_j} = \bar{p}$. It follows from Lemma 4 that $\lim_{k \rightarrow \infty} t_k \|\text{grad } f(p_k)\|^2 = 0$, and due to $\{t_k\}$ has a cluster point $\bar{t} \in [0, \Gamma]$, we analyze the following two possibilities

$$\text{(a)} \quad \bar{t} > 0, \quad \text{(b)} \quad \bar{t} = 0.$$

Assume that (a) holds. In this case, considering that $\lim_{k \rightarrow \infty} t_k \|\text{grad } f(p_k)\|^2 = 0$ and $\text{grad } f$ is continuous, we conclude that

$$0 = \lim_{j \rightarrow \infty} t_{k_j} \|\text{grad } f(p_{k_j})\| = \bar{t} \|\text{grad } f(\bar{p})\|.$$

Hence, $\text{grad } f(\bar{p}) = 0$ and then $\bar{p} \in \Omega^*$. Note that if Strategy 1 is used, then \bar{t} satisfies only (a). Now, we assume that (b) holds. In this case Strategies 2 or 3 is used. First assume Algorithm 1 with Strategy 2. Since $\{t_{k_j}\}$ converges to $\bar{t} = 0$ and $\{t_k\}$ is non-increasing, it follows that $\{t_k\}$ converges to $\bar{t} = 0$. Hence, taking $r \in \mathbb{N}$ we can conclude that $t_k < (\eta^r L_0)^{-1}$ for k sufficiently large. Considering k being the smallest natural number that satisfies $t_k < (\eta^r L_0)^{-1}$, by (6), we have

$$f(\exp_{p_k}((\eta^{r-1} L_0)^{-1}[-\text{grad } f(p_{k_j})])) > f(p_k) - (\eta^{r-1} L_0)^{-1} \beta \|\text{grad } f(p_k)\|^2.$$

Letting k go to $+\infty$ in the above inequality and taking into account that $\text{grad } f$ and the exponential mapping are continuous, we obtain

$$f(\exp_{\bar{p}}((\eta^{r-1} L_0)^{-1}[-\text{grad } f(\bar{p})])) \geq f(\bar{p}) - (\eta^{r-1} L_0)^{-1} \beta \|\text{grad } f(\bar{p})\|^2.$$

The last inequality is equivalent to

$$-\frac{f(\exp_{\bar{p}}((\eta^{r-1} L_0)^{-1}[-\text{grad } f(\bar{p})])) - f(\bar{p})}{(\eta^{r-1} L_0)^{-1}} \leq \beta \|\text{grad } f(\bar{p})\|^2.$$

Thus, letting r goes to $+\infty$ we obtain $\|\text{grad } f(\bar{p})\|^2 \leq \beta \|\text{grad } f(\bar{p})\|^2$ which implies $\text{grad } f(\bar{p}) = 0$, i.e., $\bar{p} \in \Omega^*$. Therefore, since $\{p_k\}$ is quasi-Fejér convergent to Ω^* , we conclude from Theorem 1 that $\{p_k\}$ converges to \bar{p} . Finally, assume that Strategy 3 is used. Since $\{t_{k_j}\}$ converges to $\bar{t} = 0$, taking $r \in \mathbb{N}$, we conclude that $t_{k_j} < 2^{-r}$ for j sufficiently large. Thus Armijo's condition (7) is not satisfied for $t = 2^{-r+1}$, i.e.,

$$f(\exp_{p_{k_j}}(2^{-r+1}[-\text{grad } f(p_{k_j})])) > f(p_{k_j}) - 2^{-r+1} \beta \|\text{grad } f(p_{k_j})\|^2.$$

Letting j go to $+\infty$ in the above inequality and taking into account that $\text{grad } f$ and the exponential mapping are continuous, we obtain

$$f(\exp_{\bar{p}}(2^{-r+1}[-\text{grad } f(\bar{p})])) \geq f(\bar{p}) - 2^{-r+1} \beta \|\text{grad } f(\bar{p})\|^2.$$

The last inequality is equivalent to

$$-\frac{f(\exp_{\bar{p}}(2^{-r+1}[-\text{grad } f(\bar{p})])) - f(\bar{p})}{2^{-r+1}} \leq \beta \|\text{grad } f(\bar{p})\|^2.$$

Thus, letting r goes to $+\infty$ we obtain $\|\text{grad } f(\bar{p})\|^2 \leq \beta \|\text{grad } f(\bar{p})\|^2$ which implies $\text{grad } f(\bar{p}) = 0$, i.e., $\bar{p} \in \Omega^*$. Therefore, since $\{p_k\}$ is quasi-Fejér convergent to Ω^* , we conclude from Theorem 1 that $\{p_k\}$ converges to \bar{p} and the proof is completed. \square

3.2 Iteration-Complexity Analysis

In this section we present an iteration-complexity bound related to the gradient method for minimizing a convex functions with Lipschitz continuous gradient with constant $L > 0$. In the following, as an application of Lemma 6, we obtain the iteration-complexity bound for the gradient method with Strategy 2.

Theorem 3. *Let $\{p_k\}$ be generated by by Algorithm 1 with Strategy 2. Then, for every $N \in \mathbb{N}$, there holds*

$$f(p_N) - f^* \leq \eta L \frac{L_0 d^2(p_0, q) + 2 (\mathcal{C}_{\rho, \kappa}^q - 1) [f(p_0) - f^*]}{2NL_0}, \quad (15)$$

for each $q \in \Omega^*$. As a consequence, given a tolerance $\epsilon > 0$, the number of iterations required to obtain $p_N \in \mathcal{M}$ such that $f(p_N) - f^* < \epsilon$, is bounded by

$$\eta L [L_0 d^2(p_0, q) + 2 (\mathcal{C}_{\rho, \kappa}^q - 1) [f(p_0) - f^*]] / (2L_0\epsilon) = \mathcal{O}(1/\epsilon).$$

Proof. Take $q \in \Omega^*$. After some simple algebraic manipulations and taking into account that $f^* = f(q)$ for each $q \in \Omega^*$, Lemma 6 becomes

$$2t_k (f(p_{k+1}) - f^*) \leq [d^2(p_k, q) - d^2(p_{k+1}, q)] + 2t_k [\mathcal{C}_{\rho, \kappa}^q - 1] [f(p_k) - f(p_{k+1})],$$

for all $k = 0, 1, \dots$. Using (9) and taking into account that $\mathcal{C}_{\rho, \kappa}^q \geq 1$, $f(p_{k+1}) - f^* \geq 0$ and $f(p_k) - f(p_{k+1}) \geq 0$, for all $k = 0, 1, \dots$, it follows that

$$\frac{2}{\eta L} [f(p_{k+1}) - f^*] \leq [d^2(p_k, q) - d^2(p_{k+1}, q)] + \frac{2}{L_0} [\mathcal{C}_{\rho, \kappa}^q - 1] [f(p_k) - f(p_{k+1})],$$

Summing both sides of the above inequality for $k = 0, 1, \dots, N-1$, we obtain

$$\frac{2}{\eta L} \sum_{i=0}^{N-1} [f(p_{i+1}) - f^*] \leq [d^2(p_0, q) - d^2(p_N, q)] + \frac{2}{L_0} [\mathcal{C}_{\rho, \kappa}^q - 1] [f(p_0) - f(p_N)].$$

Since $\{f(x_k)\}$ is a decreasing sequence, we conclude that

$$\frac{2}{\eta L} N (f(p_N) - f^*) \leq [d^2(p_0, q) - d^2(p_N, q)] + \frac{2}{L_0} [\mathcal{C}_{\rho, \kappa}^q - 1] [f(p_0) - f(p_N)],$$

which is equivalent to (15). The second statement of the theorem follows as an immediate consequence of the first part. \square

Whenever the Lipschitz constant $L > 0$ is computable, we can take a constant stepsize and Theorem 3 trivially implies the following result.

Theorem 4. *Let $\{p_k\}$ be generated by by Algorithm 1 with Strategy 1. Then, for every $N \in \mathbb{N}$, there holds*

$$f(p_N) - f^* \leq \frac{L d^2(p_0, q) + 2 (\mathcal{C}_{\rho, \kappa}^q - 1) [f(p_0) - f^*]}{2N}, \quad (16)$$

for each $q \in \Omega^*$. As a consequence, given a tolerance $\epsilon > 0$, the number of iterations required by the gradient method to obtain $p_N \in \mathcal{M}$ such that $f(p_N) - f^* < \epsilon$, is bounded by

$$[L d^2(p_0, q) + 2 (\mathcal{C}_{\rho, \kappa}^q - 1) [f(p_0) - f^*]] / (2\epsilon) = \mathcal{O}(1/\epsilon).$$

We remark that, if $\kappa = 0$ then $\mathcal{C}_{\rho, \kappa}^q = 1$. As a consequence, Theorem 4 merges into [6, Theorem 3.2].

Corollary 1. *Let $\{p_k\}$ be generated by Algorithm 1 with Strategy 1. Then, for every $N \in \mathbb{N}$, there holds*

$$\min \{ \|\text{grad } f(p_k)\| : k = 0, 1, \dots, N \} \leq \frac{2\sqrt{L [L d^2(p_0, q) + 2 (\mathcal{C}_{\rho, \kappa}^q - 1) [f(p_0) - f^*]]}}{N}, \quad (17)$$

for each $q \in \Omega^*$. As a consequence, given a tolerance $\epsilon > 0$, the number of iterations required by the gradient method to obtain $p_N \in \mathcal{M}$ such that $\|\text{grad } f(p_N)\| < \epsilon$, is bounded by $\mathcal{O}(\sqrt{L [L d^2(p_0, q) + 2 (\mathcal{C}_{\rho, \kappa}^q - 1) [f(p_0) - f^*]]}/\epsilon)$.

Proof. Let $N \in \mathbb{N}$. Using the notation $\lceil N/2 \rceil$ for the least integer that is greater than or equal to $N/2$, we have

$$f(p_{N+1}) - f^* + \sum_{j=\lceil N/2 \rceil}^N [f(p_j) - f(p_{j+1})] = f(p_{\lceil N/2 \rceil}) - f^*. \quad (18)$$

Thus, combining the last inequality with Theorem 4, we conclude that

$$f(p_{N+1}) - f^* + \sum_{j=\lceil N/2 \rceil}^N [f(p_j) - f(p_{j+1})] \leq \frac{L d^2(p_0, q) + 2 (\mathcal{C}_{\rho, \kappa}^q - 1) [f(p_0) - f^*]}{2\lceil N/2 \rceil}. \quad (19)$$

On the other hand, using Lemma 4 and considering that $t_k = 1/L$, we obtain

$$\frac{1}{2L} \sum_{j=\lceil N/2 \rceil}^N \|\text{grad } f(p_j)\|^2 \leq \sum_{j=\lceil N/2 \rceil}^N [f(p_j) - f(p_{j+1})] \leq f(p_{\lceil N/2 \rceil}) - f^*.$$

In view of $N/2 \leq \lceil N/2 \rceil$, the above inequality together with (18) and (19) yield

$$\frac{1}{2L} \sum_{j=\lceil N/2 \rceil}^N \|\text{grad } f(p_j)\|^2 \leq \frac{L d^2(p_0, q) + 2 (\mathcal{C}_{\rho, \kappa}^q - 1) [f(p_0) - f^*]}{N}.$$

Therefore,

$$\min \{ \|\text{grad } f(p_k)\|^2 ; k = \lceil N/2 \rceil, \dots, N \} \leq \frac{4L [L d^2(p_0, q) + 2 (\mathcal{C}_{\rho, \kappa}^q - 1) [f(p_0) - f^*]]}{N^2},$$

which implies the desired inequality. The second statement of the corollary follows as an immediate consequence of the first one. \square

We end this section by recalling an iteration-complexity bound for non-convex functions defined in a general Riemannian manifolds, which appeared in [8].

Theorem 5. *Let $\{p_k\}$ be generated by by Algorithm 1 with Strategy 1. Then, for every $N \in \mathbb{N}$, there holds*

$$\min \{ \|\text{grad } f(p_k)\| : k = 0, 1, \dots, N \} \leq \frac{\sqrt{2L(f(p_0) - f^*)}}{\sqrt{N+1}}.$$

As a consequence, given a tolerance $\epsilon > 0$, the number of iterations required to obtain $p_N \in \mathcal{M}$ such that $\|\text{grad } f(p_N)\| < \epsilon$ is bounded by $\mathcal{O}(L(f(p_0) - f^)/\epsilon^2)$.*

Under the assumption of convexity and lower boundedness of curvature, we conclude that Corollary 1 improves Theorem 5. It is worth to point out that results on iteration-complexity bound to the gradient method on Riemannian manifold with non-negative curvature and in Hadamard manifolds with lower bound curvature has already appeared [6, 38, 39]. The result of this section present a contribution to the systematic study of the iteration-complexity of the gradient methods in the Riemannian setting.

4 Examples

In the following sections, we present some examples of functions satisfying the assumptions of our results in the previous sections. In particular we show that, by endowing the constrained set with a suitable Riemannian metric, a constrained Euclidean optimization problem with non-convex objective function having non-Lipschitz gradient can be seen as unconstrained Riemannian optimization problem with convex objective function having Lipschitz gradient. *Throughout the next sections we denote*

$$\mathbb{R}_{++}^n := \{x := (x_1, \dots, x_n)^T \in \mathbb{R}^{n \times 1} : x_i > 0, i = 1, \dots, n\},$$

the positive orthant, \mathbb{P}^n the set of symmetric matrices of order $n \times n$ and \mathbb{P}_{++}^n the cone of symmetric positive definite matrices.

4.1 Examples in the positive orthant

In this section, we present examples in the positive orthant endowed with a new Riemannian metric. To present this examples we need some definitions and results of Riemannian geometry. Endowing \mathbb{R}_{++}^n with the Riemannian metric $\langle \cdot, \cdot \rangle$ defined by $\langle u, v \rangle := u^T G(x)v$, for all $x \in \mathbb{R}_{++}^n$ and $u, v \in T_x \mathbb{R}_{++}^n \equiv \mathbb{R}^n$, where $G : \mathbb{R}_{++}^n \rightarrow \mathbb{P}_{++}^n$ is given by

$$G(x) := \text{diag}(x_1^{-2}, \dots, x_n^{-2}) \in \mathbb{R}^{n \times n}, \quad (20)$$

we obtain a complete Riemannian manifold with zero curvature, which will be denoted by $\mathcal{M} := (\mathbb{R}_{++}^n, G)$. Let $f : \mathcal{M} \rightarrow \mathbb{R}$ be a twice differentiable function. We denote by

$f'(x)$ and $f''(x)$ the Euclidean gradient and hessian of f at x , respectively. Thus, (20) implies that the Riemannian *gradient* and *hessian* of f are given, respectively, by

$$\text{grad} f(x) = \text{diag}(x)^2 f'(x), \quad x \in \mathcal{M}, \quad (21)$$

$$\text{hess} f(x)v = [\text{diag}(x)^2 f''(x) + \text{diag}(x)\text{diag}(f'(x))]v, \quad x \in \mathcal{M}, \quad (22)$$

where $\text{diag}(x) := \text{diag}(x_1, \dots, x_n) \in \mathbb{R}^{n \times n}$. Next we present two examples of convex functions with Lipschitz gradient in $\mathcal{M} := (\mathbb{R}_{++}^n, G)$.

Example 2. Consider the function $f : \mathbb{R}_{++}^n \rightarrow \mathbb{R}$ defined by

$$f(x) := \sum_{i=1}^n f_i(x_i), \quad f_i(x_i) := -a_i e^{-b_i x_i} + c_i \ln(x_i)^2 + d_i \ln(x_i), \quad i = 1, \dots, n, \quad (23)$$

where $a_i, b_i, d_i \in \mathbb{R}_+$ and $c_i \in \mathbb{R}_{++}$ satisfy $c_i > a_i$. Since f is coercive, it has a minimum. By using (23) the first and second derivative of f at $x \in \mathbb{R}_{++}^n$ are given by $f'(x) := (f'_1(x_1), \dots, f'_n(x_n))$ and $f''(x) := \text{diag}(f''_1(x_1), \dots, f''_n(x_n))$, where

$$f'_i(x_i) = a_i b_i e^{-b_i x_i} + 2c_i \frac{\ln(x_i)}{x_i} + \frac{d_i}{x_i}, \quad f''_i(x_i) = -a_i b_i^2 e^{-b_i x_i} + 2c_i \left[\frac{1 - \ln(x_i)}{x_i^2} \right] - \frac{d_i}{x_i^2}, \quad (24)$$

for all $i = 1, \dots, n$. Note that $f''_i(1) < 0$, for all $i = 1, \dots, n$, and then f is not Euclidean convex. Using (22) and (24) the hessian of f in $\mathcal{M} := (\mathbb{R}_{++}^n, G)$ is given by

$$\text{Hess} f(x)v := \left(a_1 b_1 e^{-b_1 x_1} (x_1 - b_1 x_1^2) + 2c_1, \dots, a_n b_n e^{-b_n x_n} (x_n - b_n x_n^2) + 2c_n \right) v.$$

Since $c_i > a_i$, we have $a_i b_i e^{-b_i x_i} (x_i - b_i x_i^2) + 2c_i \geq 0$, for all $i = 1, \dots, n$. Hence, by using the definition of the metric, for $v = (v_1, \dots, v_n)^T \in \mathbb{R}^n$ and $x \in \mathbb{R}_{++}$, we have

$$\langle \text{Hess} f(x)v, v \rangle = \sum_{i=1}^n \left[a_i b_i e^{-b_i x_i} (x_i - b_i x_i^2) + 2c_i \right] \frac{v_i^2}{x_i^2} \geq 0,$$

concluding that f is convex in \mathcal{M} . Since $\|v\| = v^T G(x)v = 1$, we have $v_i^2 \leq x_i^2$ and owing that $(a_i b_i e^{-b_i x_i} (x_i - b_i x_i^2) + 2c_i) < a_i + 2c_i$, for all $i = 1, \dots, n$, we obtain

$$\|\text{Hess} f(x)v\|^2 = \sum_{i=1}^n \left[a_i b_i e^{-b_i x_i} (x_i - b_i x_i^2) + 2c_i \right]^2 \frac{v_i^2}{x_i^2} < \sum_{i=1}^n (a_i + 2c_i)^2, \quad x \in \mathbb{R}_{++}.$$

Therefore, (3) and Lemma 2 imply that $\text{grad} f$ is Lipschitz with $L < \sum_{i=1}^n (a_i + 2c_i)^2$.

Example 3. Consider the function $f : \mathbb{R}_{++}^n \rightarrow \mathbb{R}$ defined by

$$f(x) := \sum_{i=1}^n h_i(x_i), \quad h_i(x_i) := a_i \ln(x_i^{d_i} + b_i) - c_i \ln(x_i), \quad i = 1, \dots, n, \quad (25)$$

where $a_i, b_i, c_i, d_i \in \mathbb{R}_{++}$ satisfy $c_i < a_i d_i$ and $d_i \geq 2$, for all $i = 1, \dots, n$. The minimizer of f is $x^* = (x_1^*, \dots, x_n^*)$, where $x_i^* = \sqrt[d_i]{b_i c_i / (a_i d_i - c_i)}$, for $i = 1, \dots, n$. Function f in (25) is not Euclidean convex. However, by following the same steps as in the Example 2, we can show that f is convex and has gradient Lipschitz with constant $L < \sum_{i=1}^n a_i^2 d_i^4$ in $\mathcal{M} = (\mathbb{R}_{++}^n, G)$.

We end this section by presenting, without giving the details, two more examples of convex functions with Lipschitz gradients in $\mathcal{M} := (\mathbb{R}_{++}^n, G)$.

Remark 4. Let $a, b, c \in \mathbb{R}_{++}$. Define $h_1 : \mathbb{R}_{++}^n \rightarrow \mathbb{R}$ by $h_1(x) := a \ln(x^T x + b) - c \ln(x_1 \dots x_n)$, where $nc < 2a$, and $h_2 : \mathbb{R}_{++}^n \rightarrow \mathbb{R}$ by $h_2(x) = a \ln((x_1 \dots x_n)^2 + b) - c \ln(x_1 \dots x_n)$. By using similar arguments of Examples 2, we can prove that h_1 and h_2 are also convex with Lipschitz gradient in the Riemannian manifold $\mathcal{M} = (\mathbb{R}_{++}^n, G)$.

4.2 Examples in the SPD matrices cone

In this section, we present examples in the cone of symmetric positive definite matrices with new Riemannian metric. Following Rothaus [30], let $\mathcal{M} := (\mathbb{P}_{++}^n, \langle \cdot, \cdot \rangle)$ be the Riemannian manifold endowed with the Riemannian metric given by

$$\langle U, V \rangle := \text{tr}(VX^{-1}UX^{-1}), \quad X \in \mathcal{M}, \quad U, V \in T_X \mathcal{M}, \quad (26)$$

where $\text{tr}(X)$ denotes the trace of $X \in \mathbb{P}^n$ and $T_X \mathcal{M} \approx \mathbb{P}^n$. In fact, \mathcal{M} is a Hadamard manifold, see for example [19, Theorem 1.2. p. 325] and its curvature is bound below; see [20]. The *gradient* and *hessian* of $f : \mathbb{P}_{++}^n \rightarrow \mathbb{R}$ are given by

$$\text{grad} f(X) = X f'(X) X, \quad (27)$$

$$\text{hess} f(X)V = X f''(X)VX + \frac{1}{2} [V f'(X)X + X f'(X)V], \quad (28)$$

where $V \in T_X \mathcal{M}$, $f'(X)$ and $f''(X)$ are the Euclidean gradient and hessian of f at X , respectively. In the following, we present two examples of convex functions with Lipschitz gradient in $\mathcal{M} := (\mathbb{P}_{++}^n, \langle \cdot, \cdot \rangle)$.

Example 4. Consider the function $f : \mathbb{P}_{++}^n \rightarrow \mathbb{R}$ defined by

$$f(X) = a \ln(\det(X))^2 - b \ln(\det(X)). \quad (29)$$

where $a, b \in \mathbb{R}_{++}$. The Euclidean gradient and hessian of f are given, respectively, by

$$f'(X) = [2a \ln(\det(X)) - b] X^{-1}, \quad (30)$$

$$f''(X)V = 2a \text{tr}(X^{-1}V)X^{-1} - [2a \ln(\det(X)) - b] X^{-1}VX^{-1}, \quad (31)$$

for all $X \in \mathbb{P}_{++}^n$ and $V \in \mathbb{P}^n$. It follows from (30) that each $X \in \mathcal{M}$ satisfying $\det X = e^{b/(2a)}$ is a critical point of f . Thus, letting $V = I_n$ and $X = tI_n$ with $t \in \mathbb{R}_{++}$ in (31) we obtain that $f''(tI_n)I_n = [2ant^{-2} - 2an \ln t + b]I_n$. Thus $f''(tI_n)$ is not positive definite for t sufficiently large. Hence, f is not Euclidean convex. Moreover, f'' is not bounded and consequently f' is not Lipschitz. On the other hand, combining (28), (30) and (31), after some calculation we obtain

$$\text{Hess} f(X)V = 2a \text{tr}(X^{-1}V)X, \quad \langle \text{Hess} f(X)V, V \rangle = 2a \text{tr}(X^{-1}V)^2 \geq 0, \quad (32)$$

for all $X \in \mathcal{M}$ and $V \in T_X \mathcal{M}$. Thus, f is convex in \mathcal{M} . Moreover, (26) with (32) yield $\|\text{Hess } f(X)V\| = 2a\text{tr}(X^{-1}V)$, for all $X \in \mathcal{M}$ and $V \in T_X \mathcal{M}$. If we assume that $\|V\|^2 = \text{tr}(VX^{-1}VX^{-1}) = 1$ then $\text{tr}(X^{-1}V) \leq \sqrt{n}$. Hence,

$$\|\text{Hess } f(X)V\| \leq 2a\sqrt{n}, \quad X \in \mathcal{M}, \quad V \in T_X \mathcal{M}, \quad \|V\| = 1.$$

Therefore, (3) and Lemma 2 imply that $\text{grad } f$ is Lipschitz with constant $L \leq 2a\sqrt{n}$.

Example 5. Consider the function $f : \mathbb{P}_{++}^n \rightarrow \mathbb{R}$ defined by

$$f(X) = a \ln \left(\det(X)^{b_1} + b_2 \right) - c \ln(\det X), \quad (33)$$

where $a, b_1, b_2, c \in \mathbb{R}_{++}$ with $c < ab_1$. Function f in (33) is not Euclidean convex. On the other hand, by using similar arguments as in the Example 4, we can see that f is convex and has Lipschitz gradient with constant $L < ab_1^2 n$ in $\mathcal{M} = (\mathbb{P}_{++}^n, \langle \cdot, \cdot \rangle)$.

5 Numerical Experiments

In this section, we present some numerical experiments to illustrate the behavior of the Riemannian gradient method for minimizing convex functions onto the positive orthant and the cone of symmetric positive definite matrices. We implemented Algorithm 1 with Strategies 1, 2 and 3, and tested it on the examples of Section 4. Additionally, we consider the application of the method to compute the Riemannian center of mass, which is a specific instance of a geometric mean for points in a Riemannian manifold. In due course, we will describe this problem in more detail.

For the positive orthant, the *exponential mapping* $\exp_x : T_x \mathcal{M} \rightarrow \mathcal{M}$ in the Riemannian manifold $\mathcal{M} := (\mathbb{R}_{++}^n, G)$ is assigned by

$$\exp_x(v) = \left(x_1 e^{\frac{v_1}{x_1}}, \dots, x_n e^{\frac{v_n}{x_n}} \right), \quad (34)$$

for each $v := (v_1, \dots, v_n)^T \in \mathbb{R}^{n \times 1}$ and $x := (x_1, \dots, x_n)^T \in \mathbb{R}_{++}^n$, see [27]. By using the gradient in (21) and the definition of metric we obtain

$$\|\text{grad } f(x)\|^2 = \text{grad } f(x)^T G(x) \text{grad } f(x) = \sum_{i=1}^n \left[x_i \frac{\partial f}{\partial x_i}(x) \right]^2,$$

for each $x := (x_1, \dots, x_n) \in \mathcal{M}$. Considering the cone of symmetric positive definite matrices, the *exponential mapping* $\exp_X : T_X \mathcal{M} \rightarrow \mathcal{M}$ in the Riemannian manifold $\mathcal{M} := (\mathbb{P}_{++}^n, \langle \cdot, \cdot \rangle)$, is given by

$$\exp_X(V) = X^{1/2} e^{(X^{-1/2} V X^{-1/2})} X^{1/2}, \quad (35)$$

for each $V \in \mathbb{P}^n$ and $X \in \mathbb{P}_{++}^n$. By using (27), we have $\|\text{grad } f(X)\|^2 = \text{tr} \left([X f'(X)]^2 \right)$, for each $X \in \mathcal{M}$. In both cases, although (1) is a constrained optimization problem,

by (34) and (35), Algorithm 1 generates only feasible points without using projections or any other procedure to remain the feasibility. Hence, problem (1) can be seen as unconstrained Riemannian optimization problem.

Our numerical experience indicates that it is advantageous to perform a reasonably stringent line search. Therefore, we used $\beta = 1/2$ for Strategies 2 and 3. Additionally, we set $L_0 = 1$ and $\eta = 2$ for Strategy 2. We stopped the execution of Algorithm 1 at p_k declaring convergence if

$$\|f'(p_k)\|_\infty \leq 10^{-5}.$$

Since, by (21) and (27), $\text{grad}f(p_k) = 0$ if only if $f'(p_k) = 0$, this is a reasonable stopping criterion. The maximum number of allowed iterations was set to 1000. Codes are written in Matlab and are freely available at <https://orizon.mat.ufg.br/>.

5.1 Academic problems

We begin the numerical experiments by testing the Riemannian gradient method on the problems of minimizing the functions of the examples in Sections 4. We call these problems by Problem 1, 2, 3 and 4, respectively.

5.1.1 Academic problems in the positive orthant

In this section, we compare the performance of the Riemannian with the Euclidian gradient methods on Problems 1 and 2. We considered Algorithm 1 with Strategy 3 and implemented the Euclidian gradient method also using the Armijo rule with the same algorithmic parameters. It is worth mentioning that, in principle, the Euclidian method can generate iterates out of the positive orthant. Thus, in order to keep the feasibility, in each iteration we simply determine the maximum step size to remain within the feasible set and perform a convenient linear search by shrinking the step size until the Armijo condition is satisfied.

We generated several instances of Problems 1 and 2 by considering functions (23) and (25), respectively, with $n = 100$ and different parameters. In all cases, for each $i = 1, \dots, n$, we set parameters a_i with the same value. Equivalently for parameters b_i , c_i , and d_i .

Problem 1. First, parameters a_i , b_i , and d_i were randomly generated between 0 and 10. Then, in order to guarantee that $c_i > a_i$, we randomly generated parameters c_i between $1.1a_i$ and $5.0a_i$. All problems were solved 100 times using starting points from a uniform random distribution inside the box $[0, 20]^n$. For each method, Table 1 informs the percentage of runs that has reached a critical point (%), the average numbers of iterations (it) and functions evaluations (nfev) of the successful runs.

As can be seen, the Riemannian gradient method is clearly more efficient than the Euclidian gradient method in this set of problems. In *all* 18 problem instances considered, the Riemannian version required fewer iterations and function evaluations. Overall, on

#	a_i	b_i	c_i	d_i	Riemannian Gradient method			Euclidian Gradient method		
					%	it	nfev	%	it	nfev
1	3.77	8.17	11.10	5.92	100.0	14.1	85.5	100.0	72.3	255.4
2	7.88	5.49	17.95	3.01	100.0	21.1	148.5	100.0	56.9	208.2
3	8.96	1.72	42.11	7.18	100.0	17.0	137.1	100.0	56.0	203.3
4	3.14	1.30	13.77	9.32	100.0	9.0	55.0	100.0	76.3	232.6
5	5.49	1.72	6.82	0.83	100.0	10.0	51.0	100.0	65.1	227.3
6	4.59	4.25	13.31	8.11	100.0	11.0	67.0	100.0	71.2	228.5
7	2.10	3.80	4.31	0.10	100.0	21.2	107.0	100.0	54.1	184.7
8	8.69	7.47	28.54	4.77	100.0	8.0	57.1	100.0	61.1	255.3
9	9.85	2.24	44.60	0.57	100.0	16.0	129.0	100.0	52.0	201.9
10	2.60	1.71	9.65	2.07	100.0	18.0	109.2	100.0	57.1	185.6
11	6.03	1.40	13.57	8.94	100.0	9.0	55.0	100.0	79.2	238.1
12	5.71	4.99	9.37	3.22	100.0	20.1	121.7	100.0	59.2	191.2
13	1.38	6.07	6.78	4.86	100.0	9.0	46.1	100.0	73.5	219.6
14	2.22	0.24	5.58	9.04	100.0	14.0	71.0	100.0	141.8	408.6
15	4.19	6.24	7.73	9.48	100.0	7.0	36.0	100.0	105.8	315.4
16	8.27	2.42	10.96	3.02	100.0	17.0	103.0	100.0	66.3	237.4
17	4.72	0.64	19.35	0.62	100.0	18.0	127.0	100.0	55.6	204.1
18	2.99	1.63	11.15	6.44	100.0	14.0	85.1	100.0	75.8	250.8

Table 1: Parameters of function (23) as well as the performance of the Riemannian and Euclidian gradient methods.

average, the Riemannian gradient method performed 19.8% of iterations and 37.5% of function evaluations required by the Euclidian method.

Figure 1 (a) shows a typical behavior of the methods on Problem 1. This case corresponds to $n = 2$, $a_i = 1$, $b_i = c_i = d_i = 2$ for $i = 1, 2$, and the initial point $p_0 = [5, 1]^T$. The stopping criterion was satisfied with 4 and 14 iterations for the Riemannian and Euclidian gradient methods, respectively. The *zig-zag* path of the Euclidian gradient method can be seen clearly. In contrast, the Riemannian method rapidly approaches the minimizer. In Figure 1 (b), the sup-norm of the euclidean gradient is displayed as a function of the iteration number, which clearly shows the distinction between the methods. While the Euclidian method required 10 iterations for $\|f'(p_k)\|_\infty$ to reach order of 10^{-2} , the Riemannian algorithm required only 2 iterations.

Problem 2. We tested the algorithms on a set of 100 instances of Problem 2. We randomly generated parameters a_i and b_i between 0 and 10, parameters d_i between 2 and 10, and a constant μ_i belonging to the interval $(0, 1)$. Then, we set $c_i = \mu_i a_i d_i$, fulfilling the conditions $c_i < a_i d_i$ and $d_i \geq 2$, for all $i = 1, \dots, n$. As for Problem 1, each instance was solved 100 times using starting points from a uniform random distribution inside the box $[0, 20]^n$. The results are given in the following form: for each problem instance, Figure 2 (a) informs the average number of iterations, and Figure 2 (b) informs the average number of functions evaluations. As a matter of aesthetics, the graphs are independent and were organized in an increasing way.

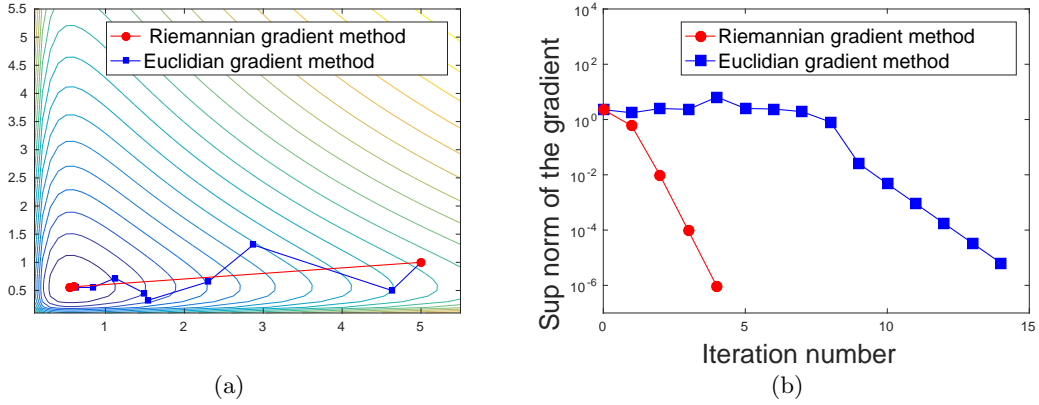


Figure 1: (a) A typical behavior of the Riemannian and the Euclidian gradient methods for which the *zigzag* pattern appears for the Euclidian algorithm. (b) Sup-norm of the euclidean gradients per iteration.

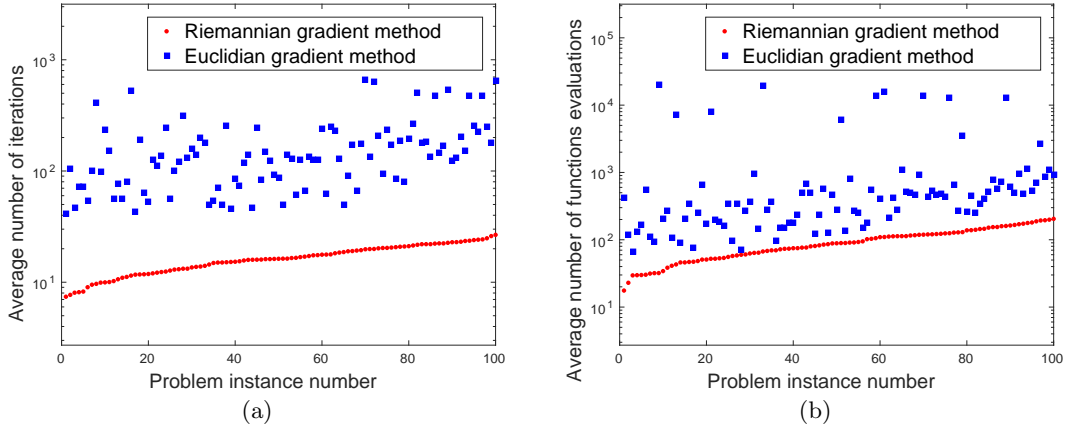


Figure 2: (a) Average number of iterations and (b) average number of functions evaluations required for each of 100 instances of Problem 2 for the Riemannian and the Euclidian gradient methods.

Figure 2 shows that the Riemannian gradient method required fewer iterations and function evaluations than the Euclidian gradient method in *all* problem instances. In terms of percentages, on average, the Riemannian algorithm performed 9.7% and 5.6% of the number of iterations and functions evaluations required by the Euclidian algorithm, respectively.

The results of this section allow us to conclude that there are problems for which the introduction of a suitable metric makes it possible to explore its geometric and alge-

braic structures, resulting in a large reduction in the computational cost of obtaining its solution. In fact, by introducing a suitable Riemannian metric, a constrained optimization problem with non-convex objective function and non-Lipschitz gradient can be transformed into an optimization problem with convex objective function and Lipschitz gradient.

5.1.2 Academic problems in the SPD matrices cone

In this section we illustrate the practical applicability of the Riemannian gradient method for minimizing convex functions onto the cone of symmetric positive definite matrices. We used Problem 3 to test the Riemannian gradient method varying the dimension and the domain of the starting points, while Problem 4 was used to compare the different linear search strategies. For Problem 3, we adopted Strategy 3.

Problem 3. We set $a = b = 1$ in function (29). In the first set of tests, we assigned the following values to the dimension: $n = 10, 20, 50, 100$, and 150 . For each specific value of n , we run the Riemannian gradient method 100 times using random starting points with eigenvalues belonging to the interval $(0, 20)$. In the second set of tests, we set $n = 50$ and varied the interval that contains the eigenvalues of the starting points. Again, for each combination, the method was run 100 times using random starting points. The results for the first and second set of tests are in Table 2 (a) and (b), respectively. First column of Table 2 (a) informs the considered dimension, while the first column of Table 2 (b) contains the interval for the eigenvalues of the starting points. Columns “%”, “it”, and “nfev” are as in Table 1.

n	%	it	nfev
10	100.0	18.2	110.2
20	100.0	19.9	140.4
50	100.0	14.2	114.9
100	100.0	15.2	138.2
150	100.0	27.1	271.5

(a)

$\lambda_i(X_0)$	%	it	nfev
(0 10)	98.0	14.2	114.6
(0 100)	99.0	14.6	117.6
(0 500)	99.0	15.0	121.0
(0 1000)	100.0	15.1	121.6
(0 2000)	100.0	15.2	122.4

(b)

Table 2: Performance of the Riemannian gradient method in Problem 3 varying: (a) the dimension; (b) the domain of the starting points.

The highlight of Table 2 is that the Riemannian gradient method was robust with respect to the dimension and to the choice of the starting points. Furthermore, except for the case where $n = 150$, it was not very sensitive to the variation of the dimension or to the domain of the starting points.

For comparative purposes, we implemented and tested the Euclidean method in Problem 3. For $n = 5$ (respectively, $n = 10$), 15 (respectively, 96) out of the 100 considered starting points resulted in an iteration history that reached the maximum number of iterations allowed. Finally, we observe that, by using (35) and the function (29), the

Riemannian and the Euclidian gradient iteration becomes, respectively,

$$X_{k+1} = \left[\det(X_k)^{2a} e^b \right]^{-t_k} X_k \quad k = 0, 1, \dots,$$

and

$$X_{k+1} = X_k - t_k [2a \ln(\det(X_k)) - b] X_k^{-1}, \quad k = 0, 1, \dots,$$

where the steep-size $t_k > 0$ is computed according to the adopted line search strategy. Thus, we can see that the Riemannian gradient iterations are simpler and have a lower computational cost to be performed.

Problem 4. We set $n = 100$, $a = b_1 = b_2 = 1$ and $c = 0.5$ in function (33), fulfilling $c < ab_1$. We tested the Riemannian gradient method with each of the three strategies by running each combination 100 times using random starting points with eigenvalues belonging to the interval $(0, 20)$. The results in Table 3 are given as in the previous tables.

Strategy 1			Strategy 2			Strategy 3		
%	it	nfev	%	it	nfev	%	it	nfev
100.0	452.5	453.5	99.0	15.3	21.3	100.0	15.3	70.9

Table 3: Performance of the Riemannian gradient method with the different line search strategies.

For Strategy 1, since the Lipschitz gradient constant satisfies $L < ab_1^2 n$, we used the Lipschitz stepsize $t_k = 1/(ab_1^2 n) < 1/L$, for all $k = 1, 2, \dots$. Overall, as can be seen in Table 3, the Riemannian method with Lipschitz stepsizes is clearly the least efficient, requiring an exceedingly large number of iterations. In this case the method performs one function evaluation per iteration. The poor performance is due to the short stepsizes in all iterations. On the other hand, we point out that the efficiency of the Riemannian gradient method with Lipschitz stepsize is closely related to an accurate estimate of the Lipschitz gradient constant.

Remark 2 helps to explain the results of Table 3 for Strategies 2 and 3. Regardless of the starting point, Algorithm 1 with both strategies performed exactly the same number of iterations. Additionally, in a typical run, the stepsizes were non-increasing. Therefore, overall, by Remark 2, the adaptive scheme in Strategy 2 required fewer function evaluations per iteration than the Armijo line search of Strategy 3.

Despite the simple linesearch mechanisms employed here, the numerical results indicate that, as it has to be expected, the efficient implementation of linear search algorithms can significantly improve the Riemannian gradient method.

5.2 The Riemannian center of mass

The Riemannian center of mass and so called Karcher mean is a specific instance of a geometric mean for points in Riemannian manifolds. It has several practical applications

and has appeared in many papers, we refer the reader to [7, 18, 33] and the references therein.

5.2.1 The center of mass on the SPD matrices cone

Denotes by $\|\cdot\|_F$ the Frobenius norm associated to the inner product $\langle U, V \rangle_F := \text{tr}(VU)$, for all $U, V \in \mathbb{P}_{++}^n$. Let d be the Riemannian distance defined in $\mathcal{M} := (\mathbb{P}_{++}^n, \langle \cdot, \cdot \rangle)$, i.e.,

$$d(A, X) = \left\| \ln \left(X^{-1/2} A X^{-1/2} \right) \right\|_F, \quad A, X \in \mathbb{P}_{++}^n, \quad (36)$$

see [27]. The Karcher mean of m positive definite matrices $A_1, \dots, A_m \in \mathbb{P}_{++}^n$ is the unique solution of the optimization problem

$$\min \left\{ f(x) := \frac{1}{2} \sum_{j=1}^m \left\| \ln \left(X^{-1/2} A_j X^{-1/2} \right) \right\|_F^2 : X \in \mathbb{P}_{++}^n \right\}. \quad (37)$$

Indeed, f is a strong convex function in \mathcal{M} due to the square of the distance (36) be strongly convex in \mathcal{M} , see for example [12]. Since f is a strong convex function, all sub-level sets of f are bounded. As a consequence, f has Riemannian Lipschitz gradient on each sublevel set of f . Finally, we remark that (36) is not an Euclidean convex function. By [18] and using (27), we conclude that

$$\text{grad } f(X) = \sum_{i=1}^m X^{1/2} \ln \left(X^{1/2} A_i^{-1} X^{1/2} \right) X^{1/2}. \quad (38)$$

Thus, by using (35) and (38), the Riemannian gradient iteration for solving (37) is

$$X_{k+1} = X_k^{1/2} e^{-t_k \sum_{i=1}^n \ln \left(X_k^{1/2} A_i^{-1} X_k^{1/2} \right)} X_k^{1/2}, \quad k = 0, 1, \dots$$

see, for example, [39].

In [2], Afsari *et al.* studied the convergence of the Riemannian gradient method with a Lipschitz stepsize for the center of mass problem in a manifold with curvature bounded from above and below. The stepsize is defined from a local estimate for the Lipschitz gradient constant. Consider problem (37), and let $r > 0$ be such that $A_1, \dots, A_m \subset B(X_0, r)$, where $B(X_0, r)$ is the open ball with center X_0 and radius r . They showed that it is possible to achieve convergence with $t_k = t$ for all $k = 0, 1, \dots$, where $t \in (0, 2\bar{t})$ and

$$\bar{t} = \frac{1}{4r \coth(4r)}. \quad (39)$$

Recently, Bento *et al.* [5] extended the convergence of the gradient method to the Hadamard setting for continuously differentiable functions which satisfy the Kurdyka-Lojasiewicz inequality. In particular, they proposed a Riemannian gradient method with Armijo line search for problem (37). Basically, their proposal coincides with Algorithm 1 with Strategy 3.

We tested Algorithm 1 with each strategy on a set of 200 randomly generated problems (37) with $n = 200$ and $m = 5, 10, 20$ or 50 . For each value of m we considered 50 problem instances. Let us clarify how a matrix A was defined. First, we randomly generated an orthonormal matrix U and a diagonal matrix D with elements belonging to $(0, 100)$. Then, we set $A = UDU^T$ ensuring that $A \in \mathbb{P}_{++}^n$. Given a problem instance with data $A_1, \dots, A_m \in \mathbb{P}_{++}^n$, we defined the starting point X_0 as the *explog* geometric mean given by

$$X_0 := \exp \left(\frac{1}{m} \sum_{i=1}^m \ln(A_i) \right),$$

see, for example, [3]. For Strategy 1 the Lipschitz stepsize t was defined according to [2]. We set $t = 1.99\bar{t}$, where \bar{t} is given by (39). Radius r can be calculated by computing the maximum distance of X_0 to each matrix A_i , $i = 1, \dots, m$. Numerical comparisons are reported in Figure 3 using performance profiles [14]. We adopted the number of functions evaluations and CPU time as performance measurements.

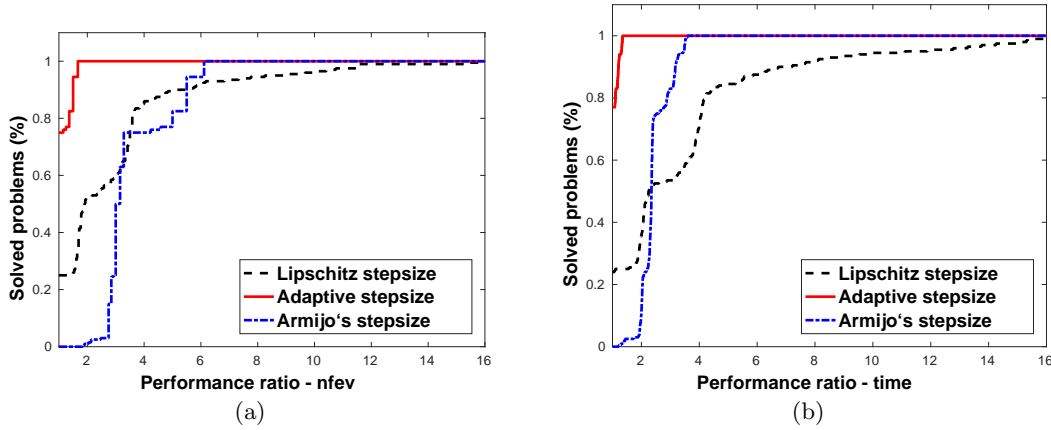


Figure 3: Performance profile comparing the Riemannian gradient method with different line search strategies using as performance measurement: (a) number of function evaluations; (b) CPU time.

As can be seen, Algorithm 1 with Strategy 2 is the most efficient on the chosen set of test problems. Efficiencies of the methods are 25.0% (respectively, 24.0%), 75.0% (respectively, 76.0%), and 0.0% (respectively, 0.0%) respectively, considering the number of function evaluations (respectively, CPU time) as performance measurement. Efficiency of Algorithm 1 with Strategy 3 is 0.0% because Strategy 2 outperformed Strategy 3 in all considered instances. Curiously, $m = 20$ in all problems for which Strategy 1 was the most efficient. Robustness are 99.5%, 100.0%, and 100.0% respectively, see Table 5.2.1. Only in a problem instance Algorithm 1 with Strategy 1 reached the maximum number of iterations allowed.

The similarity of the Figures 3 (a) and (b) suggests that the number of function evaluations is a good indicator of performance. Indeed, evaluating function f is compu-

	Efficiency (nfev – CPU time) (%)	Robustness (%)
Strategy 1	25.0 – 24.0	99.5
Strategy 2	75.0 – 76.0	100.0
Strategy 3	0.0 – 0.0	100.0

Table 4:

tationally expensive, since it involves inverting X and computing m matrix logarithms. This implies that linear search schemes must be carefully formulated for the center of mass problem. Overall, the naive implementation of the Armijo line search in Strategy 3 was overcome by the method with Lipschitz stepsize. On the other hand, the results indicate that the adaptive search proposed in Strategy 2 is a promising scheme worth to consider.

5.2.2 The center of mass on the positive orthant

Let $\mathcal{M} := (\mathbb{R}_{++}^n, G)$ be the Riemannian manifolds defined in Section 4.1 and d the Riemannian distance associated. Hence, we have

$$d^2(y, x) = \sum_{i=1}^n \ln^2 \left(\frac{y_i}{x_i} \right), \quad y = (y_1, \dots, y_n), \quad x = (x_1, \dots, x_n) \in \mathbb{R}_{++}^n. \quad (40)$$

The center of mass of m points $w^1, \dots, w^m \in \mathbb{R}_{++}^n$ is the unique solution of the optimization problem

$$\min \left\{ f(x) := \frac{1}{2} \sum_{j=1}^m d^2(w^j, x) : x \in \mathbb{R}_{++}^n \right\}. \quad (41)$$

Since the square of the distance (40) is strongly convex in \mathcal{M} , then f is a strong convex function in \mathcal{M} , see for example [12]. By using (21), we conclude that

$$\text{grad } f(x) = \left(x_1 \sum_{j=1}^m \ln \left(\frac{x_1}{w_1^j} \right), \dots, x_n \sum_{j=1}^m \ln \left(\frac{x_n}{w_n^j} \right) \right),$$

where $x = (x_1, \dots, x_n) \in \mathbb{R}_{++}^n$. Problem (41) has closed solution $x^* = (x_1^*, \dots, x_n^*) \in \mathbb{R}_{++}^n$ given by

$$x_i^* = \left(\prod_{j=1}^m w_i^j \right)^{\frac{1}{m}},$$

for all $i = 1, \dots, m$. Indeed, direct calculations show that $\text{grad } f(x^*) = 0$.

Due to the closed-form solution, we use problem (41) to illustrate the results on iteration-complexity bound of Section 3.2. We consider the Riemannian gradient algorithm with Lipschitz stepsize. Note that the set of positive definite diagonal matrices

can be identified with \mathbb{R}_{++}^n . Thus, problem (41) can be seen as a particular case of problem (37) for positive definite diagonal matrices. Given $w^1, \dots, w^m \in \mathbb{R}_{++}^n$ and defining $A_i = \text{diag}(w^i)$ for all $i = 1, \dots, m$, we defined the Lipschitz stepsize as in Section 5.2.2.

We set $n = 100$, $m = 5$ and randomly generated the elements of w^1, \dots, w^m and initial point x_0 from a uniform distribution on $(0, 100)$. The computed Lipschitz stepsize was set to $t \approx 0.06$. The Riemannian gradient algorithm stopped declaring “solution was found” with 30 iterations. Figures 4 (a) and (b) report the function values of the left and right hand sides of inequalities (16) and (17), respectively.

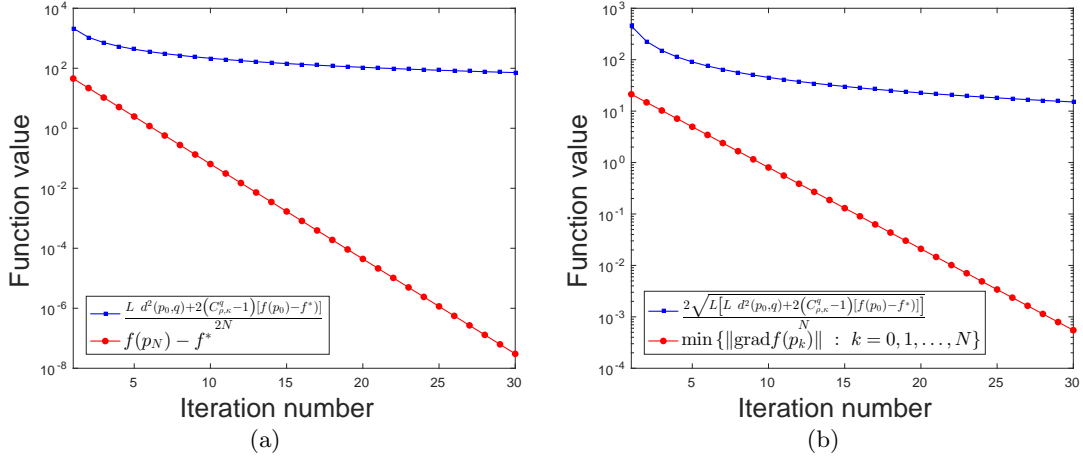


Figure 4: Iteration-complexity bound for the Riemannian gradient method with Lipschitz stepsize related to: (a) objective function value – Theorem 4 ; (b) norm of the Riemannian gradient – Corollary 1.

As can be seen in Figure 4, the iteration-complexity bounds related to the objective function value and the norm of the Riemannian gradient are always respected, see Theorem 4 and Corollary 1. This illustrates the practical reliability of our iteration-complexity results.

6 Conclusions

In this paper, the behavior of the gradient method for convex optimization problems on Riemannian manifolds with lower bounded sectional curvature were analyzed. We considered three different finite procedures for determining the stepsize, namely, constant stepsize, adaptive procedure and Armijo’s procedure. As far as we know, the full convergence of the sequence generated by this method with these three strategies is a new contribution of this paper, which adds important results in the available convergence theory. Besides, under mild assumptions, we showed that the iteration-complexity bound related to the method is $\mathcal{O}(1/\epsilon)$ for finding a point $p_N \in \mathcal{M}$ such that $f(p_N) - f^* < \epsilon$. The numerical experiments provided illustrate the effectiveness of

the method in this new setting and certify the conclusions suggested by the theoretical results. Despite the simple linesearch mechanisms employed here, the numerical results indicate that, as it has to be expected, the efficient implementation of linear search algorithms can significantly improve the Riemannian gradient method. In particular, the effectiveness of the method to find the Riemannian mass center and the so-called Karcher’s mean is presented, indicating that the adaptive procedure is a promising scheme that is worth considering. We expect that this paper will contribute to the development of studies of optimization methods in the Riemannian setting. Finally, it would be interesting to analyze stochastic versions of the the gradient method by using adaptive procedures.

References

- [1] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, Princeton, NJ, 2008. With a foreword by Paul Van Dooren.
- [2] B. Afsari, R. Tron, and R. Vidal. On the convergence of gradient descent for finding the Riemannian center of mass. *SIAM J. Control Optim.*, 51(3):2230–2260, 2013.
- [3] T. Ando, C.-K. Li, and R. Mathias. Geometric means. *Linear Algebra and its Applications*, 385:305–334, 2004.
- [4] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):183–202, 2009.
- [5] G. C. Bento, S. D. B. Bitar, J. X. Cruz Neto, P. R. Oliveira, and J. C. Souza. The steepest descent method for computing riemannian center of mass on hadamard manifolds. *Technical report, submitted*, 2017.
- [6] G. C. Bento, O. P. Ferreira, and J. G. Melo. Iteration-Complexity of Gradient, Subgradient and Proximal Point Methods on Riemannian Manifolds. *J. Optim. Theory Appl.*, 173(2):548–562, 2017.
- [7] D. A. Bini and B. Iannazzo. Computing the Karcher mean of symmetric positive definite matrices. *Linear Algebra Appl.*, 438(4):1700–1710, 2013.
- [8] N. Boumal, P.-A. Absil, and C. Cartis. Global rates of convergence for nonconvex optimization on manifolds. *ArXiv e-prints*, 1(1):1–31, 2016.
- [9] R. Burachik, L. M. G. Drummond, A. N. Iusem, and B. F. Svaiter. Full convergence of the steepest descent method with inexact line searches. *Optimization*, 32(2):137–146, 1995.
- [10] J. da Cruz Neto, L. De Lima, and P. Oliveira. Geodesic algorithms in riemannian geometry. *Balkan J. Geom. Appl.*, 3(2):89–100, 1998.

- [11] J. X. da Cruz Neto, L. L. de Lima, and P. R. Oliveira. Geodesic algorithms in Riemannian geometry. *Balkan J. Geom. Appl.*, 3(2):89–100, 1998.
- [12] J. X. da Cruz Neto, O. P. Ferreira, and L. R. Lucambio Pérez. Contributions to the study of monotone vector fields. *Acta Math. Hungar.*, 94(4):307–320, 2002.
- [13] M. P. do Carmo. *Riemannian geometry*. Mathematics: Theory & Applications. Birkhäuser Boston, Inc., Boston, MA, 1992. Translated from the second Portuguese edition by Francis Flaherty.
- [14] E. D. Dolan and J. J. Moré. Benchmarking optimization software with performance profiles. *Mathematical programming*, 91(2):201–213, 2002.
- [15] A. Edelman, T. A. Arias, and S. T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl.*, 20(2):303–353, 1999.
- [16] O. P. Ferreira, A. N. Iusem, and S. Z. Németh. Concepts and techniques of optimization on the sphere. *TOP*, 22(3):1148–1170, 2014.
- [17] D. Gabay. Minimizing a differentiable function over a differential manifold. *J. Optim. Theory Appl.*, 37(2):177–219, 1982.
- [18] B. Jeuris, R. Vandebril, and B. Vandereycken. A survey and comparison of contemporary algorithms for computing the matrix geometric mean. *Electron. Trans. Numer. Anal.*, 39:379–402, 2012.
- [19] S. Lang. *Fundamentals of differential geometry*, volume 191 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1999.
- [20] C. Lenglet, M. Rousson, R. Deriche, and O. Faugeras. Statistics on the manifold of multivariate normal distributions: theory and application to diffusion tensor MRI processing. *J. Math. Imaging Vision*, 25(3):423–444, 2006.
- [21] C. Li, B. S. Mordukhovich, J. Wang, and J.-C. Yao. Weak sharp minima on Riemannian manifolds. *SIAM J. Optim.*, 21(4):1523–1560, 2011.
- [22] C. Li and J.-C. Yao. Variational inequalities for set-valued vector fields on Riemannian manifolds: convexity of the solution set and the proximal point algorithm. *SIAM J. Control Optim.*, 50(4):2486–2514, 2012.
- [23] D. G. Luenberger. The gradient projection method along geodesics. *Management Sci.*, 18:620–631, 1972.
- [24] J. H. Manton. A framework for generalising the Newton method and other iterative methods from Euclidean space to manifolds. *Numer. Math.*, 129(1):91–125, 2015.
- [25] Y. Nesterov. *Introductory lectures on convex optimization*, volume 87 of *Applied Optimization*. Kluwer Academic Publishers, Boston, MA, 2004. A basic course.

- [26] Y. Nesterov. Gradient methods for minimizing composite functions. *Math. Program.*, 140(1, Ser. B):125–161, 2013.
- [27] Y. E. Nesterov and M. J. Todd. On the Riemannian geometry defined by self-concordant barriers and interior-point methods. *Found. Comput. Math.*, 2(4):333–361, 2002.
- [28] T. Rapcsák. *Smooth nonlinear optimization in \mathbf{R}^n* , volume 19 of *Nonconvex Optimization and its Applications*. Kluwer Academic Publishers, Dordrecht, 1997.
- [29] M. Raydan. The Barzilai and Borwein gradient method for the large scale unconstrained minimization problem. *SIAM J. Optim.*, 7(1):26–33, 1997.
- [30] O. S. Rothaus. Domains of positivity. *Abh. Math. Sem. Univ. Hamburg*, 24:189–235, 1960.
- [31] T. Sakai. *Riemannian geometry*, volume 149 of *Translations of Mathematical Monographs*. American Mathematical Society, Providence, RI, 1996. Translated from the 1992 Japanese original by the author.
- [32] S. T. Smith. Optimization techniques on Riemannian manifolds. In *Hamiltonian and gradient flows, algorithms and control*, volume 3 of *Fields Inst. Commun.*, pages 113–136. Amer. Math. Soc., Providence, RI, 1994.
- [33] S. Sra and R. Hosseini. Conic geometric optimization on the manifold of positive definite matrices. *SIAM J. Optim.*, 25(1):713–739, 2015.
- [34] C. Udrişte. *Convex functions and optimization methods on Riemannian manifolds*, volume 297 of *Mathematics and its Applications*. Kluwer Academic Publishers Group, Dordrecht, 1994.
- [35] X. Wang, C. Li, J. Wang, and J.-C. Yao. Linear convergence of subgradient algorithm for convex feasibility on Riemannian manifolds. *SIAM J. Optim.*, 25(4):2334–2358, 2015.
- [36] X. M. Wang, C. Li, and J. C. Yao. Subgradient projection algorithms for convex feasibility on Riemannian manifolds with lower bounded curvatures. *J. Optim. Theory Appl.*, 164(1):202–217, 2015.
- [37] Y.-x. Yuan. Step-sizes for the gradient method. In *Third International Congress of Chinese Mathematicians. Part 1, 2*, volume 2 of *AMS/IP Stud. Adv. Math.*, 42, pt. 1, pages 785–796. Amer. Math. Soc., Providence, RI, 2008.
- [38] H. Zhang, S. J. Reddi, and S. Sra. Fast stochastic optimization on Riemannian manifolds. *ArXiv e-prints*, pages 1–17, 2016.
- [39] H. Zhang and S. Sra. First-order methods for geodesically convex optimization. *JMLR: Workshop and Conference Proceedings*, 49(1):1–21, 2016.