SIAM J. OPTIM.
Vol. 27, No. 1, pp. 173–189

A RIEMANNIAN GRADIENT SAMPLING ALGORITHM FOR NONSMOOTH OPTIMIZATION ON MANIFOLDS*

SEYEDEHSOMAYEH HOSSEINI† AND ANDRÉ USCHMAJEW†

Abstract. In this paper, an optimization method for nonsmooth locally Lipschitz functions on complete Riemannian manifolds is presented. The method is based on approximating the subdifferential of the cost function at every iteration by the convex hull of transported gradients from tangent spaces at randomly generated nearby points to the tangent space of the current iterate and can hence be seen as a generalization of the well known gradient sampling algorithm to a Riemannian setting. A convergence result is obtained under the assumption that the cost function is bounded below and continuously differentiable on an open set of full measure and that the employed vector transport and retraction satisfy certain conditions, which hold, for instance, for the exponential map and parallel transport. Then with probability one the algorithm produces iterates at which the cost function is differentiable, and each cluster point of the iterates is a Clarke stationary point. Modifications yielding only ε-stationary points are also possible.

Key words. Riemannian optimization, locally Lipschitz functions, descent direction, Clarke subdifferential

AMS subject classifications. 49J52, 65K05, 58C05

DOI. 10.1137/16M1069298

1. Introduction.

We consider the optimization problem

$$\min f(x), \quad x \in M,$$

where M is a complete Riemannian manifold of dimension n and $f\colon M \to \mathbb{R}$ is locally Lipschitz on M. Many problems in machine learning, computer vision, pattern recognition, and signal processing are formulated as optimization problems on Riemannian manifolds; see [2, 3, 11, 12, 16, 17]. In particular, Stiefel and Grassmann manifolds arise naturally for eigenvalue problems [32, 37] and low-rank matrix and tensor optimization tasks [22, 28, 33, 34].

In nonlinear optimization on linear spaces, line search methods, which are based on updating the iterate by finding a descent direction and then adding a multiple of the obtained direction to the previous iterate, can be used. Burke, Lewis, and Overton [7] proposed and analyzed the gradient sampling (GS) algorithm for minimizing an objective function f that is locally Lipschitz. At each iteration, their proposed algorithm computes the gradients of the objective function $f\colon \mathbb{R}^n \to \mathbb{R}$ at the current iterate and at $m \geq n+1$ randomly generated nearby points. Since a locally Lipschitz function is almost everywhere differentiable, this step is successful with probability one. The convex hull of computed gradients is used to find an approximate ε-steepest descent direction by solving a quadratic program, where ε denotes the sampling radius. A backtracking Armijo line search along this direction then produces a candidate for the next iterate. This candidate is possibly further perturbed, if necessary, to stay in the set on which the objective function f is differentiable; this perturbation is random and small enough to retain the Armijo sufficient descent property. The sampling radius either is fixed for all iterations or is reduced dynamically; see [7] for all

*Received by the editors April 6, 2016; accepted for publication (in revised form) November 23, 2016; published electronically February 8, 2017.
http://www.siam.org/journals/siopt/27-1/M106929.html
†Hausdorff Center for Mathematics & Institute for Numerical Simulation, University of Bonn, 53115 Bonn, Germany (hosseini@ins.uni-bonn.de, uschmajew@ins.uni-bonn.de).

details. Regarding convergence, the essential statement is that if the GS algorithm converges to a point, this limit point is a Clarke stationary point for the cost function f with probability one, but to prove this an additional assumption is required that f is in fact continuously differentiable on an open set of full measure. The convergence analysis for the GS algorithm has been considerably strengthened and simplified by Kiwiel [23], whose reasoning we attempted to follow here. In particular, it is shown in [23] that every cluster point of the sequence of iterates is a Clarke stationary point for the cost function with probability one without assuming that the whole sequence is convergent.[1]

The subject of the present paper is the extension of the GS algorithm to optimization problems on Riemannian manifolds. Due to the nonlinear structure of the domain, we cannot use the usual techniques of optimization on linear spaces to study such problems, which means we need new tools. For example, the extension of line search methods to manifolds is achieved by the notion of retractions. The most important issue in such a method is finding a descent direction in the tangent space at the current iterate and then using the retractions to move along the descent direction to get the next iterate. In this way, further classical methods, such as Newton-type and trust-region methods, have been successfully generalized to problems on Riemannian manifolds for optimizing smooth objective functions; see [1, 2, 20, 21] and references therein. In a nondifferentiable problem, gradient information cannot be generally used to determine a direction in which the function is decreasing. Therefore, techniques of nonsmooth analysis have to be employed. Recently, some research has been started on nonsmooth optimization algorithms in a manifold setting; see [4, 5, 10, 14, 15, 24] and references therein. In [4], a cyclic proximal point algorithm on a Hadamard manifold is employed to minimize a nonsmooth function in this setting. Riemannian versions of the subgradient method have been studied in [5, 10]. In [14], a nonsmooth trust region method is generalized on Riemannian manifolds. The work [24] is concerned with the manifold alternating directions method of multipliers (MADMM), which can be seen as an extension of the classical ADMM scheme for manifold constrained nonsmooth optimization problems. In [15] the ε-subdifferential is defined on Riemannian manifolds by using the inverse of the derivative of the exponential map as a vector transport. Then, the ε-subdifferential is approximated by an iterative algorithm which starts with one element of the ε-subdifferential as a first iteration, and in every subsequent iteration, a new element of the ε-subdifferential is computed following some rule and added to the working set to improve the approximation. Afterward, this approximation is used in a nonsmooth minimization algorithm on Riemannian manifolds.

In this paper, we define the ε-subdifferential using isometric vector transports which satisfy a locking condition; see (2.3) below. Using the updated version of the ε-subdifferential we propose a Riemannian gradient sampling algorithm and present a convergence result for this algorithm under the assumptions that (i) the manifold has positive injectivity radius with respect to the used retraction (see section 2) and that (ii) the cost function is continuously differentiable on an open set of full measure.

[1] We note that in both papers [7] and [23] it is only assumed that f is continuously differentiable on an open and dense set D. In private communication, the authors of [7] have indicated that it was an oversight that the full measure assumption on D was not made, since open and dense does not imply full measure in general. Hence it cannot be stated that the sampled points are in D with probability one. However, as an alternative, if D has only positive measure in the neighborhood of a point, the algorithm could replace any sampled points that are not in D by continuing to sample points until they are in D, a process that must terminate with probability one.

We have been informed recently that a very similar Riemannian GS algorithm had been previously proposed in [19], but no convergence analysis of the method had been conducted.

Regarding the cost function, our convergence result is somewhat stronger than the one obtained in [15]; in particular following [23] we can dispense with the assumption of [15] that the objective function has compact level sets. On the other hand, this comes at the price that the Riemannian GS algorithm is considerably more expensive per iteration on manifolds of large dimension, since it requires by construction $\dim M + 1$ additional Riemannian gradient samples, whereas the algorithm in [15] often needs very few in practice (say, two or three) and hence has a complexity comparable to the Riemannian steepest descent method. As for the assumption on a positive injectivity radius, it excludes some interesting geometries like manifolds of fixed rank matrices or tensors, which have shown to be amenable to Riemannian optimization for some smooth problems [9, 25, 29, 35, 36]. We hope to overcome this current limitation in subsequent work.

Nevertheless, we believe that our generalization of the GS algorithm and its convergence analysis to a Riemannian setting closes a conceptual gap between nonsmooth optimization on linear spaces and manifolds and contributes to a deeper understanding of the mechanisms behind it. While the general reasoning in the convergence proof tries to mimic the arguments in linear space [7, 23], some nontrivial modifications are necessary. For instance, Kiwiel's analogue to our Lemma 4.3, namely, [23, Lemma 3.2], proves a concrete lower bound for the step sizes selected by the Armijo backtracking, which we were not able to achieve here. Instead, we use a modified nonconstructive argument.

The paper is organized as follows. In section 2 we collect some basic definitions and properties regarding Riemannian manifolds and locally Lipschitz functions on them. Moreover, two general lemmas taken from [23] are presented to be used in proving the convergence result. In section 3, the Riemannian gradient sampling algorithm is presented. The main part is section 4, which contains the convergence analysis. Finally in section 5, a simple numerical experiment on the so-called sparse vector problem is presented to illustrate the convergence result, as well as to compare the Riemannian GS algorithm with some other methods.

2. Preliminaries. We denote by $\operatorname{cl} N$, $\operatorname{int} N$, and $\operatorname{conv} N$ the closure, the interior, and the convex hull of a set N.

Riemannian manifolds. In this paper we consider a smooth manifold M of dimension

$$n := \dim M \geq 1$$

endowed with a Riemannian metric $\langle \cdot, \cdot \rangle$ on the tangent space $T_x M$ and assume that the manifold is *complete* with respect to the induced metric. This metric is the Riemannian distance, denoted by $\operatorname{dist}(x, y)$ for two points $x, y \in M$, and induces the same topology as the smooth manifold structure. We identify (via the Riemannian metric) $T_x M$ with the cotangent space at x, in particular, we will consider Riemannian subgradients at x as elements of the tangent space $T_x M$. The tangent bundle is TM. If we write $\xi_x \in TM$, it means in particular that $\xi_x \in T_x M$.

We denote by $B(x, \varepsilon) := \{y \in M : \operatorname{dist}(x, y) < \varepsilon)$ an open ball centered at x with radius ε.

Retractions. A smooth mapping $R\colon TM \to M$ is called a retraction if it has the following properties.

(i) $R_x(0_x) = x$ for every $x \in M$, where R_x is the restriction of R to T_xM and $0_x \in T_xM$ is the zero element of the tangent space of M at x.
(ii) $dR_x(0_x) = \mathrm{Id}_{T_xM}$, where Id_{T_xM} denotes the identity mapping on T_xM.

We further assume that there exists κ such that we have

$$\mathrm{dist}(R_x(\xi_x), x) \leq \kappa \, \|\xi_x\| \tag{2.1}$$

for all $x \in M$ and $\xi_x \in T_xM$. This poses no restriction in most situations of interest.

Vector transports. Of crucial importance will be a notion of tangent vector transport that is compatible with the retraction. A vector transport associated with a retraction R is defined as a continuous function $\mathcal{T} \colon TM \times TM \to TM$, $(\eta_x, \xi_x) \mapsto \mathcal{T}_{\eta_x}(\xi_x)$, which for all (η_x, ξ_x) satisfies the following conditions:

(i) $\mathcal{T}_{\eta_x} \colon T_xM \to T_{R(\eta_x)}M$ is a linear *invertible* map,
(ii) $\mathcal{T}_{0_x}(\xi_x) = \xi_x$.

In short, if $\eta_x \in T_xM$ and $R(\eta_x) = y$, then $\mathcal{T}_{\eta_x}$ transports vectors from the tangent space of M at x to the tangent space at y.

Two additional properties will be needed for the convergence result. First, the vector transport should preserve inner products, that is,

$$\langle \mathcal{T}_{\eta_x}(\xi_x), \mathcal{T}_{\eta_x}(\zeta_x) \rangle = \langle \xi_x, \zeta_x \rangle. \tag{2.2}$$

In particular, $\xi_x \mapsto \mathcal{T}_{\eta_x}(\xi_x)$ is then an isometry.

Second, we will assume that $\mathcal{T}$ satisfies the following condition, called *locking condition* in [21], for transporting vectors along their own direction:

$$\mathcal{T}_{\xi_x}(\xi_x) = \beta_{\xi_x} dR_x(\xi_x)(\xi_x), \quad \beta_{\xi_x} := \frac{\|\xi_x\|}{\|dR_x(\xi_x)(\xi_x)\|}, \tag{2.3}$$

where

$$dR_x(\xi_x)(\xi_x) = \frac{d}{dt} R_x(t\xi_x)|_{t=1}.$$

Geometrically, this condition states that the transport of a vector ξ_x along itself must be isometric and parallel to the velocity of the curve $t \mapsto R_x(t\xi_x)$. The locking condition will be crucial in the proof of Lemma 4.3 below, since it allows one to formulate a mean value theorem for the function $f \circ R_x$ in terms of transported tangent vectors.

The conditions (2.2) and (2.3) can be difficult to verify but are in particular satisfied for the (at least in theory) most natural choices of R and $\mathcal{T}$: the exponential map and the parallel transport. In this case $\beta_{\xi_x} = 1$. For further discussion, especially on construction of vector transport satisfying the locking condition, we refer to [21, section 4].

Injectivity radius. Since we aim at transporting subgradients from tangent spaces at nearby points of $x \in M$ to the tangent space at x, it will be important to know the range of R_x. Let

$$\iota(x) := \sup\{\varepsilon > 0 : B(x, \varepsilon) \subseteq R_x(T_xM) \text{ and } R_x \text{ is injective on } R_x^{-1}(B(x, \varepsilon))\}.$$

Then the *injectivity radius of M with respect to the retraction R* is defined as

$$\iota(M) := \inf_{x \in M} \iota(x).$$

When using the exponential map as a retraction, this definition coincides with the usual one.

Formally, the Riemannian gradient sampling algorithm presented in section 3 requires $\iota(M) > 0$. In particular, it needs an explicit positive lower bound on $\iota(M)$

as an input, which is hence assumed to be available. When M is compact, we at least know that $\iota(M) > 0$. As another example, we mention Hadamard manifolds for which $\iota(M) = \infty$ when using the exponential map as a retraction.

Using the injectivity radius with respect to R, we can introduce a more intuitive notation for vector transports:

$$\mathcal{T}_{x\to y}(\xi_x) := \mathcal{T}_{\eta_x}(\xi_x), \quad \text{and} \quad \mathcal{T}_{x\leftarrow y}(\xi_y) := (\mathcal{T}_{\eta_x})^{-1}(\xi_y) \quad \text{whenever } y = R_x(\eta_x).$$

It is clear that $\mathcal{T}_{x\leftarrow y}(\xi_y)$ is well defined for all $y \in B(x, \iota(x))$ and, in particular, for all $y \in B(x, \iota(M))$. In the following, when using the notation $\mathcal{T}_{x\leftarrow y}(\xi_y)$, it will always be ensured that it is well defined.

Lebesgue measure. The concept of Lebesgue measurability extends from $\mathbb{R}^n$ to smooth manifolds. If a maximal atlas of coordinate charts $\{(U_\alpha, \phi_\alpha)\}$, where $\phi_\alpha \colon U_\alpha \to V_\alpha$, is given, a set E in M is called Lebesgue measurable if we can find a covering of E using the given charts and $\phi_\alpha(U_\alpha \cap E)$ is Lebesgue measurable for each α. This concept is independent of the particular choice of covering. It is worth mentioning that the transition function is of course not necessarily preserving the measure of a set, but the nice point is that nullsets are mapped to nullsets; see [13]. In this article, we always consider the Lebesgue measure on M.

Locally Lipschitz functions. Let $f \colon M \to \mathbb{R}$ be a function defined on a Riemannian manifold M, then f is said to satisfy a Lipschitz condition of constant L on a given subset S of M if

$$|f(x) - f(y)| \le L \operatorname{dist}(x, y)$$

for all $x, y \in S$. A function f is said to be Lipschitz near $x \in M$ if it satisfies the Lipschitz condition of some constant 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 $x \in M$. Throughout this paper we consider locally Lipschitz functions on M.

Riemannian subdifferential. Let $f \colon M \to \mathbb{R}$ be a locally Lipschitz function, and

$$\Omega_f := \{x \in M : f \text{ is differentiable at } x\}.$$

Every locally Lipschitz function defined on a Riemannian manifold M is almost everywhere differentiable with respect to the Lebesgue measure on M, that is, $M \setminus \Omega_f$ is of measure zero. This follows from the Rademacher theorem in linear spaces, see [13, Theorem 2, p. 81], and the local equivalence of the Riemannian distance with the Euclidean distance in a chart.

We define the Riemannian subdifferential (in the sense of Clarke) of f at x, denoted by $\partial f(x)$, as

$$\partial f(x) := \operatorname{conv}\left\{ \lim_{\ell\to\infty} \operatorname{grad} f(x_\ell) : x_\ell \to x,\ x_\ell \in \Omega_f \right\} \subset T_xM, \tag{2.4}$$

where grad denotes the Riemannian gradient. Recall that the Riemannian gradient of f at x is the unique tangent vector $\operatorname{grad} f(x)$ at x such that $df(x)(\xi) = \langle \operatorname{grad} f(x), \xi\rangle$ for all $\xi \in T_xM$. It is also worthwhile to recall that $\lim_{\ell\to\infty} \operatorname{grad} f(x_\ell)$ in (2.4) has the following meaning. Let $(\xi_\ell) \subseteq TM$, $\xi_\ell \in T_{x_\ell}M$, be a sequence of vectors in TM. We say ξ_ℓ converges to ξ, denoted as $\lim_{\ell\to\infty} \xi_\ell = \xi$, if $x_\ell \to x$ and if for any smooth vector field X on M it holds that $\langle \xi_\ell, X(x_\ell)\rangle \to \langle \xi, X(x)\rangle$. Every element of the Riemannian subdifferential is called a (Riemannian) subgradient. One can prove that

the set $\partial f(x)$ is compact in T_xM [18]. Further, we have

$$\partial f(x) = \partial(f \circ R_x)(0_x) \tag{2.5}$$

for any retraction R, which may be seen as an alternative definition of $\partial f(x)$ relying on the definition of subdifferential on linear spaces.

A point $x \in M$ is called a stationary point of f if $0_x \in \partial f(x)$. A necessary condition that f achieves a local minimum at x is that x is a stationary point of f; see [15].

Riemannian ε-subdifferential. The notion of vector transport allows us to define the ε-subdifferential of a locally Lipschitz function f at a point $x \in M$ as follows:

$$\partial_\varepsilon f(x) := \operatorname{cl}\operatorname{conv}\{\mathcal{T}_{x\leftarrow y}(\partial f(y)) : y \in \operatorname{cl} B(x,\varepsilon)\}.$$

In this definition it is assumed that $\varepsilon < \iota(x)$. The set $\partial_\varepsilon f(x)$ is compact and convex in T_xM. For more details and properties of the ε-subdifferential see [15]. A point $x \in M$ is called an ε-stationary point of f if $0_x \in \partial_\varepsilon f(x)$.

Riemannian generalized directional derivative. For $x \in M$, let $\hat{f}_x = f \circ R_x$ denote the restriction of the pullback $\hat{f} = f \circ R$ to T_xM. The Clarke generalized directional derivative of f at x in the direction $w \in T_xM$, denoted by $f^\circ(x;w)$, is defined by $f^\circ(x;w) = \hat{f}_x^\circ(0_x;w)$, where $\hat{f}_x^\circ(0_x;w)$ denotes the Clarke generalized directional derivative of $\hat{f}_x \colon T_xM \to \mathbb{R}$ at 0_x in the direction $w \in T_xM$; see [8, 18]. The relation between the Clarke generalized directional derivative of f and its subdifferential reads as follows:

$$f^\circ(x;w) = \sup_{\xi\in\partial f(x)} \langle \xi, w\rangle = \max_{\xi\in\partial f(x)} \langle \xi, w\rangle.$$

This holds due to (2.5), since the corresponding result in linear spaces is well known [8]. Motivated by this formula, an ε-version (for $\varepsilon < \iota(x)$) of the generalized directional derivative is defined by

$$f_\varepsilon^\circ(x;w) := \sup_{\xi\in\partial_\varepsilon f(x)} \langle \xi, w\rangle = \max_{\xi\in\partial_\varepsilon f(x)} \langle \xi, w\rangle.$$

Descent direction. A direction $g \in T_xM$ is a descent direction at $x \in M$, if there exists $\alpha > 0$ such that for every $t \in (0,\alpha)$, we have

$$f(R_x(tg)) - f(x) = \hat{f}_x(tg) - \hat{f}_x(0_x) < 0.$$

It is known that if $f^\circ(x;g) = \hat{f}_x^\circ(0_x;g) < 0$, then g is a descent direction [27, Theorem 5.2.5]. Under some assumptions, the same can be shown for the ε-version.

PROPOSITION 2.1. *At $x \in M$, assume that the retraction R satisfies* (2.1) *for some $\kappa > 0$, the associated vector transport $\mathcal{T}$ satisfies the locking condition* (2.3)*, and $\varepsilon < \iota(x)$. If $f_\varepsilon^\circ(x;g) < 0$, then g is a descent direction at x.*

Proof. By Lebourg's mean value theorem for nonsmooth functions on Riemannian manifolds [18], there exists for every $t > 0$ a $t_0 \in [0,t]$ and $\xi \in \partial f(R_x(t_0 g))$, such that

$$f(R_x(tg)) - f(x) = \langle \xi, dR_x(t_0 g)(g)\rangle = \frac{1}{\beta_{t_0 g}} \langle \mathcal{T}_{x\leftarrow R_x(t_0 g)}(\xi), g\rangle,$$

where $\beta_{t_0 g}$ is the constant from the locking condition (2.3). Assuming that t is small enough, we can conclude that $R_x(t_0 g) \in B(x,\varepsilon)$ and $\mathcal{T}_{x\leftarrow R_x(t_0 g)}(\xi) \in \partial_\varepsilon f(x)$. Indeed,

it is enough to assume that $t < \frac{\varepsilon}{\|g\|\kappa}$ with κ from (2.1). Therefore,

$$f(R_x(tg)) - f(x) \leq \frac{1}{\beta_{t_0 g}} f_\varepsilon^\circ(x; g) < 0. \tag{2.6}$$

Hence g is a descent direction. ∎

The estimate (2.6) suggests that the best possible descent direction may be obtained by solving

$$\begin{aligned}\min_{\|g\|\leq 1, g\in T_xM} f_\varepsilon^\circ(x; g) &= \min_{\|g\|\leq 1, g\in T_xM} \max_{\xi\in\partial_\varepsilon f(x)} \langle \xi, g\rangle \\ &= \max_{\xi\in\partial_\varepsilon f(x)} \min_{\|g\|\leq 1, g\in T_xM} \langle \xi, g\rangle = -\min_{\xi\in\partial_\varepsilon f(x)} \|\xi\|,\end{aligned}$$

where the second equality is due to the minimax theorem. Indeed, in analogy to optimization in linear spaces, the following holds.

Proposition 2.2. *For $\varepsilon < \iota(x)$, let*

$$w = \operatorname{argmin}_{\xi\in\partial_\varepsilon f(x)} \|\xi\|, \quad g := -\frac{w}{\|w\|}. \tag{2.7}$$

Then $f_\varepsilon^\circ(x; g) = -\|w\|^2$. In particular, under the assumptions of Proposition 2.1, *if $w \neq 0$, then g is a descent direction.*

Proof. On the one hand, $f_\varepsilon^\circ(x, g) = \max_{\xi\in\partial_\varepsilon f(x)} \langle \xi, g\rangle \geq \langle w, g\rangle = -\|w\|$. On the other hand, we have the variational inequality $\langle w, w\rangle \leq \langle \xi, w\rangle$ for every $\xi \in \partial_\varepsilon f(x)$, which implies $f_\varepsilon^\circ(x, g) \leq -\|w\|$. In conclusion, $f_\varepsilon^\circ(x, g) = -\|w\|$. ∎

Although $-w$ obtained from (2.7) provides a descent direction, computing it poses a formidable task. Therefore, we need to find an approximation for the ε-subdifferential at x which can be computed algorithmically. The Riemannian gradient sampling algorithm finds a descent direction in the convex hull of randomly sampled Riemannian gradients at nearby points transported to the tangent space at x which serves as such an approximation.

Two lemmas. We conclude the preliminaries with two lemmas that will be crucial for the convergence proof but are of a more general nature. The first statement, which is a continuous perturbation of the variational characterization of the minimum norm element in a convex set, is taken from [23].

Lemma 2.3. *Assume that a nonempty compact convex set C in a Euclidean space does not contain zero. Then for every $\beta \in (0, 1)$ there exists $\nu > 0$ such that if $u, v \in C$ and $\|u\| \leq \min\{\|w\| : w \in C\} + \nu$, we deduce that $\langle v, u\rangle > \beta \|u\|^2$.*

The second lemma, which is an interesting fact about divergent sequences, is implicitly proved in $\mathbb{R}^n$ in the demonstration of [23, Theorem 3.3]. For convenience, we make it an explicit statement here and formulate it for metric spaces.

Lemma 2.4. *Let $(x_\ell)_{\ell\in\mathbb{N}}$ be a divergent sequence in a metric space, and let* dist *denote the metric. Then for every infinite convergent subsequence $(x_\ell)_{\ell\in\mathcal{L}}$, $\mathcal{L} \subset \mathbb{N}$, it holds that $\sum_{\ell\in\mathcal{L}} \operatorname{dist}(x_\ell, x_{\ell+1}) = \infty$.*

Proof. Let $(x_\ell)_{\ell\in\mathcal{L}}$ converge to $\bar{x}$. Then for any $\eta > 0$ the set $\mathcal{L}_\eta = \{\ell \in \mathcal{L} : \operatorname{dist}(x_\ell, \bar{x}) \leq \eta\}$ has infinitely many elements. But since $\bar{x}$ is not the limit of the whole sequence (x_ℓ), there exists η_0 such that $\mathbb{N}\setminus\mathcal{L}_{\eta_0}$ also has infinitely many elements.

Assume now $\sum_{\ell\in\mathcal{L}} \operatorname{dist}(x_\ell, x_{\ell+1}) < \infty$, then in particular

$$\sum_{\ell\in\mathcal{L}_{\eta_0}} \operatorname{dist}(x_{\ell+1}, x_\ell) < \infty. \tag{2.8}$$

For every $k \in \mathcal{L}_{\eta_0/2}$ there exists a smallest $k' \geq \ell$ such that $k' \in \mathbb{N} \setminus \mathcal{L}_{\eta_0}$. In other words, $\ell \in \mathcal{L}_{\eta_0}$ for all $k \leq \ell < k'$. Using the triangle inequality, we get

$$\eta_0/2 \leq \operatorname{dist}(x_k, x_{k'}) \leq \sum_{\ell=k}^{k'-1} \operatorname{dist}(x_\ell, x_{\ell+1}) \quad \text{for all } k \in \mathcal{L}_{\eta_0/2}.$$

This is a contradiction, since by (2.8) the right-hand side tends to zero for $k \to \infty$. ∎

3. Gradient sampling algorithm on Riemannian manifolds. The proposed Riemannian gradient sampling algorithm is given as Algorithm 1. Recall

Algorithm 1: Gradient sampling algorithm on manifolds.

Require: *Retraction R; injectivity radius $\iota(M)$; κ satisfying (2.1); vector transport $\mathcal{T}$; **measurable subset $D \subseteq \Omega_f$ such that $M \setminus D$ is of measure zero**.*
Input: $x_0 \in M \cap D$; $\delta_{opt} \geq 0$; $\delta_0 > 0$; $\varepsilon_{opt} \geq 0$; $0 < \varepsilon_0 < \iota(M)$; $\beta \in (0,1)$; $\theta_\varepsilon \in (0,1)$; $\theta_\delta \in (0,1)$; sampling size $m \geq n+1$.
for $\ell = 0, 1, 2, \ldots$ **do**
  Choose m points $\{x_\ell^i\}_{i=1}^m$ independently and uniformly from $B(x_\ell, \varepsilon_\ell)$.
  `// gradient sampling`
  if $\{x_\ell^i\}_{i=1}^m \not\subset D$ **then**
    return `// abort`
  end
  Let

$$G_\ell := \operatorname{conv}\{\operatorname{grad} f(x_\ell), \mathcal{T}_{x_\ell \leftarrow x_\ell^1}(\operatorname{grad} f(x_\ell^1)), \ldots, \mathcal{T}_{x_\ell \leftarrow x_\ell^m}(\operatorname{grad} f(x_\ell^m))\},$$

  and find

$$w_\ell = \operatorname{argmin}\{\|w\| : w \in G_\ell\}.$$

  if $\|w_\ell\| \leq \delta_{opt}$ *and* $\varepsilon_\ell \leq \varepsilon_{opt}$ **then**
    return `// success`
  end
  if $\|w_\ell\| \leq \delta_\ell$ **then**
    $\varepsilon_{\ell+1} := \theta_\varepsilon \varepsilon_\ell$, $\delta_{\ell+1} := \theta_\delta \delta_\ell$
    $x_{\ell+1} := x_\ell$
  else
    $\varepsilon_{\ell+1} = \varepsilon_\ell$, $\delta_{\ell+1} = \delta_\ell$, $g_\ell := -\dfrac{w_\ell}{\|w_\ell\|}$ `// descent direction`
    $t_\ell := \max\{t : f(R_{x_\ell}(tg_\ell)) - f(x_\ell) < -\beta t \|w_\ell\|, t \in \{1, \gamma, \gamma^2, \ldots\}\}$ `// line search`
    if $R_{x_\ell}(t_\ell g_\ell) \in D$ **then**
      $x_{\ell+1} := R_{x_\ell}(t_\ell g_\ell)$
    else
      Find $x_{\ell+1} \in D$ such that $f(x_{\ell+1}) - f(x_\ell) < -\beta t_\ell \|w_\ell\|$ `// stay in D`
      and $\operatorname{dist}(R_{x_\ell}(t_\ell g_\ell), x_{\ell+1}) \leq \kappa t_\ell$.
    end
  end
end

$n = \dim M \geq 1$. It is assumed that $\iota(M) > 0$. Note that the algorithm is formulated for an arbitrary subset $D \subseteq \Omega_f$ of full measure. The convergence analysis below assumes that D is additionally open and that $\operatorname{grad} f$ is continuous on D; see Assumption 4.1.

Let us make some comments. First, we should mention that the line search in line 16 of the algorithm is well defined and t_ℓ can be found using a finite process. To see this, observe that for $w_\ell = \operatorname{argmin}\{\|w\| : w \in G_\ell\}$ and $g_\ell = \frac{-w_\ell}{\|w_\ell\|}$ it holds that

$$\langle \operatorname{grad} f(x_\ell), g_\ell \rangle \leq \sup_{w \in G_\ell} \langle w, g_\ell \rangle = \langle w_\ell, g_\ell \rangle = -\|w_\ell\|,$$

the first equality being due to convexity of G_ℓ as defined in line 7 of Algorithm 1. Since $x_\ell \in D$, we deduce that $(f \circ R_{x_\ell})'(0_{x_\ell}, g_\ell)$, which is the directional derivative of $f \circ R_{x_\ell}$ at 0_{x_ℓ} in the direction g_ℓ, exists and is equal to $\langle \operatorname{grad} f(x_\ell), g_\ell \rangle$. Therefore, there exists $\alpha > 0$ such that

$$f \circ R_{x_\ell}(tg_\ell) - f \circ R_{x_\ell}(0_{x_\ell}) < t\beta \langle \operatorname{grad} f(x_\ell), g_\ell \rangle \leq -t\beta \|w_\ell\|$$

for all $t \in (0, \alpha)$.

Next, we note that, in line with standard gradient sampling in linear spaces [7], the algorithm keeps the iterates in D by construction. However, the case that $R_{x_\ell}(t_\ell g_\ell) \notin D$ that would require adjustment is very unlikely to ever happen in practice, although we are not able to prove rigorously that it is a zero probability event. In any case the adjustment in lines 20 and 21 can be implemented easily as follows. Assuming $R_{x_\ell}(t_\ell g_\ell) \notin D$, we simply continue drawing $x_{\ell+1}$ uniformly from $B(R_{x_\ell}(t_\ell g_\ell), \kappa t_\ell / k)$, $k = 1, 2 \ldots$, until $x_{\ell+1} \in D$ and $f(x_{\ell+1}) - f(x_\ell) < -\beta t_\ell \|w_\ell\|$. By continuity of f and the inequality $f(R_{x_\ell}(tg_\ell)) - f(x_l) < -\beta t_\ell \|w_\ell\|$, this process terminates with probability one.

Finally, we note for later reference that the iterates satisfy

$$\operatorname{dist}(x_\ell, x_{\ell+1}) \leq 2\kappa t_\ell, \tag{3.1}$$

since either $x_{\ell+1} = x_\ell$, $x_{\ell+1} = R_{x_\ell}(t_\ell g_\ell)$, or $x_{\ell+1}$ is given from the lines 20 and 21 of Algorithm 1. The inequality is clear for the first case, while for the second case it follows from (2.1). For the last case, we note that (2.1) and line 21 of Algorithm 1 yield

$$\operatorname{dist}(x_\ell, x_{\ell+1}) \leq \operatorname{dist}(R_{x_\ell}(tg_\ell), x_{\ell+1}) + \operatorname{dist}(R_{x_\ell}(tg_\ell), x_\ell) \leq 2\kappa t_\ell.$$

4. Convergence result. The convergence result for Algorithm 1 as stated in Theorem 4.4 below requires the following main assumptions.

Assumption 4.1.

(i) The function f is locally Lipschitz and continuously differentiable on an open subset $D \subseteq \Omega_f \subseteq M$, which is of full measure (see footnote 1).

(ii) The employed vector transport associated with the given retraction preserves inner products in the sense of (2.2) and satisfies the locking condition (2.3).

With these assumptions, we introduce some further notation, always restricting to $\varepsilon < \iota(M)$.

First, we define the following sets

$$G_\varepsilon(x) := \operatorname{cl} \operatorname{conv}\{\mathcal{T}_{x \leftarrow y}(\operatorname{grad} f(y)) : y \in \operatorname{cl} B(x, \varepsilon) \cap D\}.$$

It is easy to see that $G_\varepsilon(x) \subseteq \partial_\varepsilon f(x)$ for every $x \in M$. Moreover, $\partial_{\varepsilon_0} f(x) \subseteq G_\varepsilon(x)$ for $\varepsilon_0 < \varepsilon$. We further denote

$$\rho_\varepsilon(x) := \min\{\|w\| : w \in G_\varepsilon(x)\}.$$

Next, let

$$D_\varepsilon(x) := \operatorname{cl} B(x,\varepsilon) \cap D, \quad D_\varepsilon^m(x) := \prod_1^m D_\varepsilon(x),$$

where $\prod_1^m$ denotes an m-fold Cartesian product, then for $x, \bar{x} \in M$ and $\nu > 0$ we define sets

$$V_\varepsilon(\bar{x}, x, \nu) := \{y = (y^1, \dots, y^m) \in D_\varepsilon^m(x) : \tilde{\rho}_\varepsilon(y) \le \rho_\varepsilon(\bar{x}) + \nu\},$$

where

$$\tilde{\rho}_\varepsilon(y) := \min\{\|w\| : w \in \operatorname{conv}\{\mathcal{T}_{\bar{x}\leftarrow y^i}(\operatorname{grad} f(y^i))\}_{i=1}^m\}.$$

The following lemma states that if x is chosen in a small neighborhood of $\bar{x}$, then $V_\varepsilon(\bar{x}, x, \nu)$ contains a nonempty open set.

LEMMA 4.2. *Let $0 < \varepsilon < \iota(M)$, $\bar{x} \in M$, and let $f : M \to \mathbb{R}$ satisfy Assumption* 4.1(i). *For any $\nu > 0$, there exist $\tau > 0$ and a nonempty open set $\hat{V} = \hat{V}(\bar{x}, \varepsilon, \tau)$ such that $\operatorname{cl} \hat{V} \subseteq V_\varepsilon(\bar{x}, x, \nu)$ for all $x \in B(\bar{x}, \tau)$.*

Proof. Let $w \in G_\varepsilon(\bar{x})$ such that $\rho_\varepsilon(\bar{x}) = \|w\|$. The set $\operatorname{conv}\{\mathcal{T}_{\bar{x}\leftarrow y}(\operatorname{grad} f(y)) : y \in \operatorname{cl} B(\bar{x}, \varepsilon) \cap D\}$ is dense in $G_\varepsilon(\bar{x})$ by definition. Hence, using Carathéodory's theorem and the continuity of $y \mapsto \mathcal{T}_{\bar{x}\leftarrow y}(\operatorname{grad} f(y))$ on D, we can find $\tilde{y} = (\tilde{y}^1, \dots, \tilde{y}^m) \in \prod_1^m B(\bar{x}, \varepsilon) \cap D$ and nonnegative $\lambda_1, \dots, \lambda_m$ with $\lambda_1 + \cdots + \lambda_m = 1$ such that

$$u = \sum_{i=1}^m \lambda_i \mathcal{T}_{\bar{x}\leftarrow \tilde{y}^i}(\operatorname{grad} f(\tilde{y}^i))$$

satisfies $\|u\| \le \rho_\varepsilon(\bar{x}) + \nu/3$.

Now there is an $0 < \bar{\varepsilon} < \varepsilon$, such that

$$\tilde{V} := \prod_{i=1}^m B(\tilde{y}^i, \bar{\varepsilon}) \subseteq D_{\varepsilon-\bar{\varepsilon}}^m(\bar{x}), \tag{4.1}$$

and moreover

$$\left\| \sum_{i=1}^m \lambda_i \mathcal{T}_{\bar{x}\leftarrow y^i}(\operatorname{grad} f(y^i)) \right\| \le \rho_\varepsilon(\bar{x}) + \nu \tag{4.2}$$

for all $y = (y^1, \dots, y^m) \in \tilde{V}$. Now pick any $0 < \tau \le \bar{\varepsilon}$ and $x \in B(\bar{x}, \tau)$. Then, by (4.1), $\tilde{V} \subseteq D_\varepsilon(x)$, and therefore $\tilde{V} \subseteq V_{\bar{\varepsilon}}(\bar{x}, x, \nu)$ by (4.2). Since $\tilde{V}$ is open, there exists an open subset $\hat{V}$ such that $\operatorname{cl} \hat{V} \subset \tilde{V}$, which implies that $\operatorname{cl} \hat{V} \subseteq V_\varepsilon(\bar{x}, x, \nu)$. ∎

The next observation is key to the convergence result.

LEMMA 4.3. *Assume that $(x_\ell)_{\ell \in \mathcal{L}}$ is a subsequence of iterates constructed by Algorithm* 1 *that converges to some $\bar{x} \in M$. Suppose that*

(i) *$\varepsilon_\ell = \varepsilon$ is constant for all $\ell \in \mathcal{L}$,*
(ii) *$0 \notin G_\varepsilon(\bar{x})$,*
(iii) *$\liminf_{\ell \in \mathcal{L}} t_\ell = 0$.*

Let $\nu > 0$ be taken from Lemma 2.3 for $C = G_\varepsilon(\bar{x})$ and τ and $\hat{V}$ be obtained from Lemma 4.2 for this ν. Then $(x_\ell^1, \dots, x_\ell^m) \in \hat{V}(\bar{x}, \varepsilon, \tau)$ can be true only for finitely many $\ell \in \mathcal{L}$.

Proof. For notational convenience we denote $x_\ell^0 := x_\ell$, so that

$$w_\ell := \sum_{i=0}^{m} \lambda_\ell^i \mathcal{T}_{x_\ell \leftarrow x_\ell^i}(\operatorname{grad} f(x_\ell^i))$$

is obtained at the ℓth iteration of the algorithm. We argue by contradiction. Without loss of generality we can hence assume that $(x_\ell^1, \dots, x_\ell^m) \in \hat{V}(\bar{x}, \varepsilon, \tau)$ for all $\ell \in \mathcal{L}$. We may also assume $t_\ell < 1$ for all $\ell \in \mathcal{L}$ and $t_\ell \to 0$ for $\ell \to \infty$, $\ell \in \mathcal{L}$. In the following we only consider $\ell \in \mathcal{L}$ without explicitly noting it anymore.

By construction, since $t_\ell < 1$ is assumed, $\gamma^{-1} t_\ell$ fails the Armijo condition, that is,

$$-\beta\gamma^{-1} t_\ell \|w_\ell\| \le f(R_{x_\ell}(\gamma^{-1} t_\ell g_\ell)) - f(x_\ell). \tag{4.3}$$

By Lebourg's mean value Theorem, there exists $t \in [0, \gamma^{-1} t_\ell]$ and $v_\ell \in \partial f(R_{x_\ell}(t g_\ell))$, such that

$$f(R_{x_\ell}(\gamma^{-1} t_\ell g_\ell)) - f(x_\ell) = \langle v_\ell, dR_{x_\ell}(t g_\ell)(g_\ell) \rangle.$$

Hence, using the locking condition (2.3),

$$f(R_{x_\ell}(\gamma^{-1} t_\ell g_\ell)) - f(x_\ell) = \frac{t}{\beta_{t g_\ell}} \langle \mathcal{T}_{x_\ell \leftarrow R_{x_\ell}(t g_\ell)}(v_\ell), g_\ell \rangle.$$

In combination with (4.3), we conclude that

$$\langle \mathcal{T}_{x_\ell \leftarrow R_{x_\ell}(t g_\ell)}(v_\ell), w_\ell \rangle \le \beta \beta_{t g_\ell} \|w_\ell\|^2. \tag{4.4}$$

Since $(x_\ell^1, \dots x_\ell^m) \in \hat{V} \subseteq V_\varepsilon(\bar{x}, \bar{x}, \nu)$, there exists a subsequence of $(x_\ell^1, \dots x_\ell^m)$ converging to some $(z^1, \dots, z^m) \in \operatorname{cl} \hat{V} \subseteq V_\varepsilon(\bar{x}, \bar{x}, \nu)$ (as M is complete). Denoting $\xi^i = \mathcal{T}_{\bar{x} \leftarrow z^i}(\operatorname{grad} f(z^i))$ we have

$$\min\{\|w\| : w \in \operatorname{conv}\{\xi^i\}_{i=1}^m\} \le \rho_\varepsilon(\bar{x}) + \nu,$$

since $\operatorname{grad} f$ is continuous on $\operatorname{cl} \hat{V}$. We extract a further subsequence for which $\operatorname{grad} f(x_\ell)$ converges to some $\xi^0 \in \partial f(\bar{x}) \subseteq G_\varepsilon(\bar{x})$. Then,

$$\min\{\|w\| : w \in \operatorname{conv}\{\xi^i\}_{i=0}^m\} \le \rho_\varepsilon(\bar{x}) + \nu. \tag{4.5}$$

We assume that the minimum is attained at some

$$\tilde{w} = \sum_{i=0}^{m} \tilde{\lambda}^i \xi^i.$$

Since $z^i \in \operatorname{cl} B(\bar{x}, \varepsilon)$, $i = 1, \dots, m$, it holds that $\tilde{w} \in G_\varepsilon(\bar{x})$ and $\|\tilde{w}\| \le \rho_\varepsilon(\bar{x}) + \nu$. Hence by Lemma 2.3,

$$\langle v, \tilde{w} \rangle > \beta \|\tilde{w}\|^2 \tag{4.6}$$

for every $v \in G_\varepsilon(\bar{x})$. Now, we claim that (w_ℓ) has a subsequence convergent to $\tilde{w}$. Then, since $t_\ell \to 0$, $x_\ell \to \bar{x}$, and $v_\ell \in \partial f(R_{x_\ell}(tg_\ell))$ has a convergent subsequence to some $v \in \partial f(\bar{x}) \subset G_\varepsilon(\bar{x})$, limiting (4.4) in subsequences gives us

$$\langle v, \tilde{w} \rangle \leq \beta \left\| \tilde{w} \right\|^2,$$

which contradicts (4.6) and hence proves the lemma.

To prove the existence of such a subsequence of (w_ℓ), we take a common subsequence of the subsequences considered above for which the λ_ℓ^i converge to some $\lambda_\star^i$, $i = 0, 1, \ldots, m$. Then w_ℓ converges to

$$w_\star = \sum_{i=0}^{m} \lambda_\star^i \xi^i,$$

which is an element of the set for which $\tilde{w}$ is the minimum norm element (see (4.5)). Since this minimum norm element is unique, it is enough to show that $\|w_\star\| \leq \|\tilde{w}\|$. Assume the opposite, that $\|\tilde{w}\| \leq \|w_\star\| - \eta$ for some $\eta > 0$. Then consider ℓ large enough such that

$$\left\| \sum_{i=0}^{m} \tilde{\lambda}^i \left(\xi^i - \mathcal{T}_{\bar{x} \leftarrow x_\ell} \mathcal{T}_{x_\ell \leftarrow x_\ell^i} (\operatorname{grad} f(x_\ell^i)) \right) \right\| \leq \eta/3$$

and

$$\|w_\ell\| \geq \|w_\star\| - \eta/3.$$

Using the isometry property of $\mathcal{T}_{\bar{x} \leftarrow x_\ell}$ and the triangle inequality, it follows that

$$\left\| \sum_{i=0}^{m} \tilde{\lambda}^i \mathcal{T}_{x_\ell \leftarrow x_\ell^i} (\operatorname{grad} f(x_\ell^i)) \right\| \leq \|\tilde{w}\| + \eta/3 \leq \|w_\star\| - 2\eta/3 \leq \|w_\ell\| - \eta/3,$$

which contradicts the choice of w_ℓ. ∎

We are now in the position to prove subsequential convergence of Algorithm 1 to Clarke stationary points under Assumption 4.1.

THEOREM 4.4. *At any iteration of Algorithm* 1, *the event that it terminates due to activation of the if-clause in line* 4 *has zero probability. Let* (x_ℓ) *be an infinite sequence generated by the algorithm with* $\delta_{opt} = \varepsilon_{opt} = 0$. *Then either* $f(x_\ell) \downarrow -\infty$ *or* $\delta_\ell \downarrow 0$, $\varepsilon_\ell \downarrow 0$ *and every cluster point of the sequence of iterations is a stationary point for* f.

Proof. It is clear that a termination due to line 4 has zero probability, since D is assumed to have full measure. We consider the case that an infinite sequence (x_ℓ) is generated and $\liminf_\ell f(x_\ell) > -\infty$.

By construction, we have that $f(x_{\ell+1}) - f(x_\ell) < -\beta t_\ell \|w_\ell\|$. Therefore, using telescopic sums,

$$\sum_{\ell=1}^{\infty} t_\ell \|w_\ell\| < \infty. \tag{4.7}$$

Due to estimate (3.1), this also implies

$$\sum_{\ell=1}^{\infty} \operatorname{dist}(x_{\ell+1}, x_\ell) \|w_\ell\| < \infty. \tag{4.8}$$

We first prove that $\delta_\ell \downarrow 0$. Assume to the contrary that there exists ℓ^* such that $\delta_\ell = \delta$ remain fixed for all $\ell \geq \ell^*$. This only happens if line 11 in Algorithm 1 is not activated anymore. In particular, this means that $\varepsilon_\ell = \varepsilon$ also remains fixed and $\|w_\ell\| > \delta$ for $\ell \geq \ell^*$. By (4.7) and (4.8), the latter implies $t_\ell \to 0$ and $\sum_{\ell=1}^{\infty} \operatorname{dist}(x_{\ell+1}, x_\ell) < \infty$. So x_ℓ is a Cauchy sequence and has a limit $\bar{x} \in M$. We consider two cases. In the first case, $0 \notin G_\varepsilon(\bar{x})$. Let ν, τ, and $\hat{V} = \hat{V}(\bar{x}, \varepsilon, \tau)$ be chosen as in Lemma 4.3. Since $t_\ell \to 0$, this lemma states that we can have $(x_\ell^1, \dots, x_\ell^m) \in \hat{V}$ only a finite number of times. This is a contradiction, because $(x_\ell^1, \dots, x_\ell^m)$ are sampled independently and uniformly from $D_\varepsilon^m(x_\ell)$, and $\hat{V}$ is a nonempty open subset of $D_\varepsilon^m(x_\ell)$. In the second case, $0 \in G_\varepsilon(\bar{x})$. Then $\rho_\varepsilon(\bar{x}) = 0$. We consider $\nu := \delta/2$ and choose τ and $\hat{V} = \hat{V}(\bar{x}, \tau, \nu)$ according to Lemma 4.2. By the same argument as before, we must have $(x_\ell^1, \dots, x_\ell^m) \in \hat{V}$ infinitely often. Also, we have $x_\ell \in B(\bar{x}, \tau)$ for ℓ large enough. Then

$$\begin{aligned}
&\min\{\|w\| : w \in \operatorname{conv}\{\mathcal{T}_{\bar{x} \leftarrow x_\ell^i}(\operatorname{grad} f(x_\ell^i))\}_{i=1}^m\} \\
&\quad \leq \rho_\varepsilon(\bar{x}) + \nu = \delta/2 \leq \|w_\ell\| - \delta/2 \\
&\quad \leq \min\{\|w\| : w \in \operatorname{conv}\{\mathcal{T}_{\bar{x} \leftarrow x_\ell^i}(\operatorname{grad} f(x_\ell^i))\}_{i=1}^m\} - \delta/2.
\end{aligned}$$

However, along similar lines as in the proof of Lemma 4.3 (by considering a convergent subsequence $(x_\ell^1, \dots, x_\ell^m) \to (z^1, \dots, z^m)$), we can show that both sequences of minima have the same limit inferior, and hence obtain a contradiction. In summary, we conclude that line 11 in Algorithm 1 is activated infinitely, that is, $\delta_\ell \downarrow 0$ and $\|w_\ell\|$ has a subsequence converging to 0. Since $\theta_\varepsilon \in (0, 1)$, this also implies $\varepsilon_\ell \downarrow 0$.

Suppose now that (x_ℓ) has a cluster point $\bar{x}$. If $x_\ell \to \bar{x}$, then a subsequence of w_ℓ converges to $0 \in T_{\bar{x}}M$. Since $w_\ell \in \partial_{\varepsilon_\ell} f(x_\ell)$ and $\partial_{\varepsilon_\ell} f(x_\ell)$ has a closed graph, we deduce that $0 \in \partial f(\bar{x})$. If x_ℓ does not converge to $\bar{x}$, let $(x_\ell)_{\ell \in \mathcal{L}}$ be a subsequence that does. Repeating a reasoning by Kiwiel [23], it follows from (4.8) that $\liminf_{\ell \in \mathcal{L}} \|w_\ell\| = 0$, since otherwise $\sum_{\ell \in \mathcal{L}} \operatorname{dist}(x_\ell, x_{\ell+1}) < \infty$ in contradiction to Lemma 2.4. Hence we get the same conclusion. ∎

We note that the previous convergence theorem yields a Clarke stationary point in the exact sense. If only ε-optimality is required, the algorithm can be modified as follows to keep the sampling radius fixed.

THEOREM 4.5. *Let (x_ℓ) be a sequence generated by the algorithm with $\delta_0 = \delta_{opt} = 0$, $0 < \varepsilon_0 = \varepsilon_{opt} = \varepsilon < \iota(M)$. Then with probability one either the algorithm stops at some iteration ℓ with $\|w_\ell\| = 0$, or $f(x_\ell) \downarrow -\infty$, or there exists a subsequence $(x_\ell)_{\ell \in \mathcal{L}}$ such that $w_\ell \to 0$ for $\ell \in \mathcal{L}$ and every cluster point of $(x_\ell)_{\ell \in \mathcal{L}}$ is an ε-stationary point.*

The proof is almost verbatim to the one of Theorem 4.4. The only difference is that due to $\delta_0 = \delta_{opt} = 0$ line 11 will never be activated in Algorithm 1, which is obvious by induction ($\delta_\ell = \delta_0 = 0$ for all ℓ). So in fact, lines 11–13 could be entirely removed in this modification. However, the arguments showing that the assumption $\|w_\ell\| > \delta$ for all $\ell \geq \ell^*$ and some $\delta > 0$ yields a contradiction are not affected. Hence there exists a subsequence $(w_\ell)_{\ell \in \mathcal{L}}$ converging to zero. In turn, noting $\varepsilon_\ell = \varepsilon$ is fixed, $w_\ell \in \partial_\varepsilon f(x_\ell)$ implies $0 \in \partial_\varepsilon f(\bar{x})$ for any cluster point $\bar{x}$ of the sequence $(x_\ell)_{\ell \in \mathcal{L}}$.

5. Numerical experiment. As an application we consider a problem that was recently discussed in [31]. The goal is to find the sparsest vector in an n-dimensional linear subspace W of $\mathbb{R}^m$. Letting $Q \in \mathbb{R}^{m \times n}$ denote a matrix whose columns form an orthonormal basis for W, the *sparse vector problem* reads

$$\min \|Qx\|_0, \quad x \in S,$$

where S is the Euclidean unit sphere in $\mathbb{R}^n$ and $\|Qx\|_0$ is the number of nonzero elements of Qx. Since this function is not locally Lipschitz, we replace it with the 1-norm as a surrogate. This leads to the problem

$$\min \|Qx\|_1, \quad x \in S, \tag{5.1}$$

which fits into our framework with $M = S$ and $f(x) = \|Qx\|_1$.

The natural Riemannian metric on S is induced from the ambient space $\mathbb{R}^n$, and the function f is easily shown to be locally Lipschitz with respect to the corresponding Riemannian distance $\operatorname{dist}(x, y) = \arccos\langle x, y\rangle$. Hence it is almost everywhere differentiable on S. Riemannian gradients are then obtained by orthogonal projections of Euclidean gradients on the tangent spaces. In our example,

$$\operatorname{grad} f(x) = (I_n - xx^T)Q^T \operatorname{sign}(Qx),$$

where sign is the elementwise sign function, I_n is the $n \times n$ identity matrix, and x is considered a column vector. Note that f is differentiable at x if and only if the entries of Qx do not change their sign in a whole neighborhood of $x \in S$ (which includes the case of being constantly zero). So, if f is differentiable at x, it is actually continuously differentiable in a whole neighborhood of $x \in S$. In other words, the dense set Ω_f of differentiable points is open and since f is also continuously differentiable on Ω_f, we may take $D = \Omega_f$.

In our implementation of Algorithm 1 we use the exponential map as a retraction,

$$R_x(\xi) := \exp_x(\xi) = \cos(\|\xi\|)x + \sin(\|\xi\|)\frac{\xi}{\|\xi\|},$$

and parallel transport as vector transport,

$$\mathcal{T}_{x\to\gamma(t)}(\xi) := (I_n + (\cos(\|\dot\gamma(0)t\|) - 1)uu^T - \sin(\|\dot\gamma(0)t\|)xu^T)\xi,$$

where γ is a geodesic on S with $\gamma(0) = x$, and $u = \frac{\dot\gamma(0)}{\|\dot\gamma(0)\|}$. Note that $\mathcal{T}_{x\leftarrow y}(\xi_y) = \mathcal{T}_{y\to\sigma(1)}(\xi_y)$, where $\sigma(t) = \exp_y(tv)$ denotes the geodesic connecting y to x. It is obtained using

$$v = \frac{\operatorname{dist}(x, y)}{\|(I - xx^T)(y - x)\|}(I - xx^T)(y - x).$$

So explicitly, we have

$$\mathcal{T}_{x\leftarrow y}(\xi_y) = (I_n + (\cos(\|v\|) - 1)uu^T - \sin(\|v\|)yu^T)\xi_y, \quad u = \frac{v}{\|v\|}.$$

This is well defined for all $y \neq -x$. (It holds that $\iota(M) = \pi$.) We have implemented all these operations on the sphere using the Manopt toolbox [6]. So, Assumption 4.1 is satisfied in this setup, and our convergence result applies.

The algorithm is implemented in MATLAB using IEEE double precision arithmetic. In order to solve the convex quadratic program in line 7 of Algorithm 1, we use the QP solver `quadprog` from the MATLAB optimization toolbox. The sample size is considered to be $\dim M + 1$. We use $\varepsilon_0 = 1$, $\theta_\varepsilon = 0.1$, $\delta = 10^{-6}$, $\theta_\delta = 0.1$. We set the backtracking reduction factor equal to 0.5 and the Armijo parameter $\beta = 10^{-4}$. The maximum number of iterations is set to be 5000. As Algorithm 1 is implemented in finite precision, two changes must be made. Specifically, we do not check in line 4

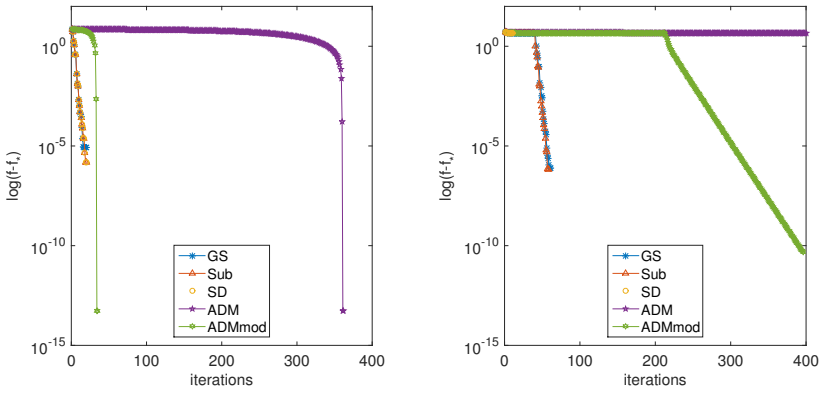

FIG. 1. *Results for a ten-dimensional subspace in* $\mathbb{R}^{100}$. *Left: minimum function value is* $f_* = 1$. *Right: minimum function value is* $f_* = \sqrt{7} \approx 2.646$.

whether the sample points lie in the set D on which f is differentiable and do not implement lines 17–21 either.

We compare Algorithm 1 (denoted as "GS") with two other basic methods. The ε-subdifferential algorithm from [15] (denoted as "Sub" in the plots) uses a systematic way to add subgradients to the approximation of the ε-subdifferential. The method proposed in [31] for solving (5.1) (denoted as "ADM") is an alternating optimization method, in which one iteration consists in applying a soft thresholding operator to Qx with a fixed shift λ to obtain $y = \max(Qx - \lambda\mathbf{1}, 0)$ with fewer nonzero entries (here $\mathbf{1}$ is the vector containing all ones) and then projecting back to the sphere by setting $x = Q^T y/\|Q^T y\|$. A modification suggested in [30] is to choose the shift λ adaptively according to the number nnz of nonzero entries that remained after the previous application of soft thresholding. (We used $\lambda = 0.1/\sqrt{nnz}$.) Furthermore, soft thresholding is replaced by hard thresholding once the number nnz remained unchanged sufficiently often (for 60 iterations in our experiments). This method is denoted as "ADMmod." Finally, we also plot the result for the simple Riemannian steepest method (denoted as "SD"), which is possible since with probability one f is differentiable at all iterates.

In Figure 1 we see the outcome of two experiments in terms of produced function values. In both cases $m = 100$ and $n = 10$. In the first example, the subspace W is the linear hull of the unit vector $e_1 = (1, 0, \ldots, 0)$ and nine random vectors. Hence the minimum function value of f is one, and with probability one it is only attained at $\pm e_1$. All methods recover $\pm e_1$; the error curves are omitted since they look similar. In the second example, W is generated in the same way but using $e = (1, 1, 1, 1, 1, 1, 1, 0 \ldots, 0)$ instead of e_1. The minimum value for f on the sphere is likely to equal $\sqrt{7}$. All methods used the same randomly generated initial guess.

The GS, Sub, and ADMmod methods are successful in both scenarios. As can be seen in the right plot, the unmodified ADM method is typically unable to recover vectors with more than one nonzero entry (see [30] for an explanation), whereas ADMmod succeeds by switching to hard thresholding after a phase of stagnation. The SD method stagnated here as well. This behavior of SD was not observed very often, but we included this plot to emphasize that it can happen, whereas GS and Sub always succeeded.

We note that while all five methods have been put in a single plot for convenience, the comparison by iteration number is not necessarily meaningful since the methods

are quite different. Timings of the algorithms are not given since no emphasis was put on efficient implementation. However, we can comment that the ADM methods were by far the fastest, which is clear since they involve no gradients and no line searches. Among the Riemannian methods, GS was, as expected, the slowest, since it computes the minimum in a convex hull of $1+(n+1)$ Riemannian gradients, whereas SD uses only one gradient, and Sub typically one to three, without being offset too much by the additional cost of the more systematic sampling procedure. On the other hand, there is no convergence result for SD available (in fact, simple counter-examples to convergence exist already for linear spaces [26]), and the one for Sub is weaker, as explained in the introduction. In conclusion, the Riemannian GS algorithm is a conceptually simple generalization of the GS algorithm to Riemannian manifolds with a strong theoretical backup, but at the price of higher computational cost.

Acknowledgments. We thank Michael Overton and an anonymous referee for their helpful comments.

REFERENCES

[1] P.-A. Absil and K. A. Gallivan, *Accelerated line-search and trust-region methods*, SIAM J. Numer. Anal., 47 (2009), pp. 997–1018.

[2] P.-A. Absil, R. Mahony, and R. Sepulchre, *Optimization Algorithms on Matrix Manifolds*, Princeton University Press, Princeton, NJ, 2008.

[3] R. L. Adler, J.-P. Dedieu, J. Y. Margulies, M. Martens, and M. Shub, *Newton's method on Riemannian manifolds and a geometric model for the human spine*, IMA J. Numer. Anal., 22 (2002), pp. 359–390.

[4] M. Bačák, R. Bergmann, G. Steidl, and A. Weinmann, *A second order nonsmooth variational model for restoring manifold-valued images*, SIAM J. Sci. Comput., 38 (2016), pp. A567–A597.

[5] P. B. Borckmans, S. Easter Selvan, N. Boumal, and P.-A. Absil, *A Riemannian subgradient algorithm for economic dispatch with valve-point effect*, J. Comput. Appl. Math., 255 (2014), pp. 848–866.

[6] N. Boumal, B. Mishra, P.-A. Absil, and R. Sepulchre, *Manopt, a MATLAB toolbox for optimization on manifolds*, J. Mach. Learn. Res., 15 (2014), pp. 1455–1459.

[7] J. V. Burke, A. S. Lewis, and M. L. Overton, *A robust gradient sampling algorithm for nonsmooth, nonconvex optimization*, SIAM J. Optim., 15 (2005), pp. 751–779.

[8] F. H. Clarke, *Optimization and Nonsmooth Analysis*, 2nd ed., SIAM, Philadelphia, 1990.

[9] C. Da Silva and F. J. Herrmann, *Optimization on the hierarchical Tucker manifold—applications to tensor completion*, Linear Algebra Appl., 481 (2015), pp. 131–173.

[10] G. Dirr, U. Helmke, and C. Lageman, *Nonsmooth Riemannian optimization with applications to sphere packing and grasping*, in Lagrangian and Hamiltonian Methods for Nonlinear Control 2006, Lecture Notes in Control and Inform. Sci. 366, Springer, Berlin, 2007, pp. 29–45.

[11] X. Dong, P. Frossard, P. Vandergheynst, and N. Nefedov, *Clustering on multi-layer graphs via subspace analysis on Grassmann manifolds*, IEEE Trans. Signal Process., 62 (2014), pp. 905–918.

[12] A. Edelman, T. A. Arias, and S. T. Smith, *The geometry of algorithms with orthogonality constraints*, SIAM J. Matrix Anal. Appl., 20 (1999), pp. 303–353.

[13] L. C. Evans and R. F. Gariepy, *Measure Theory and Fine Properties of Functions*, CRC Press, Boca Raton, FL, 1992.

[14] P. Grohs and S. Hosseini, *Nonsmooth trust region algorithms for locally Lipschitz functions on Riemannian manifolds*, IMA J. Numer. Anal., 36 (2016), pp. 1167–1192.

[15] P. Grohs and S. Hosseini, *ε-subgradient algorithms for locally Lipschitz functions on Riemannian manifolds*, Adv. Comput. Math., 42 (2016), pp. 333–360.

[16] P. Grohs and M. Sprecher, *Total variation regularization on Riemannian manifolds by iteratively reweighted minimization*, Inf. Inference, 5 (2016), pp. 353–378.

[17] S. Hosseini, W. Huang, and R. Yousefpour, *Line search algorithms for locally Lipschitz functions on Riemannian manifolds*, INS preprint 1626, 2016.

[18] S. Hosseini and M. R. Pouryayevali, *Generalized gradients and characterization of epi-Lipschitz sets in Riemannian manifolds*, Nonlinear Anal., 74 (2011), pp. 3884–3895.
[19] W. Huang, *Optimization Algorithms on Riemannian Manifolds with Applications*, Ph.D. thesis, Department of Mathematics, Florida State University, Tallahassee, 2014.
[20] W. Huang, P.-A. Absil, and K. A. Gallivan, *A Riemannian BFGS Method for Nonconvex Optimization Problems*, Technical report UCL-INMA-2015.11, UC Louvain, 2015.
[21] W. Huang, K. A. Gallivan, and P.-A. Absil, *A Broyden class of quasi-Newton methods for Riemannian optimization*, SIAM J. Optim., 25 (2015), pp. 1660–1685.
[22] M. Ishteva, P.-A. Absil, S. Van Huffel, and L. De Lathauwer, *Best low multilinear rank approximation of higher-order tensors, based on the Riemannian trust-region scheme*, SIAM J. Matrix Anal. Appl., 32 (2011), pp. 115–135.
[23] K. C. Kiwiel, *Convergence of the gradient sampling algorithm for nonsmooth nonconvex optimization*, SIAM J. Optim., 18 (2007), pp. 379–388.
[24] K. Kovnatsky, A. Glasho and M. M. Bronstein, *MADMM: A generic algorithm for non-smooth optimization on manifolds*, European Conference on Computer Vision, Lecture Notes in Comput. Sci. 9909, Springer, NY, 2016.
[25] D. Kressner, M. Steinlechner, and B. Vandereycken, *Low-rank tensor completion by Riemannian optimization*, BIT, 54 (2014), pp. 447–468.
[26] A. S. Lewis and M. L. Overton, *Nonsmooth optimization via quasi-Newton methods*, Math. Program., 141 (2013), pp. 135–163.
[27] M. M. Mäkelä and P. Neittaanmäki, *Nonsmooth optimization. Analysis and algorithms with applications to optimal control*, World Scientific, River Edge, NJ, 1992.
[28] B. Mishra, G. Meyer, F. Bach, and R. Sepulchre, *Low-rank optimization with trace norm penalty*, SIAM J. Optim., 23 (2013), pp. 2124–2149.
[29] B. Mishra, G. Meyer, S. Bonnabel, and R. Sepulchre, *Fixed-rank matrix factorizations and Riemannian low-rank optimization*, Comput. Statist., 29 (2014), pp. 591–621.
[30] Y. Nakatsukasa, T. Soma, and A. Uschmajew, *Finding a low-rank basis in a matrix subspace*, Math. Program., to appear.
[31] Q. Qu, J. Sun, and J. Wright, *Finding a sparse vector in a subspace: Linear sparsity using alternating directions*, in Adv. Neural Inform. Process. Syst. 27, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, eds., Curran, Red Hook, NY, 2014, pp. 3401–3409.
[32] R. C. Riddell, *Minimax problems on Grassmann manifolds. Sums of eigenvalues*, Adv. Math., 54 (1984), pp. 107–199.
[33] H. Sato and T. Iwai, *A Riemannian optimization approach to the matrix singular value decomposition*, SIAM J. Optim., 23 (2013), pp. 188–212.
[34] B. Savas and L.-H. Lim, *Quasi-Newton methods on Grassmannians and multilinear approximations of tensors*, SIAM J. Sci. Comput., 32 (2010), pp. 3352–3393.
[35] B. Vandereycken, *Low-rank matrix completion by Riemannian optimization*, SIAM J. Optim., 23 (2013), pp. 1214–1236.
[36] B. Vandereycken and S. Vandewalle, *A Riemannian optimization approach for computing low-rank solutions of Lyapunov equations*, SIAM J. Matrix Anal. Appl., 31 (2010), pp. 2553–2579.
[37] Z. Zhao, Z.-J. Bai, and X.-Q. Jin, *A Riemannian Newton algorithm for nonlinear eigenvalue problems*, SIAM J. Matrix Anal. Appl., 36 (2015), pp. 752–774.