Journal of Scientific Computing (2025) 105:69
https://doi.org/10.1007/s10915-025-03099-2

REVIEW

An Adaptive Cubic Regularization Inexact-Newton Method on Riemannian Manifolds

Mauricio S. Louzeiro[1] · Gilson N. Silva[2] · Jinyun Yuan[1] · Daoping Zhang[3]

Received: 10 February 2024 / Revised: 10 March 2025 / Accepted: 15 October 2025

Abstract
An inexact-Newton method with cubic regularization is designed for solving Riemannian unconstrained nonconvex optimization problems. The proposed algorithm is fully adaptive with at most $\mathcal{O}(\epsilon_g^{-3/2})$ iterations to achieve the norm of the gradient smaller than ϵ_g for given $\epsilon_g > 0$, and at most $\mathcal{O}(\max\{\epsilon_g^{-3/2}, \epsilon_H^{-3}\})$ iterations to reach a second-order stationary point respectively. Notably, the proposed algorithm remains applicable even in cases of the gradient and Hessian of the objective function are unknown. Numerical experiments are performed with gradient and Hessian being approximated by forward finite-differences to illustrate the theoretical results and numerical comparison.

Keywords Cubic Regularization · Optimization on Riemannian Manifolds · Derivative-Free · Retraction · Complexity

Mathematics Subject Classification 90C33 · 49M37 · 65K05

The work of this author was partially supported by National Natural Science Foundation of China (No. 12171087). The work of this author was partially supported by CNPq, Brazil (401864/2022-7 and 306593/2022-0). The work of this author was partially supported by National Natural Science Foundation of China (No. 12171087), Dongguan University of Technology, China (221110093), and Fudan Scholar Program, Fudan University, China. The work of this author was supported by National Natural Science Foundation of China (No. 12201320) and the Fundamental Research Funds for the Central Universities, Nankai University (No. 63221039 and 63231144).

✉ Mauricio S. Louzeiro
mauriciolouzeiro@ufg.br

Gilson N. Silva
gilson.silva@ufpi.edu.br

Jinyun Yuan
yuanjy@gmail.com

Daoping Zhang
daopingzhang@nankai.edu.cn

[1] School of Computer Science and Technology, Dongguan University of Technology, Dongguan, Guangdong, China

[2] Departamento de Matemática, Universidade Federal do Piauí, Teresina, Piauí 64049-550, Brazil

[3] School of Mathematical Sciences and LPMC, Nankai University, Tianjin 300071, China

Published online: 03 November 2025

1 Introduction

The main objective of this paper is to develop a Riemannian inexact-Newton method with cubic regularization (R-NMCR, for short) for solving the smooth unconstrained (possibly nonconvex) optimization problems

$$\min_{p\in\mathcal{M}} f(p), \tag{1}$$

where $\mathcal{M}$ represents a given Riemannian manifold, and $f\colon \mathcal{M}\to\mathbb{R}$ is a sufficiently smooth cost function.

Before proceeding, we will briefly review the literature on the Newton method with cubic regularization (E-NMCR) in Euclidean spaces, that is, $\mathcal{M}=\mathbb{R}^n$. It is well-known that Nesterov and Polyak [24] proposed the E-NMCR method to obtain an approximate solution of (1) starting from every point $x_0\in\mathbb{R}^n$ by solving the subproblem

$$u_{k+1}=\operatorname{argmin}_{u\in\mathbb{R}^n} f(x_k)+\langle\nabla f(x_k),u\rangle+\frac{1}{2}\langle\nabla^2 f(x_k)u,u\rangle+\frac{L_f}{6}\|u\|^3. \tag{2}$$

Then, the next point x_{k+1} is defined by $x_{k+1}:=x_k+u_{k+1}$ for all $k\geq 0$. Here, $\nabla^2 f$ is assumed to be L-Lipschitz continuous, and $L_f\geq L>0$ is an estimate for L.

A fact is that saddle points in nonconvex problems may still pose challenges. Due to the absence of higher-order knowledge, first-order methods can only guarantee convergence to stationary points and lack control over the possibility of getting stuck at saddle points. Alternatively, second-order algorithms typically excel at avoiding saddle points by leveraging curvature information. It is known that standard assumptions allow E-NMCR to escape strict saddle points, as seen in [1, 6, 17, 20, 28, 33]. This serves as one of the motivations to continue studying NMCR methods.

It has been shown that E-NMCR produces an iterate x_k with $\|\nabla f(x_k)\|\leq\epsilon$, for some given $\epsilon>0$, in at most $\mathcal{O}(\epsilon^{-3/2})$ iterations. Thanks to this optimal complexity result, Newton's method with cubic regularization was proposed [3, 17]. As we can see, E-NMCR solves a cubic model approximating f in each iteration, wherein the full Hessian matrix must be calculated. This may render E-NMCR less competitive a priori or even infeasible if the Hessian is unavailable. To overcome these drawbacks, an adaptive regularization was established for E-NMCR. In these adaptive schemes, subproblem (2) is solved inexactly to reduce computational costs, as highlighted in [3, 5, 11, 17]. To clarify, in adaptive schemes, the subproblem is addressed as follows:

$$u_{k+1}=\operatorname{argmin}_{u\in\mathbb{R}^n} f(x_k)+\langle\nabla f(x_k),u\rangle+\frac{1}{2}\langle H_k u,u\rangle+\frac{\sigma_k}{6}\|u\|^3, \tag{3}$$

where H_k satisfies some form of inexact condition, and $\sigma_k>0$ can be chosen in various ways. In [9, 10], it is proposed that

$$\left\|\left(H_k-\nabla^2 f(x_k)\right)u_{k+1}\right\|\leq\eta_1\|u_{k+1}\|^2 \tag{4}$$

holds for some matrix H_k and $\eta_1\geq 0$.

It is evident that at iteration k of subproblem (3), knowledge of H_k is necessary. However, obtaining H_k itself requires knowledge of x_{k+1} because H_k must satisfy the inexact condition in (4). Thus, the implementation of methods involving conditions like (4) demands additional computational effort. This is most clearly observed in the complexity result derived in [11], which is $\mathcal{O}(m[\epsilon^{-3/2}+|\log(\epsilon)|])$, where m is the dimension of the domain of the objective function. To enhance this complexity result, [17, 31] have proposed the following inexactness

condition:

$$\left\| H_k - \nabla^2 f(x_k) \right\| \leq \eta_2 \|u_k\|, \tag{5}$$

with $\eta_2 \geq 0$, which no longer involves the subsequent iteration.

For cubic model (3), an E-NMCR algorithm [17] was recently proposed based on the combination of inexact condition (5), approximated Hessian computed by the finite difference method and nonmonotonic line search with the complexity $\mathcal{O}(m\epsilon^{-3/2})$, where m is the dimension of the domain of the objective function. Furthermore, the E-NMCR with finite difference updates on the Hessian approximation requires at most $\mathcal{O}(m \max\{\epsilon_g^{-3/2}, \epsilon_H^{-3}\})$ iterations to find an approximate second-order stationary point, that is, an iterate x_k such that

$$\|\nabla f(x_k)\| \leq \epsilon_g \quad \text{and} \quad \lambda_{\min}(\nabla^2 f(x_k)) \geq -\epsilon_H, \tag{6}$$

where $\lambda_{\min}(\nabla^2 f(x_k))$ denotes the smallest eigenvalue of $\nabla^2 f(x_k)$.

Related works on manifolds: Similarly, in [34], the following cubic subproblem is proposed to analyze the optimization problem (1)

$$\begin{cases} v_k := \operatorname{argmin}_v \hat{f}_k(0) + \langle \nabla \hat{f}_k(0), v\rangle + \dfrac{1}{2}\langle \nabla^2 \hat{f}_k(0)[v], v\rangle + \dfrac{\sigma}{6}\|v\|^3 \\ p_{k+1} := R_{p_k}(v_k), \end{cases}$$

where $v \in \mathrm{T}_{p_k}\mathcal{M}$, $\hat{f}_k = f \circ \mathrm{R}_{p_k} \colon \mathrm{T}_{p_k}\mathcal{M} \to \mathbb{R}$ is the pullback associated with f, $\mathrm{R}_{(\cdot)}(\cdot)$ is a retraction on a Riemannian submanifold $\mathcal{M}$ of a Euclidean space $\mathcal{E}$ and $\sigma > 0$ is an estimate for the Lipschitz Hessian constant. To achieve the same complexity as in Euclidan space, some conditions are assumed in [34], and consequently, some constants need to be known, namely:

$$\begin{cases} \mathcal{M} \text{ must be compact,} \\ \quad \|\mathrm{R}_x(p) - x - p\| \leq L_2\|p\|^2, \quad \forall\, x \in \mathcal{M}, \quad p \in \mathrm{T}_p\mathcal{M}, \\ \left|\langle (\nabla^2_\xi \hat{f}_x(\eta) - \nabla^2_\xi \hat{f}_x(0))[v], v\rangle\right| \leq L_H^{\mathcal{R}}\|\eta\|, \ \forall\, \eta \in \mathrm{T}_x\mathcal{M}, \quad \|\eta\| \leq \mathcal{R}, \\ \quad \forall v \in \mathrm{T}_x\mathcal{M}, \quad \|v\| = 1, \\ G := \max_{x\in\mathcal{M}} \|\nabla f(x)\|_F, \\ \kappa_B := \max_{x\in\mathcal{M}} \max_{\xi\in\mathrm{T}_x\mathcal{M}, \|\xi\|=1} \|\operatorname{Hess} f(x)[\xi]\|, \\ \mathcal{R} = 3\kappa_B + 3\sqrt{G}. \end{cases} \tag{7}$$

Moreover, to execute the algorithm proposed in [34], it is necessary to choose σ such that

$$\sigma > \max\left\{ \left(\sqrt{10L_2\kappa_B + \frac{2}{3}L_H^{\mathcal{R}} + 9L_2^2 G} + 3L_2\sqrt{G} \right)^2, 1 \right\}, \tag{8}$$

where $L_2, \kappa_B, L_H^{\mathcal{R}}, G$ are the constants defined in (7). Thus, it is not difficult to see that the algorithm proposed in [34] can become impractical. As in the Euclidean context, it was proved that the R-NMCR finds an approximate second-order stationary point within $\mathcal{O}(\epsilon^{-3/2})$ iterations [34].

In [3], a more general algorithm was proposed to approximately solve problem (1). Specifically, (i) the two main results in [3] can be applied to every complete Riemannian manifold such that the exponential and retraction maps can be used, (ii) the subproblem in [3] is the same as the one studied in [34], but σ is adaptively chosen and does not depend on any constant, as in (8), (iii) the complexity order $\mathcal{O}(\epsilon^{-3/2})$ is guaranteed when using both the exponential map and a general retraction, (iv) the subproblem in [3] requires computing

the exact gradient and Hessian of the objective function at each iteration, (v) the Riemannian adaptive regularization with cubics (ARC) poposed in [3] computes, at each iteration, a regularized ratio of actual improvement over model improvement, i.e., it computes

$$\rho_k := \frac{f(x_k) - f(R_{x_k}(s_k))}{m_k(0) - m_k(s_k) + \frac{\sigma_k}{3}\|s_k\|^3}, \tag{9}$$

where $f(\cdot)$ is the objective function, $s_k \in \mathrm{T}_{p_k}\mathcal{M}$ must satisfy two first-order progress conditions, $m_k(\cdot)$ is the cubic regularization of $f(\cdot)$, σ_k is the parameter of regularization, and x_k is the sequence generated by ARC. Once ρ_k is computed, x_k is updated as $x_{k+1} := R_{x_k}(s_k)$ only if $\rho_k \geq \eta_1$, where $\eta_1 \in (0, 1)$ is given. Hence, ARC requires a monotone decrease in f to update x_k.

Consider subsample and cubic regularization techiques to approximately solve the problem [12]

$$\min_{p\in\mathcal{M}} f(p) := \frac{1}{n}\sum_{i=1}^{n} f_i(p),$$

where $f_i : \mathcal{M} \to \mathbb{R}$ is a sufficiently smooth cost or loss function for each $i \in \{1, \ldots, n\}$. In this case, the suproblem takes the following structure:

$$\begin{cases} v_k := \operatorname{argmin}_v \hat{f}_k(0) + \langle \mathcal{G}_k, v\rangle + \frac{1}{2}\langle \mathcal{H}_k[v], v\rangle + \frac{\sigma_k}{3}\|v\|^3 \\ p_{k+1} := R_{p_k}(v_k), \end{cases} \tag{10}$$

with $v \in \mathrm{T}_{p_k}\mathcal{M}$ and $\mathcal{G}_k$ and $\mathcal{H}_k[v]$ being, respectively, the approximated Riemannian gradient and Hessian calculated using the subsampling technique, i.e.,

$$\mathcal{G}_k = \frac{1}{|S_g|}\sum_{i\in S_g} \operatorname{grad} f_i(p_k) \quad \text{and} \quad \mathcal{H}_k[v] := \frac{1}{|S_H|}\sum_{i\in S_H} \operatorname{Hess} f_i(p_k)[v], \tag{11}$$

where $S_g, S_H \in \{1, \ldots, n\}$ are the sets of the subsampled indices used for estimating the Riemannian gradient and Hessian, respectively. It is straighforward to see that if $n = 1$, then $i = 1$, $\mathcal{G}_k = \operatorname{grad} f(p_k)$, $\mathcal{H}_k[v] = \operatorname{Hess} f(p_k)[v]$, and hence subproblem (10) is the same as in [3, 34]. To prove the main results in [12], some strong assumptions are made, for instance: (i) the knowledge of the Lipschitz Hessian constant, (ii) the knowledge of a constant that bounds the inexact Hessian defined in (11), which consequently implies that σ_k depends on these constants. Under these conditions, the best second-order complexity achievement obtained in [12] is $\mathcal{O}(\max\{\epsilon_g^{-2}, \epsilon_H^{-3}\})$, where ϵ_g and ϵ_H are as in (6). To improve this complexity results, an additional condition was assumed on the solution of the subproblem, namely, v_k must satisfy the following system:

$$\begin{cases} \langle G_k, v_k\rangle + \langle H_k[v_k], v_k\rangle + \sigma_k\|v_k\|^3 = 0 \\ \langle H_k[v_k], v_k\rangle + \sigma_k\|v_k\|^3 \geq 0. \end{cases}$$

Thus, the new second-order complexity result in [12] is $\mathcal{O}(\max\{\epsilon_g^{-3/2}, \epsilon_H^{-3}\})$.

Our contributions: A new R-NMCR is proposed to approximately solve (1) that is entirely adaptive, which means that Lipschitz gradient or Hessian constants are not necessarily known in advance, neither the regularization parameter of the cubic models and the accuracy of the Hessian approximations jointly adjusted by using a nonmonotone line search criterion. The main results obtained here are applicable to all complete Riemannian manifold where the exponential and retraction maps can be utilized. Our subproblem is also

Table 1 Comparison between R-NMCR and ARC

Method	Complexity	Derivative-Free	Monotone Line Search Test	Adaptive	Inexact
R-NMCR	$\mathcal{O}(\max\{\epsilon_g^{-3/2}, \epsilon_H^{-3}\})$	Yes	No	Yes	Yes
ARC [3]	$\mathcal{O}(\max\{\epsilon_g^{-3/2}, \epsilon_H^{-3}\})$	No	Yes	Yes	No

inexactly solved in the sense of approximated Riemannian gradient and Hessian. Moreover, under standard assumptions, the proposed algorithm requires at most $\mathcal{O}(\max\{\epsilon_g^{-3/2}, \epsilon_H^{-3}\})$ iterations to achieve a second-order stationary point. This means obtaining a point p_k that satisfies a condition similar to (6) but in the Riemannian context. Finally, the new algorithm can be applied when the gradient and Hessian approximations are computed by using forward finite-differences. In this case, the overall complexity of our proposed R-NMCR depends on the dimension of the domain of the objective function Table 1.

We sumarize in table below the main difference between the proposed method in [3] and our method R-NMCR.

The complexity analysis of ARC proposed in [3], in addition to being based on [4] and [10], introduces assumptions that explore the geometry of the manifold and the retraction used in the algorithm. In particular, the first-order analysis in [3] is divided into two cases: (i) R is a general retraction, and (ii) R is the exponential map. Our complexity analysis is based on [3], but with specific adaptations to accommodate the differences between the algorithms, as illustrated in Table 1, and to unify the first-order complexity analysis for cases (i) and (ii).

The subsequent sections of this paper are structured as follows. Section 2 provides a concise review of the preliminaries. In Section 3, the primary derivative-free algorithm proposed are introduced. The worst-case complexity of the proposed algorithm is analyzed in Section 4. Section 5 detailed insights into the computation of the approximated Riemannian gradient and Hessian are given. In Section 6, the results of numerical tests conducted to showcase the effectiveness of the proposed algorithm are displayed. Finally, a summary and concluding remarks are given in the last section.

2 Preliminary

In this section, we review notations, definitions, and basic properties related to Riemannian manifolds, which are utilized throughout the paper. These concepts can be found in introductory books on Riemannian geometry and optimization on manifolds, such as [13, 21, 22, 30] and [2, 6].

Suppose that $\mathcal{M}$ is an n-dimensional connected, smooth manifold. The tangent space at $p \in \mathcal{M}$ is a n-dimensional vector space denoted by $\mathrm{T}_p\,\mathcal{M}$ whose origin is 0_p. The disjoint union of all tangent spaces $\mathrm{T}\,\mathcal{M} := \cup_{p\in\mathcal{M}}(\{p\} \times \mathrm{T}_p\,\mathcal{M})$ is called the *tangent bundle* of $\mathcal{M}$. The Riemannian metric at $p \in \mathcal{M}$ is denoted by $\langle\cdot,\cdot\rangle_p$: $\mathrm{T}_p\,\mathcal{M}\times\mathrm{T}_p\,\mathcal{M} \to \mathbb{R}$ and $\|\cdot\|_p$ for the associated norm in $\mathrm{T}_p\,\mathcal{M}$. For simplicity we shall omit all these indices when no ambiguity arises. Assume that $\mathcal{M}$ is equipped with a Riemannian metric, that is a *Riemannian manifold*.

A *vector field* V on $\mathcal{M}$ is a correspondence associated to each point $p \in \mathcal{M}$ a vector $V(p) \in \mathrm{T}_p\,\mathcal{M}$. Let us denote the smooth vector fields on $\mathcal{M}$ by $\mathcal{X}(\mathcal{M})$ and $\bar{\nabla}\colon \mathcal{X}(\mathcal{M}) \times \mathcal{X}(\mathcal{M}) \to \mathcal{X}(\mathcal{M})$ for the Levi-Civita connection associated to $\mathcal{M}$.

Let $\gamma : I \to \mathcal{M}$ be a smooth curve on an open interval $I \subseteq \mathbb{R}$. A vector field V along γ is said to be *parallel* with respect to $\bar{\nabla}$ if $(D/\,\mathrm{d}t)V \equiv 0$, where $(D/\,\mathrm{d}t)V$ denotes the covariant

derivative associated with $\bar{\nabla}$. For each $t \in I$, the tangent vector of γ at $\gamma(t)$, also known as the velocity of γ at t, is denoted by $\gamma'(t)$. The acceleration of γ, written as γ'', is the vector field $(D/\,\mathrm{d}t)\gamma'$. The curve γ is called a *geodesic* with respect to $\bar{\nabla}$ if $\gamma'' \equiv 0$. When the geodesic equation $(D/\,\mathrm{d}t)\gamma' = 0$ is a second-order nonlinear ordinary differential equation, the geodesic $\gamma = \gamma_v(\cdot, p)$ is determined by its position p and velocity v at p. A Riemannian manifold is *complete* if the geodesics are defined for all values of $t \in \mathbb{R}$. Owing to the completeness of the Riemannian manifold $\mathcal{M}$, the *exponential map* $\exp_p \colon \mathrm{T}_p\mathcal{M} \to \mathcal{M}$ is given by $\exp_p v = \gamma_v(1, p)$, for each $p \in \mathcal{M}$. Next, a detailed definition is provided for a map that generalizes the exponential map and plays a crucial role in the approach presented in this paper.

Definition 1 ([2, Definition 4.1.1 and Sect. 5]) A retraction on $\mathcal{M}$ is a smooth map

$$\mathrm{R} \colon \mathrm{T}\mathcal{M} \to \mathcal{M} \colon (p, v) \mapsto \mathrm{R}_p(v)$$

such that each curve $c(t) = \mathrm{R}_p(tv)$ satisfies $c(0) = p$ and $c'(0) = v$. Retractions that additionally satisfy $c''(0) = 0$ are termed second-order retractions.

The set of smooth scalar functions on $\mathcal{M}$ is denoted by $\mathcal{F}(\mathcal{M})$. The *differential* of a function $f \in \mathcal{F}(\mathcal{M})$ at p is the linear map $\mathcal{D}f(p) \colon \mathrm{T}_p\mathcal{M} \to \mathbb{R}$ which assigns to each $v \in \mathrm{T}_p\mathcal{M}$ the value

$$\mathcal{D}f(p)[v] = \gamma'(t_0)[f] = \frac{\mathrm{d}}{\mathrm{d}t}(f \circ \gamma)\Big|_{t=t_0},$$

for every smooth curve $\gamma \colon I \to \mathcal{M}$ satisfying $\gamma(t_0) = p$ and $\gamma'(t_0) = v$. The *Riemannian gradient* at p of f, $\operatorname{grad} f(p)$, is defined by the unique tangent vector at p such that $\langle \operatorname{grad} f(p), v\rangle_p = \mathcal{D}f(p)[v]$ for all $v \in \mathrm{T}_p\mathcal{M}$. The *Riemannian Hessian* of $f \in \mathcal{F}(\mathcal{M})$ at $p \in \mathcal{M}$ is a linear operator $\operatorname{Hess} f(p) \colon \mathrm{T}_p\mathcal{M} \to \mathrm{T}_p\mathcal{M}$ defined as $\operatorname{Hess} f(p)[u] = \bar{\nabla}_u \operatorname{grad} f$. For real functions on vector spaces (such as $\mathrm{T}_p\mathcal{M}$), we let ∇ and ∇^2 denote the usual gradient and Hessian. The norm of a linear map $A \colon \mathrm{T}_p\mathcal{M} \to \mathrm{T}_p\mathcal{M}$ is defined by $\|A\|_{\mathrm{op}} := \sup\{\|Av\| \colon v \in \mathrm{T}_p\mathcal{M},\ \|v\| = 1\}$.

For each $t_0, t \in I$, $t_0 < t$, the connection $\bar{\nabla}$ induces an isometry $\mathrm{P}_{\gamma,t_0,t} \colon \mathrm{T}_{\gamma(t_0)}\mathcal{M} \to \mathrm{T}_{\gamma(t)}\mathcal{M}$ relative to Riemannian metric on $\mathcal{M}$ given by $\mathrm{P}_{\gamma,t_0,t}\ v = V(\gamma(t))$, where V is the unique vector field on γ such that $\bar{\nabla}_{\gamma'(t)} V(\gamma(t)) = 0$ and $V(\gamma(t_0)) = v$.

The isometry $\mathrm{P}_{\gamma,t_0,t}$ is called *parallel transport* along of γ joining $\gamma(t_0)$ to $\gamma(t)$. For simplicity $\mathrm{P}_v \colon \mathrm{T}_p\mathcal{M} \to \mathrm{T}_{\exp_p v}\mathcal{M}$ denotes parallel transport along the geodesic $\gamma(t) = \exp_p tv$ from $t_0 = 0$ to $t = 1$.

Definition 2 ([3, Definition 2]) A function $f \colon \mathcal{M} \to \mathbb{R}$ has an L-Lipschitz continuous Hessian if it is twice differentiable and if

$$\left\|\mathrm{P}_v^{-1} \circ \operatorname{Hess} f(\exp_p v) \circ \mathrm{P}_v - \operatorname{Hess} f(p)\right\|_{\mathrm{op}} \le L\|v\|,\ \ \forall\ (p, v) \in \mathrm{T}\mathcal{M}\,.$$

The following Lemma provides classic inequalities that will be explored extensively throughout this paper.

Lemma 1 *[[3, Proposition 2]] Let $f \colon \mathcal{M} \to \mathbb{R}$ be twice differentiable on a complete Riemannian manifold $\mathcal{M}$. If $f \colon \mathcal{M} \to \mathbb{R}$ has an L-Lipschitz continuous Hessian then*

$$\left| f(\exp_p v) - f(p) - \langle \operatorname{grad} f(p), v\rangle - \frac{1}{2}\langle \operatorname{Hess} f(p)[v], v\rangle \right| \le \frac{L}{6}\|v\|^3,$$

and

$$\left\|\mathrm{P}_v^{-1} \operatorname{grad} f(\exp_p v) - \operatorname{grad} f(p) - \operatorname{Hess} f(p)[v]\right\| \le \frac{L}{2}\|v\|^2.$$

Lemma 2 *([17, Lemma 4]) Given two real constants $a, b > 0$ and a set $\{z_k : k = 1, \ldots, \bar{N}\}$ of nonnegative real numbers, with $\bar{N} \geq 2$ natural, let $\bar{k} := \operatorname{argmin}_{k \in \{1,\ldots,\bar{N}-1\}}((z_k)^a + (z_{k+1})^a)$. If $\sum_{k=1}^{\bar{N}} (z_k)^a \leq b$ then the inequality $\max\{z_{\bar{k}}, z_{\bar{k}+1}\} \leq \left(2b/(\bar{N}-1)\right)^{1/a}$ holds.*

In this paper, all manifolds $\mathcal{M}$ are assumed to be Riemannian, connected, finite-dimensional, and complete.

3 The Riemannian Inexact-Newton Method

In this section, our aim is to introduce a comprehensive Riemannian inexact-Newton method with cubic regularization and demonstrate its convergence. The algorithm presented below is applicable when both the gradient and Hessian are known. Alternatively, it can be employed with any approximation of the gradient and Hessian that satisfies the assumptions outlined in Step 1.1.

Algorithm 1: General R-NMCR

Step 0. Choose a retraction R, a point $(p_0, v_0) \in \mathrm{T}\mathcal{M}$ $(v_0 \neq 0)$ and constants $\sigma_1 > 0$ and $\theta \geq 0$. Set $k = 1$.
Step 1. Find the smallest integer $\alpha \geq 0$ such that $2^{\alpha-1}\sigma_k \geq \sigma_1$.
Step 1.1. Choose a vector $g_{k,\alpha} \in \mathrm{T}_{p_k}\mathcal{M}$ and an operator $\mathrm{B}_{k,\alpha} : \mathrm{T}_{p_k}\mathcal{M} \to \mathrm{T}_{p_k}\mathcal{M}$ that satisfy

$$\|\operatorname{grad} f(p_k) - g_{k,\alpha}\| \leq \frac{\kappa_g}{2^{\alpha-1}}\|v_{k-1}\|^2, \qquad \|\operatorname{Hess} f(p_k) - \mathrm{B}_{k,\alpha}\|_{\mathrm{op}} \leq \frac{\kappa_{\mathrm{B}}}{2^{\alpha-1}}\|v_{k-1}\|, \tag{12}$$

for fixed constants $\kappa_g, \kappa_{\mathrm{B}} \geq 0$ that are independent of k and α.
Step 1.2. Consider the cubic model $m_{k,\alpha}$ on $\mathrm{T}_{p_k}\mathcal{M}$ given by

$$m_{k,\alpha}(v) = f(p_k) + \langle g_{k,\alpha}, v\rangle + \frac{1}{2}\langle \mathrm{B}_{k,\alpha}[v], v\rangle + \frac{2^{\alpha}\sigma_k}{6}\|v\|^3, \tag{13}$$

and compute an approximate minimizer $v_{k,\alpha}$ of $m_{k,\alpha}$ over $\mathrm{T}_{p_k}\mathcal{M}$ that satisfies

$$m_{k,\alpha}(v_{k,\alpha}) \leq f(p_k) \quad \text{and} \quad \|\nabla m_{k,\alpha}(v_{k,\alpha})\| \leq \theta\|v_{k,\alpha}\|^2. \tag{14}$$

Optionally, if second-order criticality is targeted, $v_{k,\alpha}$ must also satisfy

$$\lambda_{\min}(\mathrm{B}_{k,\alpha}) \geq -2^{\alpha-1}\sigma_k\|v_{k,\alpha}\| - \theta\|v_{k-1}\|. \tag{15}$$

Step 1.3. If

$$f\left(\mathrm{R}_{p_k}(v_{k,\alpha})\right) \leq f(p_k) + \frac{\sigma_k}{24}\|v_{k-1}\|^3 - \frac{2^{\alpha}\sigma_k}{24}\|v_{k,\alpha}\|^3 \tag{16}$$

hold, set $\alpha_k = \alpha$, $v_k = v_{k,\alpha_k}$ and go to Step 2. Otherwise, set $\alpha := \alpha + 1$ and return to Step 1.1.
Step 2. Set $p_{k+1} = \mathrm{R}_{p_k}(v_k)$, $\sigma_{k+1} = 2^{\alpha_k - 1}\sigma_k$, $k := k + 1$, and return to Step 1.

Remark 1 Note that Algorithm 1 works well when $g_{k,\alpha} = \operatorname{grad} f(p_k)$ and $\mathrm{B}_{k,\alpha} = \operatorname{Hess} f(p_k)$ are known for all $\alpha \geq 0$ because the inequalities in (12) are satisfied naturally. Furthermore, closed-form expressions of the approximations $g_{k,\alpha}$ and $\mathrm{B}_{k,\alpha}$ not only satisfy (12), but also eliminate evaluations of the gradient and Hessian. Hence, a Derivative-Free R-NMCR algorithm can be developed from the general R-NMCR algorithm with these closed-form expressions for $g_{k,\alpha}$ and $\mathrm{B}_{k,\alpha}$.

Remark 2 For implementation of Algorithm 1, instead of knowing constants κ_g and κ_{B}, it is possible to choose $g_{k,\alpha}$ and $\mathrm{B}_{k,\alpha}$ satisfying (12) for κ_g and κ_{B} unknown. This flexibility is particularly crucial when κ_g and κ_{B} depend on the Lipschitz constant of Hess f. Some examples of approximations of $g_{k,\alpha}$ and $\mathrm{B}_{k,\alpha}$ satisfying (12), with κ_g and κ_{B} dependent on the Lipschitz constant of Hess f, are given in Section 5.

In the next Remark, we highlight the main differences between Algorithm 1 and the ARC algorithm proposed in [3].

Remark 3 Since, for each iteration $k \geq 1$, Step 1.2 of Algorithm 1 defines a cubic model $m_{k,\alpha}(v_{k,\alpha})$, which must be minimized over $\mathrm{T}_{p_k}\mathcal{M}$ until the condition in Step 1.3 is satisfied, the subproblem in Step 1.2 may need to be solved multiple times within a single iteration. Taking this into account, we now discuss the two main differences between Algorithm 1 and the ARC algorithm proposed in [3]. First, Step 1.1 of Algorithm 1 states that the subproblem in Step 1.2 can be solved using any approximation for grad $f(\cdot)$ and Hess $f(\cdot)$ that satisfies (12), implying that Algorithm 1 can operate in a derivative-free manner. In contrast, the subproblem in ARC [3], which is identical to the one in Step 1.2 of Algorithm 1, requires the computation of grad $f(\cdot)$ and Hess $f(\cdot)$ at least once in each of its iterations. Thus, if grad $f(\cdot)$ and Hess $f(\cdot)$ are computationally expensive or unavailable, Algorithm 1 is a more attractive alternative to the ARC method proposed in [3]. Second, since the ratio ρ_k defined in (9) is nonnegative for each iteration $k \geq 1$, and ARC requires the condition $\rho_k \geq \eta_1$ for some $\eta_1 \in (0, 1)$ to update its current iterate x_k, it follows that the sequence $\{f(x_k)\}$ generated by ARC is monotonically decreasing. In contrast, our proposed algorithm is more flexible. Specifically, as indicated in Step 1.3 and Step 2, Algorithm 1 updates its iterate p_k before ensuring the monotonic decrease of $\{f(p_k)\}$.

For theoretical analysis, some basic assumptions for the cost function f of problem (1) are given as follows. The first one is very common and says that the cost function f is lower bounded.

Assumption 1 There exists $f_{\mathrm{low}} \in \mathbb{R}$ such that $f(p) \geq f_{\mathrm{low}}$ for all $p \in \mathcal{M}$.

Before presenting the second assumption, we shall introduce a notation that will be used throughout this paper. For a given cost function f and a specified retraction R, at each iterate p_k of Algorithm 1, we will consider the following notation:

$$\hat{f}_k \coloneqq f \circ \mathrm{R}_{p_k} \colon \mathrm{T}_{p_k}\mathcal{M} \to \mathbb{R}. \tag{12}$$

The function $\hat{f}_k$ is often called the pullback of the cost function f to the tangent space $\mathrm{T}_{p_k}\mathcal{M}$.

Assumption 2 The function f is twice continuously differentiable and there exists a constant L such that, at each iteration k, the inequality

$$\left| \hat{f}_k(v) - f(p_k) - \langle \mathrm{grad}\, f(p_k), v\rangle - \frac{1}{2}\langle \mathrm{Hess}\, f(p_k)[v], v\rangle \right| \leq \frac{L}{6}\|v\|^3 \quad \text{holds}, \forall\, v \in \mathrm{T}_{p_k}\mathcal{M}. \tag{13}$$

Remark 4 It follows from the definition of $\hat{f}_k$ in (12) that

$$f(p_k) = \hat{f}_k(0) \quad \text{and} \quad \mathrm{grad}\, f(p_k) = \nabla \hat{f}_k(0). \tag{14}$$

Moreover, if the retraction R employed in the definition of $\hat{f}_k$ is a second-order retraction, there is

$$\mathrm{Hess}\, f(p_k) = \nabla^2 \hat{f}_k(0), \tag{15}$$

as demonstrated in [2, Proposition 5.45]. Therefore, whenever R is a second-order retraction (e.g., R = exp), (13) can be reformulated as

$$\left| \hat{f}_k(v) - \hat{f}_k(0) - \langle \nabla \hat{f}_k(0), v \rangle - \frac{1}{2} \langle \nabla^2 \hat{f}_k(0)[v], v \rangle \right| \leq \frac{L}{6} \|v\|^3, \ \forall\, v \in \mathrm{T}_{p_k} \mathcal{M}. \tag{16}$$

On the other hand, as $\hat{f}_k$ is defined on the vector space $\mathrm{T}_{p_k} \mathcal{M}$, (16) holds for L-Lipschitz $\nabla^2 \hat{f}_k$. Overall, it can be asserted that a sufficient condition for Assumption 2 to be satisfied is that R and $\nabla^2 \hat{f}_k$ are a second-order retraction and L-Lipschitz respectively.

Under a reasonable assumption (Assumption 2, to be more specific), the next result guarantees that Algorithm 1 is well-defined, that is, the existence of $\alpha \in [0, +\infty)$ satisfying condition (16).

Theorem 1 *Suppose that Assumption 2 holds. For every iteration k, if $\alpha \geq 0$ satisfies*

$$2^{\alpha-1} \sigma_k \geq 12(2\kappa_g + \kappa_{\mathrm{B}}) + L \tag{17}$$

then

$$\hat{f}_k(v_{k,\alpha}) \leq f(p_k) + \frac{\sigma_k}{24} \|v_{k-1}\|^3 - \frac{2^\alpha \sigma_k}{24} \|v_{k,\alpha}\|^3. \tag{18}$$

Proof Take an arbitrary iteration k and a constant $\alpha \geq 0$ satisfying (17). It follows from Assumption 2 with $v = v_{k,\alpha}$, the definition of $m_{k,\alpha}$ (given in (13)), the first inequality of (14), and (12) that

$$\begin{aligned}
&\hat{f}_k(v_{k,\alpha}) \\
&\leq f(p_k) + \langle \operatorname{grad} f(p_k), v_{k,\alpha} \rangle + \frac{1}{2} \langle \operatorname{Hess} f(p_k)[v_{k,\alpha}], v_{k,\alpha} \rangle + \frac{L}{6} \|v_{k,\alpha}\|^3 \\
&= m_{k,\alpha}(v_{k,\alpha}) + \langle \operatorname{grad} f(p_k) - g_{k,\alpha}, v_{k,\alpha} \rangle + \frac{1}{2} \left\langle (\operatorname{Hess} f(p_k) - \mathrm{B}_{k,\alpha})[v_{k,\alpha}], v_{k,\alpha} \right\rangle \\
&\quad + \frac{L - 2^\alpha \sigma_k}{6} \|v_{k,\alpha}\|^3 \\
&\leq f(p_k) + \| \operatorname{grad} f(p_k) - g_{k,\alpha} \| \|v_{k,\alpha}\| + \frac{1}{2} \left\| \operatorname{Hess} f(p_k) - \mathrm{B}_{k,\alpha} \right\|_{\mathrm{op}} \|v_{k,\alpha}\|^2 \\
&\quad + \frac{L - 2^\alpha \sigma_k}{6} \|v_{k,\alpha}\|^3 \\
&\leq f(p_k) + \frac{\kappa_g}{2^{\alpha-1}} \|v_{k-1}\|^2 \|v_{k,\alpha}\| + \frac{\kappa_{\mathrm{B}}}{2^\alpha} \|v_{k-1}\| \|v_{k,\alpha}\|^2 + \frac{L - 2^\alpha \sigma_k}{6} \|v_{k,\alpha}\|^3.
\end{aligned}$$

Since $\|v_{k-1}\|^2 \|v_{k,\alpha}\| \leq \|v_{k-1}\|^3 + \|v_{k,\alpha}\|^3$ and $\|v_{k-1}\| \|v_{k,\alpha}\|^2 \leq \|v_{k-1}\|^3 + \|v_{k,\alpha}\|^3$, it follows that

$$\hat{f}_k(v_{k,\alpha}) \leq f(p_k) + \frac{2\kappa_g + \kappa_{\mathrm{B}}}{2^\alpha} \|v_{k-1}\|^3 + \frac{2\kappa_g + \kappa_{\mathrm{B}}}{2^\alpha} \|v_{k,\alpha}\|^3 + \frac{L - 2^\alpha \sigma_k}{6} \|v_{k,\alpha}\|^3.$$

By means of (17), we can ensure that $(2\kappa_g + \kappa_{\mathrm{B}})/2^\alpha \leq \sigma_k/24$ holds for every $\alpha \geq 0$. Thus, the previous inequality leads to the following

$$\hat{f}_k(v_{k,\alpha}) \leq f(p_k) + \frac{\sigma_k}{24} \|v_{k-1}\|^3 + \frac{\sigma_k + 4L - 2^{\alpha+2} \sigma_k}{24} \|v_{k,\alpha}\|^3.$$

By using (17) again, one can easily conclude that

$$\sigma_k + 4L - 2^{\alpha+2} \sigma_k = (\sigma_k - 2^\alpha \sigma_k) + (4L - 2^{\alpha+1} \sigma_k) - 2^\alpha \sigma_k \leq -2^\alpha \sigma_k$$

for all $\alpha \geq 0$. Finally, the previous inequality implies that (18) is true. □

The next result shows that the sequence $\{\sigma_k\}$ of regularization parameters is bounded, and further provides lower and upper bounds for this sequence.

Corollary 1 *Under Assumption 2, the sequence of regularization parameters $\{\sigma_k\}$ in Algorithm 1 satisfies*

$$\sigma_1 \leq \sigma_k \leq 24(2\kappa_g + \kappa_B) + 2L + \sigma_1 := \sigma_{\max}, \qquad k = 1, 2, \ldots. \tag{19}$$

Proof Clearly, (19) is true for $k = 1$, and thus our induction base holds. Suppose that (19) holds for some $k \geq 1$. If $\alpha_k = 0$, then by Step 1 and the induction hypothesis, we have

$$\sigma_1 \leq \sigma_{k+1} = 2^{-1}\sigma_k \leq \sigma_k \leq \sigma_{\max},$$

that is, (19) holds for $k + 1$. On the other hand, if $\alpha_k \geq 1$, then there is

$$2^{\alpha_k - 1}\sigma_k \leq \sigma_{\max}. \tag{20}$$

Indeed, by assuming that (20) is not true, it follows that

$$2^{\alpha_k - 2}\sigma_k > 2^{-1}\sigma_{\max} > 12\left(2\kappa_g + \kappa_B\right) + L.$$

In this case, by Theorem 1, inequality (16) would have been satisfied for $\alpha = \alpha_k - 1$, contradicting the minimality of α_k. Thus, (20) is true. Consequently, it follows from Step 1 and (20) that

$$\sigma_1 \leq \sigma_{k+1} = 2^{\alpha_k - 1}\sigma_k \leq \sigma_{\max},$$

that is, (19) also holds for $k + 1$ in this case. This completes the induction argument. □

4 Worst-Case Iteration Complexity Analysis

In this section, we provide first- and second-order analysis of Algorithm 1 in different subsections. The next result will support both analyses.

Lemma 3 *Let $N \geq 3$ be a natural number and define*

$$\bar{k} := \operatorname{argmin}_{k \in \{1, \ldots, N-2\}} \left\{ \|v_k\|^3 + \|v_{k+1}\|^3 \right\}. \tag{21}$$

If Assumptions 1 and 2 hold then

$$\max\left\{\|v_{\bar{k}}\|, \|v_{\bar{k}+1}\|\right\} \leq \left[\frac{48(f(p_1) - f_{low})}{\sigma_1} + 2\|v_0\|^3\right]^{\frac{1}{3}} \frac{1}{(N-2)^{\frac{1}{3}}}. \tag{22}$$

Proof Consider (16) with $\alpha = \alpha_k$. Since $v_{k,\alpha_k} = v_k$, $p_{k+1} = \mathrm{R}_{p_k}(v_k)$, $2\sigma_{k+1} = 2^{\alpha_k}\sigma_k$ and, by Corollary 1, $\sigma_k \geq \sigma_1$ for all $k \geq 1$, we can conclude that

$$\begin{aligned} f(p_k) - f(p_{k+1}) &\geq \frac{2\sigma_{k+1}}{24}\|v_k\|^3 - \frac{\sigma_k}{24}\|v_{k-1}\|^3 \\ &= \frac{\sigma_{k+1}}{24}\|v_k\|^3 + \frac{1}{24}\left(\sigma_{k+1}\|v_k\|^3 - \sigma_k\|v_{k-1}\|^3\right) \\ &\geq \frac{\sigma_1}{24}\|v_k\|^3 + \frac{1}{24}\left(\sigma_{k+1}\|v_k\|^3 - \sigma_k\|v_{k-1}\|^3\right), \end{aligned}$$

for all $k = 1, \ldots, N-1$. Summing over $k = 1, \ldots, N-1$ and using Assumption 1, we get

$$f(p_1) - f_{low} \geq \sum_{k=1}^{N-1} f(p_k) - f(p_{k+1})$$

$$\geq \frac{\sigma_1}{24}\sum_{k=1}^{N-1}\|v_k\|^3 + \frac{1}{24}\sum_{k=1}^{N-1}\left(\sigma_{k+1}\|v_k\|^3 - \sigma_k\|v_{k-1}\|^3\right)$$
$$\geq \frac{\sigma_1}{24}\sum_{k=1}^{N-1}\|v_k\|^3 - \frac{\sigma_1}{24}\|v_0\|^3,$$

which is equivalent to

$$\sum_{k=1}^{N-1}\|v_k\|^3 \leq \frac{24(f(p_1) - f_{low})}{\sigma_1} + \|v_0\|^3. \tag{23}$$

Thus, (22) follows from (23) and Lemma 2 with $z_k = \|v_k\|$, $\bar{N} = N-1$ and $a = 3$. □

4.1 First-order analysis

For our first-order analysis, we need to impose the following additional assumption.

Assumption 3 Let f be a twice continuously differentiable function, and let $\mathcal{G}\colon \mathrm{T}\mathcal{M} \to \mathrm{T}\mathcal{M}$ be a mapping whose image of $\mathcal{G}(p,\cdot)\colon \mathrm{T}_p\mathcal{M} \to \mathrm{T}\mathcal{M}$ belongs to $\mathrm{T}_p\mathcal{M}$ for all $p \in \mathcal{M}$. Suppose that there exists a constant L' such that, at each iteration k, the inequality

$$\|\mathcal{G}(p_k, v) - \operatorname{grad} f(p_k) - \operatorname{Hess} f(p_k)[v]\| \leq \frac{L'}{2}\|v\|^2 \ \text{ holds, } \forall\, v \in \mathrm{T}_{p_k}\mathcal{M}. \tag{24}$$

Remark 5 Due to Lemma 1, we can assert that if Hess f is L'-Lipshitz then Assumption 3 holds for $\mathcal{G}\colon \mathrm{T}\mathcal{M} \to \mathrm{T}\mathcal{M}$ defined by $\mathcal{G}(p,v) = \mathrm{P}_v^{-1}\operatorname{grad} f(\exp_p v)$ for all $(p,v) \in \mathrm{T}\mathcal{M}$. Furthermore, assume that R is an arbitrary second-order retraction (not necessarily equal to the exponential map). Hence, it follows from (14) and (15) that (24) can be rewritten as

$$\left\|\mathcal{G}(p_k, v) - \nabla\hat{f}_k(0) - \nabla^2\hat{f}_k(0)[v]\right\| \leq \frac{L'}{2}\|v\|^2, \ \ \forall\, v \in \mathrm{T}_{p_k}\mathcal{M},$$

which ensures that if $\nabla^2\hat{f}_k$ is L'-Lipschitz then Assumption 3 is satisfied for $\mathcal{G}\colon \mathrm{T}\mathcal{M} \to \mathrm{T}\mathcal{M}$ defined by $\mathcal{G}(p,v) = \nabla(f \circ \mathrm{R}_p)(v)$ for all $(p,v) \in \mathrm{T}\mathcal{M}$.

The following lemma provides an important inequality for the next two theorems.

Lemma 4 *Suppose that f and $\mathcal{G}$ satisfy Assumption 3. Under Assumption 2, for every iteration k one has*

$$\|\mathcal{G}(p_k, v_k)\| \leq \tau \max\left\{\|v_{k-1}\|, \|v_k\|\right\}^2, \qquad \tau := \frac{L' + 2(\theta + \sigma_{\max}) + 4(\kappa_g + \kappa_{\mathrm{B}})}{2}, \tag{25}$$

where $\sigma_{\max}$ is defined in (19).

Proof It follows from (13) with $\alpha = \alpha_k$ and $\sigma_{k+1} = 2^{\alpha_k - 1}\sigma_k$ that

$$\begin{aligned}\nabla m_{k,\alpha_k}(v_k) &= g_{k,\alpha_k} + \mathrm{B}_{k,\alpha_k}[v_k] + 2^{\alpha_k-1}\sigma_k\|v_k\|v_k \\ &= \mathcal{G}(p_k, v_k) - (\mathcal{G}(p_k, v_k) - \operatorname{grad} f(p_k) - \operatorname{Hess} f(p_k)[v_k]) \\ &\quad + \left(g_{k,\alpha_k} - \operatorname{grad} f(p_k)\right) + \left(\mathrm{B}_{k,\alpha_k} - \operatorname{Hess} f(p_k)\right)[v_k] + \sigma_{k+1}\|v_k\|v_k.\end{aligned}$$

Taking norms on both sides and also using the second inequality of (14) with $\alpha = \alpha_k$, we find by triangle inequality that

$$\theta\|v_k\|^2 \geq \left\|\nabla m_{k,\alpha_k}(v_k)\right\| \geq \|\mathcal{G}(p_k, v_k)\| - \|\mathcal{G}(p_k, v_k) - \operatorname{grad} f(p_k) - \operatorname{Hess} f(p_k)[v_k]\|$$

$$- \left\| g_{k,\alpha_k} - \operatorname{grad} f(p_k) \right\| - \left\| (\mathrm{B}_{k,\alpha_k} - \operatorname{Hess} f(p_k))[v_k] \right\| - \sigma_{k+1} \|v_k\|^2.$$

Rearranging, using (24) with $v = v_k$ and (12) with $\alpha = \alpha_k$, we get

$$\begin{aligned} \|\mathcal{G}(p_k, v_k)\| &\leq \frac{L' + 2(\theta + \sigma_{k+1})}{2} \|v_k\|^2 + \frac{\kappa_g}{2^{\alpha_k - 1}} \|v_{k-1}\|^2 + \left\| \mathrm{B}_{k,\alpha_k} - \operatorname{Hess} f(p_k) \right\|_{\mathrm{op}} \|v_k\| \\ &\leq \frac{L' + 2(\theta + \sigma_{k+1})}{2} \|v_k\|^2 + \frac{\kappa_g}{2^{\alpha_k - 1}} \|v_{k-1}\|^2 + \frac{\kappa_{\mathrm{B}}}{2^{\alpha_k - 1}} \|v_{k-1}\| \|v_k\| \\ &\leq \frac{L' + 2(\theta + \sigma_{k+1})}{2} \|v_k\|^2 + 2\kappa_g \|v_{k-1}\|^2 + 2\kappa_{\mathrm{B}} \|v_{k-1}\| \|v_k\| \\ &\leq \frac{L' + 2(\theta + \sigma_{k+1}) + 4(\kappa_g + \kappa_{\mathrm{B}})}{2} \max \{\|v_{k-1}\|, \|v_k\|\}^2 . \end{aligned}$$

Therefore, the proof conclusion follows from the previous inequality together with Corollary 1. □

Now we can establish our first complexity result for Algorithm 1. Here, we are concerned only with the particular case where the retraction chosen is the exponential map, i.e., $\mathrm{R} = \exp$, and Assumption 3 for $\mathcal{G} \colon \mathrm{T}\mathcal{M} \to \mathrm{T}\mathcal{M}$ defined by

$$\mathcal{G}(p, v) = \mathrm{P}_v^{-1} \operatorname{grad} f(\exp_p v), \ \ \forall\, (p, v) \in \mathrm{T}\mathcal{M}. \tag{26}$$

Theorem 2 *Let* $\mathrm{R} = \exp$. *Under Assumption 1, 2, and 3 with* $\mathcal{G}$ *given in* (26), *let* $p_0, p_1, p_2, \ldots$ *be the iterates produced by Algorithm 1. For arbitrary* $\epsilon > 0$, *if* $\| \operatorname{grad} f(p_k)\| > \epsilon$ *for all* $k \in \{1, \ldots, N\}$ *then*

$$N \leq 2 + \left\lfloor \frac{48(f(p_1) - f_{low})}{\sigma_1} + 2\|v_0\|^3 \right\rfloor \left(\frac{\epsilon}{\tau}\right)^{-\frac{3}{2}} \tag{27}$$

where τ *is defined in* (25).

Proof Inequality (27) is trivially satisfied for $N = 1$ and $N = 2$. Then assume $N \geq 3$. Taking into account the hypothesis of this theorem and the fact that $\bar{k}$ defined in (21) belongs to $\{1, \ldots, N-2\}$, we conclude that $\| \operatorname{grad} f(p_{\bar{k}+2})\| > \epsilon$. Hence, since $\mathrm{P}_{v_{\bar{k}+1}}^{-1}$ is an isometry, Lemma 4 with $k = \bar{k} + 1$ and $\mathcal{G}$ as in (26) yields

$$\epsilon < \| \operatorname{grad} f(p_{\bar{k}+2})\| = \| \mathrm{P}_{v_{\bar{k}+1}}^{-1} \operatorname{grad} f(\exp_{p_{\bar{k}+1}} v_{\bar{k}+1})\| = \|\mathcal{G}(p_{\bar{k}+1}, v_{\bar{k}+1})\| \leq \tau \max \left\{\|v_{\bar{k}}\|, \|v_{\bar{k}+1}\|\right\}^2,$$

with τ given in (25). By using this inequality together with Lemma 3 we find

$$\epsilon < \tau \left\lfloor \frac{48(f(p_1) - f_{low})}{\sigma_1} + 2\|v_0\|^3 \right\rfloor^{\frac{2}{3}} \frac{1}{(N-2)^{\frac{2}{3}}}.$$

After rearranging the terms of this inequality, the proof will be complete. □

The proof of the next result follows directly from the previous theorem and the fact that Assumption 2 and 3 are satisfied for $\mathrm{R} = \exp$, $L' = L$ and $\mathcal{G}$ given in (26) for L-Lipschitz Hess f.

Corollary 2 *Under Assumption 1, let $p_0, p_1, p_2, \ldots$ be the iterates produced by Algorithm 1 with* R = exp. *For arbitrary $\epsilon > 0$, with L-Lipschitz* Hess f $\|\operatorname{grad} f(p_k)\| \leq \epsilon$ *for all*

$$k > 2 + \left[\frac{48(f(p_1) - f_{low})}{\sigma_1} + 2\|v_0\|^3\right]\left(\frac{\epsilon}{\tau}\right)^{-\frac{3}{2}},$$

where τ is considered with $L' = L$ in (25). In particular, $\lim_{k\to\infty} \|\operatorname{grad} f(p_k)\| = 0$.

Our second complexity result will be proposed for a general retraction R whose proof requires the use of an additional assumption that relates $\nabla \hat{f}_k$ and grad f for each iteration k. This assumption is detailed below.

Assumption 4 There exist constants $a \in (0, +\infty]$ and $b \in (0, 1]$ such that $\|v\| \leq a$, $v \in \mathrm{T}_{p_k}\mathcal{M}$, implies $\|\nabla \hat{f}_k(v)\| \geq b\|\operatorname{grad} f(\mathrm{R}_{p_k}(v))\|$.

Remark 6 For the first-order analysis of a general retraction the following assumption was used in [3]:

(A) There exist constants $a \in (0, +\infty]$ and $b \in (0, 1]$ such that $\|v\| \leq a$, $v \in \mathrm{T}_{p_k}\mathcal{M}$, implies $\varsigma_{\min}(\mathrm{DR}_{p_k}(v)) \geq b$, where $\varsigma_{\min}$ extracts the smallest singular value of an operator.

Since the calculations in [3, Sect. 4] guarantee the inequality

$$\|\nabla \hat{f}_k(v)\| \geq \varsigma_{\min}(\mathrm{DR}_{p_k}(v))\|\operatorname{grad} f(\mathrm{R}_{p_k}(v))\|,$$

for all $v \in \mathrm{T}_{p_k}\mathcal{M}$, it is easy to show that if **(A)** holds then Assumption 4 also holds. In view of this, we can state that [3, Sect. 7] secures Assumption 4 for a large family of manifolds and retractions.

We now prepare to present our first-order complexity result for a general retraction R. Here, we consider the function $\mathcal{G}\colon \mathrm{T}\mathcal{M} \to \mathrm{T}\mathcal{M}$ defined by

$$\mathcal{G}(p, v) = \nabla(f \circ \mathrm{R}_p)(v), \ \forall\, (p, v) \in \mathrm{T}\mathcal{M}. \tag{28}$$

Theorem 3 *Under Assumptions 1, 2, 3 with $\mathcal{G}$ given in (28), and 4, let $p_0, p_1, p_2, \ldots$ be the iterates produced by Algorithm 1. Choose $a \in (0, +\infty]$ and $b \in (0, 1]$ satisfying Assumption 4. For every $\epsilon > 0$, if $\|\operatorname{grad} f(p_k)\| > \epsilon$ for all $k \in \{1, \ldots, N\}$ then*

$$N \leq 2 + \left[\frac{48(f(p_1) - f_{low})}{\sigma_1} + 2\|v_0\|^3\right]\max\left\{a^{-3}, \left(\frac{b\epsilon}{\tau}\right)^{-\frac{3}{2}}\right\}, \tag{29}$$

where τ is defined in (25).

Proof Inequality (29) is trivially satisfied when $N = 1$ and $N = 2$. Assume $N \geq 3$. Throughout the proof consider the definition of $\bar{k}$ given in (21). Our analysis will be divided into the following two cases:

(i) $\max\{\|v_{\bar{k}}\|, \|v_{\bar{k}+1}\|\} \in [0, a)$;
(ii) $\max\{\|v_{\bar{k}}\|, \|v_{\bar{k}+1}\|\} \geq a$.

If case (i) holds then $\|v_{\bar{k}+1}\| \leq a$ and, by Assumption 4 with $v = v_{\bar{k}+1}$ and $k = \bar{k} + 1$, we have

$$\|\nabla \hat{f}_{\bar{k}+1}(v_{\bar{k}+1})\| \geq b\|\operatorname{grad} f(\mathrm{R}_{p_{\bar{k}+1}}(v_{\bar{k}+1}))\| = b\|\operatorname{grad} f(p_{\bar{k}+2})\|.$$

By hypothesis and the fact that $\bar{k} \in \{1, \ldots, N-2\}$ we conclude that $\|\operatorname{grad} f(p_{\bar{k}+2})\| > \epsilon$ and, therefore, the above inequality leads to $\|\nabla \hat{f}_{\bar{k}+1}(v_{\bar{k}+1})\| > b\epsilon$. Using this, (28) with $(p, v) = (p_{\bar{k}+1}, v_{\bar{k}+1})$ and Lemma 4 with $k = \bar{k}+1$ and $\mathcal{G}$ as in (28), one can conclude that

$$\left(\frac{b\epsilon}{\tau}\right)^{\frac{1}{2}} < \left(\frac{\|\nabla \hat{f}_{\bar{k}+1}(v_{\bar{k}+1})\|}{\tau}\right)^{\frac{1}{2}} = \left(\frac{\|\mathcal{G}(p_{\bar{k}+1}, v_{\bar{k}+1})\|}{\tau}\right)^{\frac{1}{2}} \leq \max\left\{\|v_{\bar{k}}\|, \|v_{\bar{k}+1}\|\right\}.$$

On the other hand, if case (ii) holds then $a \leq \max\left\{\|v_{\bar{k}-1}\|, \|v_{\bar{k}}\|\right\}$. Thus, in all cases, one can conclude that

$$\min\left\{a, \left(\frac{b\epsilon}{\tau}\right)^{\frac{1}{2}}\right\} \leq \max\left\{\|v_{\bar{k}-1}\|, \|v_{\bar{k}}\|\right\} \leq \left[\frac{48(f(p_1) - f_{low})}{\sigma_1} + 2\|v_0\|^3\right]^{\frac{1}{3}} \frac{1}{(N-2)^{\frac{1}{3}}},$$

where the second inequality follows from Lemma 3. By rearranging the terms in a convenient way we can get (29). □

As previously discussed in Remarks 4 and 5, if $\nabla^2 \hat{f}_k$ is L-Lipschitz for each iteration k, and R is a second-order retraction, then Assumptions 2 and 3 are satisfied with $L' = L$ and $\mathcal{G}$ given in (28). Considering this and the previous theorem, we can derive the following corollary.

Corollary 3 *Under Assumptions 1 and 4, let $p_0, p_1, p_2, \ldots$ be the iterates produced by Algorithm 1 with a second-order retraction chosen. In each iteration k assume that $\nabla^2 \hat{f}_k$ is L-Lipschitz. Choose $a \in (0, +\infty]$ and $b \in (0, 1]$ satisfying Assumption 4. Then, for every $\epsilon > 0$, one has $\|\operatorname{grad} f(p_k)\| \leq \epsilon$ for all*

$$k > 2 + \left[\frac{48(f(p_1) - f_{low})}{\sigma_1} + 2\|v_0\|^3\right] \max\left\{a^{-3}, \left(\frac{b\epsilon}{\tau}\right)^{-\frac{3}{2}}\right\},$$

where τ is considered with $L' = L$ in (25). In particular, $\lim_{k\to\infty} \|\operatorname{grad} f(p_k)\| = 0$.

4.2 Second-order analysis

In this subsection we shall give a second-order complexity result for Algorithm 1 with second-order progress condition (15) enforced. This condition is similar to one given in [17] for the same purpose in the Euclidean case. Unlike the first-order analysis in the previous subsection, here we give second-order analysis only for general retraction R, but not for the particular $\mathrm{R} = \exp$.

Theorem 4 *Under Assumptions 1 and 2, let $p_0, p_1 \ldots$ be the iterates produced by Algorithm 1 with second-order progress (15) enforced. For every $\epsilon > 0$, if $\lambda_{\min}(\operatorname{Hess} f(p_k)) < -\epsilon$ for all $k \in \{1, \ldots, N\}$ then*

$$N \leq 2 + \left[\frac{48(f(p_1) - f_{low})}{\sigma_1} + 2\|v_0\|^3\right]\left[\frac{\epsilon}{\sigma_{\max} + \theta + \kappa_{\mathrm{B}}}\right]^{-3}, \tag{30}$$

where $\sigma_{\max}$ is defined in (19).

Proof It is clear that (30) holds for $N = 1$ and $N = 2$. Assume $N \geq 3$. Let $\bar{k}$ be the constant defined in (21). By using the second inequality of (12) with $k = \bar{k}+1$ and $\alpha = \alpha_{\bar{k}+1}$, we

obtain

$$\begin{aligned}\left\langle \mathrm{B}_{\bar{k}+1,\alpha_{\bar{k}+1}}\left[\frac{v}{\|v\|}\right], \frac{v}{\|v\|}\right\rangle &= \left\langle \mathrm{Hess}\, f(p_{\bar{k}+1})\left[\frac{v}{\|v\|}\right], \frac{v}{\|v\|}\right\rangle \\ &\quad + \left\langle \left(\mathrm{B}_{\bar{k}+1,\alpha_{\bar{k}+1}} - \mathrm{Hess}\, f(p_{\bar{k}+1})\right)\left[\frac{v}{\|v\|}\right], \frac{v}{\|v\|}\right\rangle \\ &\leq \left\langle \mathrm{Hess}\, f(p_{\bar{k}+1})\left[\frac{v}{\|v\|}\right], \frac{v}{\|v\|}\right\rangle + \left\| \mathrm{B}_{\bar{k}+1,\alpha_{\bar{k}+1}} - \mathrm{Hess}\, f(p_{\bar{k}+1})\right\|_{\mathrm{op}} \\ &\leq \left\langle \mathrm{Hess}\, f(p_{\bar{k}+1})\left[\frac{v}{\|v\|}\right], \frac{v}{\|v\|}\right\rangle + \frac{\kappa_{\mathrm{B}}}{2^{\alpha-1}}\|v_{\bar{k}}\| \\ &\leq \left\langle \mathrm{Hess}\, f(p_{\bar{k}+1})\left[\frac{v}{\|v\|}\right], \frac{v}{\|v\|}\right\rangle + 2\kappa_{\mathrm{B}}\|v_{\bar{k}}\|,\end{aligned}$$

for all $v \in \mathrm{T}_{p_{\bar{k}+1}}\mathcal{M}\backslash\{0\}$. After calculating the infimum with respect to $v \in \mathrm{T}_{p_{\bar{k}+1}}\mathcal{M}$ on both sides and use the hypothesis of this theorem, we arrive at

$$\begin{aligned}\lambda_{\min}\left(\mathrm{B}_{\bar{k}+1,\alpha_{\bar{k}+1}}\right) &\leq \lambda_{\min}\left(\mathrm{Hess}\, f(p_{\bar{k}+1})\right) + 2\kappa_{\mathrm{B}}\|v_{\bar{k}}\| < -\epsilon + 2\kappa_{\mathrm{B}}\|v_{\bar{k}}\| \leq -\epsilon + 2\kappa_{\mathrm{B}} \\ &\quad \max\{\|v_{\bar{k}}\|, \|v_{\bar{k}+1}\|\}.\end{aligned}$$

It follows from $\sigma_{\bar{k}+2} = 2^{\alpha_{\bar{k}+1}-1}\sigma_{\bar{k}+1}$, $v_{\bar{k}+1} = v_{\bar{k}+1,\alpha_{\bar{k}+1}}$, Corollary 1 with the previous inequalities, (15) with $k = \bar{k}+1$ and $\alpha = \alpha_{\bar{k}+1}$ that

$$\begin{aligned}-(\sigma_{\max}+\theta)\max\{\|v_{\bar{k}}\|, \|v_{\bar{k}+1}\|\} &\leq -(\sigma_{\bar{k}+2}+\theta)\max\{\|v_{\bar{k}}\|, \|v_{\bar{k}+1}\|\} \\ &= -(2^{\alpha_{\bar{k}+1}-1}\sigma_{\bar{k}+1}+\theta)\max\{\|v_{\bar{k}}\|, \|v_{\bar{k}+1,\alpha_{\bar{k}+1}}\|\} \\ &\leq -2^{\alpha_{\bar{k}+1}-1}\sigma_{\bar{k}+1}\|v_{\bar{k}+1,\alpha_{\bar{k}+1}}\| - \theta\|v_{\bar{k}}\| \\ &\leq \lambda_{\min}\left(\mathrm{B}_{\bar{k}+1,\alpha_{\bar{k}+1}}\right) < -\epsilon + 2\kappa_{\mathrm{B}}\max\{\|v_{\bar{k}}\|, \|v_{\bar{k}+1}\|\}.\end{aligned}$$

Hence, it follows that

$$\frac{\epsilon}{\sigma_{\max}+\theta+2\kappa_{\mathrm{B}}} < \max\{\|v_{\bar{k}}\|, \|v_{\bar{k}+1}\|\} \leq \left[\frac{48(f(p_1)-f_{low})}{\sigma_1} + 2\|v_0\|^3\right]^{\frac{1}{3}} \frac{1}{(N-2)^{\frac{1}{3}}},$$

where the second inequality comes from Lemma 3. Starting from this inequality, we arrive at (30) by using simple algebraic manipulations. Therefore, the proof is complete. □

The next result is an immediate consequence of the previous theorem. Looking at this result and at Corollary 2 or 3, depending on the Assumptions required, it is possible to find a number of iterations N such that p_N is an approximate second-order critical point, that is p_N satisfies $\|\mathrm{grad}\, f(p_N)\| \leq \epsilon_g$ and $\lambda_{\min}(\mathrm{Hess}\, f(p_N)) \geq -\epsilon_H$, for $\epsilon_g > 0$ and $\epsilon_H > 0$ chosen arbitrarily.

Corollary 4 *Under Assumptions 1 and 2, let $p_0, p_1 \ldots$ be the iterates produced by Algorithm 1 with second-order progress (15) enforced. Then, for every $\epsilon > 0$, one has $\lambda_{\min}(\mathrm{Hess}\, f(p_k)) \geq -\epsilon$ for all*

$$k > 2 + \left[\frac{48(f(p_1)-f_{low})}{\sigma_1} + 2\|v_0\|^3\right]\left[\frac{\epsilon}{\sigma_{\max}+\theta+\kappa_{\mathrm{B}}}\right]^{-3},$$

where $\sigma_{\max}$ is defined in (19). In particular, $\liminf_{k\to\infty}\lambda_{\min}(\mathrm{Hess}\, f(p_k)) \geq 0$.

5 Finite-Difference Approximations for Gradients and Hessians

In this section we borrow finite-differences to generate approximations $g_{k,\alpha}$ and $\mathrm{B}_{k,\alpha}$ satisfying (12). These approximations and the strategies employed in the proofs of this section are frequently explored in the study of derivative-free optimization; see [11, 14, 25] for the Euclidean case and [2, 6, 23] for the Riemannian case, for instance.

For results established in this section, we assume that, for each iteration k and constant $\alpha \geq 0$, $\mathfrak{B}^k := \{e_1^k, \ldots, e_n^k\}$ forms an orthonormal basis for $\mathrm{T}_{p_k}\mathcal{M}$ and $h_{k,\alpha}$ is a real number defined by

$$h_{k,\alpha} := \frac{\|v_{k-1}\|}{2^{\alpha-1}\sigma_k}. \tag{31}$$

5.1 Approximation for Gradients

The following result provides an approximation $g_{k,\alpha}$ for $\operatorname{grad} f(p_k)$ that depends only on evaluations of the objective function f and satisfies the first inequality in (12) for every iteration k and constant $\alpha \geq 0$. The analysis presented here is developed for a general retraction R. Recall that $\hat{f}_k$ refers to the notation introduced in (12).

Proposition 1 *For every iteration k and constant $\alpha \geq 0$, let $g_{k,\alpha} \in \mathrm{T}_{p_k}\mathcal{M}$ be defined by*

$$g_{k,\alpha} = \sum_{i=1}^{n} \frac{\hat{f}_k(h_{k,\alpha}e_i^k) - \hat{f}_k(-h_{k,\alpha}e_i^k)}{2h_{k,\alpha}} e_i^k. \tag{32}$$

Under Assumption 2, the first inequality of (12) is always satisfied with $\kappa_g = L\sqrt{n}/(3\sigma_1^2)$.

Proof By leveraging the orthonormality of the basis $\mathfrak{B}^k = \{e_1^k, \ldots, e_n^k\}$, we can assert that

$$\begin{aligned} \left\| g_{k,\alpha} - \operatorname{grad} f(p_k) \right\| &\leq \sqrt{n} \max_{i=1,\ldots,n} |\langle g_{k,\alpha} - \operatorname{grad} f(p_k), e_i^k \rangle| \\ &= \frac{\sqrt{n}}{2h_{k,\alpha}} \max_{i=1,\ldots,n} |2h_{k,\alpha}\langle g_{k,\alpha} - \operatorname{grad} f(p_k), e_i^k \rangle|. \end{aligned} \tag{33}$$

Use (32) to rewrite the term within the modulus that appears in (33) as follows:

$$\begin{aligned} &2h_{k,\alpha}\left\langle g_{k,\alpha} - \operatorname{grad} f(p_k), e_i^k \right\rangle \\ &= \hat{f}_k(h_{k,\alpha}e_i^k) - f(p_k) - \langle \operatorname{grad} f(p_k), h_{k,\alpha}e_i^k \rangle - \frac{1}{2}\langle \operatorname{Hess} f(p_k)[h_{k,\alpha}e_i^k], h_{k,\alpha}e_i^k \rangle \\ &\quad - \Big(\hat{f}_k(-h_{k,\alpha}e_i^k) - f(p_k) - \langle \operatorname{grad} f(p_k), (-h_{k,\alpha}e_i^k) \rangle \\ &\quad - \frac{1}{2}\langle \operatorname{Hess} f(p_k)[-h_{k,\alpha}e_i^k], (-h_{k,\alpha}e_i^k) \rangle \Big). \end{aligned}$$

Applying the triangle inequality and utilizing Assumption 2 with $v = h_{k,\alpha}e_i^k$ and $v = -h_{k,\alpha}e_i^k$, we obtain

$$\begin{aligned} &\left| 2h_{k,\alpha}\left\langle g_{k,\alpha} - \operatorname{grad} f(p_k), e_i^k \right\rangle \right| \\ &\leq \left| \hat{f}_k(h_{k,\alpha}e_i^k) - f(p_k) - \langle \operatorname{grad} f(p_k), h_{k,\alpha}e_i^k \rangle - \frac{1}{2}\langle \operatorname{Hess} f(p_k)[h_{k,\alpha}e_i^k], h_{k,\alpha}e_i^k \rangle \right| \\ &\quad + \left| \hat{f}_k(-h_{k,\alpha}e_i^k) - f(p_k) - \langle \operatorname{grad} f(p_k), (-h_{k,\alpha}e_i^k) \rangle - \frac{1}{2}\langle \operatorname{Hess} f(p_k)[-h_{k,\alpha}e_i^k], (-h_{k,\alpha}e_i^k) \rangle \right| \end{aligned}$$

$$\leq \frac{L}{6}\|h_{k,\alpha}e_i^k\|^3 + \frac{L}{6}\|-h_{k,\alpha}e_i^k\|^3 = \frac{L}{3}(h_{k,\alpha})^3.$$

From this inequality, along with (33), Corollary 1, and (31), it follows that

$$\|g_{k,\alpha} - \operatorname{grad} f(p_k)\| \leq \frac{L\sqrt{n}}{6}(h_{k,\alpha})^2 = \frac{L\sqrt{n}}{6(2^{\alpha-1}\sigma_k^2)}\frac{\|v_{k-1}\|^2}{2^{\alpha-1}} \leq \frac{L\sqrt{n}}{3\sigma_1^2}\frac{\|v_{k-1}\|^2}{2^{\alpha-1}}.$$

This completes the proof. □

5.2 Approximations for Hessians

Here, we present some approximations $\mathrm{B}_{k,\alpha}$ satisfying the second inequality in (12). The following result provides a $\mathrm{B}_{k,\alpha}$ inspired by the finite difference approximation of the Hessian proposed in [2]. This approximation relies on the evaluation of grad f and the choice of a mapping $\mathcal{G}\colon \mathrm{T}\mathcal{M} \to \mathrm{T}\mathcal{M}$ associated with Assumption 3. In Remark 4, two natural options for choosing $\mathcal{G}$ are proposed, and their convenience of use depends on the retraction chosen in Algorithm 1.

Proposition 2 *Let $\mathcal{G}\colon \mathrm{T}\mathcal{M} \to \mathrm{T}\mathcal{M}$ be a mapping such that the image of $\mathcal{G}(p,\cdot)\colon \mathrm{T}_p\mathcal{M} \to \mathrm{T}\mathcal{M}$ belongs to $\mathrm{T}_p\mathcal{M}$ for all $p \in \mathcal{M}$. Suppose that Assumption 3 is satisfied. For each iteration k and constant $\alpha \geq 0$, if $\mathrm{B}_{k,\alpha}\colon \mathrm{T}_{p_k}\mathcal{M} \to \mathrm{T}_{p_k}\mathcal{M}$ is defined as*

$$\mathrm{B}_{k,\alpha} = \frac{\mathrm{A}_{k,\alpha} + \mathrm{A}_{k,\alpha}^*}{2}, \tag{34}$$

where $\mathrm{A}_{k,\alpha}\colon \mathrm{T}_{p_k}\mathcal{M} \to \mathrm{T}_{p_k}\mathcal{M}$ is the operator that, for each $i \in \{1,\ldots,n\}$, assumes

$$\mathrm{A}_{k,\alpha}[e_i^k] = \frac{\mathcal{G}(p_k, h_{k,\alpha}e_i^k) - \operatorname{grad} f(p_k)}{h_{k,\alpha}}, \tag{35}$$

then the second inequality in (12) is always satisfied with $\kappa_{\mathrm{B}} = \sqrt{n}L'/(2\sigma_1)$.

Proof It follows from the definition of $\|\cdot\|_{\mathrm{op}}$ along with the orthonormality of the basis $\mathfrak{B}^k$, self-adjoint Hess $f(p_k)$ and $\mathrm{B}_{k,\alpha}$ and (34) that

$$\begin{aligned}
&\left\|\mathrm{B}_{k,\alpha} - \operatorname{Hess} f(p_k)\right\|_{\mathrm{op}} \\
&= \sup\{\|\left(\mathrm{B}_{k,\alpha} - \operatorname{Hess} f(p_k)\right)[v]\| : v \in \mathrm{T}_{p_k}\mathcal{M},\ \|v\| = 1\} \\
&= \sup\left\{\left\|\sum_{i=1}^n \left\langle \left(\mathrm{B}_{k,\alpha} - \operatorname{Hess} f(p_k)\right)[v], e_i^k\right\rangle e_i^k\right\| : v \in \mathrm{T}_{p_k}\mathcal{M},\ \|v\| = 1\right\} \\
&= \sup\left\{\left(\sum_{i=1}^n \left\langle \left(\mathrm{B}_{k,\alpha} - \operatorname{Hess} f(p_k)\right)[v], e_i^k\right\rangle^2\right)^{\frac{1}{2}} : v \in \mathrm{T}_{p_k}\mathcal{M},\ \|v\| = 1\right\} \\
&= \sup\left\{\left(\sum_{i=1}^n \left\langle \left(\mathrm{B}_{k,\alpha} - \operatorname{Hess} f(p_k)\right)[e_i^k], v\right\rangle^2\right)^{\frac{1}{2}} : v \in \mathrm{T}_{p_k}\mathcal{M},\ \|v\| = 1\right\} \\
&\leq \sup\left\{\left(\sum_{i=1}^n \left\|\left(\mathrm{B}_{k,\alpha} - \operatorname{Hess} f(p_k)\right)[e_i^k]\right\|^2 \|v\|^2\right)^{\frac{1}{2}} : v \in \mathrm{T}_{p_k}\mathcal{M},\ \|v\| = 1\right\} \\
&\leq \sqrt{n}\max_{i=1,\ldots,n}\left\|\left(\mathrm{B}_{k,\alpha} - \operatorname{Hess} f(p_k)\right)[e_i^k]\right\|.
\end{aligned}$$

Furthermore, from (34), it comes that

$$\begin{aligned}\left\|\left(\mathrm{B}_{k,\alpha}-\operatorname{Hess} f(p_k)\right)[e_i^k]\right\| &= \frac{1}{2}\left\|\left(\mathrm{A}_{k,\alpha}-\operatorname{Hess} f(p_k)\right)[e_i^k]+\left(\mathrm{A}_{k,\alpha}-\operatorname{Hess} f(p_k)\right)^*[e_i^k]\right\| \\ &= \frac{1}{2}\left(\left\|\left(\mathrm{A}_{k,\alpha}-\operatorname{Hess} f(p_k)\right)[e_i^k]\right\|\right. \\ &\quad \left.+\left\|\left(\mathrm{A}_{k,\alpha}-\operatorname{Hess} f(p_k)\right)^*[e_i^k]\right\|\right) \\ &= \left\|\left(\mathrm{A}_{k,\alpha}-\operatorname{Hess} f(p_k)\right)[e_i^k]\right\|,\end{aligned}$$

for all $i \in \{1,\dots,n\}$. Therefore, considering Assumption 3 with $v = h_{k,\alpha}e_i^k$, (31), (35) and Corollary 1, we obtain

$$\begin{aligned}\left\|\mathrm{B}_{k,\alpha}-\operatorname{Hess} f(p_k)\right\|_{\mathrm{op}} &\le \sqrt{n}\max_{i=1,\dots,n}\left\|(\mathrm{A}_{k,\alpha}-\operatorname{Hess} f(p_k))[e_i^k]\right\| \\ &= \frac{\sqrt{n}}{h_{k,\alpha}}\max_{i=1,\dots,n}\left\|\mathcal{G}(p_k,h_{k,\alpha}e_i^k)-\operatorname{grad} f(p_k)-\operatorname{Hess} f(p_k)[h_{k,\alpha}e_i^k]\right\| \\ &\le \frac{\sqrt{n}L'}{2h_{k,\alpha}}\max_{i=1,\dots,n}\left\|h_{k,\alpha}e_i^k\right\|^2 = \frac{\sqrt{n}L'}{2\sigma_k}\frac{\|v_{k-1}\|}{2^{\alpha-1}} \le \frac{\kappa_{\mathrm{B}}}{2^{\alpha-1}}\|v_{k-1}\|.\end{aligned}$$

This completes the proof. □

In the following remark, we provide an alternative way to express $B_{k,\alpha}$ given in (34) that does not depend on the adjoint operator.

Remark 7 By using the orthonormality of the basis $\mathfrak{B}^k = \{e_1^k,\dots,e_n^k\}$ and (34), we obtain

$$\begin{aligned}\mathrm{B}_{k,\alpha}\,v &= \sum_{i=1}^n\langle \mathrm{B}_{k,\alpha}\,v,e_i\rangle e_i = \sum_{i=1}^n\sum_{j=1}^n\langle v,e_j\rangle\left\langle \mathrm{B}_{k,\alpha}\,e_j,e_i\right\rangle e_i \\ &= \frac{1}{2}\sum_{i=1}^n\sum_{j=1}^n\langle v,e_j\rangle\left\langle(\mathrm{A}_{k,\alpha}+\mathrm{A}_{k,\alpha}^*)e_j,e_i\right\rangle e_i \\ &= \frac{1}{2}\sum_{i=1}^n\sum_{j=1}^n\langle v,e_j\rangle\left(\left\langle \mathrm{A}_{k,\alpha}\,e_j,e_i\right\rangle+\left\langle \mathrm{A}_{k,\alpha}\,e_i,e_j\right\rangle\right)e_i,\end{aligned}$$

for all $v \in \mathrm{T}_{p_k}\mathcal{M}$.

Unlike the previous result, the approximation $\mathrm{B}_{k,\alpha}$ provided in the following result relies solely on evaluations of the objective function f but not the evaluation of the gradient of f.

Proposition 3 *Consider Algorithm 1 with a general second-order retraction (case 1) and with* $\mathrm{R} = \exp$ *(case 2). For each iteration k and constant* $\alpha \ge 0$*, let* $\mathrm{B}_{k,\alpha}\colon \mathrm{T}_{p_k}\mathcal{M} \to \mathrm{T}_{p_k}\mathcal{M}$ *be the operator defined in* (34) *with* $\mathrm{A}_{k,\alpha}\colon \mathrm{T}_{p_k}\mathcal{M} \to \mathrm{T}_{p_k}\mathcal{M}$ *characterized by*

$$\langle \mathrm{A}_{k,\alpha}[e_i^k],e_j^k\rangle = \begin{cases}\dfrac{\hat{f}_k(h_{k,\alpha}e_i^k+h_{k,\alpha}e_j^k)-\hat{f}_k(h_{k,\alpha}e_i^k)-\hat{f}_k(h_{k,\alpha}e_j^k)+\hat{f}_k(0)}{(h_{k,\alpha})^2}, & \text{for case 1,}\\[2ex] \dfrac{f\left(\exp_{q_{k,\alpha}^i}\left(\mathrm{P}_{v_{k,\alpha}^i}v_{k,\alpha}^j\right)\right)-f(q_{k,\alpha}^i)-f(q_{k,\alpha}^j)+f(p_k)}{(h_{k,\alpha})^2}, & \text{for case 2,}\end{cases} \tag{36}$$

for every $i, j \in \{1,\dots,n\}$*, where the notations* $q_{k,\alpha}^i \coloneqq \exp_{p_k} v_{k,\alpha}^i$ *and* $v_{k,\alpha}^i \coloneqq h_{k,\alpha}e_i^k$ *are assumed for every* $i \in \{1,\dots,n\}$*. Additionally, assume that* $\nabla^2\hat{f}_k$ *is L-Lipschitz for case*

1, and assume that Hess f *is* L*-Lipschitz for case 2. Then the second inequality of* (12) *is always satisfied for* $\kappa_{\mathrm{B}} = L(5n + 3\sqrt{n})/6\sigma_1$ *in both cases.*

Proof Take an iteration k and a constant $\alpha \geq 0$. By following the idea of the proof of Proposition 2, we can conclude that

$$\begin{aligned}\left\|\mathrm{B}_{k,\alpha} - \mathrm{Hess}\, f(p_k)\right\|_{\mathrm{op}} &\leq \sqrt{n} \max_{i=1,\ldots,n} \left\|(\mathrm{A}_{k,\alpha} - \mathrm{Hess}\, f(p_k))[e_i^k]\right\| \\ &= \frac{\sqrt{n}}{h_{k,\alpha}} \max_{i=1,\ldots,n} \left\|h_{k,\alpha}\, \mathrm{A}_{k,\alpha}[e_i^k] - \mathrm{Hess}\, f(p_k)[v_{k,\alpha}^i]\right\| \end{aligned} \tag{37}$$

for both cases. Given that $\nabla^2 \hat{f}_k$ is L-Lipschitz for case 1 and Hess f is L-Lipschitz for case 2, it follows from Remark 5 with $v = v_{k,\alpha}^i$ that

$$\left\|\mathcal{G}(p_k, v_{k,\alpha}^i) - \mathrm{grad}\, f(p_k) - \mathrm{Hess}\, f(p_k)[v_{k,\alpha}^i]\right\| \leq \frac{L}{2}\|v_{k,\alpha}^i\|^2, \tag{38}$$

where

$$\mathcal{G}(p_k, v_{k,\alpha}^i) = \begin{cases} \nabla \hat{f}_k(v_{k,\alpha}^i), & \text{for case 1,} \\ \mathrm{P}_{v_{k,\alpha}^i}^{-1}\, \mathrm{grad}\, f(\exp_{p_k} v_{k,\alpha}^i), & \text{for case 2.} \end{cases} \tag{39}$$

Then, by applying (38), along with the definition of $v_{k,\alpha}^i$ and the orthonormality of $\mathfrak{B}^k$, we find

$$\begin{aligned}&\left\|h_{k,\alpha}\, \mathrm{A}_{k,\alpha}[e_i^k] - \mathrm{Hess}\, f(p_k)[v_{k,\alpha}^i]\right\| \\ &= \left\|h_{k,\alpha}\, \mathrm{A}_{k,\alpha}[e_i^k] - \mathcal{G}(p_k, v_{k,\alpha}^i) + \mathrm{grad}\, f(p_k) + \mathcal{G}(p_k, v_{k,\alpha}^i) - \mathrm{grad}\, f(p_k) - \mathrm{Hess}\, f(p_k)[v_{k,\alpha}^i]\right\| \\ &\leq \left\|h_{k,\alpha}\, \mathrm{A}_{k,\alpha}[e_i^k] - \mathcal{G}(p_k, v_{k,\alpha}^i) + \mathrm{grad}\, f(p_k)\right\| + \left\|\mathcal{G}(p_k, v_{k,\alpha}^i) - \mathrm{grad}\, f(p_k) - \mathrm{Hess}\, f(p_k)[v_{k,\alpha}^i]\right\| \\ &\leq \left\|\sum_{j=1}^{n} \left\langle h_{k,\alpha}\, \mathrm{A}_{k,\alpha}[e_i^k] - \mathcal{G}(p_k, v_{k,\alpha}^i) + \mathrm{grad}\, f(p_k), e_j^k\right\rangle e_j^k\right\| + \frac{L}{2}\|v_{k,\alpha}^i\|^2 \\ &\leq \sqrt{n} \max_{j=1,\ldots,n} \left|\left\langle h_{k,\alpha}\, \mathrm{A}_{k,\alpha}[e_i^k] - \mathcal{G}(p_k, v_{k,\alpha}^i) + \mathrm{grad}\, f(p_k), e_j^k\right\rangle\right| + \frac{L}{2}(h_{k,\alpha})^2 \\ &= \frac{\sqrt{n}}{h_{k,\alpha}} \max_{j=1,\ldots,n} \left|\underbrace{(h_{k,\alpha})^2 \left\langle \mathrm{A}_{k,\alpha}[e_i^k], e_j^k\right\rangle - \left\langle \mathcal{G}(p_k, v_{k,\alpha}^i), v_{k,\alpha}^j\right\rangle + \left\langle \mathrm{grad}\, f(p_k), v_{k,\alpha}^j\right\rangle}_{(i)}\right| + \frac{L}{2}(h_{k,\alpha})^2. \end{aligned} \tag{40}$$

Note that, as a consequence of (36) and (39), the expression (i) can be rephrased as

$$\begin{aligned}&(h_{k,\alpha})^2 \left\langle \mathrm{A}_{k,\alpha}[e_i^k], e_j^k\right\rangle - \left\langle \nabla \hat{f}_k(h_{k,\alpha} e_i^k), h_{k,\alpha} e_j^k\right\rangle + \left\langle \nabla \hat{f}_k(0), h_{k,\alpha} e_j^k\right\rangle \\ &= \hat{f}_k(h_{k,\alpha} e_i^k + h_{k,\alpha} e_j^k) - \hat{f}_k(h_{k,\alpha} e_i^k) - \left\langle \nabla \hat{f}_k(h_{k,\alpha} e_i^k), h_{k,\alpha} e_j^k\right\rangle \\ &\quad - \frac{1}{2}\left\langle \nabla^2 \hat{f}_k(h_{k,\alpha} e_i^k)[h_{k,\alpha} e_j^k], h_{k,\alpha} e_j^k\right\rangle \\ &\quad - \left[\hat{f}_k(h_{k,\alpha} e_j^k) - \hat{f}_k(0) - \left\langle \nabla \hat{f}_k(0), h_{k,\alpha} e_j^k\right\rangle - \frac{1}{2}\left\langle \nabla^2 \hat{f}_k(0)[h_{k,\alpha} e_j^k], h_{k,\alpha} e_j^k\right\rangle\right] \\ &\quad + \frac{1}{2}\left\langle \left(\nabla^2 \hat{f}_k(h_{k,\alpha} e_i^k) - \nabla^2 \hat{f}_k(0)\right)[h_{k,\alpha} e_j^k], h_{k,\alpha} e_j^k\right\rangle, \end{aligned}$$

for case 1, and as

$$
\begin{aligned}
&(h_{k,\alpha})^2 \left\langle \mathrm{A}_{k,\alpha}[e_i^k], e_j^k \right\rangle - \left\langle \mathrm{P}^{-1}_{v^i_{k,\alpha}} \operatorname{grad} f(q^i_{k,\alpha}), v^j_{k,\alpha} \right\rangle + \left\langle \operatorname{grad} f(p_k), v^j_{k,\alpha} \right\rangle \\
&= f\left(\exp_{q^i_{k,\alpha}}\left(\mathrm{P}_{v^i_{k,\alpha}} v^j_{k,\alpha}\right)\right) - f(q^i_{k,\alpha}) - f(q^j_{k,\alpha}) + f(p_k) - \left\langle \mathrm{P}^{-1}_{v^i_{k,\alpha}} \operatorname{grad} f(q^i_{k,\alpha}), v^j_{k,\alpha} \right\rangle \\
&+ \left\langle \operatorname{grad} f(p_k), v^j_{k,\alpha} \right\rangle \\
&= f\left(\exp_{q^i_{k,\alpha}}\left(\mathrm{P}_{v^i_{k,\alpha}} v^j_{k,\alpha}\right)\right) - f(q^i_{k,\alpha}) - \left\langle \operatorname{grad} f(q^i_{k,\alpha}), \mathrm{P}_{v^i_{k,\alpha}} v^j_{k,\alpha} \right\rangle \\
&- \frac{1}{2}\left\langle \operatorname{Hess} f(q^i_{k,\alpha})[\mathrm{P}_{v^i_{k,\alpha}} v^j_{k,\alpha}], \mathrm{P}_{v^i_{k,\alpha}} v^j_{k,\alpha} \right\rangle \\
&\quad - \left[f(q^j_{k,\alpha}) - f(p_k) - \left\langle \operatorname{grad} f(p_k), v^j_{k,\alpha} \right\rangle - \frac{1}{2}\left\langle \operatorname{Hess} f(p_k)[v^j_{k,\alpha}], v^j_{k,\alpha} \right\rangle \right] \\
&\quad + \frac{1}{2}\left\langle \left(\mathrm{P}^{-1}_{v^i_{k,\alpha}} \circ \operatorname{Hess} f(q^i_{k,\alpha}) \circ \mathrm{P}_{v^i_{k,\alpha}} - \operatorname{Hess} f(p_k) \right) v^j_{k,\alpha}, v^j_{k,\alpha} \right\rangle,
\end{aligned}
$$

for case 2. Applying the modulus to both sides, utilizing the triangular inequality, considering that both $\nabla^2 \hat{f}_k$ and Hess f are L-Lipschitz, and employing Lemma 1, we obtain

$$
\begin{aligned}
&\left|(h_{k,\alpha})^2 \left\langle \mathrm{A}_{k,\alpha}[e_i^k], e_j^k \right\rangle - \left\langle \nabla \hat{f}_k(h_{k,\alpha} e_i^k), h_{k,\alpha} e_j^k \right\rangle + \left\langle \nabla \hat{f}_k(0), h_{k,\alpha} e_j^k \right\rangle\right| \\
&\le \left| \hat{f}_k(h_{k,\alpha} e_i^k + h_{k,\alpha} e_j^k) - \hat{f}_k(h_{k,\alpha} e_i^k) - \left\langle \nabla \hat{f}_k(h_{k,\alpha} e_i^k), h_{k,\alpha} e_j^k \right\rangle \right. \\
&\left. - \frac{1}{2}\left\langle \nabla^2 \hat{f}_k(h_{k,\alpha} e_i^k)[h_{k,\alpha} e_j^k], h_{k,\alpha} e_j^k \right\rangle \right| \\
&\quad + \left| \hat{f}_k(h_{k,\alpha} e_j^k) - \hat{f}_k(0) - \left\langle \nabla \hat{f}_k(0), h_{k,\alpha} e_j^k \right\rangle - \frac{1}{2}\left\langle \nabla^2 \hat{f}_k(0)[h_{k,\alpha} e_j^k], h_{k,\alpha} e_j^k \right\rangle \right| \\
&\quad + \frac{1}{2}\left| \left\langle \left(\nabla^2 \hat{f}_k(h_{k,\alpha} e_i^k) - \nabla^2 \hat{f}_k(0) \right)[h_{k,\alpha} e_j^k], h_{k,\alpha} e_j^k \right\rangle \right| \\
&\le \frac{L}{6}\|h_{k,\alpha} e_j^k\|^3 + \frac{L}{6}\|h_{k,\alpha} e_j^k\|^3 + \frac{1}{2}\left\| \nabla^2 \hat{f}_k(h_{k,\alpha} e_i^k) - \nabla^2 \hat{f}_k(0) \right\|_{\mathrm{op}} \|h_{k,\alpha} e_j^k\| \|h_{k,\alpha} e_j^k\| \\
&\le \frac{L}{3}\|h_{k,\alpha} e_j^k\|^3 + \frac{L}{2}\|h_{k,\alpha} e_j^k\|^3 = \frac{5L}{6}(h_{k,\alpha})^3
\end{aligned}
$$

and

$$
\begin{aligned}
&\left|(h_{k,\alpha})^2 \left\langle \mathrm{A}_{k,\alpha}[e_i^k], e_j^k \right\rangle - \left\langle \mathrm{P}^{-1}_{v^i_{k,\alpha}} \operatorname{grad} f(q^i_{k,\alpha}), v^j_{k,\alpha} \right\rangle + \left\langle \operatorname{grad} f(p_k), v^j_{k,\alpha} \right\rangle\right| \\
&\le \left| f\left(\exp_{q^i_{k,\alpha}}\left(\mathrm{P}_{v^i_{k,\alpha}} v^j_{k,\alpha}\right)\right) - f(q^i_{k,\alpha}) - \left\langle \operatorname{grad} f(q^i_{k,\alpha}), \mathrm{P}_{v^i_{k,\alpha}} v^j_{k,\alpha} \right\rangle \right. \\
&\left. - \frac{1}{2}\left\langle \operatorname{Hess} f(q^i_{k,\alpha})[\mathrm{P}_{v^i_{k,\alpha}} v^j_{k,\alpha}], \mathrm{P}_{v^i_{k,\alpha}} v^j_{k,\alpha} \right\rangle \right| \\
&\quad + \left| f(q^j_{k,\alpha}) - f(p_k) - \left\langle \operatorname{grad} f(p_k), v^j_{k,\alpha} \right\rangle - \frac{1}{2}\left\langle \operatorname{Hess} f(p_k)[v^j_{k,\alpha}], v^j_{k,\alpha} \right\rangle \right| \\
&\quad + \frac{1}{2}\left| \left\langle \left(\mathrm{P}^{-1}_{v^i_{k,\alpha}} \circ \operatorname{Hess} f(q^i_{k,\alpha}) \circ \mathrm{P}_{v^i_{k,\alpha}} - \operatorname{Hess} f(p_k) \right) v^j_{k,\alpha}, v^j_{k,\alpha} \right\rangle \right|
\end{aligned}
$$

$$\leq \frac{L}{6}\|\mathrm{P}_{v^i_{k,\alpha}} v^j_{k,\alpha}\|^3 + \frac{L}{6}\|v^j_{k,\alpha}\|^3 + \frac{1}{2}\left\|\mathrm{P}^{-1}_{v^i_{k,\alpha}} \circ \mathrm{Hess}\, f(q^i_{k,\alpha}) \circ \mathrm{P}_{v^i_{k,\alpha}} - \mathrm{Hess}\, f(p_k)\right\|_{\mathrm{op}} \|v^j_{k,\alpha}\|^2$$
$$\leq \frac{L}{6}\|v^j_{k,\alpha}\|^3 + \frac{L}{6}\|v^j_{k,\alpha}\|^3 + \frac{L}{2}\|v^j_{k,\alpha}\|^3 = \frac{5L}{6}(h_{k,\alpha})^3.$$

By combining the above calculations with (37) and (40), in any case, we have

$$\left\|\mathrm{B}_{k,\alpha} - \mathrm{Hess}\, f(p_k)\right\|_{\mathrm{op}} \leq \frac{\sqrt{n}}{h_{k,\alpha}}\left(\frac{\sqrt{n}}{h_{k,\alpha}}\frac{5L}{6}(h_{k,\alpha})^3 + \frac{L}{2}(h_{k,\alpha})^2\right) = \frac{L(5n+3\sqrt{n})}{6\sigma_k}\frac{\|v_{k-1}\|}{2^{\alpha-1}}.$$

Therefore, the conclusion of the proof follows from Corollary 1. □

6 Numerical experiments

In this section, the numerical performance of Algorithm 1 is illustrated by implementing the derivative-free form with the Manopt package [7], where the approximated gradient $g_{k,\alpha}$ and approximated Hessian $B_{k,\alpha}$ are computed by (32) and (36), respectively. The Riemannian conjugate gradient method with Polak-Ribiere update formula [7] was used to solve the cubic subproblem in Algorithm 1. The parameters are initialized as $\sigma_1 = 1$ and $\theta = 1$ and the outer iteration terminates when $\|g_{k,\alpha}\| \leq 10^{-6}$, unless stated otherwise. For the exactness of the derivative-free form of Algorithm 1 and comparison, the ARC method [3] is implemented as well. The cubic subproblem in the ARC method is also solved by using the Riemannian conjugate gradient method with Polak-Ribiere update formula. The ARC method terminates when $\|\mathrm{grad}\, f(p_k)\| \leq 10^{-6}$. All the codes executions are carried out on a MacbookPro running macOS Sequoia, 15.1.1, with 16 GB RAM, an Apple M1 Pro CPU, and Matlab R2022a. Additionally, to ensure reproducibility, the randomness is fixed by using the Matlab built-in command `rng(2024)`.

We first consider nine Riemannian optimization problems on nine different Riemannian manifolds:

1. Top eigenvalue is to solve
$$\max_{X\in \mathrm{Sp}(r)} X^T A X,$$
where $A \in \mathbb{R}^{r\times r}$ is symmetric (randomly generated from i.i.d Gaussian entries) and $\mathrm{Sp}(r) = \{X | X \in \mathbb{R}^r, X^T X = 1\}$ is a sphere manifold. Optimal objective value corresponds to the largest eigenvalue of A.
2. Dominant invariant subspace [15] is to solve
$$\max_{X\in \mathrm{Gr}(r,t)} \frac{1}{2}\mathrm{Trace}(X^T A X),$$
where $A \in \mathbb{R}^{r\times r}$ is symmetric (randomly generated from i.i.d Gaussian entries) and $\mathrm{Gr}(r,t) = \{\mathrm{span}(X) : X \in \mathbb{R}^{r\times t}, X^T X = I_t\}$ is a Grassmann manifold. Optima correspond to dominant invariant subspaces of A .
3. Dominant complex invariant subspace [7] is to solve
$$\max_{X\in \mathrm{cGr}(r,t)} \mathrm{Re}(\mathrm{Trace}(X^H A X)),$$
where $A \in \mathbb{C}^{r\times r}$ is Hermitian (randomly generated from i.i.d Gaussian entries) and $\mathrm{cGr}(r,t) = \{\mathrm{span}(X) : X \in \mathbb{C}^{r\times t}, X^H X = I_t\}$ is a complex Grassmann manifold. Optima correspond to dominant complex invariant subspaces of A.

4. Generalized eigenvalue computation [2] is to solve

$$\max_{X \in \mathrm{gGr}(r,t)} \frac{1}{2}\mathrm{Trace}(X^T A X),$$

where $A \in \mathbb{R}^{r\times r}$ is symmetric (randomly generated from i.i.d Gaussian entries) and $\mathrm{gGr}(r,t) = \{\mathrm{span}(X) : X \in \mathbb{R}^{r\times t}, X^T B X = I_t\}$ for some $B \succ 0$ is a generalized Grassmann manifold.
5. Elliptopt SDP problem [6] is to solve

$$\min_{X \in \mathrm{Ob}(r,t)} \frac{1}{2}\mathrm{Trace}(X^T A X),$$

where $A \in \mathbb{R}^{r\times r}$ is symmetric (randomly generated from i.i.d Gaussian entries) and $\mathrm{Ob}(r,t) = \{X | X \in \mathbb{R}^{r\times t}, (XX^T)_{ii} = 1, i = 1, 2, \cdots, r\}$ is an oblique manifold. The above problem is equivalent to the following SDP problem,

$$\min_{Y \in \mathbb{R}^{r\times r}} AY, \text{ s.t. } \mathrm{diag}(Y) = 1 \text{ and } Y \text{ is positive semidefinite.}$$

6. Elliptopt SDP complex problem [7] is to solve

$$\min_{X \in \mathrm{cOb}(r,t)} \frac{1}{2}\mathrm{Re}(\mathrm{Trace}(X^H A X)),$$

where $A \in \mathbb{C}^{r\times r}$ is Hermitian (randomly generated from i.i.d Gaussian entries) and $\mathrm{cOb}(r,t) = \{X | X \in \mathbb{C}^{r\times t}, (XX^H)_{ii} = 1, i = 1, 2, \cdots, r\}$ is a complex oblique manifold.
7. Truncated SVD is to solve

$$\max_{U \in \mathrm{St}(r,t), V \in \mathrm{St}(s,t)} \mathrm{Tr}(U^T A V N),$$

where $A \in \mathbb{R}^{r\times s}$ has i.i.d. random Gaussian entries, $N = \mathrm{diag}(t, t-1, \cdots, 1)$ and $\mathrm{St}(r,t) = \{X | X \in \mathbb{R}^{r\times t}, X^T X = I_t\}$ is a Stiefel manifold. Global optima correspond to the t dominant left and right singular vectors of A [27].
8. ShapeFit aims to reconstruct a rigid structure comprising r point $x_1, \cdots, x_r$ in $\mathbb{R}^t$ through a least-squares formulation. This reconstruction is derived from noisy measurements of selected pairwise directions $\frac{x_i - x_j}{\|x_i - x_j\|}$, where the pairs are uniformly chosen at random [18]. The set of points is centered and obeys one extra linear constraint to fix scaling ambiguity, so that the search space is a manifold, represent by $\mathrm{Sf}(t,r) = \{X | X \in \mathbb{R}^{t\times r}, XE = 0, \mathrm{Tr}(M^T X) = 1\}$, where every entry of $E \in \mathbb{R}^r$ is 1 and $M \in \mathbb{R}^{t\times r}$ is a given matrix. Refer [7] to construct the objective function and M.
9. Synchronization of rotations is to estimate t rotation matrices $Q_1, \cdots, Q_t$ in the orthogonal group $\mathrm{SO}(r)$ from noisy relative measurements $H_{ij} \approx Q_i Q_j^T$ for an Erdos-Renyi random set of pairs (i, j) following a maximum likelihood formulation [8]. The details regarding the distribution of measurements and the associated objective function can be found in the referenced material. Additionally, the initial guess is utilzed as the reference to prevent convergence to an undesirable local optimum.

For each of the above mentioned nine problems, we create five instances and generate five random initial points, except for problem 9, which is initialized deterministically. Subsequently, we execute each algorithm starting from the same initial guess on that specific instance. From Table 2, it is evident that, for all the given problems, the norms of the approximated Riemannian gradients and exact Riemannian gradients produced by both Algorithm 1

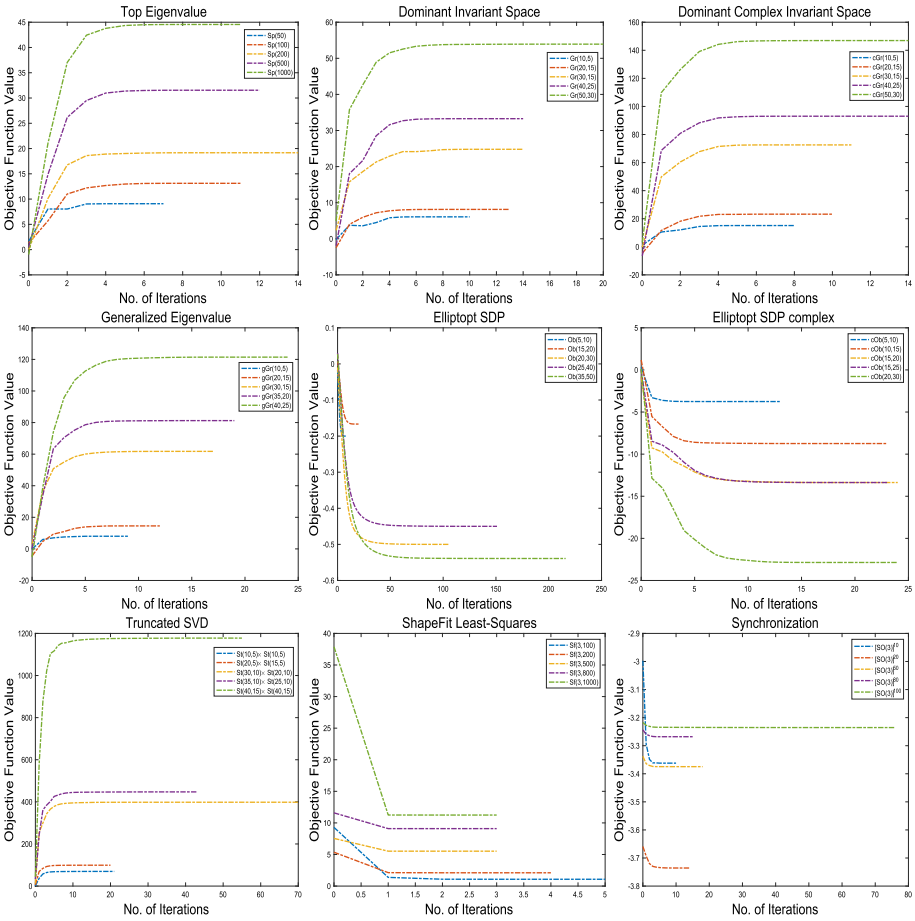

Fig. 1 Objective function value at each iteration of the nine Riemannian optimization problems with different manifolds

and ARC are consistently smaller than 10^{-6}. Notably, Algorithm 1 and ARC yield identical objective function values, which affirms that Algorithm 1 successfully achieves a minimum for these nine problems. Figures 1 and 2 visually represent the objective function values and the norms of the approximated Riemannian gradients of Algorithm 1 at each iterate across the nine Riemannian optimization problems with nine different manifolds. The objective function values exhibit monotonic behavior on the tested problems, and as the point approaches the minimum, the corresponding norm of the approximated Riemannian gradient rapidly converges, aligning with the convergence theory of adaptive regularization with cubic.

Recent advancements in adversarial machine learning have focused on developing black-box attack strategies against deep neural networks (DNNs). Despite their remarkable performance, DNNs remain vulnerable to subtle perturbations in input data, which can lead to significant misclassifications [16]. Addressing this vulnerability necessitates a dual approach: enhancing the robustness of DNNs to withstand such attacks and designing more sophisticated attack methods to thoroughly evaluate their limitations. In real-world scenarios, attackers often lack access to the internal architecture of the target model, prompting the

Table 2 The results of Algorithm 1 and ARC on nine Riemannian optimization problems with different Riemannian manifolds. OFV and #It represent the objective function value and number of iterations, respectively. $\|g_{k,\alpha}\|$ is the norm of the approximated Riemannian gradient in Algorithm 1 and $\|\text{ grad } f(p_k)\|$ is the norm of the exact Riemannian gradient in ARC.

		Algorithm 1			ARC		
problem	manifold	OFV	$\|g_{k,\alpha}\|$	#It	OFV	$\|\text{ grad } f(p_k)\|$	#It
1	Sp(50)	9.0821	5.4×10^{-9}	7	9.0821	3.3×10^{-8}	6
	Sp(100)	13.1192	5.9×10^{-9}	11	13.1192	2.5×10^{-7}	7
	Sp(200)	19.1694	3.9×10^{-7}	14	19.1694	2.1×10^{-8}	8
	Sp(500)	31.5403	5.6×10^{-10}	12	31.5403	5.1×10^{-7}	6
	Sp(1000)	44.5548	8.2×10^{-8}	11	44.5548	7.5×10^{-8}	5
2	Gr(10, 5)	6.0754	1.1×10^{-8}	9	6.0754	2.1×10^{-7}	7
	Gr(20, 15)	8.1365	7.1×10^{-11}	13	8.1365	1.9×10^{-13}	9
	Gr(30, 15)	24.7934	8.9×10^{-10}	14	24.7934	4.2×10^{-11}	8
	Gr(40, 25)	33.2728	3.7×10^{-10}	14	33.2728	1.3×10^{-13}	8
	Gr(50, 30)	53.9229	6.8×10^{-9}	20	53.9229	1.1×10^{-10}	10
3	cGr(10, 5)	15.1420	1.2×10^{-9}	8	15.1420	2.3×10^{-8}	6
	cGr(20, 15)	23.2484	4.3×10^{-8}	9	23.2484	7.7×10^{-9}	7
	cGr(30, 15)	72.5870	3.1×10^{-7}	10	72.5870	7.3×10^{-8}	9
	cGr(40, 25)	93.0367	2.5×10^{-9}	14	93.0367	3.0×10^{-11}	8
	cGr(50, 30)	146.9301	3.3×10^{-8}	14	146.9301	1.0×10^{-13}	9
4	gGr(10, 5)	8.0148	4.4×10^{-8}	9	8.0148	1.5×10^{-10}	8
	gGr(20, 15)	14.5517	6.6×10^{-10}	12	14.5517	2.8×10^{-7}	7
	gGr(30, 15)	61.7585	7.2×10^{-9}	17	61.7585	1.7×10^{-9}	10
	gGr(35, 20)	81.2347	1.1×10^{-7}	19	81.2347	2.7×10^{-8}	12
	gGr(40, 25)	121.4336	1.2×10^{-7}	24	121.4336	3.1×10^{-12}	12
5	Ob(5, 10)	-0.2000	7.2×10^{-10}	9	-0.2000	8.6×10^{-12}	6
	Ob(15, 20)	-0.1667	2.3×10^{-7}	20	-0.1667	7.2×10^{-13}	8
	Ob(20, 30)	-0.5000	2.0×10^{-7}	104	-0.5000	8.5×10^{-7}	31
	Ob(25, 40)	-0.4500	9.5×10^{-7}	147	-0.4500	6.5×10^{-7}	41
	Ob(35, 50)	-0.5391	9.3×10^{-7}	210	-0.5391	3.8×10^{-7}	56

Table 2 continued

problem	manifold	Algorithm 1			ARC		
		OFV	$\|g_{k,\alpha}\|$	#It	OFV	$\|\text{grad } f(p_k)\|$	#It
6	cOb(5, 10)	-3.7542	4.2×10^{-7}	13	-3.7542	3.2×10^{-7}	6
	cOb(10, 15)	-8.7412	1.8×10^{-7}	23	-8.7412	1.4×10^{-9}	8
	cOb(15, 20)	-13.3691	5.6×10^{-8}	24	-13.3691	3.8×10^{-9}	8
	cOb(15, 25)	-13.3691	2.5×10^{-7}	23	-13.3691	9.1×10^{-9}	8
	cOb(20, 30)	-22.8624	2.5×10^{-7}	24	-22.8624	2.2×10^{-8}	8
7	$\text{St}(10, 5) \times \text{St}(10, 5)$	69.6192	8.1×10^{-8}	20	69.6192	2.9×10^{-11}	13
	$\text{St}(20, 5) \times \text{St}(15, 5)$	99.0874	1.6×10^{-10}	20	99.0874	1.4×10^{-8}	10
	$\text{St}(30, 10) \times \text{St}(20, 10)$	398.2388	1.6×10^{-9}	70	398.2388	1.1×10^{-11}	18
	$\text{St}(35, 10) \times \text{St}(25, 10)$	447.2950	2.8×10^{-8}	42	447.2950	6.7×10^{-8}	14
	$\text{St}(40, 15) \times \text{St}(40, 15)$	1177.7358	9.2×10^{-9}	55	1177.7358	1.5×10^{-9}	18
8	Sf(3, 100)	1.0783	4.4×10^{-9}	5	1.0783	3.5×10^{-7}	4
	Sf(3, 200)	2.1040	3.7×10^{-10}	4	2.1040	5.1×10^{-14}	4
	Sf(3, 500)	5.5249	1.6×10^{-11}	3	5.5249	1.0×10^{-12}	3
	Sf(3, 800)	9.0949	1.2×10^{-10}	3	9.0949	8.0×10^{-15}	3
	Sf(3, 1000)	11.2434	4.2×10^{-11}	3	11.2434	7.3×10^{-13}	3
9	$\underbrace{[\text{SO}(3)] \times \cdots \times [\text{SO}(3)]}_{10}$	-3.3621	3.4×10^{-7}	10	-3.3621	1.3×10^{-7}	5
	$\underbrace{[\text{SO}(3)] \times \cdots \times [\text{SO}(3)]}_{20}$	-3.7356	5.4×10^{-7}	14	-3.7356	1.9×10^{-9}	6
	$\underbrace{[\text{SO}(3)] \times \cdots \times [\text{SO}(3)]}_{50}$	-3.3746	7.3×10^{-7}	18	-3.3746	3.4×10^{-7}	5
	$\underbrace{[\text{SO}(3)] \times \cdots \times [\text{SO}(3)]}_{80}$	-3.2680	9.1×10^{-7}	15	-3.2680	4.5×10^{-8}	6
	$\underbrace{[\text{SO}(3)] \times \cdots \times [\text{SO}(3)]}_{100}$	-3.2354	4.0×10^{-7}	76	-3.2354	6.8×10^{-7}	8

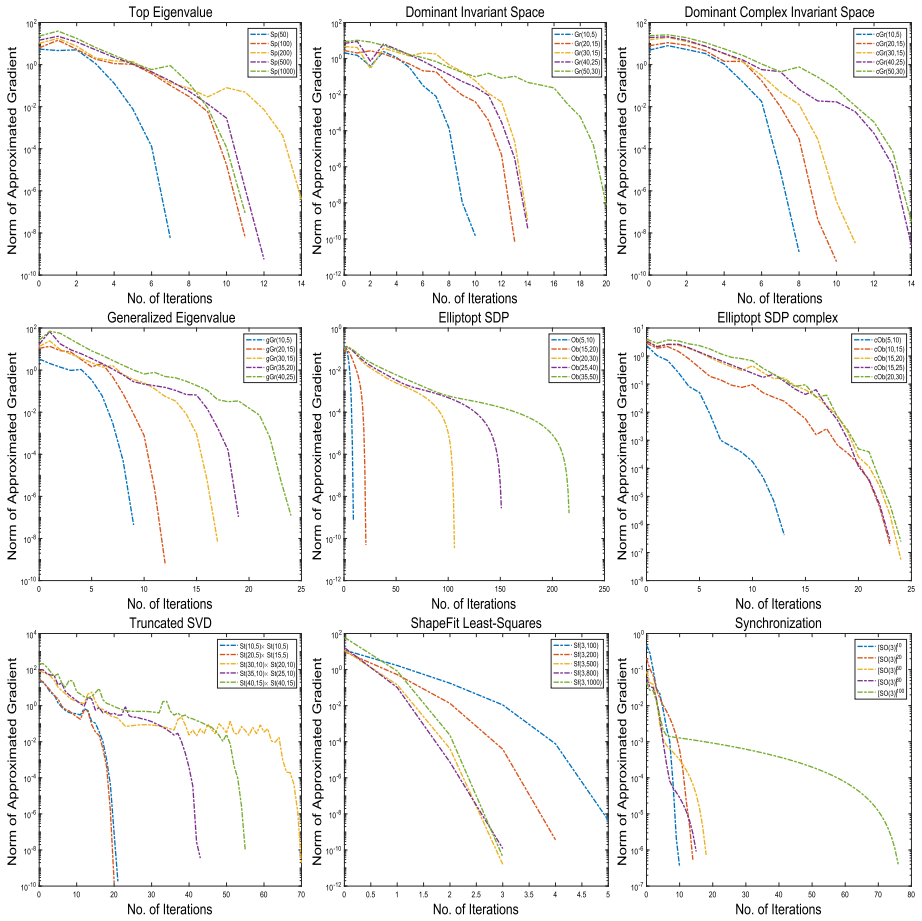

Fig. 2 Norm of approximated Riemannian gradient at each iteration of the nine Riemannian optimization problems with different manifolds

use of derivative-free optimization techniques to craft adversarial examples [29]. However, a notable limitation of existing methods is that the generated perturbations often deviate from the natural structure of the input data. For example, while natural images typically reside on a specific manifold [32], the adversarial modifications applied to them frequently fail to respect this inherent geometric structure. This insight highlights the potential of leveraging derivative-free Riemannian optimization techniques to create adversarial examples that not only deceive DNNs but also maintain the intrinsic manifold properties of the original data [23].

Next, we consider the following problem.

10. Minimizing a composite function on a sphere manifold is to solve

$$\min_{X\in \mathrm{Sp}(r_1)} \frac{1}{r_4}\| f_3 \circ f_2 \circ f_1(X)\|, \tag{41}$$

where $f_i(X) = \mathrm{Swish}(A_i X + b_i)$, $A_i \in \mathbb{R}^{r_{i+1}\times r_i}$, $b_i \in \mathbb{R}^{r_{i+1}}$, $i = 1, \cdots, 3$. Here, $(\mathrm{Swish}(b))_j = \mathrm{swish}(b_j)$ and $\mathrm{swish}(x) = \frac{x}{1+\exp(-x)}$, namely, the function Swish is to

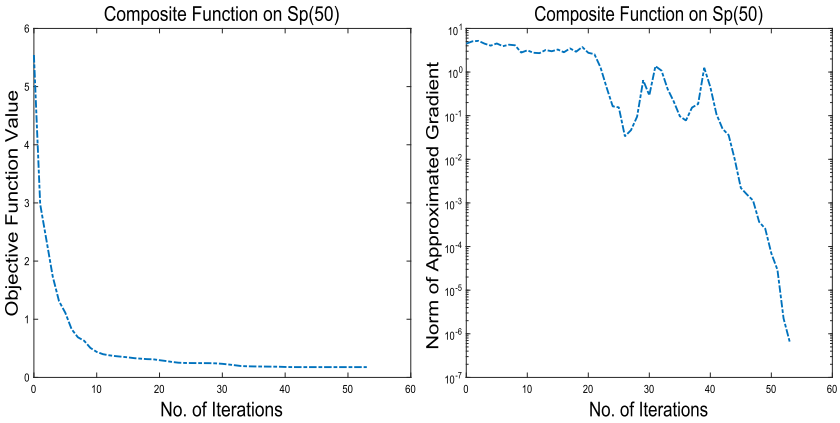

Fig. 3 Objective function value and norm of approximated Riemannian gradient at each iteration for minimizing the composite function on a sphere manifold

use function swish to act on each component. Here, A_i and $b_i, i = 1, \cdots, 3$ are randomly generated from i.i.d Gaussian entries and $r_i, i = 1, \cdots, 4$ are set as 50, 100, 90 and 80, respectively.

The objective function in problem 10 usually appears in the deep learning, which can be considered as a three-layer fully connected neural network. The function swish is an active function [26]. Next, we reinterpret the objective function in problem 10 as a pre-trained neural network. Under this framework, problem 10 transforms into a neural network adversarial attack scenario, with the underlying data distribution adhering to a spherical manifold. Given that the internal architecture of the trained model is inaccessible, the objective function in problem 10 effectively operates as a black-box function. Consequently, it can be addressed using derivative-free optimization techniques, specifically Algorithm 1. After applying Algorithm 1, we obtain a point that corresponding objective function value is 0.1770 and norm of the approximated Riemannian gradient $\|g_{k,\alpha}\|$ is 6.6×10^{-7}. Combing with the discussion of previous 9 problems, Algorithm 1 indeed gives a minimum for problem 10. In addition, Figure 3 displays the objective function value and norm of the approximated Riemannian gradient at each iteration for problem 10.

We also consider another real world application.

11. Minimizing a function on the stiffness matrix is to solve

$$\min_{X^{\mathcal{P}} \in \mathcal{S}^3_{++}} w_p \|\hat{p} - p\|^2 + w_d \det(X^{\mathcal{P}}) + w_c \text{cond}(X^{\mathcal{P}}), \tag{42}$$

where $\mathcal{S}^3_{++} = \{X | X \in \mathbb{R}^{3\times 3}, X = X^T, X \succ 0\}$, $\det(X)$ is the determinant of X, $\text{cond}(X)$ is the condition number of X, and w_p, w_d, and w_c are positive parameters.

This problem is derived from the control of robotics [19]. With a constant external force f^e applied to the system, we have the following identity which solves p by $f^e = X^{\mathcal{P}}(\hat{p} - p) - X^{\mathcal{D}}\dot{p}$, where the damping matrix $X^{\mathcal{D}} = X^{\mathcal{P}}$ for the critical damped case. Here, we use the same parameters as in [19]. As the above problem is over the positive definite manifold and it is hard to compute its corresponding gradient and Hessian, we employ Algorithm 1. For this problem, the algorithmic parameter σ_1 is set 0.0005 and the outer iteration terminates when $\|g_{k,\alpha}\| \leq 10^{-10}$. We then obtain a point that corresponding objective function value is

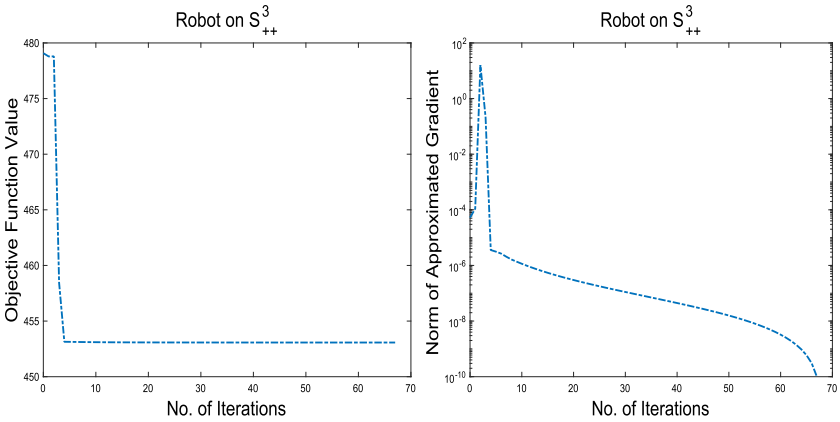

Fig. 4 Objective function value and norm of approximated Riemannian gradient at each iteration for minimizing the function on positive definite manifold

453.0721 and norm of the approximated Riemannian gradient $\|g_{k,\alpha}\|$ is 9.9×10^{-11}, which shows that it is a minima for problem 11. In addition, Figure 4 displays the objective function value and norm of the approximated Riemannian gradient at each iteration for problem 11.

Finally, we note that Algorithm 1, as a derivative-free method involving the computation of orthonormal bases for the tangent space, generally requires longer running times compared to ARC when the gradient and Hessian of the Riemannian optimization problem are easily computable. Additionally, for the various instances of problems 1-9, the iteration number of Algorithm 1 is often comparable to that of ARC, though in several cases, Algorithm 1 requires more iterations. This discrepancy may stem from two factors: first, Algorithm 1 approximates the gradient and Hessian, introducing potential errors; second, the updating rule for the penalty parameter σ in Algorithm 1 is less flexible than in ARC, as σ_k demonstrates a strictly monotonically increasing behavior. Consequently, the choice of σ_1 significantly impacts the convergence rate for each specific problem. It would be valuable to design an effective strategy for updating σ_k in future work.

7 Conclusion

In this paper, we present an inexact-Newton algorithm with cubic regularization tailored specifically for Riemannian manifolds. The distinctive feature of this algorithm lies in its ability to operate without prior knowledge of the gradient and Hessian of the objective function. Instead, it only requires approximations that satisfy a condition analogous to those mandated by inexact algorithms. Importantly, we demonstrate that approximations obtained through finite-differences meet this condition, allowing us to assert that the algorithm can be considered derivative-free.

Regarding future research directions, we find the exploration and application of the inexactness outlined in this paper in algorithms with non-differentiable objective functions to be intriguing. This exploration could potentially lead to the development of a subdifferential-free algorithm, a prospect of particular significance given the often challenging nature of calculating a subgradient.

Acknowledgements The authors would like to thank the anonymous referees for their valuable comments and suggestions, which have significantly improved the quality of this paper.

Data Availability Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

Declarations

Conflicts of Interest The authors declare that they have no conflict of interest.

References

1. Absil, P.-A., Baker, C.G., Gallivan, K.A.: Trust-region methods on Riemannian manifolds. Found. Comput. Math. **7**(3), 303–330 (2007)
2. Absil, P.-A., Mahony, R., Sepulchre, R.: Optimization algorithms on matrix manifolds. Princeton University Press, (2008)
3. Agarwal, N., Boumal, N., Bullins, B., Cartis, C.: Adaptive regularization with cubics on manifolds. Mathematical Programming, Series A **188**(1), 85–134 (2021)
4. Birgin, E.G., Gardenghi, J., Martínez, J.M., Santos, S.A., Toint, P.L.: Worst-case evaluation complexity for unconstrained nonlinear optimization using high-order regularized models. Math. Program. **163**, 359–368 (2017)
5. Birgin, E.G., Krejić, N., Martínez, J.M.: Iteration and evaluation complexity for the minimization of functions whose computation is intrinsically inexact. Math. Comput. **89**, 253–278 (2020)
6. Boumal, N.: An introduction to optimization on smooth manifolds. Cambridge University Press, (2023)
7. Boumal, N., Mishra, B., Absil, P.-A., Sepulchre, R.: Manopt, a matlab toolbox for optimization on manifolds. J. Mach. Learn. Res. **15**, 1455–1459 (2014)
8. Boumal, N., Singer, A., Absil, P.-A.: Robust estimation of rotations from relative measurements by maximum likelihood. In IEEE 52nd Annual Conference on Decision and Control (CDC), pages 1156–1161, (2013)
9. Cartis, C., Gould, N., Toint, P.: Adaptive cubic regularisation methods for unconstrained optimization. part i: motivation, convergence and numerical results. Mathematical Programming, Series A, 127:245–295, (2011)
10. worst-case function and derivative evaluation complexity: C. Cartis, N. Gould, P. Toint. Adaptive cubic regularisation methods for unconstrained optimization. part ii. Mathematical Programming, Series A **130**, 295–319 (2011)
11. Cartis, C., Gould, N.I., Toint, P.L.: On the oracle complexity of first-order and derivative-free algorithms for smooth nonconvex minimization. SIAM J. Optim. **22**(1), 66–86 (2012)
12. Deng, Y., Mu, T.: Faster Riemannian Newton-type optimization by subsampling and cubic regularization. Mach. Learn. **112**, 3527–3589 (2023)
13. do Carmo, M.P.: Riemannian geometry. Mathematics: Theory & Applications. Birkhäuser Boston, Inc., Boston, MA, (1992). Translated from the second Portuguese edition by Francis Flaherty
14. Doikov, N., Grapiglia, G.N.: First and zeroth-order implementations of the regularized newton method with lazy approximated hessians. arXiv preprint arXiv:2309.02412, (2023)
15. Edelman, A., Arias, T.A., Smith, S.T.: The geometry of algorithms with orthogonality constraints. SIAM J. Matrix Anal. Appl. **20**(2), 303–353 (1999)
16. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572, (2014)
17. Grapiglia, G.N., Goncalves, M.L., Silva, G.: A cubic regularization of newtons method with finite difference hessian approximations. Numerical Algorithms **90**(2), 607–630 (2022)
18. Hand, P., Lee, C., Voroninski, V.: Shapefit: Exact location recovery from corrupted pairwise directions. Commun. Pure Appl. Math. **71**(1), 3–50 (2018)
19. Jaquier, N., Rozo, L., Calinon, S., Bürger, M.: Bayesian optimization meets riemannian manifolds in robot learning. In Conference on Robot Learning, pages 233–246. PMLR, (2020)
20. Kohler, J.M., Lucchi, A.: Sub-sampled cubic regularization for non-convex optimization. Proceedings of the 34th International Conference on Machine Learning, 70
21. Lee, J.M.: Introduction to Smooth Manifolds, volume 218 of Graduate Texts in Mathematics. Springer-Verlag, New York

22. Lee, J.M.: Riemannian manifolds: an introduction to curvature, volume 176. Springer Science & Business Media, (2006)
23. Li, J., Balasubramanian, K., Ma, S.: Stochastic zeroth-order riemannian derivative estimation and optimization. Math. Oper. Res. **48**(2), 1183–1211 (2023)
24. Nesterov, Y., Polyak, B.T.: Cubic regularization of Newton method and its global performance. Math. Program. **108**, 177–205 (2006)
25. Nocedal, J., Wright, S.J.: Numerical optimization. Springer, (1999)
26. Ramachandran, P., Zoph, B., Le, Q.V.: Searching for activation functions. arXiv:1710.05941, (2017)
27. Sato, H., Iwai, T.: A Riemannian optimization approach to the matrix singular value decomposition. SIAM J. Optim. **23**(1), 188–212 (2013)
28. Tripuraneni, N., Stern, M.,x Jin, M., Regier, J., Jordan, M.I.: Stochastic cubic regularization for fast nonconvex optimization. Advances in Neural Information Processing Systems, (2018)
29. Tu, C.-C., Ting, P., Chen, P.-Y., Liu, S., Zhang, H., Yi, J., Hsieh, C.-J., Cheng, S.-M.: Autozoom: Autoencoder-based zeroth order optimization method for attacking black-box neural networks. In Proceedings of the AAAI Conference on Artificial Intelligence **33**, 742–749 (2019)
30. Tu, L.W.: An Introduction to Manifolds. Universitext. Springer, New York, 2 edition
31. Wang, Z., Zhou, Y., Liang, Y., Lan, G.: A note on inexact gradient and Hessian conditions for cubic regularized Newton's method. Oper. Res. Lett. **47**(2), 146–149 (2019)
32. Weinberger, K.Q., Saul, L.K.: Unsupervised learning of image manifolds by semidefinite programming. Int. J. Comput. Vision **70**, 77–90 (2006)
33. Yang, W., Yang, Y., Zhang, C., Cao, M.: A Newton-like trust region method for large-scale unconstrained nonconvex minimization. Abstract and Applied Analysis, (2013)
34. Zhang, J., Zhang, S.: A cubic regularized Newton's method over Riemannian manifolds. arXiv:1805.05565