Adv Comput Math (2016) 42:333–360
DOI 10.1007/s10444-015-9426-z

ε-subgradient algorithms for locally lipschitz functions on Riemannian manifolds

P. Grohs[1] · **S. Hosseini**[1]

Received: 17 October 2014 / Accepted: 27 June 2015 /
Published online: 11 July 2015

Abstract This paper presents a descent direction method for finding extrema of locally Lipschitz functions defined on Riemannian manifolds. To this end we define a set-valued mapping $x \to \partial_\varepsilon f(x)$ named ε-subdifferential which is an approximation for the Clarke subdifferential and which generalizes the Goldstein-ε-subdifferential to the Riemannian setting. Using this notion we construct a steepest descent method where the descent directions are computed by a computable approximation of the ε-subdifferential. We establish the global convergence of our algorithm to a stationary point. Numerical experiments illustrate our results.

Keywords Riemannian manifolds · Lipschitz function · Descent direction · Clarke subdifferential

Mathematics Subject Classifications (2010) 49J52 · 65K05 · 58C05

1 introduction

This paper is concerned with the numerical solution of optimization problems defined on Riemannian manifolds where the objective function may be nonsmooth. Such problems arise in a variety of applications, e.g., in computer vision, signal processing,

Communicated by: A. Zhou

✉ S. Hosseini
seyedehsomayeh.hosseini@math.ethz.ch

P. Grohs
pgrohs@sam.math.ethz.ch

[1] ETH Zürich, Seminar for Applied Mathematics, Rämistrasse 101, 8092 Zürich, Switzerland

motion and structure estimation, or numerical linear algebra; see for instance [2, 3, 27, 35].

In the linear case is well known that ordinary gradient descent, when applied to nonsmooth functions, typically fails by converging to a non-optimal point. The fundamental difficulty is that most interesting nonsmooth objective functions assume their extrema at points where the gradient is not defined. This has led to the introduction of the generalized gradient of convex functions defined on a linear space by Rockafellar in 1961 and subsequently for locally Lipschitz functions by Clarke in 1975; [13, 36]. Their use in optimization algorithms began soon after their appearance. Since the Clarke generalized gradient is in general difficult to compute numerically, most of algorithms which are based on it can be efficient only for certain types of functions; see for instance [7, 23, 45].

The paper [19] is among the first works on optimization of Lipschitz functions on Euclidean spaces. In that article a new set valued mapping named $\varepsilon-$subdifferential $\partial_\varepsilon f$ of a function f was introduced, and several properties of this map, which are useful for building optimization algorithms of locally Lipschitz functions on linear spaces, were presented. For the numerical computation of the $\varepsilon-$subdifferential various strategies have been proposed in the literature.

The gradient sampling algorithm (GS), introduced and analyzed by Burke, Lewis and Overton [11, 12], is a method for minimizing an objective function f that is locally Lipschitz and continuously differentiable in an open dense subset of $\mathbb{R}^n$. At each iteration, the GS algorithm computes the gradient of f at the current iterate and at $m \geq n + 1$ randomly generated nearby points. This bundle of gradients is used to find an approximate ε-steepest descent direction as the solution of a quadratic program, where ε denotes the sampling radius. A standard Armijo line search along this direction produces a candidate for the next iterate, which is obtained by perturbing the candidate, if necessary, to stay in the set Ω where f is differentiable; the perturbation is random and small enough to maintain the Armijo sufficient descent property. The sampling radius may be fixed for all iterations or may be reduced dynamically.

The discrete gradient method (DG) approximates $\partial_\varepsilon f(x)$ by a set of discrete gradients. In this algorithm, the descent direction is iteratively computed, and in every iteration the approximation of $\partial_\varepsilon f(x)$ is improved by adding a discrete gradient to the set of discrete gradients; see [7].

In [31], $\partial_\varepsilon f(x)$ is approximated by an iterative algorithm. The algorithm starts with one element of $\partial_\varepsilon f(x)$ in the first iteration, and in every subsequent iteration, a new element of $\partial_\varepsilon f(x)$ is computed and added to the working set to improve the approximation of $\partial_\varepsilon f(x)$. The results of the algorithm presented in [31] as compared to those obtained by the GS is more efficient, and as compared to those by the DG is more accurate, [31].

The extension of the aforementioned optimization techniques to Riemannian manifolds are the subject of the present paper. A manifold, in general, does not have a linear structure, hence the usual techniques, which are often used to study optimization problems on linear spaces cannot be applied and new techniques need to be developed.

The development of smooth and nonsmooth Riemannian optimization algorithms is primarily motivated by their large-scale applications in robust, sparse, structured principal component analysis, statistics on manifolds (e.g. median calculation of positive semidefinite tensors), and low-rank optimization (matrix completion, collaborative filtering, source separation); see [25, 41–43]. Furthermore, these algorithms have a lot of applications in image processing, computer vision, constrained optimization problems on linear spaces; [4, 10, 16].

Contributions Our main contributions are twofold. First, we define a Riemannian generalization of the ε-subdifferential defined in [19]. This is nontrivial since the linear definition of $\partial_\varepsilon f(x)$, $x \in \mathbb{R}^n$ involves subgradients of f at points $y \in \mathbb{R}^n$ different from x. In the linear case this is not an issue since tangent spaces at different points can be identified. In the nonlinear case, with M being a Riemannian manifold and $f : M \to \mathbb{R}$, we move these subgradients at points $y \in M$ to the tangent space in x via the derivative of the logarithm mapping in order to obtain a workable definition of the ε-subdifferential; see Definition 3.1 below. In Section 3.1, we prove several basic properties of the novel Riemannian ε-subdifferential which subsequently enables us to formulate conditions for descent directions in Section 3.2. Using these basic properties of the ε-subdifferential, we are able to generalize (GS) and the algorithm in [31] to the Riemannian setting. In Section 3.3, we present the details for the generalization of [31] which yields the second main contribution of the present paper, namely a proof of global convergence of the proposed algorithm. Finally, our proposed algorithm is implemented in MATLAB environment and applied to some nonsmooth problems with locally Lipschitz objective functions.

Previous work For the optimization of smooth objective functions many classical methods for unconstrained minimization, such as Newton-type and trust-region methods have been successfully generalized to problems on Riemannian manifolds [1, 3, 14, 30, 34, 39, 40, 46]. The recent monograph by Absil, Mahony and Sepulchre discusses, in a systematic way, the framework and many numerical first-order and second-order manifold-based algorithms for minimization problems on Riemannian manifolds with an emphasis on applications to numerical linear algebra, [2].

In considering optimization problems with nonsmooth objective functions on Riemannian manifolds, it is necessary to generalize concepts of nonsmooth analysis to Riemannian manifolds. In the past few years a number of results have been obtained on numerous aspects of nonsmooth analysis on Riemannian manifolds, [5, 6, 20–22, 29].

Recently, some mathematicians have started developing nonsmooth optimization algorithms to manifold settings. It is worth noting that while they presented gradient based and proximal point algorithms on manifolds, their numerical experiments are limited to some special test functions whose subdifferential either are singleton or can be computed explicitly. This might be because of the difficulty of finding the subdifferential of the functions; see [8, 9, 15, 17, 33]. Finally, it is worth mentioning the paper [16], which presents a survey on Riemannian geometry methods for smooth and nonsmooth constrained optimization as well as gradient and subgradient descent algorithms on a Riemannian manifold. In that paper, the methods are illustrated by applications from robotics and multi antenna communication.

2 Preliminaries

In this paper, we use the standard notations and known results of Riemannian manifolds; see, e.g. [26]. Throughout this paper, M is an n-dimensional complete manifold endowed with a Riemannian metric $\langle ., . \rangle$ on the tangent space T_xM. We identify (via the Riemannian metric) the tangent space of M at a point x, denoted by T_xM, with the cotangent space at x, denoted by T_xM^*. As usual we denote by $B(x, \delta)$ the open ball centered at x with radius δ, by $\text{int}N(\text{cl}N)$ the interior (closure) of the set N. Also, let S be a nonempty closed subset of a Riemannian manifold M, we define $\text{dist}_S : M \longrightarrow \mathbb{R}$ by

$$\text{dist}_S(x) := \inf\{\text{dist}(x, s) : s \in S\},$$

where dist is the Riemannian distance on M. Recall that the set S in a Riemannian manifold M is called convex if every two points $p_1, p_2 \in S$ can be joined by a unique geodesic whose image belongs to S. For the point $x \in M$, $\exp_x : U_x \to M$ will stand for the exponential function at x, where U_x is an open subset of T_xM. Recall that $\exp_x$ maps straight lines of the tangent space T_xM passing through $0_x \in T_xM$ into geodesics of M passing through x.

We will also use the parallel transport of vectors along geodesics. Recall that, for a given curve $\gamma : I \to M$, number $t_0 \in I$, and a vector $V_0 \in T_{\gamma(t_0)}M$, there exists a unique parallel vector field $V(t)$ along $\gamma(t)$ such that $V(t_0) = V_0$. Moreover, the map defined by $V_0 \mapsto V(t_1)$ is a linear isometry between the tangent spaces $T_{\gamma(t_0)}M$ and $T_{\gamma(t_1)}M$, for each $t_1 \in I$. In the case when γ is a minimizing geodesic and $\gamma(t_0) = x$, $\gamma(t_1) = y$, we will denote this map by L_{xy}, and we will call it the parallel transport from T_xM to T_yM along the curve γ. Note that, L_{xy} is well defined when the minimizing geodesic which connects x to y, is unique. For example, the parallel transport L_{xy} is well defined when x and y are contained in a convex neighborhood. In what follows, L_{xy} will be used wherever it is well defined. The isometry L_{yx} induces another linear isometry L^*_{yx} between T_xM^* and T_yM^*, such that for every $\sigma \in T_xM^*$ and $v \in T_yM$, we have $\langle L^*_{yx}(\sigma), v\rangle = \langle \sigma, L_{yx}(v)\rangle$. We will still denote this isometry by $L_{xy} : T_xM^* \to T_yM^*$.

By $i_M(x)$ we denote the injectivity radius of M at x, that is the supremum of the radius r of all balls $B(0_x, r)$ in T_xM for which $\exp_x$ is a diffeomorphism from $B(0_x, r)$ onto $B(x, r)$. Note that if U is a compact subset of a Riemannian manifold M and $i(U) := \inf\{i_M(x) : x \in U\}$, then $0 < i(U)$; see [24].

A retraction on a manifold M is a continuously differentiable map $R : TM \to M$ with the following properties. Let R_x denote the restriction of R to T_xM.

- $R_x(0_x) = x$, where 0_x denotes the zero element of T_xM.

- With the canonical identification $T_{0_x}T_xM \approx T_xM$, $dR_x(0_x) = id_{T_xM}$ where id_{T_xM} denotes the identity map on T_xM.

In the present paper, we are concerned with the minimization of locally Lipschitz functions which we now define.

Definition 2.1 (Lipschitz condition) Recall that a real valued function f defined on a Riemannian manifold M is said to satisfy a Lipschitz condition of rank k on a given subset S of M if $\mid f(x) - f(y) \mid \leq k\mathrm{dist}(x, y)$ for every $x, y \in S$, where dist is the Riemannian distance on M. A function f is said to be Lipschitz near $x \in M$ if it satisfies the Lipschitz condition of some rank on an open neighborhood of x. A function f is said to be locally Lipschitz on M if f is Lipschitz near x, for every $x \in M$.

Let us continue with the definition of the Clarke generalized directional derivative for locally Lipschitz functions on Riemannian manifolds; see [20, 22].

Definition 2.2 (Clarke generalized directional derivative) Suppose $f : M \to \mathbb{R}$ is a locally Lipschitz function on a Riemannian manifold M. Let $\phi_x : U_x \to T_xM$ be an exponential chart at x. Given another point $y \in U_x$,consider $\sigma_{y,v}(t) := \phi_y^{-1}(tw)$, a geodesic passing through y with derivative w, where (ϕ_y, y) is an exponential chart around y and $d(\phi_x \circ \phi_y^{-1})(0_y)(w) = v$. Then, the Clarke generalized directional derivative of f at $x \in M$ in the direction $v \in T_xM$, denoted by $f^\circ(x, v)$, is defined as

$$f^\circ(x; v) = \limsup_{y \to x, t \downarrow 0} \frac{f(\sigma_{y,v}(t)) - f(y)}{t}.$$

If f is differentiable in $x \in M$, we define the gradient of f as the unique vector $\mathrm{grad}\, f(x) \in T_xM$ which satisfies

$$\langle \mathrm{grad}\, f(x), \xi \rangle = df(x)(\xi) \quad \text{for all } \xi \in T_xM.$$

Using the previous definition of a Riemannian Clarke generalized directional derivative we can also generalize the notion of the subdifferential to a Riemannian context.

Definition 2.3 (Subdifferential) We define the subdifferential of f at x, denoted by $\partial f(x)$, as the subset of T_xM with support function given by $f^\circ(x; .)$, i.e., for every $v \in T_xM$,

$$f^\circ(x; v) = \sup \{\langle \xi, v \rangle : \xi \in \partial f(x)\}.$$

It can be proved [20] that

$$\partial f(x) = \mathrm{conv} \left\{ \lim_{i \to \infty} \mathrm{grad}\, f(x_i) : \{x_i\} \subseteq \Omega_f, x_i \to x \right\},$$

where Ω_f is a dense subset of M on which f is differentiable.

It is worthwhile to mention that $\lim \mathrm{grad}\, f(x_i)$ in the previous definition is obtained as follows. Let $\xi_i \in T_{x_i}M$, $i = 1, 2, \ldots$ be a sequence of tangent vectors of M and $\xi \in T_xM$. We say ξ_i converges to ξ, denoted by $\lim \xi_i = \xi$, provided that $x_i \to x$ and, for any smooth vector field X, $\langle \xi_i, X(x_i)\rangle \to \langle \xi, X(x)\rangle$.

Using the notion of subdifferential, we can now define stationary points of a locally Lipschitz mapping f.

Definition 2.4 (Stationary point, Stationary set) A point x is a stationary point of f if $0 \in \partial f(x)$. Z is a stationary set if each $z \in Z$ is a stationary point.

Proposition 2.5 *A necessary condition that f achieve a local minimum at x is that* $0 \in \partial f(x)$.

Proof If f has a local minimum at x, then for every $v \in T_xM$, $f^\circ(x; v) \geq 0$ which implies $0 \in \partial f(x)$. □

3 The Riemannian ε-subdifferential

In smooth optimization, there exist minimization methods, which, instead of using the gradient, use its approximations through finite differences (forward, backward, and central differences). In [28], a very simple convex nondifferentiable function was presented, for which these finite differences may give no information about the subdifferential. It follows that these finite-difference estimates of the gradient cannot be used for the approximation of the subgradient of the nonsmooth functions. In [19] a set valued mapping named ε-subdifferential, to approximate the subdifferential of locally Lipschitz functions defined on $\mathbb{R}^n$ was introduced.

The present section generalizes the concept of the ε-subdifferential of locally Lipschitz functions to functions defined on a Riemannian manifold, generalizing the corresponding Euclidean concept introduced in [19]. The definition is as follows.

Definition 3.1 (ε-subdifferential) Let $f : M \to \mathbb{R}$ be a locally Lipschitz function on a Riemannian manifold M, and θ_k be any sequence of positive numbers converging downward to zero. For each $\varepsilon > 0$ with $\varepsilon + \theta_k < i_M(x)$ for almost every k, the $\varepsilon-$subdifferential at x is defined by

$$\partial_\varepsilon f(x) := \operatorname{conv} \bigcap_k \operatorname{cl} \left\{ d\exp_x^{-1}(y)(\operatorname{grad} f(y)) : y \in \operatorname{cl}B(x, \varepsilon + \theta_k) \cap \Omega_f \right\},$$

where the intersection is taken over all k for which $\varepsilon + \theta_k < i_M(x)$.

Clearly this definition is independent of the choice of the sequence θ_k.

3.1 Basic properties

In the present subsection, we establish some basic properties of the ε-subdifferential as defined above in Definition 3.1; see [19] for similar results in the linear case. We select ε small enough that f is Lipschitz on $B(x, 2\varepsilon)$ and $\exp_x$ is a diffeomorphism from $B(0_x, 2\varepsilon)$ onto $B(x, 2\varepsilon)$.

Lemma 3.2 *For every $y \in B(x, \varepsilon)$,*

$$d\exp_x^{-1}(y)(\partial f(y)) \subset \partial_\varepsilon f(x).$$

Proof For every $\xi = \lim_{i\to\infty} \operatorname{grad} f(y_i)$, where $\operatorname{grad} f(y_i)$ exists and $y_i \to y$, we have

$$d\exp_x^{-1}(y)(\xi) = \lim_{i\to\infty} d\exp_x^{-1}(y_i)(\operatorname{grad} f(y_i)),$$

hence there exists $N \in \mathbb{N}$, such that for every $i \geq N$, $y_i \in B(x, \varepsilon)$ and

$$\lim_{i\to\infty} d\exp_x^{-1}(y_i)(\operatorname{grad} f(y_i)) \in \bigcap_k \operatorname{cl}\left\{d\exp_x^{-1}(y)(\operatorname{grad} f(y)) : y \in \operatorname{cl}B(x, \varepsilon + \theta_k) \cap \Omega_f\right\},$$

which means

$$d\exp_x^{-1}(y)\left(\left\{\lim_{i\to\infty} \operatorname{grad} f(y_i) : \{y_i\} \subseteq \Omega_f, y_i \to y\right\}\right)$$

is a subset of $\bigcap_k \operatorname{cl}\{d\exp_x^{-1}(y)(\operatorname{grad} f(y)) : y \in \operatorname{cl}B(x, \varepsilon+\theta_k)\cap\Omega_f\}$, which implies

$$d\exp_x^{-1}(y)(\partial f(y)) \subset \partial_\varepsilon f(x).$$

□

Lemma 3.3 *$\partial_\varepsilon f(x)$ is a nonempty compact and convex subset of T_xM.*

Proof By the Lipschitzness of f and the smoothness of the exponential map,

$$S_k = \operatorname{cl}\left\{d\exp_x^{-1}(y)(\operatorname{grad} f(y)) : y \in \operatorname{cl}B(x, \varepsilon + \theta_k) \cap \Omega_f\right\}$$

is a closed bounded subset of T_xM and $S_{k+1} \subset S_k$. Hence $\bigcap_k S_k$ is compact and nonempty, and convex hull of a compact set in T_xM is compact. The convexity of $\partial_\varepsilon f(x)$ is deduced by the definition. □

Lemma 3.4

$$\partial_\varepsilon f(x) = \operatorname{conv}\left\{\lim_{i\to\infty} d\exp_x^{-1}(y_i)(\operatorname{grad} f(y_i)) : \lim_{i\to\infty} y_i = y \in clB(x, \varepsilon),\ (y_i) \in \Omega_f\right\}.$$

Proof We start with the inclusion

$$\partial_\varepsilon f(x) \supset \operatorname{conv}\left\{\lim_{i\to\infty} d\exp_x^{-1}(y_i)(\operatorname{grad} f(y_i)) : \lim_{i\to\infty} y_i = y \in \operatorname{cl}B(x, \varepsilon),\ (y_i) \in \Omega_f\right\}.$$

Let y_i be a sequence in Ω_f converging to some point $y \in \operatorname{cl}B(x, \varepsilon)$ and $v = \lim_{i\to\infty} d\exp_x^{-1}(y_i)(\operatorname{grad} f(y_i))$. For any sequence of positive numbers θ_k converging downward to zero, we have $\operatorname{cl}B(x, \varepsilon) \subset B(x, \theta_k + \varepsilon)$. Therefore, for any k, there exists N_k such that for $i \geq N_k$, $y_i \in B(x, \theta_k + \varepsilon)$. We set $(z_j^k)_j = (y_{N_k+j})_j \in B(x, \theta_k+\varepsilon)$. Now it is clear that for every k, $v = \lim_{j\to\infty} d\exp_x^{-1}(z_j^k)(\operatorname{grad} f(z_j^k)) \in \operatorname{cl}\{d\exp_x^{-1}(y)(\operatorname{grad} f(y)) : y \in \operatorname{cl}B(x, \varepsilon+\theta_k)\cap\Omega_f\}$, which proves the first inclusion.

For the converse, let

$$w \in \bigcap_k \mathrm{cl}\left\{d\exp_x^{-1}(y)(\mathrm{grad}\, f(y)) : y \in \mathrm{cl}B(x, \varepsilon + \theta_k) \cap \Omega_f\right\}.$$

Then, for every $k \in \mathbb{N}$ with $\theta_k + \varepsilon < i_M(x)$, we have

$$w \in \mathrm{cl}\left\{d\exp_x^{-1}(y)(\mathrm{grad}\, f(y)) : y \in \mathrm{cl}B(x, \varepsilon + \theta_k) \cap \Omega_f\right\},$$

Therefore, we can find a sequence $y_i \in \mathrm{cl}B(x, \varepsilon + \theta_i) \cap \Omega_f$ such that

$$\lim_{i\to\infty} \|d\exp_x^{-1}(y_i)(\mathrm{grad}\, f(y_i)) - w\| = 0$$

and (if necessary after passing to a subsequence),

$$\lim_{i\to\infty} y_i = y \in \mathrm{cl}B(x, \varepsilon),$$

as required. □

Using the previous lemma one can prove the following characterization of the Riemannian ε-subdifferential.

Lemma 3.5 *We have*

$$\partial_\varepsilon f(x) = \mathrm{conv}\left\{d\exp_x^{-1}(y)(\partial f(y)) : y \in clB(x, \varepsilon)\right\}.$$

Proof Assume that $\eta \in \partial_\varepsilon f(x)$, Lemma 3.4 implies $\eta = \Sigma_{k=1}^n t_k \xi_k$ where

$$\xi_k = \lim_{i_k\to\infty} d\exp_x^{-1}(y_{i_k})(\mathrm{grad}\, f(y_{i_k})),$$

$y_{i_k} \in \Omega_f$, $\lim_{i_k\to\infty} y_{i_k} = y_k \in \mathrm{cl}B(x, \varepsilon)$. Hence

$$\xi_k = d\exp_x^{-1}(y_k)\left(\lim_{i_k\to\infty} \mathrm{grad}\, f(y_{i_k})\right).$$

Set $\eta_k = (d\exp_x^{-1}(y_k))^{-1}(\xi_k)$ in $\partial f(y_k)$, then

$$\eta = \Sigma_{k=1}^n t_k d\exp_x^{-1}(y_k)(\eta_k) \in \mathrm{conv}\left\{d\exp_x^{-1}(y)(\partial f(y)) : y \in \mathrm{cl}B(x, \varepsilon)\right\}.$$

For the converse, let

$$A = \left\{d\exp_x^{-1}(y)(\partial f(y)) : y \in \mathrm{cl}B(x, \varepsilon)\right\}$$

and $\xi \in A$, then $\xi = d\exp_x^{-1}(y)(\eta)$, where $\eta = \lim_{i\to\infty} \mathrm{grad}\, f(y_i)$, $y_i \in \Omega_f$, $\lim_{i\to\infty} y_i = y$. Hence

$$\xi = \lim_{i\to\infty} d\exp_x^{-1}(y_i)(\mathrm{grad}\, f(y_i))$$

which implies $A \subset \partial_\varepsilon f(x)$, and the property of convex hull completes the proof. □

The following remark is required in the sequel.

Remark 3.6 Let M be a Riemannian manifold. An easy consequence of the definition of the parallel translation along a curve as a solution to an ordinary linear differential equation, implies that the mapping

$$C : TM \to T_{x_0}M, C(x, \xi) = L_{xx_0}(\xi),$$

when x is in a neighborhood U of x_0, is well defined and continuous at (x_0, ξ_0), that is, if $(x_n, \xi_n) \to (x_0, \xi_0)$ in TM then $L_{x_n x_0}(\xi_n) \to L_{x_0 x_0}(\xi_0) = \xi_0$, for every $(x_0, \xi_0) \in TM$; see [5, Remark 6.11].

Remark 3.7 Note that for small enough $\varepsilon > 0$, $\partial f(x) \subset \partial_\varepsilon f(x)$. If $\varepsilon_1 > \varepsilon_2$, then $\partial_{\varepsilon_2} f(x) \subset \partial_{\varepsilon_1} f(x)$. Therefore, $\partial f(x) \subseteq \lim_{\varepsilon_k \downarrow 0} \partial_{\varepsilon_k} f(x) = \bigcap_{\varepsilon_k} \partial_{\varepsilon_k} f(x)$. We claim that $\bigcap_{\varepsilon_k} \partial_{\varepsilon_k} f(x) \subseteq \partial f(x)$. To prove the claim, we assume on the contrary that there exists $\xi \in \bigcap_{\varepsilon_k} \partial_{\varepsilon_k} f(x) \setminus \partial f(x)$. Since $\partial_{\varepsilon_k} f(x)$ is a sequence of compact and nested subsets of T_xM, we have

$$\xi \in \cap_{\varepsilon_k} \partial_{\varepsilon_k} f(x) = \text{conv} \cap_{\varepsilon_k} \left\{ d\exp_x^{-1}(y)(\partial f(y)) : \text{dist}(y, x) \leq \varepsilon_k \right\}.$$

Hence, $\xi = \Sigma_{k=1}^m t_k \xi_k$, where $\xi_k \in \cap_{\varepsilon_k} \{d\exp_x^{-1}(y)(\partial f(y)) : \text{dist}(y, x) \leq \varepsilon_k\}$, and $\Sigma_{k=1}^m t_k = 1$. Therefore, there exists $w_{k_i} \in \partial f(y_{k_i})$ such that $\text{dist}(y_{k_i}, x) \leq \varepsilon_k$ with $\xi_k = d\exp_x^{-1}(y_{k_i})(w_{k_i})$. Since M is complete, it follows that $\{y_{k_i}\}$ has a subsequence convergent to x in M. By Theorem 2.9 of [20], $L_{y_{k_i} x}(w_{k_i})$ has a subsequence convergent to some vector $\tilde{\xi} \in \partial f(x)$. Since $L_{y_{k_i} x}(w_{k_i}) = L_{y_{k_i} x}((d\exp_x^{-1}(y_{k_i}))^{-1}(\xi_k))$ converges to ξ_k, then $\tilde{\xi} = \xi_k \in \partial f(x)$ and since $\partial f(x)$ is convex, $\xi \in \partial f(x)$ which is a contradiction.

We recall that a set valued function $F : X \to Y$, where X, Y are topological spaces, is said to be upper semicontinuous at x,if for every open neighborhood U of $F(x)$, there exits an open neighborhood V of x,such that

$$y \in V \Longrightarrow F(y) \subseteq U.$$

Assume that F has compact values, then there is a sequential characterization for the set valued upper semicontinuity as follows: F is upper semicontinuous at x,if and only if for each sequence $\{x_n\} \subset X$ converging to x and each sequence $\{y_n\} \subset F(x_n)$ converging to y; $y \in F(x)$.

Lemma 3.8 *Let U be a compact subset of M and $\varepsilon < i(U)$, then for every open neighborhood Win U, the set valued mapping $\partial_\varepsilon f : W \to TM$ is upper semicontinuous.*

Proof For every arbitrary fixed $x \in W$, let r be a positive number with $r < \varepsilon$. We define $F : B(x, r) \cap W \to T_xM$ by

$$F(z) = L_{zx} \left(\left\{ d\exp_z^{-1}(y)(\partial f(y)) : y \in \text{cl}B(z, \varepsilon) \right\} \right).$$

First, we prove F is upper semicontinuous at x.

Let $\{x_k\} \subset B(x,r) \cap W$ and $\{v_k\} \subset T_xM$ be two sequences converging, respectively, to x and v, where $v_k \in F(x_k)$. Hence $v_k = L_{x_k x}(d\exp_{x_k}^{-1}(y_k)(\xi_k))$ where $\xi_k \in \partial f(y_k)$ and $y_k \in \mathrm{cl}B(x_k, \varepsilon)$.

Note that M is complete, therefore $\{y_k\}$ has a subsequence convergent to some point y in M. Moreover, f is Lipschitz on $B(x,\varepsilon)$, by Theorem 2.9 of [20] we deduce that $L_{y_k y}(\xi_k)$ has a subsequence convergent to some vector $\xi \in \partial f(y)$. Thus, $v = d\exp_x^{-1}(y)(\xi)$, where $\xi \in \partial f(y)$. Since $\mathrm{dist}(x_k, y_k) \leq \varepsilon$ by the continuity of the distance function $\mathrm{dist}(x,y) \leq \varepsilon$, which means $v \in F(x)$ and F is upper semicontinuous at x. Note that F has compact values, consequently the set valued function $\mathrm{conv}F : B(x,r) \cap W \to T_xM$ defined by

$$\mathrm{conv}F(z) = L_{zx}\left(\mathrm{conv}\left\{d\exp_z^{-1}(y)(\partial f(y)) : y \in \mathrm{cl}B(z,\varepsilon)\right\}\right).$$

is upper semicontinuous at x.

Now, we prove the upper semicontinuity of $\partial_\varepsilon f$ at x. Let $\{x_k\} \subset B(x,r) \cap W$ and $\{v_k\} \subset TM$ be two sequences converging, respectively, to x and v, where $v_k \in \partial_\varepsilon f(x_k)$. Then $L_{x_k x}(v_k) \in \mathrm{conv}F(x_k)$ and by Remark 3.6, $L_{x_k x}(v_k)$ converges to v and by upper semicontinuity of $\mathrm{conv}F$ at x, $v \in \mathrm{conv}F(x) = \partial_\varepsilon f(x)$. □

Lemma 3.9 *Let B be a closed ball in a complete Riemannian manifold M, $f : M \to \mathbb{R}$ be locally Lipschitz, Z be the set of all stationary points of f in B and $B_\delta := \{x \in B : dist_Z(x) \geq \delta > 0\}$. Then there exist $\varepsilon > 0$ and $\sigma > 0$ such that $0 \notin \partial_\varepsilon f(x)$ and $\min\{\|v\| : v \in \partial_\varepsilon f(x)\} \geq \sigma$,for all $x \in B_\delta$.*

Proof Since B is compact, it follows that there exists $\varepsilon > 0$ such that $\partial_\varepsilon f$ is well-defined on B_δ. Assume that $x \in B_\delta$, consequently $0 \notin \partial f(x)$. We claim that there exists $\varepsilon > 0$ such that $0 \notin \partial_\varepsilon f(x)$. On the contrary, suppose that $0 \in \partial_{\frac{1}{i}} f(x)$, for $i = N, N+1, \ldots, 1/N < i_M(x)$. Since $\partial_{\frac{1}{i}} f(x)$ is a sequence of compact and nested subsets of T_xM, we have

$$0 \in \cap_{i=N}^{\infty} \partial_{\frac{1}{i}} f(x) = \mathrm{conv} \cap_{i=N}^{\infty} \left\{d\exp_x^{-1}(y)(\partial f(y)) : \mathrm{dist}(y,x) \leq \frac{1}{i}\right\}.$$

Hence, $0 = \Sigma_{k=1}^m t_k \xi_k$, where $\xi_k \in \cap_{i=N}^{\infty}\left\{d\exp_x^{-1}(y)(\partial f(y)) : \mathrm{dist}(y,x) \leq \frac{1}{i}\right\}$. Therefore, there exists $w_{k_i} \in \partial f(y_{k_i})$ such that $\mathrm{dist}(y_{k_i}, x) \leq \frac{1}{i+N}$ with $\xi_k = d\exp_x^{-1}(y_{k_i})(w_{k_i})$. Since M is complete, it follows that $\{y_{k_i}\}$ has a subsequence convergent to x in M. By Theorem 2.9 of [20], $L_{y_{k_i} x}(w_{k_i})$ has a subsequence convergent to some vector $\xi \in \partial f(x)$. Since $L_{y_{k_i} x}(w_{k_i}) = L_{y_{k_i} x}((d\exp_x^{-1}(y_{k_i}))^{-1}(\xi_k))$ converges to ξ_k, then $\xi = \xi_k \in \partial f(x)$ and since $\partial f(x)$ is convex, $0 \in \partial f(x)$ which is a contradiction.

To prove the second part of the lemma; note that $\partial_\varepsilon f(x)$ is a compact subset of T_xM, and the norm function is continuous, therefore there exists $0 \neq w \in \partial_\varepsilon f(x)$ such that $\|w\| = \min\{\|v\| : v \in \partial_\varepsilon f(x)\}$. Assume on the contrary, that for every $i \in \mathbb{N}$, there exists $x_i \in B_\delta$ provided that $\|w_i\| = \min\{\|v\| : v \in \partial_\varepsilon f(x_i)\}$ and $0 < \|w_i\| < 1/i$. Therefore, there exist convergent subsequences of x_i and w_i with respective limits $x \in B_\delta$ and $0 \in \partial_\varepsilon f(x)$, which is a contradiction. □

3.2 Descent directions

In the present section, we treat the problem of finding directions $w_0 \in \partial_\varepsilon f(x)$ such that with suitable step lengths $t > 0$ the objective function f affords a decrease along the geodesic $\exp_x\left(\frac{-tw_0}{\|w_0\|}\right)$. The next result shows that, whenever one has full knowledge of the ε-subdifferential, a suitable descent direction can be obtained by solving a simple quadratic program. We will use the following theorem; for its proof see [20].

Theorem 3.10 (**Lebourg's Mean Value Theorem**) *Let M be a finite dimensional Riemannian manifold, $x, y \in M$ and $\gamma : [0, 1] \longrightarrow M$ be a smooth path joining x and y. Let f be a Lipschitz function around $\gamma[0, 1]$. Then there exist $0 < t_0 < 1$ and $\xi \in \partial f(\gamma(t_0))$ such that*

$$f(y) - f(x) = \langle \xi, \gamma'(t_0)\rangle.$$

The next theorem can be proved using a property of the exponential map called radially isometry.

Definition 3.11 Assume that $R : TM \to M$ is a retraction on M, we say R is a radial isometry if for each $x \in M$ there exists $\epsilon > 0$ such that for all $v, w \in B(0_x, \epsilon)$, we have

$$\langle dR_x(v)(v), dR_x(v)(w)\rangle = \langle v, w\rangle.$$

By Gauss's lemma, we know that the exponential map is a radial isometry.

Theorem 3.12 *Assume $\varepsilon > 0$ and δ are given from Lemma 3.9 so that $0 \notin \partial_\varepsilon f(x)$ for all $x \in B_\delta$. Let $x \in B_\delta$ and consider an element of $\partial_\varepsilon f(x)$ with minimum norm,*

$$w_0 := \mathit{argmin}\,\{\|v\| : v \in \partial_\varepsilon f(x)\},$$

and get $g_0 := -\frac{w_0}{\|w_0\|}$. Then g_0 affords a uniform decrease of f over $B(x, \varepsilon)$, i.e.,

$$f(\exp_x(\varepsilon g_0)) - f(x) \le -\varepsilon\|w_0\|.$$

Proof By Lebourg's mean value theorem [20], there exist $0 < t_0 < 1$ and $\xi \in \partial f(\gamma(t_0))$ such that $f(\exp_x(\varepsilon g_0)) - f(x) = \langle \xi, \gamma'(t_0)\rangle$, where $\gamma(t) := \exp_x(t\varepsilon g_0)$ is a geodesic starting at x by initial speed εg_0. Since $\exp_x$ is a radial isometry, therefore we have that

$$\begin{aligned} f(\exp_x(\varepsilon g_0)) - f(x) &= \langle \xi, d\exp_x(\varepsilon t_0 g_0)(\varepsilon g_0)\rangle \\ &= \varepsilon\langle d\exp_x(\varepsilon t_0 g_0)(d\exp_x^{-1}(\exp_x(\varepsilon t_0 g_0))(\xi)), d\exp_x(\varepsilon t_0 g_0)(g_0)\rangle \\ &= \varepsilon\langle d\exp_x^{-1}(\exp_x(\varepsilon t_0 g_0))(\xi), g_0\rangle. \end{aligned}$$

Since $\mathrm{dist}(\exp_x(\varepsilon t_0 g_0), x) = t_0\varepsilon \le \varepsilon$, it follows that $d\exp_x^{-1}(\exp_x(\varepsilon t_0 g_0))(\xi) \in \partial_\varepsilon f(x)$. We claim that $\|w_0\|^2 \le \langle \phi, w_0\rangle$ for every $\phi \in \partial_\varepsilon f(x)$, which implies $\langle \phi, g_0\rangle \le -\|w_0\|$. Hence, we can deduce that $f(\exp_x(\varepsilon g_0)) - f(x) \le -\varepsilon\|w_0\|$.

Proof of the claim: assume on the contrary; there exists $\phi \in \partial_\varepsilon f(x)$ such that $\langle \phi, w_0 \rangle < \|w_0\|^2$ and consider $w := w_0 + t(\phi - w_0) \in \partial_\varepsilon f(x)$, then

$$\|w_0\|^2 - \|w\|^2 = -t(2\langle w_0, \phi - w_0 \rangle + t\langle \phi - w_0, \phi - w_0 \rangle),$$

we can assume that t is small enough such that $\|w_0\|^2 > \|w\|^2$, which is a contradiction. □

Remark 3.13 Instead of using the differential of the inverse exponential map to transport the subgradients of the cost function at the point y to the point x, we can choose the differential of the inverse of a retraction $R : TM \to M$, however by the proof of the previous theorem we have to restrict ourselves to the class of retractions that are radially isometric; see [1]. For example as parallel transport can be considered as the differential of the inverse of a retraction, therefore it might be used to transport the subgradients of the cost function at the point y to the point x; see [34]. It is worth mentioning that if the differential of the inverse of a retraction $R : TM \to M$ is selected to transport the vectors, then the retraction R must also be used to take a step in the direction of a tangent vector . Using a good retraction amounts to finding an approximation of the exponential mapping that can be computed with low computational cost while not adversely affecting the behavior of the optimization algorithm.

Definition 3.14 (Descent direction) Let $f : M \to \mathbb{R}$ be a locally Lipschitz function on a complete Riemannian manifold M, $w \in T_x M$, $g = -\frac{w}{\|w\|}$ is called a decent direction at x, if there exists $\alpha > 0$ such that

$$f(\exp_x(tg)) - f(x) \leq -t\|w\|, \forall t \in (0, \alpha). \tag{3.1}$$

In the construction of the previous theorem, g_0 is a descent direction of f at x, because along the same lines as the proof, we can deduce that for g_0 and every $t \in (0, \varepsilon)$,

$$f(\exp_x(tg_0)) - f(x) \leq -t\|w\|.$$

It is clear that we can choose the mentioned descent direction in order to move along a geodesic starting from an initial point toward a neighborhood of a minimum point.

3.3 Approximation of the ε-subdifferential

For general nonsmooth optimization problems it may be difficult to give an explicit description of the full subdifferential set. In the present section, we generalize ideas of [31] to obtain an iterative procedure to approximate the ε-subdifferential. We start with the subgradient of an arbitrary point nearby x and move the subgradient to the tangent space in x via the derivative of the logarithm mapping, and in every subsequent iteration, the subgradient of a new point nearby x is computed and moved to the tangent space in x to add to the working set to improve the approximation of $\partial_\varepsilon f(x)$. Indeed, we do not want to provide a description of the entire ε-subdifferential set at each iteration, what we do is to approximate $\partial_\varepsilon f(x)$ by the convex hull of its elements. In this way, let $W_k := \{v_1, ..., v_k\} \subseteq \partial_\varepsilon f(x)$, then we define

$$w_k := \underset{v \in \operatorname{conv} W_k}{\operatorname{argmin}} \|v\|.$$

Now if we have

$$f(\exp_x(\varepsilon g_k)) - f(x) \leq -c\varepsilon\|w_k\|, c \in (0, 1) \tag{3.2}$$

where $g_k = -\frac{w_k}{\|w_k\|}$, then we can say $\operatorname{conv} W_k$ is an acceptable approximation for $\partial_\varepsilon f(x)$. Otherwise, we add a new element of $\partial_\varepsilon f(x) \setminus \operatorname{conv} W_k$ to W_k.

Lemma 3.15 *Let $v \in \partial_\varepsilon f(x)$ such that $\langle v, g_k \rangle > -\|w_k\|$, then $v \notin \operatorname{conv} W_k$.*

Proof It can be proved along the same lines as the proof of the claim of Theorem 3.12. □

The following lemma proves that if W_k is not an acceptable approximation for $\partial_\varepsilon f(x)$, then there exists $v_{k+1} \in \partial_\varepsilon f(x)$ such that $\langle v_{k+1}, g_k \rangle \geq -c\|w_k\| > -\|w_k\|$, therefore we have from the previous lemma that $v_{k+1} \in \partial_\varepsilon f(x) \setminus \operatorname{conv} W_k$.

Lemma 3.16 *Let $W_k = \{v_1, ..., v_k\} \subset \partial_\varepsilon f(x)$, $0 \notin \operatorname{conv} W_k$ and*

$$w_k = \operatorname{argmin}\{\|v\| : v \in \operatorname{conv} W_k\}.$$

If we have $f(\exp_x(\varepsilon g_k)) - f(x) > -c\varepsilon\|w_k\|$, where $c \in (0, 1)$ and $g_k = \frac{-w_k}{\|w_k\|}$, then there exist $\theta_0 \in (0, \varepsilon]$ and $\bar{v}_{k+1} \in \partial f(\exp_x(\theta_0 g_k))$ such that

$$\langle d\exp_x^{-1}(\exp_x(\theta_0 g_k))(\bar{v}_{k+1}), g_k \rangle \geq -c\|w_k\|,$$

and $v_{k+1} := d\exp_x^{-1}(\exp_x(\theta_0 g_k))(\bar{v}_{k+1}) \notin \operatorname{conv} W_k$.

Proof We prove this lemma using Lemma 3.1 and Proposition 3.1 in [31]. Define

$$h(t) := f(\exp_x(t g_k)) - f(x) + ct\|w_k\|, t \in \mathbb{R},$$

and a new locally Lipschitz function $G : B(0_x, i_M(x)) \subset T_x M \to \mathbb{R}$ by $G(g) = f(\exp_x(g))$, then $h(t) = G(t g_k) - G(0) + ct\|w_k\|$. Assume that $h(\varepsilon) > 0$, then by Proposition 3.1 of [31], there exists $\theta_0 \in [0, \varepsilon]$ such that h is increasing in a neighborhood of θ_0. Therefore, by Lemma 3.1 of [31] for every $\xi \in \partial h(\theta_0)$, one has $\xi \geq 0$. By [20, Proposition 3.1]

$$\partial h(\theta_0) \subseteq \big\langle \partial f(\exp_x(\theta_0 g_k)), d\exp_x(\theta_0 g_k)(g_k) \big\rangle + c\|w_k\|.$$

If $\bar{v}_{k+1} \in \partial f(\exp_x(\theta_0 g_k))$ such that

$$\langle \bar{v}_{k+1}, d\exp_x(\theta_0 g_k)(g_k) \rangle + c\|w_k\| \in \partial h(\theta_0),$$

then

$$\Big\langle d\exp_x^{-1}(\exp_x(\theta_0 g_k))(\bar{v}_{k+1}), g_k \Big\rangle + c\|w_k\| \geq 0.$$

Now, Lemma 3.15 implies that

$$v_{k+1} := d\exp_x^{-1}(\exp_x(\theta_0 g_k))(\bar{v}_{k+1}) \notin \operatorname{conv} W_k,$$

which proves our claim. □

Now we present Algorithm 1 to find a vector $v_{k+1} \in \partial_\varepsilon f(x)$ which can be added to the set W_k in order to improve the approximation of $\partial_\varepsilon f(x)$. It is easy to prove by Proposition 3.2 and Proposition 3.3 of [31] that this algorithm terminates after finitely many iterations. Practically, applying Algorithm 1 is not costly. We have observed that, as h does not usually have a local extremum on $(0, \varepsilon)$ for ε small, the algorithm terminates after one iteration.

Algorithm 1 An h-increasing point algorithm; $v = Increasing(x, g, a, b)$

```
Input $x \in M, g \in T_x M, a, b \in \mathbb{R}$.
Let $t = b$.
repeat
    select $v \in \partial f(\exp_x(tg))$ such that $\langle v, d\exp_x(tg)(g)\rangle + c\|w\| \in \partial h(t)$
    if $\langle v, d\exp_x(tg)(g)\rangle + c\|w\| < 0$ then
        $t = \frac{a+b}{2}$
        if $h(b) > h(t)$ then
            $a = t$
        else
            $b = t$
        end if
    end if
until $\langle v, d\exp_x(tg)(g)\rangle + c\|w\| \geq 0$
```

Then we give Algorithm 2 for finding a descent direction. Moreover, Theorem 3.17 proves that Algorithm 2 terminates after finitely many iterations.

Theorem 3.17 *Let for the point $x_1 \in M$, the level set $N = \{x : f(x) \leq f(x_1)\}$ be bounded, then for each $x \in N$, Algorithm 2 terminates after finitely many iterations.*

Proof Now we claim that either after a finite number of iterations the stopping condition is satisfied or for some m,

$$\|w_m\| \leq \delta,$$

and the algorithm terminates. If the stopping condition is not satisfied and $\|w_k\| > \delta$, then by Lemma 3.16 we find $v_{k+1} \notin \text{conv} W_k$ such that

$$\langle v_{k+1}, w_k\rangle \leq c\|w_k\|^2.$$

Note that $d\exp_x^{-1}$ on $\text{cl}B(x, \varepsilon)$ is bounded by some $m_1 \geq 0$ and by the Lipschitzness of f of the constant L, Theorem 2.9 of [20] implies that for every $\xi \in \partial_\varepsilon f(x)$,

Algorithm 2 A descent direction algorithm; $(g_k, k) = Decent(x, \delta, c, \varepsilon)$

```
Input $x \in M$; $\delta, c, \varepsilon \in (0, 1)$.
Let $g_1 \in T_x M$ such that $\|g_1\| = 1$.
if $f$ is differentiable on $\exp_x(\varepsilon g_1)$, then
        $v = d\exp_x^{-1}(\exp_x(\varepsilon g_1))(\operatorname{grad} f(\exp_x(\varepsilon g_1)))$
else select arbitrary $v \in d\exp_x^{-1}(\exp_x(\varepsilon g_1))(\partial f(\exp_x(\varepsilon g_1)))$
    Set $W_1 = \{v\}$ and let $k = 1$
end if
Step 1: (Compute a descent direction)
Solve the following minimization problem and let $w_k$ be its solution:
        $\min_{v \in \operatorname{conv} W_k} \|v\|.$
if $\|w_k\| \leq \delta$ then stop
else let $g_{k+1} = -\frac{w_k}{\|w_k\|}$.
end if
Step 2: (Stopping condition)
if $f(\exp_x(\varepsilon g_{k+1})) - f(x) \leq -c\varepsilon\|w_k\|$, then stop.
end if
Step 3: $v = Increasing(x, g_{k+1}, 0, \varepsilon)$.
Set $v_{k+1} = v$, $W_{k+1} = W_k \cup \{v_{k+1}\}$ and $k = k + 1$. Go to step 1.
```

$\|\xi\| \leq m_1 L$. Now, $w_{k+1} \in \operatorname{conv}\{v_{k+1}\} \cup W_k$ has the minimum norm, so for all $t \in (0, 1)$,

$$\begin{aligned}
\|w_{k+1}\|^2 &\leq \|tv_{k+1} + (1-t)w_k\|^2 \\
&\leq \|w_k\|^2 + 2t\langle w_k, (v_{k+1} - w_k)\rangle + t^2\|v_{k+1} - w_k\|^2 \\
&\leq \|w_k\|^2 - 2t(1-c)\|w_k\|^2 + 4t^2L^2m_1^2 \\
&\leq \left(1 - [(1-c)(2Lm_1)^{-1}\delta]^2\right)\|w_k\|^2, \qquad (3.3)
\end{aligned}$$

where the last inequality is obtained by assuming $t = (1-c)(2Lm_1)^{-2}\|w_k\|^2 \in (0, 1)$, $\delta \in (0, Lm_1)$ and $\|w_k\| > \delta$. Now considering $r = 1 - [(1-c)(2Lm_1)^{-1}\delta]^2$, it follows that

$$\|w_{k+1}\|^2 \leq r\|w_k\|^2 \leq ... \leq r^k(Lm_1)^2.$$

Therefore, after a finite number of iterations $\|w_{k+1}\| \leq \delta$. □

Finally, Algorithm 3 is the minimization algorithm which finds a descent direction in any iteration.

Theorem 3.18 *If $f : M \to \mathbb{R}$ is a locally Lipschitz function on a complete Riemannian manifold M, and*

$$N = \{x : f(x) \leq f(x_1)\}$$

Algorithm 3 A minimization algorithm; $x_k = Min(f, x_1, \theta_\varepsilon, \theta_\delta, \varepsilon_1, \delta_1, c)$

```
Input: $f$ (A locally Lipschitz function defined on a complete Riemannian manifold $M$); $x_1 \in M$ (a starting point); $c, \theta_\varepsilon, \theta_\delta, \varepsilon_1, \delta_1 \in (0, 1)$; $k = 1$.
Step 1 (Set new parameters) $s = 1$ and $x_k^s = x_k$.
Step 2. (Descent direction) $(g_k^s, n_k^s) = Decent(x_k^s, \delta_k, c, \varepsilon_k)$
    $$\|w_k^s\| = \min\{\|w\| : w \in \mathrm{conv} W_k^s\}.$$
if $\|w_k^s\| = 0$ then stop
else let $g_k^s = -\frac{w_k^s}{\|w_k^s\|}$ be the descent direction.
end if
if $\|w_k^s\| \leq \delta_k$ then set $\varepsilon_{k+1} = \varepsilon_k \theta_\varepsilon$, $\delta_{k+1} = \delta_k \theta_\delta$, $x_{k+1} = x_k^s$, $k = k + 1$. Go to Step 1.
else
    $$\sigma = \mathrm{argmax} \left\{\sigma \geq \varepsilon_k : f(\exp_{x_k^s}(\sigma g_k^s)) - f(x_k^s) \leq -c\sigma \|w_k^s\|\right\}$$
    and construct the next iterate $x_k^{s+1} = \exp_{x_k^s}\left(\sigma g_k^s\right)$. Set $s = s+1$ and go to Step 2.
end if
```

is bounded, then either Algorithm 3 terminates after a finite number of iterations with $\|w_k^s\| = 0$*, or every accumulation point of the sequence* $\{x_k\}$ *belongs to the set*

$$X = \{x \in M : 0 \in \partial f(x)\}.$$

Proof Note that there exists $\varepsilon < i(N)$ such that $\partial_\varepsilon f$ on N is well-defined. If the algorithm terminates after finite number of iterations, then x_k^s is an $\varepsilon-$stationary point of f. Suppose that the algorithm does not terminate after finitely many iterations. Assume that g_k^s is a descent direction, since $\sigma \geq \varepsilon_k$, we have

$$f\left(x_k^{s+1}\right) - f\left(x_k^s\right) \leq -c\varepsilon_k \|w_k^s\| < 0,$$

for $s = 1, 2, ...,$therefore, $f\left(x_k^{s+1}\right) < f(x_k^s)$ for $s = 1, 2, ...$ Since f is Lipschitz and N is bounded, it follows that f has a minimum in N. Therefore, $f(x_k^s)$ is a bounded decreasing sequence in $\mathbb{R}$, so is convergent. Thus $f(x_k^s) - f\left(x_k^{s+1}\right)$ is convergent to zero and there exists s_k such that

$$f\left(x_k^s\right) - f\left(x_k^{s+1}\right) \leq c\varepsilon_k \delta_k,$$

for all $s \geq s_k$. Thus

$$\|w_k^s\| \leq \frac{f\left(x_k^s\right) - f\left(x_k^{s+1}\right)}{c\varepsilon_k} \leq \delta_k, s \geq s_k. \tag{3.4}$$

Hence after finitely many iterations, there exists s_k such that

$$x_{k+1} = x_k^{s_k},$$

and

$$\min\left\{\|v\| : v \in \operatorname{conv} W_k^{s_k}\right\} \leq \delta_k.$$

Since M is a complete Riemannian manifold and $\{x_k\} \subset N$ is bounded, there exists a subsequence $\{x_{k_i}\}$ converging to a point $x^* \in M$. Since $\operatorname{conv} W_{k_i}^{s_{k_i}}$ is a subset of $\partial_{\varepsilon_{k_i}} f\left(x_{k_i}^{s_{k_i}}\right)$, then

$$\|w_{k_i}\| = \min\left\{\|v\| : v \in \partial_{\varepsilon_{k_i}} f\left(x_{k_i}^{s_{k_i}}\right)\right\} \leq \delta_{k_i}.$$

Hence $\lim_{k_i\to\infty} \|w_{k_i}\| = 0$. Note that $w_{k_i} \in \partial_{\varepsilon_{k_i}} f\left(x_{k_i}^{s_{k_i}}\right)$, hence by Lemma 3.8 and Remark 3.7, $0 \in \partial f(x^*)$. □

4 Numerical experiments

We close this article by giving several numerical experiments. We set the parameters as follows: $c = 0.2$, $\delta_1 = 10^{-5}$, $\varepsilon_1 = 0.1$, and $\theta_\delta = 1$. In Algorithm 3, for all values of $k \leq 4$, we set $\theta_\varepsilon = 0.1$ and for $k > 4$, we set $\theta_\varepsilon = 0.8$. Algorithm 3 terminates when $\varepsilon_k < 10^{-7}$. We assume that the Armijo parameter $c = 0.2$, and use the simple line search strategy,

$$\sigma = \operatorname{argmax}\left\{\sigma \geq \varepsilon_k : f\left(\exp_{x_k^s}\left(\sigma g_k^s\right)\right) - f\left(x_k^s\right) \leq -c\sigma \left\|w_k^s\right\|\right\}.$$

Indeed, we start with $\sigma = 1$ and backtrack with a factor $\gamma = 0.5$. It is worth pointing out that in Algorithm 2, we generate g_1 randomly, therefore our algorithm has a stochastic behavior.

4.1 Denoising on a Riemannian manifold

We are going to solve the one dimensional total variation problem for functions which map into a manifold. Therefore, assume that M is a manifold, consider the minimization problem

$$\min_{u\in BV([0,1];M)}\left\{F(u) := \operatorname{dist}_2(f,u)^2 + \lambda\|\nabla u\|_1\right\}, \tag{4.1}$$

where $f : [0,1] \to M$ is the given (noisy) function, u is a function of bounded variation from $[0, 1]$ to M, dist_2 is the distance on the function space $L^2([0, 1]; M)$, and $\lambda > 0$ is a Lagrangian parameter, [37]. Note that for every $w \in [0, 1]$, $\nabla u(w) : \mathbb{R} \to T_{u(w)}M$ and $\|\nabla u\|_1 = \int_{[0,1]} \|\nabla u(w)\| dw$. Now we can formulate a discrete version of the problem (4.1) by restricting the space of functions to V_h^M which is the space of all geodesic finite element functions for M associated with a regular grid on [0, 1]; see [18, 38]. We refer to [38] for the definition of geodesic finite element spaces V_h^M.

Using the nodal evaluation operator $\varepsilon : V_h^M \to M^n$, $(\varepsilon(v_h))_i = v_h(x_i)$, where x_i is the i-th vertex of the simplicial grid on [0, 1], one can find an equivalent problem defined on M^n as follows,

$$\min_{u \in M^n} \left\{ F_*(u) := \operatorname{dist}_*(\varepsilon(f), u)^2 + \lambda \|\nabla(\varepsilon^{-1}(u))\|_1 \right\}, \tag{4.2}$$

where dist_* is the Riemannian distance on M^n.

Theorem 4.1 *Let M be a Hadamard manifold. If F_* is defined as in (4.2), then F_* is convex as a function defined on M^n.*

Proof It is enough to prove that $\|\nabla(\varepsilon^{-1}(u))\|_1$ is convex. Thus, we should prove that $\int_{[0,1]} \|\nabla v_{hu}(w)\| dw$, where v_{hu} is the geodesic finite element function corresponding to u, is convex. To do this, assume that u_1, u_2 are two arbitrary points in M^n and $\gamma = (\gamma_1, ..., \gamma_n)$ is a geodesic connecting them. We first show that for every arbitrary fix $w \in [0, 1]$, $f(t) = \|\nabla v_{h\gamma(t)}(w)\|$, as a function of t, is convex. Define

$$g(t) = 1/2\langle \nabla v_{h\gamma(t)}(w), \nabla v_{h\gamma(t)}(w)\rangle.$$

Assume that Γ is a grid on [0, 1] and $(s_i, s_{i+1}) \in \Gamma$ is such that $w \in (s_i, s_{i+1})$, moreover, σ_{it} is a minimizing geodesic parametrized by arc length connecting $\gamma_i(t)$ and $\gamma_{(i+1)}(t)$. Since σ_{it} is a geodesic with a constant speed, we have that

$$g(t) = 1/2\langle \nabla\sigma_{it}(w), \nabla\sigma_{it}(w)\rangle = \frac{1}{2}\int_0^1 \langle \nabla\sigma_{it}(x), \nabla\sigma_{it}(x)\rangle dx.$$

Now we define another smooth function $G : [0, 1] \times [0, 1] \to M$ by

$$G(t, x) = \sigma_{it}(x).$$

We put $V(t, x) := \frac{\partial G}{\partial t}(t, x)$ and usually write $\nabla\sigma_{it}(x) = \nabla G = \frac{\partial G}{\partial x}dx$. Consider the vector bundle $T([0, 1] \times [0, 1])^* \otimes G^{-1}TM$ over $[0, 1] \times [0, 1]$, which admits a natural fiber metric and a standard connection ∇ compatible with the metric. Under the natural identification, we denote $\nabla_x = \nabla_{(0, \frac{\partial}{\partial x})}$ and $\nabla_t = \nabla_{(\frac{\partial}{\partial t}, 0)}$. Therefore,

$$\begin{aligned}
1/2\frac{\partial^2}{\partial^2 t}\left\langle \frac{\partial G}{\partial x}dx, \frac{\partial G}{\partial x}dx \right\rangle &= \frac{\partial}{\partial t}\left\langle \nabla_t \frac{\partial G}{\partial x}dx, \frac{\partial G}{\partial x}dx \right\rangle \\
&= \frac{\partial}{\partial t}\left\langle \nabla_x \frac{\partial G}{\partial t}dx, \frac{\partial G}{\partial x}dx \right\rangle \\
&= \left\langle \nabla_t \nabla_x \frac{\partial G}{\partial t}dx, \frac{\partial G}{\partial x}dx \right\rangle + \left\langle \nabla_x \frac{\partial G}{\partial t}dx, \nabla_x \frac{\partial G}{\partial t}dx \right\rangle \\
&= \left\langle \nabla_x \nabla_t \frac{\partial G}{\partial t}dx, \frac{\partial G}{\partial x}dx \right\rangle + \left\langle R(\frac{\partial G}{\partial t}, \frac{\partial G}{\partial x})\frac{\partial G}{\partial t}dx, \frac{\partial G}{\partial x}dx \right\rangle \\
&\quad + \langle \nabla_x V dx, \nabla_x V dx\rangle
\end{aligned}$$

Since ∇ is metric ,

$$
\begin{aligned}
0 &= \int_0^1 \frac{\partial}{\partial x}\left\langle \nabla_t \frac{\partial G}{\partial t}, \frac{\partial G}{\partial x}dx\right\rangle dx\\
&= \int_0^1 \left\langle \nabla_x \nabla_t \frac{\partial G}{\partial t}dx, \frac{\partial G}{\partial x}dx\right\rangle dx.
\end{aligned}
$$

Hence,

$$g''(t) = \int_0^1 \|\nabla V\|^2 - trace\langle R(\nabla G, V)V, \nabla G\rangle.$$

Since M is a Hadamard manifold, it follows that $g''(t) \geq 0$ which implies g is convex. By definition of g, it is clear that $g(t) = 1/2f^2(t)$. We assume that $f(t) \neq 0$, then

$$f''(t) = \frac{g''(t)f^2(t) - (g'(t))^2}{f^3(t)}.$$

Hence,

$$
\begin{aligned}
f''(t) = \frac{1}{f^3(t)}\Bigg\{ &\int_0^1 (\|\nabla V\|^2)\left\langle \nabla v_{h\gamma(t)}(w), \nabla v_{h\gamma(t)}(w)\right\rangle\\
&-trace\langle R(\nabla G, V)V, \nabla G\rangle\langle \nabla v_{h\gamma(t)}(w), \nabla v_{h\gamma(t)}(w)\rangle\\
&-\langle \nabla V, \nabla v_{h\gamma(t)}(w)\rangle^2\Bigg\} \geq 0,
\end{aligned}
$$

which is obtained by the Cauchy-Schwarz inequality and the negativity of the sectional curvature. Thus, we proved that for every $w \in [0, 1]$, $f(t) = \|\nabla v_{h\gamma}(w)\|$ is convex, hence

$$\|\nabla v_{h\gamma(t)}(w)\| \leq t\|\nabla v_{hu_1}(w)\| + (1-t)\|\nabla v_{hu_2}(w)\|,$$

which implies

$$\int_{[0,1]} \|\nabla v_{h\gamma(t)}(w)\|dw \leq t\int_{[0,1]} \|\nabla v_{hu_1}(w)\|dw + (1-t)\int_{[0,1]} \|\nabla v_{hu_2}(w)\|dw,$$

which means $\int_{[0,1]} \|\nabla v_{hu}(w)\| = \|\nabla(\varepsilon^{-1}(u))\|_1$ is convex.

Note that if M is a Hadamard manifold, dist^2 is also a convex function on M^n. Hence we can conclude that F_* is convex on M^n. □

Let $\varepsilon(f) = (p_1, ..., p_n)$,then $F_* : M^n \to \mathbb{R}$ can be defined by

$$F_*(u_1, ..., u_n) = \Sigma_{i=1}^n \text{dist}(p_i, u_i)^2 + \lambda\Sigma_{i=1}^{n-1}\text{dist}(u_i, u_{i+1}),$$

where dist is the Riemannian distance on M. In order to find the subdifferential of F, we have to find the subdifferential of the distance and squared distance functions. The distance function is differentiable at $(p, q) \in M \times M$ if and only if there is a unique length minimizing geodesic from p to q. Furthermore, the distance function is smooth in a neighborhood of (p, q) if and only if p and q are not conjugate points along this minimizing geodesic. Consequently, the distance function is nondifferentiable at (p, q) if and only if $p = q$ or p and q are the conjugate points. Let the distance function be differentiable at (p, q), then

$$\frac{\partial \mathrm{dist}}{\partial p}(p, q) = \frac{-\exp_p^{-1}(q)}{\mathrm{dist}(p, q)}, \ \frac{\partial \mathrm{dist}^2}{\partial p}(p, q) = -2\exp_p^{-1}(q).$$

The following lemma is a direct consequence of Definition 2.3, the fact that dist_p is smooth on $U \setminus \{p\}$, where U is a sufficiently small open neighborhood of p, and $\exp_p$ is near $0 \in T_pM$ a radial isometry.

Lemma 4.2 *Let M be a complete Riemannian manifold. If $dist_p : M \to \mathbb{R}$ is defined by $dist_p(q) = dist(p, q)$,then*

$$\partial dist_p(p) = B,$$

where B is the closed unit ball of T_pM.

In our numerical examples, we consider a two dimensional sphere S^2 and the space of positive-definite matrices which is known as a Hadamard manifold. Therefore, F_* is convex on the space of positive definite matrices, while F_* is not a convex function on every sphere; see [44]. We first recall the properties of these two Riemannian manifolds.

The unit sphere S^2 is the smooth compact manifold

$$S^2 = \{x \in \mathbb{R}^3 : \|x\| = 1\},$$

and the global coordinates on S^2 are naturally given by this embedding into $\mathbb{R}^3$. The tangent space at a point $x \in S^2$ is

$$T_xS^2 = \{v \in \mathbb{R}^3 : \langle x, v\rangle = 0\}.$$

The inner product on T_xS^2 is defined by

$$\langle v, w\rangle_{T_xS^2} = \langle v, w\rangle_{\mathbb{R}^3}.$$

The exponential map

$$\exp_x : T_xS^2 \to S^2$$

is defined by

$$\exp_x(v) = \cos(\|v\|)x + \sin(\|v\|)\frac{v}{\|v\|}.$$

Moreover, if $x \in S^2$, then

$$\exp_x^{-1} : S^2 \to T_xS^2$$

is defined by

$$\exp_x^{-1}(y) = \frac{\theta}{\sin(\theta)}(y - x\cos(\theta)),$$

where $\theta = \arccos\langle x, y\rangle$. The Riemannian distance between two points x, y in S^2 is given by

$$\mathrm{dist}(x, y) = \arccos\langle x, y\rangle.$$

Let $t \to \gamma(t)$ be a geodesic on S^2, and let $u = \frac{\gamma^\circ(0)}{\|\gamma^\circ(0)\|}$. The parallel translation of a vector $v \in T_{\gamma(0)}S^2$, along the geodesic γ, is given by [2]

$$L_{\gamma(0)\gamma(t)}(v) = -\gamma(0)\sin(\|\gamma^\circ(0)\|t)u'v + u\cos(\|\gamma^\circ(0)\|t)u'v + (I - uu')v.$$

Utilizing the properties of the exponential map on a Riemannian manifold, for fixed point $x \in S^2$, and for each $\varepsilon > 0$, we may find number $\delta_x > 0$ such that

$$\|d(\exp_x^{-1})(y) - L_{yx}\| \leq \varepsilon, \text{provided thatdist}(x, y) < \delta_x.$$

It is worthwhile to mention that on any sphere antipodal points are conjugate points, but without loss of generality we can assume that u_i and u_{i+1} are not conjugate. In fact we use more than two nodal points for discretization of the function F, therefore it can be assumed that there is a nodal point between every two antipodal points.

The set of symmetric positive definite matrices, as a Riemannian manifold, is the most studied example of manifolds of nonpositive curvature. The space of all $n \times n$ symmetric, positive definite matrices will be denoted by $P(n)$. The tangent space to $P(n)$ at any of its points P is the space $T_P P(n) = \{P\} \times S(n)$, where $S(n)$ is the space of symmetric $n \times n$ matrices. On each tangent space $T_P P(n)$, the inner product is defined by

$$\langle A, B\rangle_{T_P P(n)} = \text{tr}(P^{-1}AP^{-1}B).$$

The Riemannian distance between $P, Q \in P(n)$ is given by

$$\text{dist}(P, Q) = \left(\Sigma_{i=1}^n \ln^2(\lambda_i)\right)^{(1/2)},$$

where λ_i, $i = 1, ..., n$ are eigenvalues of $P^{-1}Q$. The exponential map

$$\exp_P : S(n) \to P(n)$$

is defined by

$$\exp_P(v) = P^{1/2}\exp\left(P^{-1/2}vP^{-1/2}\right)P^{1/2}.$$

Moreover, if $P \in P(n)$, then

$$\exp_P^{-1} : P(n) \to S(n)$$

is defined by

$$\exp_P^{-1}(Q) = P^{1/2}\log\left(P^{-1/2}QP^{-1/2}\right)P^{1/2},$$

where log, exp, denote the logarithm and exponential functions on matrix space; for more details see [32].

First, we assume that $M = S^2$. We need to define a function from [0, 1] to S^2 to get the original image. Afterward, we add a Gaussian noise to the image to get the noisy image. Finally we apply algorithm 5 to the function F_* defined on M^{100} to get the denoised image, see Fig. 1.

For another example, we assume that $M = P(2)$. We add a Guassian noise to an original image on $P(2)$. Then we apply algorithm 5 to F_* on M^{100} to denoise the noisy image. In Fig. 2, we present the results regarding to the minimization of F_* on M^{100}. Note that every symmetric positive definite matrix $A \in P(2)$ defines an ellipse. The principal axes are given by the eigenvectors of A and the square root of the eigenvalues are the radii of the corresponding axes.

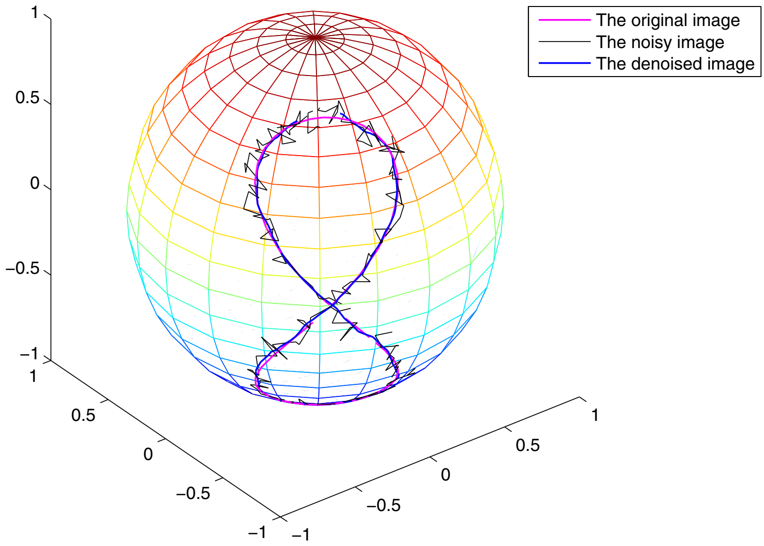

Fig. 1 TV regularization on S^2

4.2 Riemannian geometric median on the Sphere S^2

Let M be a Riemannian manifold. Given points $p_1, ..., p_m$ in M and corresponding positive real weights $w_1, ..., w_m$, with $\sum_{i=1}^m w_i = 1$, define the weighted sum of distances function

$$f(q) = \sum_{i=1}^{m} w_i \text{dist}(p_i, q),$$

where dist is the Riemannian distance function on M. We define the weighted geometric median x, as the minimizer of f. When all the weights are equal, $w_i = 1/m$, we call x simply the geometric median. Now, we assume that $M = S^2$. In Fig. 3 the results of the ε-subgradient algorithm for Riemannian geometric median on S^2 are plotted. Algorithm 2 terminated after only one iteration.

4.3 Rayleigh quotients on S^2

We consider the maximum of m Rayleigh quotients on the sphere S^2, i.e.,

$$f(x) = \max_{i=1,...,m} \frac{1}{2} x' A_i x, \tag{4.3}$$

$A_i \in S(3)$. Our aim is to find a minimum of f. In Fig. 4, the results of the ε-subgradient algorithm for Rayleigh quotients on S^2 are plotted. We have seen that for $\varepsilon > 0$ small, Algorithm 2 terminates after one iteration. Indeed, generally we don't meet points where several $x' A_i x$ achieve the maximum simultaneously.

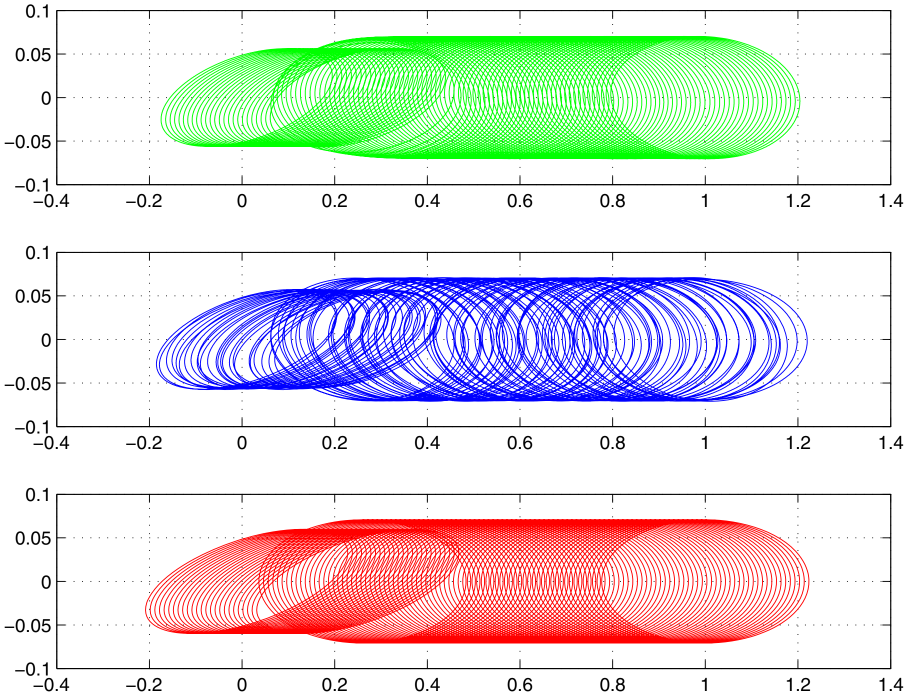

Fig. 2 TV regularization on $P(2)$. *Down-to-up*: the original image, the noisy image, the denoised image

4.4 Sphere packing on Grassmannians

We assume that the Grassmannian $\mathrm{Gr}(n, k)$ is the set of all k-dimensional linear subspaces of $\mathbb{R}^n$. In this section, we consider the problem of the packing of m spherical balls on $\mathrm{Gr}(n, k)$ with respect to the chordal distance. Let $B(P, r)$ denote the ball in $\mathrm{Gr}(n, k)$ with respect to chordal distance. Then we would like to find m points $P_1, \ldots, P_m$ in $\mathrm{Gr}(n, k)$ such that

$$\max \left\{ r \mid \forall i \neq j : B(P_i, r) \cap B(P_j, r) = \emptyset \right\}, \tag{4.4}$$

is maximized. This problem has been solved in [16] using a subgradient method.

Indeed, $\mathrm{Gr}(n, k)$ can be identified with the set $\{P \in S(n) | P^2 = P, \mathrm{tr}(P) = k\}$. Moreover, the tangent space of the Grassmannian at the point P, denoted by $T_P\mathrm{Gr}(n, k)$, is the following set

$$T_P\mathrm{Grass}(n, k) = \{P\Omega - \Omega P | \Omega \in so(n)\},$$

where

$$so(n) = \{\Omega \in \mathbb{R}^{n\times n} | \Omega' = -\Omega\}.$$

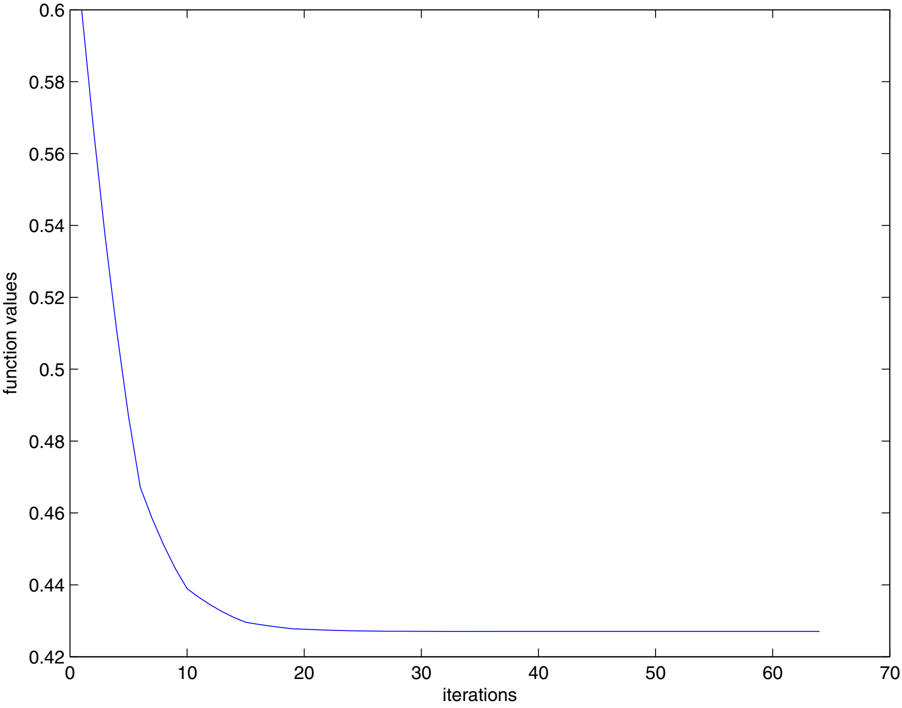

Fig. 3 ε-subgradient descent for Riemannian geometric median on S^2

As $\mathrm{Gr}(n, k)$ is a subset of the Euclidean vector space $S(n)$, the scalar product $\langle P, Q\rangle := \mathrm{tr}(PQ)$ induces a Riemannian metric on it. Therefore the chordal distance on $\mathrm{Gr}(n, k)$, denoted by $\mathrm{dist}(P, Q)$, is defined by

$$\mathrm{dist}(P, Q) = \sqrt{\frac{1}{2}}\|P - Q\|_F,$$

where $\|.\|_F$ denotes the Frobenius norm. On $\mathrm{Gr}(n, k)$ with the induced Riemannian metric, the geodesic γ emanating from P in the direction $\eta \in T_P\mathrm{Grass}(n, k)$ is defined by

$$\gamma(t) = \exp(t(\eta P - P\eta))P\exp(-t(\eta P - P\eta)).$$

The problem (4.4) is equivalent to the minimizing the following nonsmooth function;

$$F(P_1, ..., P_m) := \max_{i\neq j} \mathrm{tr}(P_i P_j), \tag{4.5}$$

on $\mathrm{Gr}(n, k) \times ... \times \mathrm{Gr}(n, k)$; see [16].

In Table 1, we illustrate the results of the nonsmooth subgradient (SB) method and our method for the sphere packing in $\mathrm{Gr}(16, 2)$ with $m = 10$ after 80 iterations with the same arbitrary starting points for both methods and the same step lengths. The computation time for both methods is the same. Our results show that the both methods have similar performance for this example. However, the ε-subgradient algorithm

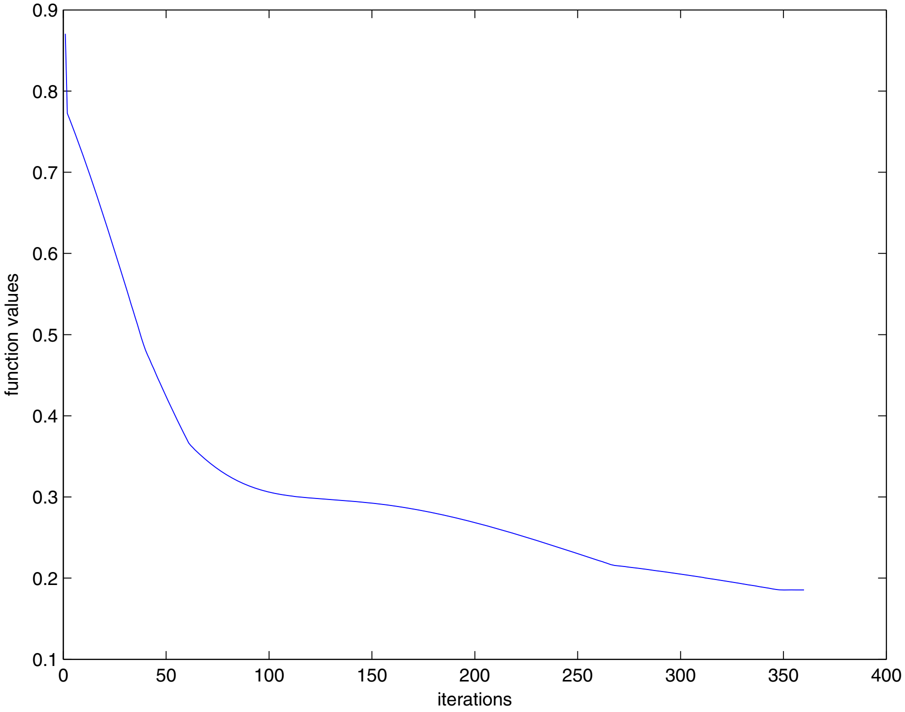

Fig. 4 ε-subgradient descent for Rayleigh quotients on S^2

Table 1 Numerical results in terms of number of function evaluations and the final obtained value of the function for sphere packings on Grassmannians

No.	Minimal distance in our method	nfeval	Minimal distance in SB	nfeval
1	1.853	605	1.853	601
2	1.682	230	1.683	231
3	1.775	1101	1.775	1109
4	1.872	1128	1.872	1123
5	1.801	237	1.801	234
6	1.813	1608	1.813	1609
7	1.872	1205	1.872	1209
8	1.874	891	1.874	890
9	1.719	1400	1.718	1409
10	1.789	989	1.789	980
11	1.708	425	1.708	430
12	1.804	1236	1.804	1230
13	1.897	1674	1.897	1679
14	1.676	2344	1.676	2340
15	1.631	1300	1.631	1306

is more general, because in this algorithm we do not need to have an explicit formula for the subdifferential and it can be computed approximately.

5 Conclusions

We have presented a practical algorithm in the context of ε-subgradient methods for nonsmooth problems on Riemannian manifolds. To the best of our knowledge, this is the first practical paper on approximating the subdifferential of locally Lipschitz functions on Riemannian manifolds. Using the algorithms presented in this paper, one can solve all nonsmooth locally Lipschitz minimization problems on Riemannian manifolds. The main result is the global convergence property of our minimization algorithm which is stated in Theorem 3.8. Moreover, comparing with subgradient algorithm [16], the ε-subgradient algorithm is much more general, because in this algorithm we do not need to have an explicit formula for the subdifferential and it can be computed approximately. An implementation of our proposed minimization algorithm is given in Matlab environment and tested on some problems.

References

1. Absil, P.A., Baker, C.G.: Trust-region methods on Riemannian manifolds. Found. Comput. Math. **7**, 303–330 (2007)
2. Absil, P.A., Mahony, R., Sepulchre, R.: Optimization Algorithm on Matrix Manifolds. Princeton University Press (2008)
3. Adler, R.L., Dedieu, J.P., Margulies, J.Y., Martens, M., Shub, M.: Newton's method on Riemannian manifolds and a geometric model for the human spine. IMA J. Numer. Anal. **22**, 359–390 (2002)
4. Afsari, B., Tron, R., Vidal, R.: On the convergence of gradient descent for finding the Riemannian center of mass. SIAM J. Control Optim. **51**, 2230–2260 (2013)
5. Azagra, D., Ferrera, J., López-Mesas, F.: Nonsmooth analysis and Hamilton-Jacobi equations on Riemannian manifolds. J. Funct. Anal. **220**, 304–361 (2005)
6. Azagra, D., Ferrera, J.: Applications of proximal calculus to fixed point theory on Riemannian manifolds. Nonlinear. Anal. **67**, 154–174 (2007)
7. Bagirov, A.M.: Continuous subdifferential approximations and their applications. J. Math. Sci. **115**, 2567–2609 (2003)
8. Bento, G.C., Ferreira, O.P., Oliveira, P.R.: Unconstrained steepest descent method for multicriteria optimization on Riemannian manifolds. J. Optim. Theory Appl. **154**, 88–107 (2012)
9. Bento, G.C., Melo, J.G.: A subgradient method for convex feasibility on Riemannian manifolds. J. Optim. Theory Appl. **152**, 773–785 (2012)
10. Borckmans, P.B., Easter Selvan, S., Boumal, N., Absil, P.A.: A Riemannian subgradient algorithm for economic dispatch with valve-point effect. J. Comput. Appl. Math. **255**, 848–866 (2014)
11. Burke, J.V., Lewis, A.S., Overton, M.L.: Approximating subdifferentials by random sampling of gradients. Math. Oper. Res. **27**, 567–584 (2002)
12. Burke, J.V., Lewis, A.S., Overton, M.L.: A robust gradient sampling algorithm for nonsmooth, nonconvex optimization. SIAM J. Optim. **15**, 751–779 (2005)
13. Clarke, F.H.: Necessary Conditions for Nonsmooth Problems in Optimal Control and the Calculus of Variations, Thesis, University of Washington, Seattle (1973)
14. da Cruz Neto, J.X., Lima, L.L., Oliveira, P.R.: Geodesic algorithm in Riemannian manifolds. Balkan J. Geom. Appl. **2**, 89–100 (1998)
15. da Cruz Neto, J.X., Ferreira, O.P., Lucambio Perez, L.R.: A proximal regularization of the steepest descent method in Riemannian manifold. Balkan J. Geom. Appl. **2**, 1–8 (1999)

16. Dirr, G., Helmke, U., Lageman, C.: Nonsmooth Riemannian optimization with applications to sphere packing and grasping. In Lagrangian and Hamiltonian Methods for Nonlinear Control 2006: Proceedings from the 3rd IFAC Workshop, Nagoya, Japan, 2006, Lecture Notes in Control and Information Sciences, vol. 366. Springer, Berlin (2007)
17. Ferreira, O.P., Oliveira, P.R.: Subgradient algorithm on Riemannian manifolds. J. Optim. Theory Appl. **97**, 93–104 (1998)
18. Grohs, P., Hardering, H., Sander, O.: Optimal a priori discretization error bounds for geodesic finite elements, SAM Report 2013–16, ETH Zürich. Submitted (2013)
19. Goldstein, A.A.: Optimization of Lipschitz continuous functions. Math. Program. **13**, 14–22 (1977)
20. Hosseini, S., Pouryayevali, M.R.: Generalized gradients and characterization of epi-Lipschitz sets in Riemannian manifolds. Nonlinear Anal. **74**, 3884–3895 (2011)
21. Hosseini, S., Pouryayevali, M.R.: Euler characterization of epi-Lipschitz subsets of Riemannian manifolds. J. Convex. Anal. **20**(1), 67–91 (2013)
22. Hosseini, S., Pouryayevali, M.R.: On the metric projection onto prox-regular subsets of Riemannian manifolds. Proc. Amer. Math. Soc. **141**, 233–244 (2013)
23. Kiwiel, K.C.: Methods of descent for nondifferentiable optimization. Lecture Notes in Mathematics, vol. 1133. Springer, Berlin (1985)
24. Klingenberg, W.: Riemannian Geometry. Walter de Gruyter Studies in Mathematics, vol. 1. Walter de Gruyter, Berlin (1995)
25. Kressner, D., Steinlechner, M., Vandereycken, B.: Low-rank tensor completion by Riemannian optimization. BIT Numer. Math. **54**, 447–468 (2014)
26. Lang, S.: Fundamentals of Differential Geometry. Graduate Texts in Mathematics, vol. 191. Springer, New York (1999)
27. Lee, P.Y.: Geometric Optimization for Computer Vision. PhD thesis, Australian National University (2005)
28. Lemarechal, C.: Nondifferentiable optimization. In: Nemhauser, G.L., et al. (eds.) Handbook in Operations Research and Management Science, vol. 1, pp. 529–572. North Holland, Amsterdam (1989)
29. Li, C., Mordukhovich, B.S., Wang, J., Yao, J.C.: Weak sharp minima on Riemannian manifolds. SIAM J. Optim. **21**(4), 1523–1560 (2011)
30. Mahony, R.E.: The constrained Newton method on a Lie group and the symmetric eigenvalue problem. Linear Algebra. Appl. **248**, 67–89 (1996)
31. Mahdavi-Amiri, N., Yousefpour, R.: An effective nonsmooth optimization algorithm for locally Lipschitz functions. J. Optim. Theory Appl. **155**, 180–195 (2012)
32. Moakher, M., Zerai, M.: The Riemannian geometry of the space of positive-definite matrices and its application to the regularization of positive-definite matrix-valued data. J. Math. Imaging Vision **40**, 171–187 (2011)
33. Papa Quiroz, E.A., Quispe, E.M., Oliveira, P.R.: Steepest descent method with a generalized Armijo search for quasiconvex functions on Riemannian manifolds. J. Math. Anal. Appl. **341**, 467–477 (2008)
34. Ring, W., Wirth, B.: Optimization methods on Riemannian manifolds and their application to shape space. SIAM J. Optim. **22**(2), 596–627 (2012)
35. Riddell, R.C.: Minimax problems on Grassmann manifolds. Sums of eigenvalues. Adv. Math. **54**, 107–199 (1984)
36. Rockafellar, R.T.: Convex Functions and Dual Extremum Problems, Thesis, Harvard (1963)
37. Rudin, L., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Phys. D **60**(1), 259–268 (1992)
38. Sander, O.: Geodesic finite elements for Cosserat rods. Internat. J. Numer. Methods Engrg. **82**, 1645–1670 (2010)
39. Smith, S.T.: Optimization techniques on Riemannian manifolds. Fields Institute Communications **3**, 113–146 (1994)
40. Udriste, C.: Convex Functions and Optimization Methods on Riemannian Manifolds. Kluwer Academic Publishers, Dordrecht (1994)
41. Usevich, K., Markovsky, I.: Optimization on a Grassmann manifold with application to system identification. Submitted, (http://homepages.vub.ac.be/imarkovs/t-abstracts.html)
42. Vandereycken, B., Vandewalle, S.: A Riemannian optimization approach for computing low-rank solutions of Lyapunov equations. SIAM. J. Matrix Anal. Appl. **31**(5), 2553–2579 (2010)

43. Vandereycken, B.: Low-rank matrix completion by Riemannian optimization. SIAM J. Optim. **23**(2), 1214–1236 (2013)
44. Yau, S.T.: Non-existence of continuous convex functions on certain Riemannian manifolds. Math. Ann. **207**, 269–270 (1974)
45. Zhang, L.S., Sun, X.L.: An algorithm for minimizing a class of locally Lipschitz functions. J. Optim. Theory Appl. **90**, 203–212 (1996)
46. Zhang, L.H.: Riemannian Newton method for the multivariate eigenvalue problem. SIAM J. Matrix Anal. Appl. **31**, 2972–2996 (2010)