Noname manuscript No.
(will be inserted by the editor)

Adaptive regularization with cubics on manifolds

Naman Agarwal · Nicolas Boumal · Brian Bullins · Coralia Cartis

Received: date / Accepted: date

Abstract Adaptive regularization with cubics (ARC) is an algorithm for unconstrained, non-convex optimization. Akin to the trust-region method, its iterations can be thought of as approximate, safe-guarded Newton steps. For cost functions with Lipschitz continuous Hessian, ARC has optimal iteration complexity, in the sense that it produces an iterate with gradient smaller than ε in $O(1/\varepsilon^{1.5})$ iterations. For the same price, it can also guarantee a Hessian with smallest eigenvalue larger than $-\sqrt{\varepsilon}$. In this paper, we study a generalization of ARC to optimization on Riemannian manifolds. In particular, we generalize the iteration complexity results to this richer framework. Our central contribution lies in the identification of appropriate manifold-specific assumptions that allow us to secure these complexity guarantees both when using the exponential map and when using a general retraction. A substantial part of the paper is devoted to studying these assumptions—relevant beyond ARC—and providing user-friendly sufficient conditions for them. Numerical experiments are encouraging.

Authors are listed alphabetically. NB was partially supported by NSF award DMS-1719558. CC acknowledges support from NERC through grant NE/L012146/1. NA and BB were supported by Elad Hazan's NSF grant IIS-1523815.

N. Agarwal
Google AI, Princeton, NJ
E-mail: namanagarwal@google.com

N. Boumal
Department of Mathematics, Princeton University, NJ
E-mail: nboumal@math.princeton.edu

B. Bullins
Toyota Technological Institute at Chicago, IL
E-mail: bbullins@ttic.edu

C. Cartis
Mathematical Institute, University of Oxford, UK
E-mail: coralia.cartis@maths.ox.ac.uk

Keywords Optimization on manifolds · Complexity · Lipschitz regularity · Cubic regularization · Newton's method

1 Introduction

Adaptive regularization with cubics (ARC) is an iterative algorithm used to solve unconstrained optimization problems of the form

$$\min_{x \in \mathbb{R}^n} \; f(x),$$

where $f\colon \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable (Griewank, 1981). Given any initial iterate $x_0 \in \mathbb{R}^n$, assuming f is lower-bounded and has a Lipschitz continuous Hessian, ARC produces an iterate x_k with small gradient, namely, $\|\nabla f(x_k)\| \leq \varepsilon$, in at most $O(1/\varepsilon^{1.5})$ iterations (Nesterov and Polyak, 2006; Cartis et al., 2011a; Birgin et al., 2017). This improves upon the worst-case iteration complexity of steepest descent and classical trust-region methods. In fact, this iteration complexity is optimal under those assumptions (Carmon et al., 2019), contributing to renewed interest in this method.

In this paper, we study a generalization of ARC to optimization on manifolds, that is,

$$\min_{x \in \mathcal{M}} \; f(x), \tag{P}$$

where $\mathcal{M}$ is a given Riemannian manifold and $f\colon \mathcal{M} \to \mathbb{R}$ is a (sufficiently smooth) cost function. The practical interest in optimization on manifolds stems from its ubiquity: it comes up naturally in numerical linear algebra (spectral decompositions, low-rank Lyapunov equations), signal and image processing (shape analysis, diffusion tensor imaging, community detection on graphs, rotational video stabilization), statistics and machine learning (matrix/tensor completion, metric learning, Gaussian mixtures, activity recognition, independent component analysis), robotics and computer vision (simultaneous localization and mapping, structure from motion, pose estimation) and various other fields. The theoretical interest comes from the fact that Riemannian geometry is arguably to "right" setting for *unconstrained* optimization—indeed, it is the minimal mathematical structure required to have comfortable notions of gradients and Hessians, which are the basic building blocks of smooth, unconstrained optimization algorithms. See for example (Absil et al., 2008) and (Boumal, 2020) for book-length introductions to this topic. See *related work* below for further references.

Building upon the existing literature for the Euclidean case, we generalize the worst-case iteration complexity analysis of ARC to manifolds, obtaining essentially the same guarantees but with a wider application range: see numerical experiments in Section 9 for some examples.

In particular, with the appropriate assumptions discussed in Sections 3 and 4, we find that ε-critical points of f on $\mathcal{M}$ can be computed in $O(1/\varepsilon^{1.5})$ iterations. We also show an iteration complexity bound for the computation of

approximate second-order critical points in Section 5. Key differences with the Euclidean setting lie in the particular assumptions we make. We further study these assumptions in Sections 6 and 7. A subproblem solver—necessary to run ARC—is detailed in Section 8. Our algorithm is implemented in the Manopt framework (Boumal et al., 2014) and distributed as part of that toolbox. In Section 9, we close with numerical comparisons to existing solvers, in particular the related Riemannian trust-region method (RTR) (Absil et al., 2007).

Main results

An important ingredient of ARC on manifolds (Algorithm 1) is the *retraction* R, which allows one to move around the manifold by following tangent vectors. This notion is defined in Section 2. Our results depend on the choice of retraction.

For a twice continuously differentiable cost function $f : \mathcal{M} \to \mathbb{R}$, the first- and second-order necessary optimality conditions at x read (Yang et al., 2014):

$$\|\mathrm{grad} f(x)\|_x = 0, \qquad \lambda_{\min}(\mathrm{Hess} f(x)) \geq 0,$$

where $\mathrm{grad} f$ and $\mathrm{Hess} f$ are the Riemannian gradient and Hessian of f—see Section 3 for definitions; $\|\cdot\|_x$ is the Riemannian norm at x and $\lambda_{\min}$ extracts the smallest eigenvalue of a symmetric operator.

Our first main result applies to *complete* Riemannian manifolds, for which we can use the so-called *exponential map* as retraction R. The statement below summarizes more explicit results of Sections 3 and 5, stating iterates of ARC eventually satisfy the necessary optimality conditions up to some tolerance, with a bound on the number of iterations this may require.

Theorem 1 *Consider a cost function f on a complete Riemannian manifold $\mathcal{M}$. If*

a) f is lower bounded (A1), and
b) the Riemannian Hessian of f is Lipschitz continuous (A2, A3),

then, for any $x_0 \in \mathcal{M}$ and $\varepsilon > 0$, Algorithm 1 with the exponential retraction produces an iterate $x_k \in \mathcal{M}$ such that $f(x_k) \leq f(x_0)$, $\|\mathrm{grad} f(x_k)\|_{x_k} \leq \varepsilon$ and (if condition (3) is enforced) $\lambda_{\min}(\mathrm{Hess} f(x_k)) \geq -\sqrt{\varepsilon}$, with $k = \tilde{O}(1/\varepsilon^{1.5})$. The bound is dimension- and curvature-free.

Our second main result is an extension of the above which allows us to use other retractions, the main motivation being that the exponential map may be unavailable or expensive to compute. We state it as a summary of results in Sections 4 and 5.

Theorem 2 *Consider a cost function f on a Riemannian manifold $\mathcal{M}$ equipped with a retraction* R*. If*

a) f is lower bounded (A1),

b) the pullbacks $f \circ \mathrm{R}_x$ satisfy a type of second-order Lipschitz condition (A2, A4), and
c) the differential of the retraction is well behaved (A5),

then, for any $x_0 \in \mathcal{M}$ and sufficiently small $\varepsilon > 0$, Algorithm 1 with retraction R *produces an iterate $x_k \in \mathcal{M}$ such that $f(x_k) \leq f(x_0)$, $\|\mathrm{grad} f(x_k)\|_{x_k} \leq \varepsilon$ and (if condition (3) is enforced and* R *is second order) $\lambda_{\min}(\mathrm{Hess} f(x_k)) \geq -\sqrt{\varepsilon}$, with $k = \tilde{O}(1/\varepsilon^{1.5})$.*

We further provide sufficient conditions for the assumptions on the pullbacks and the retraction to be satisfied, in Sections 6 and 7 respectively. For example, A5 is satisfied if the sublevel set of x_0 is compact.

Related work

Numerous algorithms for unconstrained optimization have been generalized to Riemannian manifolds (Luenberger, 1972; Gabay, 1982; Smith, 1994; Edelman et al., 1998; Absil et al., 2008), among them gradient descent, nonlinear conjugate gradients, stochastic gradients (Bonnabel, 2013; Zhang et al., 2016), BFGS (Ring and Wirth, 2012), Newton's method (Adler et al., 2002) and trust-regions (Absil et al., 2007). See these references and also our numerical experiments in Section 9 for a discussion of numerous applications.

ARC in particular was extended to manifolds in the PhD thesis of Qi (2011). There, under a different set of regularity assumptions, asymptotic convergence analyses are proposed, in the same spirit as the analyses presented in the aforementioned references for other methods. Qi also presents local convergence analyses, showing superlinear local convergence under some assumptions.

In contrast, we here favor a global convergence analysis with explicit bounds on iteration complexity to reach approximate criticality. Such bounds are standard in optimization on Euclidean spaces. Around the same time, they have been generalized to Riemannian gradient descent and other algorithms by Zhang and Sra (2016) (focusing on geodesic convexity), by Bento et al. (2017) (looking also at proximal point methods), and by Boumal et al. (2018) (also analyzing RTR). In the first two works, the regularity assumptions on the cost function are close in spirit to those we lay out in Section 3, whereas in the third work the assumptions are closer to our Section 4.

Closest to our work, Zhang and Zhang (2018) recently proposed a convergence analysis of a cubically regularized method on manifolds, also establishing an $O(1/\varepsilon^{1.5})$ iteration complexity. Their analysis (independent from ours: early versions of our results appeared on public repositories around the same time, theirs two weeks before ours) focuses on compact submanifolds of a Euclidean space and uses a fixed regularization parameter (which must be set properly by the user). Subproblems are assumed to be solved to global optimality, though it appears this could be relaxed within their framework. We improve on these points as follows: our analysis is intrinsic (no embedding space is ever referenced), we do not need $\mathcal{M}$ to be compact, our regularization parameter ς_k is

dynamically adapted (which is both easier for the user and more efficient), and the subproblem solver only needs to meet weak requirements to reach approximate criticality. These improvements lead to implementable, competitive algorithms. Zhang and Zhang (2018) also study superlinear local convergence rates, in line with Qi (2011) but with different assumptions.

The work by Zhang and Zhang (2018) is also related to adaptive *quadratic* regularization on embedded submanifolds of Euclidean space recently studied by Hu et al. (2018), where the quadratic model is written in terms of the Euclidean gradient and Hessian.

More recently, two independent papers generalize work by Jin et al. (2019) to provide iteration complexity bounds for a Riemannian version of perturbed gradient descent, allowing to reach approximate second-order criticality without looking at the Hessian, and with logarithmic dependence in the dimension of the manifold. Sun et al. (2019) provide an analysis based on regularity assumptions akin to the ones we lay out in Section 3, while Criscitiello and Boumal (2019) base their analysis on regularity assumptions closer to the ones we lay out in Section 4.

Our complexity analysis builds on prior work for the Euclidean case by Cartis et al. (2011a) and Birgin et al. (2017). Complexity lower bounds given by Cartis et al. (2018) and Carmon et al. (2019) show that the bounds in (Nesterov and Polyak, 2006; Cartis et al., 2011a; Birgin et al., 2017) are optimal in ε-dependency for the appropriate class of functions. A variant of ARC that is closely related to trust-region methods was presented in (Dussault, 2018).

Recently, various works have focused on efficiently solving the ARC subproblem (that is, minimizing m_k as defined in (1)) in the Euclidean setting. Agarwal et al. (2017) propose an efficient method to solve the subproblem leading to fast algorithms for converging to second-order local minima in the Euclidean setting. Carmon and Duchi (2019) and Tripuraneni et al. (2018b) propose gradient descent–based methods to solve the subproblem. Several recent papers consider the effect of subsampling on the subproblem (Tripuraneni et al., 2018b; Kohler and Lucchi, 2017; Zhou et al., 2018; Zhang et al., 2018; Wang et al., 2019).

In the Riemannian case, the subproblem is posed on a tangent space, which is a linear subspace. Hence, all of the above methods are applicable in the Riemannian setting as well. In particular, we use the Krylov subspace method originally proposed in (Cartis et al., 2011a). Recently, Carmon and Duchi (2018) and Gould and Simoncini (2019) provided a bound on the amount of work this method may require to provide sufficient progress (see also Remark 1).

On a technical note, in Definition 4 we formulate second-order assumptions on the retraction to disentangle the requirements on f from those on the retraction. These are related to (but differ from) the assumptions and discussions in (Ring and Wirth, 2012), specifically Lemma 6, Propositions 5 and 7, and Remarks 2 and 3 in that reference.

Algorithm 1 Riemannian adaptive regularization with cubics (ARC)

Parameters: $\theta > 0$, $\varsigma_{\min} > 0$, $0 < \eta_1 \leq \eta_2 < 1$, $0 < \gamma_1 < 1 < \gamma_2 < \gamma_3$
Input: $x_0 \in \mathcal{M}$, $\varsigma_0 \geq \varsigma_{\min}$
for $k = 0, 1, 2 \ldots$ **do**
Consider the pullback $\hat{f}_k = f \circ \mathrm{R}_{x_k} : \mathrm{T}_{x_k}\mathcal{M} \to \mathbb{R}$. Define the model m_k on $\mathrm{T}_{x_k}\mathcal{M}$:

$$m_k(s) = \hat{f}_k(0) + \langle s, \nabla \hat{f}_k(0) \rangle + \frac{1}{2}\langle s, \nabla^2 \hat{f}_k(0)[s] \rangle + \frac{\varsigma_k}{3}\|s\|^3. \tag{1}$$

Compute a step $s_k \in \mathrm{T}_{x_k}\mathcal{M}$ satisfying first-order progress conditions (see Section 8):

$$m_k(s_k) \leq m_k(0), \qquad \text{and} \qquad \|\nabla m_k(s_k)\| \leq \theta \|s_k\|^2. \tag{2}$$

Optionally, if second-order criticality is targeted, s_k must also satisfy this condition:

$$\lambda_{\min}(\nabla^2 m_k(s_k)) \geq -\theta \|s_k\|, \tag{3}$$

where $\lambda_{\min}$ extracts the smallest eigenvalue of a symmetric operator.
If $s_k = 0$, terminate (see Lemma 1).
Compute the regularized ratio of actual improvement over model improvement:

$$\rho_k = \frac{f(x_k) - f(\mathrm{R}_{x_k}(s_k))}{m_k(0) - m_k(s_k) + \frac{\varsigma_k}{3}\|s_k\|^3}. \tag{4}$$

If $\rho_k \geq \eta_1$, accept the step: $x_{k+1} = \mathrm{R}_{x_k}(s_k)$. Otherwise, reject it: $x_{k+1} = x_k$.
Update the regularization parameter:

$$\varsigma_{k+1} \in \begin{cases} [\max(\varsigma_{\min}, \gamma_1 \varsigma_k), \varsigma_k] & \text{if } \rho_k \geq \eta_2 \quad \text{(very successful)}, \\ [\varsigma_k, \gamma_2 \varsigma_k] & \text{if } \rho_k \in [\eta_1, \eta_2) \quad \text{(successful)}, \\ [\gamma_2 \varsigma_k, \gamma_3 \varsigma_k] & \text{if } \rho_k < \eta_1 \quad \text{(unsuccessful)}. \end{cases} \tag{5}$$

end for

2 ARC on manifolds

ARC on manifolds is listed as Algorithm 1. It is a direct adaptation from (Cartis et al., 2011a; Birgin et al., 2017). Like many other optimization algorithms, its generalization to manifolds relies on a chosen retraction (Shub, 1986; Absil et al., 2008). For some $x \in \mathcal{M}$, let $\mathrm{T}_x\mathcal{M}$ denote the tangent space at x: this is a linear space. Intuitively, a retraction R on a manifold provides a means to move away from x along a tangent direction $s \in \mathrm{T}_x\mathcal{M}$ while remaining on the manifold, producing $\mathrm{R}_x(s) \in \mathcal{M}$. For a formal definition, we use the *tangent bundle*,

$$\mathrm{T}\mathcal{M} = \{(x, s) : x \in \mathcal{M} \text{ and } s \in \mathrm{T}_x\mathcal{M}\},$$

which is itself a smooth manifold.

Definition 1 (Retraction (Absil et al., 2008, Def. 4.1.1)) A *retraction* on a manifold $\mathcal{M}$ is a smooth mapping R from the tangent bundle $\mathrm{T}\mathcal{M}$ to $\mathcal{M}$ with the following properties. Let $\mathrm{R}_x : \mathrm{T}_x\mathcal{M} \to \mathcal{M}$ denote the restriction of R to $\mathrm{T}_x\mathcal{M}$ through $\mathrm{R}_x(s) = \mathrm{R}(x, s)$. Then,

(i) $\mathrm{R}_x(0) = x$, where 0 is the zero vector in $\mathrm{T}_x\mathcal{M}$; and

(ii) The differential of R_x at 0, $\mathrm{DR}_x(0)$, is the identity map on $\mathrm{T}_x\mathcal{M}$.

In other words: retraction curves $c(t) = \mathrm{R}_x(ts)$ are smooth and pass through $c(0) = x$ with velocity $c'(0) = \mathrm{DR}_x(0)[s] = s$. For the special case where $\mathcal{M}$ is a linear space, the canonical retraction is $\mathrm{R}_x(s) = x + s$. For the unit sphere, a typical retraction is $\mathrm{R}_x(s) = \frac{x+s}{\|x+s\|}$.

Importantly, the retraction R chosen to optimize over a particular manifold $\mathcal{M}$ is part of the algorithm specification. For a given cost function f and a specified retraction R, at iterate x_k, we define the *pullback* of the cost function to the tangent space $\mathrm{T}_{x_k}\mathcal{M}$:

$$\hat{f}_k = f \circ \mathrm{R}_{x_k} : \mathrm{T}_{x_k}\mathcal{M} \to \mathbb{R}. \tag{6}$$

This operation lifts f to a linear space. We then define a model $m_k : \mathrm{T}_{x_k}\mathcal{M} \to \mathbb{R}$, obtained as a truncated second-order Taylor expansion of the pullback with cubic regularization: see (1). We use the notation $\langle \cdot, \cdot \rangle_x$ to denote the Riemannian metric on $\mathrm{T}_x\mathcal{M}$, and we usually simplify this notation to $\langle \cdot, \cdot \rangle$ when the base point is clear from context. Likewise, $\|s\|_x = \sqrt{\langle s, s \rangle_x}$ is the norm of $s \in \mathrm{T}_x\mathcal{M}$ induced by the Riemannian metric, and we usually omit the subscript, writing $\|s\|$. Furthermore, for real functions on linear spaces (such as $\hat{f}_k$ and m_k), we let ∇ and ∇^2 denote the (usual) gradient and Hessian operators.

At iteration k, a *subproblem solver* is used to approximately minimize the model m_k, producing a *trial step* s_k: specific requirements are listed as (2) and (3); for the first-order condition, we follow the lead of Birgin et al. (2017). Section 8 discusses a practical algorithm.

The quality of the trial step s_k is evaluated by computing ρ_k (4): the *regularized* ratio of actual to anticipated cost improvement, also following Birgin et al. (2017). Note that the denominator of ρ_k is equal to the difference between $\hat{f}_k(0)$ and the second-order Taylor expansion of $\hat{f}_k$ around 0 evaluated at s_k. If $\rho_k \geq \eta_1$, we accept the trial step and set $x_{k+1} = \mathrm{R}_{x_k}(s_k)$: such steps are called *successful*. Among them, we further identify *very successful* steps, for which $\rho_k \geq \eta_2$; for those, not only is the step accepted, but the regularization parameter ς_k is (usually) decreased. Otherwise, we reject the step and set $x_{k+1} = x_k$: these steps are *unsuccessful*, and we necessarily increase ς_k.

We expect Algorithm 1 to produce an infinite sequence of iterates. In practice of course, one would terminate the algorithm as soon as approximate criticality is achieved within some prescribed tolerance. This paper bounds the number of iterations this may require. In the unlikely event that the subproblem solver produces the trivial step $s_k = 0$, the algorithm cannot proceed. Fortunately, this only happens if we reached exact criticality of appropriate order. Proofs are in Appendix A.

Lemma 1 *If second-order progress* (3) *is not enforced, then the first-order condition* (2) *allows the subproblem solver to return* $s_k = 0$ *if and only if* $\mathrm{grad} f(x_k) = 0$. *If both* (2) *and* (3) *are enforced, the subproblem solver is allowed to return* $s_k = 0$ *if and only if* $\mathrm{grad} f(x_k) = 0$ *and* $\mathrm{Hess} f(x_k)$ *is positive semidefinite.*

We now introduce two basic assumptions about the cost function f, affording us two supporting lemmas. The first common assumption is that the cost function f is lower bounded.

A1. *There exists a finite f_{low} such that $f(x) \geq f_{\text{low}}$ for all $x \in \mathcal{M}$.*

The second assumption is that f is sufficiently differentiable so that the models m_k are well defined, and that second-order Taylor expansions of $\hat{f}_k$ in the tangent space at x_k are sufficiently accurate. In the Euclidean case, the latter follows from a Lipschitz condition on the Hessian of f. In the next two sections, we discuss how this generalizes to manifolds.

A2. *The cost function f is twice continuously differentiable. Furthermore, there exists a constant L such that, at each iteration k, for the trial step s_k selected by the subproblem solver, the pullback $\hat{f}_k = f \circ \mathrm{R}_{x_k}$ satisfies*

$$\hat{f}_k(s_k) - \left[\hat{f}_k(0) + \langle s_k, \nabla \hat{f}_k(0)\rangle + \frac{1}{2}\langle s_k, \nabla^2 \hat{f}_k(0)[s_k]\rangle\right] \leq \frac{L}{6}\|s_k\|^3. \tag{7}$$

The two supporting lemmas below follow the standard Euclidean analysis. The first lemma establishes that the regularization parameter ς_k does not grow unbounded.

Lemma 2 (Birgin et al. (2017, Lem. 2.2)) *Under A2, the regularization parameter remains bounded: for all k, it holds that $\varsigma_k \leq \varsigma_{\max}$, with*

$$\varsigma_{\max} = \max\left(\varsigma_0, \frac{L\gamma_3}{2(1-\eta_2)}\right). \tag{8}$$

Conditioned on the conclusions of this lemma, the next lemma states that among the first $\bar{k}$ iterations of ARC, a certain number are sure to be successful.

Lemma 3 (Cartis et al. (2011a, Thm. 2.1)) *If $\varsigma_k \leq \varsigma_{\max}$ for all k (as provided by Lemma 2), then the number K of successful iterations among $0, \ldots, \bar{k}-1$ satisfies*

$$\bar{k} \leq \left(1 + \frac{|\log(\gamma_1)|}{\log(\gamma_2)}\right) K + \frac{1}{\log(\gamma_2)} \log\left(\frac{\varsigma_{\max}}{\varsigma_0}\right).$$

In other words, in order to bound the *total* number of iterations ARC may require to attain a certain goal, it is sufficient to bound the number of *successful* iterations that goal may require. The following proposition (extracted from the main proof in (Birgin et al., 2017)) further states that this can be done by showing successful steps are not too short.

Proposition 1 *Let $\{(x_0, s_0), (x_1, s_1), \ldots\}$ be the set of iterates and trial steps generated by Algorithm 1. If A1 holds, we have*

$$\sum_{k \in \mathcal{S}} \|s_k\|^3 \leq \frac{3(f(x_0) - f_{\text{low}})}{\eta_1 \varsigma_{\min}},$$

where $\mathcal{S}$ is the set of successful iterations.

Proof. By definition, if iteration k is successful, then $\rho_k \geq \eta_1$ (4). Combining with the first part of the first-order progress condition (2) yields

$$f(x_k) - f(x_{k+1}) \geq \eta_1 \left(m_k(0) - m_k(s_k) + \frac{\varsigma_k}{3}\|s_k\|^3 \right) \geq \frac{\eta_1 \varsigma_{\min}}{3}\|s_k\|^3.$$

On the other hand, for unsuccessful iterations, $x_{k+1} = x_k$ and the cost does not change. Using A1, a telescoping sum yields:

$$f(x_0) - f_{\text{low}} \geq \sum_{k=0}^{\infty} f(x_k) - f(x_{k+1}) = \sum_{k \in \mathcal{S}} f(x_k) - f(x_{k+1}) \geq \frac{\eta_1 \varsigma_{\min}}{3} \sum_{k \in \mathcal{S}} \|s_k\|^3,$$

as announced. □

3 First-order analysis with the exponential map

In this section, we provide a first-order analysis of Algorithm 1 for the case where $\mathcal{M}$ is a complete manifold and we use the exponential retraction $\mathrm{R} = \mathrm{Exp}$—we define these terms momentarily. This notably encompasses the Euclidean case where $\mathcal{M} = \mathbb{R}^n$, with $\mathrm{Exp}_x(s) = x + s$, as well as all compact or Hadamard manifolds. As such, the results in this section offer a strict generalization of the Euclidean analysis proposed in (Birgin et al., 2017) under the assumption of Lipschitz continuous Hessian. An in-depth reference for the Riemannian geometry tools we use is the monograph by Lee (2018), while Absil et al. (2008) offer an optimization-focused treatment.

On a complete Riemannian manifold $\mathcal{M}$, for any point x and tangent vector $v \in \mathrm{T}_x\mathcal{M}$, there exists a unique smooth curve $\gamma_v \colon \mathbb{R} \to \mathcal{M}$ such that $\gamma_v(0) = x$, $\gamma_v'(0) = v$ and, for $t < t'$ close enough, $\gamma_v|_{[t,t']}$ is the shortest path connecting $\gamma_v(t)$ to $\gamma_v(t')$. This curve is called a *geodesic*. The *exponential map* is built from these geodesics as the map

$$\mathrm{Exp} \colon \mathrm{T}\mathcal{M} \to \mathcal{M} \colon (x, v) \mapsto \mathrm{Exp}_x(v) = \gamma_v(1).$$

This is a smooth map. Because $\mathrm{Exp}_x(tv) = \gamma_{tv}(1) = \gamma_v(t)$, we also find that $\mathrm{Exp}_x(0) = x$ and $\mathrm{DExp}_x(0)[v] = v$, so that the exponential map is indeed a retraction (Definition 1). If the manifold is not complete, then Exp is only defined on an open subset of $\mathrm{T}\mathcal{M}$: when we need $\mathcal{M}$ to be complete, we say so explicitly.

The Riemannian gradient of $f \colon \mathcal{M} \to \mathbb{R}$, denoted by $\mathrm{grad} f$, is the vector field on $\mathcal{M}$ such that $\mathrm{D}f(x)[s] = \langle \mathrm{grad} f(x), s \rangle$, where $\mathrm{D}f(x)[s]$ is the directional derivative of f at x along the tangent direction s. One can show that

$$\mathrm{grad} f(x) = \nabla(f \circ \mathrm{Exp}_x)(0) = \nabla \hat{f}_x(0), \tag{9}$$

so that the Riemannian gradient of f at x is nothing but the Euclidean gradient of the pullback $\hat{f}_x = f \circ \mathrm{Exp}_x$ at the origin of the tangent space $\mathrm{T}_x\mathcal{M}$ (see also Lemma 5 for a similar statement with retractions).

The Riemannian Hessian of f is the covariant derivative of the gradient vector field, with respect to the Riemannian connection. Denoted by $\mathrm{Hess} f$, it defines a tensor field as follows: $\mathrm{Hess} f(x)$ is a linear operator from $\mathrm{T}_x\mathcal{M}$ into itself, self-adjoint with respect to the Riemannian metric on that tangent space. Analogously to (9), one can show that

$$\mathrm{Hess} f(x) = \nabla^2 \hat{f}_x(0), \tag{10}$$

which expresses the Riemannian Hessian of f at x as the Euclidean Hessian of the pullback $\hat{f}_x$ at the origin of $\mathrm{T}_x\mathcal{M}$. (Here too, see Lemma 5 below.)

These two statements show that the model m_k (1) can be written equivalently as

$$m_k(s) = f(x_k) + \langle \mathrm{grad} f(x_k), s\rangle + \frac{1}{2}\langle \mathrm{Hess} f(x_k)[s], s\rangle + \frac{\varsigma_k}{3}\|s\|^3 \tag{11}$$

and that A2 requires

$$f(\mathrm{Exp}_{x_k}(s_k)) - \left[f(x_k) + \langle s_k, \mathrm{grad} f(x_k)\rangle + \frac{1}{2}\langle s_k, \mathrm{Hess} f(x_k)[s_k]\rangle\right] \leq \frac{L}{6}\|s_k\|^3 \tag{12}$$

for each (x_k, s_k) produced by Algorithm 1.

In particular, if $\mathcal{M}$ is a Euclidean space with the exponential map $\mathrm{Exp}_x(s) = x + s$, it is well known that we can secure (12) if we assume that the Hessian of f is L-Lipschitz continuous. This can be written as

$$\left\|\nabla^2 f(x) - \nabla^2 f(y)\right\|_{\mathrm{op}} \leq L\|x - y\|,$$

where the norm on the left-hand side is the operator norm. Generalizing this to the Riemannian setting, we face the issue that $\mathrm{Hess} f(x)$ and $\mathrm{Hess} f(y)$ are linear operators defined on distinct tangent spaces (if $x \neq y$): in order to compare them, we need one more tool to compare tangent vectors in distinct tangent spaces.

Given a smooth curve $c\colon [0,1] \to \mathcal{M}$ connecting $c(0) = x$ to $c(1) = y$, consider a tangent vector $v \in \mathrm{T}_x\mathcal{M}$ and a smooth vector field $Z\colon [0,1] \to \mathrm{T}\mathcal{M}$ along c—that is, $Z(t) \in \mathrm{T}_{c(t)}\mathcal{M}$—such that $Z(0) = v$. If the covariant derivative of Z with respect to the Riemannian connection vanishes identically, then we say that Z is a *parallel vector field* along c. (For example, the velocity vector field γ' of a geodesic γ is parallel.) This vector field exists and is unique. We call $Z(1)$ the *parallel transport* of v from x to y along c. Parallel transports are linear isometries with respect to the Riemannian metric, and they depend on the chosen path.

Using parallel transports, we can formulate a standard notion of Lipschitz continuity for Riemannian Hessians. We use the following notation often: given a tangent vector $s \in \mathrm{T}_x\mathcal{M}$,

$$P_s\colon \mathrm{T}_x\mathcal{M} \to \mathrm{T}_{\mathrm{Exp}_x(s)}\mathcal{M} \tag{13}$$

denotes parallel transport along the geodesic $\gamma(t) = \mathrm{Exp}_x(ts)$ from $t = 0$ to $t = 1$.

Definition 2 A function $f: \mathcal{M} \to \mathbb{R}$ on a Riemannian manifold $\mathcal{M}$ has an *L-Lipschitz continuous Hessian* if it is twice differentiable and if, for all (x, s) in the domain of Exp,

$$\left\| P_s^{-1} \circ \mathrm{Hess} f(\mathrm{Exp}_x(s)) \circ P_s - \mathrm{Hess} f(x) \right\|_{\mathrm{op}} \leq L\|s\|.$$

For f three times continuously differentiable, this property holds if and only if the covariant derivative of the Riemannian Hessian is uniformly bounded by L (we omit a proof). In particular, this holds with some L for any smooth function on a compact manifold. In the Euclidean case, parallel transports are identity maps (independent of the transport curve), so that this is equivalent to the usual definition.

Crucially, for cost functions with Lipschitz Hessian, we recover familiar-looking bounds on Taylor expansions of both f itself and, as will be instrumental momentarily, of $\mathrm{grad} f$. Results of this nature are standard: they appear frequently in complexity analyses for Riemannian optimization, see for example (Ferreira and Svaiter, 2002; Bento et al., 2017; Sun et al., 2019). Proofs are in Appendix B.

Proposition 2 *Let $f: \mathcal{M} \to \mathbb{R}$ be twice differentiable on a Riemannian manifold $\mathcal{M}$. Given (x, s) in the domain of* Exp*, assume there exists $L \geq 0$ such that, for all $t \in [0, 1]$,*

$$\left\| P_{ts}^{-1} \circ \mathrm{Hess} f(\mathrm{Exp}_x(ts)) \circ P_{ts} - \mathrm{Hess} f(x) \right\|_{\mathrm{op}} \leq L\|ts\|.$$

Then, the two following inequalities hold:

$$\left| f(\mathrm{Exp}_x(s)) - f(x) - \langle s, \mathrm{grad} f(x) \rangle - \frac{1}{2} \langle s, \mathrm{Hess} f(x)[s] \rangle \right| \leq \frac{L}{6}\|s\|^3, \text{ and}$$

$$\left\| P_s^{-1} \mathrm{grad} f(\mathrm{Exp}_x(s)) - \mathrm{grad} f(x) - \mathrm{Hess} f(x)[s] \right\| \leq \frac{L}{2}\|s\|^2.$$

This justifies the introduction of the following assumption, still with notation (13) for P_{s_k}.

A3. *There exists a constant L' such that, at each successful iteration k, for the step s_k selected by the subproblem solver, we have*

$$\left\| P_{s_k}^{-1} \mathrm{grad} f(\mathrm{Exp}_{x_k}(s_k)) - \mathrm{grad} f(x_k) - \mathrm{Hess} f(x_k)[s_k] \right\| \leq \frac{L'}{2}\|s_k\|^2. \tag{14}$$

Corollary 1 *If $f: \mathcal{M} \to \mathbb{R}$ has L-Lipschitz continuous Hessian, then, for any sequence $\{(x_k, s_k)\}_{k=0,1,2,\ldots}$ in the domain of* Exp*, f satisfies A2 with the same L (and $\mathrm{R} = \mathrm{Exp}$), and it satisfies A3 with $L' = L$. In particular, if f is smooth and $\mathcal{M}$ is compact, assumptions A1, A2 and A3 hold.*

We can now state our main result regarding the complexity of running Algorithm 1 on a complete manifold with the exponential retraction, for the purpose of computing approximate first-order critical points. We require $\mathcal{M}$ to be complete so that Exp is indeed a retraction, defined on the whole tangent

bundle. The proof follows (Birgin et al., 2017, Thm. 2.5) up to the fact that we bound the *total number* of successful iterations that map to points with large gradient (as opposed to bounding the number of such iterations among the first $\bar{k}$), more in the spirit of (Cartis et al., 2012): this enables us to make a statement about the limit of $\|\mathrm{grad} f(x_k)\|$. Recall that $\varsigma_{\max}$ is provided by Lemma 2.

Theorem 3 *Let $\mathcal{M}$ be a complete Riemannian manifold and let* $\mathrm{R} = \mathrm{Exp}$. *Under A1, A2 and A3, for an arbitrary $x_0 \in \mathcal{M}$, let $x_0, x_1, x_2 \ldots$ be the iterates produced by Algorithm 1. For any $\varepsilon > 0$, the* total number *of* successful *iterations k such that $\|\mathrm{grad} f(x_{k+1})\| > \varepsilon$ is bounded above by*

$$K_1(\varepsilon) \triangleq \frac{3(f(x_0) - f_{\mathrm{low}})}{\eta_1 \varsigma_{\min}} \left(\frac{L'}{2} + \theta + \varsigma_{\max} \right)^{1.5} \frac{1}{\varepsilon^{1.5}}.$$

Furthermore, $\lim_{k \to \infty} \|\mathrm{grad} f(x_k)\| = 0$ *(that is, limit points are critical).*

Proof. If iteration k is successful, we have $x_{k+1} = \mathrm{Exp}_{x_k}(s_k)$. The gradient of the model m_k (11) at s_k (in the tangent space at x_k) is given by

$$\begin{aligned} \nabla m_k(s_k) &= \mathrm{grad} f(x_k) + \mathrm{Hess} f(x_k)[s_k] + \varsigma_k \|s_k\| s_k \\ &= P_{s_k}^{-1} \mathrm{grad} f(x_{k+1}) \\ &\quad + \big(\mathrm{grad} f(x_k) + \mathrm{Hess} f(x_k)[s_k] - P_{s_k}^{-1} \mathrm{grad} f(x_{k+1}) \big) + \varsigma_k \|s_k\| s_k, \end{aligned}$$

with P_{s_k} as defined by (13). Owing to the first-order progress condition (2), by the triangle inequality and also using A3, we find

$$\theta \|s_k\|^2 \geq \|\nabla m_k(s_k)\| \geq \|P_{s_k}^{-1} \mathrm{grad} f(x_{k+1})\| - \frac{L'}{2} \|s_k\|^2 - \varsigma_k \|s_k\|^2.$$

Rearranging and using that P_{s_k} is an isometry, we get for all successful k that

$$\|\mathrm{grad} f(x_{k+1})\| = \|P_{s_k}^{-1} \mathrm{grad} f(x_{k+1})\| \leq \left(\frac{L'}{2} + \theta + \varsigma_{\max} \right) \|s_k\|^2, \tag{15}$$

where we also called upon Lemma 2 to claim $\varsigma_k \leq \varsigma_{\max}$. Define a subset of the successful steps based on the tolerance ε:

$$\mathcal{S}_\varepsilon = \{k : \rho_k \geq \eta_1 \text{ and } \|\mathrm{grad} f(x_{k+1})\| > \varepsilon\}.$$

For $k \in \mathcal{S}_\varepsilon$, we can lower-bound $\|s_k\|^3$ using (15) since $\|\mathrm{grad} f(x_{k+1})\| > \varepsilon$. Then, calling upon Proposition 1, we find

$$\frac{3(f(x_0) - f_{\mathrm{low}})}{\eta_1 \varsigma_{\min}} \geq \sum_{k \in \mathcal{S}_\varepsilon} \|s_k\|^3 \geq \frac{\varepsilon^{1.5}}{\left(\frac{L'}{2} + \theta + \varsigma_{\max} \right)^{1.5}} |\mathcal{S}_\varepsilon|.$$

This proves the main claim. The claim regarding limit points is proved in Appendix B. □

(Above, it is natural to consider the sequence $\{x_{k+1}\}$ for successful iterations k, as this enumerates each distinct point in the whole sequence once.) A key consequence of Theorem 3 is that, if the number of *successful* iterations among $0, \ldots, \bar{k}-1$ strictly exceeds $K_1(\varepsilon)$, then it must be that $\|\mathrm{grad} f(x_k)\| \leq \varepsilon$ for some k in $0, \ldots, \bar{k}$. Combining this with Lemma 3 yields the first main result: a bound on the total number of iterations it may take ARC to produce an approximate critical point on a complete manifold, using the exponential map, and (essentially) assuming a Lipschitz continuous Hessian.

Corollary 2 *Under the assumptions of Theorem 3, Algorithm 1 produces a point $x_k \in \mathcal{M}$ such that $f(x_k) \leq f(x_0)$ and $\|\mathrm{grad} f(x_k)\| \leq \varepsilon$ in at most*

$$\left(1 + \frac{|\log(\gamma_1)|}{\log(\gamma_2)}\right) K_1(\varepsilon) + \frac{1}{\log(\gamma_2)} \log\left(\frac{\varsigma_{\max}}{\varsigma_0}\right) + 1$$

iterations.

In the Euclidean case, this recovers the result of (Birgin et al., 2017) exactly. Note also that this complexity result is unaffected by the curvature of the manifold. Moreover, if L is known and $L' = L$ (which holds under the Lipschitz Hessian assumption), then we can set $\varsigma_0 = \varsigma_{\min} = \frac{1}{2}L$ (so that $\varsigma_{\max} = \frac{\gamma_3}{2(1-\eta_2)}L$) and $\theta = \frac{1}{2}L$. With those choices, we find that

$$K_1(\varepsilon) = \frac{6}{\eta_1}\left(1 + \frac{\gamma_3}{2(1-\eta_2)}\right)^{1.5} (f(x_0) - f_{\mathrm{low}})\sqrt{L}\frac{1}{\varepsilon^{1.5}}. \tag{16}$$

This exhibits a complexity scaling with $\sqrt{L}$ when L is known, as in (Nesterov and Polyak, 2006) and in the lower bound discussed in (Carmon et al., 2019).

4 First-order analysis with a general retraction

The results of the previous section provide a strict, lossless generalization of a known result in the Euclidean case. However, we note two practical shortcomings:

- For the analysis to apply, the algorithm must compute the exponential map.
- The Lipschitz condition (Definition 2) may be difficult to assess as it involves parallel transports or bounding the covariant derivative of the Riemannian Hessian.

Regarding the first point, we organized proofs in Appendix B to highlight why it is not clear how to generalize Proposition 2 to general retractions. In a nutshell, it is because parallel transports and geodesics interact particularly nicely through the fact that the velocity vector field of a geodesic is a parallel vector field.

To address both points, we propose alternate regularity conditions which (a) allow for any retraction, and (b) involve conceptually simpler objects. We

do this by focusing on the pullbacks $\hat{f}_x = f \circ \mathrm{R}_x$, which have the merit of being scalar functions on linear spaces—this is in the spirit of prior work (Boumal et al., 2018). We offer justification for these assumptions below, and in section 6.

The first regularity assumption, A2, is readily phrased in terms of the pullback. We focus on providing a replacement for the second condition: A3. Translating this condition to pullbacks by analogy, we aim to bound the difference between $\nabla \hat{f}_x(s)$—which, conveniently, is a vector tangent at x—and a classical truncated Taylor expansion for it: $\nabla \hat{f}_x(0) + \nabla^2 \hat{f}_x(0)[s]$. In so doing, it is useful to note that $\nabla \hat{f}_x(s)$ is related to $\mathrm{grad} f(\mathrm{R}_x(s))$ by a linear operator, as follows:

$$\nabla \hat{f}_x(s) = T_s^* \mathrm{grad} f(\mathrm{R}_x(s)), \quad \text{with} \quad T_s = \mathrm{DR}_x(s) \colon \mathrm{T}_x\mathcal{M} \to \mathrm{T}_{\mathrm{R}_x(s)}\mathcal{M}, \tag{17}$$

where the star indicates the adjoint with respect to the Riemannian metric. Indeed,

$$\begin{aligned} \forall s, \dot{s} \in \mathrm{T}_x\mathcal{M}, \qquad \langle \nabla \hat{f}_x(s), \dot{s} \rangle_x &= \mathrm{D}\hat{f}_x(s)[\dot{s}] \\ &= \mathrm{D}f(\mathrm{R}_x(s))[\mathrm{DR}_x(s)[\dot{s}]] \\ &= \langle \mathrm{grad} f(\mathrm{R}_x(s)), \mathrm{DR}_x(s)[\dot{s}] \rangle_{\mathrm{R}_x(s)} \\ &= \left\langle (\mathrm{DR}_x(s))^* \left[\mathrm{grad} f(\mathrm{R}_x(s))\right], \dot{s} \right\rangle_x . \end{aligned} \tag{18}$$

(This also plays a role in (Ring and Wirth, 2012, p599).)

Considering for a moment how an estimate for $\nabla \hat{f}_x(s)$ might look like if we use the exponential retraction and under the Lipschitz continuous Hessian assumption A3 as above, we find by triangular inequality that

$$\begin{aligned} \|\nabla \hat{f}_x(s) - \nabla \hat{f}_x(0) - \nabla^2 \hat{f}_x(0)[s]\| &\leq \|\nabla \hat{f}_x(s) - P_s^{-1} \mathrm{grad} f(\mathrm{Exp}_x(s))\| \\ &\quad + \|P_s^{-1} \mathrm{grad} f(\mathrm{Exp}_x(s)) - \mathrm{grad} f(x) - \mathrm{Hess} f(x)[s]\| \\ &\leq \left\|T_s^* - P_s^{-1}\right\|_{\mathrm{op}} \|\mathrm{grad} f(\mathrm{Exp}_x(s))\| + \frac{L'}{2}\|s\|^2, \end{aligned}$$

using (18) with $T_s = \mathrm{DExp}_x(s)$. Since parallel transport P_s (13) is an isometry, $P_s^{-1} = P_s^*$ and $\left\|T_s^* - P_s^{-1}\right\|_{\mathrm{op}} = \|T_s - P_s\|_{\mathrm{op}}$. For small s, we expect T_s (the differential of the exponential map) and P_s (parallel transport) to be nearly the same. Indeed, $\|T_s - P_s\|_{\mathrm{op}}$ is a continuous function of s and $T_0 = P_0 = \mathrm{Id}$. How much they differ for nonzero s is related to the curvature of the manifold. As a result, we conclude that

$$\left\|\nabla \hat{f}_x(s) - \nabla \hat{f}_x(0) - \nabla^2 \hat{f}_x(0)[s]\right\| \leq \frac{L'}{2}\|s\|^2 + q(\|s\|) \cdot \|\mathrm{grad} f(\mathrm{Exp}_x(s))\| \tag{19}$$

for some continuous function $q \colon \mathbb{R}^+ \to \mathbb{R}^+$ such that $q(0) = 0$.

As an illustration, consider the special case where $\mathcal{M}$ has constant sectional curvature C. In this case, it can be shown using Jacobi fields (see the proof of Lemma 10 in the appendix) that

$$T_s\dot{s} = \mathrm{DExp}_x(s)[\dot{s}] = P_s\dot{s} + h(\|s\|)P_s\left(\dot{s} - \frac{\langle s, \dot{s}\rangle}{\|s\|^2}s\right), \tag{20}$$

where

$$h(\|s\|) = \begin{cases} 0 & \text{if } C = 0, \\ \frac{\sin(\|s\|/R)}{\|s\|/R} - 1 & \text{if } C = \frac{1}{R^2} > 0, \\ \frac{\sinh(\|s\|/R)}{\|s\|/R} - 1 & \text{if } C = -\frac{1}{R^2} < 0. \end{cases}$$

Thus, for manifolds with constant sectional curvature, inequality (19) holds with $q(\|s\|) = |h(\|s\|)|$, independent of x. This function behaves as $\frac{1}{6}\left(\frac{\|s\|}{R}\right)^2 = \frac{C}{6}\|s\|^2$ for small $\|s\|$. This derivation generalizes to manifolds with sectional curvature bounded both from above and from below (Tripuraneni et al., 2018a), (Waldmann, 2012, Thm. A.2.9).

Returning to retractions in general, the above motivates us to introduce the following assumption on the pullbacks, meant to replace A3.

A4. *There exists a constant L' such that, at each successful iteration k, for the step s_k selected by the subproblem solver, the pullback $\hat{f}_k = f \circ \mathrm{R}_{x_k}$ obeys*

$$\left\|\nabla\hat{f}_k(s_k) - \nabla\hat{f}_k(0) - \nabla^2\hat{f}_k(0)[s_k]\right\| \leq \frac{L'}{2}\|s_k\|^2 + q(\|s_k\|)\|\mathrm{grad} f(\mathrm{R}_{x_k}(s_k))\|, \tag{21}$$

where $q\colon \mathbb{R}^+ \to \mathbb{R}^+$ is some continuous function satisfying $q(0) = 0$.

Notice how this assumption involves simple tools compared to A3, which relies on the exponential map and parallel transports. Furthermore, if we strengthen the condition by forcing $q \equiv 0$, we get a Lipschitz-type condition on the pullback: see Section 6.

Looking at the proof of Theorem 3, specifically equation (15), we anticipate the need to lower-bound the norm of $\nabla\hat{f}_k(s_k)$. Owing to (17), it holds that

$$\|\nabla\hat{f}_k(s)\| \geq \sigma_{\min}(\mathrm{DR}_{x_k}(s))\,\|\mathrm{grad} f(\mathrm{R}_{x_k}(s))\|, \tag{22}$$

where $\sigma_{\min}$ extracts the smallest singular value of an operator. For our purpose, it is important that this least singular value remains bounded away from zero. This is only a concern for small steps (as large successful steps provide sufficient improvement for other reasons.) Providentially, for $s = 0$, Definition 1 ensures $\sigma_{\min}(\mathrm{DR}_{x_k}(0)) = 1$, so that by continuity we expect that it should be possible to meet this requirement. We summarize this discussion in the following assumption.

A5. *There exist constants $a > 0$ and $b > 0$ such that, at each successful iteration k,*

$$\text{if } \|s_k\| \leq a, \text{ then } \sigma_{\min}(\mathrm{DR}_{x_k}(s_k)) \geq b. \tag{23}$$

(The constant a is allowed to be $+\infty$, while b is necessarily at most 1.)

In the Euclidean case with $\mathrm{R}_x(s) = x + s$, $\mathrm{DR}_x(s)$ is an isometry and one can set $a = +\infty$ and $b = 1$. We secure A5 in Section 7 for a large family of manifolds and retractions.

With these new assumptions, we can adapt Theorem 3 to general retractions. The main change in the proof consists in treating short and long steps separately. This induces a condition that ε must be small enough for the rate $O(1/\varepsilon^{1.5})$ to materialize. We stress that it is not necessary to know L, L', q, a and b as they appear in A2, A4 and A5 to run Algorithm 1 in practice: they are only used for the analysis. Recall that $\varsigma_{\max}$ is provided by Lemma 2.

Theorem 4 *Let $\mathcal{M}$ be a Riemannian manifold equipped with a retraction* R*. Under A1, A2, A4 and A5, for an arbitrary $x_0 \in \mathcal{M}$, let $x_0, x_1, x_2 \ldots$ be the iterates produced by Algorithm 1. For any $\varepsilon > 0$, the* total number *of* successful *iterations k such that $\|\mathrm{grad} f(x_{k+1})\| > \varepsilon$ is bounded above by*

$$K_1(\varepsilon) \triangleq \frac{3(f(x_0) - f_{\text{low}})}{\eta_1 \varsigma_{\min}} \max\left(\left(\frac{\frac{L'}{2} + \theta + \varsigma_{\max}}{b - q(r)}\right)^{1.5} \frac{1}{\varepsilon^{1.5}}, \frac{1}{r^3}\right)$$

for any $r \in (0, a]$ such that $q(r) < b$. Furthermore, $\lim_{k\to\infty} \|\mathrm{grad} f(x_k)\| = 0$.

Proof. If iteration k is successful, we have $x_{k+1} = \mathrm{R}_{x_k}(s_k)$. The gradient of the model m_k (1) at s_k is given by

$$\begin{aligned}\nabla m_k(s_k) &= \nabla \hat{f}_k(0) + \nabla^2 \hat{f}_k(0)[s_k] + \varsigma_k \|s_k\| s_k \\ &= \nabla \hat{f}_k(s_k) + \left(\nabla \hat{f}_k(0) + \nabla^2 \hat{f}_k(0)[s_k] - \nabla \hat{f}_k(s_k)\right) + \varsigma_k \|s_k\| s_k.\end{aligned}$$

Owing to the first-order progress condition (2), using A4 and (22) with $T_{s_k} = \mathrm{DR}_{x_k}(s_k)$,

$$\begin{aligned}\theta \|s_k\|^2 \geq \|\nabla m_k(s_k)\| &\geq \sigma_{\min}(T_{s_k}) \|\mathrm{grad} f(x_{k+1})\| \\ &\quad - \frac{L'}{2}\|s_k\|^2 - q(\|s_k\|) \cdot \|\mathrm{grad} f(x_{k+1})\| - \varsigma_k \|s_k\|^2.\end{aligned}$$

Rearranging and calling upon Lemma 2, we get for all successful iterations k that

$$\left(\sigma_{\min}(T_{s_k}) - q(\|s_k\|)\right) \cdot \|\mathrm{grad} f(x_{k+1})\| \leq \left(\frac{L'}{2} + \theta + \varsigma_{\max}\right) \|s_k\|^2. \tag{24}$$

If k is a successful step and $\|s_k\| \leq a$, then A5 guarantees $\sigma_{\min}(T_{s_k}) \geq b > 0$. Additionally, since q is continuous and satisfies $q(0) = 0$ there necessarily

exists $r \in (0, a]$ such that $q(r) < b$. This motivates the following. Recall this subset of the successful steps:

$$\mathcal{S}_\varepsilon = \{k : \rho_k \geq \eta_1 \text{ and } \|\mathrm{grad} f(x_{k+1})\| > \varepsilon\}.$$

Further partition this subset in two, based on step length: *short* steps in $\mathcal{S}_{\text{short}}$ and *long* steps in $\mathcal{S}_{\text{long}}$. The partition is based on r as constructed above:

$$\mathcal{S}_{\text{short}} = \{k \in \mathcal{S}_\varepsilon : \|s_k\| \leq r\}, \qquad \text{and} \qquad \mathcal{S}_{\text{long}} = \mathcal{S}_\varepsilon \backslash \mathcal{S}_{\text{short}}.$$

For $k \in \mathcal{S}_{\text{short}}$, we can lower-bound $\|s_k\|^3$ using (24) since $\sigma_{\min}(T_{s_k}) - q(\|s_k\|) \geq b - q(r) > 0$. For $k \in \mathcal{S}_{\text{long}}$, we have $\|s_k\|^3 > r^3$ by definition. Then, calling upon Proposition 1, we find

$$\begin{aligned}
\frac{3(f(x_0) - f_{\text{low}})}{\eta_1 \varsigma_{\min}} &\geq \sum_{k \in \mathcal{S}_{\text{short}}} \|s_k\|^3 + \sum_{k \in \mathcal{S}_{\text{long}}} \|s_k\|^3 \\
&\geq \frac{(b - q(r))^{1.5} \varepsilon^{1.5}}{\left(\frac{L'}{2} + \theta + \varsigma_{\max}\right)^{1.5}} |\mathcal{S}_{\text{short}}| + r^3 |\mathcal{S}_{\text{long}}| \\
&\geq \min\left(\left(\frac{(b - q(r))\varepsilon}{\frac{L'}{2} + \theta + \varsigma_{\max}}\right)^{1.5}, r^3\right) |\mathcal{S}_\varepsilon|.
\end{aligned}$$

This proves the main claim. For limit points, see the matching argument in Theorem 3. □

A corollary identical to Corollary 2 holds for Theorem 4 as well. In the Euclidean case with $\mathrm{R}_x(s) = x + s$ and a Lipschitz continuous Hessian, we can set $a = +\infty$, $b = 1$, $q \equiv 0$ and $r = +\infty$, thus also recovering the result of (Birgin et al., 2017) exactly.

5 Second-order analysis

The two previous sections show how to meet first-order necessary optimality conditions approximately. To further satisfy second-order necessary optimality conditions approximately, we also require second-order progress in the subproblem solver, through condition (3).

This condition is similar to one proposed by Cartis et al. (2017) for the same purpose in the Euclidean case. A direct extension of the proof in that reference would involve the Hessian of the pullback at the trial step s_k rather than at the origin. As we have seen for gradients, this leads to technical difficulties. We provide a proof that achieves the same complexity bound while avoiding such issues. As a result, there is no need to distinguish between the exponential and the general retraction cases for second-order analysis.

We have the following bound on the total number of successful iterations which can produce points where the Hessian is far from positive semidefinite, akin to Theorems 3 and 4.

Theorem 5 *Under A1 and A2, for an arbitrary $x_0 \in \mathcal{M}$, let $x_0, x_1, x_2 \ldots$ be the iterates produced by Algorithm 1 with second-order progress (3) enforced. For any $\varepsilon > 0$, the* total number *of* successful *iterations k such that $\lambda_{\min}(\nabla^2 \hat{f}_k(0)) < -\varepsilon$ is bounded above by*

$$K_2(\varepsilon) \triangleq \frac{3(f(x_0) - f_{\text{low}})}{\eta_1 \varsigma_{\min}} (\theta + 2\varsigma_{\max})^3 \frac{1}{\varepsilon^3}.$$

Furthermore, $\liminf_{k\to\infty} \lambda_{\min}(\nabla^2 \hat{f}_k(0)) \geq 0$.

Proof. The second-order condition (3) implies a lower-bound on step-sizes related to the minimal eigenvalue of the Hessian of $\hat{f}_k = f \circ \mathrm{R}_{x_k}$. Indeed, by definition of the model m_k (1),

$$\forall s, \dot{s} \in \mathrm{T}_{x_k}\mathcal{M}, \qquad \nabla^2 m_k(s)[\dot{s}] = \nabla^2 \hat{f}_k(0)[\dot{s}] + \varsigma_k \left(\|s\| \dot{s} + \frac{\langle s, \dot{s} \rangle}{\|s\|} s \right).$$

It follows that

$$\begin{aligned}
\lambda_{\min}(\nabla^2 \hat{f}_k(0)) &= \min_{\|\dot{s}\|=1} \left\langle \dot{s}, \nabla^2 \hat{f}_k(0)[\dot{s}] \right\rangle \\
&= \min_{\|\dot{s}\|=1} \left\langle \dot{s}, \nabla^2 m_k(s)[\dot{s}] \right\rangle - \varsigma_k \left(\|s\| \|\dot{s}\|^2 + \frac{\langle s, \dot{s} \rangle^2}{\|s\|} \right) \\
&\geq \lambda_{\min}(\nabla^2 m_k(s)) - 2\varsigma_k \|s\|.
\end{aligned}$$

In particular, with $s = s_k$, the second-order progress condition (3) and Lemma 2 yield

$$-\lambda_{\min}(\nabla^2 \hat{f}_k(0)) \leq (\theta + 2\varsigma_{\max}) \|s_k\|. \tag{25}$$

Consider this particular subset of the successful iterations:

$$\mathcal{S}_\varepsilon = \{k : \rho_k \geq \eta_1 \text{ and } \lambda_{\min}(\nabla^2 \hat{f}_k(0)) < -\varepsilon\}.$$

Using Proposition 1 with (25) on this set leads to:

$$\frac{3(f(x_0) - f_{\text{low}})}{\eta_1 \varsigma_{\min}} \geq |\mathcal{S}_\varepsilon| \left(\frac{\varepsilon}{\theta + 2\varsigma_{\max}} \right)^3,$$

which is the desired bound on the number of steps in $\mathcal{S}_\varepsilon$. The limit inferior result follows from an argument similar to that at the end of the proof of Theorem 3. □

Here too, if L is known we can set $\varsigma_0 = \varsigma_{\min} = \frac{1}{2}L$ (so that $\varsigma_{\max} = \frac{\gamma_3}{2(1-\eta_2)}L$) and $\theta = \frac{1}{2}L$. With those choices, we find that for a target depending on L we have:

$$K_2(\sqrt{L\varepsilon}) = \frac{6}{\eta_1} \left(\frac{1}{2} + \frac{\gamma_3}{(1-\eta_2)} \right)^3 (f(x_0) - f_{\text{low}}) \sqrt{L} \frac{1}{\varepsilon^{1.5}}. \tag{26}$$

This exhibits a complexity scaling with $\sqrt{L}$ when L is known.

Theorem 5 is a statement about the Hessian of the pullbacks, $\nabla^2 \hat{f}_k(0)$, whereas we would more naturally desire a statement about the Hessian of the cost function itself, $\mathrm{Hess} f(x_k)$. For the exponential retraction, these two objects are the same (10). More generally, they are the same for any *second-order retraction* (of which the exponential map is one example) (Absil et al., 2008, §5): we defer their (standard) definition to Section 6. For now, accepting the claim that for second-order retractions we have $\mathrm{Hess} f(x_k) = \nabla^2 \hat{f}_k(0)$, we get a more directly useful corollary: a complexity result for the computation of approximate second-order critical points on manifolds. The proof is in Appendix C.

Corollary 3 *Under A1 and A2, for an arbitrary $x_0 \in \mathcal{M}$ and for any $\varepsilon_g, \varepsilon_H > 0$, if either*

(a) we use the exponential retraction and A3 holds, or
(b) we use a second-order retraction and both A4 and A5 hold,

then Algorithm 1 with second-order progress (3) enforced produces a point $x_k \in \mathcal{M}$ such that $f(x_k) \leq f(x_0)$, $\|\mathrm{grad} f(x_k)\| \leq \varepsilon_g$ and $\lambda_{\min}(\mathrm{Hess} f(x_k)) \geq -\varepsilon_H$ in at most

$$\left(1 + \frac{|\log(\gamma_1)|}{\log(\gamma_2)}\right)(K_1(\varepsilon_g) + K_2(\varepsilon_H) + 1) + \frac{1}{\log(\gamma_2)} \log\left(\frac{\varsigma_{\max}}{\varsigma_0}\right) + 1$$

iterations, with K_1 as provided by Theorem 3 or 4, depending on assumptions.

If the retraction is not second order, we still get a bound on the eigenvalues of the Riemannian Hessian if the retraction has bounded acceleration at the origin at x_k: see (Boumal et al., 2018, §3.5) and also Lemma 5 below.

6 Regularity assumptions

The regularity assumptions A2 and A4 pertain to the pullbacks $f \circ \mathrm{R}_{x_k}$. As such, they mix the roles of f and R. The purpose of this section is to shed some light on these assumptions, specifically in a way that disentangles the roles of f and R. In that respect, the main result is Theorem 6 below. Proofs are in Appendix D.

Each pullback is a function from a Euclidean space $\mathrm{T}_{x_k}\mathcal{M}$ to $\mathbb{R}$, so that standard calculus applies. Since the retraction is smooth by definition, pullbacks are as many times differentiable as f. This leads to the following simple fact.

Lemma 4 *Assume $f : \mathcal{M} \to \mathbb{R}$ is twice continuously differentiable. If there exists L such that, for all (x, s) among the sequence of iterates and trial steps $\{(x_0, s_0), (x_1, s_1), \ldots\}$ produced by Algorithm 1, with $\hat{f} = f \circ \mathrm{R}_x$, it holds that*

$$\left\| \nabla^2 \hat{f}(ts) - \nabla^2 \hat{f}(0) \right\|_{\mathrm{op}} \leq tL\|s\| \tag{27}$$

for all $t \in [0,1]$*, then A2 holds with this* L *and A4 holds with* $L' = L$ *and* $q \equiv 0$*.*

We call this a *Lipschitz-type* assumption on $\nabla^2 \hat{f}$ because it compares the Hessians at ts and 0, rather than comparing them at two arbitrary points on the tangent space. On the other hand, we require this to hold on several tangent spaces with the same constant L.

In light of Lemma 4, one way to understand our regularity assumptions is to understand the Hessian of the pullback at points which are not the origin. The following lemma provides the necessary identities. The Hessian formula we have not seen elsewhere. Notation-wise, recall that T^* denotes the adjoint of a linear operator T; furthermore, the *intrinsic acceleration* $c''(t)$ of a smooth curve $c(t)$ is the covariant derivative of its velocity vector field $c'(t)$ on the Riemannian manifold $\mathcal{M}$.

Lemma 5 *Given* $f \colon \mathcal{M} \to \mathbb{R}$ *twice continuously differentiable and* $x \in \mathcal{M}$*, the gradient and Hessian of the pullback* $\hat{f} = f \circ \mathrm{R}_x$ *at* $s \in \mathrm{T}_x\mathcal{M}$ *are given by*

$$\nabla \hat{f}(s) = T_s^* \mathrm{grad} f(\mathrm{R}_x(s)), \tag{28}$$

$$\nabla^2 \hat{f}(s) = T_s^* \circ \mathrm{Hess} f(\mathrm{R}_x(s)) \circ T_s + W_s, \tag{29}$$

where

$$T_s = \mathrm{DR}_x(s) \colon \mathrm{T}_x\mathcal{M} \to \mathrm{T}_{\mathrm{R}_x(s)}\mathcal{M} \tag{30}$$

is linear, and W_s *is a symmetric linear operator on* $\mathrm{T}_x\mathcal{M}$ *defined through polarization by*

$$\langle W_s[\dot{s}], \dot{s} \rangle = \langle \mathrm{grad} f(\mathrm{R}_x(s)), c''(0) \rangle \,, \tag{31}$$

with $c''(0) \in \mathrm{T}_{\mathrm{R}_x(s)}\mathcal{M}$ *the intrinsic acceleration on* $\mathcal{M}$ *of* $c(t) = \mathrm{R}_x(s + t\dot{s})$ *at* $t = 0$*.*

The particular case $s = 0$ connects to the comments around Corollary 3: since T_0 is identity by definition of retractions, we find that

$$\nabla^2 \hat{f}(0) = \mathrm{Hess} f(x) + W_0, \tag{32}$$

where W_0 is zero in particular if the initial acceleration of retraction curves is zero (or if $\mathrm{grad} f(x) = 0$). As in (Absil et al., 2008, §5), this motivates the definition of second-order retractions, for which $\nabla^2 \hat{f}(0) = \mathrm{Hess} f(x)$.

Definition 3 (Second-order retraction) A retraction R on $\mathcal{M}$ is *second order* if, for any $x \in \mathcal{M}$ and $\dot{s} \in \mathrm{T}_x\mathcal{M}$, the curve $c(t) = \mathrm{R}_x(t\dot{s})$ has zero initial acceleration: $c''(0) = 0$.

Practical second-order retractions are often available (Absil and Malick, 2012, Ex. 23). To prove our main result, we further restrict retractions.

Definition 4 (Second-order nice retraction) Let S be a subset of the tangent bundle $\mathrm{T}\mathcal{M}$. A retraction R on $\mathcal{M}$ is *second-order nice on* S if there exist constants c_1, c_2, c_3 such that, for all $(x, s) \in S$ and for all $\dot{s} \in \mathrm{T}_x\mathcal{M}$, all of the following hold:

1. $\forall t \in [0,1]$, $\|T_{ts}\|_{\mathrm{op}} \leq c_1$ where $T_{ts} = \mathrm{DR}_x(ts)$;
2. $\forall t \in [0,1]$, the covariant derivative of $U(t) = T_{ts}\dot{s}$ satisfies $\left\|\frac{\mathrm{D}}{\mathrm{d}t}U(t)\right\| \leq c_2\|s\|\|\dot{s}\|$; and
3. $\|c''(0)\| \leq c_3\|s\|\|\dot{s}\|^2$ where $c(t) = \mathrm{R}_x(s + t\dot{s})$.

If this holds for $S = \mathrm{T}\mathcal{M}$, we say R is *second-order nice*.

Second-order nice retractions are, in particular, second-order retractions (consider $s = 0$ in the last condition). For $\mathcal{M}$ a Euclidean space, the canonical retraction $\mathrm{R}_x(s) = x + s$ is second-order nice with $c_1 = 1$ and $c_2 = c_3 = 0$. The classical retraction on the sphere is also second-order nice, with small constants c_1, c_2, c_3. We expect this to be the case for many usual retractions on compact or flat manifolds. For manifolds with negative curvature, the exponential retraction would lead $\|T_s\|_{\mathrm{op}}$ to grow arbitrarily large with s going to infinity, hence it is important to consider restrictions to appropriate subsets, or to use another retraction.

Proposition 3 *For the unit sphere* $\mathcal{M} = \{x \in \mathbb{R}^n : \|x\| = 1\}$ *as a Riemannian submanifold of* $\mathbb{R}^n$*, the retraction* $\mathrm{R}_x(s) = \frac{x+s}{\|x+s\|}$ *is second-order nice with* $c_1 = 1, c_2 = \frac{4\sqrt{3}}{9}, c_3 = 2$.

Theorem 6 *Let* $f\colon \mathcal{M} \to \mathbb{R}$ *be three times continuously differentiable. Assume the retraction is second-order nice on the set* $\{(x_0, s_0), (x_1, s_1), \ldots\}$ *of points and steps generated by Algorithm 1 (see Definition 4). If the sequence* $x_0, x_1, x_2, \ldots$ *remains in a compact subset of* $\mathcal{M}$*, then A2 and A4 are satisfied with a same* L *(related to the Lipschitz properties of* f*,* $\mathrm{grad}f$ *and* $\mathrm{Hess}f$*: see the proof for an explicit expression) and* $q \equiv 0$.

Algorithm 1 is a descent method, so that the last condition holds in particular if the sublevel set $\{x \in \mathcal{M} : f(x) \leq f(x_0)\}$ *is compact, and a fortiori if* $\mathcal{M}$ *is compact.*

This general result shows existence of (loose) bounds for the Lipschitz constants. For specific optimization problems, it is sometimes easy to derive more accurate constants by direct computation: see for example (Criscitiello and Boumal, 2019, App. D) for PCA.

7 Controlling the differentiated retraction

In Theorem 4, we control the worst-case running time of ARC via the differential of the retraction R, through A5. This assumption, which does not come up in the Euclidean case, involves constants a, b to control $\sigma_{\min}(\mathrm{DR}_{x_k}(s_k))$. Our first result shows A5 is satisfied for some a and b for any retraction on any manifold, provided the sublevel set of $f(x_0)$ is compact (see also Theorem 6). This is mostly a topological argument. Proofs are in Appendix E.

Theorem 7 *Let* R *be a retraction on a Riemannian manifold* $\mathcal{M}$*, and let* $\mathcal{U}$ *be a nonempty compact subset of* $\mathcal{M}$*. For any* $b \in (0,1)$ *there exists* $a > 0$ *such that, for all* $x \in \mathcal{U}$ *and* $s \in \mathrm{T}_x\mathcal{M}$ *with* $\|s\|_x \leq a$*, we have* $\sigma_{\min}(\mathrm{DR}_x(s)) \geq b$*. In particular, A5 is satisfied with such* (a, b) *provided the iterates* $x_0, x_1, x_2, \ldots$ *remain in* $\mathcal{U}$*.*

We further quantify the constants a and b in two cases of interest:

1. For the Stiefel manifold $\mathrm{St}(n,p) = \{X \in \mathbb{R}^{n\times p} : X^\top X = I_p\}$ as a Riemannian submanifold of $\mathbb{R}^{n\times p}$ with the usual inner product $\langle A, B\rangle = \mathrm{Tr}(A^\top B)$, we explicitly control (a, b) for the popular Q-factor retraction ($\mathrm{R}_X(S)$ is obtained by Gram–Schmidt orthonormalization of the columns of $X + S$). Special cases include the sphere ($p = 1$) and the orthogonal group ($p = n$).
2. For complete manifolds with bounded sectional curvature, we control (a, b) for the case of the exponential retraction. Important special cases include Euclidean spaces (flat manifolds), manifolds with nonpositive curvature (Hadamard manifolds, including the manifold of positive definite matrices (Moakher and Batchelor, 2006; Bhatia, 2007)), and compact manifolds (Bishop and Crittenden, 1964, §9.3, p166).

The first result follows from a direct calculation.

Proposition 4 *For the Stiefel manifold with the Q-factor retraction, for any* $a > 0$*, define* $b = 1 - 3a - \frac{1}{2}a^2$*. If* b *is positive, then A5 holds with these* a *and* b*. Moreover, for the sphere, we have that A5 is satisfied for any* $a > 0$ *and* $b = \frac{1}{1+a^2}$*.*

The second result follows from the connection between the differential of the exponential map and certain *Jacobi fields* on $\mathcal{M}$, together with standard comparison theorems from Riemannian geometry (Lee, 2018, Ch. 10, 11).

Proposition 5 *Let* $\mathcal{M}$*, complete, have sectional curvature upper bounded by* C*, and let the retraction* R *be the exponential retraction* Exp*:*

If $C \leq 0$*, then A5 is satisfied for any* $a > 0$ *and* $b = 1$*;*
If $C > 0$*, then A5 is satisfied for any* $0 < a < \frac{\pi}{\sqrt{C}}$ *and* $b = \frac{\sin(a\sqrt{C})}{a\sqrt{C}}$*.*

8 Solving the subproblem

At each iteration, Algorithm 1 requires the approximate minimization of the model m_k (1) in the tangent space $\mathrm{T}_{x_k}\mathcal{M}$. Since the latter is a linear space, this is the same subproblem as in the Euclidean case. In contrast to working simply over $\mathbb{R}^n$ however, one practical difference is that we do not usually have access to a preferred basis for $\mathrm{T}_{x_k}\mathcal{M}$, so that it is preferable to resort to basis-free solvers. To this end, we describe a Lanczos method as Algorithm 2.

Let us phrase the subproblem in a general context. Given a vector space $\mathcal{X}$ of dimension n with an inner product $\langle\cdot,\cdot\rangle$ (and associated norm $\|\cdot\|$), an

element $g \in \mathcal{X}$, a self-adjoint linear operator $H \colon \mathcal{X} \to \mathcal{X}$ and a real $\varsigma > 0$, define the function $m \colon \mathcal{X} \to \mathbb{R}$ as

$$m(s) = \langle g, s \rangle + \frac{1}{2} \langle s, H(s) \rangle + \frac{\varsigma}{3} \|s\|^3. \tag{33}$$

We wish to compute an element $s \in \mathcal{X}$ such that

$$m(s) \leq m(0) \qquad \text{and} \qquad \|\nabla m(s)\| \leq \theta \|s\|^2. \tag{34}$$

This corresponds to satisfying condition (2) at iteration k, where $\mathcal{X} = \mathrm{T}_{x_k}\mathcal{M}$ is endowed with the Riemannian inner product at x_k, $g = \nabla \hat{f}_k(0)$, $H = \nabla^2 \hat{f}_k(0)$ and $\varsigma = \varsigma_k$. If $g = 0$, then $s = 0$ satisfies the condition: henceforth, we assume $g \neq 0$.

Certainly, a global minimizer of (33) meets our requirements (it would also satisfy the equivalent of the second-order condition (3)). Such a minimizer can be computed, but known procedures for this task involve a diagonalization of H, which may be expensive. Instead, we use the Lanczos-based method proposed in (Cartis et al., 2011b, §6): the latter iteratively produces a sequence of orthonormal vectors $\{q_1, \ldots, q_n\}$ and a symmetric tridiagonal matrix T of size n such that (Trefethen and Bau, 1997, Lec. 36)[1]

$$q_1 = \frac{g}{\|g\|} \qquad \text{and} \qquad T_{ij} = \langle q_i, H(q_j) \rangle \text{ for all } i, j \text{ in } 1 \ldots n. \tag{35}$$

Let T_k denote the $k \times k$ principal submatrix of T: producing $q_1, \ldots, q_k$ and T_k requires exactly k calls to H. Consider $m(s)$ (33) restricted to the subspace spanned by $q_1, \ldots, q_k$:

$$\begin{aligned} \forall y \in \mathbb{R}^k, \text{ with } s = \sum_{i=1}^{k} y_i q_i, \quad m(s) &= \langle g, y_1 q_1 \rangle + \frac{1}{2} \sum_{i,j=1}^{k} y_i y_j \langle q_i, H(q_j) \rangle + \frac{\varsigma}{3} \|y\|^3 \\ &= y_1 \|g\| + \frac{1}{2} y^\top T_k y + \frac{\varsigma}{3} \|y\|^3, \end{aligned} \tag{36}$$

where $\|\cdot\|$ is also the 2-norm over $\mathbb{R}^k$. Since T_k is tridiagonal, it can be diagonalized efficiently. As a result, it is inexpensive to compute a global minimizer of $m(s)$ restricted to the subspace spanned by $q_1, \ldots, q_k$ (Cartis et al., 2011b, §6.1). Furthermore, since the Lanczos basis is constructed incrementally, we can minimize the restricted cubic at $k = 1$, check the stopping criterion (34), and proceed to $k = 2$ only if necessary, etc. The hope (borne out in experiments) is that the algorithm stops well before k reaches n (at which point it necessarily succeeds.) In this way, we limit the number of calls to H, which is typically the most expensive part of the process. This general strategy was first proposed in the context of trust-region subproblems by Gould et al. (1999). Algorithm 2 is set up to target first order progress only. In order to ensure

[1] In case of so-called breakdown in the Lanczos iteration at step k, we follow the standard procedure which is to generate q_k as a random unit vector orthogonal to $q_1, \ldots, q_{k-1}$, then to proceed as normal. This does not jeopardize the desired properties (35).

satisfaction of second-order progress, one may have to force the execution of n iterations (which is rarely done in practice).

We draw attention to a technical point. Upon minimizing (36), we obtain a vector $y \in \mathbb{R}^k$. To check the stopping criterion (34), we must compute $\|\nabla m(s)\|$, where $s = \sum_{i=1}^{k} y_i q_i$. Since

$$\nabla m(s) = g + H(s) + \varsigma \|s\| s, \tag{37}$$

one approach involves computing s (that is, form the linear combination of q_i's) and applying H to s: both operations may be expensive in high dimension. An alternative (shown in Algorithm 2) is to recognize that, due to the inner workings of Lanczos iterations, $\nabla m(s)$ lies in the subspace spanned by $\{q_1, \ldots, q_{k+1}\}$ (if $k < n$). Explicitly,

$$\nabla m(s) = \|g\| q_1 + \sum_{i=1}^{k+1} (T_{1:k+1,1:k}\, y)_i q_i + \varsigma \|y\| \sum_{i=1}^{k} y_i q_i, \tag{38}$$

where $T_{1:k+1,1:k}$ is the submatrix of T containing the first $k+1$ rows and first k columns. This expression gives a direct way to compute $\|\nabla m(s)\|$ without forming s and without calling H, simply by running the Lanczos iteration one step ahead.

Remark 1 Carmon and Duchi (2018) analyzed the number of Lanczos iterations that may be required to reach approximate solutions to the subproblem. For the Euclidean case, they conclude that the overall complexity of ARC to compute an ε-critical point in terms of Hessian-vector products (which dominate the number of cost and gradient computations) is $O(\varepsilon^{-7/4})$. Furthermore, the dependence on the dimension of the search space is only logarithmic. (The logarithmic terms are caused by the need to randomize for the so-called hard case—see the reference for important details in that regard.) Their conclusions should extend to the Riemannian setting as well. Gould and Simoncini (2019) extend these results to study the decrease of the norm of the gradient specifically, as required here.

9 Numerical experiments

We implement Algorithm 1 within the Manopt framework (Boumal et al., 2014) (our code is part of that toolbox) and compare the performance of our implementation against some existing solvers in that toolbox, namely, the Riemannian trust-region method (RTR) (Absil et al., 2007) and the Riemannian conjugate gradients method with Hestenes–Stiefel update formula (CG-HS) (Absil et al., 2008, §8.3). All algorithms terminate when $\|\mathrm{grad} f(x_k)\| \leq 10^{-9}$. CG-HS also terminates if it is unable to produce a step of size more than 10^{-10}. Code to reproduce the experiments is available at `https://github.com/NicolasBoumal/arc`. We report results with randomness fixed by `rng(2019)` within Matlab R2019b.

We consider a suite of six Riemannian optimization problems:

Algorithm 2 Lanczos-based cubic model subsolver

Parameters: $\theta, \varsigma > 0$, vector space $\mathcal{X}$ with inner product $\langle \cdot, \cdot \rangle$ and norm $\| \cdot \|$
Input: $g \in \mathcal{X}$ nonzero, a self-adjoint linear operator $H : \mathcal{X} \to \mathcal{X}$

$k \leftarrow 1$
Obtain q_1, T_1 via a Lanczos iteration (35)
Solve $y^{(1)} = \operatorname{argmin}_{y \in \mathbb{R}} \ \|g\|y + \frac{1}{2}T_1 y^2 + \frac{1}{3}\varsigma|y|^3$
Obtain q_2, T_2 via a Lanczos iteration (35)
Compute $\|\nabla m(s_1)\|$ via (38), where $s_1 = y^{(1)} q_1$

while $\|\nabla m(s_k)\| > \theta \|s_k\|^2$ **do**
    $k \leftarrow k + 1$
    Solve $y^{(k)} = \operatorname{argmin}_{y \in \mathbb{R}^k} \ \|g\|y_1 + \frac{1}{2}y^\top T_k y + \frac{1}{3}\varsigma\|y\|^3$ following (Cartis et al., 2011b, §6.1)
    Obtain q_{k+1}, T_{k+1} via a Lanczos iteration (35)
    Compute $\|\nabla m(s_k)\|$ via (38), where $s_k = \sum_{i=1}^{k} y_i^{(k)} q_i$
end while

Output: $s_k \in \mathcal{X}$

1. Dominant invariant subspace: $\max_{X \in \mathrm{Gr}(n,p)} \frac{1}{2}\mathrm{Tr}(X^\top A X)$, where $A \in \mathbb{R}^{n \times n}$ is symmetric (randomly generated from i.i.d. Gaussian entries) and $\mathrm{Gr}(n, p)$ is the Grassmann manifold of subspaces of dimension p in $\mathbb{R}^n$, represented by orthonormal matrices in $\mathbb{R}^{n \times p}$. Optima correspond to dominant invariant subspaces of A (Edelman et al., 1998).
2. Truncated SVD: $\max_{U \in \mathrm{St}(m,p), V \in \mathrm{St}(n,p)} \mathrm{Tr}(U^\top A V N)$, where $\mathrm{St}(n, p)$ is the set of matrices in $\mathbb{R}^{n \times p}$ with orthonormal columns, $A \in \mathbb{R}^{m \times n}$ has i.i.d. random Gaussian entries and $N = \mathrm{diag}(p, p-1, \ldots, 1)$. Global optima correspond to the p dominant left and right singular vectors of A (Sato and Iwai, 2013). For this and the previous problem, the random matrices have small eigen or singular value gap, which makes them challenging.
3. Low-rank matrix completion via optimization on one Grassmann manifold, as in (Boumal and Absil, 2011). The target matrix $A \in \mathbb{R}^{m \times n}$ has rank r: it is fully specified by $r(m+n-r)$ parameters. We observe this many entries of A picked uniformly at random, times an oversampling factor (osf). The task is to recover A from those samples. A is generated from two random Gaussian factors as $A = LR^\top$ to have rank r exactly. The variable is $U \in \mathrm{Gr}(m, r)$, and the cost function minimizes the sum of squared errors between UW_U and the observed entries of A, where $W_U \in \mathbb{R}^{r \times n}$ is the optimal matrix for that purpose (which has an explicit expression once U is fixed, efficiently computable).
4. Max-cut: given the adjacency matrix $A \in \mathbb{R}^{n \times n}$ of a graph with n nodes, we solve the semidefinite relaxation of the Max-Cut graph partitioning problem via the Burer–Monteiro formulation (Burer and Monteiro, 2005) on the oblique manifold (Journée et al., 2010): $\min_{X \in \mathrm{OB}(n,p)} \frac{1}{2}\mathrm{Tr}(X^\top A X)$, where $\mathrm{OB}(n, p)$ is the set of matrices in $\mathbb{R}^{n \times p}$ with unit-norm rows. Here, we pick graph #22 from a collection of graphs called Gset (see any of the

references): it has $n = 2000$ nodes and 19990 edges, and p is set close to $\sqrt{2n}$ as justified in (Boumal et al., 2019).

5. Synchronization of rotations: m rotation matrices $Q_1, \ldots, Q_m$ in the special orthogonal group $\mathrm{SO}(d)$ are estimated from noisy relative measurements $H_{ij} \approx Q_i Q_j^\top$ for an Erdős–Rényi random set of pairs (i, j) following a maximum likelihood formulation, as in (Boumal et al., 2013). The specific distribution of the measurements and the corresponding cost function are described in the reference. All algorithms are initialized with the technique proposed in the reference, to avoid convergence to a poor local optimum.
6. ShapeFit: least-squares formulation of the problem of recovering a rigid structure of n points $x_1, \ldots, x_n$ in $\mathbb{R}^d$ from noisy measurements of some of the pairwise directions $\frac{x_i - x_j}{\|x_i - x_j\|}$ picked uniformly at random, following (Hand et al., 2018). The set of points is centered and obeys one extra linear constraint to fix scaling ambiguity, so that the search space is effectively a linear subspace of $\mathbb{R}^{n \times d}$: this is the manifold $\mathcal{M}$. The cost function as spelled out in the reference makes this a structured linear least-squares problem.

For each problem, we generate one instance and one random initial guess (except for problem 5 which is initialized deterministically). Then, we run each algorithm from that same initial guess on that same instance. Figure 1 displays the progress of each algorithm on each problem as the gradient norm of iterates (on a log scale) as a function of elapsed computation time to reach each iterate (in seconds) on a laptop from 2016. For the same run, Figure 2 reports the number of gradient calls and Hessian-vector products (summed) issued by all algorithms along the way. Figure 3 reports the number of outer iterations for ARC and RTR, that is, excluding work done by subsolvers.

For ARC, we report results with $\theta = 0.25$ and $\theta = 2$; this is used in the stopping criterion for subproblem solves following (2) (we do not check (3)). To initialize ς_0, we use 100 divided by the initial trust-region radius of RTR (Δ_0) chosen by Manopt. Other parameters of ARC are set as follows: $\varsigma_{\min} = 10^{-10}, \eta_1 = 0.1, \eta_2 = 0.9, \gamma_1 = 0.1, \gamma_2 = \gamma_3 = 2$, with update rule:

$$\varsigma_{k+1} = \begin{cases} \max(\varsigma_{\min}, \gamma_1 \varsigma_k) & \text{if } \rho_k \geq \eta_2 \quad \text{(very successful)}, \\ \varsigma_k & \text{if } \rho_k \in [\eta_1, \eta_2) \quad \text{(successful)}, \\ \gamma_2 \varsigma_k & \text{if } \rho_k < \eta_1 \quad \text{(unsuccessful)}. \end{cases} \tag{39}$$

Standard safeguards to account for numerical round-off errors are included in the code (not described here). Parameters for the other methods have the default values given by Manopt.

We experiment with two subproblem solvers for ARC. The first one is the Lanczos-based method as described in Section 8 (ARC Lanczos). The second one is a (Euclidean) nonlinear, nonnegative-Polak–Ribière conjugate gradients method run on the model m_k (1) in the tangent space $\mathrm{T}_{x_k}\mathcal{M}$ (ARC NLCG), implemented by Bryan Zhu (Zhu, 2019). This solver uses the initialization recommended by Carmon and Duchi (2019) for gradient descent (and from which they proved convergence to a global optimizer, despite non-convexity

of the model), and exact line-search. We find that this subproblem solver performs well in practice. It is simpler to implement, and uses less memory than the Lanczos method.

We find that ARC's performance is in the same ballpark as RTR's, with the caveat that ARC's best performance requires tuning (choosing the right subproblem solver and θ for the problem class), whereas RTR is more robust. Since RTR's code has been refined over many years, we expect that further work can help reduce the gap. For example, we expect that the performance of ARC could be improved with further tuning of the regularization parameter update rule. In particular, we find that it is important to reduce regularization fast when close to convergence (yet not earlier), to allow ARC to make steps similar to Newton's method. Work by Gould et al. (2012) could be a good starting point for such exploration.

Acknowledgments We thank Pierre-Antoine Absil for numerous insightful and technical discussions, Stephen McKeown for directing us to, and guiding us through the relevance of Jacobi fields for our study of A5, Chris Criscitiello and Eitan Levin for many discussions regarding regularity assumptions on manifolds, and Bryan Zhu for contributing his nonlinear CG subproblem solver to Manopt, and related discussions.

References

P.-A. Absil and J. Malick. Projection-like retractions on matrix manifolds. *SIAM Journal on Optimization*, 22(1):135–158, 2012. doi:10.1137/100802529.

P.-A. Absil, C. G. Baker, and K. A. Gallivan. Trust-region methods on Riemannian manifolds. *Foundations of Computational Mathematics*, 7(3):303–330, 2007. doi:10.1007/s10208-005-0179-9.

P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, Princeton, NJ, 2008. ISBN 978-0-691-13298-3.

R. Adler, J. Dedieu, J. Margulies, M. Martens, and M. Shub. Newton's method on Riemannian manifolds and a geometric model for the human spine. *IMA Journal of Numerical Analysis*, 22(3):359–390, 2002. doi:10.1093/imanum/22.3.359.

N. Agarwal, Z. Allen-Zhu, B. Bullins, E. Hazan, and T. Ma. Finding approximate local minima faster than gradient descent. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1195–1199. ACM, 2017.

G. Bento, O. Ferreira, and J. Melo. Iteration-complexity of gradient, subgradient and proximal point methods on Riemannian manifolds. *Journal of Optimization Theory and Applications*, 173(2):548–562, 2017. doi:10.1007/s10957-017-1093-4.

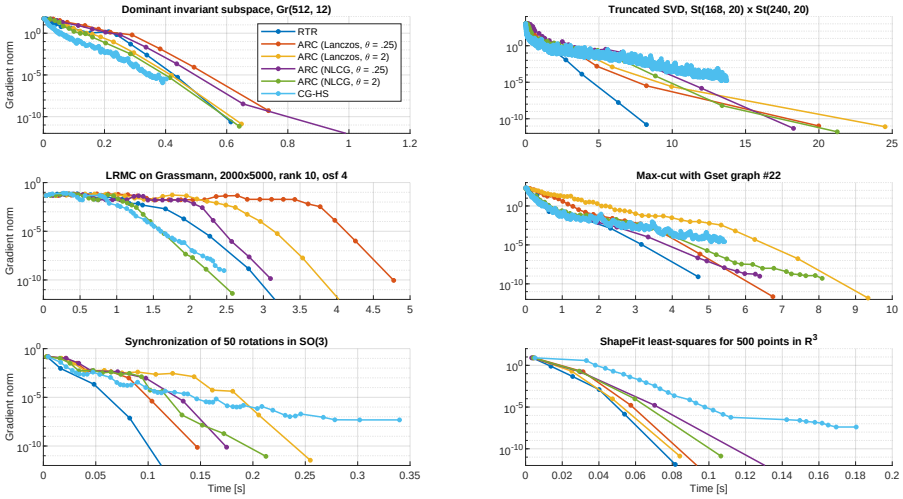

Fig. 1: Gradient norm at each iterate for the three competing solvers on the six benchmark problems, as a function of computation time needed by those solvers to reach those iterates (in seconds). Our algorithm is ARC (tested with two different subproblem solvers, each with two parameter settings); RTR is Riemannian trust-regions; CG-HS is Riemannian conjugate gradients with Hestenes–Stiefel step selection.

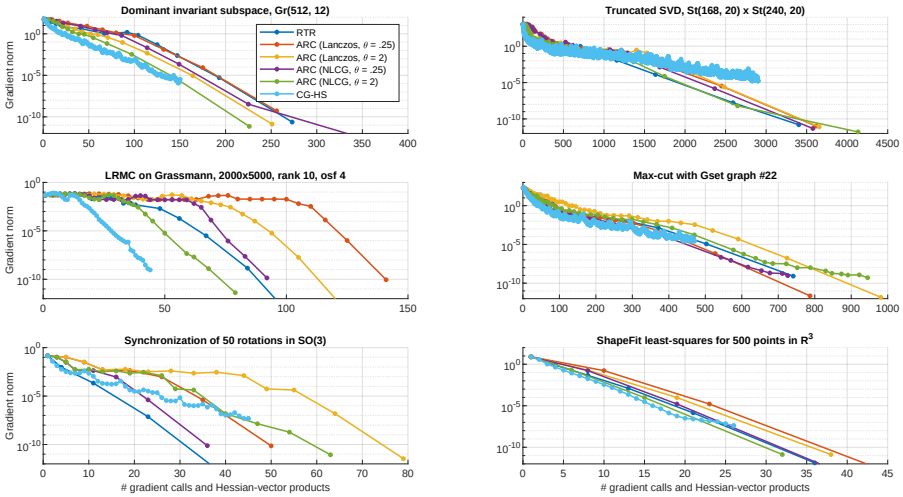

Fig. 2: Gradient norm at each iterate, as a function of the number of gradient calls and Hessian-vector calls (the sum of both) issued by those solvers to reach those iterates.

C. Bergé. *Topological Spaces: including a treatment of multi-valued functions, vector spaces, and convexity*. Oliver and Boyd, Ltd, 1963.

R. Bhatia. *Positive definite matrices*. Princeton University Press, 2007.

E. Birgin, J. Gardenghi, J. Martínez, S. Santos, and P. Toint. Worst-case evaluation complexity for unconstrained nonlinear optimization using high-order regularized models. *Mathematical Programming*, 163(1):359–368, May 2017. doi:10.1007/s10107-016-1065-8.

R. Bishop and R. Crittenden. *Geometry of manifolds*, volume 15. Academic press, 1964.

S. Bonnabel. Stochastic gradient descent on Riemannian manifolds. *Automatic Control, IEEE Transactions on*, 58(9):2217–2229, 2013. doi:10.1109/TAC.2013.2254619.

N. Boumal. An introduction to optimization on smooth manifolds. To appear, 2020.

N. Boumal and P.-A. Absil. RTRMC: A Riemannian trust-region method for low-rank matrix completion. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 24 (NIPS)*, pages 406–414. 2011.

N. Boumal, A. Singer, and P.-A. Absil. Robust estimation of rotations from relative measurements by maximum likelihood. In *Decision and Control (CDC), 2013 IEEE 52nd Annual Conference on*, pages 1156–1161, Dec 2013. doi:10.1109/CDC.2013.6760038.

N. Boumal, B. Mishra, P.-A. Absil, and R. Sepulchre. Manopt, a Matlab toolbox for optimization on manifolds. *Journal of Machine Learning Research*, 15:1455–1459, 2014. URL `http://www.manopt.org`.

N. Boumal, P.-A. Absil, and C. Cartis. Global rates of convergence for nonconvex optimization on manifolds. *IMA Journal of Numerical Analysis*, 2018. doi:10.1093/imanum/drx080.

N. Boumal, V. Voroninski, and A. Bandeira. Deterministic guarantees for Burer-Monteiro factorizations of smooth semidefinite programs. *Communications on Pure and Applied Mathematics*, 73(3):581–608, 2019. doi:10.1002/cpa.21830.

S. Burer and R. Monteiro. Local minima and convergence in low-rank semidefinite programming. *Mathematical Programming*, 103(3):427–444, 2005.

Y. Carmon and J. Duchi. Gradient descent finds the cubic-regularized nonconvex Newton step. *SIAM Journal on Optimization*, 29(3):2146–2178, 2019. doi:10.1137/17M1113898.

Y. Carmon and J. C. Duchi. Analysis of Krylov subspace solutions of regularized nonconvex quadratic problems. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 10728–10738. Curran Associates, Inc., 2018.

Y. Carmon, J. Duchi, O. Hinder, and A. Sidford. Lower bounds for finding stationary points I. *Mathematical Programming*, 2019. doi:10.1007/s10107-019-01406-y.

C. Cartis, N. Gould, and P. Toint. Adaptive cubic regularisation methods for unconstrained optimization. Part II: worst-case function- and derivative-evaluation complexity. *Mathematical Programming*, 130:295–319, 2011a. doi:10.1007/s10107-009-0337-y.

C. Cartis, N. Gould, and P. Toint. Adaptive cubic regularisation methods for unconstrained optimization. Part I: motivation, convergence and numerical results. *Mathematical Programming*, 127(2):245–295, 2011b. doi:10.1007/s10107-009-0286-5.

C. Cartis, N. Gould, and P. Toint. Complexity bounds for second-order optimality in unconstrained optimization. *Journal of Complexity*, 28(1):93–108, 2012. doi:10.1016/j.jco.2011.06.001.

C. Cartis, N. Gould, and P. Toint. Improved second-order evaluation complexity for unconstrained nonlinear optimization using high-order regularized models. *arXiv preprint arXiv:1708.04044*, 2017.

C. Cartis, N. I. Gould, and P. L. Toint. Worst-case evaluation complexity and optimality of second-order methods for nonconvex smooth optimization. *arXiv preprint arXiv:1709.07180. To appear in Proceedings of the ICM*, 2018.

C. Criscitiello and N. Boumal. Efficiently escaping saddle points on manifolds. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 5985–5995. Curran Associates, Inc., 2019. URL http://papers.nips.cc/paper/8832-efficiently-escaping-saddle-points-on-manifolds.

M. do Carmo. *Riemannian geometry*. Mathematics: Theory & Applications. Birkhäuser Boston Inc., Boston, MA, 1992. ISBN 0-8176-3490-8. Translated from the second Portuguese edition by Francis Flaherty.

J.-P. Dussault. ARCq: A new adaptive regularization by cubics. *Optimization Methods and Software*, 33(2):322–335, 2018. doi:10.1080/10556788.2017.1322080.

A. Edelman, T. Arias, and S. Smith. The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications*, 20 (2):303–353, 1998.

O. Ferreira and B. Svaiter. Kantorovich's theorem on Newton's method in Riemannian manifolds. *Journal of Complexity*, 18(1):304–329, 2002. doi:https://doi.org/10.1006/jcom.2001.0582.

D. Gabay. Minimizing a differentiable function over a differential manifold. *Journal of Optimization Theory and Applications*, 37(2):177–219, 1982.

N. Gould and V. Simoncini. Error estimates for iterative algorithms for minimizing regularized quadratic subproblems. *Optimization Methods and Software*, 0(0):1–25, 2019. doi:10.1080/10556788.2019.1670177.

N. Gould, S. Lucidi, M. Roma, and P. Toint. Solving the trust-region subproblem using the Lanczos method. *SIAM Journal on Optimization*, 9(2): 504–525, 1999. doi:10.1137/S1052623497322735.

N. I. M. Gould, M. Porcelli, and P. L. Toint. Updating the regularization parameter in the adaptive cubic regularization algorithm. *Computational*

Optimization and Applications, 53(1):1–22, Sep 2012. doi:10.1007/s10589-011-9446-7.

A. Griewank. The modification of Newton's method for unconstrained optimization by bounding cubic terms. Technical Report Technical report NA/12, Department of Applied Mathematics and Theoretical Physics, University of Cambridge, 1981.

P. Hand, C. Lee, and V. Voroninski. ShapeFit: Exact location recovery from corrupted pairwise directions. *Communications on Pure and Applied Mathematics*, 71(1):3–50, 2018.

J. Hu, A. Milzarek, Z. Wen, and Y. Yuan. Adaptive quadratically regularized Newton method for Riemannian optimization. *SIAM Journal on Matrix Analysis and Applications*, 39(3):1181–1207, 2018. doi:10.1137/17M1142478.

C. Jin, P. Netrapalli, R. Ge, S. Kakade, and M. Jordan. Stochastic gradient descent escapes saddle points efficiently. *arXiv:1902.04811*, 2019.

M. Journée, F. Bach, P.-A. Absil, and R. Sepulchre. Low-rank optimization on the cone of positive semidefinite matrices. *SIAM Journal on Optimization*, 20(5):2327–2351, 2010. doi:10.1137/080731359.

J. Kohler and A. Lucchi. Sub-sampled cubic regularization for non-convex optimization. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pages 1895–1904. JMLR.org, 2017.

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

D. Luenberger. The gradient projection method along geodesics. *Management Science*, 18(11):620–631, 1972.

M. Moakher and P. Batchelor. *Symmetric Positive-Definite Matrices: From Geometry to Applications and Visualization*, pages 285–298. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006. doi:10.1007/3-540-31272-2_17.

Y. Nesterov and B. T. Polyak. Cubic regularization of Newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

B. O'Neill. *Semi-Riemannian geometry: with applications to relativity*, volume 103. Academic Press, 1983.

C. Qi. *Numerical Optimization Methods On Riemannian Manifolds*. PhD thesis, Department of Mathematics, Florida State University, Tallahassee, FL, 2011. URL https://diginole.lib.fsu.edu/islandora/object/fsu:180485/datastream/PDF/view.

W. Ring and B. Wirth. Optimization methods on Riemannian manifolds and their application to shape space. *SIAM Journal on Optimization*, 22(2):596–627, 2012. doi:10.1137/11082885X.

H. Sato and T. Iwai. A Riemannian optimization approach to the matrix singular value decomposition. *SIAM Journal on Optimization*, 23(1):188–212, 2013. doi:10.1137/120872887.

M. Shub. Some remarks on dynamical systems and numerical analysis. In L. Lara-Carrero and J. Lewowicz, editors, *Proc. VII ELAM.*, pages 69–92. Equinoccio, U. Simón Bolívar, Caracas, 1986.

S. Smith. Optimization techniques on Riemannian manifolds. *Fields Institute Communications*, 3(3):113–135, 1994.

Y. Sun, N. Flammarion, and M. Fazel. Escaping from saddle points on Riemannian manifolds. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 7276–7286. Curran Associates, Inc., 2019. URL http://papers.nips.cc/paper/8948-escaping-from-saddle-points-on-riemannian-manifolds.pdf.

L. Trefethen and D. Bau. *Numerical linear algebra*. Society for Industrial and Applied Mathematics, 1997. ISBN 978-0898713619.

N. Tripuraneni, N. Flammarion, F. Bach, and M. Jordan. Averaging stochastic gradient descent on riemannian manifolds. In *Conference On Learning Theory*, pages 650–687, 2018a.

N. Tripuraneni, M. Stern, C. Jin, J. Regier, and M. Jordan. Stochastic cubic regularization for fast nonconvex optimization. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 2899–2908. Curran Associates, Inc., 2018b. URL http://papers.nips.cc/paper/7554-stochastic-cubic-regularization-for-fast-nonconvex-optimization.pdf.

S. Waldmann. Geometric wave equations. *arXiv preprint arXiv:1208.4706*, 2012.

Z. Wang, Y. Zhou, Y. Liang, and G. Lan. Stochastic variance-reduced cubic regularization for nonconvex optimization. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2731–2740, 2019.

W. Yang, L.-H. Zhang, and R. Song. Optimality conditions for the nonlinear programming problems on Riemannian manifolds. *Pacific Journal of Optimization*, 10(2):415–434, 2014.

H. Zhang and S. Sra. First-order methods for geodesically convex optimization. In *Conference on Learning Theory*, pages 1617–1638, 2016.

H. Zhang, S. Reddi, and S. Sra. Riemannian SVRG: Fast stochastic optimization on Riemannian manifolds. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 4592–4600. Curran Associates, Inc., 2016.

J. Zhang and S. Zhang. A cubic regularized Newton's method over Riemannian manifolds. *arXiv preprint arXiv:1805.05565*, 2018.

J. Zhang, L. Xiao, and S. Zhang. Adaptive stochastic variance reduction for subsampled Newton method with cubic regularization. *arXiv preprint arXiv:1811.11637*, 2018.

D. Zhou, P. Xu, and Q. Gu. Stochastic variance-reduced cubic regularized Newton methods. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5990–5999, Stockholmsmassan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL http://proceedings.mlr.press/v80/zhou18d.html.

B. Zhu. Algorithms for optimization on manifolds using adaptive cubic regularization. Bachelor's thesis, Princeton University, Mathematics Department, 2019.

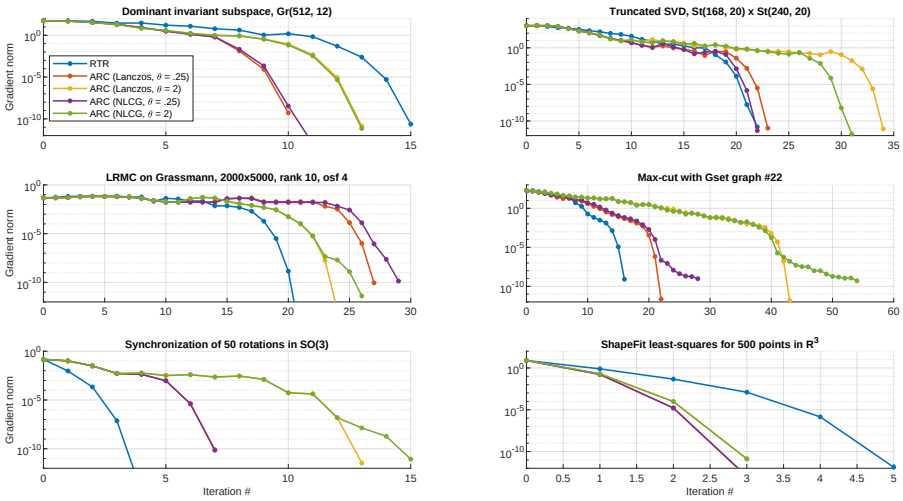

Fig. 3: Gradient norm at each iterate, as a function of the number of outer iterations for ARC and RTR: both of these solvers rely on a subproblem solver. This plot compares the behavior of the algorithms separately from their subproblem solvers' work. As a result, this hides effects related to how stringent the stopping criterion of the subproblem solver is, hence of how costly the subproblem solves are. For example, one can (usually) reduce the number of outer iterations of ARC by reducing θ. As the subproblems of RTR and ARC are similar, we expect that (in principle) it should be possible to solve them equally well in about the same time.

A Proofs from Section 2: mechanical lemmas

Lemma 1 characterizes the conditions under which the subproblem solver is allowed to return $s_k = 0$ at iteration k.

Proof of Lemma 1. By definition of the model m_k (1) and by properties of retractions (17),

$$\nabla m_k(0) = \nabla \hat{f}_k(0) = \mathrm{grad} f(x_k),$$

where $\hat{f}_k = f \circ \mathrm{R}_{x_k}$. Thus, if $\mathrm{grad} f(x_k) = 0$, the first-order condition (2) allows $s_k = 0$. The other way around, if $s_k = 0$ is allowed, then $\|\nabla m_k(0)\| = 0$, so that $\mathrm{grad} f(x_k) = 0$.

Now assume the second-order condition (3) is enforced. If $s_k = 0$ is allowed, then we already know that $\mathrm{grad} f(x_k) = 0$. Combined with (32), we deduce that

$$\nabla^2 m_k(0) = \nabla^2 \hat{f}_k(0) = \mathrm{Hess} f(x_k),$$

for any retraction. Then, condition (3) at $s_k = 0$ indicates $\nabla^2 m_k(0)$ is positive semidefinite, hence $\mathrm{Hess} f(x_k)$ is positive semidefinite. The other way around, if $\mathrm{grad} f(x_k) = 0$ and $\mathrm{Hess} f(x_k)$ is positive semidefinite, then $\nabla m_k(0) = \mathrm{grad} f(x_k)$ and $\nabla^2 m_k(0) = \mathrm{Hess} f(x_k)$, so that indeed $s_k = 0$ is allowed. □

The two supporting lemmas presented in Section 2 follow from the regularization parameter update mechanism of Algorithm 1. The standard proofs are not affected by the fact we here work on a manifold. We provide them for the sake of completeness.

Proof of Lemma 2. Using the definition of ρ_k (4), $m_k(0) = f(x_k)$ (1) and $m_k(0) - m_k(s_k) \geq 0$ by condition (2):

$$1 - \rho_k = 1 - \frac{f(x_k) - f(\mathrm{R}_{x_k}(s_k))}{m_k(0) - m_k(s_k) + \frac{\varsigma_k}{3}\|s_k\|^3} \leq \frac{f(\mathrm{R}_{x_k}(s_k)) - m_k(s_k) + \frac{\varsigma_k}{3}\|s_k\|^3}{\frac{\varsigma_k}{3}\|s_k\|^3}.$$

Owing to A2, the numerator is upper bounded by $(L/6)\|s_k\|^3$. Hence, $1 - \rho_k \leq \frac{L}{2\varsigma_k}$. If $\varsigma_k \geq \frac{L}{2(1-\eta_2)}$, then $1 - \rho_k \leq 1 - \eta_2$ so that $\rho_k \geq \eta_2$, meaning step k is very successful. The regularization mechanism (5) then ensures $\varsigma_{k+1} \leq \varsigma_k$. Thus, ς_{k+1} may exceed ς_k only if $\varsigma_k < \frac{L}{2(1-\eta_2)}$, in which case it can grow at most to $\frac{L\gamma_3}{2(1-\eta_2)}$, but cannot grow beyond that level in later iterations. □

Proof of Lemma 3. Partition iterations $0, \ldots, \bar{k} - 1$ into successful or very successful ($\mathcal{S}_{\bar{k}}$) and unsuccessful ($\mathcal{U}_{\bar{k}}$) ones. Following the update mechanism (5), for $k \in \mathcal{S}_{\bar{k}}$, $\varsigma_{k+1} \geq \gamma_1 \varsigma_k$, while for $k \in \mathcal{U}_{\bar{k}}$, $\varsigma_{k+1} \geq \gamma_2 \varsigma_k$. Thus, by induction, $\varsigma_{\bar{k}} \geq \varsigma_0 \gamma_1^{|\mathcal{S}_{\bar{k}}|} \gamma_2^{|\mathcal{U}_{\bar{k}}|}$. By assumption, $\varsigma_{\bar{k}} \leq \varsigma_{\max}$ so that

$$\log\left(\frac{\varsigma_{\max}}{\varsigma_0}\right) \geq |\mathcal{S}_{\bar{k}}| \log(\gamma_1) + |\mathcal{U}_{\bar{k}}| \log(\gamma_2) = |\mathcal{S}_{\bar{k}}| \left[\log(\gamma_1) - \log(\gamma_2)\right] + \bar{k} \log(\gamma_2),$$

where we also used $|\mathcal{S}_{\bar{k}}| + |\mathcal{U}_{\bar{k}}| = \bar{k}$. Isolating $\bar{k}$ using $\gamma_2 > 1 > \gamma_1$ allows to conclude. □

B Proofs from Section 3: first-order analysis, exponentials

Certain tools from Riemannian geometry are useful throughout the appendices—see for example (O'Neill, 1983, pp59–67). To fix notation, let ∇ denote the Riemannian connection on $\mathcal{M}$ (not to be confused with ∇ and ∇^2 which denote gradient and Hessian of functions on linear spaces, such as pullbacks). With this notation, the Riemannian Hessian (Absil et al., 2008, Def. 5.5.1) is defined by $\mathrm{Hess} f = \nabla \mathrm{grad} f$. Furthermore, $\frac{\mathrm{D}}{\mathrm{d}t}$ denotes the covariant derivative of vector fields along curves on $\mathcal{M}$, induced by ∇. With this notation, given

a smooth curve $c\colon \mathbb{R} \to \mathcal{M}$, the intrinsic acceleration is defined as $c''(t) = \frac{\mathrm{D}^2}{\mathrm{d}t^2}c(t)$. For example, for a Riemannian submanifold of a Euclidean space, $c''(t)$ is obtained by orthogonal projection of the classical acceleration of c in the embedding space to the tangent space at $c(t)$. Geodesics are those curves which have zero intrinsic acceleration.

We first state and prove a partial version of Proposition 2 which applies for general retractions. Right after this, we prove Proposition 2. The purpose of this detour is to highlight how crucial properties of geodesics and of their interaction with parallel transports allow for the more direct guarantees of Section 3. In turn, this serves as motivation for the developments in Section 4.

Proposition 6 *Let $f\colon \mathcal{M} \to \mathbb{R}$ be twice differentiable on a Riemannian manifold $\mathcal{M}$ equipped with a retraction* R. *Given $(x, s) \in \mathrm{T}\mathcal{M}$, assume there exists $L \geq 0$ such that, for all $t \in [0, 1]$,*

$$\left\| P_{ts}^{-1}\big(\mathrm{Hess} f(c(t))[c'(t)]\big) - \mathrm{Hess} f(x)[s] \right\| \leq L\|s\| \cdot \ell(c|_{[0,t]}),$$

where P_{ts} is parallel transport along $c(t) = \mathrm{R}_x(ts)$ from $c(0)$ to $c(t)$ (note the retraction instead of the exponential) and $\ell(c|_{[0,t]}) = \int_0^t \|c'(\tau)\|\mathrm{d}\tau$ is the length of c restricted to the interval $[0, t]$. Then,

$$\left\| P_s^{-1}\mathrm{grad} f(\mathrm{R}_x(s)) - \mathrm{grad} f(x) - \mathrm{Hess} f(x)[s] \right\| \leq L\|s\| \int_0^1 \ell(c|_{[0,t]})\,\mathrm{d}t.$$

Proof. Pick a basis $v_1, \ldots, v_d$ for $\mathrm{T}_x\mathcal{M}$, and define the parallel vector fields $V_i(t) = P_{ts}(v_i)$ along $c(t)$. Since parallel transport is an isometry, $V_1(t), \ldots, V_d(t)$ form a basis for $\mathrm{T}_{c(t)}\mathcal{M}$ for each $t \in [0, 1]$. As a result, we can express the gradient of f along $c(t)$ in these bases,

$$\mathrm{grad} f(c(t)) = \sum_{i=1}^{d} \alpha_i(t) V_i(t), \tag{40}$$

with $\alpha_1(t), \ldots, \alpha_d(t)$ differentiable. Using properties of the Riemannian connection ∇ and its associated covariant derivative $\frac{\mathrm{D}}{\mathrm{d}t}$ (O'Neill, 1983, pp59–67), we find on one hand that

$$\frac{\mathrm{D}}{\mathrm{d}t}\mathrm{grad} f(c(t)) = \nabla_{c'(t)}\mathrm{grad} f = \mathrm{Hess} f(c(t))[c'(t)],$$

and on the other hand that

$$\frac{\mathrm{D}}{\mathrm{d}t}\sum_{i=1}^{d} \alpha_i(t) V_i(t) = \sum_{i=1}^{d} \alpha_i'(t) V_i(t) = P_{ts}\sum_{i=1}^{d} \alpha_i'(t) v_i,$$

where we used that $\frac{\mathrm{D}}{\mathrm{d}t}V_i(t) = 0$, by definition of parallel transport. Furthermore,

$$c'(t) = \mathrm{DR}_x(ts)[s] = T_{ts}(s),$$

where $T_{ts} = \mathrm{DR}_x(ts)$ is a linear operator from the tangent space at x to the tangent space at $c(t)$—just like P_{ts}. Combining, we deduce that

$$\sum_{i=1}^{d} \alpha_i'(t) v_i = \Big(P_{ts}^{-1} \circ \mathrm{Hess} f(c(t)) \circ T_{ts}\Big)[s].$$

Going back to (40), we also see that

$$G(t) \triangleq P_{ts}^{-1}\mathrm{grad} f(c(t)) = \sum_{i=1}^{d} \alpha_i(t) v_i$$

is a map from (a subset of) $\mathbb{R}$ to $\mathrm{T}_x\mathcal{M}$—two linear spaces—so that we can differentiate it in the usual way:

$$G'(t) = \sum_{i=1}^{d} \alpha_i'(t) v_i.$$

We conclude that

$$G'(t) = \frac{\mathrm{d}}{\mathrm{d}t}\Big[P_{ts}^{-1}\mathrm{grad}f(c(t))\Big] = \Big(P_{ts}^{-1} \circ \mathrm{Hess}f(c(t)) \circ T_{ts}\Big)[s]. \tag{41}$$

Since G' is continuous,

$$\begin{aligned} P_{ts}^{-1}\mathrm{grad}f(c(t)) = G(t) &= G(0) + \int_0^t G'(\tau)\mathrm{d}\tau \\ &= \mathrm{grad}f(x) + \int_0^t \big(P_{\tau s}^{-1} \circ \mathrm{Hess}f(c(\tau)) \circ T_{\tau s}\big)[s]\mathrm{d}\tau. \end{aligned}$$

Moving $\mathrm{grad}f(x)$ to the left-hand side and subtracting $\mathrm{Hess}f(x)[ts]$ on both sides, we find

$$P_{ts}^{-1}\mathrm{grad}f(c(t)) - \mathrm{grad}f(x) - \mathrm{Hess}f(x)[ts] = \int_0^t \big(P_{\tau s}^{-1} \circ \mathrm{Hess}f(c(\tau)) \circ T_{\tau s} - \mathrm{Hess}f(x)\big)[s]\mathrm{d}\tau.$$

Using the main assumption on $\mathrm{Hess}f$ along c, it easily follows that

$$\Big\|P_{ts}^{-1}\mathrm{grad}f(c(t)) - \mathrm{grad}f(x) - \mathrm{Hess}f(x)[ts]\Big\| \leq \|s\| L \int_0^t \ell(c|_{[0,\tau]})\mathrm{d}\tau. \tag{42}$$

For $t = 1$, this is the announced inequality. □

Proof of Proposition 2. In this proposition we work with the exponential retraction, so that instead of a general retraction curve $c(t)$ we work along a geodesic $\gamma(t) = \mathrm{Exp}_x(ts)$. By definition, the velocity vector field $\gamma'(t)$ of a geodesic $\gamma(t)$ is parallel, meaning

$$\gamma'(t) = P_{ts}(\gamma'(0)) = P_{ts}(s). \tag{43}$$

This elegant interplay of geodesics and parallel transport is crucial. In particular,

$$\ell(\gamma|_{[0,t]}) = \int_0^t \|\gamma'(\tau)\|\mathrm{d}\tau = t\|s\|,$$

and the condition in Proposition 6 becomes

$$\Big\|P_{ts}^{-1}(\mathrm{Hess}f(\gamma(t))[P_{ts}(s)]) - \mathrm{Hess}f(x)[s]\Big\| \leq tL\|s\|^2,$$

which is indeed guaranteed by our own assumptions. We deduce that (42) holds:

$$\Big\|P_{ts}^{-1}\mathrm{grad}f(\gamma(t)) - \mathrm{grad}f(x) - \mathrm{Hess}f(x)[ts]\Big\| \leq \|s\| L \int_0^t \ell(\gamma|_{[0,\tau]})\mathrm{d}\tau = \frac{L}{2}\|s\|^2 t^2. \tag{44}$$

The relation (43) also yields the scalar inequality. Indeed, since $f \circ \gamma \colon [0,1] \to \mathbb{R}$ is continuously differentiable,

$$\begin{aligned} f(\mathrm{Exp}_x(s)) = f(\gamma(1)) &= f(\gamma(0)) + \int_0^1 (f \circ \gamma)'(t)\mathrm{d}t \\ &= f(x) + \int_0^1 \big\langle \mathrm{grad}f(\gamma(t)), \gamma'(t)\big\rangle \,\mathrm{d}t \\ &= f(x) + \int_0^1 \Big\langle P_{ts}^{-1}\mathrm{grad}f(\gamma(t)), s\Big\rangle \,\mathrm{d}t, \end{aligned}$$

where on the last line we used (43) and the fact that P_{ts} is an isometry. For a general retraction curve $c(t)$, instead of s as the right-most term we would find $P_{ts}^{-1}(c'(t))$ which may vary with t: this would make the next step significantly more difficult. Move $f(x)$ to the left-hand side and subtract terms on both sides to get

$$f(\mathrm{Exp}_x(s)) - f(x) - \langle \mathrm{grad} f(x), s\rangle - \frac{1}{2}\langle s, \mathrm{Hess} f(x)[s]\rangle \\ = \int_0^1 \left\langle P_{ts}^{-1}\mathrm{grad} f(\gamma(t)) - \mathrm{grad} f(x) - \mathrm{Hess} f(x)[ts], s\right\rangle \mathrm{d}t.$$

Using (44) and Cauchy–Schwarz, it follows immediately that

$$\left| f(\mathrm{Exp}_x(s)) - f(x) - \langle s, \mathrm{grad} f(x)\rangle - \frac{1}{2}\langle s, \mathrm{Hess} f(x)[s]\rangle \right| \leq \int_0^1 \frac{L}{2}\|s\|^3 t^2 \mathrm{d}t = \frac{L}{6}\|s\|^3,$$

as announced. □

Next, we provide an argument for the last claim in Theorem 3.

Proof of Theorem 3. We argue that $\lim_{k\to\infty} \|\mathrm{grad} f(x_k)\| = 0$. The first claim of the theorem states that, for every $\varepsilon > 0$, there is a finite number of successful steps k such that x_{k+1} has gradient larger than ε. Thus, for any $\varepsilon > 0$, there exists K: the last successful step such that x_{K+1} has gradient larger than ε. Furthermore, there is a finite number of unsuccessful steps directly after $K+1$. Indeed, $\varsigma_{K+1} \geq \varsigma_{\min}$, and failures increase ς exponentially; additionally, ς cannot outgrow $\varsigma_{\max}$ by Lemma 2. Thus, after a finite number of failures, a new success arises, necessarily producing an iterate with gradient norm at most ε since K was the last successful step to produce a larger gradient. By the same argument, all subsequent iterates have gradient norm at most ε. In other words: for any $\varepsilon > 0$, there exists K' finite such that for all $k \geq K'$, $\|\mathrm{grad} f(x_k)\| \leq \varepsilon$, that is: $\lim_{k\to\infty} \|\mathrm{grad} f(x_k)\| = 0$. □

C Proofs from Section 5: second-order analysis

Proof of Corollary 3. Consider these subsets of the set of successful iterations $\mathcal{S}$:

$$\mathcal{S}^1 \triangleq \{k \in \mathcal{S} : \|\mathrm{grad} f(x_{k+1})\| > \varepsilon_g\}, \quad \text{and} \quad \mathcal{S}^2 \triangleq \{k \in \mathcal{S} : \lambda_{\min}(\mathrm{Hess} f(x_k)) < -\varepsilon_H\}.$$

These sets are finite: for $K_1 = K_1(\varepsilon_g)$ as provided by either Theorem 3 or Theorem 4, and for $K_2 = K_2(\varepsilon_H)$ as provided by Theorem 5, we know that

$$|\mathcal{S}^1| \leq K_1, \qquad \text{and} \qquad |\mathcal{S}^2| \leq K_2.$$

Note that successful steps are in one-to-one correspondence with the distinct points in the sequence of iterates $x_1, x_2, x_3, \ldots$[2] The first inequality states at most K_1 of the distinct points in that list have large gradient. The second inequality states at most K_2 of the distinct points in that same list have significantly negative Hessian eigenvalues. Thus, if more than $K_1 + K_2 + 1$ distinct points appear among $x_0, x_1, \ldots, x_{\bar{k}}$ (note the $+1$ as we added x_0 to the list), then at least one of these points has both a small gradient and an almost positive semidefinite Hessian. In particular, as long as the number of successful iterations among $0, \ldots, \bar{k}-1$ exceeds $K_1 + K_2 + 1$ (strictly), there must exist $k \in \{0, \ldots, \bar{k}\}$ such that

$$\|\mathrm{grad} f(x_k)\| \leq \varepsilon_g \qquad \text{and} \qquad \lambda_{\min}(\mathrm{Hess} f(x_k)) \geq -\varepsilon_H.$$

Lemma 3 allows to conclude. □

[2] This is true because the cost function is strictly decreasing when successful, so that any x_k can only be repeated in one contiguous subset of iterates. Hence, if k is a successful iteration, match it to x_{k+1} (this is why we omitted x_0 from the list.)

D Proofs from Section 6: regularity assumptions

Proof of Lemma 4. Since $\hat{f}$ is a real function on a linear space, standard calculus applies:

$$\hat{f}(s) - \left[\hat{f}(0) + \langle s, \nabla\hat{f}(0)\rangle + \frac{1}{2}\langle s, \nabla^2\hat{f}(0)[s]\rangle\right] = \int_0^1\int_0^1 t_1 \left\langle \left[\nabla^2\hat{f}(t_1t_2s) - \nabla^2\hat{f}(0)\right][s], s\right\rangle \mathrm{d}t_1\mathrm{d}t_2,$$

$$\nabla\hat{f}(s) - \left[\nabla\hat{f}(0) + \nabla^2\hat{f}(0)[s]\right] = \int_0^1 \left[\nabla^2\hat{f}(ts) - \nabla^2\hat{f}(0)\right][s]\,\mathrm{d}t.$$

Taking norms on both sides, by a triangular inequality to pass the norm through the integral and integrating respectively $t_1^2t_2$ and t, we find using our main assumption (27) that

$$\left|\hat{f}(s) - \left[\hat{f}(0) + \langle s, \nabla\hat{f}(0)\rangle + \frac{1}{2}\langle s, \nabla^2\hat{f}(0)[s]\rangle\right]\right| \leq \frac{1}{6}L\|s\|^3, \text{ and}$$

$$\left\|\nabla\hat{f}(s) - \left[\nabla\hat{f}(0) + \nabla^2\hat{f}(0)[s]\right]\right\| \leq \frac{1}{2}L\|s\|^2. \qquad \square$$

Proof of Lemma 5. For an arbitrary $\dot{s} \in \mathrm{T}_x\mathcal{M}$, consider the curve $c(t) = \mathrm{R}_x(s + t\dot{s})$, and let $g = f \circ c\colon \mathbb{R} \to \mathbb{R}$. We compute the derivatives of g in two different ways. On the one hand, $g(t) = \hat{f}(s + t\dot{s})$ so that

$$g'(t) = \mathrm{D}\hat{f}(s + t\dot{s})[\dot{s}] = \langle \nabla\hat{f}(s + t\dot{s}), \dot{s}\rangle,$$

$$g''(t) = \left\langle \frac{\mathrm{d}}{\mathrm{d}t}\nabla\hat{f}(s + t\dot{s}), \dot{s}\right\rangle = \langle \nabla^2\hat{f}(s + t\dot{s})[\dot{s}], \dot{s}\rangle.$$

On the other hand, $g(t) = f(c(t))$ so that, using properties of $\frac{\mathrm{D}}{\mathrm{d}t}$ (O'Neill, 1983, pp59–67):

$$g'(t) = \mathrm{D}f(c(t))[c'(t)] = \langle \mathrm{grad}f(c(t)), c'(t)\rangle,$$

$$\begin{aligned} g''(t) &= \frac{\mathrm{d}}{\mathrm{d}t}\left\langle (\mathrm{grad}f \circ c)(t)), c'(t)\right\rangle \\ &= \left\langle \nabla_{c'(t)}\mathrm{grad}f, c'(t)\right\rangle + \left\langle (\mathrm{grad}f \circ c)(t)), \frac{\mathrm{D}}{\mathrm{d}t}c'(t)\right\rangle \\ &= \left\langle \mathrm{Hess}f(c(t))[c'(t)], c'(t)\right\rangle + \left\langle \mathrm{grad}f(c(t)), c''(t)\right\rangle. \end{aligned}$$

Equating the different identities for $g'(t)$ and $g''(t)$ at $t = 0$ while using $c'(0) = T_s\dot{s}$, we find for all $\dot{s} \in \mathrm{T}_x\mathcal{M}$:

$$\langle \nabla\hat{f}(s), \dot{s}\rangle = \left\langle \mathrm{grad}f(\mathrm{R}_x(s)), T_s\dot{s}\right\rangle,$$

$$\langle \nabla^2\hat{f}(s)[\dot{s}], \dot{s}\rangle = \left\langle \mathrm{Hess}f(\mathrm{R}_x(s))[T_s\dot{s}], T_s\dot{s}\right\rangle + \left\langle \mathrm{grad}f(\mathrm{R}_x(s)), c''(0)\right\rangle.$$

The last term, $\langle \mathrm{grad}f(\mathrm{R}_x(s)), c''(0)\rangle$, is seen to be the difference of two quadratic forms in $\dot{s}$, so that it is itself a quadratic form in $\dot{s}$. This justifies the definition of W_s through polarization. The announced identities follow by identification. $\square$

Proof of Proposition 3. With $\mathrm{R}_x(s) = \frac{x+s}{\sqrt{1+\|s\|^2}}$, it is easy to derive

$$\begin{aligned} T_s\dot{s} \triangleq \mathrm{DR}_x(s)[\dot{s}] &= \left[\frac{1}{\sqrt{1+\|s\|^2}}I_n - \frac{1}{\sqrt{1+\|s\|^2}^3}(x+s)s^\top\right]\dot{s} \\ &= \frac{1}{\sqrt{1+\|s\|^2}}\left[I_n - \mathrm{R}_x(s)\mathrm{R}_x(s)^\top\right]\dot{s}, \end{aligned} \tag{45}$$

where we used $x^\top\dot{s} = 0$ in between the two steps to replace $s^\top$ with $(x+s)^\top$. The matrix between brackets is the orthogonal projector from $\mathbb{R}^n$ to $\mathrm{T}_{\mathrm{R}_x(s)}\mathcal{M}$. Thus, its singular values are upper bounded by 1. Since T_s is an operator on $\mathrm{T}_x\mathcal{M} \subset \mathbb{R}^n$,

$$\|T_s\|_{\mathrm{op}} \leq \frac{1}{\sqrt{1+\|s\|^2}} \leq 1.$$

This secures the first property with $c_1 = 1$.

For the second property, consider $U(t) = T_{ts}\dot{s}$ and

$$U'(t) \triangleq \frac{\mathrm{D}}{\mathrm{d}t}U(t) = \mathrm{Proj}_{c(t)}\frac{\mathrm{d}}{\mathrm{d}t}U(t),$$

where $\mathrm{Proj}_y(v) = v - y(y^\top v)$ is the orthogonal projector to $\mathrm{T}_y\mathcal{M}$ and $c(t) = \mathrm{R}_x(ts)$. Define $g(t) = \frac{1}{\sqrt{1+t^2\|s\|^2}}$. Then, from (45), we have

$$U(t) = \left[g(t)I_n - tg(t)^3(x+ts)s^\top\right]\dot{s}. \tag{46}$$

This is easily differentiated in the embedding space $\mathbb{R}^n$:

$$\frac{\mathrm{d}}{\mathrm{d}t}U(t) = \left[g'(t)I_n - (tg(t)^3)'(x+ts)s^\top - tg(t)^3 ss^\top\right]\dot{s}.$$

The projection at $c(t)$ zeros out the middle term, as it is parallel to $x + ts$. This offers a simple expression for $U'(t)$, where in the last equality we use $g'(t) = -tg(t)^3\|s\|^2$:

$$U'(t) = \mathrm{Proj}_{c(t)}\left(\left[g'(t)I_n - tg(t)^3 ss^\top\right]\dot{s}\right) = -tg(t)^3 \cdot \mathrm{Proj}_{c(t)}\left(\left[\|s\|^2 I_n + ss^\top\right]\dot{s}\right).$$

The norm can only decrease after projection, so that, for $t \in [0,1]$,

$$\|U'(t)\| \leq 2tg(t)^3\|s\|^2\|\dot{s}\|.$$

Let $h(t) = 2tg(t)^3\|s\|^2 = \frac{2t\|s\|^2}{(1+t^2\|s\|^2)^{1.5}}$. For $s = 0$, h is identically zero. Otherwise, h attains its maximum $h\left(t = \frac{1}{\sqrt{2}\|s\|}\right) = \frac{4\sqrt{3}}{9}\|s\|$. It follows that $\|U'(t)\| \leq c_2\|s\|\|\dot{s}\|$ for all $t \in [0,1]$ with $c_2 = \frac{4\sqrt{3}}{9}$.

Finally, we establish the last property. Given $s, \dot{s} \in \mathrm{T}_x\mathcal{M}$, consider $c(t) = \mathrm{R}_x(s + t\dot{s})$. Simple calculations yield:

$$c'(t) = \frac{\mathrm{d}}{\mathrm{d}t}c(t) = \frac{1}{\sqrt{1+\|s+t\dot{s}\|^2}}\left[\dot{s} - \langle \dot{s}, c(t)\rangle\, c(t)\right] = \frac{1}{\sqrt{1+\|s+t\dot{s}\|^2}}\mathrm{Proj}_{c(t)}\dot{s}. \tag{47}$$

This is indeed in the tangent space at $c(t)$. The classical derivative of $c'(t)$ is given by

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}c'(t) &= -\frac{1}{\sqrt{1+\|s+t\dot{s}\|^2}}\left[\langle \dot{s}, c'(t)\rangle\, c(t) + \langle \dot{s}, c(t)\rangle\, c'(t) + \frac{\langle s+t\dot{s}, \dot{s}\rangle}{1+\|s+t\dot{s}\|^2}\mathrm{Proj}_{c(t)}\dot{s}\right] \\ &= -\frac{1}{\sqrt{1+\|s+t\dot{s}\|^2}}\left[\langle \dot{s}, c'(t)\rangle\, c(t) + 2\frac{\langle s+t\dot{s}, \dot{s}\rangle}{1+\|s+t\dot{s}\|^2}\mathrm{Proj}_{c(t)}\dot{s}\right],\end{aligned}$$

where we used (47) and orthogonality of x and $\dot{s}$ in $\langle c(t), \dot{s}\rangle = \frac{1}{\sqrt{1+\|s+t\dot{s}\|^2}}\langle x + s + t\dot{s}, \dot{s}\rangle$. The acceleration of c is $c''(t) = \frac{\mathrm{D}}{\mathrm{d}t}c'(t) = \mathrm{Proj}_{c(t)}\left(\frac{\mathrm{d}}{\mathrm{d}t}c'(t)\right)$. The first term vanishes after projection, while the second term is unchanged. Overall,

$$c''(t) = -\frac{2\langle s+t\dot{s}, \dot{s}\rangle}{\sqrt{1+\|s+t\dot{s}\|^2}^3}\mathrm{Proj}_{c(t)}\dot{s} = -\frac{2\langle c(t), \dot{s}\rangle}{1+\|s+t\dot{s}\|^2}\mathrm{Proj}_{c(t)}\dot{s}. \tag{48}$$

In particular, $c''(0) = -2\frac{\langle s, \dot{s}\rangle}{\sqrt{1+\|s\|^2}^3}\mathrm{Proj}_{c(0)}\dot{s}$, so that $\|c''(0)\| \leq 2\min(\|s\|, 0.4)\|\dot{s}\|^2$ and the property holds with $c_3 = 2$. (Peculiarly, if s and $\dot{s}$ are orthogonal, $c''(0) = 0$.) □

In order to prove Theorem 6, we introduce two supporting lemmas (needed only for the case where $\mathcal{M}$ is not compact) and one key lemma. The first lemma below is similar in spirit to (Cartis et al., 2011b, Lem. 2.2).

Lemma 6 *Let $f\colon \mathcal{M} \to \mathbb{R}$ be twice continuously differentiable. Let $\{(x_0, s_0), (x_1, s_1), \ldots\}$ be the points and steps generated by Algorithm 1. Each step has norm bounded as:*

$$\|s_k\| \leq \sqrt{\frac{3\|\nabla \hat{f}_k(0)\|}{\varsigma_{\min}}} + \frac{3}{2\varsigma_{\min}} \max\left(0, -\lambda_{\min}(\nabla^2 \hat{f}_k(0))\right), \tag{49}$$

where $\hat{f}_k = f \circ \mathrm{R}_{x_k}$ is the pullback, as in (6).

Proof. Owing to the first-order progress condition (2), using Cauchy–Schwarz and the fact that $\varsigma_k \geq \varsigma_{\min}$ for all k by design of the algorithm, we find

$$\begin{aligned}\varsigma_{\min}\|s_k\|^3 \leq \varsigma_k\|s_k\|^3 &\leq -3\left\langle s_k, \nabla \hat{f}_k(0) + \frac{1}{2}\nabla^2 \hat{f}_k(0)[s_k]\right\rangle \\ &\leq 3\|s_k\|\left(\|\nabla \hat{f}_k(0)\| + \frac{1}{2}\max\left(0, -\lambda_{\min}(\nabla^2 \hat{f}_k(0))\right)\|s_k\|\right).\end{aligned}$$

This defines a quadratic inequality in $\|s_k\|$:

$$\varsigma_{\min}\|s_k\|^2 - h_k\|s_k\| - g_k \leq 0,$$

where to simplify notation we let $h_k = \frac{3}{2}\max(0, -\lambda_{\min}(\nabla^2 \hat{f}_k(0)))$ and $g_k = 3\|\nabla \hat{f}_k(0)\|$. Since $\|s_k\|$ must lie between the two roots of this quadratic, we know in particular that

$$\|s_k\| \leq \frac{h_k + \sqrt{h_k^2 + 4\varsigma_{\min} g_k}}{2\varsigma_{\min}} \leq \frac{h_k + \sqrt{\varsigma_{\min} g_k}}{\varsigma_{\min}},$$

where in the last step we used $\sqrt{u+v} \leq \sqrt{u} + \sqrt{v}$ for any $u, v \geq 0$. □

Lemma 7 *Let $f\colon \mathcal{M} \to \mathbb{R}$ be twice continuously differentiable. Let $\{(x_0, s_0), (x_1, s_1), \ldots\}$ be the points and steps generated by Algorithm 1. Consider the following subset of $\mathcal{M}$, obtained by collecting all curves generated by retracted steps (both accepted and rejected):*

$$\mathcal{N} = \bigcup_k \{\mathrm{R}_{x_k}(ts_k) : t \in [0, 1]\}. \tag{50}$$

If the sequence $\{x_0, x_1, x_2, \ldots\}$ remains in a compact subset of $\mathcal{M}$, then $\mathcal{N}$ is included in a compact subset of $\mathcal{M}$.

Proof. If $\mathcal{M}$ is compact, the claim is clear since $\mathcal{N} \subseteq \mathcal{M}$. Otherwise, we use Lemma 6. Specifically, considering the upper bound in that lemma, define

$$\alpha(x) = \sqrt{\frac{3\|\nabla \hat{f}_x(0)\|}{\varsigma_{\min}}} + \frac{3}{2\varsigma_{\min}} \max\left(0, -\lambda_{\min}(\nabla^2 \hat{f}_x(0))\right),$$

where $\hat{f}_x = f \circ \mathrm{R}_x$. This is a continuous function of x, and $\|s_k\| \leq \alpha(x_k)$. Since by assumption $\{x_0, x_1, \ldots\} \subseteq \mathcal{K}$ with $\mathcal{K}$ compact, we find that

$$\forall k, \quad \|s_k\| \leq \sup_{k'} \alpha(x_{k'}) \leq \max_{x \in \mathcal{K}} \alpha(x) \triangleq r,$$

where r is a finite number. Consider the following subset of the tangent bundle $\mathrm{T}\mathcal{M}$:

$$\mathcal{K}' = \{(x, s) \in \mathrm{T}\mathcal{M} : x \in \mathcal{K}, \|s\|_x \leq r\}.$$

Since $\mathcal{K}$ is compact, $\mathcal{K}'$ is compact. Furthermore, since the retraction is a continuous map, $\mathrm{R}(\mathcal{K}')$ is compact, and it contains $\mathcal{N}$. □

Lemma 8 *Let $f\colon \mathcal{M} \to \mathbb{R}$ be three times continuously differentiable, and consider the points and steps $\{(x_0, s_0), (x_1, s_1), \ldots\}$ generated by Algorithm 1. Assume the retraction is second-order nice on this set (see Definition 4). If the set $\mathcal{N}$ as defined by (50) is contained in a compact set $\mathcal{K}$, then A2 and A4 are satisfied.*

Proof. For some k and $\bar{t} \in [0, 1]$, let $(x, s) = (x_k, \bar{t}s_k)$ and define the pullback $\hat{f} = f \circ \mathrm{R}_x$. Notice in particular that $\mathrm{R}_x(s) \in \mathcal{N} \subseteq \mathcal{K}$. Combine the expression for the Hessian of the pullback (29) with (27) to get:

$$\left\|\nabla^2 \hat{f}(s) - \nabla^2 \hat{f}(0)\right\|_{\mathrm{op}} \leq \|T_s^* \circ \mathrm{Hess} f(\mathrm{R}_x(s)) \circ T_s - \mathrm{Hess} f(x)\|_{\mathrm{op}} + \|W_s - W_0\|_{\mathrm{op}}\,.$$

By definition of W_s (31), using the third condition on the retraction, we find that $W_0 = 0$ and

$$\|W_s\|_{\mathrm{op}} = \max_{\substack{\dot{s} \in \mathrm{T}_x\mathcal{M} \\ \|\dot{s}\| \leq 1}} |\langle W_s[\dot{s}], \dot{s}\rangle| \leq \|\mathrm{grad} f(\mathrm{R}_x(s))\| \cdot \max_{\substack{\dot{s} \in \mathrm{T}_x\mathcal{M} \\ \|\dot{s}\| \leq 1}} \|c''(0)\| \leq c_3 G \|s\|,$$

where $G = \max_{y \in \mathcal{K}} \|\mathrm{grad} f(y)\|$ is finite by compactness of $\mathcal{K}$ and continuity of the gradient norm. Thus, it remains to show that

$$\|T_s^* \circ \mathrm{Hess} f(\mathrm{R}_x(s)) \circ T_s - \mathrm{Hess} f(x)\|_{\mathrm{op}} \leq c' \|s\|$$

for some constant c'. For an arbitrary $\dot{s} \in \mathrm{T}_x\mathcal{M}$, owing to differentiability properties of f,

$$\left\langle \left[T_s^* \circ \mathrm{Hess} f(\mathrm{R}_x(s)) \circ T_s - \mathrm{Hess} f(x)\right][\dot{s}], \dot{s}\right\rangle = \int_0^1 \frac{\mathrm{d}}{\mathrm{d}t} \left\langle T_{ts}^* \circ \mathrm{Hess} f(\mathrm{R}_x(ts)) \circ T_{ts}[\dot{s}], \dot{s}\right\rangle \mathrm{d}t. \tag{51}$$

We aim to upper bound the above by $c'\|s\|\|\dot{s}\|^2$. Consider the curve $c(t) = \mathrm{R}_x(ts)$ and a tangent vector field $U(t) = T_{ts}\dot{s}$ along c. Then, define

$$\begin{aligned} h(t) &= \langle T_{ts}^* \circ \mathrm{Hess} f(c(t)) \circ T_{ts}[\dot{s}], \dot{s}\rangle \\ &= \langle \mathrm{Hess} f(c(t))[T_{ts}\dot{s}], T_{ts}\dot{s}\rangle \\ &= \langle \mathrm{Hess} f(c(t))[U(t)], U(t)\rangle\,. \end{aligned}$$

The integrand in (51) is the derivative of the real function h:

$$\begin{aligned} h'(t) &= \frac{\mathrm{d}}{\mathrm{d}t} \langle \mathrm{Hess} f(c(t))[U(t)], U(t)\rangle \\ &= \left\langle \frac{\mathrm{D}}{\mathrm{d}t}\Big[\mathrm{Hess} f(c(t))[U(t)]\Big], U(t)\right\rangle + \left\langle \mathrm{Hess} f(c(t))[U(t)], \frac{\mathrm{D}}{\mathrm{d}t} U(t)\right\rangle \\ &= \left\langle \left(\nabla_{c'(t)} \mathrm{Hess} f\right)[U(t)], U(t)\right\rangle + 2\left\langle \mathrm{Hess} f(c(t))[U(t)], U'(t)\right\rangle, \end{aligned}$$

where $U'(t) \triangleq \frac{\mathrm{D}}{\mathrm{d}t} U(t)$ and we used that the Hessian is symmetric. Here, $\nabla_{c'(t)} \mathrm{Hess} f$ is the Levi–Civita derivative of the Hessian tensor field at $c(t)$ along $c'(t)$—see (do Carmo, 1992, Def. 4.5.7, p102) for the notion of derivative of a tensor field. For every t, the latter is a symmetric linear operator on the tangent space at $c(t)$. By Cauchy–Schwarz,

$$|h'(t)| \leq \|\nabla_{c'(t)} \mathrm{Hess} f\|_{\mathrm{op}} \|U(t)\|^2 + 2\|\mathrm{Hess} f(c(t))\|_{\mathrm{op}} \|U(t)\| \|U'(t)\|.$$

By compactness of $\mathcal{K}$ and continuity of the Hessian, we can define

$$H = \max_{y \in \mathcal{K}} \|\mathrm{Hess} f(y)\|_{\mathrm{op}}.$$

By linearity of the connection ∇, if $c'(t) \neq 0$,

$$\nabla_{c'(t)} \mathrm{Hess} f = \|c'(t)\| \cdot \nabla_{\frac{c'(t)}{\|c'(t)\|}} \mathrm{Hess} f.$$

Furthermore, $c'(t) = T_{ts}s$ has norm bounded by the first assumption on the retraction: $\|c'(t)\| \leq c_1\|s\|$. Thus, in all cases, by compactness of $\mathcal{K}$ and continuity of the function $v \to \nabla_v \mathrm{Hess} f$ on the tangent bundle $\mathrm{T}\mathcal{M}$, there is a finite J as follows:

$$\|\nabla_{c'(t)}\mathrm{Hess} f\|_{\mathrm{op}} \leq c_1\|s\| \cdot \overbrace{\max_{\substack{y\in\mathcal{K}, v\in \mathrm{T}_y\mathcal{M} \\ \|v\|\leq 1}} \|\nabla_v \mathrm{Hess} f\|_{\mathrm{op}}}^{J}.$$

Of course, $\|U(t)\| \leq c_1\|\dot{s}\|$. Finally, we bound $\|U'(t)\|$ using the second property of the retraction: $\|U'(t)\| \leq c_2\|s\|\|\dot{s}\|$. Collecting what we learned about $|h'(t)|$ and injecting in (51),

$$\left|\left\langle \left[T_s^* \circ \mathrm{Hess} f(\mathrm{R}_x(s)) \circ T_s - \mathrm{Hess} f(x)\right][\dot{s}], \dot{s}\right\rangle\right| \leq \int_0^1 |h'(t)|\mathrm{d}t \leq \left[c_1^3 J + 2c_1c_2H\right]\|s\|\|\dot{s}\|^2.$$

Finally, it follows from Lemma 4 that A2 and A4 hold with $L = L' = c_3G + 2c_1c_2H + c_1^3J$ and $q \equiv 0$. We note in closing that the constants G, H, J can be related to the Lipschitz properties of f, $\mathrm{grad} f$ and $\mathrm{Hess} f$, respectively. □

The theorem we wanted to prove now follows as a direct corollary.

Proof of Theorem 6. For the main result, simply combine Lemmas 7 and 8. To support the closing statement, it is sufficient to verify that Algorithm 1 is a descent method owing to the step acceptance mechanism and the first part of condition (2). □

E Proofs from Section 7: differential of retraction

Stiefel manifold

Proposition 4 regarding the Stiefel manifold is a corollary of the following statement.

Lemma 9 *For the Stiefel manifold $\mathcal{M} = \mathrm{St}(n, p)$ with the Q-factor retraction R, for all $X \in \mathcal{M}$ and $S \in \mathrm{T}_X\mathcal{M}$,*

$$\sigma_{\min}(\mathrm{DR}_X(S)) \geq 1 - 3\|S\|_{\mathrm{F}} - \frac{1}{2}\|S\|_{\mathrm{F}}^2,$$

where $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm. Moreover, for the special case $p = 1$ (the unit sphere in $\mathbb{R}^n$), the retraction reduces to $\mathrm{R}_x(s) = \frac{x+s}{\|x+s\|}$ and we have for all $x \in \mathcal{M}, s \in \mathrm{T}_x\mathcal{M}$:

$$\sigma_{\min}(\mathrm{DR}_x(s)) = \frac{1}{1 + \|s\|^2}.$$

Proof. Let $X \in \mathrm{St}(n, p)$ and $S \in \mathrm{T}_X\mathrm{St}(n, p) = \{\dot{X} \in \mathbb{R}^{n\times p} : \dot{X}^\top X + X^\top \dot{X} = 0\}$ be fixed. Define Q, R as the thin QR-decomposition of $X + S$, that is, Q is an $n \times p$ matrix with orthonormal columns and R is a $p \times p$ upper triangular matrix with positive diagonal entries such that $X + S = QR$: this decomposition exists and is unique since $X + S$ has full column rank, as shown below (53). By definition, we have that $\mathrm{R}_X(S) = Q$.

For a matrix M, define $\mathrm{tril}(M)$ as the lower triangular portion of the matrix M, that is, $\mathrm{tril}(M)_{ij} = M_{ij}$ if $i \geq j$ and 0 otherwise. Further define $\rho_{\mathrm{skew}}(M)$ as

$$\rho_{\mathrm{skew}}(M) \triangleq \mathrm{tril}(M) - \mathrm{tril}(M)^\top.$$

As derived in (Absil et al., 2008, Ex. 8.1.5) (see also the erratum for the reference) we have a formula for the directional derivative of the retraction along any $Z \in \mathrm{T}_X\mathrm{St}(n, p)$:

$$\mathrm{DR}_X(S)[Z] = Q\rho_{\mathrm{skew}}(Q^\top ZR^{-1}) + (I - QQ^\top)ZR^{-1}. \tag{52}$$

We first confirm that R is always invertible. To see this, note that S being tangent at X means $S^\top X + X^\top S = 0$ and therefore

$$R^\top R = \underbrace{(X+S)^\top (X+S)}_{\text{start reading here}} = X^\top X + X^\top S + S^\top X + S^\top S = I_p + S^\top S, \tag{53}$$

which shows R is invertible. Moreover the above expression also implies that:

$$\sigma_k(R) = \sigma_k(X+S) = \sqrt{\lambda_k((X+S)^\top (X+S))} = \sqrt{1+\lambda_k(S^\top S)} = \sqrt{1+\sigma_k(S)^2},$$

where $\sigma_k(M)$ represents the kth singular value of M and λ_k likewise extracts the kth eigenvalue (in decreasing order for symmetric matrices). In particular we have that

$$\begin{aligned}
\sigma_{\min}(R^{-1}) &= \frac{1}{\sqrt{1+\sigma_{\max}(S)^2}} \geq \frac{1}{\sqrt{1+\|S\|_{\mathrm{F}}^2}} \geq 1 - \frac{1}{2}\|S\|_{\mathrm{F}}^2, \\
\sigma_{\max}(R^{-1}) &= \frac{1}{\sqrt{1+\sigma_{\min}(S)^2}} \leq 1.
\end{aligned} \tag{54}$$

Further note that since $QR = X + S$, we have that $Q = (X+S)R^{-1}$ and therefore

$$\begin{aligned}
Q^\top Z R^{-1} &= (R^{-1})^\top (X+S)^\top Z R^{-1} \\
&= (R^{-1})^\top X^\top Z R^{-1} + (R^{-1})^\top S^\top Z R^{-1}.
\end{aligned}$$

The first term above is always skew-symmetric since Z is tangent at X, so that $X^\top Z + Z^\top X = 0$. Furthermore, for any skew-symmetric matrix M, $\rho_{\mathrm{skew}}(M) = M$. Therefore, using (52),

$$\begin{aligned}
\mathrm{DR}_X(S)[Z] &= Q\rho_{\mathrm{skew}}(Q^\top Z R^{-1}) + (I - QQ^\top) Z R^{-1} \\
&= Q\left(\rho_{\mathrm{skew}}(Q^\top Z R^{-1}) - Q^\top Z R^{-1}\right) + Z R^{-1} \\
&= Q\left(\rho_{\mathrm{skew}}((R^{-1})^\top S^\top Z R^{-1}) - (R^{-1})^\top S^\top Z R^{-1}\right) + Z R^{-1},
\end{aligned} \tag{55}$$

where in the last step we used $XR^{-1} - Q = -SR^{-1}$. Further note that for any matrix M of size $p \times p$,

$$\|Q(\rho_{\mathrm{skew}}(M) - M)\|_{\mathrm{F}} = \|\mathrm{tril}(M) - \mathrm{tril}(M)^\top - M\|_{\mathrm{F}} \leq 3\|M\|_{\mathrm{F}}. \tag{56}$$

Hence, we have that,

$$\begin{aligned}
\|\mathrm{DR}_X(S)[Z]\|_{\mathrm{F}} &\geq \|Z R^{-1}\|_{\mathrm{F}} - 3\|(R^{-1})^\top S^\top Z R^{-1}\|_{\mathrm{F}} \\
&\geq \|Z\|_{\mathrm{F}} \left(\sigma_{\min}(R^{-1}) - 3\sigma_{\max}(R^{-1})^2 \sigma_{\max}(S)\right),
\end{aligned} \tag{57}$$

where we have used $\|A\|_{\mathrm{F}}\sigma_{\min}(B) \leq \|AB\|_{\mathrm{F}} \leq \|A\|_{\mathrm{F}}\sigma_{\max}(B)$ multiple times. Using the bounds on the singular values of R^{-1} (derived in (54)) we get that

$$\|\mathrm{DR}_X(S)[Z]\|_{\mathrm{F}} \geq \|Z\|_{\mathrm{F}} \left(1 - \frac{1}{2}\|S\|_{\mathrm{F}}^2 - 3\|S\|_{\mathrm{F}}\right).$$

Since this holds for all tangent vectors Z, we get that

$$\sigma_{\min}(\mathrm{DR}_X(S)) \geq 1 - 3\|S\|_{\mathrm{F}} - \frac{1}{2}\|S\|_{\mathrm{F}}^2.$$

To prove a better bound for the case of $p = 1$ (the sphere), we improve the analysis of the expression derived in (55). Note that for $p = 1$, the matrix inside the ρ_{skew} operator is a scalar, whose skew-symmetric part is necessarily zero. Also note that Q is a single column

matrix with value $\frac{x+s}{\|x+s\|}$ and $R = \|x+s\|$. Also, $X^\top S X^\top Z = 0$ since S, Z are tangent. Therefore,

$$\begin{aligned}
\mathrm{DR}_X(S)[Z] &= ZR^{-1} - Q(R^{-1})^\top S^\top Z R^{-1} \\
&= \frac{1}{\|x+s\|}\left(z - \frac{s^\top z}{1+\|s\|^2}(x+s)\right) \\
&= \frac{1}{\|x+s\|}\left(z - \frac{s^\top z}{1+\|s\|^2}s - \frac{s^\top z}{1+\|s\|^2}x\right).
\end{aligned}$$

Since x is orthogonal to s and z,

$$\begin{aligned}
\|\mathrm{DR}_x(s)[z]\|^2 &= \frac{1}{1+\|s\|^2}\left(\left\|z - \frac{s^\top z}{1+\|s\|^2}s\right\|^2 + \left(\frac{s^\top z}{1+\|s\|^2}\right)^2\right) \\
&= \frac{1}{1+\|s\|^2}\left(\|z\|^2 - 2\frac{(s^\top z)^2}{1+\|s\|^2} + \left(\frac{s^\top z}{1+\|s\|^2}\right)^2(1+\|s\|^2)\right) \\
&= \frac{1}{1+\|s\|^2}\left(\|z\|^2 - \frac{(s^\top z)^2}{1+\|s\|^2}\right) \\
&\geq \|z\|^2\frac{1}{1+\|s\|^2}\left(1 - \frac{\|s\|^2}{1+\|s\|^2}\right) \\
&= \|z\|^2\frac{1}{(1+\|s\|^2)^2}.
\end{aligned}$$

The worst-case scenario is achieved when z and s are aligned. Overall, we get

$$\|\mathrm{DR}_x(s)[z]\| \geq \|z\|\frac{1}{1+\|s\|^2},$$

which establishes the bound for the sphere. □

Differential of exponential map for manifolds with bounded curvature

Proposition 5 regarding the differential of the exponential map on complete manifolds with bounded sectional curvature follows as a corollary of the following statement.

Lemma 10 *Assume all sectional curvatures of $\mathcal{M}$, complete, are bounded above by C:*

If $C \leq 0$, then $\sigma_{\min}(\mathrm{DExp}_x(s)) = 1$;
If $C = \frac{1}{R^2} > 0$ and $\|s\| \leq \pi R$, then $1 \geq \sigma_{\min}(\mathrm{DExp}_x(s)) \geq \frac{\sin(\|s\|/R)}{\|s\|/R}$.

As usual, we use the convention $\sin(t)/t = 1$ at $t = 0$.

Proof. This results from a combination of few standard facts in Riemannian geometry:

1. (Lee, 2018, Prop. 10.10) Given any two tangent vectors $s, \dot{s} \in \mathrm{T}_x\mathcal{M}$, $J(t) = \mathrm{DExp}_x(ts)[t\dot{s}]$ is the unique Jacobi field along the geodesic $\gamma(t) = \mathrm{Exp}_x(ts)$ satisfying $J(0) = 0$ and $\frac{\mathrm{D}}{\mathrm{d}t}J(0) = \dot{s}$.
2. In particular, if $\dot{s} = \alpha s$ for some $\alpha \in \mathbb{R}$ so that $\dot{s}$ and s are parallel, then

$$J(t) = \mathrm{DExp}_x(ts)][t\dot{s}] = \frac{\mathrm{d}}{\mathrm{d}q}\mathrm{Exp}_x(ts + qt\dot{s})\Big|_{q=0} = \frac{\mathrm{d}}{\mathrm{d}q}\gamma(t + q\alpha t)\Big|_{q=0} = \alpha t\gamma'(t) = tP_{ts}(\dot{s}),$$

using $\gamma'(t) = P_{ts}(s)$. It remains to understand the case where $\dot{s}$ is orthogonal to s.

3. (Lee, 2018, Prop. 10.12) If $\mathcal{M}$ has constant sectional curvature C, $\|s\| = 1$ and $\langle s, \dot{s} \rangle = 0$, the Jacobi field above is given by:

$$J(t) = s_C(t) P_{ts}(\dot{s}),$$

where P_{ts} denotes parallel transport along γ as in (13) and

$$s_C(t) = \begin{cases} t & \text{if } C = 0, \\ R \sin(t/R) & \text{if } C = \frac{1}{R^2} > 0, \text{ and} \\ R \sinh(t/R) & \text{if } C = -\frac{1}{R^2}. \end{cases}$$

This can be reparameterized to allow for $\|s\| \neq 1$. Evaluating at $t = 1$ and using linearity in $\dot{s}$, we find for any $s, \dot{s} \in \mathrm{T}_x\mathcal{M}$ that

$$\mathrm{DExp}_x(s)[\dot{s}] = P_s\left(\dot{s}_{\parallel} + \frac{s_C(\|s\|)}{\|s\|}\dot{s}_{\perp}\right), \tag{58}$$

where $\dot{s}_{\perp}$ is the part of $\dot{s}$ which is orthogonal to s and $\dot{s}_{\parallel}$ is the part of $\dot{s}$ which is parallel to s—this corresponds to expression (20). By isometry of parallel transport, it is a simple exercise in linear algebra to deduce that

$$\sigma_{\min}(\mathrm{DExp}_x(s)) = \min\left(1, \frac{s_C(\|s\|)}{\|s\|}\right).$$

4. (Lee, 2018, Thm. 11.9(a)) Consider the case where $\dot{s}$ is orthogonal to s of unit norm once again: the Jacobi field comparison theorem states that if the sectional curvatures of $\mathcal{M}$ are upper-bounded by C, then $\|J(t)\|$ is at least as large as what it would be if $\mathcal{M}$ had constant sectional curvature C—with the additional condition that $\|s\| \leq \pi R$ if $C = 1/R^2 > 0$. This leads to the conclusion through similar developments as above, using also (Lee, 2018, Prop. 10.7) to separate the components of $J(t)$ that are parallel or orthogonal to $\gamma'(t)$. □

Extending to general retractions

In order to prove Theorem 7, we first introduce a result from topology. We follow Bergé (1963), including the blanket assumption that all encountered topological spaces are Hausdorff (page 65 in that reference)—this is the case for us so long as the topology of $\mathcal{M}$ itself is Hausdorff, which most authors require as part of the definition of a smooth manifold. Products of topological spaces are equipped with the product topology. Neighborhoods are open. A *correspondence* $\Gamma\colon Y \to Z$ maps points in Y to subsets of Z.

Definition 5 (Upper semicontinuous (u.s.c.) mapping) A correspondence $\Gamma\colon Y \to Z$ between two topological spaces Y, Z is a *u.s.c. mapping* if, for all y in Y, $\Gamma(y)$ is a compact subset of Z and, for any neighborhood V of $\Gamma(y)$, there exists a neighborhood U of y such that, for all $u \in U$, $\Gamma(u) \subseteq V$.

Theorem 8 (Bergé (1963, Thm. VI.2, p116)) *If ϕ is an upper semicontinuous, real-valued function in $Y \times Z$ and Γ is a u.s.c. mapping of Y into Z (two topological spaces) such that $\Gamma(y)$ is nonempty for each y, then the real-valued function M defined by*

$$M(y) = \max_{z \in \Gamma(y)} \phi(y, z)$$

is upper semicontinuous. (Under the assumptions, the maximum is indeed attained.)

We use the above theorem to establish our result. Manifolds (including tangent bundles) are equipped with the natural topology inherited from their smooth structure.

Proof of Theorem 7. It is sufficient to show that the function

$$t(r) = \inf_{(x,s)\in \mathrm{T}\mathcal{M}: x\in\mathcal{U}, \|s\|_x \leq r} \sigma_{\min}(\mathrm{DR}_x(s)) \tag{59}$$

is lower semicontinuous from $\mathbb{R}^+ = \{r \in \mathbb{R} : r \geq 0\}$ to $\mathbb{R}$, with respect to their usual topologies. Indeed, $t(0) = 1$ owing to the fact that $\mathrm{DR}_x(0)$ is the identity map for all x, and t being lower semicontinuous means that it cannot "jump down". Explicitly, lower semicontinuity at $r = 0$ implies that, for all $\delta > 0$, there exists $a > 0$ such that for all $r \leq a$ we have $t(r) \geq t(0) - \delta = 1 - \delta \triangleq b$.

To this end, consider the correspondence $\Gamma\colon \mathbb{R}^+ \to \mathrm{T}\mathcal{M}$ defined by

$$\Gamma(r) = \{(x, s) \in \mathrm{T}\mathcal{M} : x \in \mathcal{U} \text{ and } \|s\|_x \leq r\}. \tag{60}$$

Further consider the function $\phi\colon \mathbb{R}^+ \times \mathrm{T}\mathcal{M} \to \mathbb{R}$ defined by $\phi(r, (x, s)) = -\sigma_{\min}(\mathrm{DR}_x(s))$. Then, $t(r) = -M(r)$, where

$$M(r) = \sup_{(x,s)\in\Gamma(r)} \phi(r, (x, s)). \tag{61}$$

Thus, we must show M is upper semicontinuous. By Theorem 8, this is the case if

1. ϕ is upper semicontinuous,
2. $\Gamma(r)$ is nonempty and compact for all $r \geq 0$, and
3. For any $r \geq 0$ and any neighborhood $\mathcal{V}$ of $\Gamma(r)$ in $\mathrm{T}\mathcal{M}$, there exists a neighborhood I of r in $\mathbb{R}^+$ such that, for all $r' \in I$, we have $\Gamma(r') \subseteq \mathcal{V}$.

The first condition holds a fortiori since ϕ is continuous, owing to smoothness of $\mathrm{R}\colon \mathrm{T}\mathcal{M} \to \mathcal{M}$. The second condition holds since $\mathcal{U}$ is nonempty and compact. For the third condition, we show in Lemma 11 below that there exists a continuous function $\Delta\colon \mathcal{U} \to \mathbb{R}$ (continuous with respect to the subspace topology) such that $\{(x, s) \in \mathrm{T}\mathcal{M} : x \in \mathcal{U} \text{ and } \|s\|_x \leq \Delta(x)\} \subseteq \mathcal{V}$ and $\Delta(x) > r$ for all $x \in \mathcal{U}$ (if $\mathcal{M}$ is not connected, apply the lemma to each connected component which intersects with $\mathcal{U}$). As a result, $\min_{x\in\mathcal{U}} \Delta(x) = r + \varepsilon$ for some $\varepsilon > 0$ (using $\mathcal{U}$ compact), and $\Gamma(r + \varepsilon)$ is included in $\mathcal{V}$. We conclude that $I = [0, r + \varepsilon)$ is a suitable neighborhood of r to verify the condition. □

We now state and prove the last piece of the puzzle, which applies above with $r(x)$ constant ($L = 0$). Although the context is quite different, the first part of the proof is inspired by that of the tubular neighborhood theorem in (Lee, 2018, Thm. 5.25).

Lemma 11 *Let $\mathcal{U}$ be any subset of a connected Riemannian manifold $\mathcal{M}$ and let $r\colon \mathcal{U} \to \mathbb{R}^+$ be L-Lipschitz continuous with respect to the Riemannian distance* dist *on $\mathcal{M}$, that is,*

$$\forall x, x' \in \mathcal{U}, \quad |r(x) - r(x')| \leq L\mathrm{dist}(x, x').$$

Consider this subset of the tangent bundle:

$$\{(x, s) \in \mathrm{T}\mathcal{M} : x \in \mathcal{U} \text{ and } \|s\|_x \leq r(x)\}.$$

For any neighborhood $\mathcal{V}$ of this set in $\mathrm{T}\mathcal{M}$, there exists an $(L + 1)$-Lipschitz continuous function $\Delta\colon \mathcal{U} \to \mathbb{R}^+$ such that $\Delta(x) > r(x)$ for all $x \in \mathcal{U}$ and

$$\{(x, s) \in \mathrm{T}\mathcal{M} : x \in \mathcal{U} \text{ and } \|s\|_x \leq \Delta(x)\} \subseteq \mathcal{V}.$$

Proof. Consider the following open subsets of the tangent bundle, defined for each $x \in \mathcal{M}$ and $\delta \in \mathbb{R}$:

$$V_\delta(x) = \left\{(x', s') \in \mathrm{T}\mathcal{M} : \mathrm{dist}(x, x') < \delta - r(x) \text{ and } \|s'\|_{x'} < \delta\right\}.$$

Referring to these sets, define the function $\Delta\colon \mathcal{U} \to \mathbb{R}$ as:

$$\Delta(x) = \sup\{\delta \in \mathbb{R} : V_\delta(x) \subseteq \mathcal{V}\}.$$

This is well defined since $V_{r(x)}(x) = \emptyset$, so that $\Delta(x) \geq r(x)$ for all x. If $\Delta(x) = \infty$ for some x, then $\mathcal{V} = \mathrm{T}\mathcal{M}$ and the claim is clear (for example, redefine $\Delta(x) = r(x) + 1$ for all x). Thus, we assume $\Delta(x)$ finite for all x. The rest of the proof is in two parts.

Step 1: Δ is Lipschitz continuous. Pick $x, x' \in \mathcal{U}$, arbitrary. We must show

$$\Delta(x) - \Delta(x') \leq (L+1)\mathrm{dist}(x, x').$$

Then, by reversing the roles of x and x', we get $|\Delta(x) - \Delta(x')| \leq (L+1)\mathrm{dist}(x, x')$, as desired. If $\Delta(x) \leq (L+1)\mathrm{dist}(x, x')$, the claim is clear since $\Delta(x') \geq 0$. Thus, we now assume $\Delta(x) > (L+1)\mathrm{dist}(x, x')$. Define $\delta = \Delta(x) - (L+1)\mathrm{dist}(x, x') > 0$. It is sufficient to show that $V_\delta(x') \subseteq \mathcal{V}$, as this implies $\Delta(x') \geq \delta = \Delta(x) - (L+1)\mathrm{dist}(x, x')$, allowing us to conclude. To this end, we show the first inclusion in:

$$V_\delta(x') \subseteq V_{\Delta(x)}(x) \subseteq \mathcal{V}.$$

Consider an arbitrary $(x'', s'') \in V_\delta(x')$. This implies two things: first, $\|s''\|_{x''} < \delta \leq \Delta(x)$, and second:

$$\begin{aligned}
\mathrm{dist}(x'', x) &\leq \mathrm{dist}(x'', x') + \mathrm{dist}(x', x) \\
&< \delta - r(x') + \mathrm{dist}(x', x) \\
&= \Delta(x) - r(x) + r(x) - r(x') - L\mathrm{dist}(x, x') \\
&\leq \Delta(x) - r(x),
\end{aligned}$$

where in the last step we used $r(x) - r(x') \leq L\mathrm{dist}(x, x')$ since r is L-Lipschitz continuous on $\mathcal{U}$. As a result, (x'', s'') is in $V_{\Delta(x)}(x)$, which concludes this part of the proof.

Step 2: $\Delta(x) > r(x)$ for all $x \in \mathcal{U}$. Pick $x \in \mathcal{U}$, arbitrary: $\mathcal{V}$ is a neighborhood of

$$\{(x, s) \in \mathrm{T}\mathcal{M} : \|s\|_x \leq r(x)\}. \tag{62}$$

The claim is that there exists $\varepsilon > 0$ such that

$$\left\{(x', s') \in \mathrm{T}\mathcal{M} : \mathrm{dist}(x, x') \leq \varepsilon \text{ and } \|s'\|_{x'} \leq r(x) + \varepsilon\right\} \tag{63}$$

is included in $\mathcal{V}$. Indeed, that would show that $\Delta(x) \geq r(x) + \varepsilon > r(x)$. To show this, we construct special coordinates on $\mathrm{T}\mathcal{M}$ around x.

The (inverse of the) exponential map at x restricted to tangent vectors of norm strictly less than $\mathrm{inj}(x)$ (the injectivity radius at x) provides a diffeomorphism φ from $\mathcal{W} \subseteq \mathcal{M}$ (the open geodesic ball of radius $\mathrm{inj}(x)$ around x) to $B(0, \mathrm{inj}(x))$: the open ball centered around the origin in the Euclidean space $\mathbb{R}^d$, where $d = \dim \mathcal{M}$. Additionally, from the chart $(\mathcal{W}, \varphi)$, we extract coordinate vector fields on $\mathcal{W}$: a set of smooth vector fields $W_1, \ldots, W_d$ on $\mathcal{W}$ such that, at each point in $\mathcal{W}$, the corresponding tangent vectors form a basis for the tangent space. We further orthonormalize this local frame (see (Lee, 2018, Prop. 2.8)) into a new local frame, $E_1, \ldots, E_d$, so that for each $x' \in \mathcal{W}$ we have that $E_1(x'), \ldots, E_d(x')$ form an orthonormal basis for $\mathrm{T}_{x'}\mathcal{M}$ (with respect to the Riemannian metric at x'). Then, the map

$$\psi(x', s') = \big(\varphi(x'), \zeta(x', s')\big) \qquad \text{with} \qquad \zeta(x', s') = \big(\langle E_1(x'), s' \rangle_{x'}, \ldots, \langle E_d(x'), s' \rangle_{x'}\big)$$

establishes a diffeomorphism between $\mathrm{T}\mathcal{W}$ and $B(0, \mathrm{inj}(x)) \times \mathbb{R}^d$, with the following properties:

1. $\mathrm{dist}(x, x') = \|\varphi(x')\|$ (in particular, $\varphi(x) = 0$), and
2. For any $s', v' \in \mathrm{T}_{x'}\mathcal{M}$, it holds $\langle s', v' \rangle_{x'} = \langle \zeta(x', s'), \zeta(x', v') \rangle$.

(Here, $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$ denote the Euclidean inner product and norm in $\mathbb{R}^d$.)

Expressed in these coordinates (that is, mapped through ψ), the set in (62) becomes:

$$D_0 = \{0\} \times \bar{B}(0, r(x)),$$

where $\bar{B}(0, r(x))$ denotes the closed Euclidean ball of radius $r(x)$ around the origin in $\mathbb{R}^d$. Of course, $\mathcal{V} \cap \mathrm{T}\mathcal{W}$ maps to a neighborhood of D_0 in $\mathbb{R}^d \times \mathbb{R}^d$: call it O. Similarly, the set in (63) maps to:

$$D_\varepsilon = \bar{B}(0, \varepsilon) \times \bar{B}(0, r(x) + \varepsilon).$$

It remains to show that there exists $\varepsilon > 0$ such that D_ε is included in O.

Use this distance on $\mathbb{R}^d \times \mathbb{R}^d$: $\mathrm{dist}((y, z), (y', z')) = \max(\|y - y'\|, \|z - z'\|)$. This distance is compatible with the usual topology. For each $(0, z)$ in D_0, there exists $\varepsilon_z > 0$ such that

$$C(z, \varepsilon_z) = \left\{(y', z') \in \mathbb{R}^d \times \mathbb{R}^d : \|y'\| < \varepsilon_z \text{ and } \|z - z'\| < \varepsilon_z\right\}$$

is included in O (this is where we use the fact that $\mathcal{V}$—hence O—is open). The collection of open sets $C(z, \varepsilon_z/2)$ forms an open cover of D_0. Since D_0 is compact, we may extract a finite subcover, that is, we select $z_1, \ldots, z_n$ such that the sets $C(z_i, \varepsilon_{z_i}/2)$ cover D_0. Now, define $\varepsilon = \min_{i=1,\ldots,n} \varepsilon_{z_i}/2$ (necessarily positive), and consider any point $(y, z) \in D_\varepsilon$. We must show that (y, z) is in O. To this end, let $\bar{z}$ denote the point in $\bar{B}(0, r(x))$ which is closest to z. Since $(0, \bar{z})$ is in D_0, there exists i such that $(0, \bar{z})$ is in $C(z_i, \varepsilon_{z_i}/2)$. As a result,

$$\|z - z_i\| \leq \|z - \bar{z}\| + \|\bar{z} - z_i\| < \varepsilon + \varepsilon_{z_i}/2 \leq \varepsilon_{z_i}.$$

Likewise, $\|y\| \leq \varepsilon \leq \varepsilon_{z_i}/2 < \varepsilon_{z_i}$. Thus, we conclude that (y, z) is in $C(z_i, \varepsilon_{z_i})$, which is included in O. This confirms D_ε is in O, so that the set in (63) is in $\mathcal{V}$ for some $\varepsilon > 0$. □