J. Nonsmooth Anal. Optim. 6 (2026), 16115, DOI: 10.46298/jnsao-2026-16115
Submitted: 2025-07-25, accepted: 2026-09-22

© the authors, CC-BY-SA 4.0

# THE RIEMANNIAN CONVEX BUNDLE METHOD

Ronny Bergmann* Roland Herzog† Hajg Jasa‡

Abstract We introduce the convex bundle method to solve convex, non-smooth optimization problems on Riemannian manifolds of bounded sectional curvature. Each step of our method is based on a model that involves the convex hull of previously collected subgradients, parallelly transported to the tangent space at the current serious iterate. This generalizes the dual approach to nonsmooth optimization via classical bundle subproblems in Euclidean space. We prove that, under mild conditions, the convex bundle method converges to a minimizer. Several numerical examples implemented using `Manopt.jl` illustrate the performance of the proposed method and compare it to the subgradient method, the cyclic proximal point algorithm, as well as the proximal bundle method. We showcase how the proposed method outperforms these algorithms in terms of CPU runtime and number of iterations on examples for finding Riemannian medians, while it is comparable in performance on large signal denoising tasks.

## 1 INTRODUCTION

Research on optimization problems defined on Riemannian manifolds has experienced steadily increasing activity in the past decades. Among these, non-smooth, unconstrained optimization defines a class of particularly interesting problems. For instance, tasks such as image denoising [5, 33, 46], inpainting [7], and interpolation [8, 48] are often expressed in terms of high-dimensional, non-smooth optimization tasks defined on Riemannian manifolds.

We consider problems of the form

$$\text{Minimize } f(p), \quad p \in \mathcal{M}, \tag{1.1}$$

where $\mathcal{M}$ is a Riemannian manifold and $f$ is geodesically convex and potentially non-smooth. Our setting will be that of a manifold $\mathcal{M}$ with bounded sectional curvature. The precise requirements are given in the beginning of Section 3.

In the Euclidean case, various algorithms lend themselves to the minimization of non-smooth convex functions with additional structure. We mention the Douglas-Rachford splitting algorithm [20], the Alternating Direction Method of Multipliers (ADMM) [23], and primal-dual splitting methods such as the Chambolle-Pock algorithm [15]. On Riemannian manifolds, some of the first non-smooth optimization algorithms include the subgradient method [21] and the proximal point algorithm [22]. Further algorithmic options exist on Hadamard manifolds, e. g., adaptations of Douglas-Rachford splitting [10] or, quite recently, the Chambolle-Pock algorithm [9] based on a newly introduced concept of duality.

*Norwegian University of Science and Technology, Department of Mathematical Sciences, NO-7041 Trondheim, Norway, ronny.bergmann@ntnu.no, https://www.ntnu.edu/employees/ronny.bergmann, ORCID: 0000-0001-8342-7218

†Interdisciplinary Center for Scientific Computing, Heidelberg University, 69120 Heidelberg, Germany *and* Institute for Mathematics Heidelberg University, 69120 Heidelberg, Germany, roland.herzog@iwr.uni-heidelberg.de, https://scoop.iwr.uni-heidelberg.de, ORCID: 0000-0003-2164-6575

‡Simula Research Laboratory, Department of Numerical Analysis and Scientific Computing, 0164 Oslo, Norway, hajg@simula.no, https://www.simula.no/people/hajg, ORCID: 0009-0002-7917-0530

A class of algorithms that have proved to be effective in optimizing generic, convex and non-convex, non-smooth functions in the Euclidean setting is that of bundle methods; see, e. g., [35] for a survey of bundle methods, and [11, 25, 37] for textbook presentations on variants thereof. These algorithms are characterized by the fact that the descent strategy is employed within a framework that guarantees stability by keeping track of past iterates and subgradients. One subcategory of stable bundle methods is that of proximal bundle algorithms, where a proximal term is added to the model of the objective function. A Riemannian bundle method for possibly non-convex, Lipschitz functions, that makes use of a proximal term and models the objective function on the tangent space was presented in [27]. To obtain a search direction, this method solves a proximal subproblem that is formally different from the subproblem we solve in the convex case, most prominently due to the introduction of a proximal parameter. The next candidate point is then obtained through the search direction. This search direction may have to be adapted according to a restricted-step procedure that depends on the injectivity radius of the manifold. A related method is the Riemannian restricted memory quasi-Newton bundle method [41] for minimizing locally Lipschitz functions. This method approximates the curvature information of the objective function by applying a Riemannian quasi-Newton update. A subgradient aggregation technique is then used to reduce the computations in the quadratic subproblem for procuring the candidate descent direction. The stepsize selection entails a line-search that leverages a Riemannian semismoothness concept.

We are further interested in numerical applications and examples for non-smooth optimization problems on Riemannian manifolds, such as typical signal and manifold-valued image denoising problems. In this context, in parallel to the aforementioned algorithmic advancements, ample software has been developed. Examples include the `Manopt` family, consisting of the inaugural MATLAB version `Manopt` [12], the PYTHON version `Pymanopt` [42], and the JULIA version `Manopt.jl` [6]. Another popular software related to differential geometry and to some extent to optimization is the PYTHON package `Geomstats` [36]. In this work we focus on the JULIA package `Manopt.jl`.

### CONTRIBUTIONS

The aim of this paper is to generalize the *convex* bundle method (CBM) for convex, non-smooth functions from Euclidean spaces to Riemannian manifolds with bounded sectional curvature. A Euclidean version of this method is described, e. g., in [25, Chapter 6.7]. We generalize this method by exploring properties of geodesically convex functions, for which a Riemannian generalization of the convex subdifferential (also known as the Fenchel subdifferential) is defined. At its core, our method approximates the $\varepsilon$-subdifferential of the objective function at the current serious iterate by the convex hull of the subgradients that were collected at the previous iterates. This is shown in Theorem 3.2, where the main challenge arises from the fact that these subgradients lie in different tangent spaces and need to be parallely transported, which impacts their ability to be part of an approximate subdifferential at the target point. The search direction in our algorithm is obtained by solving a dual subproblem arising from an orthogonal projection of the zero tangent vector onto the aforementioned inner approximation of the convex $\varepsilon$-subdifferential. To procure the next candidate point, the search direction is shortened if necessary. This is done in accordance with the function's effective domain, using a backtracking procedure. If the manifold is strictly positively curved, a restriction of the domain of the objective may be needed for this set to be geodesically convex. On the other hand, if the manifold is strictly negatively curved, we require the objective's domain to be bounded to prove that the sequence of serious iterates is bounded in Lemma 5.7. Theorem 5.8 provides the same type of result as its Euclidean counterpart presented, e. g., in [25, Theorem 6.80], under the additional assumption that the domain of the objective be bounded whenever the manifold is not flat. The convergence to a (global) minimizer of the objective is guaranteed whenever the set of interior solutions to Problem 1.1 is nonempty, and the sequence of serious iterates produced by the proposed algorithm does not have accumulation points

on the boundary of dom $f$. This result hinges on the following observation. Inspired by [3, p.15] and motivated in Item (ii), in case the algorithm produces a null step, we look for a step size that satisfies a geometric condition involving the linearization errors and curvature-dependent remainder terms defined in (3.1) and (3.7), respectively. This ensures that the last serious iterate output by our method is a minimizer in case the algorithm produces only a finite number of serious steps.

In summary, we

- approximate the $\varepsilon$-subdifferential in Theorem 3.2;
- introduce Algorithm 1 and analyze its basic properties;
- show the asymptotic convergence of this method in the case where infinitely many serious steps occur in Lemma 5.4;
- show the asymptotic convergence the proposed algorithm in the case where finitely many serious iterates are produced in Lemma 5.5;
- prove the full convergence of Algorithm 1 in Theorem 5.8.

### ORGANIZATION

The remainder of this paper is organized as follows. In Section 2 we recall some notions from Riemannian geometry and fix the notation. Section 3 introduces the main problem, the convex bundle method model, and an approximation to the $\varepsilon$-subdifferential tailored to Riemannian manifolds with bounded sectional curvature. The derivation and convergence analysis of our method is motivated and inspired by [25, Chapter 6.7]. Proofs that directly carry over to the Riemannian case without any sort of modification are omitted, whereas proofs that require additional steps will be given. Section 4 is dedicated to the description of the algorithm. Section 5 shows the analysis of the convergence properties of the proposed algorithm. The same observations regarding the Euclidean case made for Section 3 apply to this section as well. In Section 6 we illustrate the performance of the Riemannian convex bundle method (RCBM) on several numerical examples on Riemannian manifolds. Section 7 concludes the paper.

## 2 PRELIMINARIES

In this section, we recall some notions from Riemannian geometry. For more details, the reader may wish to consult [31, 19]. Let $\mathcal{M}$ be a smooth manifold with Riemannian metric $(\cdot\,,\,\cdot)$ and let D be its Levi-Civita connection. Given $p, q \in \mathcal{M}$, we denote by

$$C_{pq} = \{\gamma\colon [0,1] \to \mathcal{M} \mid \gamma \text{ is piecewise smooth and } \gamma(0) = p,\ \gamma(1) = q\}$$

the set of all piecewise smooth curves from $p$ to $q$. The *arc length* of $\gamma \in C_{pq}$, $L\colon C_{pq} \to \mathbb{R}_{\geq 0}$ is defined as

$$L(\gamma) = \int_0^1 \|\dot\gamma\|_{\gamma(t)}\ \mathrm{d}t,$$

where $\|\cdot\|_p$ is the norm induced by the Riemannian metric in the tangent space $\mathcal{T}_p\mathcal{M}$ at the point $p$. The inner product given by the Riemannian metric at $p$ is $(\cdot\,,\,\cdot)_p\colon \mathcal{T}_p\mathcal{M} \times \mathcal{T}_p\mathcal{M} \to \mathbb{R}$. The explicit reference to the base point will be omitted whenever the base point is clear from context. Tangent vectors in $\mathcal{T}_p\mathcal{M}$ will be generically denoted by $X_p$ or $Y_p$ etc. where the subindex merely serves to indicate the base point.

The sectional curvature at $p \in \mathcal{M}$ for linearly independent tangent vectors $X_p, Y_p \in \mathcal{T}_p\mathcal{M}$ is

$$K_p(X_p, Y_p) = \frac{(\mathrm{R}_p(X_p, Y_p)Y_p\,,\, X_p)}{\|X_p\|^2\,\|Y_p\|^2 - (X_p\,,\, Y_p)^2}, \tag{2.1}$$

where R is the Riemann curvature tensor.

The Riemannian distance between two points, $\operatorname{dist}\colon \mathcal{M}\times\mathcal{M}\to\mathbb{R}_{\geq 0}\cup\{+\infty\}$, is given by

$$\operatorname{dist}(p,q)=\inf_{\gamma\in C_{pq}} L(\gamma).$$

A *geodesic arc* from $p\in\mathcal{M}$ to $q\in\mathcal{M}$ is a smooth curve $\gamma\in C_{pq}$ that is parallel along itself, i. e., $\mathrm{D}_{\dot\gamma(t)}\dot\gamma(t)=0$ for all $t\in[0,1]$. A finite-length geodesic arc is *minimal* if its arc length coincides with the Riemannian distance between its extreme points. Given two points $p,q\in\mathcal{M}$, we denote by $\gamma_{p,q}$ a minimal geodesic arc that connects $p=\gamma_{p,q}(0)$ to $q=\gamma_{p,q}(1)$.

On Riemannian manifolds, there exist several notions of convex sets, which differ in strength; see for instance [44, Section 2] or [45, Definition 2.1]. We will only speak of sets referred to as *strongly* convex sets in the literature.

Definition 2.1. A set $\mathcal{U}\subseteq\mathcal{M}$ is said to be *strongly convex* if for any $p,q\in\mathcal{U}$, there exists a unique minimal geodesic arc $\gamma_{p,q}$, and, moreover, $\gamma_{p,q}$ lies entirely in $\mathcal{U}$.

The exponential map at $p\in\mathcal{M}$ is denoted by $\exp_p\colon\mathcal{U}_p\to\mathcal{M}$, where $\mathcal{U}_p\subseteq\mathcal{T}_p\mathcal{M}$ is a suitable neighborhood of $0_p\in\mathcal{T}_p\mathcal{M}$ that depends on the injectivity radius of $\mathcal{M}$ at the point $p$, which is the quantity given by

$$i_{\mathcal{M}}(p):=\sup\left\{r>0\,\middle|\,\exp_p\colon B_r(0_p)\to\exp_p(B_r(0_p))\text{ is a diffeomorphism}\right\},$$

where $B_r(0_p)\subseteq\mathcal{T}_p\mathcal{M}$; see, e. g., [32, Chapter 6, p.165]. The (global) injectivity radius of $\mathcal{M}$ is $i_{\mathcal{M}}:=\inf_{p\in\mathcal{M}} i_{\mathcal{M}}(p)$. The inverse of the exponential map at $p\in\mathcal{M}$ is the logarithmic map, and it is denoted by $\log_p$. It might not be defined on all of $\mathcal{M}$, but it is guaranteed to be defined on some open ball of radius $r\leq i_{\mathcal{M}}(p)$ centered at $p$, denoted by $B_r(p)=\{q\in\mathcal{M}\,|\,\operatorname{dist}(p,q)<r\}$. Note that given a strongly convex set $\mathcal{U}$ and $p,q\in\mathcal{U}$, $\log_p q$ is the unique tangent vector that realizes the initial velocity of the unique minimal geodesic arc $\gamma_{p,q}$ that lies in $\mathcal{U}$.

We denote the *parallel transport* from $p$ to $q$ along $\gamma_{p,q}$ with respect to the connection D by $\mathrm{P}_{q\leftarrow p}\colon\mathcal{T}_p\mathcal{M}\to\mathcal{T}_q\mathcal{M}$. It is defined by

$$\mathrm{P}_{q\leftarrow p}Y_p=X_q\quad\text{for } Y_p\in\mathcal{T}_p\mathcal{M},$$

where $X$ is the unique smooth vector field along $\gamma_{p,q}$ which satisfies

$$X_p=Y_p\text{ and }\mathrm{D}_{\dot\gamma}X=0.$$

Notice that $\mathrm{P}_{q\leftarrow p}$ depends on the choice of a minimal geodesic $\gamma_{p,q}$ but in our setting it will not cause any ambiguity since $\gamma_{p,q}$ is unique in a strongly convex set. Denoting by $\mathrm{P}_{p\leftarrow q}$ the parallel transport along the curve $\gamma_{p,q}$ traveled in the opposite direction, we have

$$(\mathrm{P}_{q\leftarrow p}Y_p\,,\,Z_q)=(Y_p\,,\,\mathrm{P}_{p\leftarrow q}Z_q)\tag{2.2}$$

for all $Y_p\in\mathcal{T}_p\mathcal{M}$ and all $Z_q\in\mathcal{T}_q\mathcal{M}$. Moreover, we have

$$\mathrm{P}_{q\leftarrow p}\log_p q=-\log_q p.\tag{2.3}$$

Definition 2.2. A function $f\colon\mathcal{M}\to\overline{\mathbb{R}}:=\mathbb{R}\cup\{+\infty\}$ is said to be *lower semi-continuous* at $p\in\mathcal{M}$ if

$$\liminf_{q\to p} f(q)\geq f(p).$$

Definition 2.3. Let $\mathcal{U}\subseteq\mathcal{M}$ be a strongly convex set. A function $f\colon\mathcal{M}\to\overline{\mathbb{R}}$ is said to be *geodesically convex on* $\mathcal{U}$ if for any $p,q\in\mathcal{U}$,

$$(f\circ\gamma_{p,q})(t)\leq(1-t)f(p)+tf(q)\quad\text{holds for all } t\in[0,1].\tag{2.4}$$

A function as in Definition 2.3 can be obtained, for instance, by starting with a strongly convex set $\mathcal{U}$ and a function $f\colon \mathcal{U} \to \mathbb{R}$ satisfying (2.4) for any $p, q \in \mathcal{U}$, and by extending it to $+\infty$ outside of $\mathcal{U}$. Then $f$ will be geodesically convex on $\operatorname{dom} f \coloneqq \{p \in \mathcal{M} \mid f(p) \neq +\infty\} = \mathcal{U}$.

Let $f\colon \mathcal{M} \to \overline{\mathbb{R}}$ be a function such that $\operatorname{dom} f$ is strongly convex and $f$ is a geodesically convex function on $\operatorname{dom} f$. The *subdifferential* of $f$ at $p \in \operatorname{dom} f$ is defined as

$$\partial f(p) \coloneqq \big\{X_p \in \mathcal{T}_p\mathcal{M} \,\big|\, f(q) \geq f(p) + (X_p\,,\, \log_p q) \text{ for all } q \in \operatorname{dom} f\big\}.$$

For any $\varepsilon > 0$, the $\varepsilon$-subdifferential of $f$ at $p \in \operatorname{dom} f$ is defined as

$$\partial_\varepsilon f(p) = \big\{X_p \in \mathcal{T}_p\mathcal{M} \,\big|\, f(q) \geq f(p) + (X_p\,,\, \log_p q) - \varepsilon \text{ for all } q \in \operatorname{dom} f\big\}.$$

Clearly, the inclusion

$$\partial f(p) \subseteq \partial_\varepsilon f(p)$$

holds for all $\varepsilon > 0$.

## 3 APPROXIMATION OF THE $\varepsilon$-SUBDIFFERENTIAL

We begin by specifying the standing assumptions for the remainder of the paper.

Assumption 3.1 (standing assumptions).

(i) $\mathcal{M}$ is a complete Riemannian manifold of finite dimension.

(ii) $f\colon \mathcal{M} \to \overline{\mathbb{R}}$ is a lower semi-continuous function.

(iii) $\operatorname{dom} f \subseteq \mathcal{M}$ is a strongly convex set with nonempty interior in $\mathcal{M}$, and $f$ is a geodesically convex function on $\operatorname{dom} f$.

(iv) There exist two numbers $\omega, \Omega \in \mathbb{R}$ such that the sectional curvature satisfies $\omega \leq K_p(X_p, Y_p) \leq \Omega$ for all $p \in \operatorname{dom} f$ and all linearly independent unit vectors $X_p, Y_p \in \mathcal{T}_p\mathcal{M}$.

(v) In case $\omega < 0$ or $\Omega > 0$, $\operatorname{dom} f$ is bounded and $\delta \coloneqq \operatorname{diam}(\operatorname{dom} f)$ denotes its diameter. In case $\Omega > 0$, we assume $\delta < \frac{\pi}{\sqrt{\Omega}}$. If the manifold is flat ($\omega = \Omega = 0$), then $\operatorname{dom} f$ is allowed to be unbounded.

Notice that the third assumption in particular implies that the logarithmic map is well-defined on $\operatorname{dom} f$. The last assumption is due to the fact that in case of $\Omega > 0$, sets of diameter larger than $\frac{\pi}{\sqrt{\Omega}}$ cannot be strongly convex; see, e. g., [34, Proposition 4.2 (iii)], [45, Proposition 4.1] or [13, Proposition II.1.4].

As usual in bundle methods see, e. g., [11, Chapter 10], we aim to approximate the objective function through a sequence of models $\varphi^{(k)}\colon \operatorname{dom} f \to \mathbb{R}$, $k \in \mathbb{N}_0$, defined by

$$\varphi^{(k)}(q) \coloneqq \max\big\{f(p^{(j)}) + (X_{p^{(j)}}\,,\, \log_{p^{(j)}} q) \,\big|\, j = 0, \dots, k\big\}, \quad q \in \operatorname{dom} f,$$

where $X_{p^{(j)}} \in \partial f(p^{(j)})$ is a *subgradient* of $f$ at $p^{(j)} \in \mathcal{M}$ for all $j = 0, \dots, k \in \mathbb{N}_0$. Define now the *linearization errors* as

$$e_j^{(k)} \coloneqq f(p^{(k)}) - f(p^{(j)}) - (X_{p^{(j)}}\,,\, \log_{p^{(j)}} p^{(k)}) \tag{3.1}$$

and note that $e_j^{(k)} \geq 0$ holds for all $j = 0, \dots, k$ by the definition of subgradient. We can therefore rewrite

$$\varphi^{(k)}(q) = f(p^{(k)}) + \max\big\{-e_j^{(k)} + (X_{p^{(j)}}\,,\, \log_{p^{(j)}} q - \log_{p^{(j)}} p^{(k)}) \,\big|\, j = 0, \dots, k\big\} \tag{3.2}$$

for $q \in \operatorname{dom} f$.

The linearization errors defined in (3.1) and the representation of the model in (3.2) are the exact analogs of their counterparts for convex bundle methods in vector spaces. However, it turns out that, contrary to the vector space situation, the linearization errors cannot be used directly to define approximations of the $\varepsilon$-subdifferential of $f$. This is because it is not clear a priori that such a choice of linearization errors ensures that a subgradient at an initial point $p$ can be parallely transported to a different point $q$ and remain an element of the $\varepsilon$-subdifferential at $q$, as is the case when dealing with vector spaces. This is an effect caused by curvature.

For $p \in \operatorname{dom} f$ we introduce the following quantities,

$$r_j^{(k)}(p) \coloneqq \varrho_{\omega,\Omega}(\operatorname{dist}(\tilde{p},\, p))\, \|X_{p^{(j)}}\|\, \|\log_{p^{(j)}} p^{(k)}\|, \tag{3.3}$$

where $\tilde{p}$ lies on the minimal geodesic arc that connects $p^{(j)}$ to $p^{(k)}$,

$$\varrho_{\omega,\Omega}(s) \coloneqq \max\bigl\{\zeta_{1,\omega}(s) - 1,\; 1 - \zeta_{2,\Omega}(s)\bigr\}, \tag{3.4}$$

and

$$\begin{aligned} \zeta_{1,\omega}(s) &\coloneqq \begin{cases} 1 & \text{if } \omega \geq 0, \\ \sqrt{-\omega}\, s \coth(\sqrt{-\omega}\, s) & \text{if } \omega < 0, \end{cases} \\ \zeta_{2,\Omega}(s) &\coloneqq \begin{cases} 1 & \text{if } \Omega \leq 0, \\ \sqrt{\Omega}\, s \cot(\sqrt{\Omega}\, s) & \text{if } \Omega > 0, \end{cases} \end{aligned} \tag{3.5}$$

for $s \in \mathbb{R}$. We refer the reader to Figure 3.1 for an illustration. Observe that

$$r_j^{(k)}(p) \leq \varrho_{\omega,\Omega}(\delta)\, \|X_{p^{(j)}}\|\, \|\log_{p^{(j)}} p^{(k)}\| \tag{3.6}$$

holds for all $p \in \operatorname{dom} f$. For more details, see Appendix A.

As the curvature bounds $\omega$ and $\Omega$, as well as the diameter $\delta$ are fixed throughout the manuscript, we will use $\varrho$ as a shorthand notation for $\varrho_{\omega,\Omega}(\delta)$. Then $r_j^{(k)}(p)$ simplifies to

$$r_j^{(k)} \coloneqq \varrho\, \|X_{p^{(j)}}\|\, \|\log_{p^{(j)}} p^{(k)}\|. \tag{3.7}$$

The curvature-dependent quantities $e_j^{(k)} + r_j^{(k)}$ allow for an approximation of the $\varepsilon$-subdifferential of $f$ at a given point on the manifold using a convex combination of parallely transported subgradients from other points. This approximation depends on bounds that involve the sectional curvature of the manifold.

We are now in a position to prove a result parallel to [25, Theorem 6.68], which shows that we obtain an inner approximation of the $\varepsilon$-subdifferential, using convex combinations of parallely transported subgradients. In what follows, we will occasionally use the symbol $\Delta_k$ to denote the probability simplex in $\mathbb{R}^{k+1}$, i. e., $\Delta_k \coloneqq \bigl\{\lambda \in \mathbb{R}^{k+1} \bigm| \sum_{j=0}^{k} \lambda_j = 1,\; \lambda_j \geq 0,\; j = 0, \ldots, k\bigr\}$.

Theorem 3.2. *Suppose that $k \in \mathbb{N}_0$, $p^{(j)} \in \operatorname{dom} f$ and $X_{p^{(j)}} \in \partial f(p^{(j)})$ holds for all $j = 0, \ldots, k$. Then, for any $\varepsilon > 0$, the subset of $\mathcal{T}_{p^{(k)}}\mathcal{M}$ defined by*

$$G_\varepsilon^{(k)} \coloneqq \Bigl\{\sum_{j=0}^{k} \lambda_j \operatorname{P}_{p^{(k)} \leftarrow p^{(j)}} X_{p^{(j)}} \Bigm| \sum_{j=0}^{k} \lambda_j \bigl(e_j^{(k)} + r_j^{(k)}\bigr) \leq \varepsilon,\; \lambda \in \Delta_k\Bigr\} \tag{3.8}$$

*satisfies*

$$G_\varepsilon^{(k)} \subseteq \partial_\varepsilon f(p^{(k)}).$$

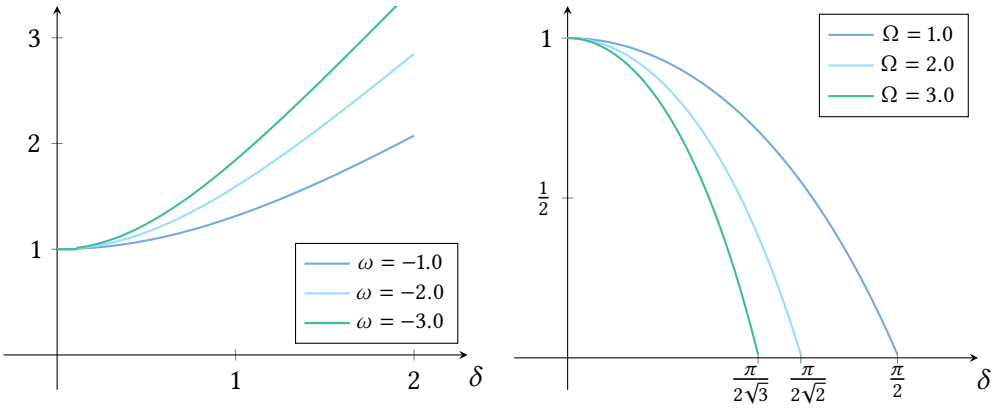



Figure 3.1: Illustration of the functions $\zeta_{1,\omega}$ (left) and $\zeta_{2,\Omega}$ (right) defined in (3.5).

*Proof.* Let $p \in \operatorname{dom} f$, then

$$\begin{aligned}
&\left(X_{p^{(j)}}, \log_{p^{(j)}} p - \log_{p^{(j)}} p^{(k)}\right) \\
&\quad = \left(X_{p^{(j)}}, \log_{p^{(j)}} p\right) - \left(X_{p^{(j)}}, \log_{p^{(j)}} p^{(k)}\right) \\
&\quad \leq f(p) - f(p^{(j)}) - \left(X_{p^{(j)}}, \log_{p^{(j)}} p^{(k)}\right) \\
&\quad = f(p) - f(p^{(k)}) + e_j^{(k)},
\end{aligned}$$

where we used the subdifferential inequality at $p^{(j)}$ and the definition of linearization errors given in (3.1). Therefore

$$f(p) \geq f(p^{(k)}) + \left(X_{p^{(j)}}, \log_{p^{(j)}} p - \log_{p^{(j)}} p^{(k)}\right) - e_j^{(k)} \quad \text{for all } p \in \operatorname{dom} f.$$

Hence, we see that the model $\varphi^{(k)}$ approximates the objective function $f$ from below on $\operatorname{dom} f$ due to the linearization errors (3.1) satisfying $e_j^{(k)} \geq 0$.

By adding and subtracting $\left(\mathrm{P}_{p^{(k)} \leftarrow p^{(j)}} X_{p^{(j)}}, \log_{p^{(k)}} p\right)$ to the right-hand side of the previous inequality, we get

$$\begin{aligned}
f(p) \geq f(p^{(k)}) &+ \left(\mathrm{P}_{p^{(k)} \leftarrow p^{(j)}} X_{p^{(j)}}, \log_{p^{(k)}} p\right) \\
&+ \left(X_{p^{(j)}}, \log_{p^{(j)}} p - \mathrm{P}_{p^{(j)} \leftarrow p^{(k)}} \log_{p^{(k)}} p - \log_{p^{(j)}} p^{(k)}\right) - e_j^{(k)},
\end{aligned}$$

for all $p \in \operatorname{dom} f$. Following an argument presented in [2, Appendix C], see also Appendix A, we can write

$$f(p) \geq f(p^{(k)}) + \left(\mathrm{P}_{p^{(k)} \leftarrow p^{(j)}} X_{p^{(j)}}, \log_{p^{(k)}} p\right) - r_j^{(k)} - e_j^{(k)}, \tag{3.9}$$

for all $p \in \operatorname{dom} f$, with $r_j^{(k)}$ from (3.7).

Let $\lambda \in \Delta_k$, multiply (3.9) by $\lambda_j$ and sum up to obtain

$$f(p) \geq f(p^{(k)}) + \sum_{j=0}^{k} \lambda_j \left(\mathrm{P}_{p^{(k)} \leftarrow p^{(j)}} X_{p^{(j)}}, \log_{p^{(k)}} p\right) - \sum_{j=0}^{k} \lambda_j \left(e_j^{(k)} + r_j^{(k)}\right), \tag{3.10}$$

for all $p \in \operatorname{dom} f$. Hence, by bi-linearity of the Riemannian metric, we can take the sum and multiplication by $\lambda_j$ inside the inner product. Therefore

$$\sum_{j=0}^{k} \lambda_j \, \mathrm{P}_{p^{(k)} \leftarrow p^{(j)}} X_{p^{(j)}} \in \partial_\varepsilon f(p^{(k)})$$

holds if and only if

$$\sum_{j=0}^{k} \lambda_j \left( e_j^{(k)} + r_j^{(k)} \right) \leq \varepsilon,$$

which concludes the argument. □

Note that the set $G_\varepsilon^{(k)}$ is a subset of the inner-product space $\mathcal{T}_{p^{(k)}}\mathcal{M}$, and it is nonempty, convex, and compact. The orthogonal projection $\operatorname{proj}\colon \mathcal{T}_{p^{(k)}}\mathcal{M} \to G_\varepsilon^{(k)}$ of $Y_{p^{(k)}}$, given by

$$\|Y_{p^{(k)}} - \operatorname{proj}_{G_\varepsilon^{(k)}}(Y_{p^{(k)}})\| \leq \|Y_{p^{(k)}} - Z_{p^{(k)}}\| \quad \text{for all } Z_{p^{(k)}} \in G_\varepsilon^{(k)},$$

is thus well-defined; see for example [25, Lemma 2.17] for a proof of this fact. We can therefore define

$$g^{(k)} \coloneqq \operatorname{proj}_{G_\varepsilon^{(k)}}(0_{p^{(k)}}). \tag{3.11}$$

The negative of this tangent vector will serve as search direction in our bundle method.

(3.11) means $\|g^{(k)}\| \leq \|Z_{p^{(k)}}\|$ for all $Z_{p^{(k)}} \in G_\varepsilon^{(k)}$, or equivalently

$$g^{(k)} \coloneqq \operatorname*{arg\,min}_{Z_{p^{(k)}} \in G_\varepsilon^{(k)}} \|Z_{p^{(k)}}\|.$$

Using the characterization of $G_\varepsilon^{(k)}$ from Theorem 3.2, we can therefore write

$$g^{(k)} = \sum_{j=0}^{k} \lambda_j^{(k)} \operatorname{P}_{p^{(k)} \leftarrow p^{(j)}} X_{p^{(j)}}, \tag{3.12}$$

where the tuple $\lambda^{(k)} = (\lambda_0^{(k)}, \ldots, \lambda_k^{(k)})$ is a solution to the following convex quadratic program,

$$\begin{aligned} \operatorname*{arg\,min}_{\lambda \in \mathbb{R}^{k+1}} \quad & \frac{1}{2} \left\| \sum_{j=0}^{k} \lambda_j \operatorname{P}_{p^{(k)} \leftarrow p^{(j)}} X_{p^{(j)}} \right\|^2 \\ \text{s. t.} \quad & \sum_{j=0}^{k} \lambda_j \left( e_j^{(k)} + r_j^{(k)} \right) \leq \varepsilon, \quad \text{and} \quad \lambda \in \Delta_k. \end{aligned} \tag{P1}$$

Closely related to Problem P1 is another quadratic program, in which the weighted sum of the components of $\lambda^{(k)}$ is not constrained but rather penalized:

$$\begin{aligned} \operatorname*{arg\,min}_{\lambda \in \mathbb{R}^{k+1}} \quad & \frac{1}{2} \left\| \sum_{j=0}^{k} \lambda_j \operatorname{P}_{p^{(k)} \leftarrow p^{(j)}} X_{p^{(j)}} \right\|^2 + \sum_{j=0}^{k} \lambda_j \left( e_j^{(k)} + r_j^{(k)} \right) \\ \text{s. t.} \quad & \lambda \in \Delta_k. \end{aligned} \tag{P2}$$

The relation between Problem P1 and Problem P2 can be studied through the following result, the proof of which is analogous to [25, Lemma 6.70].

Lemma 3.3. *The vector $\lambda^{(k)} = (\lambda_0^{(k)}, \ldots, \lambda_k^{(k)}) \in \mathbb{R}^{k+1}$ is a solution to Problem P2 if and only if there exist $g^{(k)} \in \mathcal{T}_{p^{(k)}}\mathcal{M}$ and a Lagrange multiplier $\xi^{(k)} \in \mathbb{R}$ such that the triple $(\lambda^{(k)}, g^{(k)}, \xi^{(k)})$ fulfills the*

*following KKT conditions*

$$\begin{aligned} g &= \sum_{j=0}^{k} \lambda_j \, \mathrm{P}_{p^{(k)} \leftarrow p^{(j)}} X_{p^{(j)}}, \\ \sum_{j=0}^{k} \lambda_j &= 1 \quad \textit{and} \quad \lambda_j \geq 0 \quad \textit{for all } j = 0, \dots, k, \\ (g, \, \mathrm{P}_{p^{(k)} \leftarrow p^{(j)}} X_{p^{(j)}}) + e_j^{(k)} + r_j^{(k)} + \xi &\geq 0 \quad \textit{for all } j = 0, \dots, k, \\ \lambda_j \big( (g, \, \mathrm{P}_{p^{(k)} \leftarrow p^{(j)}} X_{p^{(j)}}) + e_j^{(k)} + r_j^{(k)} + \xi \big) &= 0 \quad \textit{for all } j = 0, \dots, k. \end{aligned}$$

Let now $\lambda^{(k)} = (\lambda_0^{(k)}, \dots, \lambda_k^{(k)}) \in \mathbb{R}^{k+1}$ be a solution to Problem P2 and choose $\varepsilon \coloneqq \sum_{j=0}^{k} \lambda_j^{(k)} \left( e_j^{(k)} + r_j^{(k)} \right)$ analogously to [25, p.389]. We find that the KKT conditions of Problem P1, given by

$$\begin{aligned} g &= \sum_{j=0}^{k} \lambda_j \, \mathrm{P}_{p^{(k)} \leftarrow p^{(j)}} X_{p^{(j)}}, \\ \sum_{j=0}^{k} \lambda_j &= 1 \quad \text{and} \quad \lambda_j \geq 0 \quad \text{for all } j = 0, \dots, k, \\ (g, \, \mathrm{P}_{p^{(k)} \leftarrow p^{(j)}} X_{p^{(j)}}) + \zeta (e_j^{(k)} + r_j^{(k)}) + \xi &\geq 0 \quad \text{for all } j = 0, \dots, k, \\ \lambda_j \big( (g, \, \mathrm{P}_{p^{(k)} \leftarrow p^{(j)}} X_{p^{(j)}}) + \zeta (e_j^{(k)} + r_j^{(k)}) + \xi \big) &= 0 \quad \text{for all } j = 0, \dots, k, \\ \zeta \Big( \sum_{j=0}^{k} \lambda_j^{(k)} \left( e_j^{(k)} + r_j^{(k)} \right) - \varepsilon \Big) &= 0 \quad \text{and} \quad \zeta \geq 0, \\ \sum_{j=0}^{k} \lambda_j^{(k)} \left( e_j^{(k)} + r_j^{(k)} \right) &\leq \varepsilon, \end{aligned}$$

are also satisfied by solutions of Problem P2. Thus, solutions of Problem P2 are solutions to Problem P1 for this choice of $\varepsilon$. Lastly, we can compute the unique Lagrange multiplier of Problem P2 from $\lambda^{(k)}$ as follows:

$$\begin{aligned} \xi^{(k)} = \xi^{(k)} \sum_{j=0}^{k} \lambda_j^{(k)} &= \sum_{j=0}^{k} \lambda_j^{(k)} \left( -\big( g^{(k)}, \, \mathrm{P}_{p^{(k)} \leftarrow p^{(j)}} X_{p^{(j)}} \big) - e_j^{(k)} - r_j^{(k)} \right) \\ &= -\big( g^{(k)}, \, \sum_{j=0}^{k} \lambda_j^{(k)} \, \mathrm{P}_{p^{(k)} \leftarrow p^{(j)}} X_{p^{(j)}} \big) - \sum_{j=0}^{k} \lambda_j^{(k)} \, e_j^{(k)} - \sum_{j=0}^{k} \lambda_j^{(k)} \, r_j^{(k)} \\ &= -\| g^{(k)} \|^2 - \sum_{j=0}^{k} \lambda_j^{(k)} \, e_j^{(k)} - \sum_{j=0}^{k} \lambda_j^{(k)} \, r_j^{(k)}. \end{aligned}$$

Remark 3.4. It is possible to utilize more general tools than the exponential map and parallel transport in Theorem 3.2. We introduce two ways this could be achieved.

First, let $\operatorname{retr}_p X_p$ be a retraction (see, e. g., [1, Definition 4.1.1]) of $X_p \in \mathcal{T}_p \mathcal{M}$ on $\mathcal{M}$. Let $\mathrm{T}_{p, Y_p} X_p$ be a continuous vector transport (see, e. g., [1, Definition 8.1.1]) of $X_p \in \mathcal{T}_p \mathcal{M}$ from $p \in \mathcal{M}$ to the point $\operatorname{retr}_p Y_p \in \mathcal{M}$, and let $\mathrm{T}^{\mathrm{retr}}_{p, Y_p} X = \mathrm{d}_{Y_p} \operatorname{retr}_p [X_p]$ be the differential of the retraction.

As in [30, Assumption 1.3], or, alternatively, [28, (2.5)], assume that there exists a constant $C_0 > 0$ such that T satisfies

$$\| \mathrm{T}_{p, Y_p} - \mathrm{T}^{\mathrm{retr}}_{p, Y_p} \| \leq C_0 \| Y_p \|,$$

for all $p, q \in \operatorname{dom} f$ with $\operatorname{retr}_p Y_p =: q$. We remark that this inequality is satisfied for all such $p, q$ as long as the vector transport $\mathrm{T}_{q \leftarrow p} \coloneqq \mathrm{T}_{p,Y_p}$ is continuous. Then, as in [30, Lemma 3.7], or, alternatively, [28, Lemma 3.5], there exists a constant $C_t > 0$ such that, for all $p, q \in \operatorname{dom} f$

$$\|\mathrm{T}_{q \leftarrow p} X_p - \mathrm{P}_{q \leftarrow p} X_p\| \leq C_t \|Y_p\| \|X_p\|,$$

where $X_p, Y_p \in \mathcal{T}_p \mathcal{M}$ and $\operatorname{retr}_p Y_p = q$. Similarly, assume as in [30, Assumption 1.6] that there exists a constant $C_r > 0$ such that

$$\|\log_p q - \operatorname{retr}_p^{-1} q\| \leq C_r \, \|\operatorname{retr}_p^{-1} q\|^2, \tag{3.13}$$

for all $p, q \in \operatorname{dom} f$. It is also necessary to assume that the retraction is invertible in $\operatorname{dom} f$. Note that (3.13) is quite a natural assumption as retractions are first order approximations of the exponential map. Let now

$$\tilde{e}_j^{(k)} \coloneqq f(p^{(k)}) - f(q^{(j)}) - (X_{q^{(j)}}, \operatorname{retr}_{q^{(j)}}^{-1} p^{(k)}), \tag{3.14}$$

and

$$\tilde{r}_j^{(k)} \coloneqq \varrho \left\| X_{q^{(j)}} \right\| \left\| \operatorname{retr}_{q^{(j)}}^{-1} p^{(k+1)} \right\|. \tag{3.15}$$

Combining these facts with (3.10) we get

$$\sum_{j=0}^{k} \lambda_j \, \mathrm{T}_{p^{(k)} \leftarrow q^{(j)}} X_{q^{(j)}} \in \partial_{\varepsilon_e + \varepsilon_r + \varepsilon_t} f(p^{(k)})$$

if and only if $\sum_{j=0}^{k} \lambda_j \big( \tilde{e}_j^{(k)} + \tilde{r}_j^{(k)} \big) \leq \varepsilon_e$, as before, and additionally

$$\sum_{j=0}^{k} \lambda_j \, C_r \, (1 + \varrho) \, \|X_{q^{(j)}}\| \, \|\operatorname{retr}_{q^{(j)}}^{-1} p^{(k)}\|^2 \leq \varepsilon_r$$

$$\text{and} \quad \sum_{j=0}^{k} \lambda_j \, C_t \, \delta \, \|X_{q^{(j)}}\| \, \|\log_{q^{(j)}} p^{(k)}\| \leq \varepsilon_t,$$

for all $k \in \mathbb{N}$ and arbitrary $\varepsilon_e, \varepsilon_r, \varepsilon_t > 0$. Hence, for $\varepsilon \coloneqq \varepsilon_e + \varepsilon_r + \varepsilon_t$, the quality of the inner approximation of the $\varepsilon$-subdifferential of $f$ at $p^{(k)}$ depends on $C_r$ and $C_t$ as well.

As a second possible approach, one may rephrase all definitions related to geodesic convexity using *retraction convexity* instead, as introduced, e. g., in [29, Definition 2]. This would entail replacing geodesic arcs $\gamma_{p,q}(t) = \exp_p(t \log_p q)$ with retraction curves $c_{p,q}(t) = \operatorname{retr}_p(t \operatorname{retr}_p^{-1} q)$. Most notably, the definition of $\varepsilon$-subdifferential would then include the inverse of the chosen retraction, which is thus evidently required to be invertible in $\operatorname{dom} f$. Furthermore, for the derivation in Appendix A to hold in this case, it is necessary to use a vector transport that is parallel along the aforementioned retraction curves.

## 4 PRESENTATION OF THE METHOD

In this section we present the proposed Riemannian convex bundle method as Algorithm 1. The notation is close to the one in [25, Chapter 6.7].

After an initialization step, in Line 3 the method computes a solution $\lambda^{(k)}$ of the convex, quadratic Problem P2 formed using the subgradients currently held in the bundle, indexed by $J^{(k)}$. This vector of convex coefficients is then used in Line 4 to procure a search direction $d^{(k)}$, as well as the numbers $\varepsilon^{(k)}$ and $\sigma^{(k)}$ that quantify how well the subdifferential is approximated at the $k$-th step. The tolerance given by the user prescribes whether such approximation is satisfactory, and whether the zero vector is close to being an element thereof, in which case the algorithm is stopped. A new candidate point is constructed by moving from the last serious iterate in the new search direction with a suitable step size (Line 14 to Line 16). To assess whether the new candidate is a serious iterate, a check is carried out (Line 18) to verify whether the objective decreased enough. If this check is successful, the candidate is labeled as the new serious iterate. Otherwise, the algorithm performs a null step and looks for a new candidate via computing a smaller step size until a certain geometric condition is met. The new candidate and a subgradient of the objective sampled at this point are appended to the bundle irrespectively of whether a serious or a null step is being performed. Finally, the algorithm purges unused subgradients from the bundle and updates the linearization errors and curvature-dependent remainder terms (Line 26 to Line 28).

We would now like to point out some more details concerning the proposed method.

(i) The backtracking loop in Line 14 to Line 16 of Algorithm 1 serves a dual purpose. On the one hand, it ensures that the trial and serious iterates remain inside the interior of the domain of $f$. On the other hand, on a compact manifold, it may happen that a geodesic "wraps around", so that $\exp_{p^{(k)}}(t^{(k)}d^{(k)})$ is closer to $p^{(k)}$ than $t^{(k)}\|d^{(k)}\|$. This situation is avoided with a strategy inspired by [9, Definition 2.8]. To see that the loop in Line 14 of Algorithm 1 is finite, we argue as follows. Given that $p^{(k)} \in \operatorname{int}(\operatorname{dom} f)$ for each $k$, and that $\operatorname{int}(\operatorname{dom} f)$ is an open set, there exists a ball $B_r(p^{(k)}) \subseteq \operatorname{int}(\operatorname{dom} f)$ for some $r > 0$. Hence, as long as the minimizer does not lie on the boundary of $\operatorname{dom} f$, the loop stops as soon as $t^{(k)}\|d^{(k)}\| < r$. This is guaranteed to happen as the backtracking factor $\beta$ decreases the step size $t^{(k)}$.

(ii) The loop from Line 22 to Line 24 is inspired by [3, Inequality (1.30), p.15]. This remark is dedicated to give further motivation for introducing such loop. Let $q^{(k+1)}$ be a null step, i. e.

$$f(q^{(k+1)}) > f(p^{(k)}) + m\, t^{(k)}\, \xi^{(k)}.$$

Then, by the definition of linearization errors one obtains that

$$\begin{aligned} e_{k+1}^{(k)} &= f(p^{(k)}) - f(q^{(k+1)}) - \big(X_{q^{(k+1)}}\,,\, \log_{q^{(k+1)}} p^{(k)}\big) \\ &= f(p^{(k)}) - f(q^{(k+1)}) + \big(\mathrm{P}_{p^{(k)} \leftarrow q^{(k+1)}} X_{q^{(k+1)}}\,,\, \log_{p^{(k)}} q^{(k+1)}\big). \end{aligned}$$

Thus,

$$\begin{aligned} f(q^{(k+1)}) - f(p^{(k)}) &= \big(\mathrm{P}_{p^{(k)} \leftarrow q^{(k+1)}} X_{q^{(k+1)}}\,,\, \log_{p^{(k)}} q^{(k+1)}\big) - e_{k+1}^{(k)} \\ &= -\big(\mathrm{P}_{p^{(k)} \leftarrow q^{(k+1)}} X_{q^{(k+1)}}\,,\, t^{(k)} g^{(k)}\big) - e_{k+1}^{(k)}, \end{aligned} \tag{4.1}$$

because $q^{(k+1)} = \exp_{p^{(k)}}(t^{(k)}d^{(k)})$ with $d^{(k)} = -g^{(k)}$ and $t^{(k)} \in (0, 1]$. Notice that the second equality above is ensured by the second condition in the backtracking loop; see Line 14 in Algorithm 1. Combining the two previous expressions, we get the following strict inequality

$$\big(\mathrm{P}_{p^{(k)} \leftarrow q^{(k+1)}} X_{q^{(k+1)}}\,,\, t^{(k)}\, g^{(k)}\big) < -m\, t^{(k)}\, \xi^{(k)} - e_{k+1}^{(k)}.$$

Hence, for null steps, there exists a slack $\alpha \geq 0$ such that

$$\big(\mathrm{P}_{p^{(k)} \leftarrow q^{(k+1)}} X_{q^{(k+1)}}\,,\, t^{(k)}\, g^{(k)}\big) = -m\, t^{(k)}\, \xi^{(k)} - e_{k+1}^{(k)} - \alpha.$$

**Algorithm 1** Riemannian Convex Bundle Method (RCBM)

**Require:** $f$, oracle for $\partial f$, initial point $p^{(0)} \in \operatorname{int}\operatorname{dom} f$
**Require:** parameters $\varrho$, $m \in (0,1)$, $\beta \in (0,1)$, tol $> 0$
**Ensure:** last serious iterate $p^{(k_*)}$ for some $k_* \in \mathbb{N}_0$

Set $q^{(0)} \coloneqq p^{(0)}$, $X_{q^{(0)}} \in \partial f(p^{(0)})$, $g^{(0)} \coloneqq X_{q^{(0)}}$, $\varepsilon^{(0)} \coloneqq 0$, $e_0^{(0)} \coloneqq 0$, $\sigma^{(0)} \coloneqq 0$, $r_0^{(0)} \coloneqq 0$, $k \coloneqq 0$, $K \coloneqq \emptyset$, $J^{(0)} \coloneqq \{0\}$.
**while** true **do**
  Compute a solution $\lambda^{(k)} \in \mathbb{R}^{|J^{(k)}|}$ of

$$\begin{aligned} \underset{\lambda \in \mathbb{R}^{|J^{(k)}|}}{\arg\min} \quad & \frac{1}{2}\Big\| \sum_{j \in J^{(k)}} \lambda_j \, \mathrm{P}_{p^{(k)} \leftarrow q^{(j)}} X_{q^{(j)}} \Big\|^2 + \sum_{j \in J^{(k)}} \lambda_j \, e_j^{(k)} + \sum_{j \in J^{(k)}} \lambda_j^{(k)} \, r_j^{(k)} \\ \text{s. t.} \quad & \sum_{j \in J^{(k)}} \lambda_j = 1 \quad \text{and } \lambda_j \geq 0 \quad \text{for all } j \in J^{(k)}. \end{aligned}$$

  Set $g^{(k)} \coloneqq \sum_{j \in J^{(k)}} \lambda_j^{(k)} \, \mathrm{P}_{p^{(k)} \leftarrow q^{(j)}} X_{q^{(j)}}$
  Set $\varepsilon^{(k)} \coloneqq \sum_{j \in J^{(k)}} \lambda_j^{(k)} \, e_j^{(k)}$
  Set $\sigma^{(k)} \coloneqq \sum_{j \in J^{(k)}} \lambda_j^{(k)} \, r_j^{(k)}$
  Set $d^{(k)} \coloneqq -g^{(k)}$ ▷ search direction
  Set $\xi^{(k)} \coloneqq -\|g^{(k)}\|^2 - \varepsilon^{(k)} - \sigma^{(k)}$
  **if** $-\xi^{(k)} \leq$ tol **then** ▷ stopping criterion
    **return** $p^{(k)}$
  **end if**
  Set $t^{(k)} \coloneqq 1$
  **while** $q^{(k+1)} \coloneqq \exp_{p^{(k)}}(t^{(k)} d^{(k)}) \notin \operatorname{int}\operatorname{dom} f$,
      or $\operatorname{dist}(p^{(k)}, q^{(k+1)}) < t^{(k)} \|d^{(k)}\|$ **do** ▷ trial step
    Set $t^{(k)} \coloneqq \beta \, t^{(k)}$ ▷ backtracking
  **end while**
  Evaluate $X_{q^{(k+1)}} \in \partial f(q^{(k+1)})$
  **if** $f(q^{(k+1)}) \leq f(p^{(k)}) + m \, t^{(k)} \, \xi^{(k)}$ **then**
    Set $\bar{t}^{(k)} \coloneqq t^{(k)}$ and $p^{(k+1)} \coloneqq q^{(k+1)}$ and $K \coloneqq K \cup \{k\}$ ▷ serious step
  **else**
    Set $\bar{t}^{(k)} \coloneqq 0$ and $p^{(k+1)} \coloneqq p^{(k)}$
    **while** $m \, t^{(k)} \, \xi^{(k)} \geq \big(\mathrm{P}_{p^{(k)} \leftarrow q^{(k+1)}} X_{q^{(k+1)}}, \, t^{(k)} \, d^{(k)}\big) - e_{k+1}^{(k)} - r_{k+1}^{(k)}$ **do**
      Set $t^{(k)} \coloneqq \beta \, t^{(k)}$ and $q^{(k+1)} \coloneqq \exp_{p^{(k)}}(t^{(k)} d^{(k)})$
    **end while** ▷ null step
  **end if**
  Set $J_+^{(k)} \coloneqq \big\{ j \in J^{(k)} \,\big|\, \lambda_j^{(k)} > 0 \big\}$ and $J^{(k+1)} \coloneqq J_+^{(k)} \cup \{k+1\}$
  Set $e_j^{(k+1)} \coloneqq f(p^{(k+1)}) - f(q^{(j)}) - \big(X_{q^{(j)}}, \log_{q^{(j)}} p^{(k+1)}\big)$ for all $j \in J^{(k+1)}$
  Set $r_j^{(k+1)} \coloneqq \varrho \, \big\|X_{q^{(j)}}\big\| \, \big\|\log_{q^{(j)}} p^{(k+1)}\big\|$ for all $j \in J^{(k+1)}$
  Set $k \coloneqq k+1$
**end while**

Thus, the backtracking loop in question aims to find a small enough $\bar{t}^{(k)} \leq t^{(k)}$ such that

$$\left(\mathrm{P}_{p^{(k)} \leftarrow \bar{q}^{(k+1)}} X_{\bar{q}^{(k+1)}}\,,\ \bar{t}^{(k)}\, g^{(k)}\right) < -m\,\bar{t}^{(k)}\,\xi^{(k)} - e_{k+1}^{(k)} - r_{k+1}^{(k)}, \tag{4.2}$$

for $\bar{q}^{(k+1)} \coloneqq \exp_{p^{(k)}}(-\bar{t}^{(k)} g^{(k)})$, $X_{\bar{q}^{(k+1)}} \in \partial f(\bar{q}^{(k+1)})$, and updated $e_{k+1}^{(k)}$ and $r_{k+1}^{(k)}$ accordingly. Observe now that, by the definition of linearization errors, this last inequality implies

$$f(\bar{q}^{(k+1)}) > f(p^{(k)}) + m\,\bar{t}^{(k)}\,\xi^{(k)} + r_{k+1}^{(k)} \geq f(p^{(k)}) + m\,\bar{t}^{(k)}\,\xi^{(k)},$$

which means that the newly found $\bar{q}^{(k+1)}$ is still a null iterate. Finally, we note here that the loop in Line 24 in Algorithm 1 has always stopped within a finite number of iterations in our numerical experiments.

(iii) Following Remark 3.4, the exponential map in Line 14 may be replaced by a retraction. Analogously, the logarithmic map in Line 27 and Line 26 can be replaced by the (local) inverse retraction. Moreover, the parallel transport in Line 3, and Line 4 can be replaced by a continuous vector transport T. This amounts to using the quantities $\tilde{e}_j^{(k)}$ defined in (3.14) in place of $e_j^{(k)}$ in Line 3, Line 5, and Line 27. Analogously, $\tilde{r}_j^{(k)}$ defined in (3.15) is to be used in place of $r_j^{(k)}$ in Line 3, Line 6, and Line 28. Additional conditions have to be worked out that ensure $\tilde{e}_j^{(k)} \geq 0$. We can obtain an approximate Riemannian convex bundle method along these lines, but we leave the details to future research.

## 5 CONVERGENCE OF THE METHOD

The convergence analysis of Algorithm 1 follows along the lines of [25, Chapter 6.7], and we point out where adjustments are necessary due to the nonzero curvature of the space and the fact that the objective may take the value $\infty$. Such adjustments are already necessary in the first step in the analysis, which is an adaptation of [25, Lemmas 6.73 and 6.74]. Here we have to work with the linearization errors $e_j^{(k)}$, the curvature-dependent remainders $r_j^{(k)}$, and with parallel transport. Throughout, we assume Theorem 3.1 to hold. As is customary, we carry out the convergence analysis with the stopping criterion disabled.

Lemma 5.1. *The sequences $(p^{(k)})$, $(q^{(k)})$, $(g^{(k)})$, $(\varepsilon^{(k)})$, $(\sigma^{(k)})$, $(\xi^{(k)})$ generated by Algorithm 1 possess the following properties:*

*(1) $e_j^{(k)}, r_j^{(k)} \geq 0$, and $\mathrm{P}_{p^{(k)} \leftarrow q^{(j)}} X_{q^{(j)}} \in \partial_{e_j^{(k)} + r_j^{(k)}} f(p^{(k)})$ for all $j \in J^{(k)}$ and $k \in \mathbb{N}_0$;*

*(2) $\varepsilon^{(k)}, \sigma^{(k)} \geq 0$, and $g^{(k)} \in \partial_{\varepsilon^{(k)} + \sigma^{(k)}} f(p^{(k)})$ for all $k \in \mathbb{N}_0$;*

*(3) $\xi^{(k)} \leq 0$ for all $k \in \mathbb{N}_0$;*

*(4) $p^{(k)}$ belongs to $\operatorname{int}\operatorname{dom} f$ for all $k \in \mathbb{N}_0$;*

*(5) if $\xi^{(k)} = 0$ for some $k \in \mathbb{N}_0$, then $p^{(k)}$ is a global minimizer of $f$.*

*Proof.* (1): The first statement follows directly by applying the subgradient inequality to $e_j^{(k)}$. Furthermore, since $\varrho \geq 0$ holds, it follows that $r_j^{(k)} \geq 0$ for all $j = 0, \ldots, k$ and $k \in \mathbb{N}_0$.

For the second statement, consider first the case $k = 0$. Then $p^{(0)} = q^{(0)}$ and

$$e_0^{(0)} = 0.$$

Hence

$$X_{p^{(0)}} \in \partial f(p^{(0)}) = \partial_0 f(p^{(0)}) = \partial_{e_0^{(0)}} f(p^{(0)}).$$

For $k \geq 0$ we have

$$e_j^{(k+1)} = f(p^{(k+1)}) - f(q^{(j)}) - \left(X_{q^{(j)}}, \log_{q^{(j)}} p^{(k+1)}\right),$$

for all $j \in J^{(k+1)}$, where $X_{q^{(j)}} \in \partial f(q^{(j)})$. Hence, applying (3.9), we find

$$f(p) \geq f(p^{(k+1)}) + \left(\mathrm{P}_{p^{(k+1)} \leftarrow q^{(j)}} X_{q^{(j)}}, \log_{p^{(k+1)}} p\right) - e_j^{(k+1)} - r_j^{(k+1)}$$

for all $p \in \operatorname{dom} f$, so that $\mathrm{P}_{p^{(k+1)} \leftarrow q^{(j)}} X_{q^{(j)}} \in \partial_{e_j^{(k+1)} + r_j^{(k+1)}} f(p^{(k+1)})$ for all $j \in J^{(k+1)}$.

(2): Similarly, by the definitions of $\varepsilon^{(k)}$, $\sigma^{(k)}$ and $\lambda_j$, we obtain $\varepsilon^{(k)} \geq 0$, and $\sigma^{(k)} \geq 0$ for all $k \in \mathbb{N}_0$. Now let $p \in \operatorname{dom} f$ and note that, thanks to the previous point,

$$\begin{aligned} f(p) &= \sum_{j \in J^{(k)}} \lambda_j^{(k)} f(p) \\ &\geq \sum_{j \in J^{(k)}} \lambda_j^{(k)} \left[ f(p^{(k)}) + \left(\mathrm{P}_{p^{(k)} \leftarrow q^{(j)}} X_{q^{(j)}}, \log_{p^{(k)}} p\right) - e_j^{(k)} - r_j^{(k)} \right] \\ &= f(p^{(k)}) + \left(g^{(k)}, \log_{p^{(k)}} p\right) - \varepsilon^{(k)} - \sigma^{(k)}. \end{aligned}$$

Therefore $g^{(k)} \in \partial_{\varepsilon^{(k)} + \sigma^{(k)}} f(p^{(k)})$ for all $k \in \mathbb{N}_0$.

(3): Since $\xi^{(k)} = -\|g^{(k)}\|^2 - \varepsilon^{(k)} - \sigma^{(k)}$, we obtain the statement from the previous point.

(4): This property follows immediately from the condition $p^{(0)} \in \operatorname{int} \operatorname{dom} f$ and the way subsequent serious iterates are constructed in Algorithm 1.

(5): If $\xi^{(k)} = 0$ for some $k \in \mathbb{N}_0$, then $g^{(k)} = 0$, $\varepsilon^{(k)} = 0$, and $\sigma^{(k)} = 0$. Therefore

$$0 = g^{(k)} \in \partial_0 f(p^{(k)}) = \partial f(p^{(k)}).$$

□

Remark 5.2. Lemma 5.1 item (5) gives a condition for stopping Algorithm 1 motivated by $0 \in \partial f(p^*)$ being a sufficient condition for global optimality, that is also necessary when $p^* \in \operatorname{int} \operatorname{dom} f$. Given a real number $\operatorname{tol} > 0$, the stopping criterion in Line 9 reads $-\xi^{(k)} \leq \operatorname{tol}$. This means $\|g^{(k)}\|^2 + \varepsilon^{(k)} + \sigma^{(k)} \leq \operatorname{tol}$, implying that $p^{(k)}$ is close to satisfying $0 \in \partial f(p^{(k)})$.

The following result, which is an adaptation of [25, Lemma 6.75], shows that the convergence of $\sum_{k=0}^{\infty} -\bar{t}^{(k)} \xi^{(k)}$, together with the fact that, if infinitely many serious steps occur, the search directions approach the zero tangent vector and since $\varepsilon^{(k)} + \sigma^{(k)} \to 0$, the $\varepsilon$-subdifferential approaches the subdifferential in the following sense

$$\lim_{\varepsilon \to 0} \partial_\varepsilon f(p) \coloneqq \bigcap \{\partial_\varepsilon f(p) \mid \varepsilon > 0\} = \partial_0 f(p) = \partial f(p).$$

Lemma 5.3. *Let $f_* \in \mathbb{R}$ be such that $f(p^{(k)}) \geq f_*$ for all $k \in \mathbb{N}_0$. Then the sequences $(p^{(k)})$, $(g^{(k)})$, $(\varepsilon^{(k)})$, $(\sigma^{(k)})$ generated by Algorithm 1 are such that*

*(1) $f(p^{(k)}) - f(p^{(k+1)}) \to 0$ for $k \to +\infty$*

*(2) $\sum_{k=0}^{\infty} \bar{t}^{(k)} \left(\|g^{(k)}\|^2 + \varepsilon^{(k)} + \sigma^{(k)}\right) \leq \dfrac{f(p^{(0)}) - f_*}{m}$*

*(3) if there are infinitely many serious steps $k_n \in K$, then $g^{(k_n)} \to 0$, $\varepsilon^{(k_n)} \to 0$, and $\sigma^{(k)} \to 0$ for $n \to \infty$.*

*Proof.* By construction, the sequence $\left(f(p^{(k)})\right)$ is monotone non-increasing because $\xi^{(k)} \leq 0$. Since it is also bounded from below by assumption, it converges, which shows item (1).

(2): For all $k \in \mathbb{N}_0$, hence for serious and null steps, by rearranging the inequality in Line 18 of Algorithm 1 we get

$$f(p^{(k)}) - f(p^{(k+1)}) \geq m\, \bar{t}^{(k)}(-\xi^{(k)}),$$

since $\bar{t}^{(k)} \in [0,1]$ and $\bar{t}^{(k)} \leq t^{(k)}$. By summation we obtain

$$f(p^{(0)}) - f(p^{(k+1)}) \geq m \sum_{j=0}^{k} \bar{t}^{(j)}\,(-\xi^{(j)}).$$

Since $f(p^{(k)}) \geq f_*$ for all $k \in \mathbb{N}_0$, then

$$f(p^{(0)}) - f_* \geq m \sum_{j=0}^{k} \bar{t}^{(j)}\,(-\xi^{(j)}),$$

and by passing to the limit

$$f(p^{(0)}) - f_* \geq m \sum_{j=0}^{\infty} \bar{t}^{(j)}\,(-\xi^{(j)}).$$

The claim follows because $\xi^{(j)} = -\|g^{(j)}\|^2 - \varepsilon^{(j)} - \sigma^{(j)}$.

(3): Let $K$ be the set of serious steps defined Line 19 of Algorithm 1 and suppose $|K| = +\infty$. Then $0 < \bar{t}^{(k)} \leq 1$ if $k \in K$, namely if $k = k_n$ with $k_n < k_{n+1} < \dots$ for some $n \in \mathbb{N}_0$. The claim follows from the previous statement, since the series $\sum_{n=1}^{\infty} \bar{t}^{(k_n)}(-\xi^{(k_n)})$ converges and $\|g^{(k_n)}\|^2 \geq 0$, $\varepsilon^{(k_n)} \geq 0$, and $\sigma^{(k_n)} \geq 0$ by item (2) of Lemma 5.1. □

Next, we show that if Algorithm 1 generates *infinitely* many serious steps, every accumulation point of the sequence of serious iterates is a minimizer of the objective. This is an adaptation of [25, Lemma 6.76].

Lemma 5.4. *Let $\left(p^{(k)}\right)$ be a sequence generated by Algorithm 1 and suppose that infinitely many serious steps occur. Then every accumulation point $p^* \in \operatorname{int}(\operatorname{dom} f)$ of $\left(p^{(k)}\right)$ is a minimizer of $f$.*

*Proof.* Let $p^* \in \operatorname{int}(\operatorname{dom} f)$ be an accumulation point for the sequence $\left(p^{(k)}\right)$, with subsequence $p^{(k_n)} \to p^*$. Since $f(p^{(k)})$ is monotone non-increasing and there exists a subsequence $f(p^{(k_n)})$ that converges with $\lim_{n\to\infty} f(p^{(k_n)}) \geq f(p^*)$ due to lower semi-continuity, we can conclude $f(p^{(k)}) \geq f(p^*) =: f_*$ for all $k \in \mathbb{N}_0$ and $f(p^{(k)}) \to f_*$ for $k \to +\infty$. Since $p^*$ is an accumulation point for $\left(p^{(k)}\right)$ and $p^{(k)}$ does not change through a null step, $p^*$ is an accumulation point for the subsequence $\left(p^{(k_n)}\right)_{k_n \in K}$ as well, with $K$ being the set of indices for the serious steps defined in Line 19 of Algorithm 1. By assumption, there exist infinitely many $k_n \in K$ with $k_n < k_{n+1}$ for all $n \in \mathbb{N}_0$, so that $p^{(k_n)} \to p^*$ for $n \to \infty$. From Lemma 5.1 item (2) we have $g^{(k)} \in \partial_{\varepsilon^{(k)}+\sigma^{(k)}} f(p^{(k)})$ for all $k \in \mathbb{N}_0$, so that

$$f(q) \geq f(p^{(k_n)}) + \left(g^{(k_n)}, \log_{p^{(k_n)}} q\right) - \varepsilon^{(k_n)} - \sigma^{(k_n)} \quad \text{for all } q \in \operatorname{dom} f.$$

Because $f(p^{(k)}) \geq f_*$ holds for all $k \in \mathbb{N}_0$, from Lemma 5.3 item (3) we get $g^{(k_n)} \to 0$, $\varepsilon^{(k_n)} \to 0$, and $\sigma^{(k_n)} \to 0$ for $n \to \infty$. By passing to the limit inferior in the previous inequality,

$$\begin{aligned} f(q) &\geq \liminf_{n\to\infty} \left[f(p^{(k_n)}) + \left(g^{(k_n)}, \log_{p^{(k_n)}} q\right) - \varepsilon^{(k_n)} - \sigma^{(k_n)}\right] \\ &= \liminf_{n\to\infty} f(p^{(k_n)}) \geq f_* \end{aligned}$$

for all $q \in \operatorname{dom} f$, where the second inequality is due to the lower semi-continuity of $f$. □

We now show that if the number of serious iterates generated by Algorithm 1 is *finite*, then the last such iterate is a minimizer of the objective.

Lemma 5.5. *Let $\left(p^{(k)}\right)$ be a sequence generated by Algorithm 1 and suppose that only a finite number of serious steps occur, so that $p^{(k)} = p^{(k_*)}$ holds for all $k \geq k_*$ for some $k_* \in \mathbb{N}_0$. Then $p^* \coloneqq p^{(k_*)}$ is a minimizer of $f$.*

*Proof.* The main idea in this proof is to bound the objective of the subproblem in Line 3 from above and below using the decomposition of the index set $J^{(k)}$, and to show that this implies the convergence to zero of $\xi^{(k)}$. Since $p^{(k+1)} = p^{(k)}$ for all $k \geq k_*$, from the update rules of Algorithm 1 we have that $e_j^{(k+1)} = e_j^{(k)}$, and $r_j^{(k+1)} = r_j^{(k)}$ for all $j \in J_+^{(k)}$ and all $k \geq k_*$. Let

$$\eta^{(k)} \coloneqq \sum_{j \in J_+^{(k)}} \lambda_j^{(k)}\, e_j^{(k+1)} = \varepsilon^{(k)},$$
$$\text{and} \quad \psi^{(k)} \coloneqq \sum_{j \in J_+^{(k)}} \lambda_j^{(k)}\, r_j^{(k+1)} = \sigma^{(k)},$$

for all $k \geq k_*$. Let us denote the objective function of the subproblem defined in Line 3 of Algorithm 1 by

$$Q^{(k)}(\lambda) \coloneqq \frac{1}{2}\Big\| \sum_{j \in J^{(k)}} \lambda_j \, \mathrm{P}_{p^{(k)} \leftarrow q^{(j)}} X_{q^{(j)}} \Big\|^2 + \sum_{j \in J^{(k)}} \lambda_j\, e_j^{(k)} + \sum_{j \in J^{(k)}} \lambda_j\, r_j^{(k)}.$$

Suppose that $\mu \in [0,1]$ is an arbitrary number and let $\overline{\lambda}_j$ be defined for all $j \in J^{(k)} = J_+^{(k-1)} \cup \{k\}$ as

$$\overline{\lambda}_j = \begin{cases} \mu & \text{if } j = k \\ (1-\mu)\,\lambda_j^{(k-1)} & \text{if } j \in J_+^{(k-1)}, \end{cases}$$

where the numbers $\lambda_j^{(k-1)}$ are the components of the solution $\lambda^{k-1}$ to the subproblem in Line 3 of Algorithm 1 at the $(k-1)$-st iteration. Note that $\overline{\lambda}_j \geq 0$ for all $j \in J^{(k)}$ and

$$\sum_{j \in J^{(k)}} \overline{\lambda}_j = \mu + \sum_{j \in J_+^{(k-1)}} (1-\mu)\,\lambda_j^{(k-1)} = \mu + (1-\mu) \sum_{j \in J^{(k-1)}} \lambda_j^{(k-1)} = 1$$

since the numbers $\lambda_j^{(k-1)}$ for $j \in J^{(k-1)} \setminus J_+^{(k-1)}$ are zero for all $k \in \mathbb{N}_0$ by definition of $J_+^{(k)}$ and $\lambda_j^{(k)}$. This shows that the vector $\overline{\lambda} \in \mathbb{R}^{|J^{(k)}|}$ is admissible for the quadratic program defined in Line 3 of Algorithm 1. Therefore

$$Q^{(k)}(\lambda^{(k)}) \leq Q^{(k)}(\overline{\lambda}) \quad \text{for all } k \geq k_*.$$

We need to estimate $Q^{(k)}(\overline{\lambda})$, and to do this we use the decomposition $J^{(k)} = J_+^{(k-1)} \cup \{k\}$. Let $X_{q^{(j)}} \in \partial f(q^{(j)})$ for all $j \in J^{(k)}$ and $k > k_*$. From the update rules of Algorithm 1, and because $p^{(k)} = p^{(k-1)}$ holds for all $k > k_*$, we get

$$\begin{aligned}
&\sum_{j \in J^{(k)}} \overline{\lambda}_j \, \mathrm{P}_{p^{(k)} \leftarrow q^{(j)}} X_{q^{(j)}} \\
&\quad = \mu\, \mathrm{P}_{p^{(k)} \leftarrow q^{(k)}} X_{q^{(k)}} + (1-\mu) \sum_{j \in J^{(k-1)}} \lambda_j^{(k-1)}\, \mathrm{P}_{p^{(k)} \leftarrow q^{(j)}} X_{q^{(j)}} \\
&\quad = \mu\, \mathrm{P}_{p^{(k)} \leftarrow q^{(k)}} X_{q^{(k)}} + (1-\mu) \sum_{j \in J^{(k-1)}} \lambda_j^{(k-1)}\, \mathrm{P}_{p^{(k-1)} \leftarrow q^{(j)}} X_{q^{(j)}} \\
&\quad = \mu\, \mathrm{P}_{p^{(k)} \leftarrow q^{(k)}} X_{q^{(k)}} + (1-\mu)\, g^{(k-1)},
\end{aligned}$$

as well as

$$\begin{aligned}\sum_{j\in J^{(k)}} \overline{\lambda}_j\, e_j^{(k)} &= \overline{\lambda}^{(k)}\, e_k^{(k)} + \sum_{j\in J_+^{(k-1)}} \overline{\lambda}_j\, e_j^{(k)} \\ &= \mu\, e_k^{(k)} + \sum_{j\in J_+^{(k-1)}} (1-\mu)\,\lambda_j^{(k-1)}\, e_j^{(k-1)} \\ &= \mu\, e_k^{(k)} + (1-\mu)\,\eta^{(k-1)}.\end{aligned}$$

Analogously, we obtain

$$\sum_{j\in J^{(k)}} \overline{\lambda}_j\, r_j^{(k)} = \mu\, r_k^{(k)} + (1-\mu)\,\psi^{(k-1)}.$$

By plugging these into the definition of $Q^{(k)}(\overline{\lambda})$ we obtain

$$\begin{aligned}Q^{(k)}(\lambda^{(k)}) &\le Q^{(k)}(\overline{\lambda}) \\ &= \frac{1}{2}\left\|\mu\, \mathrm{P}_{p^{(k)}\leftarrow q^{(k)}} X_{q^{(k)}} + (1-\mu)\, g^{(k-1)}\right\|^2 \\ &\quad + \mu\, e_k^{(k)} + (1-\mu)\,\eta^{(k-1)} + \mu\, r_k^{(k)} + (1-\mu)\,\psi^{(k-1)}.\end{aligned}$$

Since $\mu \in [0,1]$, we see that

$$Q^{(k)}(\lambda^{(k)}) \le \min_{\mu\in[0,1]} M^{(k)}(\mu),$$

where

$$\begin{aligned}M^{(k)}(\mu) \coloneqq \frac{1}{2}&\left\|\mu\, \mathrm{P}_{p^{(k)}\leftarrow q^{(k)}} X_{q^{(k)}} + (1-\mu)\, g^{(k-1)}\right\|^2 \\ &+ \mu\, e_k^{(k)} + (1-\mu)\,\eta^{(k-1)} + \mu\, r_k^{(k)} + (1-\mu)\,\psi^{(k-1)}.\end{aligned}$$

Let $\mu^{(k)}$ be a solution to the convex, quadratic problem

$$\operatorname*{arg\,min}_{\mu\in[0,1]} M^{(k)}(\mu)$$

with optimal value $w^{(k)} = M^{(k)}(\mu^{(k)})$. Then

$$w^{(k)} \le M^{(k)}(\mu) \quad \text{for all } \mu \in [0,1]. \tag{5.1}$$

Hence, for all $k > k_*$,

$$\begin{aligned}w^{(k)} &\le M^{(k)}(0) \\ &= \frac{1}{2}\|g^{(k-1)}\|^2 + \eta^{(k-1)} + \psi^{(k-1)} \\ &= \frac{1}{2}\Big\| \sum_{j\in J^{(k-1)}} \lambda_j^{(k-1)}\, \mathrm{P}_{p^{(k-1)}\leftarrow q^{(j)}} X_{q^{(j)}} \Big\|^2 \\ &\quad + \sum_{j\in J^{(k-1)}} \lambda_j^{(k-1)}\, e_j^{(k-1)} + \sum_{j\in J^{(k-1)}} \lambda_j^{(k-1)}\, r_j^{(k-1)} \\ &= Q^{(k-1)}(\lambda^{k-1}) \\ &\le w^{(k-1)}.\end{aligned}$$

We can therefore conclude

$$0 \leq w^{(k)} \leq w^{(k-1)} \leq w^{(k_*)} \quad \text{for all } k > k_* \tag{5.2}$$

and

$$\|g^{(k)}\| \leq \sqrt{2w^{(k_*)}}, \quad \eta^{(k)} \leq w^{(k_*)}, \quad \text{and } \psi^{(k)} \leq w^{(k_*)} \quad \text{for all } k \geq k_*. \tag{5.3}$$

Since the $(k-1)$-th step is a null step, from the condition in the loop in Line 24, see also Item (ii), we get

$$\begin{aligned}\left(\mathrm{P}_{p^{(k)} \leftarrow q^{(k)}} X_{q^{(k)}}\,,\, t^{(k-1)}\, g^{(k-1)}\right) < m\, t^{(k-1)} \left(\|g^{(k-1)}\|^2 + \eta^{(k-1)} + \psi^{(k-1)}\right) \\ - e_k^{(k)} - r_k^{(k)},\end{aligned}$$

which implies

$$\left(\mathrm{P}_{p^{(k)} \leftarrow q^{(k)}} X_{q^{(k)}}\,,\, g^{(k-1)}\right) < m \left(\|g^{(k-1)}\|^2 + \eta^{(k-1)} + \psi^{(k-1)}\right) - e_k^{(k)} - r_k^{(k)}. \tag{5.4}$$

Let us further estimate $M^{(k)}(\mu)$:

$$\begin{aligned}M^{(k)}(\mu) &= \frac{1}{2}\left\|\mu\, \mathrm{P}_{p^{(k-1)} \leftarrow q^{(k)}} X_{q^{(k)}} + (1-\mu)\, g^{(k-1)}\right\|^2 \\ &\quad + \mu\, e_k^{(k)} + (1-\mu)\, \eta^{(k-1)} + \mu\, r_k^{(k)} + (1-\mu)\, \psi^{(k-1)} \\ &= \frac{1}{2}\left\|\mu\left(\mathrm{P}_{p^{(k-1)} \leftarrow q^{(k)}} X_{q^{(k)}} - g^{(k-1)}\right) + g^{(k-1)}\right\|^2 \\ &\quad + \mu\, e_k^{(k)} + (1-\mu)\, \eta^{(k-1)} + \mu\, r_k^{(k)} + (1-\mu)\, \psi^{(k-1)} \\ &= \frac{1}{2}\mu^2 \|\mathrm{P}_{p^{(k-1)} \leftarrow q^{(k)}} X_{q^{(k)}} - g^{(k-1)}\|^2 + \frac{1}{2}\|g^{(k-1)}\|^2 \\ &\quad + \mu \left[\left(\mathrm{P}_{p^{(k-1)} \leftarrow q^{(k)}} X_{q^{(k)}}\,,\, g^{(k-1)}\right) - \|g^{(k-1)}\|^2\right] \\ &\quad + \mu\, e_k^{(k)} + (1-\mu)\, \eta^{(k-1)} + \mu\, r_k^{(k)} + (1-\mu)\, \psi^{(k-1)}.\end{aligned}$$

From (5.3) we see that

$$\frac{1}{2}\|g^{(k-1)}\|^2 \leq \frac{1}{2}\|g^{(k-1)}\|^2 + \eta^{(k-1)} + \psi^{(k-1)} \leq w^{(k-1)},$$

and plugging this into the previous inequality and using (5.4) yields

$$\begin{aligned}M^{(k)}(\mu) &\leq \frac{1}{2}\mu^2 \|\mathrm{P}_{p^{(k-1)} \leftarrow q^{(k)}} X_{q^{(k)}} - g^{(k-1)}\|^2 \\ &\quad + \mu \left[m\left(\|g^{(k-1)}\|^2 + \eta^{(k-1)} + \psi^{(k-1)}\right) - e_k^{(k)} - r_k^{(k)} - \|g^{(k-1)}\|^2\right] \\ &\quad + \mu\, e_k^{(k)} - \mu\, \eta^{(k-1)} + \mu\, r_k^{(k)} - \mu\, \psi^{(k-1)} + w^{(k-1)} \\ &= \frac{1}{2}\mu^2 \|\mathrm{P}_{p^{(k-1)} \leftarrow q^{(k)}} X_{q^{(k)}} - g^{(k-1)}\|^2 \\ &\quad - \mu\,(1-m)\left(\|g^{(k-1)}\|^2 + \eta^{(k-1)} + \psi^{(k-1)}\right) + w^{(k-1)}.\end{aligned} \tag{5.5}$$

The sequences $\left(g^{(k)}\right)$, $\left(\eta^{(k)}\right)$, and $\left(\psi^{(k)}\right)$ are bounded thanks to (5.3). Moreover, the subdifferential is locally bounded; see, e. g., [43, Theorem 4.6]. Hence, there exists a constant $a > 0$ such that

$$\|X_{q^{(k)}}\| = \|\mathrm{P}_{p^{(k-1)} \leftarrow q^{(k)}} X_{q^{(k)}}\| \leq a, \quad \|g^{(k)}\| \leq a, \quad \eta^{(k)} \leq a, \quad \psi^{(k)} \leq a, \tag{5.6}$$

for all $k > k_*$, so that

$$\|\mathrm{P}_{p^{(k-1)} \leftarrow q^{(k)}} X_{q^{(k)}} - g^{(k-1)}\|^2 \leq \left(\|\mathrm{P}_{p^{(k-1)} \leftarrow q^{(k)}} X_{q^{(k)}}\| + \|g^{(k-1)}\|\right)^2 \leq 4\, a^2.$$

Plugging this into (5.5) and defining

$$\theta^{(k)}(\mu) \coloneqq 2\, a^2\, \mu^2 - \left[(1-m)\left(\|g^{(k-1)}\|^2 + \eta^{(k-1)} + \psi^{(k-1)}\right)\right]\mu + w^{(k-1)},$$

we conclude that

$$M^{(k)}(\mu) \leq \theta^{(k)}(\mu) \quad \text{for all } k > k_*. \tag{5.7}$$

Since $\theta^{(k)}$ is a quadratic function of $\mu$, it is straightforward to see that it achieves its minimum value at the point

$$\begin{aligned}
\mu_*^{(k)} &= \frac{1}{4\, a^2}\left[(1-m)\left(\|g^{(k-1)}\|^2 + \eta^{(k-1)} + \psi^{(k-1)}\right)\right] \\
&\leq \frac{1}{4\, a^2}\left[\|g^{(k-1)}\|^2 + \eta^{(k-1)} + \psi^{(k-1)}\right] \\
&\leq \frac{a^2 + 2\, a}{4\, a^2},
\end{aligned}$$

where we used (5.6). Note that we can choose $a \geq \frac{2}{3}$ without loss of generality, so that $\mu_*^{(k)} \leq 1$ for all $k > k_*$. The minimal value of $\theta^{(k)}$ on $[0,1]$ is then

$$\begin{aligned}
\theta^{(k)}(\mu_*^{(k)}) = 2\, a^2\, {\mu_*^{(k)}}^2 - (1-m)\left(\|g^{(k-1)}\|^2 + \eta^{(k-1)} + \psi^{(k-1)}\right) \\
\mu_*^{(k)} + w^{(k-1)}.
\end{aligned} \tag{5.8}$$

Hence, since $\mu_*^{(k)} \in [0,1]$, through (5.1), (5.3), (5.7), and (5.8) we get

$$\begin{aligned}
w^{(k)} &= M^{(k)}(\mu^{(k)}) \leq M^{(k)}(\mu_*^{(k)}) \leq \theta^{(k)}(\mu_*^{(k)}) \\
&= -\frac{1}{8\, a^2}(1-m)^2\left(\|g^{(k-1)}\|^2 + \eta^{(k-1)} + \psi^{(k-1)}\right)^2 + w^{(k-1)},
\end{aligned}$$

for all $k > k_*$, where the last equality comes from substituting the expression for $\mu_*^{(k)}$. Hence,

$$w^{(k)} \leq -\frac{1}{8\, a^2}(1-m)^2\left(\|g^{(k-1)}\|^2 + \eta^{(k-1)} + \psi^{(k-1)}\right)^2 + w^{(k-1)}$$

for all $k > k_*$. Rearranging the terms and summing the resulting inequalities for $j = k_* + 1, \ldots, k+1$ yields

$$C \sum_{j=k_*}^{k} \left(\|g^{(j)}\|^2 + \eta^{(j)} + \psi^{(j)}\right)^2 \leq w^{(k_*)} - w^{(k+1)},$$

where $C = \dfrac{1}{8a^2}(1-m)^2$. Now, $w^{(k+1)} \geq 0$ due to (5.2), thus by passing to the limit we have

$$\sum_{j=k_*}^{\infty} \left(\|g^{(j)}\|^2 + \eta^{(j)} + \psi^{(j)}\right)^2 < +\infty.$$

Thus $g^{(k)} \to 0$, $\eta^{(k)} = \varepsilon^{(k)} \to 0$, and $\psi^{(k)} = \sigma^{(k)} \to 0$ as $k \to +\infty$, for $k \geq k_*$. The claim follows by reasoning as in the second part of Lemma 5.4. □

We can now state the following preliminary convergence result.

Theorem 5.6. *Every accumulation point of the sequence $(p^{(k)})$ generated by Algorithm 1 that lies in* $\operatorname{int}(\operatorname{dom} f)$ *is a solution to Problem 1.1.*

*Proof.* The claim follows by combining Lemma 5.4 and Lemma 5.5. □

The following lemma guarantees the existence of accumulation points of the sequence of serious iterates of Algorithm 1 under the assumption that the set of interior solutions of Problem 1.1 is nonempty. To show this, we adapted [25, Lemma 6.79], using the function $\zeta_{1,\omega}(\delta)$ defined in (3.5).

Lemma 5.7. *Let the set $S = \{p^* \in \operatorname{int}(\operatorname{dom} f) \,|\, f(p^*) = \inf_{p\in\mathcal{M}} f(p)\}$ of interior solutions to Problem 1.1 be nonempty. If $p^* \in S$ and $(p^{(k)})$ is the sequence of serious iterates generated by Algorithm 1, then the following hold:*

*(1) For all $n \in \mathbb{N}_0$ and all $k \geq n$, we have*

$$\operatorname{dist}^2(p^*, p^{(k)}) \;\leq\; \operatorname{dist}^2(p^*, p^{(n)}) \;+\; \zeta_{1,\omega}(\delta) \sum_{j=n}^{k} \left[\operatorname{dist}^2(p^{(j+1)}, p^{(j)}) \;+\; 2\,\bar{t}^{(j)}\left(\varepsilon^{(j)} \;+\; \sigma^{(j)}\right)\right].$$

*(2)* $\zeta_{1,\omega}(\delta) \sum_{j=n}^{\infty} \left[\operatorname{dist}^2(p^{(j+1)}, p^{(j)}) + 2\,\bar{t}^{(j)}\left(\varepsilon^{(j)} + \sigma^{(j)}\right)\right] < +\infty.$

*(3) The sequence $(p^{(k)})$ is bounded.*

*Proof.* Since $g^{(k)} \in \partial_{\varepsilon^{(k)}+\sigma^{(k)}} f(p^{(k)})$ for all $k \in \mathbb{N}_0$ due to item (2) of Lemma 5.1 and since $f(p^{(k)}) \geq f(p^*)$ holds by assumption, then

$$0 \geq f(p^*) - f(p^{(k)}) \geq \left(g^{(k)}\,,\, \log_{p^{(k)}} p^*\right) - \varepsilon^{(k)} - \sigma^{(k)},$$

which reads

$$\left(g^{(k)}\,,\, \log_{p^{(k)}} p^*\right) \leq \varepsilon^{(k)} + \sigma^{(k)}.$$

For $\bar{t}^{(k)} \in [0,1]$, by definition,

$$\log_{p^{(k)}} p^{(k+1)} = \bar{t}^{(k)} d^{(k)} = -\bar{t}^{(k)} g^{(k)}$$

so that

$$-\left(\log_{p^{(k)}} p^{(k+1)}\,,\, \log_{p^{(k)}} p^*\right) \leq \bar{t}^{(k)}\left(\varepsilon^{(k)} + \sigma^{(k)}\right). \tag{5.9}$$

From [47, Lemma 6] we have

$$\begin{aligned}\operatorname{dist}^2(p^*, p^{(k+1)}) \leq \operatorname{dist}^2(p^*, p^{(k)}) + \zeta_{1,\omega}\left(\operatorname{dist}(p^*, p^{(k)})\right) \operatorname{dist}^2(p^{(k)}, p^{(k+1)}) \\ - 2\left(\log_{p^{(k)}} p^*\,,\, \log_{p^{(k)}} p^{(k+1)}\right).\end{aligned} \tag{5.10}$$

Combining now (5.9) and (5.10) we get

$$\operatorname{dist}^2(p^*, p^{(k+1)}) \;\leq\; \operatorname{dist}^2(p^*, p^{(k)}) \;+\; \zeta_{1,\omega}\left(\operatorname{dist}(p^*, p^{(k)})\right) \operatorname{dist}^2(p^{(k+1)}, p^{(k)}) \;+\; 2\,\bar{t}^{(k)}\left(\varepsilon^{(k)} \;+\; \sigma^{(k)}\right).$$

Since $\zeta_{1,\omega}(\operatorname{dist}(p^*, p^{(k)})) \geq 1$ holds, we further obtain

$$\operatorname{dist}^2(p^*, p^{(k+1)}) \;\leq\; \operatorname{dist}^2(p^*, p^{(k)}) + \zeta_{1,\omega}(\operatorname{dist}(p^*, p^{(k)}))\left[\operatorname{dist}^2(p^{(k+1)}, p^{(k)}) + 2\,\bar{t}^{(k)}\left(\varepsilon^{(k)} + \sigma^{(k)}\right)\right],$$

which yields the claim after summation because $\zeta_{1,\omega}(\operatorname{dist}(p^*, p^{(k)})) \leq \zeta_{1,\omega}(\delta)$ for all $p^*, p^{(k)} \in \operatorname{dom} f$.

(2): By definition, for $\bar{t}^{(k)} \in [0,1]$ we have

$$\begin{aligned}\operatorname{dist}^2(p^{(k+1)}, p^{(k)}) &= \|\log_{p^{(k)}} p^{(k+1)}\|^2 = \left[\bar{t}^{(k)}\right]^2 \|d^k\|^2 \\ &= \left[\bar{t}^{(k)}\right]^2 \|g^{(k)}\|^2 \leq 2\left[\bar{t}^{(k)}\right]^2 \|g^{(k)}\|^2.\end{aligned}$$

We have

$$\begin{aligned}&\zeta_{1,\omega}(\delta) \sum_{j=0}^{\infty}\left(\operatorname{dist}^2(p^{(j+1)}, p^{(j)}) + 2\,\bar{t}^{(j)}\left(\varepsilon^{(j)} + \sigma^{(j)}\right)\right) \\ &\qquad \leq 2\,\zeta_{1,\omega}(\delta) \sum_{j=0}^{\infty}\left(\left[\bar{t}^{(j)}\right]^2 \|g^{(j)}\|^2 + \bar{t}^{(j)}\left(\varepsilon^{(j)} + \sigma^{(j)}\right)\right) \\ &\qquad \leq 2\,\zeta_{1,\omega}(\delta) \sum_{j=0}^{\infty}\left(\bar{t}^{(j)} \|g^{(j)}\|^2 + \bar{t}^{(j)}\left(\varepsilon^{(j)} + \sigma^{(j)}\right)\right) < +\infty,\end{aligned}$$

where the last inequality is due to item (2) of Lemma 5.3. This gives the claim since $\zeta_{1,\omega}(\delta)$ is finite as $\delta < +\infty$ if $\omega < 0$ by Theorem 3.1, and, by (3.5), $\zeta_{1,\omega}(\delta) = 1$ for $\omega \geq 0$ and any $\delta$.

item (1) and item (2) imply the fact that $\left(p^{(k)}\right)$ is bounded. □

We can finally prove the following convergence result, adapted from [25, Theorem 6.80].

Theorem 5.8. *Let the set* $S = \left\{p^* \in \operatorname{int}(\operatorname{dom} f) \,\middle|\, f(p^*) = \inf_{p \in \mathcal{M}} f(p)\right\}$ *of interior solutions to Problem 1.1 be nonempty. Every sequence of serious iterates* $\left(p^{(k)}\right)$ *generated by Algorithm 1 that does not have accumulation points on the boundary of* $\operatorname{dom} f$ *converges to a minimizer of* $f$.

*Proof.* From item (3) of Lemma 5.7, we have that the sequence $\left(p^{(k)}\right)$ is bounded, hence there exists at least one accumulation point $p^* \in \mathcal{M}$. Since the accumulation points of $\left(p^{(k)}\right)$ do not lie outside of the boundary of $\operatorname{dom} f$ by assumption, we have $p^* \in \operatorname{int}(\operatorname{dom} f)$, and it is a minimizer of $f$ because of Theorem 5.6, hence $p^* \in S$.

The idea is to show that the sequence $\left(p^{(k)}\right)$ converges to $p^*$. To this end, let $\varepsilon > 0$. Since $\left(p^{(k)}\right)$ is bounded, there exists a subsequence that converges to $p^*$, and thus there exists a number $n \in \mathbb{N}_0$ such that

$$\operatorname{dist}^2(p^*, p^{(n)}) \leq \frac{1}{2}\varepsilon$$

and, because of item (2) of Lemma 5.7,

$$\zeta_{1,\omega}(\delta) \sum_{j=n}^{\infty}\left(\operatorname{dist}^2(p^{(j+1)}, p^{(j)}) + 2\,t^{(j)}\left(\varepsilon^{(j)} + \sigma^{(j)}\right)\right) \leq \frac{1}{2}\varepsilon.$$

Substituting these inequalities in the first point of Lemma 5.7 gives

$$\operatorname{dist}^2(p^*, p^{(k)}) \leq \operatorname{dist}^2(p^*, p^{(n)}) + \frac{1}{2}\varepsilon \leq \varepsilon,$$

for all $k \geq n$. The claim follows from $\varepsilon > 0$ being arbitrary. □

## 6 NUMERICAL EXAMPLES

The Riemannian Convex Bundle Method (RCBM, Algorithm 1) has been implemented in Julia 1.10 within `Manopt.jl` 0.5. The implementation[1] is available with the level of generality provided by retractions and vector transports, as discussed in Item (iii) in Section 4. The numerical experiments were performed on a MacBook Pro M1, 16 GB RAM running macOS 14.2.

In the following, we compare the RCBM to two other algorithms for the minimization of non-smooth functions, namely the proximal bundle algorithm (PBA) introduced in [27], and the subgradient method (SGM) which was introduced for optimization on manifolds in [21]. In one of the experiments, namely the signal denoising problem described in Section 6.2, we also compared the RCBM to the cyclic proximal point algorithm (CPPA) presented in [4]. All of these methods are available in `Manopt.jl`. In particular, we provide an implementation of the algorithm described in [27], which was not publicly available before, to the best of our knowledge. All algorithms have been used with their default parameters when available, except where explicitly stated otherwise. As for the PBA, the proximal parameter is initialized at $\mu_0 = \frac{1}{2}$, thereafter being updated with $\mu_i = \log(i+1)$ where $i > 0$ is the $i$-th iteration. The RCBM, the PBA, and the SGM require a subgradient oracle, whereas the CPPA requires the proximal maps of the objective function. The tolerance parameter from Remark 5.2 for stopping the RCBM was set to $\operatorname{tol} = 10^{-8}$, and its maximal number of iterations was set to 5000, in accordance with the default iteration cap of all other algorithms. For the RCBM, we set the default values for the descent-test parameter to $m = 10^{-3}$, and for the contraction factor $\beta = 0.975$. For practical reasons, we need to limit the maximal number of elements the bundle can contain. Hence, we set a maximal bundle size cap of 25 elements, used for all examples unless specified differently. When this limit is reached, the oldest element is removed from the bundle unless it coincides with the last serious iterate, in which case it is kept.

Most examples feature $\Omega \leq 0$, i. e., the sectional curvature is everywhere non-positive. This is the case, in particular, on Hadamard manifolds. These manifolds support real-valued, globally convex functions. In this case, we may choose the diameter $\delta = \operatorname{diam}(\operatorname{dom} f)$ to be finite but arbitrarily large. In practice, we discuss the value of $\delta$ directly in the respective experiment subsection. Two of the examples featured in this section, namely the Riemannian median on the sphere and the spectral Procrustes problem, are meant to investigate the performance of the RCBM outside the setting of geodesic convexity. This is because the objective functions in these experiments are not necessarily geodesically convex, even when restricted to an arbitrary strongly convex set, although they are still convex in the Euclidean sense.

The runtimes of all experiments in this section have been measured with the `BenchmarkTools.jl` package [17]. This package performs either $10^5$ runs of some given code or stops after 5 s, whichever happens first, to then take the median of the measured runtimes. Since a first run is usually invoking some just-in-time compilation in Julia, we perform one call to the solver before starting the benchmark. We use this first run also to record the number of iterations and the objective values that are reported in the results. To solve the quadratic program Problem P2, we use the `RipQP.jl` package [38] with its default parameters, paired with the `QuadraticModels.jl` package [39] to represent the quadratic problem.

The Riemannian distance plays a key role in our examples. By defining it as $\operatorname{dist}_q \colon \mathcal{M} \to \mathbb{R}$ by $\operatorname{dist}_q(p) = \operatorname{dist}(p, q)$, the subdifferential at $p = q$ is given by

$$\partial \operatorname{dist}_p(p) = \left\{ X_p \in \mathcal{T}_p\mathcal{M} \,\middle|\, \|X_p\| \leq 1 \right\}, \tag{6.1}$$

see, e. g., [26, Lemma 4.2]. We use this fact in computing subgradients for the objective functions in our examples. In practice, we sample a random unit-length tangent vector $\bar{X}_p$ at $p$ when $p = q$, so that

[1] manoptjl.org/stable/solvers/convex_bundle_method/

we get $\bar{X}_p \in \partial \operatorname{dist}_p(p)$, and we compute

$$\partial \operatorname{dist}_q(p) = \{\operatorname{grad} \operatorname{dist}_q(p)\} = \left\{-\frac{\log_p q}{\operatorname{dist}(p,q)}\right\} \tag{6.2}$$

otherwise. Further note that, with a similar notation,

$$\partial \operatorname{dist}_q^2(p) = \{\operatorname{grad} \operatorname{dist}_q^2(p)\} = \{-2 \log_p q\}. \tag{6.3}$$

### 6.1 RIEMANNIAN MEDIAN

We discuss three experiments.[2] For the first two, let $\mathcal{M}$ be a Hadamard manifold, i. e., a complete, simply connected Riemannian manifold with non-positive curvature $\Omega \leq 0$ everywhere. Let $q^{[1]}, \dots, q^{[N]} \in \mathcal{M}$ be $N = 1000$ Gaussian random data points, and let $f\colon \mathcal{M} \to \mathbb{R}$ be defined by the globally geodesically convex function

$$f(p) = \sum_{j=1}^{N} w_j \operatorname{dist}(p, q^{[j]}),$$

where $w_j$, $j = 1, \dots, N$ are positive weights such that $\sum_{j=1}^{N} w_j = 1$. The Riemannian geometric median $p^*$ of the points

$$\left\{q^{[1]}, \dots, q^{[N]} \,\middle|\, q^{[j]} \in \mathcal{M} \text{ for all } j = 1, \dots, N\right\}$$

is then defined as

$$p^* \coloneqq \operatorname*{arg\,min}_{p \in \mathcal{M}} f(p),$$

where equality is justified since $p^*$ is uniquely determined on Hadamard manifolds; see, e. g., [4, p.10]. In our experiments, we choose weights $w_j = \frac{1}{N}$. By (6.2), the subgradients of the objective function are given by

$$\partial f(p) = \frac{1}{N} \sum_{j=1}^{N} \partial \operatorname{dist}_{q^{[j]}}(p),$$

where we choose a random subgradient from the subdifferential. The initial point for the algorithms is chosen as one of the two points that realize the maximal distance within the dataset and the diameter $\delta$ is set to two times the maximal distance between the points in the dataset. The first column of Table 6.1, Table 6.2, and Table 6.3 shows the manifold dimension for each run. Here and in all subsequent tables, we highlight the best results in this color.

The first experiment is set on the hyperbolic spaces $\mathcal{M} = \mathcal{H}^n$, $n = 2^k$ for $k = 1, 2, 5, 10, 15$, equipped with the Minkowski metric of signature $(+, +, \dots, -)$. Note that here we have $\omega = \Omega = -1$ everywhere. The results are summarized in Table 6.1. The RCBM and the SGM are largely comparable in terms of their iteration and runtime performance, with the RCBM being faster. The SGM provides the lowest objective in all runs except for the highest-dimensional experiment. The optimal objective values attained by the RCBM and the PBA are the same.

For the second experiment, we consider symmetric positive definite matrices $\mathcal{M} = \mathcal{P}(n)$, $n = 2, 5, 10, 15$, endowed with the affine-invariant metric that makes $\mathcal{P}(n)$ a Hadamard manifold. Note that here we have $\omega = -\frac{1}{2}$, and $\Omega = 0$ from [18, Appendix I, Proposition I.1]. A comparison between the algorithms is presented in Table 6.2. The PBA outperforms the RCBM in dimension 15, while the RCBM is faster than the other algorithms in the remaining experiments, both in terms of number of iterations and CPU runtime. The RCBM lies between the PBA and SGM in terms of number of iterations, while being more comparable to the former. All algorithms provide identical objective values up to the five decimal digits shown.

[2] juliamanifolds.github.io/ManoptExamples.jl/stable/examples/RCBM-Median/

| | RCBM | | | PBA | | | SGM | | |
|---|---|---|---|---|---|---|---|---|---|
| Dim. | Iter. | Time (s) | Obj. | Iter. | Time (s) | Obj. | Iter. | Time (s) | Obj. |
| 2 | 9 | $5.24 \cdot 10^{-3}$ | 1.051 92 | 251 | $1.32 \cdot 10^{-1}$ | 1.051 92 | 18 | $8.11 \cdot 10^{-3}$ | 1.047 48 |
| 4 | 8 | $4.70 \cdot 10^{-3}$ | 1.075 16 | 230 | $1.32 \cdot 10^{-1}$ | 1.075 16 | 19 | $9.53 \cdot 10^{-3}$ | 1.055 18 |
| 32 | 15 | $1.52 \cdot 10^{-2}$ | 1.085 59 | 234 | $1.80 \cdot 10^{-1}$ | 1.085 59 | 25 | $2.09 \cdot 10^{-2}$ | 1.085 59 |
| 1024 | 16 | $2.85 \cdot 10^{-1}$ | 1.097 06 | 234 | 4.01 | 1.097 06 | 23 | $4.00 \cdot 10^{-1}$ | 1.097 06 |
| 32 768 | 16 | 7.34 | 1.068 10 | 229 | $9.13 \cdot 10^{1}$ | 1.068 10 | 21 | 8.82 | 1.064 88 |

Table 6.1: Comparison between the RCBM, the PBA, and the SGM on $\mathcal{H}^n$ with varying dimensions for the Riemannian median example from Section 6.1.

| | RCBM | | | PBA | | | SGM | | |
|---|---|---|---|---|---|---|---|---|---|
| Dim. | Iter. | Time (s) | Obj. | Iter. | Time (s) | Obj. | Iter. | Time (s) | Obj. |
| 3 | 43 | $3.04 \cdot 10^{-1}$ | 0.260 85 | 57 | $4.42 \cdot 10^{-1}$ | 0.260 85 | 4629 | $4.65 \cdot 10^{1}$ | 0.260 85 |
| 15 | 49 | 2.01 | 0.436 54 | 75 | 1.75 | 0.436 54 | 1727 | $4.05 \cdot 10^{1}$ | 0.436 54 |
| 55 | 15 | 1.31 | 0.618 06 | 89 | 6.15 | 0.618 06 | 776 | $5.34 \cdot 10^{1}$ | 0.618 06 |
| 120 | 6 | 1.20 | 0.764 03 | 123 | $1.54 \cdot 10^{1}$ | 0.764 03 | 438 | $5.36 \cdot 10^{1}$ | 0.764 03 |

Table 6.2: Comparison between the RCBM, the PBA, and the SGM on $\mathcal{P}(n)$ with varying dimensions for the Riemannian median example from Section 6.1.

For the third experiment, we consider the sphere $\mathcal{S}^n$ for $n = 2^k$ for $k = 1, 2, 5, 10, 15$, endowed with the metric inherited from the embedding into $\mathbb{R}^{n+1}$. Note that a major difference here is that this manifold has constant positive sectional curvature $\omega = \Omega = 1$. In this case, we lose the global convexity of the Riemannian distance and thus of the objective. Minimizers still exist, but they may, in general, be non-unique. To circumvent these two issues, we restrict the sampling of data points to a ball of radius $\frac{\pi}{6}$ around the north pole. We restrict the objective $f$ to a strongly convex ball of diameter $\delta = \frac{\pi}{3}$ centered at the pole, and extend it to $+\infty$ outside. By [34, Proposition 4.2 (ii)], the radial distance is geodesically convex on this ball.

A comparison of the performance of the different algorithms are presented in Table 6.3. It is clear from Table 6.3 that the RCBM is fastest both in terms of number of iterations and CPU runtime for the case of dimension $n = 2$, and only in terms of CPU runtime for $n = 2^{15}$. The PBA is faster in the remaining cases. In any case, the RCBM and the PBA are largely comparable here, while the SGM was outperformed by both in this experiment. All algorithms provide essentially indistinguishable objective values.

### 6.2 DENOISING THROUGH TOTAL VARIATION

Total variation models on manifolds have been researched and implemented in signal and image denoising and restoration problems for manifold valued data [33, 46]. Let $\mathcal{I} = \{1, \dots, n\}$ for $n \geq 1$ be an index set, let $\mathcal{N}$ be a Hadamard manifold, and consider a manifold-valued signal $q\colon \mathcal{I} \to \mathcal{N}$. We can interpret $q = (q^{[1]}, \dots, q^{[n]})^\mathsf{T}$ to be a point on the power manifold $\mathcal{M} = \mathcal{N}^n$. Now let $\alpha > 0$ and define the objective $f_q\colon \mathcal{M} \to \mathbb{R}$ as the sum of a data fidelity term $g$ and a total variation term $TV$ as

$$f_q(p) = \frac{1}{n}\big(g(p, q) + \alpha\ \mathrm{TV}(p)\big), \tag{6.4}$$

| | RCBM | | | PBA | | | SGM | | |
|---|---|---|---|---|---|---|---|---|---|
| Dim. | Iter. | Time (s) | Obj. | Iter. | Time (s) | Obj. | Iter. | Time (s) | Obj. |
| 2 | 43 | $1.58 \cdot 10^{-2}$ | 0.258 90 | 71 | $1.84 \cdot 10^{-2}$ | 0.258 90 | 401 | $9.70 \cdot 10^{-2}$ | 0.258 90 |
| 4 | 74 | $2.30 \cdot 10^{-2}$ | 0.253 53 | 62 | $1.68 \cdot 10^{-2}$ | 0.253 53 | 5000 | 1.33 | 0.253 53 |
| 32 | 102 | $4.30 \cdot 10^{-2}$ | 0.259 89 | 64 | $2.73 \cdot 10^{-2}$ | 0.259 89 | 231 | $9.63 \cdot 10^{-2}$ | 0.259 89 |
| 1024 | 103 | $8.90 \cdot 10^{-1}$ | 0.266 99 | 68 | $6.23 \cdot 10^{-1}$ | 0.266 99 | 185 | 1.99 | 0.266 99 |
| 32 768 | 80 | $2.72 \cdot 10^{1}$ | 0.259 30 | 65 | $2.92 \cdot 10^{1}$ | 0.259 30 | 157 | $2.05 \cdot 10^{2}$ | 0.259 30 |

Table 6.3: Comparison between the RCBM, the PBA, and the SGM on $\mathcal{S}^n$ with varying dimensions for the Riemannian median example from Section 6.1.

where

$$g(p,q) = \frac{1}{2}\sum_{i=1}^{n} \operatorname{dist}_{\mathcal{N}}(p^{[i]}, q^{[i]})^2 \qquad \text{and} \qquad \operatorname{TV}(p) = \sum_{i=1}^{n-1} \operatorname{dist}_{\mathcal{N}}(p^{[i]}, p^{[i+1]}).$$

The non-smoothness of the model comes from the total variation part, where the distance appears with exponent one. For this, we use (6.1) to compute the subdifferential of the TV function

$$\partial f_q(p) = \frac{1}{n}\big(-\log_p q + \alpha\, \partial \operatorname{TV}(p)\big),$$

where

$$\partial \operatorname{TV}(p) = \sum_{i=1}^{n-1} \partial \operatorname{dist}_{p^{[i+1]}}(p^{[i]}).$$

As for the proximal maps of the objective function, we refer to [22]. For two signals $p, q \in \mathcal{M}$ we define the mean squared error as

$$E(p,q) \coloneqq \frac{\operatorname{dist}_{\mathcal{M}}(p,q)}{n} = \frac{1}{n}\sqrt{\sum_{i=1}^{n} \operatorname{dist}_{\mathcal{N}}^2(p^{[i]}, q^{[i]})}.$$

In this example[3], we consider the task of denoising a measured signal on the hyperbolic manifold $\mathcal{N} = \mathcal{H}^2$, a Hadamard manifold. To this end, let $T > 0$ be a real number, let $\mathcal{J} = [a,b] \subseteq \mathbb{R}$ be an interval and let $p(t) \in \mathbb{R}^2$ be (the graph of) a square wave defined by

$$p(t) = \begin{bmatrix} t \\ \operatorname{sgn}(\sin(\frac{2\pi}{T}t)) \end{bmatrix} \qquad \text{for all } t \in \mathcal{I}.$$

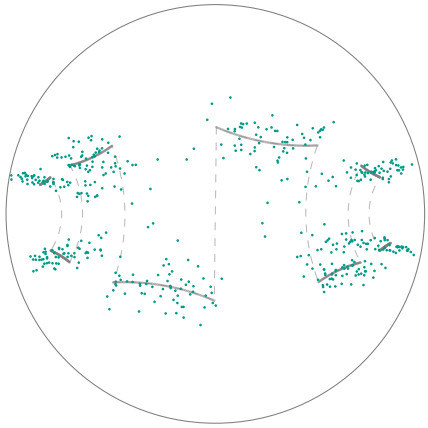
Let $N > 0$ be the number of points used for the discretization of the interval $\mathcal{I}$ and let $\tau = \frac{b-a}{N}$ be a step size, so that $\mathcal{J} = \bigcup_{i=1}^{N-1}[t_i, t_{i+1}]$ where $t_1 = a$, $t_N = b$ and $t_{i+1} = t_i + \tau$. We thus have a family of points $p^{[i]} = p(t_i)_{i=1}^{N}$ and we select

$$\mathcal{S} = \{p^{[1]}\} \cup \{p^{[i]}, p^{[i+1]} \,|\, p_2^{[i]} \neq p_2^{[i+1]} \text{ for } i = 2, \dots, N-1\} \cup \{p^{[N]}\},$$

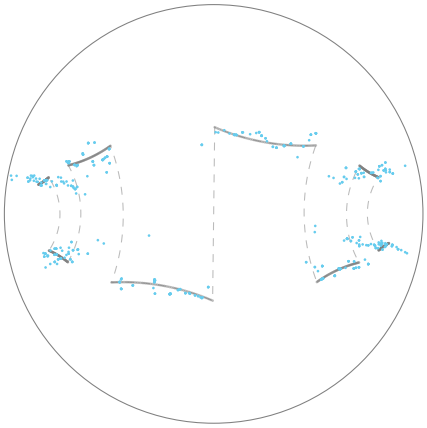
namely the first and last point, together with the points immediately next to the jumps. To construct points on $\mathcal{H}^2$ we use the isometry $\varphi\colon \mathbb{R}^2 \to \mathcal{H}^2$ given by

$$\varphi(p) = \begin{bmatrix} p_1 \\ p_2 \\ \sqrt{\|p\|^2 + 1} \end{bmatrix}.$$

[3] juliamanifolds.github.io/ManoptExamples.jl/stable/examples/H2-Signal-TV/



(a) Artificial signal $q \in (\mathcal{H}^2)^{496}$ in gray, and noisy data $\bar{q} \in (\mathcal{H}^2)^{496}$ in teal.



(b) Denoised reconstruction $q^* \in (\mathcal{H}^2)^{496}$ in cyan.

Figure 6.1: Denoising a signal on $(\mathcal{H}^2)^{496}$ using the RCBM, Algorithm 1. TV regularization example from Section 6.2.

We order the points from $\mathcal{S}$ by time $t$ into a sequence $(s^{[1]}, \dots, s^{[M]})$, where $s^{[j]} = \varphi(p^{[j]})$ for the corresponding $p^{[j]} \in \mathcal{S}$. Note that by construction, $M$ is even.

Finally, we sample $u = \left\lfloor \frac{NT}{2(b-a)} \right\rfloor$ points from each shortest geodesic connecting the points $s^{[2i-1]}$ and $s^{[2i]}$ for $i = 1, 2, \dots, \frac{M}{2}$. In our example, we set $a = -6$, $b = 6$, i. e., $\mathcal{J} = [-6, 6]$, $T = 3$, and we take $N = 496$ points, which yields a signal $q \in \mathcal{M}$ of length $n = \frac{uM}{2} = 496$. We can thus define the power manifold

$$\mathcal{M} = (\mathcal{H}^2)^{496}.$$

The manufactured signal is a point $q \in \mathcal{M}$ as shown in Figure 6.1a. We then generate a noisy signal $\bar{q} \in \mathcal{M}$ by adding Gaussian noise to the signal, i. e., we set $\bar{q}^{[i]} = \exp_{\bar{q}^{[i]}} X_{\bar{q}^{[i]}}$ for $i = 1, \dots, n$, where $X \in \mathcal{T}_q\mathcal{M}$ has a standard deviation of $\sigma = 0.1$ as in Figure 6.1a. This noisy signal is used in the data fidelity term $g(p, \bar{q})$, and it also serves as the starting point for the algorithms. The diameter $\delta$ is set to three times the maximal distance between each of the noisy points in $\mathcal{H}^2$. We fix the TV parameter to be $\alpha = 0.5$ for this experiment.

We notice that in this experiment all algorithms reached their iteration cap. The minimizer $q^* \in \mathcal{M}$ obtained from minimizing the objective function in (6.4) with the RCBM is shown in Figure 6.1b. The other algorithms yield results that are visually very close to the solution obtained by the RCBM.

Table 6.4 shows how the different algorithms compare to one another in terms of numbers of iterations, CPU runtime in seconds, final objective value, and mean squared error. Evidently, the RCBM is outperformed in terms of CPU runtime. Furthermore, the RCBM provides the second-lowest objective value, while the CPPA attains the lowest objective value. The SGM attains the lowest mean error $E(q^*, q)$, followed by the PBA and the RCBM.

### 6.3 SPECTRAL PROCRUSTES

We consider another example on a positively curved manifold and non-convex objective. Given two matrices $A, B \in \mathbb{R}^{n\times d}$ we aim to solve the spectral Procrustes problem (see, e. g., [14])

$$\operatorname*{arg\,min}_{p\in \mathrm{SO}(d)} f(p) \coloneqq \|A - B\,p\|_2, \tag{6.5}$$

where $\mathrm{SO}(d)$ is equipped with the standard bi-invariant metric, and $\|\cdot\|_2$ denotes the spectral norm of a matrix, i. e., its largest singular value. (6.5) aims to find the rotation matrix $p \in \mathbb{R}^{d\times d}$ which best

| Algorithm | Iter. | Time (sec.) | Objective | Error |
|---|---|---|---|---|
| RCBM | 5000 | 13.892 | $1.4023 \cdot 10^{-1}$ | $1.3692 \cdot 10^{-2}$ |
| PBA | 5000 | 9.519 | $1.4289 \cdot 10^{-1}$ | $1.3032 \cdot 10^{-2}$ |
| SGM | 5000 | 7.897 | $1.4622 \cdot 10^{-1}$ | $1.2460 \cdot 10^{-2}$ |
| CPPA | 5000 | 3.739 | $1.3191 \cdot 10^{-1}$ | $1.7361 \cdot 10^{-2}$ |

Table 6.4: Comparison of the four algorithms on $(\mathcal{H}^2)^{496}$ for a TV parameter of $\alpha = 0.5$. TV regularization example from Section 6.2.

aligns the columns of $B$ with those of $A$. Note that the spectral norm is convex in the Euclidean sense, but not geodesically convex on $\mathrm{SO}(d)$.

To obtain subdifferential information of $f$ in the Riemannian case, we project Euclidean convex subgradients of $f$ onto

$$\mathcal{T}_p\mathrm{SO}(d) = \left\{A \in \mathbb{R}^{d\times d} \,\middle|\, pA^{\mathsf{T}} + Ap^{\mathsf{T}} = 0,\ \operatorname{trace}(p^{-1}A) = 0\right\}.$$

We use

$$\operatorname{proj}_p(-B^{\mathsf{T}}UV^{\mathsf{T}}) \in \mathcal{T}_p\mathrm{SO}(d)$$

as a substitute for $\partial f(p)$, where $U$ and $V$ are left and right singular vectors, respectively, corresponding to the largest singular value of $A - Bp$.

The sectional curvature of the orthogonal group $\mathrm{O}(d)$ at $p \in \mathrm{O}(d)$ is given by

$$K_p(X_p, Y_p) = \frac{1}{4}\|[X_p, Y_p]\|_F^2$$

for any two orthonormal tangent vectors $X_p, Y_p \in \mathcal{T}_p\mathrm{O}(d)$, where

$$\mathcal{T}_p\mathrm{O}(d) = \left\{A \in \mathbb{R}^{d\times d} \,\middle|\, pA^{\top} + Ap^{\top} = 0\right\},$$

see, e. g., [16, Corollary 3.19]. Here $[X_p, Y_p] = X_p\,Y_p - Y_p\,X_p$ denotes the commutator of $X_p$ and $Y_p$. [24, Lemma 2.5] implies that the curvature of the orthogonal group $\mathrm{O}(d)$ is bounded below and above by $\omega = 0$ and $\Omega = \frac{1}{4}$ respectively. Hence, since $\mathrm{SO}(d)$ is the connected component of $\mathrm{O}(d)$ that contains the identity, we can use the same bounds on $\mathrm{SO}(d)$.

For this experiment[4], we set $n = 1000$, $d = 250$, and we initialize two random matrices $A, B \in \mathbb{R}^{n\times d}$. As a starting point for the algorithms, we employ the minimizer of the *orthogonal* Procrustes objective, $\|A - Bp\|_F$ , where the minimization is with respect to the Frobenius norm. The solution to this problem is given in closed form as

$$\bar{p} = UV^{\mathsf{T}},$$

see, e. g., [40]. We then set $p^{(0)} = \operatorname{proj}_{\mathrm{SO}(d)}(\bar{p})$ as the starting point for all algorithms. For RCBM, we set a diameter of $\delta = \frac{\pi}{3\sqrt{\Omega}}$. The results are summarized in Table 6.5 and Figure 6.2.

We see that in this experiment, the PBA outperforms the other two algorithms. The final objective values are all the same. From Figure 6.2, we can further evaluate the development of the cost during the iterations of all three algorithms. While the RCBM requires more time per iteration, its objective profile stays en par with the PBA, while the SGM is heavily affected by a zig-zagging behavior that leads to a higher number of iterations.

[4] juliamanifolds.github.io/ManoptExamples.jl/stable/examples/Spectral-Procrustes/

| Algorithm | Iter. | Time (sec.) | Objective |
|---|---|---|---|
| RCBM | 99 | 102.04 | 235.459 803 73 |
| PBA | 31 | 5.80 | 235.459 803 73 |
| SGM | 5000 | 402.74 | 235.459 803 73 |

Table 6.5: Comparison of the three algorithms on SO(250). Spectral Procrustes example from Section 6.3.

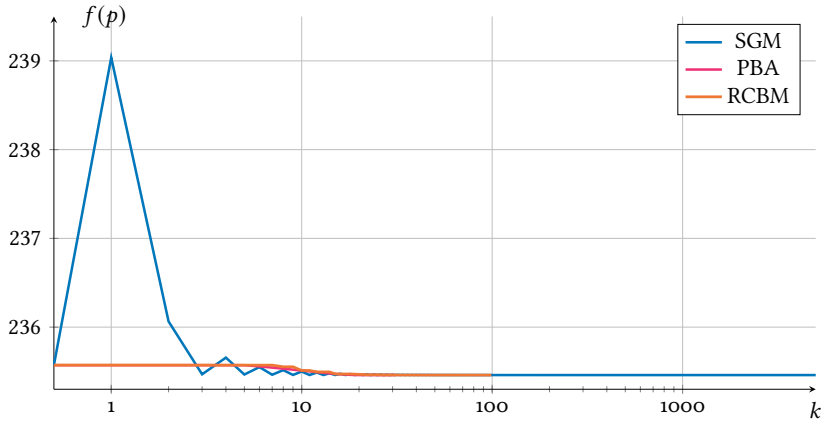



Figure 6.2: Objective versus iteration plot for the spectral Procrustes example from Section 6.3.

## 7 CONCLUSION

We introduced a new bundle method for solving smooth and non-smooth geodesically convex optimization problems on Riemannian manifolds with bounded curvature. By formulating a model based on the convex hull of previously collected subgradients, we generalized the dual formulation to minimize the objective function. We proved convergence of the method to a global minimizer under mild conditions. The performance of this new algorithm, termed the Riemannian Convex Bundle Method (RCBM), was studied and showcased through various numerical examples implemented using the Julia package `Manopt.jl`, by comparing it to other existing non-smooth optimization algorithms such as the proximal bundle algorithm (PBA), the subgradient method (SGM), and the cyclic proximal point algorithm (CPPA). For the Riemannian median, our RCBM outperforms the SGM and the PBA both in terms of the number of iterations as well as in runtime on Hadamard manifolds, and the same holds on the sphere for the experiments in dimensions 2 and $2^{15}$. Finally, the RCBM may yield meaningful results even when the objective function is not geodesically convex, as shown in the spectral Procrustes experiment. Further research directions include a better strategy to eliminate unnecessary subgradients from the bundle and investigating the necessary diameter parameter in more detail. Moreover, one further line of research is the investigation of non-asymptotic complexity bounds for the RCBM. We also aim to apply the algorithm to other structured, non-smooth optimization problems appearing in applications.

## APPENDIX A HESSIAN BOUND

We elaborate on how the argument presented in [2, Appendix C] was adapted to prove Theorem 3.2. Using the notation introduced in said theorem, let $\delta = \operatorname{diam}(\operatorname{dom} f) < +\infty$, $p \in \operatorname{dom} f$, $X_p \in \partial f(p)$, and $p^{(j)} \in \operatorname{dom} f$ and $X_{p^{(j)}} \in \partial f(p^{(j)})$ for all $j = 1, \dots, k$, $k \in \mathbb{N}$. Let $g\colon [0,1] \to \mathbb{R}$ be a function

defined by

$$g(t) = (X_{p^{(j)}}\,,\, \mathrm{P}_{p^{(j)}\leftarrow\gamma(t)} \log_{\gamma(t)} p), \tag{A.1}$$

where $\gamma(t)\colon [0,1] \to \mathcal{M}$ is the minimal geodesic arc connecting $\gamma(0) = p^{(j)}$ to $\gamma(1) = p^{(k)}$. The rest of the argument applies almost verbatim as follows. By applying the mean value theorem to $g(t)$, one can write

$$(X_{p^{(j)}}\,,\, \log_{p^{(j)}} p - \mathrm{P}_{p^{(j)}\leftarrow p^{(k)}} \log_{p^{(k)}} p) = (X_{p^{(j)}}\,,\, \mathcal{H}_{t_0} \log_{p^{(j)}} p^{(k)}) \tag{A.2}$$

for some $t_0 \in (0,1)$, where

$$\mathcal{H}_t \coloneqq -\mathrm{P}_{p^{(j)}\leftarrow\gamma(t)} \mathrm{Hess}_{\gamma(t)} \left(-\frac{1}{2}\operatorname{dist}^2(\gamma(t), p)\right) \mathrm{P}_{\gamma(t)\leftarrow p^{(j)}},$$

is a family of operators $\mathcal{H}_t\colon \mathcal{T}_{p^{(j)}}\mathcal{M} \to \mathcal{T}_{p^{(j)}}\mathcal{M}$ for $t \in [0,1]$, and $\mathrm{Hess}_{\gamma(t)}(f) : \mathcal{T}_{\gamma(t)}\mathcal{M} \to \mathcal{T}_{\gamma(t)}\mathcal{M}$ is the Riemannian Hessian of a $C^2$ function $f$ at the point $\gamma(t) \in \mathcal{M}$. From [2, Appendix D]

$$\|\mathcal{H}_t - \mathrm{id}\| \le \max\{\zeta_{1,\omega}(\operatorname{dist}(\gamma(t_0), p)) - 1, 1 - \zeta_{2,\Omega}(\operatorname{dist}(\gamma(t_0), p))\}, \tag{A.3}$$

where the functions $\zeta_{1,\omega}$ and $\zeta_{2,\Omega}$ are, respectively, upper and lower bounds to the largest and smallest eigenvalues of

$$-\mathrm{Hess}_{\gamma(t)} \left(-\frac{1}{2}\operatorname{dist}^2(\gamma(t), p)\right),$$

and they are defined in (3.5). This comes from $\mathrm{P}_{\gamma(t)\leftarrow p^{(j)}} = (\mathrm{P}_{p^{(j)}\leftarrow\gamma(t)})^{-1}$. Let now $r \in \mathbb{R}$ and consider $\varrho_{\omega,\Omega}(r)$ as defined in (3.4). For all $p \in \operatorname{dom} f$ we have

$$\begin{aligned}
&-(X_{p^{(j)}}\,,\, \log_{p^{(j)}} p - \mathrm{P}_{p^{(j)}\leftarrow p^{(k)}} \log_{p^{(k)}} p - \log_{p^{(j)}} p^{(k)}) \\
&\qquad = -(X_{p^{(j)}}\,,\, (\mathcal{H}_{t_0} - \mathrm{id})[\log_{p^{(j)}} p^{(k)}]) \\
&\qquad \le \|X_{p^{(j)}}\| \|\mathcal{H}_{t_0} - \mathrm{id}\| \|\log_{p^{(j)}} p^{(k)}\| \\
&\qquad \le \varrho_{\omega,\Omega}(\operatorname{dist}(\gamma(t_0), p)) \|X_{p^{(j)}}\| \|\log_{p^{(j)}} p^{(k)}\|,
\end{aligned}$$

where the first equality comes is due to (A.2), the first inequality comes from applying the Cauchy-Schwarz inequality, and the last inequality comes from (A.3). By rearranging the terms we get

$$\begin{aligned}
(X_{p^{(j)}}\,,\, \log_{p^{(j)}} p - \mathrm{P}_{p^{(j)}\leftarrow p^{(k)}} \log_{p^{(k)}} p) \ge (X_{p^{(j)}}\,,\, \log_{p^{(j)}} p^{(k)}) \\
- \varrho_{\omega,\Omega}(\operatorname{dist}(\gamma(t_0), p)) \|X_{p^{(j)}}\| \|\log_{p^{(j)}} p^{(k)}\|
\end{aligned} \tag{A.4}$$

for all $p \in \operatorname{dom} f$. Since $\varrho_{\omega,\Omega}(\operatorname{dist}(\gamma(t_0), p)) \le \varrho_{\omega,\Omega}(\delta)$ for all $p \in \operatorname{dom} f$, we get

$$\begin{aligned}
(X_{p^{(j)}}\,,\, \log_{p^{(j)}} p - \mathrm{P}_{p^{(j)}\leftarrow p^{(k)}} \log_{p^{(k)}} p) \ge (X_{p^{(j)}}\,,\, \log_{p^{(j)}} p^{(k)}) \\
- \varrho_{\omega,\Omega}(\delta) \|X_{p^{(j)}}\| \|\log_{p^{(j)}} p^{(k)}\|,
\end{aligned}$$

for all $p \in \operatorname{dom} f$. This is gives (3.9).

## REFERENCES


[1] P. A. Absil, R. Mahony, and R. Sepulchre, *Optimization Algorithms on Matrix Manifolds*, Princeton University Press, 2008, doi:10.1515/9781400830244.

[2] F. Alimisis, A. Orvieto, G. Becigneul, and A. Lucchi, Momentum improves optimization on Riemannian manifolds, in *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, A. Banerjee and K. Fukumizu (eds.), volume 130 of Proceedings of Machine Learning Research, PMLR, 2021, 1351–1359, https://proceedings.mlr.press/v130/alimisis21a.html.

[3] A. M. Bagirov, M. Gaudioso, N. Karmitsa, M. M. Mäkelä, and S. Taheri, *Numerical Nonsmooth Optimization: State of the Art Algorithms*, Springer International Publishing, 2020, doi:10.1007/978-3-030-34910-3.

[4] M. Bačák, Computing medians and means in Hadamard spaces, *SIAM Journal on Optimization* 24 (2014), 1542–1566, doi:10.1137/140953393.

[5] M. Bačák, R. Bergmann, G. Steidl, and A. Weinmann, A second order non-smooth variational model for restoring manifold-valued images, *SIAM Journal on Scientific Computing* 38 (2016), A567–A597, doi:10.1137/15m101988x.

[6] R. Bergmann, Manopt.jl: Optimization on manifolds in Julia, *Journal of Open Source Software* 7 (2022), 3866, doi:10.21105/joss.03866.

[7] R. Bergmann, J. H. Fitschen, J. Persch, and G. Steidl, Priors with coupled first and second order differences for manifold-valued image processing, *Journal of Mathematical Imaging and Vision* 60 (2018), 1459–1481, doi:10.1007/s10851-018-0840-y.

[8] R. Bergmann and P. Y. Gousenbourger, A variational model for data fitting on manifolds by minimizing the acceleration of a Bézier curve, *Frontiers in Applied Mathematics and Statistics* 4 (2018), doi:10.3389/fams.2018.00059, arXiv:1807.10090.

[9] R. Bergmann, R. Herzog, M. Silva Louzeiro, D. Tenbrinck, and J. Vidal-Núñez, Fenchel duality theory and a primal-dual algorithm on Riemannian manifolds, *Foundations of Computational Mathematics* 21 (2021), 1465–1504, doi:10.1007/s10208-020-09486-5, arXiv:1908.02022.

[10] R. Bergmann, J. Persch, and G. Steidl, A parallel Douglas Rachford algorithm for minimizing ROF-like functionals on images with values in symmetric Hadamard manifolds, *SIAM Journal on Imaging Sciences* 9 (2016), 901–937, doi:10.1137/15m1052858.

[11] F. Bonnans, C. Gilbert, C. Lemaréchal, and C. Sagastizábal, *Numerical Optimization*, Springer, Berlin, 2 edition, 2006, doi:10.1007/978-3-540-35447-5.

[12] N. Boumal, B. Mishra, P. A. Absil, and R. Sepulchre, Manopt, a Matlab toolbox for optimization on manifolds, *Journal of Machine Learning Research* 15 (2014), 1455–1459, https://www.jmlr.org/papers/v15/boumal14a.html.

[13] M. R. Bridson and A. Haefliger, *Metric Spaces of Non-Positive Curvature*, volume 319 of Grundlehren der Mathematischen Wissenschaften, Springer, Berlin, 1999, doi:10.1007/978-3-662-12494-9.

[14] J. Cape, Orthogonal Procrustes and norm-dependent optimality, *The Electronic Journal of Linear Algebra* 36 (2020), 158–168, doi:10.13001/ela.2020.5009.

[15] A. Chambolle and T. Pock, A first-order primal-dual algorithm for convex problems with applications to imaging, *Journal of Mathematical Imaging and Vision* 40 (2011), 120–145, doi:10.1007/s10851-010-0251-1.

[16] J. Cheeger and D. G. Ebin, *Comparison Theorems in Riemannian Geometry*, AMS Chelsea Publishing, 2008, doi:10.1090/chel/365. Revised reprint of the 1975 original.

[17] J. Chen and J. Revels, Robust benchmarking in noisy environments, in *Proceedings of the 20th Annual IEEE High Performance Extreme Computing Conference (HPEC 16)*, 2016, arXiv:1608.04295, https://ieee-hpec.org/2016/techprog2016/sept15.htm.

[18] C. Criscitiello and N. Boumal, An accelerated first-order method for non-convex optimization on manifolds, 2020, arXiv:2008.02252.

[19] M. P. a. do Carmo, *Riemannian Geometry*, Mathematics: Theory & Applications, Birkhäuser Boston, Inc., Boston, MA, 1992.

[20] J. Douglas and H. H. Rachford, On the numerical solution of heat conduction problems in two and three space variables, *Transactions of the American Mathematical Society* 82 (1956), 421–439, doi:10.2307/1993056.

[21] O. Ferreira and P. R. Oliveira, Subgradient algorithm on Riemannian manifolds, *Journal of Optimization Theory and Applications* 97 (1998), 93–104, doi:10.1023/a:1022675100677.

[22] O. Ferreira and P. R. Oliveira, Proximal point algorithm on Riemannian manifolds, *Optimization. A Journal of Mathematical Programming and Operations Research* 51 (2002), 257–270, doi:10.1080/02331930290019413.

[23] D. Gabay and B. Mercier, A dual algorithm for the solution of nonlinear variational problems via finite element approximations, *Computer and Mathematics with Applications* 2 (1976), 17–40, doi:10.1016/0898-1221(76)90003-1.

[24] J. Ge, DDVV-type inequality for skew-symmetric matrices and Simons-type inequality for Riemannian submersions, *Advances in Mathematics* 251 (2014), 62–86, doi:10.1016/j.aim.2013.10.010.

[25] C. Geiger and C. Kanzow, *Theorie und Numerik restringierter Optimierungsaufgaben*, Springer, New York, 2002, doi:10.1007/978-3-642-56004-0.

[26] P. Grohs and S. Hosseini, $\epsilon$-subgradient algorithms for locally Lipschitz functions on Riemannian manifolds, *Advances in Computational Mathematics* 42 (2015), 333–360, doi:10.1007/s10444-015-9426-z.

[27] N. Hoseini Monjezi, S. Nobakhtian, and M. R. Pouryayevali, A proximal bundle algorithm for nonsmooth optimization on Riemannian manifolds, *IMA Journal of Numerical Analysis* (2023), doi:10.1093/imanum/drab091.

[28] W. Huang, K. A. Gallivan, and P. A. Absil, A Broyden class of quasi-Newton methods for Riemannian optimization, *SIAM Journal on Optimization* 25 (2015), 1660–1685, doi:10.1137/140955483.

[29] W. Huang and K. Wei, Riemannian proximal gradient methods, *Mathematical Programming* 194 (2021), 371–413, doi:10.1007/s10107-021-01632-3.

[30] H. Kasai, H. Sato, and B. Mishra, Riemannian stochastic recursive gradient algorithm, in *Proceedings of the 35th International Conference on Machine Learning*, J. Dy and A. Krause (eds.), volume 80 of Proceedings of Machine Learning Research, PMLR, Stockholmsmässan, Stockholm, Sweden, 2018, 2516–2524, https://proceedings.mlr.press/v80/kasai18a.html.

[31] J. M. Lee, *Riemannian Manifolds. An Introduction to Curvature*, volume 176 of Graduate Texts in Mathematics, Springer, New York, 1997, doi:10.1007/b98852.

[32] J. M. Lee, *Introduction to Riemannian Manifolds*, volume 176 of Graduate Texts in Mathematics, Springer International Publishing, 2 edition, 2018, doi:10.1007/978-3-319-91755-9.

[33] J. Lellmann, E. Strekalovskiy, S. Koetter, and D. Cremers, Total variation regularization for functions with values in a manifold, in *2013 IEEE International Conference on Computer Vision*, 2013, 2944–2951, doi:10.1109/iccv.2013.366.

[34] C. Li and J. C. Yao, Variational inequalities for set-valued vector fields on Riemannian manifolds: convexity of the solution set and the proximal point algorithm, *SIAM Journal on Control and Optimization* 50 (2012), 2486–2514, doi:10.1137/110834962.

[35] M. M. Mäkelä, Survey of bundle methods for nonsmooth optimization, *Optimization Methods & Software* 17 (2002), 1–29, doi:10.1080/10556780290027828.

[36] N. Miolane, N. Guigui, A. Le Brigant, J. Mathe, B. Hou, Y. Thanwerdas, S. Heyder, O. Peltre, N. Koep, H. Zaatiti, H. Hajri, Y. Cabanes, T. Gerald, P. Chauchat, C. Shewmake, D. Brooks, B. Kainz, C. Donnat, S. Holmes, and X. Pennec, Geomstats: a Python package for Riemannian geometry in machine learning, *Journal of Machine Learning Research* 21 (2020), 1–9, arXiv:1805.08308, https://jmlr.org/papers/v21/19-027.html.

[37] D. Noll, Bundle method for non-convex minimization with inexact subgradients and function values, in *Computational and Analytical Mathematics*, volume 50 of Springer Proceedings in Mathematics & Statistics, Springer, New York, 2013, 555–592, doi:10.1007/978-1-4614-7621-4_26.

[38] D. Orban and G. Leconte, RipQP.jl. Regularized interior point solver for quadratic problems, 2020, doi:10.5281/zenodo.4309783.

[39] D. Orban, A. S. Siqueira, and Contributors, QuadraticModels.jl. Data structures for linear and quadratic optimization problems based on NLPModels.jl, 2019, doi:10.5281/zenodo.4027577.

[40] P. H. Schönemann, A generalized solution of the orthogonal Procrustes problem, *Psychometrika* 31 (1966), 1–10, doi:10.1007/bf02289451.

[41] C. Tang, S. Xing, W. Huang, and J. Jian, A restricted memory quasi-Newton bundle method for nonsmooth optimization on Riemannian manifolds, *Computational Optimization and Applications* (2026), doi:10.1007/s10589-026-00793-1, arXiv:2402.18308.

[42] J. Townsend, N. Koep, and S. Weichwald, Pymanopt: A Python toolbox for optimization on manifolds using automatic differentiation, *Journal of Machine Learning Research* 17 (2016), 1–5, arXiv:1603.03236, https://jmlr.org/papers/v17/16-177.html.

[43] C. Udrişte, *Convex Functions and Optimization Methods on Riemannian Manifolds*, volume 297 of Mathematics and its Applications, Kluwer Academic Publishers Group, Dordrecht, 1994, doi:10.1007/978-94-015-8390-9.

[44] R. Walter, On the metric projection onto convex sets in Riemannian spaces, *Archiv der Mathematik* 25 (1974), 91–98, doi:10.1007/bf01238646.

[45] X. Wang, C. Li, and J. C. Yao, On some basic results related to affine functions on Riemannian manifolds, *Journal of Optimization Theory and Applications* 170 (2016), 783–803, doi:10.1007/s10957-016-0979-x.

[46] A. Weinmann, L. Demaret, and M. Storath, Total variation regularization for manifold-valued data, *SIAM Journal on Imaging Sciences* 7 (2014), 2226–2257, doi:10.1137/130951075.

[47] H. Zhang and S. Sra, First-order methods for geodesically convex optimization, in *Proceedings of the 29th Annual Conference on Learning Theory*, V. Feldman, A. Rakhlin, and O. Shamir (eds.), volume 49 of Proceedings of Machine Learning Research, 2016, 1617–1638, arXiv:1602.06053, https://proceedings.mlr.press/v49/zhang16b.html.

[48] R. Zimmermann, Hermite interpolation and data processing errors on Riemannian matrix manifolds, *SIAM Journal on Scientific Computing* 42 (2020), A2593–A2619, doi:10.1137/19m1282878.