Global rates of convergence for nonconvex optimization on manifolds

Nicolas Boumal* P.-A. Absil† Coralia Cartis‡

May 27, 2016

Abstract

We consider the minimization of a cost function f on a manifold $\mathcal{M}$ using Riemannian gradient descent and Riemannian trust regions (RTR). We focus on satisfying necessary optimality conditions within a tolerance ε. Specifically, we show that, under Lipschitz-type assumptions on the pullbacks of f to the tangent spaces of $\mathcal{M}$, both of these algorithms produce points with Riemannian gradient smaller than ε in $\mathcal{O}(1/\varepsilon^2)$ iterations. Furthermore, RTR returns a point where also the Riemannian Hessian's least eigenvalue is larger than $-\varepsilon$ in $\mathcal{O}(1/\varepsilon^3)$ iterations. There are no assumptions on initialization. The rates match their (sharp) unconstrained counterparts as a function of the accuracy ε (up to constants) and hence are sharp in that sense.

These are the first general results for global rates of convergence to approximate first- and second-order KKT points on manifolds. They apply in particular for optimization constrained to compact submanifolds of $\mathbb{R}^n$, under simpler assumptions.

1 Introduction

Optimization on manifolds is concerned with solving nonlinear and typically nonconvex computational problems of the form

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

where $\mathcal{M}$ is a (smooth) Riemannian manifold and $f\colon \mathcal{M} \to \mathbb{R}$ is a (sufficiently smooth) cost function (Absil et al., 2008). Applications abound in machine learning, computer vision, scientific computing, numerical linear algebra, signal processing, etc. In typical applications, x is a matrix and $\mathcal{M}$ could be a Stiefel manifold of orthonormal frames (including spheres and groups of rotations), a Grassmann manifold of subspaces, a cone of positive definite matrices, or of course simply a Euclidean space such as $\mathbb{R}^n$. Appendix A summarizes useful concepts about manifolds.

The standard theory for optimization on manifolds takes the standpoint that optimizing on a manifold $\mathcal{M}$ is not fundamentally different from optimizing in $\mathbb{R}^n$. Indeed, many classical algorithms from unconstrained nonlinear optimization such as gradient descent, nonlinear

*Mathematics Department, Princeton University, Princeton, NJ, USA.
†ICTEAM Institute, Université catholique de Louvain, Louvain-la-Neuve, Belgium.
‡Mathematical Institute, University of Oxford, Oxford, UK.

conjugate gradients, BFGS, Newton's method and trust-region methods (Ruszczyński, 2006; Nocedal and Wright, 1999) have been adapted to apply to the larger framework of (P) (Absil et al., 2008; Ring and Wirth, 2012; Sato, 2014; Huang et al., 2015). Software-wise, Manopt is a general toolbox for optimization on manifolds which can be used to experiment with many of these algorithms on various manifolds (Boumal et al., 2014).

As (P) is typically nonconvex, one does not expect general purpose, efficient algorithms to converge to global optima of (P) in general. Indeed, the class of problems (P) includes known NP-hard problems. In fact, even computing *local* optima is NP-hard in general (Vavasis, 1991, §5). Nevertheless, one may still hope to compute points of $\mathcal{M}$ which satisfy first- and second-order necessary optimality conditions. These take up the same form as in unconstrained nonlinear optimization, with *Riemannian* notions of gradient and Hessian. For $\mathcal{M}$ defined by equality constraints, these conditions are equivalent to first- and second-order KKT conditions, but are simpler to manipulate because the Lagrangian multipliers are automatically determined.

The proposition below states these necessary optimality conditions. Recall that to each point x of $\mathcal{M}$ corresponds a tangent space (a linearization) $\mathrm{T}_x\mathcal{M}$. The Riemannian gradient $\mathrm{grad} f(x)$ is the unique tangent vector at x such that $\mathrm{D}f(x)[\eta] = \langle \eta, \mathrm{grad} f(x)\rangle$ for all tangent vectors η, where $\langle \cdot, \cdot \rangle$ is the Riemannian metric on $\mathrm{T}_x\mathcal{M}$, and $\mathrm{D}f(x)[\eta]$ is the directional derivative of f at x along η. The Riemannian Hessian $\mathrm{Hess} f(x)$ is a symmetric operator on $\mathrm{T}_x\mathcal{M}$, corresponding to the derivative of the gradient vector field with respect to the Levi–Civita connection—see Appendix A.

Proposition 1 (Necessary optimality conditions)**.** *Let $x \in \mathcal{M}$ be a local optimum for (P). If f is differentiable at x, then $\mathrm{grad} f(x) = 0$. If f is twice differentiable at x, then $\mathrm{Hess} f(x) \succeq 0$ (positive semidefinite).*

Proof. See (Yang et al., 2014, Rem. 4.2 and Cor. 4.2). There is no need to require differentiability or even continuity *around* x. □

A point $x \in \mathcal{M}$ which satisfies $\mathrm{grad} f(x) = 0$ is a *(first-order) critical point* (also called a stationary point). If x furthermore satisfies $\mathrm{Hess} f(x) \succeq 0$, it is a *second-order critical point*.

Existing theory for optimization algorithms on manifolds is mostly concerned with establishing global convergence to critical points without rates (where global means regardless of initialization), as well as local rates of convergence. For example, the gradient descent method is known to converge globally to critical points, and the convergence rate is linear once the iterates reach *a sufficiently small neighborhood* of the limit point (Absil et al., 2008, §4). Early work of Udriste (1994) on local convergence rates even bounds distance to optimizers as a function of iteration count, assuming initialization in a set where the Hessian of f is positive definite, with lower and upper bounds on the eigenvalues; see also (Absil et al., 2008, Thm. 4.5.6, Thm. 7.4.11). Such guarantees adequately describe the empirical behavior of those methods, but do not inform us about how many iterations are required to reach the local regime from an arbitrary initial point x_0; that is: the worst-case scenarios are not addressed.

For classical unconstrained nonlinear optimization, this caveat has been addressed by bounding the number of iterations required by known algorithms to compute points which satisfy necessary optimality conditions within some tolerance, without assumptions on the initial iterate. Among others, Nesterov (2004) gives a proof that, for $\mathcal{M} = \mathbb{R}^n$ and Lipschitz

differentiable f, gradient descent with an appropriate step-size computes a point x where $\|\mathrm{grad} f(x)\| \leq \varepsilon$ in $\mathcal{O}(1/\varepsilon^2)$ iterations. This is sharp (Cartis et al., 2010). Cartis et al. (2012) prove the same for trust-region methods, and further show that if f is twice Lipschitz continuously differentiable, then a point x where $\|\mathrm{grad} f(x)\| \leq \varepsilon$ and $\mathrm{Hess} f(x) \succeq -\varepsilon \,\mathrm{Id}$ is computed in $\mathcal{O}(1/\varepsilon^3)$ iterations, also with examples showing sharpness.

Note that optimal complexity bounds of the order $\mathcal{O}(1/\varepsilon^{1.5})$ have also been given for cubic regularization methods (Cartis et al., 2011a,b) and sophisticated trust region variants (Curtis et al., 2016) to generate x with $\|\mathrm{grad} f(x)\| \leq \varepsilon$. Bounds for regularization methods can be further improved if higher-order derivatives are available (Birgin et al., 2015).

Worst-case evaluation complexity bounds have been extended to constrained smooth problems in (Cartis et al., 2014, 2015a,b) where it is shown that the same order global rate of convergence as in the unconstrained case can be achieved for obtaining approximate KKT points by some carefully devised, albeit impractical, phase 1–phase 2 methods. We note that when the constraints are convex (but the objective may not be), practical, feasible methods have been devised (Cartis et al., 2015a) that connect to our approach below. Aside from the results presented here, no complexity bounds are yet available for reaching higher than first-order criticality for constrained nonconvex problems.

In this paper, we extend the unconstrained results to the larger class of optimization problems on manifolds (P). This work rests heavily on the original proofs (Nesterov, 2004; Cartis et al., 2012) and on existing adaptations of gradient descent and trust-region methods to manifolds (Absil et al., 2007, 2008). One key step is the identification of a set of relevant Lipschitz-type regularity assumptions which allow the proofs to carry over from $\mathbb{R}^n$ to $\mathcal{M}$ with relative ease. The version of Riemannian trust-regions (RTR) we study works exactly as usual if only first-order necessary optimality conditions are targeted, and naturally continues its work if also second-order conditions are targeted.

We state the main results here informally. We use the notion of *retraction* Retr_x (see Definition 1), which allows to map tangent vectors at x to points on $\mathcal{M}$. Iterates are related by $x_{k+1} = \mathrm{Retr}_{x_k}(\eta_k)$ for some tangent vector η_k at x_k (the step). Hence, $f \circ \mathrm{Retr}_x$ is a lift of the cost function from $\mathcal{M}$ to the tangent space at x. For $\mathcal{M} = \mathbb{R}^n$, the standard retraction gives $x_k + \eta_k$.

Main result about gradient descent (See Theorems 4 and 6.) For problem (P), if f is bounded below on $\mathcal{M}$ and $f \circ \mathrm{Retr}_x$ has Lipschitz gradient with constant L independent of x, then gradient descent with constant step size $1/L$ or with backtracking Armijo line-search returns x with $\|\mathrm{grad} f(x)\| \leq \varepsilon$ in $\mathcal{O}(1/\varepsilon^2)$ iterations.

Main result about trust regions (See Theorem 11.) For problem (P), if f is bounded below on $\mathcal{M}$ and $f \circ \mathrm{Retr}_x$ has Lipschitz gradient with constant independent of x, then RTR returns x with $\|\mathrm{grad} f(x)\| \leq \varepsilon_g$ in $\mathcal{O}(1/\varepsilon_g^2)$ iterations, under reasonably weak assumptions on the model quality. If further $f \circ \mathrm{Retr}_x$ has Lipschitz Hessian with constant independent of x, then RTR returns x with $\|\mathrm{grad} f(x)\| \leq \varepsilon_g$ and $\mathrm{Hess} f(x) \succeq -\varepsilon_H \,\mathrm{Id}$ in $\mathcal{O}(\max\{1/\varepsilon_H^3, 1/\varepsilon_g^2\varepsilon_H\})$ iterations, provided the true Hessian is used in the model and a second-order retraction is used.

Main result for compact submanifolds (See Lemmas 3 and 8.) The first-order regularity conditions above hold in particular if $\mathcal{M}$ is a compact submanifold of a Euclidean space $\mathcal{E}$ (such as $\mathbb{R}^n$) and $f\colon \mathcal{E} \to \mathbb{R}$ has a locally Lipschitz continuous gradient. The second-order regularity conditions hold if furthermore f has a locally Lipschitz continuous Hessian and the retraction is second order (Definition 2).

Since the rates $\mathcal{O}(1/\varepsilon^2)$ and $\mathcal{O}(1/\varepsilon^3)$ are sharp for gradient descent and trust regions when $\mathcal{M} = \mathbb{R}^n$ (Cartis et al., 2010, 2012), they are also sharp for $\mathcal{M}$ a generic Riemannian manifold. Below, constants are given explicitly, thus precisely bounding the total amount of work required in the worst case to attain a prescribed tolerance.

The theorems presented here are the first general results about the worst-case iteration complexity of computing (approximate) first- and second-order critical points on manifolds. The choice of analyzing Riemannian gradient descent and RTR first is guided by practical concerns, as these are among the most commonly used methods on manifolds so far. The proposed complexity bounds are particularly relevant when applied to problems for which second-order necessary optimality conditions are also sufficient. See for example (Sun et al., 2015, 2016; Boumal, 2015b, 2016; Bandeira et al., 2016) and the example in Section 4.

The complexity of Riemannian optimization is discussed in at least two recent lines of work. Zhang and Sra (2016) treat geodesically convex problems over Hadamard manifolds. This is a remarkable extension of important pieces of classic convex optimization theory to manifolds with negative curvature. Because of the focus on geodesically convex problems, those results do not apply to the more general problem (P), but have the clear advantage of guaranteeing global optimality. Sun et al. (2015, 2016) consider dictionary learning and phase retrieval, and show that these problems, when appropriately framed as optimization on a manifold, are low dimensional and have no spurious local optimizers. They derive the complexity of RTR specialized to their application. In particular, they combine the global rate with a local convergence rate, which allows them to establish an overall better complexity than $\mathcal{O}(1/\varepsilon^3)$, but with an idealized version of the algorithm and restricted to these relevant applications. In this paper, we favor a more general approach, focused on algorithms closer to the ones implemented in practice.

2 Riemannian gradient descent methods

Consider the generic Riemannian descent method described in Algorithm 1. We first prove that, provided sufficient decrease in the cost function is achieved at each iteration, the algorithm computes a point x_k such that $\|\mathrm{grad} f(x_k)\| \leq \varepsilon$ with $k = \mathcal{O}(1/\varepsilon^2)$. Then, we propose a Lipschitz-type assumption which is sufficient to guarantee that simple, popular strategies to pick the steps η_k indeed ensure sufficient decrease. The proofs mimic the standard ones, see for example (Nesterov, 2004, §1.2.3). The main novelty is the careful extension to the Riemannian setting, which requires the well-known notion of retraction (Definition 1) and the new assumption A3.

The step η_k is a tangent vector to $\mathcal{M}$ at x_k. Because $\mathcal{M}$ is nonlinear (in general), the operation $x_k + \eta_k$ is undefined. The notion of *retraction* provides a theoretically sound replacement. Informally, $x_{k+1} = \mathrm{Retr}_{x_k}(\eta_k)$ is *a point on $\mathcal{M}$ one reaches by moving away from x_k, along the direction η_k, while remaining on the manifold.* The (Riemannian) exponential

map (which generates geodesics) is a retraction. The crucial point is that many other maps are retractions, often far less difficult to compute than the exponential.

Definition 1 (Retraction, (Absil et al., 2008, Def. 4.1.1)). *A* retraction *on a manifold $\mathcal{M}$ is a smooth mapping* Retr *from the tangent bundle*[1] $\mathrm{T}\mathcal{M}$ *to $\mathcal{M}$ with the following properties. Let* $\mathrm{Retr}_x \colon \mathrm{T}_x\mathcal{M} \to \mathcal{M}$ *denote the restriction of* Retr *to* $\mathrm{T}_x\mathcal{M}$.

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

(ii) The differential of Retr_x *at* 0_x, $\mathrm{DRetr}_x(0_x)$, *is the identity map.*

These combined conditions ensure retraction curves $t \mapsto \mathrm{Retr}_x(t\eta)$ agree up to first order with geodesics passing through x with velocity η, around $t = 0$.

In linear spaces such as $\mathbb{R}^n$, the typical choice is $\mathrm{Retr}_x(\eta) = x + \eta$. On the sphere, a popular choice is $\mathrm{Retr}_x(\eta) = \frac{x+\eta}{\|x+\eta\|}$. See Remark 7 below for retractions defined only on subsets of the tangent spaces.

Algorithm 1 Generic Riemannian descent algorithm

Given: $f \colon \mathcal{M} \to \mathbb{R}$ differentiable, a retraction Retr on $\mathcal{M}$, $x_0 \in \mathcal{M}$, $\varepsilon > 0$
Init: $k \leftarrow 0$
while $\|\mathrm{grad} f(x_k)\| > \varepsilon$ **do**
    Pick $\eta_k \in \mathrm{T}_{x_k}\mathcal{M}$
    $x_{k+1} = \mathrm{Retr}_{x_k}(\eta_k)$
    $k \leftarrow k + 1$
end while
return x_k

The two central assumptions and the main theorem about Algorithm 1 follow.

A1 (Lower bound). *There exists $f^* > -\infty$ such that $f(x) \geq f^*$ for all $x \in \mathcal{M}$.*

A2 (Sufficient decrease). *There exists $c > 0$ such that, for all $k \geq 0$,*

$$f(x_k) - f(x_{k+1}) \geq c\|\mathrm{grad} f(x_k)\|^2.$$

Theorem 2. *Under A1 and A2, Algorithm 1 returns $x \in \mathcal{M}$ satisfying $f(x) \leq f(x_0)$ and $\|\mathrm{grad} f(x)\| \leq \varepsilon$ in at most*

$$\left\lceil \frac{f(x_0) - f^*}{c} \cdot \frac{1}{\varepsilon^2} \right\rceil$$

iterations. Equivalently, for $K \geq 1$,

$$\min_{k \in \{0,\ldots,K-1\}} \|\mathrm{grad} f(x_k)\| \leq \sqrt{\frac{f(x_0) - f^*}{c}} \frac{1}{\sqrt{K}}.$$

It also holds that $\|\mathrm{grad} f(x_k)\| \to 0$ as $k \to \infty$.

[1] Informally, the tangent bundle $\mathrm{T}\mathcal{M}$ is the set of all pairs (x, η_x) where $x \in \mathcal{M}$ and $\eta_x \in \mathrm{T}_x\mathcal{M}$. See the reference for a proper definition of $\mathrm{T}\mathcal{M}$ and of what it means for Retr to be smooth.

Proof. Assume Algorithm 1 executes K iterations without returning, i.e., $\|\mathrm{grad} f(x_k)\| > \varepsilon$ for all k in $0, \ldots, K-1$. Then, using boundedness of f and sufficient decrease, a classic telescoping sum argument gives

$$f(x_0) - f^* \geq f(x_0) - f(x_K) = \sum_{k=0}^{K-1} f(x_k) - f(x_{k+1}) \geq \sum_{k=0}^{K-1} c\|\mathrm{grad} f(x_k)\|^2 > cK\varepsilon^2.$$

Hence, by contradiction, the algorithm is sure to have returned if $K \geq \frac{f(x_0)-f^*}{c\varepsilon^2}$. Taking the limit for $K \to \infty$ above also implies $\|\mathrm{grad} f(x_k)\| \to 0$ as $k \to \infty$. □

To ensure that A2 can be satisfied using simple rules to choose the steps η_k, it is convenient to make the following regularity assumption about the *pullbacks*[2] $\hat{f}_x = f \circ \mathrm{Retr}_x \colon \mathrm{T}_x\mathcal{M} \to \mathbb{R}$, conveniently defined on vector spaces. We use the fact that $\nabla \hat{f}_x(0_x) = \mathrm{grad} f(x)$, which follows from the definition of retraction.[3]

A3 (Lipschitz gradient)**.** *There exists $L \geq 0$ such that, for all x among $x_0, x_1 \ldots$ generated by Algorithm 1, the pullback $\hat{f}_x = f \circ \mathrm{Retr}_x$ has Lipschitz continuous gradient with constant L, that is, for all $\eta \in \mathrm{T}_x\mathcal{M}$, it holds that*

$$\left| \hat{f}_x(\eta) - [f(x) + \langle \eta, \mathrm{grad} f(x) \rangle] \right| \leq \frac{L}{2}\|\eta\|^2. \tag{1}$$

In words, $\hat{f}_x$ is uniformly well approximated by its first-order Taylor expansion.

For the particular case $\mathcal{M} = \mathbb{R}^n$ and $\mathrm{Retr}_x(\eta) = x + \eta$, equation (1) holds for all $x \in \mathcal{M}$ and $\eta \in \mathrm{T}_x\mathcal{M}$ if f has a Lipschitz continuous gradient in $\mathbb{R}^n$—this is the classical regularity assumption required in (Nesterov, 2004; Cartis et al., 2012). Furthermore, the lemma below states that if $\mathcal{M}$ is a compact submanifold of $\mathbb{R}^n$, then a sufficient condition for A3 to hold is for $f \colon \mathbb{R}^n \to \mathbb{R}$ to have locally Lipschitz continuous gradient (so that it has Lipschitz continuous gradient on any compact subset of $\mathbb{R}^n$). The proof is in Appendix B.

Lemma 3. *Let $\mathcal{E}$ be a Euclidean space (for example, $\mathcal{E} = \mathbb{R}^n$) and let $\mathcal{M}$ be a compact Riemannian submanifold of $\mathcal{E}$. Let* Retr *be a retraction on $\mathcal{M}$. If $f \colon \mathcal{E} \to \mathbb{R}$ has Lipschitz continuous gradient in the convex hull of $\mathcal{M}$, then the pullbacks $f \circ \mathrm{Retr}_x$ have Lipschitz continuous gradient with some constant $\bar{L}$ independent of x; hence, A3 holds.*

A usual Lipschitz-type assumption, which would imply A3 in the Euclidean case, would be:

$$\forall x, y \in \mathcal{M}, \quad \|\mathrm{grad} f(x) - \mathrm{grad} f(y)\| \leq L \cdot \mathrm{dist}(x, y). \tag{2}$$

This, however, poses two difficulties. Firstly, $\mathrm{grad} f(x)$ and $\mathrm{grad} f(y)$ live in two different tangent spaces, so that their difference is not defined; instead, $\mathrm{grad} f(y)$ must be *transported* to $\mathrm{T}_x\mathcal{M}$, which requires the introduction of *parallel transports*. Secondly, the right hand side involves the *geodesic distance* on $\mathcal{M}$. Both notions involve subtle definitions; transports may even not be defined on all of $\mathcal{M}$. It is of course possible to work with (2) (see for

[2] The composition $f \circ \mathrm{Retr}_x$ is called the pullback because it, quite literally, pulls back the cost function f from the manifold $\mathcal{M}$ to the linear space $\mathrm{T}_x\mathcal{M}$.

[3] $\forall \eta, \langle \nabla \hat{f}_x(0_x), \eta \rangle = \mathrm{D}\hat{f}_x(0_x)[\eta] = \mathrm{D}f(x)[\mathrm{DRetr}_x(0_x)[\eta]] = \mathrm{D}f(x)[\eta] = \langle \mathrm{grad} f(x), \eta \rangle$.

example recent work of Zhang and Sra (2016)), but we argue that it is conceptually and computationally advantageous to avoid them when possible. The computational advantage comes from the possibility in A3 to work with any retraction, whereas parallel transport and geodesic distance are tied to the exponential map.

2.1 Fixed step-size gradient descent method

Leveraging the Lipschitz assumption A3, an easy strategy is to pick the steps η_k proportional to the negative gradient.

Theorem 4 (Riemannian gradient descent with fixed step-size)**.** *Under A1 and A3, Algorithm 1 with the explicit strategy*

$$\eta_k = -\frac{1}{L}\mathrm{grad} f(x_k)$$

returns a point $x \in \mathcal{M}$ satisfying $f(x) \leq f(x_0)$ and $\|\mathrm{grad} f(x)\| \leq \varepsilon$ in at most

$$\left\lceil 2\big(f(x_0) - f^*\big)L \cdot \frac{1}{\varepsilon^2} \right\rceil$$

iterations. Each iteration requires one cost and gradient evaluation and one retraction.

Proof. It is sufficient to ensure A2 holds with the appropriate c. The Lipschitz assumption provides a global upper bound for the pullback: for all k,

$$\forall \eta \in \mathrm{T}_{x_k}\mathcal{M}, \quad f(\mathrm{Retr}_{x_k}(\eta)) \leq f(x_k) + \langle \eta, \mathrm{grad} f(x_k)\rangle + \frac{L}{2}\|\eta\|^2. \tag{3}$$

Then, for the given choice of η_k and using $x_{k+1} = \mathrm{Retr}_{x_k}(\eta_k)$,

$$\begin{aligned} f(x_k) - f(x_{k+1}) &\geq -\langle \eta_k, \mathrm{grad} f(x_k)\rangle - \frac{L}{2}\|\eta_k\|^2 \\ &= \frac{1}{L}\|\mathrm{grad} f(x_k)\|^2 - \frac{L}{2}\frac{1}{L^2}\|\mathrm{grad} f(x_k)\|^2 \\ &= \frac{1}{2L}\|\mathrm{grad} f(x_k)\|^2. \end{aligned}$$

Thus, A2 holds with $c = \frac{1}{2L}$. We conclude by applying Theorem 2. ∎

2.2 Gradient descent with backtracking Armijo line-search

In practice, the constant L appearing in A3 is often too conservative, leading to too small steps, or not available at all to the algorithm. Accordingly, a more practical strategy is to use an inexact line-search algorithm. This has the advantage of not requiring knowledge of L, and of being adaptive. The following lemma shows that a basic Armijo-type backtracking line-search, Algorithm 2, computes a step η_k satisfying A2 in a bounded number of function calls. The statement is made more general by allowing search directions other than $-\mathrm{grad} f(x_k)$, provided they remain "related" to $-\mathrm{grad} f(x_k)$. This result is well known in the Euclidean case and carries over seamlessly under A3.

Algorithm 2 Backtracking Armijo line-search

Given: $x_k \in \mathcal{M}$, $\eta_k^0 \in \mathrm{T}_{x_k}\mathcal{M}$, $c_1 \in (0,1)$, $\bar{t} > 0$, $\tau \in (0,1)$
Init: $t \leftarrow \bar{t}$
while $f(x_k) - f(\mathrm{Retr}_{x_k}(t \cdot \eta_k^0)) < c_1 t \left\langle -\mathrm{grad} f(x_k), \eta_k^0 \right\rangle$ **do**
    $t \leftarrow \tau \cdot t$
end while
return t and $\eta_k = t\eta_k^0$.

Lemma 5. *For each iteration k of Algorithm 1, let $\eta_k^0 \in \mathrm{T}_{x_k}\mathcal{M}$ be the initial search direction to be considered for line-search. Assume there exist constants $c_2 \in (0,1]$ and $0 < c_3 \leq c_4$ such that, for all k,*

$$\left\langle -\mathrm{grad} f(x_k), \eta_k^0 \right\rangle \geq c_2 \|\mathrm{grad} f(x_k)\| \|\eta_k^0\|$$

and

$$c_3 \|\mathrm{grad} f(x_k)\| \leq \|\eta_k^0\| \leq c_4 \|\mathrm{grad} f(x_k)\|.$$

Under A3, backtracking Armijo (Algorithm 2) returns a positive t and $\eta_k = t\eta_k^0$ such that

$$f(x_k) - f(\mathrm{Retr}_{x_k}(\eta_k)) \geq c_1 c_2 c_3 t \|\mathrm{grad} f(x_k)\|^2 \qquad \text{and} \qquad t \geq \min\left(\bar{t}, \frac{2\tau c_2 (1-c_1)}{c_4 L}\right) \tag{4}$$

in

$$1 + \log_\tau (t/\bar{t}) \leq \max\left(1, 2 + \left\lceil \log_{\tau^{-1}}\left(\frac{c_4 \bar{t} L}{2c_2(1-c_1)}\right)\right\rceil\right)$$

retractions and cost evaluations (not counting evaluation of f at x_k).

Proof. By A3, upper bound (3) holds in particular with $\eta = t\eta_k^0$ (for any t):

$$f(x_k) - f(\mathrm{Retr}_{x_k}(t \cdot \eta_k^0)) \geq t \left\langle -\mathrm{grad} f(x_k), \eta_k^0 \right\rangle - \frac{Lt^2}{2} \|\eta_k^0\|^2. \tag{5}$$

We determine how small t might need to be for the stopping criterion in Algorithm 2 to surely trigger. To this end, observe that the right hand side of (5) dominates $c_1 t \left\langle -\mathrm{grad} f(x_k), \eta_k^0 \right\rangle$ if

$$t(1-c_1) \cdot \left\langle -\mathrm{grad} f(x_k), \eta_k^0 \right\rangle \geq \frac{Lt^2}{2} \|\eta_k^0\|^2.$$

Thus, the stopping criterion in Algorithm 2 is satisfied in particular for all t in

$$\left[0, \frac{2(1-c_1)\left\langle -\mathrm{grad} f(x_k), \eta_k^0 \right\rangle}{L\|\eta_k^0\|^2}\right] \supseteq \left[0, \frac{2c_2(1-c_1)\|\mathrm{grad} f(x_k)\|}{L\|\eta_k^0\|}\right] \supseteq \left[0, \frac{2c_2(1-c_1)}{c_4 L}\right].$$

Unless it equals $\bar{t}$, the returned t cannot be smaller than τ times the last upper bound. In all cases, it fulfills

$$\begin{aligned} f(x_k) - f(\mathrm{Retr}_{x_k}(t \cdot \eta_k^0)) &\geq c_1 t \left\langle -\mathrm{grad} f(x_k), \eta_k^0 \right\rangle \\ &\geq c_1 c_2 t \|\mathrm{grad} f(x_k)\| \|\eta_k^0\| \\ &\geq c_1 c_2 c_3 t \|\mathrm{grad} f(x_k)\|^2. \end{aligned}$$

To count the number of iterations, consider that checking whether $t = \bar{t}$ satisfies the stopping criterion requires one cost evaluation. Following that, t is reduced by a factor τ exactly $\log_\tau(t/\bar{t}) = \log_{\tau^{-1}}(\bar{t}/t)$ times, each followed by a cost evaluation. □

The previous discussion can be particularized to bound the amount of work required by a gradient descent method using a backtracking Armijo line-search on manifolds. The constant L appears in the bounds but needs not be known.

Theorem 6 (Riemannian gradient descent with backtracking line-search)**.** *Under A1 and A3, Algorithm 1 with Algorithm 2 for line-search using parameters $c_1, \tau, \bar{t}$ and initial search direction $\eta_k^0 = -\mathrm{grad} f(x_k)$ returns a point $x \in \mathcal{M}$ satisfying $f(x) \leq f(x_0)$ and $\|\mathrm{grad} f(x)\| \leq \varepsilon$ in at most*

$$\left\lceil \frac{f(x_0) - f^*}{c_1 \min\left(\bar{t}, \frac{2\tau(1-c_1)}{L}\right)} \cdot \frac{1}{\varepsilon^2} \right\rceil$$

iterations. After computing $f(x_0)$ and $\mathrm{grad} f(x_0)$, each iteration requires one gradient evaluation and at most $\max\left(1, 2 + \left\lceil \log_{\tau^{-1}}\left(\frac{\bar{t}L}{2(1-c_1)}\right)\right\rceil\right)$ cost evaluations and retractions.

Proof. Using $\eta_k^0 = -\mathrm{grad} f(x_k)$, one can take $c_2 = c_3 = c_4 = 1$ in Lemma 5. A2 is then fulfilled with c prescribed by the latter lemma, so that Theorem 2 applies. At iteration k, the last cost evaluation of the line-search algorithm is the cost at x_{k+1}: it needs not be recomputed. □

Remark 7. *In some applications, it may be that the retraction Retr_{x_k} is only defined in a ball of radius ρ_k around the origin in $\mathrm{T}_{x_k}\mathcal{M}$, or that the constant L in A3 does not exist globally, but only when the pullbacks $f \circ \mathrm{Retr}_{x_k}$ are restricted to balls of radius ρ_k as well. Theorems in this section can be adapted to this situation, provided $\rho = \inf_k \rho_k > 0$. It then suffices to limit the size of steps to ρ_k. If the* injectivity radius *of the manifold is positive, valid retractions exist. Compact manifolds have positive injectivity radius (Chavel, 1993, Thm. III.2.3).*

Specifically, modifying the strategy from Theorem 4, we use

$$\eta_k = -\min\left(\frac{1}{L}, \frac{\rho_k}{\|\mathrm{grad} f(x_k)\|}\right) \mathrm{grad} f(x_k).$$

Simple computations show that $f(x_k) - f(x_{k+1}) \geq \frac{1}{2}\min\left(\frac{1}{L}, \frac{\rho_k}{\|\mathrm{grad} f(x_k)\|}\right)\|\mathrm{grad} f(x_k)\|^2$. Building upon the proof of Theorem 4, and separating iterations as

$$K_1 = \left\{k \in 0 \ldots K-1 : \frac{1}{L} \leq \frac{\rho_k}{\|\mathrm{grad} f(x_k)\|}\right\}, \qquad K_2 = \{0 \ldots K-1\} \backslash K_1,$$

we get

$$2(f(x_0) - f^*) \geq \sum_{k \in K_1} \frac{1}{L}\|\mathrm{grad} f(x_k)\|^2 + \sum_{k \in K_2} \rho_k \|\mathrm{grad} f(x_k)\| \geq |K_1|\frac{1}{L}\varepsilon^2 + |K_2|\rho^2 L.$$

Under the extra assumption $\varepsilon \leq \rho L$, we have $\rho^2 L \geq \frac{1}{L}\varepsilon^2$ and we reach the same conclusion as Theorem 4. The same reasoning applies to Theorem 6, with variable initial step size $\bar{t}_k = \min(\bar{t}, \rho_k / \|\eta_k^0\|)$ and $\varepsilon \leq \frac{\rho}{\sqrt{c_3 c_4}\min\left(\bar{t}, \frac{2\tau c_2(1-c_1)}{c_4 L}\right)} = \frac{\rho}{\min\left(\bar{t}, \frac{2\tau(1-c_1)}{L}\right)}$.

3 Riemannian trust-region methods

The Riemannian trust-region method (RTR), introduced in (Absil et al., 2007), is a generalization of the classical trust-region method to manifolds (Conn et al., 2000). The main idea is as follows. At iteration k, consider the pullback of the cost function, $\hat{f}_k = f \circ \mathrm{Retr}_{x_k}$, defined on the tangent space to $\mathcal{M}$ at x_k. We wish to minimize $\hat{f}_k$, but this is as difficult as the original problem. Instead, we consider a more manageable approximation of $\hat{f}_k$, called the *model* $\hat{m}_k$. If this model is suitably simple, we may be able to minimize it. But because $\hat{m}_k$ is only a local approximation of $\hat{f}_k$, we only trust it in a ball of radius Δ_k around 0_{x_k}: the *trust region*. Hence, we minimize (or, more often, approximately minimize) $\hat{m}_k$ in that region. The output is a step η_k which is retracted to obtain $x_k^+ = \mathrm{Retr}_{x_k}(\eta_k)$: the candidate next iterate. The candidate is accepted ($x_{k+1} = x_k^+$) if the actual cost decrease $f(x_k) - f(x_k^+)$ is a sufficiently large fraction—this is controlled by parameter ρ'—of the model decrease $\hat{m}_k(0_{x_k}) - \hat{m}_k(\eta_k)$. Otherwise, the candidate is rejected ($x_{k+1} = x_k$). Depending on the level of agreement of the model decrease and actual decrease, the trust-region radius Δ_k can be reduced, kept unchanged or increased (but never above some parameter $\bar{\Delta}$). The algorithm is initialized with a point x_0 and an initial radius Δ_0. The established convergence theory for RTR, rooted in the classical analysis of trust-region methods, is remarkably robust. The essential requirements are that (i) the models $\hat{m}_k$ should agree sufficiently with the pullbacks $\hat{f}_k$ (locally); and (ii) sufficient decrease in the model should be achieved at each iteration. Provided these (and other technical conditions) hold, *global* convergence to first-order critical points is guaranteed (Absil et al., 2008, §7.4). Linear and quadratic *local* convergence rates can also be established under stronger conditions. We focus on establishing rates for global convergence.

The model at iteration k is the function

$$\begin{aligned} \hat{m}_k &\colon \mathrm{T}_{x_k}\mathcal{M} \to \mathbb{R} \\ &\colon \eta \mapsto \hat{m}_k(\eta) = f(x_k) + \langle \eta, \mathrm{grad} f(x_k) \rangle + \frac{1}{2} \langle \eta, H_k[\eta] \rangle , \end{aligned} \tag{6}$$

for some map $H_k \colon \mathrm{T}_{x_k}\mathcal{M} \to \mathrm{T}_{x_k}\mathcal{M}$. The associated trust-region subproblem which is solved approximately at each iteration is

$$\min_{\eta \in \mathrm{T}_{x_k}\mathcal{M}} \ \hat{m}_k(\eta) \quad \text{subject to} \quad \|\eta\| \leq \Delta_k. \tag{7}$$

Because the models $\hat{m}_k$ can incorporate both first- and second-order information about the cost function f, RTR can be used to compute points which approximately satisfy both first- and second-order necessary optimality conditions. In particular, we study the computation of points $x \in \mathcal{M}$ such that $\|\mathrm{grad} f(x)\| \leq \varepsilon_g$ and $\mathrm{Hess} f(x) \succeq -\varepsilon_H \,\mathrm{Id}$, where $\mathrm{Hess} f(x)$ is the Riemannian Hessian of f at x.

In this section, we generalize the global rate of convergence analysis presented in (Cartis et al., 2012) to the manifold setting. As a pragmatic modification to the standard algorithm, we distinguish between first-order and second-order steps as follows.

As long as $\|\mathrm{grad} f(x_k)\| > \varepsilon_g$, we merely seek to obtain at least a *Cauchy decrease* (see below) in the model. This requires only first-order agreement between $\hat{f}_k$ and $\hat{m}_k$: conditions on H_k are particularly mild. If $\|\mathrm{grad} f(x_k)\| \leq \varepsilon_g$ and there are no second-order requirements ($\varepsilon_H = \infty$), the algorithm returns. Otherwise, we use (for this iteration) a second-order

accurate model: H_k must be related to the Hessian of f at x_k. From the model, we obtain an escape direction u_k, and follow it to obtain sufficient decrease in the model despite the small gradient. This is called an *eigenstep* (see also below).

If $\varepsilon_H < \infty$, RTR returns once the gradient norm drops below ε_g *and* $H_k \succeq -\varepsilon_H \mathrm{Id}$. Extra conditions on H_k are needed to ensure this translates into the appropriate statements regarding $\mathrm{Hess} f(x_k)$. A straightforward if sometimes expensive condition is to set $H_k = \mathrm{Hess} f(x_k)$. Then, to ensure H_k appropriately models the Hessian of the pullback at x_k (as required by A7 below), a second-order retraction can be used—see Section 3.5.

The benefit of this distinction between types of steps is that (i) if $\varepsilon_H = \infty$, we recover the classical RTR; and (ii) if $\varepsilon_H < \infty$, second-order work (which is typically more expensive) is kept for fine convergence: the first-order steps remain cheap.

We note that the specific trust-region radius evolution mechanism in Algorithm 3 is but one possible (and popular) scheme. It is straightforward to add flexibility there too, see (Cartis et al., 2012).

3.1 Regularity assumptions

In what follows, for iteration k, we make assumptions involving the ball of radius Δ_k around 0_{x_k} in the tangent space at x_k. It may be easier to check these properties in balls of radius $\bar{\Delta}$ in the tangent spaces at all $x \in \mathcal{M}$ such that $f(x) \leq f(x_0)$, since Algorithm 3 is a descent method, and it ensures $\Delta_k \leq \bar{\Delta}$ for all k. For the same reason, it is only necessary that the retraction (Definition 1) be defined in those same balls.

A4 (Restricted Lipschitz gradient)**.** *There exists $L_g \geq 0$ such that, for all x_k among $x_0, x_1 \ldots$ generated by Algorithm 3, the composition $\hat{f}_k = f \circ \mathrm{Retr}_{x_k}$ has Lipschitz continuous gradient with constant L_g in the trust region at iteration k, that is, for all $\eta \in \mathrm{T}_{x_k}\mathcal{M}$ such that $\|\eta\| \leq \Delta_k$, it holds that*

$$\left| \hat{f}_k(\eta) - [f(x_k) + \langle \eta, \mathrm{grad} f(x_k) \rangle] \right| \leq \frac{L_g}{2} \|\eta\|^2. \tag{9}$$

A5 (Restricted Lipschitz Hessian)**.** *If $\varepsilon_H < \infty$, there exists $L_H \geq 0$ such that, for all x_k among $x_0, x_1 \ldots$ generated by Algorithm 3 and such that $\|\mathrm{grad} f(x_k)\| \leq \varepsilon_g$, $\hat{f}_k$ has Lipschitz continuous Hessian with constant L_H in the trust region at iteration k, that is, for all $\eta \in \mathrm{T}_{x_k}\mathcal{M}$ such that $\|\eta\| \leq \Delta_k$, it holds that*

$$\left| \hat{f}_k(\eta) - \left[f(x_k) + \langle \eta, \mathrm{grad} f(x_k) \rangle + \frac{1}{2} \langle \eta, \nabla^2 \hat{f}_k(0_{x_k})[\eta] \rangle \right] \right| \leq \frac{L_H}{6} \|\eta\|^3. \tag{10}$$

Note that if Retr is a second-order retraction (see Section 3.5), then $\nabla^2 \hat{f}_k(0_{x_k})$ coincides with the Riemannian Hessian of f at x_k.

In both A4 and A5, the norm $\|\eta\|$ appearing in the right hand side can be relaxed to Δ_k—we refrain from doing so as it hinders interpretation with no clear advantage.

In the previous section, Lemma 3 gives a sufficient condition for A4 to hold; we complement this statement with a sufficient condition for A5 to hold as well. In a nutshell: if $\mathcal{M}$ is a compact submanifold of $\mathbb{R}^n$ and $f \colon \mathbb{R}^n \to \mathbb{R}$ has locally Lipschitz continuous Hessian, then both assumptions hold.

Algorithm 3 Riemannian trust regions (RTR), modified to attain second-order optimality

Parameters: $\bar{\Delta} > 0$, $0 < \rho' < 1/4$, $\varepsilon_g > 0$, $\varepsilon_H > 0$
Input: $x_0 \in \mathcal{M}$, $0 < \Delta_0 \leq \bar{\Delta}$
Init: $k \leftarrow 0$
while true do

  if $\|\mathrm{grad} f(x_k)\| > \varepsilon_g$ **then** ▷ First-order step.
    Obtain $\eta_k \in \mathrm{T}_{x_k}\mathcal{M}$ satisfying A8
  else if $\varepsilon_H < \infty$ **then** ▷ Second-order step.
    if $\lambda_{\min}(H_k) < -\varepsilon_H$ **then**
      Obtain $\eta_k \in \mathrm{T}_{x_k}\mathcal{M}$ satisfying A9
    else
      return x_k ▷ $\|\mathrm{grad} f(x_k)\| \leq \varepsilon_g$ and $\lambda_{\min}(H_k) \geq -\varepsilon_H$.
    end if
  else
    return x_k ▷ $\|\mathrm{grad} f(x_k)\| \leq \varepsilon_g$.
  end if

  Compute

$$\rho_k = \frac{\hat{f}_k(0_{x_k}) - \hat{f}_k(\eta_k)}{\hat{m}_k(0_{x_k}) - \hat{m}_k(\eta_k)} \tag{8}$$

  if $\rho_k < \frac{1}{4}$ **then** ▷ Poor model-cost agreement.
    $\Delta_{k+1} = \frac{1}{4}\Delta_k$
  else if $\rho_k > \frac{3}{4}$ and $\|\eta_k\| = \Delta_k$ **then** ▷ Good agreement and limiting TR.
    $\Delta_{k+1} = \min\left(2\Delta_k, \bar{\Delta}\right)$
  else
    $\Delta_{k+1} = \Delta_k$
  end if

  if $\rho_k > \rho'$ **then** ▷ Accept the step.
    $x_{k+1} = \mathrm{Retr}_{x_k}(\eta_k)$
  else ▷ Reject the step.
    $x_{k+1} = x_k$
  end if

  $k \leftarrow k + 1$
end while

Lemma 8. *Let $\mathcal{E}$ be a Euclidean space (for example, $\mathcal{E} = \mathbb{R}^n$) and let $\mathcal{M}$ be a compact Riemannian submanifold of $\mathcal{E}$. Let* Retr *be a second-order retraction on $\mathcal{M}$. If $f : \mathcal{E} \to \mathbb{R}$ has Lipschitz continuous Hessian in the convex hull of $\mathcal{M}$, then the pullbacks $f \circ \mathrm{Retr}_x$ have Lipschitz continuous Hessian with some constant $\bar{L}$ independent of x; hence, A5 holds.*

The proof is in Appendix B. Here too, if $\mathcal{M}$ is a Euclidean space and $\mathrm{Retr}_x(\eta) = x + \eta$, A4 and A5 are satisfied if f has Lipschitz continuous Hessian in the usual sense.

3.2 Assumptions about the models

The model at iteration k is the function $\hat{m}_k$ (6) whose purpose is to approximate the pullback $\hat{f}_k = f \circ \mathrm{Retr}_{x_k}$ (the lifted cost function). It involves a map $H_k : \mathrm{T}_{x_k}\mathcal{M} \to \mathrm{T}_{x_k}\mathcal{M}$. Depending on the type of step being performed (aiming for first- or second-order optimality conditions), we require different properties of the maps H_k. Conditions for first-order optimality are particularly lax.

A6. *If $\|\mathrm{grad} f(x_k)\| > \varepsilon_g$ (so that we are only aiming for a first-order condition at this step), then H_k is radially linear. That is,*

$$\forall \eta \in \mathrm{T}_{x_k}\mathcal{M}, \forall \alpha \geq 0, \quad H_k[\alpha\eta] = \alpha H_k[\eta]. \tag{11}$$

Furthermore, there exists $c_0 \geq 0$ (the same for all first-order steps) such that

$$\|H_k\| \triangleq \sup_{\eta \in \mathrm{T}_{x_k}\mathcal{M} : \|\eta\| \leq 1} \langle \eta, H_k[\eta] \rangle \leq c_0. \tag{12}$$

Radial linearity and boundedness are sufficient to ensure first-order agreement between $\hat{m}_k$ and $\hat{f}_k$. This relaxation from complete linearity of H_k—which would be the standard assumption—notably allows the use of finite difference approximations of the Hessian (Boumal, 2015a). To reach second-order agreement, the conditions are stronger.

A7. *If $\|\mathrm{grad} f(x_k)\| \leq \varepsilon_g$ and $\varepsilon_H < \infty$ (so that we are aiming for a second-order condition), then H_k is* linear and symmetric. *Furthermore, H_k is close to $\nabla^2 \hat{f}_k(0_{x_k})$ along η_k in the sense that there exists $c_1 \geq 0$ (the same for all second-order steps) such that:*

$$\left| \left\langle \eta_k, \left(\nabla^2 \hat{f}_k(0_{x_k}) - H_k\right)[\eta_k] \right\rangle \right| \leq \frac{c_1 \Delta_k}{3} \|\eta_k\|^2. \tag{13}$$

The smaller Δ_k, the more precisely H_k must approximate the Hessian of the pullback. Lemma 13 shows Δ_k is lower-bounded in relation with ε_g and ε_H. Eq. (13) involves η_k, the ultimately chosen step which typically will depend on H_k. The stronger condition below does not reference η_k and ensures (13) is satisfied:

$$\left\| \nabla^2 \hat{f}_k(0_{x_k}) - H_k \right\| \leq \frac{c_1 \Delta_k}{3}.$$

Refer to Section 3.5 to relate H_k, $\nabla^2 \hat{f}_k(0_{x_k})$ and $\mathrm{Hess} f(x_k)$.

3.3 Assumptions about sufficient model decrease

The steps η_k can be obtained in a number of ways, leading to different local convergence rates and empirical performance. As far as global convergence guarantees are concerned though, the requirements are modest. It is only required that, at each iteration, the candidate η_k induces sufficient decrease *in the model*. Known explicit strategies achieve these decreases. In particular, solving the trust-region subproblem (7) within some tolerance (which can be done in polynomial time if H_k is linear (Vavasis, 1991, §4.3)) is certain to satisfy the assumptions. See for example the Steihaug-Toint truncated conjugate gradients method for a popular, practical choice (Steihaug, 1983; Conn et al., 2000; Absil et al., 2007). See also (Sorensen, 1982; Moré and Sorensen, 1983) for more about the trust-region subproblem. Here, we describe simpler yet satisfactory strategies.

For first-order steps, we require the following.

A8. *There exists $c_2 > 0$ such that all first-order steps η_k satisfy*

$$\hat{m}_k(0_{x_k}) - \hat{m}_k(\eta_k) \geq c_2 \min\left(\Delta_k, \frac{\varepsilon_g}{c_0}\right)\varepsilon_g. \tag{14}$$

As is well known, the explicitly computable *Cauchy step* satisfies this requirement.

Lemma 9. *Let $g_k = \operatorname{grad} f(x_k)$. If $g_k \neq 0$, define the* Cauchy step *as $\eta_k^C = -\alpha_k^C g_k$ with*

$$\alpha_k^C = \begin{cases} \min\left(\frac{\|g_k\|^2}{\langle g_k, H_k[g_k]\rangle}, \frac{\Delta_k}{\|g_k\|}\right) & \text{if } \langle g_k, H_k[g_k]\rangle > 0, \\ \frac{\Delta_k}{\|g_k\|} & \text{otherwise.} \end{cases}$$

Under A6, setting $\eta_k = \eta_k^C$ for first-order steps fulfills A8 with $c_2 = 1/2$. Computing η_k^C involves one gradient evaluation and one application of H_k.

Proof. The claim follows as an exercise from $\hat{m}_k(0_{x_k}) - \hat{m}_k(\eta_k^C) = \alpha_k^C \|g_k\|^2 - \frac{(\alpha_k^C)^2}{2}\langle g_k, H_k[g_k]\rangle$ and $\langle g_k, H_k[g_k]\rangle \leq c_0\|g_k\|^2$ (the latter follows from A6). □

The formula for α_k^C comes about as follows. Consider minimizing the model $\hat{m}_k$ (6) in the trust region along $-g_k$. Owing to A6, this reads:

$$\begin{aligned} &\min_{\alpha \geq 0} \hat{m}_k(-\alpha g_k) = f(x_k) - \alpha\|g_k\|^2 + \frac{\alpha^2}{2}\langle g_k, H_k[g_k]\rangle \\ &\text{s.t. } \alpha\|g_k\| \leq \Delta_k. \end{aligned}$$

This corresponds to minimizing a quadratic in α over the interval $[0, \Delta_k/\|g_k\|]$. The optimal point is α_k^C (Conn et al., 2000; Absil et al., 2008). The Steihaug-Toint truncated conjugate gradient method (Steihaug, 1983) is a monotonically improving iterative method for the trust-region subproblem whose first iterate is the Cauchy step; as such, it necessarily achieves the required model decrease.

For second-order steps, the requirement is as follows.

A9. *There exists $c_3 > 0$ such that all second-order steps η_k satisfy*

$$\hat{m}_k(0_{x_k}) - \hat{m}_k(\eta_k) \geq c_3 \Delta_k^2 \varepsilon_H. \tag{15}$$

This can be achieved by making a step of maximal length along a direction which certifies that $\lambda_{\min}(H_k) < -\varepsilon_H$ (Conn et al., 2000). A key theoretical feature of this strategy is that its computational complexity is independent of ε_g and ε_H.

Lemma 10. *Let $g_k = \operatorname{grad} f(x_k)$. Under A7, assume $\lambda_{\min}(H_k) < -\varepsilon_H$. There exists a tangent vector $u_k \in \mathrm{T}_{x_k}\mathcal{M}$ such that*

$$\|u_k\| = 1, \qquad \langle u_k, g_k\rangle \leq 0, \qquad \text{and} \qquad \langle u_k, H_k[u_k]\rangle < -\varepsilon_H.$$

Define the eigenstep *as $\eta_k^E = \Delta_k u_k$. Setting $\eta_k = \eta_k^E$ for second-order steps fulfills A9 with $c_3 = 1/2$. Let $v_1, \ldots, v_n$ be an orthonormal basis of $\mathrm{T}_{x_k}\mathcal{M}$. One way of computing η_k^E involves the application of H_k to $v_1, \ldots, v_n$ plus a number of arithmetic operations polynomial in $n = \dim \mathcal{M}$ and independent of $\varepsilon_g, \varepsilon_H$.*

Proof. Compute H, a symmetric matrix of size n which represents H_k in the basis $v_1, \ldots, v_n$, as $H_{ij} = \langle v_i, H_k[v_j]\rangle$. Compute a factorization $LDL^\top = H + \varepsilon_H I$ where I is the identity matrix, L is invertible and triangular, and D is block diagonal with blocks of size 1×1 and 2×2. The factorization can be computed in $\mathcal{O}(n^3)$ operations (Golub and Van Loan, 2012, §4)—see the reference for a word of caution regarding pivoting and stability. D has the same inertia as $H + \varepsilon_H I$, hence D is not positive semidefinite (otherwise $H \succeq -\varepsilon_H I$.) The structure of D makes it easy to find $x \in \mathbb{R}^n$ with $x^\top D x < 0$. Solve the triangular system $L^\top y = x$ for $y \in \mathbb{R}^n$. Now, $0 > x^\top D x = y^\top L D L^\top y = y^\top (H + \varepsilon_H I) y$. Consequently, $y^\top H y < -\varepsilon_H \|y\|^2$. We can set $u_k = \pm \sum_{i=1}^n y_i v_i / \|y\|$, where the sign is chosen to ensure $\langle u_k, g_k\rangle \leq 0$. To conclude, check that $\hat{m}_k(0_{x_k}) - \hat{m}_k(\eta_k^E) = -\langle \eta_k^E, g_k\rangle - \frac{1}{2}\langle \eta_k^E, H_k[\eta_k^E]\rangle \geq \frac{1}{2}\Delta_k^2 \varepsilon_H$. ∎

Notice from the proof that this strategy either certifies that $\lambda_{\min}(H_k) \geq -\varepsilon_H \operatorname{Id}$ (which must be checked at step 8 in Algorithm 3) or certifies the alternative by providing an escape direction. We further note that, in practice, one may prefer to use iterative methods to compute an approximate leftmost eigenvector of H_k without representing it as a matrix.

3.4 Main results and proofs for RTR

Under the discussed assumptions, we now establish our main theorem about computation of approximate first- and second-order critical points for (P) using RTR in a bounded number of iterations. The following constants will be useful:

$$\lambda_g = \frac{1}{4}\min\left(\frac{1}{c_0}, \frac{c_2}{L_g + c_0}\right) \qquad \text{and} \qquad \lambda_H = \frac{3}{4}\frac{c_3}{L_H + c_1}. \tag{16}$$

Theorem 11. *Under A1, A4, A6, A8 and assuming $\varepsilon_g \leq \frac{\Delta_0}{\lambda_g}$,[4] Algorithm 3 produces an iterate x_{N_1} satisfying $\|\operatorname{grad} f(x_{N_1})\| \leq \varepsilon_g$ with*

$$N_1 \leq \frac{3}{2}\frac{f(x_0) - f^*}{\rho' c_2 \lambda_g}\frac{1}{\varepsilon_g^2} + \frac{1}{2}\log_2\left(\frac{\Delta_0}{\lambda_g \varepsilon_g}\right) = \mathcal{O}\left(\frac{1}{\varepsilon_g^2}\right). \tag{17}$$

[4] Theorem 11 is scale invariant, in that if the cost function $f(x)$ is replaced by $\alpha f(x)$ for some positive α (which does not meaningfully change (P)), it is sensible to also multiply $L_g, L_H, c_0, c_1, \varepsilon_g$ and ε_H by α; consequently, the upper bounds on ε_g and ε_H and the upper bounds on N_1 and N_2 are invariant under this scaling. If it is desirable to always allow $\varepsilon_g, \varepsilon_H$ in, say, $(0, 1]$, one possibility is to artificially make L_g, L_H, c_0, c_1 larger (which is always allowed).

Furthermore, if $\varepsilon_H < \infty$, *then under additional assumptions A5, A7, A9 and assuming* $\varepsilon_g \leq \frac{c_2}{c_3}\frac{\lambda_H}{\lambda_g^2}$ *and* $\varepsilon_H \leq \frac{c_2}{c_3}\frac{1}{\lambda_g}$, *Algorithm 3 also produces an iterate* x_{N_2} *satisfying* $\|\mathrm{grad} f(x_{N_2})\| \leq \varepsilon_g$ *and* $\lambda_{\min}(H_{N_2}) \geq -\varepsilon_H$ *with*

$$N_1 \leq N_2 \leq \frac{3}{2}\frac{f(x_0) - f^*}{\rho' c_3 \lambda^2}\frac{1}{\varepsilon^2 \varepsilon_H} + \frac{1}{2}\log_2\left(\frac{\Delta_0}{\lambda\varepsilon}\right) = \mathcal{O}\left(\frac{1}{\varepsilon^2\varepsilon_H}\right), \tag{18}$$

where we defined $(\lambda, \varepsilon) = (\lambda_g, \varepsilon_g)$ *if* $\lambda_g\varepsilon_g \leq \lambda_H\varepsilon_H$, *and* $(\lambda, \varepsilon) = (\lambda_H, \varepsilon_H)$ *otherwise. Since the algorithm is a descent method,* $f(x_{N_2}) \leq f(x_{N_1}) \leq f(x_0)$.

Remark 12. *Theorem 11 makes a statement about* $\lambda_{\min}(H_k)$ *at termination,* not *about* $\lambda_{\min}(\mathrm{Hess} f(x_k))$. *See Section 3.5 to connect the two.*

To establish Theorem 11, we work through a few lemmas, following the proof technique in (Cartis et al., 2012). We first show Δ_k is bounded below in proportion to the tolerances ε_g and ε_H. This is used to show that the number of successful iterations in Algorithm 3 before termination (that is, iterations where $\rho_k > \rho'$ (8)) is bounded above. It is then shown that the total number of iterations is at most a constant multiple of the number of successful iterations, which implies termination in bounded time. We start by showing that the trust-region radius is bounded away from zero as long as the algorithm does not return. Essentially, this is because if Δ_k becomes too small, then the Cauchy step and eigenstep are certain to be successful owing to the quality of the model in such a small region, so that the trust-region radius could not decrease any further.

Lemma 13. *Under the assumptions of Theorem 11, if Algorithm 3 executes* N *iterations without returning, then*

$$\Delta_k \geq \min\left(\Delta_0, \lambda_g\varepsilon_g, \lambda_H\varepsilon_H\right) \tag{19}$$

for $k = 0, \ldots, N$, *where* λ_g *and* λ_H *are defined in* (16).

Proof. This follows essentially the proof of (Absil et al., 2008, Thm. 7.4.2) which itself follows classical proofs (Conn et al., 2000). The core idea is to control ρ_k (8) close to 1, to show that there cannot be arbitrarily many trust-region radius reductions. The proof is in two parts.

For the first part, assume $\|\mathrm{grad} f(x_k)\| > \varepsilon_g$. Then, consider the gap

$$|\rho_k - 1| = \left|\frac{\hat{f}_k(0_{x_k}) - \hat{f}_k(\eta_k)}{\hat{m}_k(0_{x_k}) - \hat{m}_k(\eta_k)} - 1\right| = \left|\frac{\hat{m}_k(\eta_k) - \hat{f}_k(\eta_k)}{\hat{m}_k(0_{x_k}) - \hat{m}_k(\eta_k)}\right|. \tag{20}$$

From A8, we know the denominator is not too small:

$$\hat{m}_k(0_{x_k}) - \hat{m}_k(\eta_k) \geq c_2 \min\left(\Delta_k, \frac{\varepsilon_g}{c_0}\right)\varepsilon_g.$$

Now consider the numerator:

$$\begin{aligned}
|\hat{m}_k(\eta_k) - \hat{f}_k(\eta_k)| &= \left|f(x_k) + \langle \mathrm{grad}\, f(x_k), \eta_k\rangle + \frac{1}{2}\langle \eta_k, H_k[\eta_k]\rangle - \hat{f}_k(\eta_k)\right| \\
&\leq \left|f(x_k) + \langle \mathrm{grad}\, f(x_k), \eta_k\rangle - \hat{f}_k(\eta_k)\right| + \frac{1}{2}\left|\langle \eta_k, H_k[\eta_k]\rangle\right| \\
&\leq \frac{1}{2}\left(L_g + c_0\right)\|\eta_k\|^2,
\end{aligned}$$

where we used A4 for the first term, and A6 for the second term. Assume for the time being that $\Delta_k \leq \min\left(\frac{\varepsilon_g}{c_0}, \frac{c_2\varepsilon_g}{L_g+c_0}\right) = 4\lambda_g\varepsilon_g$. Then, using $\|\eta_k\| \leq \Delta_k$, it follows that

$$|\rho_k - 1| \leq \frac{1}{2}\frac{L_g + c_0}{c_2 \min\left(\Delta_k, \frac{\varepsilon_g}{c_0}\right)\varepsilon_g}\Delta_k^2 \leq \frac{1}{2}\frac{L_g + c_0}{c_2\varepsilon_g}\Delta_k \leq \frac{1}{2}.$$

Hence, $\rho_k \geq 1/2$, and by the mechanism of Algorithm 3, it follows that $\Delta_{k+1} \geq \Delta_k$.

For the second part, assume $\|\mathrm{grad} f(x_k)\| < \varepsilon_g$ and $\lambda_{\min}(H_k) < -\varepsilon_H$. Then, by A9,

$$\hat{m}_k(0_{x_k}) - \hat{m}_k(\eta_k) \geq c_3\Delta_k^2\varepsilon_H.$$

Thus, by A5 and A7,

$$\begin{aligned}
|\hat{m}_k(\eta_k) - \hat{f}_k(\eta_k)| &= \left|f(x_k) + \langle \mathrm{grad}\, f(x_k), \eta_k\rangle + \frac{1}{2}\langle \eta_k, H_k[\eta_k]\rangle - \hat{f}_k(\eta_k)\right| \\
&\leq \frac{L_H}{6}\|\eta_k\|^3 + \frac{1}{2}\left|\left\langle \eta_k, \left(\nabla^2\hat{f}_k(0_{x_k}) - H_k\right)[\eta_k]\right\rangle\right| \\
&\leq \frac{L_H + c_1}{6}\Delta_k^3.
\end{aligned}$$

As previously, combine these observations into (20) to see that, if $\Delta_k \leq \frac{3c_3}{L_H+c_1}\varepsilon_H = 4\lambda_H\varepsilon_H$, then

$$|\rho_k - 1| \leq \frac{1}{2}\frac{L_H + c_1}{3c_3\varepsilon_H}\Delta_k \leq \frac{1}{2}. \tag{21}$$

Again, this implies $\Delta_{k+1} \geq \Delta_k$.

Now combine the two parts. We have established that, if $\Delta_k \leq 4\min(\lambda_g\varepsilon_g, \lambda_H\varepsilon_H)$, then $\Delta_{k+1} \geq \Delta_k$. To conclude the proof, consider the fact that Algorithm 3 cannot reduce the radius by more than $1/4$ in one step. □

By an argument similar to the one we have seen when studying the gradient descent methods, Lemma 13 implies an upper bound on the number of successful iterations required in Algorithm 3 to reach termination.

Lemma 14. *Under the assumptions of Theorem 11, if Algorithm 3 executes N iterations without returning, define the set of* successful steps *as*

$$S_N = \{k \in \{0, \ldots, N\} : \rho_k > \rho'\}$$

and let U_N designate the unsuccessful steps, so that S_N and U_N form a partition of $\{0, \ldots, N\}$. Assume $\varepsilon_g \leq \Delta_0/\lambda_g$. If $\varepsilon_H = \infty$, the number of successful steps obeys

$$|S_N| \leq \frac{f(x_0) - f^*}{\rho' c_2\lambda_g}\frac{1}{\varepsilon_g^2}. \tag{22}$$

Otherwise, if additionally $\varepsilon_g \leq \frac{c_2}{c_3}\frac{\lambda_H}{\lambda_g^2}$ and $\varepsilon_H \leq \frac{c_2}{c_3}\frac{1}{\lambda_g}$, it is bounded as

$$|S_N| \leq \frac{f(x_0) - f^*}{\rho' c_3}\frac{1}{\min(\lambda_g\varepsilon_g, \lambda_H\varepsilon_H)^2\varepsilon_H}. \tag{23}$$

Proof. The proof parallels (Cartis et al., 2012, Lemma 4.5). Clearly, if $k \in U_N$, then $f(x_k) = f(x_{k+1})$. On the other hand, if $k \in S_N$, then $\rho_k \geq \rho'$ (8). Combine this with A8 and A9 to see that, for $k \in S_N$,

$$\begin{aligned} f(x_k) - f(x_{k+1}) &\geq \rho' \big(\hat{m}_k(0_{x_k}) - \hat{m}_k(\eta_k)\big) \\ &\geq \rho' \min\left(c_2 \min\left(\Delta_k, \frac{\varepsilon_g}{c_0}\right)\varepsilon_g \ , \ c_3 \Delta_k^2 \varepsilon_H\right). \end{aligned}$$

By Lemma 13 and the assumption $\lambda_g \varepsilon_g \leq \Delta_0$, it holds that $\Delta_k \geq \min(\lambda_g \varepsilon_g, \lambda_H \varepsilon_H)$. Furthermore, using $\lambda_g \leq 1/c_0$ shows that $\min(\Delta_k, \varepsilon_g/c_0) \geq \min(\Delta_k, \lambda_g \varepsilon_g) \geq \min(\lambda_g \varepsilon_g, \lambda_H \varepsilon_H)$. Hence,

$$f(x_k) - f(x_{k+1}) \geq \rho' \min\left(c_2 \lambda_g \varepsilon_g^2, c_2 \lambda_H \varepsilon_g \varepsilon_H, c_3 \lambda_g^2 \varepsilon_g^2 \varepsilon_H, c_3 \lambda_H^2 \varepsilon_H^3\right). \tag{24}$$

If $\varepsilon_H = \infty$, this simplifies to

$$f(x_k) - f(x_{k+1}) \geq \rho' c_2 \lambda_g \varepsilon_g^2.$$

Sum over iterations up to N and use A1 (bounded f):

$$f(x_0) - f^* \geq f(x_0) - f(x_{N+1}) = \sum_{k \in S_N} f(x_k) - f(x_{k+1}) \geq |S_N| \rho' c_2 \lambda_g \varepsilon_g^2.$$

Hence,

$$|S_N| \leq \frac{f(x_0) - f^*}{\rho' c_2 \lambda_g} \frac{1}{\varepsilon_g^2}.$$

On the other hand, if $\varepsilon_H < \infty$, then, starting over from (24) and assuming both $c_3 \lambda_g^2 \varepsilon_g^2 \varepsilon_H \leq c_2 \lambda_H \varepsilon_g \varepsilon_H$ and $c_3 \lambda_g^2 \varepsilon_g^2 \varepsilon_H \leq c_2 \lambda_g \varepsilon_g^2$ (which is equivalent to $\varepsilon_g \leq c_2 \lambda_H / c_3 \lambda_g^2$ and $\varepsilon_H \leq c_2 / c_3 \lambda_g$), it comes with the same telescoping sum that

$$f(x_0) - f^* \geq |S_N| \rho' c_3 \min(\lambda_g \varepsilon_g, \lambda_H \varepsilon_H)^2 \varepsilon_H.$$

Solve for $|S_N|$ to conclude. ∎

Finally, we show that the total number of steps N before termination cannot be more than a fixed multiple of the number of successful steps $|S_N|$.

Lemma 15. *Under the assumptions of Theorem 11, if Algorithm 3 executes N iterations without returning, using the notation S_N and U_N of Lemma 14, it holds that*

$$|S_N| \geq \frac{2}{3}(N+1) - \frac{1}{3} \max\left(0, \log_2\left(\frac{\Delta_0}{\lambda_g \varepsilon_g}\right), \log_2\left(\frac{\Delta_0}{\lambda_H \varepsilon_H}\right)\right). \tag{25}$$

Proof. The proof rests on the lower bound for Δ_k obtained in Lemma 13. It parallels (Cartis et al., 2012, Lemma 4.6). For all $k \in S_N$, it holds that $\Delta_{k+1} \leq 2\Delta_k$. For all $k \in U_k$, it holds that $\Delta_{k+1} \leq \frac{1}{4}\Delta_k$. Hence,

$$\Delta_N \leq 2^{|S_N|} \left(\frac{1}{4}\right)^{|U_N|} \Delta_0.$$

On the other hand, Lemma 13 gives

$$\Delta_N \geq \min\left(\Delta_0, \lambda_g \varepsilon_g, \lambda_H \varepsilon_H\right).$$

Combine, divide by Δ_0 and take the log in base 2:

$$|S_N| - 2|U_N| \geq \min\left(0, \log_2\left(\frac{\lambda_g \varepsilon_g}{\Delta_0}\right), \log_2\left(\frac{\lambda_H \varepsilon_H}{\Delta_0}\right)\right).$$

Use $|S_N| + |U_N| = N + 1$ to conclude. □

We can now prove the main theorem.

Proof of Theorem 11. It is sufficient to combine Lemmas 14 and 15 in both regimes. First, we get that if $\|\mathrm{grad} f(x_k)\| > \varepsilon_g$ for $k = 0, \ldots, N$, then

$$N + 1 \leq \frac{3}{2} \frac{f(x_0) - f^*}{\rho' c_2 \lambda_g} \frac{1}{\varepsilon_g^2} + \frac{1}{2} \log_2\left(\frac{\Delta_0}{\lambda_g \varepsilon_g}\right).$$

(The term $\log_2\left(\frac{\Delta_0}{\lambda_H \varepsilon_H}\right)$ from Lemma 15 is irrelevant up to that point, as ε_H could just as well have been infinite.) Thus, after a number of iterations larger than the right hand side, an iterate with sufficiently small gradient must have been produced, to avoid a contradiction.

Second, we get that if for $k = 0, \ldots, N$ no iterate satisfies both $\|\mathrm{grad} f(x_k)\| \leq \varepsilon_g$ and $\lambda_{\min}(H_k) \geq -\varepsilon_H$, then

$$N + 1 \leq \frac{3}{2} \frac{f(x_0) - f^*}{\rho' c_3} \frac{1}{\min(\lambda_g \varepsilon_g, \lambda_H \varepsilon_H)^2 \varepsilon_H} + \frac{1}{2} \max\left(\log_2\left(\frac{\Delta_0}{\lambda_g \varepsilon_g}\right), \log_2\left(\frac{\Delta_0}{\lambda_H \varepsilon_H}\right)\right).$$

Conclude with the same argument. □

3.5 Connecting H_k and $\mathrm{Hess} f(x_k)$

Theorem 11 states termination of Algorithm 3 in terms of $\|\mathrm{grad} f(x_k)\|$ and $\lambda_{\min}(H_k)$. Ideally, the latter must be turned into a statement about $\lambda_{\min}(\mathrm{Hess} f(x_k))$, to match the second-order necessary optimality conditions of (P) more closely (recall Proposition 1). A7 itself only requires H_k to be related to the Hessian of the pullback of f at x_k, which is different from the Riemannian Hessian of f at x_k in general. Furthermore, since Δ_k may be large and A7 controls the error only along η_k, A7 is too permissive to guarantee much even in terms of the pullback's Hessian.

For a general retraction (Definition 1), the Riemannian Hessian and the pullback's Hessian agree at critical points (Absil et al., 2008, Prop. 5.5.6).

Proposition 16. *If* $\mathrm{grad} f(x_k) = 0_{x_k}$, *then* $\nabla^2 \hat{f}_k(0_{x_k}) = \mathrm{Hess} f(x_k)$.

However, this is insufficient in our setting since only approximate critical points can be reached in practice. By requiring more of the retraction, the two notions can be made to coincide globally (Absil et al., 2008, Prop. 5.5.5).

Definition 2. *A retraction* Retr *is a* second-order retraction *if*

$$\forall x \in \mathcal{M}, \forall \eta \in \mathrm{T}_x\mathcal{M}, \quad \frac{\mathrm{D}^2}{\mathrm{d}t^2}\mathrm{Retr}_x(t\eta)\Big|_{t=0} = 0_x,$$

where $\frac{\mathrm{D}^2}{\mathrm{d}t^2}\gamma$ *denotes acceleration of the curve* γ *on* $\mathcal{M}$*—see (Absil et al., 2008, §5). Thus: retracted curves locally agree with geodesics up to second order.*

Proposition 17. *If* Retr *is a second-order retraction, then* $\nabla^2 \hat{f}_k(0_{x_k}) = \mathrm{Hess} f(x_k)$.

This discussion suggests that the most direct route to ensure $\mathrm{Hess} f(x_k) \succeq -\varepsilon_H \,\mathrm{Id}$ holds when Algorithm 3 returns is to use a second-order retraction and set

$$H_k \triangleq \nabla^2 \hat{f}_k(0_{x_k}) \overset{\text{Prop. 17}}{=} \mathrm{Hess} f(x_k).$$

This recommendation is made practical by noting that retractions for submanifolds obtained as (certain types of) projections—arguably one of the most natural classes of retractions for submanifolds—are second order (Absil and Malick, 2012, Thm. 22). For example, the sphere retraction $\mathrm{Retr}_x(\eta) = (x+\eta)/\|x+\eta\|$ is second order. This also ensures A7 holds with $c_1 = 0$.

See Appendix C for a discussion in case no second-order retraction is available.

4 Example: smooth semidefinite programs

This example is based on (Boumal et al., 2016). Consider the following semidefinite program, which occurs in robust PCA (McCoy and Tropp, 2011) and as a convex relaxation of combinatorial problems such as Max-Cut, $\mathbb{Z}_2$-synchronization and community detection in the stochastic block model (Goemans and Williamson, 1995; Bandeira et al., 2016):

$$\min_{X \in \mathbb{R}^{n\times n}} \mathrm{Tr}(CX) \text{ subject to } \mathrm{diag}(X) = \mathbf{1}, X \succeq 0. \tag{26}$$

The symmetric cost matrix C depends on the application. Interior point methods solve this problem in polynomial time, but involve significant work to enforce the conic constraint $X \succeq 0$ (X symmetric, positive semidefinite). To avoid this, one possibility is to redundantly parameterize the search space as $X = YY^\top$, where Y is in $\mathbb{R}^{n\times p}$ for some well-chosen $p \geq n$ (Burer and Monteiro, 2005):

$$\min_{Y \in \mathbb{R}^{n\times p}} \mathrm{Tr}(CYY^\top) \text{ subject to } \mathrm{diag}(YY^\top) = \mathbf{1}. \tag{27}$$

This problem is of the form of (P), where $f(Y) = \mathrm{Tr}(CYY^\top)$ and the manifold is a product of n unit spheres in $\mathbb{R}^p$:

$$\mathcal{M} = \{Y \in \mathbb{R}^{n\times p} : \mathrm{diag}(YY^\top) = \mathbf{1}\} = \{Y \in \mathbb{R}^{n\times p} : \text{each row of } Y \text{ has unit norm}\}.$$

In principle, since the parameterization $X = YY^\top$ breaks convexity, the new problem could have many spurious local optimizers and saddle points. Yet, for $p = n + 1$, it is known that approximate second-order critical points Y map to approximate global optimizers as $X = YY^\top$, as stated in the following proposition. (For this particular case, there is no explicit need to control $\|\mathrm{grad} f(Y)\|$.)

Proposition 18 (Boumal et al. (2016)). *If $X^\star$ is optimal for (26) and Y is feasible for (27) with $p > n$ and $\mathrm{Hess} f(Y) \succeq -\varepsilon_H \mathrm{Id}$, then the optimality gap is bounded as*

$$0 \leq \mathrm{Tr}(CYY^\top) - \mathrm{Tr}(CX^\star) \leq \frac{n}{2}\varepsilon_H.$$

Since f is smooth in $\mathbb{R}^{n\times p}$ and $\mathcal{M}$ is a compact submanifold of $\mathbb{R}^{n\times p}$, the regularity assumptions A4 and A5 hold with any second-order retraction (Lemmas 3 and 8). In particular, they hold if $\mathrm{Retr}_Y(\dot{Y})$ is the result of normalizing each row of $Y+\dot{Y}$ (Section 3.5). Theorem 11 then implies that RTR applied to the nonconvex problem (27) computes a point $X = YY^\top$ feasible for (26) such that $\mathrm{Tr}(CX) - \mathrm{Tr}(CX^\star) \leq \delta$ in $\mathcal{O}(1/\delta^3)$ iterations. (To obtain the complexity in n, it would be necessary to bound the Lipschitz constants L_g and L_H appearing in A4 and A5 in terms of n.) Each iteration involves a product CY, potentially computing an eigenstep (the Hessian is a structured expression involving C and Y), and a retraction. The constraints are satisfied up to numerical accuracy at trivial cost. While this worst-case bound suggests it may be expensive to obtain high-accuracy solutions via this method, practice shows this is not the case. Indeed, in the numerical experiments in (Boumal, 2015b), the local behavior of RTR is typical of a superlinear local convergence rate.

We further note that, in fact, problems (26) and (27) are already equivalent (that is, have the same optimal value) for p as small as $\lceil\sqrt{2n}\rceil$ (Burer and Monteiro, 2005). This is because the SDP (26) always admits a low-rank solution. In (Boumal et al., 2016), it is shown that, generically in C, if $p \geq \lceil\sqrt{2n}\rceil$, then all second-order critical points of (27) are globally optimal (despite nonconvexity). This means RTR globally converges to global optimizers with cheaper iterations (due to reduced dimensionality); but there is no statement of quality pertaining to *approximate* second-order critical points for such small p.

5 Conclusions and perspectives

In the context of optimization on manifolds (P), we presented bounds on the number of iterations required by the Riemannian gradient descent algorithm and the Riemannian trust-region algorithm to reach points which approximately satisfy first- and second-order necessary optimality conditions, under some regularity assumptions but regardless of initialization. When the search space $\mathcal{M}$ is a Euclidean space, these bounds were already known. For the more general case of $\mathcal{M}$ being a Riemannian manifold, these bounds are new.

As a subclass of interest, we showed the regularity requirements are satisfied if $\mathcal{M}$ is a compact submanifold of $\mathbb{R}^n$ and f has locally Lipschitz continuous derivatives of appropriate order. This covers a rich class of practical optimization problems. While there are no explicit assumptions made about $\mathcal{M}$, the smoothness requirements for the pullback of the cost—A3, A4 and A5—implicitly restrict the class of manifolds to which these results apply. Indeed, for certain manifolds, even for nice cost functions f, there may not exist retractions which ensure the assumptions hold. This is the case in particular for certain incomplete manifolds, such as open Riemannian submanifolds of $\mathbb{R}^n$ and certain geometries of the set of fixed-rank matrices—see also Remark 7 about injectivity radius. For such sets, it may be necessary to adapt the assumptions. For fixed-rank matrices for example, Vandereycken (2013, §4.1) obtains convergence results assuming a kind of coercivity on the cost function: for any sequence of rank-k matrices $(X_i)_{i=1,2,\ldots}$ such that the first singular value $\sigma_1(X_i) \to \infty$ or the kth singular value $\sigma_k(X_i) \to 0$, it holds that $f(X_i) \to \infty$. The iteration bounds are sharp, but additional

information may yield more favorable bounds in more specific contexts. In particular, when the studied algorithms converge to a nondegenerate local optimizer, they do so with an at least linear rate, so that the number of iterations is merely $\mathcal{O}(\log(1/\varepsilon))$ once in the linear regime. This suggests a stitching approach: for a given application, it may be possible to show that rough approximate second-order critical points are in a local attraction basin; the iteration cost can then be bounded by the total work needed to attain such a crude point starting from anywhere, plus the total work needed to refine the crude point to high accuracy. This is, to some degree, the successful strategy in (Sun et al., 2015, 2016).

Finally, we note that it would also be interesting to study the global convergence rates of Riemannian versions of adaptive regularization algorithms using cubics (ARC), as in the Euclidean case these can achieve approximate first-order criticality in $\mathcal{O}(1/\varepsilon^{1.5})$ instead of $\mathcal{O}(1/\varepsilon^2)$ (Cartis et al., 2011a). Work in that direction could start with the convergence analyses proposed in (Qi, 2011).

Acknowledgments

NB was supported by the "Fonds Spéciaux de Recherche" (FSR) at UCLouvain and by the Chaire Havas "Chaire Economie et gestion des nouvelles données", the ERC Starting Grant SIPA and a Research in Paris grant at Inria & ENS. This paper presents research results of the Belgian Network DYSCO (Dynamical Systems, Control, and Optimization), funded by the Interuniversity Attraction Poles Programme initiated by the Belgian Science Policy Office. This work was supported by the ARC "Mining and Optimization of Big Data Models". CC acknowledges support from NERC through grant NE/L012146/1. We thank Alex d'Aspremont, Simon Lacoste-Julien, Ju Sun, Bart Vandereycken and Paul Van Dooren for helpful discussions.

A Essentials about manifolds

We give here a simplified refresher of differential geometric concepts used in the paper, restricted to Riemannian submanifolds. All concepts are illustrated with the sphere. See (Absil et al., 2008) for a more complete discussion, including quotient manifolds.

We endow $\mathbb{R}^n$ with the classical Euclidean metric: for all $x, y \in \mathbb{R}^n$, $\langle x, y\rangle = x^\top y$. Consider the smooth map $h\colon \mathbb{R}^n \mapsto \mathbb{R}^m$ with $m \leq n$ and the constraint set

$$\mathcal{M} = \{x \in \mathbb{R}^n : h(x) = 0\}.$$

This set is a submanifold of dimension $n - m$ of $\mathbb{R}^n$ if it is linearized at each point by a tangent space of dimension $n - m$ (Absil et al., 2008, Prop. 3.3.3). Translated to the origin, this subspace is the kernel of the differential of h at x (Absil et al., 2008, eq. (3.19)):

$$\mathrm{T}_x\mathcal{M} = \{\eta \in \mathbb{R}^n : \mathrm{D}h(x)[\eta] = 0\}.$$

For example, the unit sphere in $\mathbb{R}^n$ is a submanifold of dimension $n - 1$ defined by

$$\mathcal{S}^{n-1} = \{x \in \mathbb{R}^n : x^\top x = 1\},$$

and the tangent space at x is

$$\mathrm{T}_x\mathcal{S}^{n-1} = \{\eta \in \mathbb{R}^n : x^\top \eta = 0\}.$$

By endowing each tangent space with the (restricted) Euclidean metric, we turn $\mathcal{M}$ into a Riemannian submanifold of the Euclidean space $\mathbb{R}^n$. (In general, the metric could be different, and would depend on x; to disambiguate, one would write $\langle \cdot, \cdot \rangle_x$.) An obvious retraction for the sphere (see Definition 1) is to normalize:

$$\mathrm{Retr}_x(\eta) = \frac{x+\eta}{\|x+\eta\|}.$$

Being an orthogonal projection to the manifold, this is actually a second-order retraction, see Definition 2 and (Absil and Malick, 2012, Thm. 22).

The Riemannian metric leads to the notion of Riemannian gradient of a real function f defined in an open set of $\mathbb{R}^n$ containing $\mathcal{M}$.[5] The Riemannian gradient of f at x is the (unique) tangent vector $\mathrm{grad} f(x)$ at x satisfying

$$\forall \eta \in \mathrm{T}_x\mathcal{M}, \quad \mathrm{D}f(x)[\eta] = \lim_{t \to 0} \frac{f(x+t\eta) - f(x)}{t} = \langle \eta, \mathrm{grad} f(x) \rangle \,. \tag{28}$$

In this setting, the Riemannian gradient is nothing but the orthogonal projection of the Euclidean (classical) gradient $\nabla f(x)$ to the tangent space. Writing $\mathrm{Proj}_x \colon \mathbb{R}^n \to \mathrm{T}_x\mathcal{M}$ for the orthogonal projector, we have (Absil et al., 2008, eq. (3.37)):

$$\mathrm{grad} f(x) = \mathrm{Proj}_x \nabla f(x).$$

Continuing the sphere example, the orthogonal projector is $\mathrm{Proj}_x(y) = y - (x^\top y)x$, and if $f(x) = \frac{1}{2}x^\top A x$ for some symmetric matrix A, then

$$\nabla f(x) = Ax, \qquad \text{and} \qquad \mathrm{grad} f(x) = Ax - (x^\top A x)x.$$

Notice that the critical points of f on $\mathcal{S}^{n-1}$ coincide with the unit eigenvectors of A.

We can further define a notion of Riemannian Hessian as the projected differential of the Riemannian gradient:[6]

$$\mathrm{Hess} f(x)[\eta] = \mathrm{Proj}_x\Big(\mathrm{D}\big(x \mapsto \mathrm{Proj}_x \nabla f(x)\big)(x)[\eta]\Big).$$

$\mathrm{Hess} f(x)$ is a linear map from $\mathrm{T}_x\mathcal{M}$ to itself, symmetric with respect to the Riemannian metric. Given a second-order retraction (Definition 2), it is equivalently defined by:

$$\forall \eta \in \mathrm{T}_x\mathcal{M}, \quad \langle \eta, \mathrm{Hess} f(x)[\eta] \rangle = \frac{\mathrm{d}^2}{\mathrm{d}t^2} f(\mathrm{Retr}_x(t\eta))\Big|_{t=0},$$

see (Absil et al., 2008, eq. (5.35)). Continuing our sphere example,

$$\begin{aligned} \mathrm{D}\big(x \mapsto \mathrm{Proj}_x \nabla f(x)\big)(x)[\eta] &= \mathrm{D}\big(x \mapsto Ax - (x^\top A x)x\big)(x)[\eta] \\ &= A\eta - (x^\top A x)\eta - 2(x^\top A \eta)x. \end{aligned}$$

[5] It is not necessary to have f defined outside of $\mathcal{M}$, but this is often the case in applications and simplifies exposition.

[6] Proper definition of Riemannian Hessians requires the notion of Riemannian connections, which we omit here; see (Absil et al., 2008, §5)

Projection of the latter gives the Hessian:

$$\mathrm{Hess} f(x)[\eta] = \mathrm{Proj}_x(A\eta) - (x^\top A x)\eta.$$

Notice that the Hessian is positive semidefinite (on the tangent space) if and only if

$$\begin{aligned} \mathrm{Hess} f(x) \succeq 0 &\iff \langle \eta, \mathrm{Hess} f(x)[\eta] \rangle \geq 0 && \forall \eta \in \mathrm{T}_x\mathcal{S}^{n-1} \\ &\iff \eta^\top A \eta \geq x^\top A x && \forall \eta \in \mathrm{T}_x\mathcal{S}^{n-1}, \|\eta\| = 1. \end{aligned}$$

Together with first-order conditions, this implies that x is a leftmost eigenvector of A.[7] This is an example of optimization problem on a manifold for which second-order necessary optimality conditions are also sufficient. This is not the norm.

As another (very) special example, consider the case $\mathcal{M} = \mathbb{R}^n$; then, $\mathrm{T}_x\mathbb{R}^n = \mathbb{R}^n$, $\mathrm{Retr}_x(\eta) = x + \eta$ is the exponential map (a fortiori a second-order retraction), Proj_x is the identity, $\mathrm{grad} f(x) = \nabla f(x)$ and $\mathrm{Hess} f(x) = \nabla^2 f(x)$.

B Compact submanifolds of Euclidean spaces

In this appendix, we prove Lemmas 3 and 8, showing that if f has locally Lipschitz continuous gradient or Hessian in a Euclidean space $\mathcal{E}$ (in the usual sense), and it is to be minimized over a compact submanifold of $\mathcal{E}$, then the results in this paper apply, in that A3, A4 and A5 hold. We use the notation of Appendix A relative to submanifolds.

Proof of Lemma 3. By assumption, ∇f is Lipschitz continuous along any line segment in $\mathcal{E}$ joining x and y in $\mathcal{M}$. Hence, there exists L such that, for all $x, y \in \mathcal{M}$,

$$\big| f(y) - [f(x) + \langle \nabla f(x), y - x \rangle] \big| \leq \frac{L}{2}\|y - x\|^2. \tag{29}$$

In particular, this holds for all $y = \mathrm{Retr}_x(\eta)$, for any $\eta \in \mathrm{T}_x\mathcal{M}$. Writing $\mathrm{grad} f(x)$ for the Riemannian gradient of $f|_\mathcal{M}$ and using that $\mathrm{grad} f(x)$ is the orthogonal projection of $\nabla f(x)$ to $\mathrm{T}_x\mathcal{M}$, the inner product above decomposes as

$$\begin{aligned} \langle \nabla f(x), \mathrm{Retr}_x(\eta) - x \rangle &= \langle \nabla f(x), \eta + \mathrm{Retr}_x(\eta) - x - \eta \rangle \\ &= \langle \mathrm{grad} f(x), \eta \rangle + \langle \nabla f(x), \mathrm{Retr}_x(\eta) - x - \eta \rangle . \end{aligned} \tag{30}$$

Combining (29) with (30) and using the triangle inequality yields

$$\begin{aligned} &\big| f(\mathrm{Retr}_x(\eta)) - [f(x) + \langle \mathrm{grad} f(x), \eta \rangle] \big| \\ &\qquad \leq \frac{L}{2}\|\mathrm{Retr}_x(\eta) - x\|^2 + \|\nabla f(x)\| \|\mathrm{Retr}_x(\eta) - x - \eta\|. \end{aligned}$$

Since $\nabla f(x)$ is continuous on the compact set $\mathcal{M}$, there exists G finite such that $\|\nabla f(x)\| \leq G$ for all $x \in \mathcal{M}$. It remains to show there exist finite constants $\alpha, \beta \geq 0$ such that, for all $x \in \mathcal{M}$ and for all $\eta \in \mathrm{T}_x\mathcal{M}$,

$$\|\mathrm{Retr}_x(\eta) - x\| \leq \alpha\|\eta\|, \text{ and} \tag{31}$$

$$\|\mathrm{Retr}_x(\eta) - x - \eta\| \leq \beta\|\eta\|^2. \tag{32}$$

[7] Indeed, any $y \in \mathcal{S}^{n-1}$ can be written as $y = \alpha x + \beta\eta$ with $x^\top\eta = 0$ and $\alpha^2 + \beta^2 = 1$; then, $y^\top A y = \alpha^2 x^\top A x + \beta^2 \eta^\top A \eta + 2\eta^\top A x$; by first-order condition, $\eta^\top A x = (x^\top A x)\eta^\top x = 0$, and by second-order condition: $y^\top A y \geq (\alpha^2 + \beta^2) x^\top A x = x^\top A x$.

For small η, this will follow from $\mathrm{Retr}_x(\eta) = x + \eta + \mathcal{O}(\|\eta\|^2)$ by Definition 1; for large η this will follow a fortiori from compactness. This will be sufficient the conclude, as then we will have for all $x \in \mathcal{M}$ and $\eta \in \mathrm{T}_x\mathcal{M}$ that

$$\left| f(\mathrm{Retr}_x(\eta)) - [f(x) + \langle \mathrm{grad} f(x), \eta \rangle] \right| \leq \left(\frac{L}{2}\alpha^2 + G\beta \right) \|\eta\|^2.$$

More formally, Definition 1 a fortiori ensures the existence of $r > 0$ such that Retr is smooth on $K = \{\eta \in \mathrm{T}\mathcal{M} : \|\eta\| \leq r\}$, a compact subset of the tangent bundle (K consists of a ball in each tangent space). First, we determine α (31).

For all $\eta \in K$, we have

$$\begin{aligned}
\|\mathrm{Retr}_x(\eta) - x\| &\leq \int_0^1 \left\| \frac{\mathrm{d}}{\mathrm{d}t} \mathrm{Retr}_x(t\eta) \right\| \mathrm{d}t = \int_0^1 \|\mathrm{DRetr}_x(t\eta)[\eta]\| \mathrm{d}t \\
&\leq \int_0^1 \max_{\xi \in K} \|\mathrm{DRetr}(\xi)\| \|\eta\| \mathrm{d}t = \max_{\xi \in K} \|\mathrm{DRetr}(\xi)\| \|\eta\|,
\end{aligned}$$

where the max exists and is finite owing to compactness of K and smoothness of Retr on K; note that this is uniform over both x and η. (If $\xi \in \mathrm{T}_z\mathcal{M}$, the notation $\mathrm{DRetr}(\xi)$ refers to $\mathrm{DRetr}_z(\xi)$.)

For all $\eta \notin K$, we have

$$\|\mathrm{Retr}_x(\eta) - x\| \leq \mathrm{diam}(\mathcal{M}) \leq \frac{\mathrm{diam}(\mathcal{M})}{r} \|\eta\|,$$

where $\mathrm{diam}(\mathcal{M})$ is the maximal distance between any two points on $\mathcal{M}$: finite by compactness of $\mathcal{M}$. Combining, we find that (31) holds with

$$\alpha = \max\left(\max_{\xi \in K} \|\mathrm{DRetr}(\xi)\|, \frac{\mathrm{diam}(\mathcal{M})}{r} \right).$$

Inequality (32) is established along similar lines. For all $\eta \in K$, we have

$$\begin{aligned}
\|\mathrm{Retr}_x(\eta) - x - \eta\| &\leq \int_0^1 \left\| \frac{\mathrm{d}}{\mathrm{d}t} (\mathrm{Retr}_x(t\eta) - x - t\eta) \right\| \mathrm{d}t = \int_0^1 \|\mathrm{DRetr}_x(t\eta)[\eta] - \eta\| \mathrm{d}t \\
&\leq \int_0^1 \|\mathrm{DRetr}_x(t\eta) - \mathrm{Id}\| \|\eta\| \mathrm{d}t \leq \max_{\xi \in K} \|\mathrm{D}^2\mathrm{Retr}(\xi)\| \|\eta\|^2,
\end{aligned}$$

where the last inequality follows from $\mathrm{DRetr}_x(0_x) = \mathrm{Id}$ and

$$\|\mathrm{DRetr}_x(t\eta) - \mathrm{Id}\| \leq \int_0^1 \left\| \frac{\mathrm{d}}{\mathrm{d}s} \mathrm{DRetr}_x(st\eta) \right\| \mathrm{d}s \leq \|t\eta\| \int_0^1 \left\| \mathrm{D}^2\mathrm{Retr}_x(t\eta) \right\| \mathrm{d}s.$$

The case $\eta \notin K$ is treated as before:

$$\|\mathrm{Retr}_x(\eta) - x - \eta\| \leq \|\mathrm{Retr}_x(\eta) - x\| + \|\eta\| \leq \frac{\mathrm{diam}(\mathcal{M}) + r}{r^2} \|\eta\|^2.$$

Combining, we find that (32) holds with

$$\beta = \max\left(\max_{\xi \in K} \|\mathrm{D}^2\mathrm{Retr}(\xi)\|, \frac{\mathrm{diam}(\mathcal{M}) + r}{r^2} \right),$$

which concludes the proof. □

We now prove the corresponding second-order result, whose aim is to verify A5.

Proof of Lemma 8. By assumption, $\nabla^2 f$ is Lipschitz continuous along any line segment in $\mathcal{E}$ joining x and y in $\mathcal{M}$. Hence, there exists L such that, for all $x, y \in \mathcal{M}$,

$$\left| f(y) - \left[f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2} \left\langle y - x, \nabla^2 f(x)[y - x] \right\rangle \right] \right| \leq \frac{L}{6} \|y - x\|^3. \tag{33}$$

Fix $x \in \mathcal{M}$. Let Proj_x denote the orthogonal projector from $\mathcal{E}$ to $\mathrm{T}_x\mathcal{M}$. Let $\mathrm{grad} f(x)$ be the Riemannian gradient of $f|_{\mathcal{M}}$ at x and let $\mathrm{Hess} f(x)$ be the Riemannian Hessian of $f|_{\mathcal{M}}$ at x (a symmetric operator on $\mathrm{T}_x\mathcal{M}$). Following Appendix A, we have

$$\begin{aligned}
\mathrm{grad} f(x) &= \mathrm{Proj}_x \nabla f(x), \text{ and} \\
\langle \eta, \mathrm{Hess} f(x)[\eta] \rangle &= \langle \eta, \mathrm{D}\big(x \mapsto \mathrm{Proj}_x \nabla f(x)\big)(x)[\eta] \rangle \\
&= \Big\langle \eta, \Big(\mathrm{D}\big(x \mapsto \mathrm{Proj}_x\big)(x)[\eta]\Big)[\nabla f(x)] + \mathrm{Proj}_x \nabla^2 f(x)[\eta] \Big\rangle \\
&= \langle I\!I(\eta, \eta), \nabla f(x) \rangle + \left\langle \eta, \nabla^2 f(x)[\eta] \right\rangle,
\end{aligned}$$

where $I\!I$ is the second fundamental form of $\mathcal{M}$: $I\!I(\eta, \eta)$ is a normal vector to the tangent space at x, capturing the second-order geometry of $\mathcal{M}$—see (Absil et al., 2009, 2013; Monera et al., 2014) for presentations relevant to our setting. In particular, $I\!I(\eta, \eta)$ is the acceleration in $\mathcal{E}$ at x of the geodesic $\gamma(t)$ on $\mathcal{M}$ defined by $\gamma(0) = x$ and $\gamma'(0) = \eta$: $\gamma''(0) = I\!I(\eta, \eta)$.

Let $\eta \in \mathrm{T}_x\mathcal{M}$ be arbitrary; $y = \mathrm{Retr}_x(\eta) \in \mathcal{M}$. Then,

$$\begin{aligned}
\langle \nabla f(x), y - x \rangle - \langle \mathrm{grad} f(x), \eta \rangle &= \langle \nabla f(x), y - x - \eta \rangle \text{ and} \\
\left\langle y - x, \nabla^2 f(x)[y - x] \right\rangle - \langle \eta, \mathrm{Hess} f(x)[\eta] \rangle &= 2 \left\langle \eta, \nabla^2 f(x)[y - x - \eta] \right\rangle \\
&\quad + \left\langle y - x - \eta, \nabla^2 f(x)[y - x - \eta] \right\rangle \\
&\quad - \langle \nabla f(x), I\!I(\eta, \eta) \rangle.
\end{aligned}$$

Since $\mathcal{M}$ is compact and f is twice continuously differentiable, there exist G, H, independent of x, such that $\|\nabla f(x)\| \leq G$ and $\|\nabla^2 f(x)\| \leq H$ (the latter is the induced operator norm). Combining with (33) and using the triangle and Cauchy–Schwarz inequalities multiple times gives

$$\begin{aligned}
&\left| f(y) - \left[f(x) + \langle \mathrm{grad} f(x), \eta \rangle + \frac{1}{2} \langle \eta, \mathrm{Hess} f(x)[\eta] \rangle \right] \right| \\
&\qquad \leq \frac{L}{6} \|y - x\|^3 + G \left\| y - x - \eta - \frac{1}{2} I\!I(\eta, \eta) \right\| + H \|\eta\| \|y - x - \eta\| + \frac{1}{2} H \|y - x - \eta\|^2.
\end{aligned}$$

Using the same argument as in Lemma 3, we can find finite constants α, β independent of x and η such that (31) and (32) hold. Use $\|y - x - \eta\|^2 \leq \|y - x - \eta\| \left(\|y - x\| + \|\eta\|\right) \leq \beta(\alpha + 1)\|\eta\|^3$ to bound the right hand side above with

$$\left(\frac{L}{6} \alpha^3 + H\beta + \frac{H\beta(\alpha + 1)}{2} \right) \|\eta\|^3 + G \left\| y - x - \eta - \frac{1}{2} I\!I(\eta, \eta) \right\|.$$

Now, we use the fact that

$$\mathrm{Retr}_x(\eta) = x + \eta + \frac{1}{2} I\!I(\eta, \eta) + \mathcal{O}(\|\eta\|^3)$$

because Retr is a second-order retraction, and as such it agrees with geodesics up to second order (Absil et al., 2009; Monera et al., 2014). This provides a suitable bound for small η; for large η, combine with the same argument as in Lemma 3, using compactness of $\mathcal{M}$. This ensures the existence of a finite constant γ, independent of x and η, such that

$$\left\| y - x - \eta - \frac{1}{2}II(\eta,\eta) \right\| \leq \gamma \|\eta\|^3.$$

See (Absil et al., 2009, §6, eq. (44) and last eq. of p19) for more details. Combining, we find that for all $x \in \mathcal{M}$ and $\eta \in \mathrm{T}_x\mathcal{M}$,

$$\left| f(\mathrm{Retr}_x(\eta)) - \left[f(x) + \langle \mathrm{grad} f(x), \eta \rangle + \frac{1}{2} \langle \eta, \mathrm{Hess} f(x)[\eta] \rangle \right] \right| \leq \left(\frac{L}{6}\alpha^3 + \frac{H\beta(\alpha+3)}{2} + \gamma \right) \|\eta\|^3.$$

Since Retr is a second-order retraction, $\mathrm{Hess} f(x)$ coincides with the Hessian of the pullback $f \circ \mathrm{Retr}_x$ (Proposition 17). This establishes A5. □

C In the absence of second-order retraction

If $\varepsilon_H < \infty$ and Algorithm 3 returns at iteration k, then $H_k \succeq -\varepsilon_H \,\mathrm{Id}$ (Theorem 11). The goal of this appendix, as a follow up to Section 3.5, is to obtain a similar statement about the Riemannian Hessian at x_k, if only a first-order retraction is available.

It is up to the user to ensure H_k is close to the pullback Hessian in operator norm on the tangent space at x_k:

$$\left\| \nabla^2 \hat{f}_k(0_{x_k}) - H_k \right\| \leq \delta_k,$$

with $\delta_k \leq \frac{c_1 \Delta_k}{3}$ to satisfy A7. If the retraction is second order (Definition 2), then $\nabla^2 \hat{f}_k(0_{x_k}) = \mathrm{Hess} f(x_k)$ (Proposition 17). Otherwise, these two operators may differ at non-critical points (Proposition 16). We give here a bound on this difference when the retraction is first order and has bounded acceleration (rather than zero) at 0_{x_k}.

The Hessian of f and that of the pullback are related by the following formulas. See (Absil et al., 2008, §5) for the precise meanings of the differential operators D and d. For all η in $\mathrm{T}_x\mathcal{M}$, writing $\hat{f}_x = f \circ \mathrm{Retr}_x$ for convenience,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} f(\mathrm{Retr}_x(t\eta)) &= \left\langle \mathrm{grad} f(\mathrm{Retr}_x(t\eta)), \frac{\mathrm{D}}{\mathrm{d}t}\mathrm{Retr}_x(t\eta) \right\rangle, \\
\left\langle \nabla^2 \hat{f}_x(0_x)[\eta], \eta \right\rangle &= \frac{\mathrm{d}^2}{\mathrm{d}t^2} f(\mathrm{Retr}_x(t\eta))\Big|_{t=0} \\
&= \left\langle \mathrm{Hess} f(x)\left[\mathrm{DRetr}_x(0_x)[\eta]\right], \frac{\mathrm{D}}{\mathrm{d}t}\mathrm{Retr}_x(t\eta)\Big|_{t=0} \right\rangle \\
&\quad + \left\langle \mathrm{grad} f(x), \frac{\mathrm{D}^2}{\mathrm{d}t^2}\mathrm{Retr}_x(t\eta)\Big|_{t=0} \right\rangle \\
&= \langle \mathrm{Hess} f(x)[\eta], \eta \rangle + \left\langle \mathrm{grad} f(x), \frac{\mathrm{D}^2}{\mathrm{d}t^2}\mathrm{Retr}_x(t\eta)\Big|_{t=0} \right\rangle.
\end{aligned}$$

(To get the third equality, it is assumed one is working with the Levi–Civita connection, so that $\mathrm{Hess} f$ is indeed the Riemannian Hessian.) Note that Propositions 16 and 17 follow immediately from here. Provided the acceleration of the retraction is bounded, we get an approximate result via Cauchy–Schwarz on the above typeset equation.

Proposition 19. *If* Retr *is a retraction with bounded acceleration at x, that is,*

$$\forall \eta \in \mathrm{T}_x\mathcal{M} \text{ with } \|\eta\| = 1, \quad \left\| \frac{\mathrm{D}^2}{\mathrm{d}t^2} \mathrm{Retr}_x(t\eta)\Big|_{t=0} \right\| \leq a,$$

then

$$\left\| \mathrm{Hess} f(x) - \nabla^2 \hat{f}_x(0_x) \right\| \leq a \cdot \|\mathrm{grad} f(x)\|.$$

Thus, if $\varepsilon_H < \infty$, Retr has acceleration bounded by a at x_k and Algorithm 3 returns at iteration k (so that $H_k \succeq -\varepsilon_H \,\mathrm{Id}$ and $\|\mathrm{grad} f(x_k)\| \leq \varepsilon_g$), then

$$\mathrm{Hess} f(x_k) \succeq -(\varepsilon_H + a\varepsilon_g + \delta_k)\,\mathrm{Id}\,. \tag{34}$$

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.

P.-A. Absil, J. Trumpf, R. Mahony, and B. Andrews. All roads lead to Newton: Feasible second-order methods for equality-constrained optimization. Technical report, Technical Report UCL-INMA-2009.024, Departement d'ingenierie mathematique, UCLouvain, Belgium, 2009.

P.-A. Absil, R. Mahony, and J. Trumpf. An extrinsic look at the Riemannian Hessian. In F. Nielsen and F. Barbaresco, editors, *Geometric Science of Information*, volume 8085 of *Lecture Notes in Computer Science*, pages 361–368. Springer Berlin Heidelberg, 2013. ISBN 978-3-642-40019-3. doi:10.1007/978-3-642-40020-9_39. URL http://sites.uclouvain.be/absil/2013.01.

A. Bandeira, N. Boumal, and V. Voroninski. On the low-rank approach for semidefinite programs arising in synchronization and community detection. In *Proceedings of The 29th Conference on Learning Theory, COLT 2016, New York, NY, June 23–26*, 2016.

E. Birgin, J. Gardenghi, J. Martınez, S. Santos, and P. Toint. Worst-case evaluation complexity for unconstrained nonlinear optimization using high-order regularized models. Technical report, Report naXys-05-2015, University of Namur, Belgium, 2015.

N. Boumal. Riemannian trust regions with finite-difference Hessian approximations are globally convergent. In F. Nielsen and F. Barbaresco, editors, *Geometric Science of Information*, volume 9389 of *Lecture Notes in Computer Science*, pages 467–475. Springer International Publishing, 2015a. doi:10.1007/978-3-319-25040-3_50.

N. Boumal. A Riemannian low-rank method for optimization over semidefinite matrices with block-diagonal constraints. *arXiv preprint arXiv:1506.00575*, 2015b.

N. Boumal. Nonconvex phase synchronization. *arXiv preprint arXiv:1601.06114*, 2016.

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, V. Voroninski, and A. Bandeira. The non-convex Burer-Monteiro approach works on smooth semidefinite programs. In preparation, 2016.

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

C. Cartis, N. I. M. Gould, and P. L. Toint. On the complexity of steepest descent, Newton's and regularized Newton's methods for nonconvex unconstrained optimization problems. *SIAM Journal on Optimization*, 20(6):2833–2852, 2010. doi:10.1137/090774100.

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. Optimal Newton-type methods for nonconvex smooth optimization problems. Technical report, ERGO technical report 11-009, School of Mathematics, University of Edinburgh, 2011b.

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. On the complexity of finding first-order critical points in constrained nonlinear optimization. *Mathematical Programming*, 144(1–2):93–106, 2014. doi:10.1007/s10107-012-0617-9.

C. Cartis, N. Gould, and P. Toint. Evaluation complexity bounds for smooth constrained nonlinear optimization using scaled KKT conditions and high-order models. Technical report, NA Technical Report, Maths E-print Archive1912, Mathematical Institute, Oxford University., 2015a.

C. Cartis, N. Gould, and P. Toint. On the evaluation complexity of constrained nonlinear least-squares and general constrained nonlinear optimization using second-order methods. *SIAM Journal on Numerical Analysis*, 53(2):836–851, 2015b. doi:10.1137/130915546.

I. Chavel. *Riemannian geometry: a modern introduction*, volume 108 of *Cambridge Tracts in Mathematics*. Cambridge University Press, 1993.

A. Conn, N. Gould, and P. Toint. *Trust-region methods*. MPS-SIAM Series on Optimization. Society for Industrial and Applied Mathematics, 2000. ISBN 978-0-89871-460-9. doi:10.1137/1.9780898719857.

F. E. Curtis, D. P. Robinson, and M. Samadi. A trust region algorithm with a worst-case iteration complexity of $O(\epsilon^{-3/2})$ for nonconvex optimization. *Mathematical Programming*, pages 1–32, 2016. doi:10.1007/s10107-016-1026-2.

M. Goemans and D. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM (JACM)*, 42 (6):1115–1145, 1995. doi:10.1145/227683.227684.

G. Golub and C. Van Loan. *Matrix computations*, volume 3 of *Johns Hopkins Studies in the Mathematical Sciences*. Johns Hopkins University Press, 4th edition, 2012. doi:10.1137/0720042.

W. Huang, K. Gallivan, and P.-A. Absil. A Broyden class of quasi-Newton methods for Riemannian optimization. *SIAM Journal on Optimization*, 25(3):1660–1685, 2015. doi:10.1137/140955483.

M. McCoy and J. Tropp. Two proposals for robust PCA using semidefinite programming. *Electronic Journal of Statistics*, 5:1123–1160, 2011. doi:10.1214/11-EJS636.

M. G. Monera, A. Montesinos-Amilibia, and E. Sanabria-Codesal. The Taylor expansion of the exponential map and geometric applications. *Revista de la Real Academia de Ciencias Exactas, Fisicas y Naturales. Serie A. Matematicas*, 108(2):881–906, 2014. doi:10.1007/s13398-013-0149-z. URL `http://dx.doi.org/10.1007/s13398-013-0149-z`.

J. Moré and D. Sorensen. Computing a trust region step. *SIAM Journal on Scientific and Statistical Computing*, 4(3):553–572, 1983. doi:10.1137/0904038.

Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87 of *Applied optimization*. Springer, 2004. ISBN 978-1-4020-7553-7.

J. Nocedal and S. Wright. *Numerical optimization*. Springer Verlag, 1999.

C. Qi. *Numerical optimization methods on Riemannian manifolds*. PhD thesis, Florida State University, Tallahassee, FL, 2011.

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.

A. Ruszczyński. *Nonlinear optimization*. Princeton University Press, Princeton, NJ, 2006.

H. Sato. A globally convergent Riemannian conjugate gradient method with the weak Wolfe conditions. *arXiv preprint arXiv:1405.4371*, 2014.

D. Sorensen. Newton's method with a model trust region modification. *SIAM Journal on Numerical Analysis*, 19(2):409–426, 1982. doi:10.1137/0719026.

T. Steihaug. The conjugate gradient method and trust regions in large scale optimization. *SIAM Journal on Numerical Analysis*, 20(3):626–637, 1983.

J. Sun, Q. Qu, and J. Wright. Complete dictionary recovery over the sphere. *arXiv preprint arXiv:1504.06785*, 2015.

J. Sun, Q. Qu, and J. Wright. A geometric analysis of phase retrieval. *arXiv preprint arXiv:1602.06664*, 2016.

C. Udriste. *Convex functions and optimization methods on Riemannian manifolds*, volume 297 of *Mathematics and its applications*. Kluwer Academic Publishers, 1994. doi:10.1007/978-94-015-8390-9.

B. Vandereycken. Low-rank matrix completion by Riemannian optimization. *SIAM Journal on Optimization*, 23(2):1214–1236, 2013. doi:10.1137/110845768.

S. Vavasis. *Nonlinear optimization: complexity issues*. Oxford University Press, Inc., 1991.

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. *arXiv preprint arXiv:1602.06053*, 2016.