Riemannian trust regions with finite-difference Hessian approximations are globally convergent

Nicolas Boumal

Inria & D.I., UMR 8548, Ecole Normale Supérieure, Paris, France,
nicolasboumal@gmail.com

Abstract. The Riemannian trust-region algorithm (RTR) is designed to optimize differentiable cost functions on Riemannian manifolds. It proceeds by iteratively optimizing local models of the cost function. When these models are exact up to second order, RTR boasts a quadratic convergence rate to critical points. In practice, building such models requires computing the Riemannian Hessian, which may be challenging. A simple idea to alleviate this difficulty is to approximate the Hessian using finite differences of the gradient. Unfortunately, this is a nonlinear approximation, which breaks the known convergence results for RTR.

We propose RTR-FD: a modification of RTR which retains global convergence when the Hessian is approximated using finite differences. Importantly, RTR-FD reduces gracefully to RTR if a linear approximation is used. This algorithm is available in the Manopt toolbox.

In the proceedings of Geometric Science of Information, GSI 2015.

1 Introduction

The Riemannian trust-region method (RTR) is a popular algorithm designed to minimize differentiable cost functions f over Riemannian manifolds $\mathcal{M}$ [1, 2]. That is, RTR attempts to compute $\min_{x \in \mathcal{M}} f(x)$. Starting with a given initial guess $x_0 \in \mathcal{M}$, it iteratively reduces the cost $f(x_k)$ along a sequence $x_0, x_1, \ldots$

Under conditions we explicit later, the sequences of iterates produced by RTR converge to critical points regardless of the initial guess (this is called *global convergence*). A critical point $x \in \mathcal{M}$ is such that $\operatorname{grad} f(x) = 0$, where $\operatorname{grad} f(x)$ is the Riemannian gradient of f at x. Since all global optimizers are critical points, this property is highly desirable.

RTR proceeds as follows. At the current iterate $x_k \in \mathcal{M}$, it produces a candidate next iterate, $x_k^+ \in \mathcal{M}$, by (approximately) minimizing a local *model* m_k of f in a neighborhood of x_k, called a *trust region*—because this is where we trust the model. This procedure always reduces the model cost m_k, but of course, the aim is to reduce the actual cost f. RTR then computes the actual cost improvement and decides to accept or reject the proposed step x_k^+ accordingly. Furthermore, depending on how accurately the actual cost improvement was predicted by the model, the size Δ_k of the trust region is reduced, increased or left unchanged for the next iteration. See Algorithm 1.

To be precise, the *inner problem* at iteration k takes the following form:

$$\min_{\eta \in \mathrm{T}_{x_k}\mathcal{M}, \|\eta\|_{P_k^{-1}} \leq \Delta_k} m_k(\eta) := f(x_k) + \langle \eta, \mathrm{grad} f(x_k) \rangle_{x_k} + \frac{1}{2} \langle \eta, H_k[\eta] \rangle_{x_k}, \quad (1)$$

where $\mathrm{T}_x\mathcal{M}$ is the tangent space to $\mathcal{M}$ at x, $\langle \cdot, \cdot \rangle_x$ is the Riemannian metric on $\mathrm{T}_x\mathcal{M}$, $H_k \colon \mathrm{T}_{x_k}\mathcal{M} \to \mathrm{T}_{x_k}\mathcal{M}$ is an operator (conditions on H_k are the topic of this paper), $P_k \colon \mathrm{T}_{x_k}\mathcal{M} \to \mathrm{T}_{x_k}\mathcal{M}$ is a symmetric *positive definite* preconditioner, $\|\eta\|^2_{P_k^{-1}} := \langle \eta, P_k^{-1}\eta \rangle_{x_k}$ defines a norm and Δ_k is the size of the trust region at iteration k. Ideally, P_k is a cheap, positive approximation of the inverse of the Hessian of f at x_k. For a first read, it is safe to assume $P_k = \mathrm{Id}$ (identity).

An approximate solution η_k to the inner problem is computed, and the candidate next iterate is obtained as $x_k^+ = \mathrm{Retr}_{x_k}\eta_k$, where $\mathrm{Retr}_x \colon \mathrm{T}_x\mathcal{M} \to \mathcal{M}$ is a *retraction* on $\mathcal{M}$ [2, def. 4.1.1]: a relaxation of the differential geometric notion of *exponential*. For all x, it satisfies $\mathrm{Retr}_x(0) = x$, and the derivative of $t \mapsto \mathrm{Retr}_x t\eta$ at $t = 0$ equals η, for all tangent η. Furthermore, $\mathrm{Retr}_x\eta$ is smooth in both x and η. If $\mathcal{M}$ is a Euclidean space such as $\mathbb{R}^n$, the classical choice is $\mathrm{Retr}_x\eta = x + \eta$.

A remarkable feature of RTR is that it guarantees global convergence under very lax conditions on both the H_k's and how well (1) is solved [2, thm. 7.4.4]. Essentially two things are required: (a) that the H_k's be uniformly bounded, symmetric *linear* operators, and (b) that the approximate model minimizers η_k produce at least the following decrease in the *model* cost at each iteration [2, eq. (7.14)]:

$$m_k(0) - m_k(\eta_k) \geq c_1 \|\mathrm{grad} f(x_k)\|_{x_k} \min\left(\Delta_k, c_2 \|\mathrm{grad} f(x_k)\|_{x_k}\right), \quad (2)$$

where $\|\eta\|^2_{x_k} := \langle \eta, \eta \rangle_{x_k}$ and $c_1, c_2 > 0$ are constants.

Then, two things are known: (a) if η_k is produced by the *Steihaug-Toint truncated conjugate-gradients algorithm* (tCG, Algorithm 2), sufficient decrease is attained; and (b) still using tCG, if H_k is a sufficiently good approximation of the Riemannian Hessian of f at x_k, RTR achieves a *superlinear local convergence rate* [2, thm. 7.4.11], i.e., close to an isolated local minimizer, convergence is fast.

As computing the Riemannian Hessian can be cumbersome (at best), there is a need for good, generic approximations of it. Linear approximations based on finite differences of the gradient have been proposed [2, § 8.2.1], but they are impractical, since they require the computation of a full operator H_k expanded in a basis of $\mathrm{T}_{x_k}\mathcal{M}$: this is an issue if the dimension of $\mathcal{M}$ is large or if it is difficult to define natural bases of the tangent spaces, that is, in most cases. Alternatives based on transporting an approximate Hessian from tangent space to tangent space may constitute a good solution, especially in low-dimension, even if they are arguably delicate to implement and typically require extra memory [6, 7].

On the other hand, it is quite natural to propose a *nonlinear* approximation of the Hessian at x_k based on finite differences, as $H_k^{\mathrm{FD}}[0] = 0$ and

$$H_k^{\mathrm{FD}}[\eta] = \frac{\mathrm{Transp}_{x_k \leftarrow y}\mathrm{grad} f(y) - \mathrm{grad} f(x_k)}{c}, \text{ with } \begin{cases} c = \alpha / \|\eta\|_{x_k}, \\ y = \mathrm{Retr}_{x_k} c\eta, \end{cases} \quad (3)$$

where $\alpha > 0$ is a small constant (more on this later) and $\mathrm{Transp}_{x \leftarrow y}$ is a *transporter*, i.e., a linear operator from $\mathrm{T}_y\mathcal{M}$ to $\mathrm{T}_x\mathcal{M}$ whose dependence on x and y is jointly continuous and such that $\mathrm{Transp}_{x \leftarrow x} = \mathrm{Id}$ for all x.[1] Transporters allow comparing vectors in different tangent spaces. In this respect, they are loose relaxations of the concept of *parallel transport* in differential geometry. For $\mathcal{M}$ a Euclidean space, the classical choice is $\mathrm{Transp}_{x \leftarrow y} = \mathrm{Id}$ since $\mathrm{T}_x\mathcal{M} \equiv \mathrm{T}_y\mathcal{M}$.

H_k^{FD} is cheap and simple to compute: it essentially requires a single extra gradient evaluation. Unfortunately, because it is nonlinear, the known global convergence theory for RTR does not apply as is. In this paper, we show how a tiny modification to the tCG algorithm makes it possible to retain global convergence even if H_k is only *radially linear*, by which we mean:

$$\forall \eta \in \mathrm{T}_{x_k}\mathcal{M}, \forall a \geq 0, \quad H_k[a\eta] = aH_k[\eta]. \tag{4}$$

Since H_k^{FD} is radially linear, this is a good first step. We then show that H_k^{FD} satisfies the other important condition, namely, uniform boundedness, under mild extra assumptions. Lastly, we note that the modification of tCG is innocuous if H_k is linear, so that it is safe to use the modified version for all purposes.

We name RTR with the modified tCG algorithm and the finite-difference Hessian approximation $H_k := H_k^{\mathrm{FD}}$ the *RTR-FD algorithm*. The Manopt toolbox [4] implements RTR-FD as a default fall-back in case the user does not specify the Hessian. Experience shows it performs well in practice (see for example [3]).

2 Global convergence with bounded, radially linear H_k's

Let $\mathcal{M}$ be a finite-dimensional Riemannian manifold and $f\colon \mathcal{M} \to \mathbb{R}$ be a scalar field on $\mathcal{M}$. We use the notation $\mathrm{dist}(x, y)$ to denote the *Riemannian distance* between two points x and y on $\mathcal{M}$. The *injectivity radius* of $\mathcal{M}$ is defined as

$$i(\mathcal{M}) := \inf_{x \in \mathcal{M}} \sup\{\varepsilon > 0 : \mathrm{Exp}_x|_{\{\eta \in \mathrm{T}_x\mathcal{M} : \|\eta\|_x < \varepsilon\}} \text{ is a diffeomorphism}\},$$

where $\mathrm{Exp}_x\colon \mathrm{T}_x\mathcal{M} \to \mathcal{M}$ is the (geometric) *exponential map* at x; loosely, the operator that generates geodesics. In other words, for all x, y such that $\mathrm{dist}(x, y) < i(\mathcal{M})$, there exists a unique minimizing geodesic joining x to y. In particular, $i(\mathbb{R}^n) = \infty$. For such close points, there is a unique, privileged transporter $\mathrm{PTransp}_{x \leftarrow y}$, called the *parallel transporter* [2, p. 148]. Assuming $i(\mathcal{M}) > 0$, we say f is *Lipschitz continuously differentiable* [2, def. 7.4.3] if it is differentiable and there exists β_1 such that, for all x, y with $\mathrm{dist}(x, y) < i(\mathcal{M})$,

$$\|\mathrm{PTransp}_{x \leftarrow y}\mathrm{grad}f(y) - \mathrm{grad}f(x)\|_x \leq \beta_1 \mathrm{dist}(x, y). \tag{5}$$

We make the following assumptions. They differ from the standard assumptions in only two ways: (a) the H_k's are allowed to be radially linear rather than linear, which requires a slight modification of the tCG algorithm, but no modification of the proofs; and (b) preconditioners are explicitly allowed.

[1] Transporters [7, § 4.3] are mostly equivalent to *vector transports* [2, def. 8.1.1].

Assumption 1 *$\mathcal{M}$ has a positive injectivity radius, $i(\mathcal{M}) > 0$.*

Assumption 2 *f is Lipschitz continuously differentiable* (5) *and $f \circ \mathrm{Retr}$ is radially Lipschitz continuously differentiable [2, def. 7.4.1].*

Assumption 3 *f is bounded below, that is, $\inf_{x \in \mathcal{M}} f(x) > -\infty$.*

Assumption 4 *The H_k's are radially linear* (4) *and bounded, i.e., there exists $\beta < \infty$ such that $\|H_k\|_{\mathrm{op}} := \max\{\|H_k[\eta]\|_{x_k} : \eta \in \mathrm{T}_{x_k}\mathcal{M}, \|\eta\|_{x_k} = 1\} \leq \beta$, $\forall k$.*

Assumption 5 *There exist $\beta_P, \beta_{P^{-1}}$ such that, for all k, $\|P_k\|_{\mathrm{op}} \leq \beta_P < \infty$ and $1/\|P_k^{-1}\|_{\mathrm{op}} \geq \beta_{P^{-1}} > 0$.*

Assumption 6 *There exist $\mu, \delta_\mu > 0$ such that for all $x \in \mathcal{M}$ and for all $\eta \in \mathrm{T}_x\mathcal{M}$ with $\|\eta\|_x \leq \delta_\mu$, the retraction satisfies: $\mathrm{dist}(x, \mathrm{Retr}_x\eta) \leq \|\eta\|_x/\mu$.*

Theorem 1. *Under assumptions 1–5, the sequence $x_0, x_1, x_2, \ldots$ generated by the modified RTR-tCG algorithm (Algs 1–2) satisfies: $\liminf_{k\to\infty} \|\mathrm{grad} f(x_k)\|_{x_k} = 0$.*

Proof. This is essentially Theorem 7.4.2 in [2], with H_k's allowed to be radially linear rather than linear, and with the possibility to use a preconditioner P_k. Without preconditioner ($P_k = \mathrm{Id}$), the proof in [2] turns out to apply verbatim. In the more general case, it can be verified that the first step η^1 computed by the tCG algorithm at iteration k is the *preconditioned Cauchy step*:

$$\eta^1 = \operatorname*{argmin}_{\eta = -\tau P_k \mathrm{grad} f(x_k)} m_k(\eta), \text{ subject to: } \|\eta\|_{P_k^{-1}} \leq \Delta_k \text{ and } \tau > 0. \tag{6}$$

To verify this, execute the first step of tCG by hand (it is oblivious to the fact that H_k is only radially linear), and compare the results to the solution of (6). The latter is simple to solve since it is a quadratic in τ, to be minimized on an interval. Using the analytic expression for η^1, it can be seen that

$$m_k(0) - m_k(\eta^1) \geq \frac{1}{2}\|\mathrm{grad} f(x_k)\|_{P_k} \min\left(\Delta_k, \frac{\|\mathrm{grad} f(x_k)\|_{P_k}}{\|P_k^{1/2} \circ H_k \circ P_k^{1/2}\|_{\mathrm{op}}}\right),$$

where $\|\eta\|_{P_k}^2 = \langle \eta, P_k\eta \rangle_{x_k}$. By submultiplicativity of the operator norm,

$$\|P_k^{1/2} \circ H_k \circ P_k^{1/2}\|_{\mathrm{op}} \leq \|P_k\|_{\mathrm{op}} \|H_k\|_{\mathrm{op}} \leq \beta\beta_P.$$

Furthermore, $\|\mathrm{grad} f(x_k)\|_{P_k} \geq \beta_{P^{-1}}^{1/2} \|\mathrm{grad} f(x_k)\|_{x_k}$. Thus, the sufficient decrease condition (2) is fulfilled by η^1. If H_k is linear, then tCG guarantees $m_k(\eta^{j+1}) < m_k(\eta^j)$ [2, Prop. 7.3.2], so that if η^1 is a sufficiently good approximate solution to the inner problem (which it is), then certainly the solution tCG returns, η_k, is too. For nonlinear H_k, this is not guaranteed anymore, hence the proposed modified tCG, which ensures that, if η^{j+1} is worse than η^j (as per the model), then η^j is returned. The latter is at least as good as η^1, hence (2) holds. □

The proof that all accumulations points are critical points holds verbatim, even though we allow the H_k's to be merely radially linear [2, Thm. 7.4.4]:

Theorem 2. *Under assumptions 1–6, the sequence $x_0, x_1, x_2, \ldots$ generated by the modified RTR-tCG algorithm (Algs 1–2) satisfies: $\lim_{k\to\infty} \mathrm{grad} f(x_k) = 0$.*

3 H_k^{FD} is bounded and radially linear

We now show that setting $H_k := H_k^{\mathrm{FD}}$ (3) to approximate the Hessian of f using finite differences fulfills Assumption 4, under these mild additional assumptions:

Assumption 7 *There exist $\mu', \delta_{\mu'} > 0$ such that, for all x, y with $\mathrm{dist}(x,y) \leq \delta_{\mu'}$, the transporter satisfies $\|\mathrm{Transp}_{x \leftarrow y}\|_{\mathrm{op}} \leq \mu'$.*

Assumption 8 *There exist $\beta_2, \delta_{\beta_2} > 0$ such that, for all x with $f(x) \leq f(x_0)$ and y with $\mathrm{dist}(x,y) \leq \delta_{\beta_2}$, it holds that $\|\mathrm{grad} f(y)\|_y \leq \beta_2$.*

Assumption 7 is inconsequential, since ideal transporters are close to isometries: it would make little sense to violate it. Assumption 8 should be easily achieved, given that f is already assumed Lipschitz continuously differentiable. These assumptions allow to make the following statement:

Theorem 3. *Under assumptions 1–3 and 5–8, with $0 < \alpha < \min(\delta_\mu, \mu\delta_{\mu'}, \mu\delta_{\beta_2})$, the operators H_k^{FD} satisfy Assumption 4, so that the sequence $x_0, x_1, x_2, \ldots$ generated by RTR-FD satisfies:*

$$\lim_{k \to \infty} \mathrm{grad} f(x_k) = 0.$$

Proof. It is clear that H_k^{FD} is radially linear. Let us show that it is also uniformly bounded. For all $\eta \in \mathrm{T}_{x_k}\mathcal{M}$ with $\|\eta\|_{x_k} = 1$ and $y = \mathrm{Retr}_{x_k}(\alpha\eta)$, Assumption 6 ensures that $\alpha < \delta_\mu$ implies $\mathrm{dist}(x_k, y) \leq \alpha/\mu < \min(\delta_{\mu'}, \delta_{\beta_2})$, so that:

$$\begin{aligned}
\|H_k^{\mathrm{FD}}[\eta]\|_{x_k} &= \frac{1}{\alpha} \left\| \mathrm{Transp}_{x_k \leftarrow y} \mathrm{grad} f(y) - \mathrm{grad} f(x_k) \right\|_{x_k} \\
&\leq \frac{1}{\alpha} \left(\left\| \mathrm{Transp}_{x_k \leftarrow y} \right\|_{\mathrm{op}} \|\mathrm{grad} f(y)\|_y + \|\mathrm{grad} f(x_k)\|_{x_k} \right) \\
&\leq \frac{(1+\mu')\beta_2}{\alpha} =: \beta.
\end{aligned}$$

Note: The dependence on $1/\alpha$ is likely artificial and might be removed. One potential start is to argue that $g(y) = \left\| \mathrm{Transp}_{x \leftarrow y} - \mathrm{PTransp}_{x \leftarrow y} \right\|_{\mathrm{op}}$ cannot grow faster than $c_x \cdot \mathrm{dist}(x,y)$ for some constant c_x (since $g(x) = 0$ and g is continuous), and then to use Lipschitz continuous differentiablity of f. □

Corollary 1. *If $\mathcal{M}$ is a Euclidean space (for example, $\mathbb{R}^n$), equipped with the standard tools $\mathrm{Retr}_x \eta = x + \eta$ and $\mathrm{Transp}_{x \leftarrow y} = \mathrm{Id}$, under assumptions 2, 3 and 5, the sequence $x_0, x_1, x_2, \ldots$ generated by RTR-FD with $\alpha > 0$ satisfies:*

$$\lim_{k \to \infty} \mathrm{grad} f(x_k) = 0.$$

Proof. Assumptions 1 and 6 are clearly fulfilled, with $i(\mathcal{M}) = \delta_\mu = \infty$ and $\mu = 1$. Assumptions 7 and 8 are not necessary, since, by Assumption 2, for $\eta \neq 0$,

$$\|H_k^{\mathrm{FD}}[\eta]\|_{x_k} / \|\eta\|_{x_k} = \frac{1}{\alpha} \|\mathrm{grad} f(x_k + c\eta) - \mathrm{grad} f(x_k)\|_{x_k} \leq \beta_1 =: \beta.$$

□

Corollary 2. *If $\mathcal{M}$ is a compact manifold and f is twice continuously differentiable, under Assumption 5 and with the same constraint on α as in Theorem 3, the sequence $x_0, x_1, x_2, \ldots$ generated by RTR-FD satisfies:* $\lim_{k\to\infty} \operatorname{grad} f(x_k) = 0$.

Proof. $\mathcal{M}$ compact implies assumptions 1, 6 and 7. f twice continuously differentiable with $\mathcal{M}$ compact implies assumptions 2, 3 and 8. See [2, Cor. 7.4.6]. □

4 A technical point for computational efficiency

Proposition 7.3.2 in [2] ensures that, provided the operator H is linear, then the model cost strictly decreases at each iteration of tCG: $m(\eta^{j+1}) < m(\eta^j)$. This notably means that there is no need to track $m(\eta^j)$. Allowing for nonlinear H's, this property is lost. The proposed fix (the modified tCG, Algorithm 2) tracks the model cost and safely terminates if a violation (a non-decrease) is witnessed.

A direct implementation of the modified tCG algorithm evaluates the model cost $f(x) + \langle \eta^j, \operatorname{grad} f(x) \rangle_x + 1/2 \langle \eta^j, H[\eta^j] \rangle_x$ at each iteration. This is not advisable, because it requires computing $H[\eta^j]$ whereas only $H[\delta_j]$ is readily available.

If H were linear, then it would hold that $H[\eta^{j+1}] = H[\eta^j + c_j\delta_j] = H[\eta^j] + c_j H[\delta_j]$, with c_j either equal to τ_j or to α_j, as prescribed by the algorithm. This suggests a recurrence to evaluate the model cost without requiring additional applications of H, which is what we use in practice. The sequence $\zeta_0, \zeta_1, \ldots$ defined by $\zeta_0 = 0$ and $\zeta_{j+1} = \zeta_j + c_j H[\delta_j]$ coincides with $H[\eta^0], H[\eta^1], \ldots$ when H is linear. The model cost at η^j is evaluated as $f(x) + \langle \eta^j, \operatorname{grad} f(x) \rangle_x + 1/2 \langle \eta^j, \zeta_j \rangle_x$.

Of course, for nonlinear H, this does not correspond to the original model. But the convergence result still holds if it corresponds to a model using $\tilde{H}$, where the $\tilde{H}$'s are still radially linear and uniformly bounded.

Conceptually run tCG a first time as described above. Then, define $\tilde{H}$ such that it is radially linear, satisfies $\tilde{H}[\eta^j] = \zeta_j$ and $\tilde{H}[\delta_j] = H[\delta_j]$, and coincides with H otherwise. ($\tilde{H}$ is never constructed in practice; it merely serves the argument.) This is well defined as long as no two vectors among $\delta_1, \delta_2, \ldots, \delta_{\text{last}}$ and $\eta^1, \eta^2, \ldots, \eta^{\text{last}}$ are aligned on the same (positive) ray (δ_0 and η^1 are aligned by construction, in a compatible fashion). We do not prove that this property holds, but we note that it seems highly unlikely that it would not, in practical instances. Then, the operators $\tilde{H}$ remain uniformly bounded provided the $\|\zeta_j\|/\|\eta^j\|$'s are uniformly bounded. If so, RTR with the modified tCG behaves exactly as if the models were defined using the $\tilde{H}$'s which satisfy Assumption 4, with true evaluation of the model. This would ensure global convergence.

In the same spirit, the tCG algorithm requires computations of P^{-1}-norms, to ensure iterates remain in the trust region. Since the preconditioner is often only available as a black box P, these P^{-1}-norms are typically computed via recurrences that only involve applying P—see [5, eqs.(7.5.5–7)]. These recurrences make use of the fact that, for linear H, r_{j+1} is orthogonal to $\delta_0, \ldots, \delta_j$. This may not be the case for nonlinear H, so that, in general, using these recurrences may lead to iterates leaving the trust region. One possible fix is to modify the recurrences so that they do not assume the aforementioned orthogonality, but we refrain from doing so in practice, for it does not appear to affect performance.

Algorithm 1 RTR : preconditioned Riemannian trust-region method

Given: $x_0 \in \mathcal{M}$, $0 < \Delta_0 \leq \bar{\Delta}$ and $0 < \rho' < 1/4$
Init: $k = 0$
repeat
  $\eta_k = \mathrm{tCG}(x_k, \Delta_k, H_k, P_k)$ ▷ solve inner problem (1) (approximately)
  $x_k^+ = \mathrm{Retr}_{x_k}(\eta_k)$ ▷ candidate next iterate
  $\rho_1 = f(x_k) - f(x_k^+)$ ▷ actual improvement
  $\rho_2 = m_k(0) - m_k(\eta_k)$ ▷ model improvement
  if $\rho_1/\rho_2 < 1/4$ **then** ▷ if the model made a poor prediction
    $\Delta_{k+1} = \Delta_k/4$ ▷ reduce the trust region radius
  ▷ if the model is good but the region is too small
  else if $\rho_1/\rho_2 > 3/4$ **and** tCG hit the boundary **then**
    $\Delta_{k+1} = \min(2\Delta_k, \bar{\Delta})$ ▷ enlarge the radius
  else
    $\Delta_{k+1} = \Delta_k$
  end if
  if $\rho_1/\rho_2 > \rho'$ **then** ▷ if the relative decrease is sufficient
    $x_{k+1} = x_k^+$ ▷ accept the step
  else ▷ otherwise
    $x_{k+1} = x_k$ ▷ reject it
  end if
  $k = k + 1$
until a stopping criterion triggers

5 Conclusion

From extensive experience, it seems that RTR-FD achieves a superlinear local convergence rate, which is expected since H_k^{FD} is "close" to the true Hessian. See for example [3]. Unfortunately, the existing local convergence analyses rely deeply on the linearity of H_k. We do not expect that a simple modification of the argument would suffice to establish superlinear convergence of RTR-FD. A possible starting point in that direction would be work by Huang et al. on Riemannian trust regions with approximate Hessians [6, 7].

Acknowledgment. The author thanks P.-A. Absil for numerous helpful discussions.

References

1. Absil, P.A., Baker, C.G., Gallivan, K.A.: Trust-region methods on Riemannian manifolds. Foundations of Computational Mathematics 7(3), 303–330 (2007)
2. Absil, P.A., Mahony, R., Sepulchre, R.: Optimization Algorithms on Matrix Manifolds. Princeton University Press, Princeton, NJ (2008)
3. Boumal, N.: Interpolation and regression of rotation matrices. In: Nielsen, F., Barbaresco, F. (eds.) Geometric Science of Information, Lecture Notes in Computer Science, vol. 8085, pp. 345–352. Springer Berlin Heidelberg (2013)
4. Boumal, N., Mishra, B., Absil, P.A., Sepulchre, R.: Manopt, a Matlab toolbox for optimization on manifolds. Journal of Machine Learning Research 15, 1455–1459 (2014), `http://www.manopt.org`

Algorithm 2 tCG(x, Δ, H, P) : modified Steihaug-Toint truncated CG method. It is obtained from the classical tCG by adding the highlighted instructions. See Section 4 for details on how to evaluate $m(\eta)$ and P^{-1}-norms.

Given: $x \in \mathcal{M}$ and $\Delta, \theta, \kappa > 0$, H, P: $\mathrm{T}_x\mathcal{M} \to \mathrm{T}_x\mathcal{M}$,
H radially linear (4), P symmetric positive definite.
Init: $\eta^0 = 0 \in \mathrm{T}_x\mathcal{M}, r_0 = \mathrm{grad} f(x), z_0 = P[r_0], \delta_0 = -z_0$
for $j = 0 \dots \max$ inner iterations $- 1$ **do**
  $\kappa_j = \langle \delta_j, H[\delta_j] \rangle_x$
  $\alpha_j = \langle z_j, r_j \rangle_x / \kappa_j$
  if $\kappa_j \leq 0$ **or** $\|\eta^j + \alpha_j \delta_j\|_{P^{-1}} \geq \Delta$ **then**
    ▷ the model Hessian has negative curvature or TR exceeded:
    Set τ_j to be the positive root of $\|\eta^j + \tau_j \delta_j\|^2_{P^{-1}} = \Delta^2$
    $\eta^{j+1} = \eta^j + \tau_j \delta_j$ ▷ hit the boundary
    if $m(\eta^{j+1}) \geq m(\eta^j)$ **then** ▷ this never triggers if H is linear or $j = 0$
      return η^j ▷ η^j is sure to decrease the model cost
    end if
    return η^{j+1}
  end if
  $\eta^{j+1} = \eta^j + \alpha_j \delta_j$
  if $m(\eta^{j+1}) \geq m(\eta^j)$ **then** ▷ idem
    return η^j
  end if
  $r_{j+1} = r_j + \alpha_j H[\delta_j]$
  if $\|r_{j+1}\|_x \leq \|r_0\|_x \cdot \min(\|r_0\|_x^\theta, \kappa)$ **then**
    return η^{j+1} ▷ this approximate solution is good enough
  end if
  $z_{j+1} = P[r_{j+1}]$
  $\beta_j = \langle z_{j+1}, r_{j+1} \rangle_x \,/\, \langle z_j, r_j \rangle_x$
  $\delta_{j+1} = -z_{j+1} + \beta_j \delta_j$
end for
return η^{last}

5. Conn, A., Gould, N., Toint, P.: Trust-region methods. MPS-SIAM Series on Optimization, Society for Industrial and Applied Mathematics (2000)
6. Huang, W., Absil, P.A., Gallivan, K.: A Riemannian symmetric rank-one trust-region method. Tech. Rep. UCL-INMA-2013.03, Université catholique de Louvain (2013)
7. Huang, W., Gallivan, K., Absil, P.A.: A Broyden class of quasi-Newton methods for Riemannian optimization. Tech. Rep. UCL-INMA-2014.01, Université catholique de Louvain (2015)