Riemannian Adaptive Regularized Newton Methods with Hölder Continuous Hessians

Chenyu Zhang[*] Rujun Jiang[†]

Abstract

This paper presents strong worst-case iteration and operation complexity guarantees for Riemannian adaptive regularized Newton methods, a unified framework encompassing both Riemannian adaptive regularization (RAR) methods and Riemannian trust region (RTR) methods. We comprehensively characterize the sources of approximation in second-order manifold optimization methods: the objective function's smoothness, retraction's smoothness, and subproblem solver's inexactness. Specifically, for a function with a μ-Hölder continuous Hessian, when equipped with a retraction featuring a ν-Hölder continuous differential and a θ-inexact subproblem solver, both RTR and RAR with $2+\alpha$ regularization (where $\alpha = \min\{\mu, \nu, \theta\}$) locate an $(\varepsilon, \varepsilon^{\alpha/(1+\alpha)})$-approximate second-order stationary point within at most $O(\varepsilon^{-(2+\alpha)/(1+\alpha)})$ iterations and at most $\widetilde{O}(\varepsilon^{-(4+3\alpha)/(2(1+\alpha))})$ Hessian-vector products with high probability. These complexity results are novel and sharp, and reduce to an iteration complexity of $O(\varepsilon^{-3/2})$ and an operation complexity of $\widetilde{O}(\varepsilon^{-7/4})$ when $\alpha = 1$.

1 Introduction

We consider an unconstrained nonconvex manifold optimization problem:

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

where $\mathcal{M}$ is a complete Riemannian manifold, and f is bounded below, and exhibits $C^{2,\mu}$ smoothness, meaning it is twice continuously differentiable and possesses a μ-order Hölder continuous Hessian. Considering the problem's nonconvexity, we propose employing Riemannian adaptive regularized Newton (RARN) methods that incorporate Riemannian trust region (RTR) [1, 35] and adaptive regularization (RAR) methods [4, 27]. Within each iteration of RARN, we solve a *regularized* model problem *inexactly* on a tangent space:

$$\eta_k \approx \operatorname*{argmin}_{\eta\in T_x\mathcal{M}} \bar{m}_{x_k}(\eta) := m_{x_k}(\eta) + \varphi(\eta; \sigma_k),$$

where m_{x_k} is the Newton model function, i.e., the truncated second-order Taylor expansion of the objective function at $x_k \in \mathcal{M}$, and φ is the regularization, with σ_k serving as an adaptively changing regularization parameter. Subsequently, if the solution yields a significant decrease in the objective function, we pull it back onto the manifold using a *retraction* as the next iteration point:

$$x_{k+1} = R_{x_k}(\eta_k).$$

In an Euclidean space, the canonical retraction is the identity function, simplifying the iteration to $x_{k+1} = x_k + \eta_k$. To align with the smoothness characteristics of the objective function, we employ

[*] Data Science Institute, Columbia University, New York, USA. (chenyu.zhang@columbia.edu).

[†] School of Data Science, Fudan University, Shanghai, China. (rjjiang@fudan.edu.cn).

θ-inexact subproblem solvers and $C^{1,\nu}$ retractions. Detailed definitions and discussions of these elements are deferred to Sections 2 and 3.

Motivated by recent advances in worst-case iteration complexity guarantees for Newton-type methods [25, 11, 29, 17], this paper investigates the iteration complexity of RTR and RAR under the aforementioned relaxations. Iteration complexity refers to a *non-asymptotic* bound on the number of outer iterations an algorithm needs to attain a solution within a given tolerance. Concurrently, operation complexity, which bounds the number of *unit operations* required to find an approximated solution, has emerged as another focal non-asymptotic convergence measure due to its direct correspondence with computational cost [17, 4, 8]. We also provide strong operation complexity guarantees, measured in terms of the number of Hessian-vector products, for both RTR and RAR.

1.1 Main Results

Our main results and contributions unfold across several dimensions.

A unified algorithm and iteration complexity analysis framework. We begin by formulating a unified Riemannian adaptive regularized Newton (RARN) method framework, encompassing both Riemannian trust region (RTR) and adaptive regularization (RAR) methods. Within this framework, we develop a unified iteration complexity analysis approach for RARN. This versatile module allows us to derive iteration complexity guarantees for various regularization methods effortlessly.

A comprehensive approximation characterization and relaxation. We identify the three sources of approximation within second-order manifold optimization methods: 1) The quadratic model function approximates the actual objective function. 2) The retraction approximates the smooth diffeomorphism between the tangent space and the manifold; 3) The inexact model problem solution approximates the exact solution. These approximation levels can be characterized by the objective function's smoothness, retraction's smoothness, and subproblem solver's exactness. Our investigation particularly explores objective functions with a μ-Hölder continuous Hessian, retractions with a ν-Hölder continuous differential, and θ-inexact subproblem solvers. Significantly, our analysis unveils an equitable contribution from these three sources of approximation to the iteration complexity: the iteration complexity order can only be improved by enhancing the worst-performing source.

A general-order adaptive regularization method. Our work extends the realm of adaptive regularization methods by introducing a novel adaptive $2+\omega$ regularization method ($\omega \in (0,1]$), broadening the horizons beyond cubics. We pave the way for adaptivity without the constraint of $\mu \geqslant \omega$, a restriction posed by prior work on general-order adaptive regularization [5, 15, 13]. Nevertheless, our findings reveal that optimal iteration complexity is achieved if and only if $\omega = \alpha := \min\{\mu, \nu, \theta\}$.

A sharp worst-case iteration and operation complexity. Finally, we present our complexity guarantees. When $\omega = \alpha$, both RTR and RAR locate an $(\varepsilon, \varepsilon^{\alpha/(1+\alpha)})$-approximate second-order stationary point (see Definition 1) within at most

$$O\left(\varepsilon^{-\frac{2+\alpha}{1+\alpha}}\right)$$

iterations. This represents the first second-order iteration complexity result of Newton-type methods for objective functions with a Hölder continuous Hessian, even within the framework of Euclidean spaces. This complexity matches the optimal bound of $O(\varepsilon^{-(2+\mu)/(1+\mu)})$ for this class of problems [13], as well as the optimal bound of $O(\varepsilon^{-3/2})$ for $C^{2,1}$ problems [10] when $\mu = \nu = \theta = 1$. Moreover, when $\alpha = 1$, this result establishes the first optimal iteration complexity result of $O(\varepsilon^{-3/2})$ for Riemannian trust region methods targeting both first-order and second-order stationarity. Augmented by a meticulous analysis of the subproblem solvers, we also provide a Hessian-vector product operation complexity of $\widetilde{O}(\varepsilon^{-(4+3\alpha)/(2(1+\alpha))})$, which reduces to $\widetilde{O}(\varepsilon^{-7/4})$ when $\alpha = 1$. Table 1 compares our results with other related work and highlights the gaps our study addresses.

1.2 Related Work

Table 1 Comparison of worst-case iteration complexity results for adaptive regularized Newton methods, aiming to achieve ε-approximate first-order stationarity or $(\varepsilon, \varepsilon^{\mu/(1+\mu)})$-approximate second-order stationarity. Asymptotic notation has been omitted from the complexity results.

Work	Space	Newton Extension	Stationarity	Smoothness	Iteration Complexity
[20]	Euclidean	line search	first order	$C^{2,\mu}$	$\varepsilon^{-(2+\mu)/(1+\mu)}$
[29]	Euclidean	line search	second order	$C^{2,1}$	$\varepsilon^{-3/2}$
[17]	Euclidean	trust region	second order	$C^{2,1}$	$\varepsilon^{-3/2}$
[25] and [11]	Euclidean	cubic regularization	second order	$C^{2,1}$	$\varepsilon^{-3/2}$
[7]	Riemannian	trust region	second order	$C^{2,1}$	$\varepsilon^{-5/2}$
Our work (Corollary 3)	Riemannian	trust region	second order	$C^{2,1}$	$\varepsilon^{-3/2}$
[37] and [4]	Riemannian	cubic regularization	second order	$C^{2,1}$	$\varepsilon^{-3/2}$
Our work (Corollary 2)	Riemannian	$2+\mu$ regularization	second order	$C^{2,\mu}$	$\varepsilon^{-(2+\mu)/(1+\mu)}$
Our work (Corollary 3)	Riemannian	trust region	second order	$C^{2,\mu}$	$\varepsilon^{-(2+\mu)/(1+\mu)}$

Absil et al. [1] extended trust region methods to Riemannian manifolds with asymptotic convergence results for $C^{2,1}$ objective functions. Qi [27] extended adaptive regularization with cubics to Riemannian manifolds with asymptotic local convergence results. This paper focuses on non-asymptotic convergence results of RTR and RAR. We compare the related worst-case iteration complexity findings of adaptive regularized Newton methods in Table 1. For simplicity, Table 1 solely lists the smoothness requirement on the objective function. Note that our approach also introduces a more lenient smoothness requirement on the retraction ($R \in C^{1,\nu}$), distinguishing it from prevailing research on Riemannian methods where retractions are typically C^2 [7, 37, 4].

Table 1 does not aim to encompass the entire research on the iteration complexity of Newton-type methods. Given our specific focus on establishing complexity guarantees under relaxed smoothness requirements within the Riemannian setting, we only list the foundational works that initially established these complexities. For more results along this line, please refer to [9, 8, 34, 33]. In addition to the research summarized in Table 1, it is noteworthy to mention related studies with distinct emphases. Cartis et al. [15] also considered a μ-Hölder continuous Hessian for a conceptual regularization algorithm, presenting a complexity of $O(\varepsilon_H^{-(2+\mu)/\mu})$ for the attainment of an ε_H-approximate second-order stationary point, aligning with our complexity (see Corollary 2). Dedicated to functions with a μ-Hölder continuous Hessian, [13] proved that the complexity of finding an ε-approximate first-order stationary point is lower bounded by $O(\varepsilon^{-(2+\mu)/(1+\mu)})$, a bound which our methods successfully attain.

Beyond iteration complexity, recent research has introduced second-order methods capable of achieving a remarkable *operation complexity* of $\widetilde{O}(\varepsilon^{-7/4})$ with high probability, for finding an $(\varepsilon, \varepsilon^{1/2})$-approximate stationary point. Building upon adaptive regularization with cubics, Agarwal et al. [3] derived an algorithm that attains this operation complexity bound, employing a subproblem solver with $\widetilde{O}(\varepsilon^{-1/4})$ operation complexity. Subsequently, various methods have emerged that replicate the same complexity results [28, 29, 22]. Very recently, Curtis et al. [17] demonstrated that such complexity can be achieved through a variant of trust region Newton methods that inexact solve the trust region subproblem, utilizing the truncated conjugate gradient method [30, 31]. Carmon et al. [9] and Li and Lin [23] proposed variants of accelerated gradient methods that also converge to an $(\varepsilon, \varepsilon^{1/2})$-approximate stationary point with an operation complexity of $\widetilde{O}(\varepsilon^{-7/4})$.

Notation For scalars x and y, we write $[x]_+ \coloneqq \max\{0, x\}$, $x \vee y \coloneqq \max\{x, y\}$, and $x \wedge y \coloneqq \min\{x, y\}$. We write $x \gtrsim y$ if there exists a positive constant C such that $Cx \geqslant y$, and write $x \lesssim y$ if $y \gtrsim x$.

For sets $\mathcal{A}$ and $\mathcal{B}$, we denote their cardinality by $|\mathcal{A}|$ and $|\mathcal{B}|$, and their disjoint union by $\mathcal{A} \sqcup \mathcal{B}$. We sometimes write f_k, g_k, and H_k as shorthand for the function value, gradient, and Hessian, respectively, of f at x_k. We also denote f_0 as the objective function evaluated at the initial point and $f_{\min}$ as the objective function's lower bound on the manifold. Throughout our analysis, for any subscript, C_* represents absolute positive constants independent of ε_g or ε_H. In algorithm descriptions, (TC.X&Y) indicates the requirement for simultaneously satisfying both termination criteria (TC.X) and (TC.Y), while (TC.X/Y) implies that fulfilling either termination criterion (TC.X) or (TC.Y) is sufficient. We use (C.x) to tag different cases in the analysis for clarity.

2 Preliminaries

We consider an unconstrained nonconvex optimization problem on a Riemannian manifold $\mathcal{M}$, which is a smooth manifold equipped with an inner product $\langle \cdot, \cdot \rangle_x$ on each tangent space $T_x\mathcal{M}$, $x \in \mathcal{M}$, varying smoothly over $\mathcal{M}$. The Riemannian inner product further defines a norm $\|\cdot\|_x$ on $T_x\mathcal{M}$ and a distance $\operatorname{dist}(\cdot,\cdot)$ on $\mathcal{M}$. Given our focus on the Riemannian metric, we will omit its subscript x throughout the remainder of this paper. The Riemannian inner product also helps define the gradient and Hessian of a function f from $\mathcal{M}$ to $\mathbb{R}$:

$$\langle \operatorname{grad} f(x), \xi \rangle = \mathrm{d}f(x)[\xi], \quad \operatorname{Hess} f(x)[\xi] = \nabla_\xi \operatorname{grad} f(x), \quad \forall \xi \in T_x\mathcal{M},$$

where $\mathrm{d}f(x)$ is the differential of f at x, and ∇ is the Riemannian (Levi-Civita) connection on $\mathcal{M}$.

Throughout this paper, we consistently denote a parametrized smooth curve from $[0,1]$ to $\mathcal{M}$ as γ. A vector field X is said to be parallel along γ if $\nabla_{\gamma'(t)} X = 0$ for all $t \in [0,1]$. For any $\xi \in T_{\gamma(0)}\mathcal{M}$, there exists a unique parallel vector field X_ξ along γ such that $X_\xi(0) = \xi$, defining a parallel transport operator $P_\gamma^{0\to t} : \xi \mapsto X_\xi(t)$. Parallel transport is an isometry that preserves the inner product, bridging different tangent spaces. Our focus is specifically on complete Riemannian manifolds, where geodesics—smooth curves whose tangent vector is parallel along itself—are the shortest curves connecting any two points and can be extended to the entire real axis. Consequently, the exponential map $\exp_x : \xi \mapsto \gamma(1)$, where γ is the unique geodesic with $\gamma(0) = x$, $\gamma'(0) = \xi$, and $\operatorname{dist}(\gamma(0), \gamma(1)) = \|\xi\|$, is defined on the entire tangent space. When the parallel transport is along a geodesic from x to y, we denote $P_{xy} := P_\gamma^{0\to 1}$. Detailed preliminaries for Riemannian optimization can be found in monographs such as [2] and [6].

We now define the approximate second-order stationarity, which serves as the objective of our methods.

Definition 1 (Approximate second-order stationary point)**.** We say $x \in \mathcal{M}$ is an $(\varepsilon_g, \varepsilon_H)$-approximate second-order stationary point of f if

$$\|\operatorname{grad} f(x)\| \leqslant \varepsilon_g, \quad \lambda_{\min}(\operatorname{Hess} f(x)) \geqslant -\varepsilon_H,$$

where $\lambda_{\min}$ returns the smallest eigenvalue of a linear operator on $T_x\mathcal{M}$.

Our exploration in this work centers on objective functions with a Hölder continuous Hessian.

Definition 2 (Hölder continuity of objective's Hessian)**.** We say f's Hessian is Hölder continuous with order $\mu \in (0,1]$, if there exist a constant $C_H \geqslant 0$, such that for any $x, y \in \mathcal{M}$, it holds that

$$\|P_{yx} \operatorname{Hess} f(y) P_{xy} - \operatorname{Hess} f(x)\|_{\mathrm{op}} \leqslant C_H \operatorname{dist}(x,y)^\mu,$$

where $\|\cdot\|_{\mathrm{op}}$ is the operator norm of the linear operators on $T_x\mathcal{M}$. We denote $f \in C^{2,\mu}$ if $f \in C^2$ and $\operatorname{Hess} f$ is μ-order Hölder continuous.

To pull a point on the tangent space back onto the manifold, we use retractions. Specifically, a map $R : T\mathcal{M} \to \mathcal{M}$ is called a retraction if its restriction to any $x \in \mathcal{M}$, denoted as R_x, satisfies conditions $R_x(0) = x$ and $\mathrm{d}R_x(0) = \mathrm{id}(T_x\mathcal{M})$. The exponential map is a special retraction, and retractions can be viewed as a first-order approximation of the exponential map. This approximation bias becomes more significant for nonsmooth retractions. Therefore, prior work often necessitates retractions to be C^2 or even *second-order* retractions [1, 4], whose *acceleration* at the origin is zero: $\mathrm{d}^2R_x(0_x) = 0$ (see [2] or [6]). In contrast, this paper only mandates that retractions' differential be Hölder continuous at the origin.

Definition 3 (Hölder continuity of retraction's differential)**.** We say R_x's differential is Hölder continuous at origin with order $\nu \in (0, 1]$, if there exist a constant $C_R \geqslant 0$, such that for any $x \in \mathcal{M}$ and $\xi \in T_x\mathcal{M}$, it holds that

$$\|P_{yx}\mathrm{d}R_x(\xi)P_{xy} - \mathrm{d}R_x(0_x)\|_{\mathrm{op}} \leqslant C_R\|\xi\|^{\nu},$$

where $x = R_x(0_x)$ and $y \coloneqq R_x(\xi)$. We denote $R \in C^{1,\nu}$ if $R \in C^1$ and $\mathrm{d}R$ is ν-order Hölder continuous.

The following proposition delineates the accuracy of retractions concerning their approximation to the exponential map.

Proposition 1 (Retraction properties)**.** *Suppose R_x has a Hölder continuous differential with order $\nu \in (0, 1]$ and constant C_R. For any $x \in \mathcal{M}$ and $\eta \in T_x\mathcal{M}$, it holds that*

$$\mathrm{dist}(R_x(\eta), \exp_x(\eta)) \leqslant C_R\|\eta\|^{1+\nu},$$

and

$$\|\eta - \exp_x^{-1}(R_x(\eta))\| \leqslant C_R\|\eta\|^{1+\nu}.$$

Moreover, if the operator norm of Hess f *is upper bounded by β_H, then the discrepancy between their composition with the objective function is bounded by*

$$|f(R_x(\eta)) - f(\exp_x(\eta))| \leqslant ((1 + C_R\|\eta\|^{\nu}) \cdot \beta_H\|\eta\| + \|\operatorname{grad} f(x)\|) \cdot C_R\|\eta\|^{1+\nu}.$$

The third inequality presented in Proposition 1 introduces a dependence on grad f, a distinctive characteristic of the Riemannian setting involving non-second-order retractions. As we progress through subsequent sections, it will become evident that this reliance on the gradient holds significant implications for the dynamics of the regularization parameter. Its careful analysis is essential for deriving the bounds of the regularization parameter.

We conclude this section with a blanket assumption upheld throughout the paper, serving as the foundation for all subsequent results.

Assumption 1. The objective function f is bounded below on the complete Riemannian manifold $\mathcal{M}$. The approximation tolerance in Definition 1 satisfies $\varepsilon_g \vee \varepsilon_H \leqslant 1$. The objective function possesses a Hölder continuous Hessian with order $\mu \in (0, 1]$ and constant C_H. The retraction has a Hölder continuous differential with order $\nu \in (0, 1]$ and constant C_R. Finally, the Hessian has a uniformly bounded operator norm on the level set $\mathcal{M}_{f_0} \coloneqq \{x \in \mathcal{M} : f(x) \leqslant f_0\}$, i.e., there exists $\beta_H > 0$ such that $\|\operatorname{Hess} f\|_{\mathrm{op}} \leqslant \beta_H$ on $\mathcal{M}_{f_0}$.

All conditions in Assumption 1 can be relaxed to apply exclusively along the trajectory of the algorithm. Moreover, we impose separate conditions on the objective function and the retraction we use, which is more pragmatic and versatile compared to the assumption on their composition $f \circ R$ found in existing literature [7, 4].

3 Riemannian Adaptive Regularized Newton Methods

Line search, trust region, and higher-order regularization methods are popular extensions of Newton methods for unconstrained nonconvex optimization. These diverse methods and their variants can be conceptualized as applying various forms of regularization to the Newton model function, i.e., the truncated second-order Taylor expansion of the objective function:

$$m_x(\eta) \coloneqq f(x) + \langle \eta, \operatorname{grad} f(x) \rangle + \frac{1}{2} \langle \eta, \operatorname{Hess} f(x)[\eta] \rangle . \tag{1}$$

Here, we present a unified Riemannian adaptive regularized Newton (RARN) method that considers the following regularized model problem on a tangent space:

$$\min_{\eta \in T_x \mathcal{M}} \bar{m}_x(\eta; \sigma) \coloneqq m_x(\eta) + \varphi(\eta; \sigma), \tag{2}$$

where φ is the regularization and σ is the adaptive regularization parameter that are dynamically adjusted throughout the iterations. We impose $\varphi(0; \sigma_k) = 0$. This scheme encompasses adaptive versions of the previously mentioned methods:

- For trust region methods, $\varphi(\eta; \Delta) = \delta_{B(0,\Delta)}(\eta)$, where δ is the indicator function and $B(0, \Delta)$ is the trust region.
- For adaptive p-order regularization, $\varphi(\eta, \sigma) = \frac{1}{p}\sigma \|\eta\|^p$.

At each iteration of RARN, we (approximately) solve the regularized model problem (2), and then decide whether to accept the candidate iteration point or not and update the regularization parameter, according to a comparison between the objective decrease and the (unregularized) model decrease. More precisely, after the subproblem solver returns η_k, we compute the relative decrease ratio:

$$\rho_k = \frac{f(x_k) - f(R_{x_k}(\eta_k))}{m_{x_k}(0) - m_{x_k}(\eta_k)}. \tag{3}$$

One can also compute the ratio over the regularized model decrease $\bar{m}_{x_k}(0) - \bar{m}_{x_k}(\eta_k)$, while still maintaining the validity of the complexity guarantees presented in this paper, albeit with minor adjustments in the technical details. We opt for the unregularized model decrease because it is consistent across all RARN variants. We provide the RARN framework for achieving an approximate second-order stationary point in Algorithm 1.

Within the uniform algorithm framework, we also introduce termination criteria that offer a greater degree of inexactness and flexibility compared to prior work [11, 4, 17, 33]. First of all, for all iterations, we require the candidate point to outperform the Cauchy point, which is the solution to (2) restricted along the single dimension spanned by g_k:

$$\eta_k^{\mathrm{C}} = \operatorname*{argmin}_{\eta \in \operatorname{span}\{g_k\}} \bar{m}_{x_k}(\eta).$$

This leads to the Cauchy termination criterion:

$$\bar{m}_{x_k}(\eta_k) \leqslant \bar{m}_{x_k}(\eta_k^{\mathrm{C}}). \tag{TC.C}$$

If a candidate point satisfies (TC.C), by the definition of the regularized model problem (2), we have

$$m_{x_k}(0) - m_{x_k}(\eta_k) = \bar{m}_{x_k}(0) - \bar{m}_{x_k}(\eta_k) + \varphi(\eta_k; \sigma_k) \geqslant \varphi(\eta_k; \sigma_k). \tag{4}$$

If the first-order test fails, i.e., $\|g_k\| > \varepsilon_g$, the subproblem solver needs to minimize the residual $\operatorname{grad} \bar{m}(\eta)$. Thus, we employ the following first-order termination criterion:

$$\| \operatorname{grad} \bar{m}_{x_k}(\eta_k)\| \leqslant \|\eta_k\|^{1+\theta_1}, \tag{TC.1}$$

Algorithm 1: Riemannian Adaptive Regularized Newton Method

```
parameters tolorance $\varepsilon_g, \varepsilon_H \in (0,1)$, adaptation parameters, control parameters
input initial point $x_0 \in \mathcal{M}$.
for $k = 0, 1, \ldots$ do
    if $\|\operatorname{grad} f(x_k)\| > \varepsilon_g$ then                // first-order stationarity test
        obtain $\eta_k \approx \arg\min_{\eta \in T_x\mathcal{M}} \bar{m}_{x_k}(\eta; \sigma_k)$ with termination criterion (TC.1&C)
    else if $\operatorname{Hess} f(x_k) \not\succeq -\varepsilon_H I$ then              // second-order stationarity test
        obtain $\eta_k \approx \arg\min_{\eta \in T_x\mathcal{M}} \bar{m}_{x_k}(\eta; \sigma_k)$ with termination criterion (TC.2&C)
    else
        return $x_k$
    end
    compute $\rho_k$ using (3)
    shrink or expand $\sigma_{k+1}$ according to $\rho_k$
    accept $x_{k+1} = R_{x_k}(\eta_k)$ or not according to $\rho_k$
end
```

where $\theta_1 \in (0,1]$ is the parameter governing the degree of inexactness of the subproblem solution. We remark that due to the non-differentiability of the indicator function, we incorporate the trust region as a problem constraint rather than integrating it into the regularized model function $\bar{m}$. We defer the comprehensive definition of the model problem of RTR to Section 5.

If the second-order stationarity is targeted and the second-order test fails, i.e., $H_k \not\succeq -\varepsilon_H I$, the subproblem solver should aim for the following criterion:

$$\operatorname{Hess} \bar{m}_{x_k}(\eta_k) \succeq -\|\eta_k\|^{\theta_2} \cdot I, \tag{TC.2}$$

where $\theta_2 \in (0,1]$ is an inexactness parameter. For simplicity, we use the same parameter $\theta = \theta_1 = \theta_2$ to control the subproblem solver's inexactness in our analysis. The above termination conditions only serve as basic criteria for our unified framework. In specific methods, the subproblem solver can incorporate alternative suitable termination conditions, such as truncation conditions for a trust region method. These conditions also directly establish lower bounds on the step-size $\|\eta_k\|$, subsequently forming lower bounds for the successful objective decrease, as will be demonstrated in the forthcoming sections.

Proposition 2 (Termination criteria)**.** 1) *For η_k satisfying (TC.1), if kth iteration is successful and $\|\eta_k\| \leqslant 1$, we have*

$$\|g_{k+1}\| \leqslant (2C_H(C_R+1) + \beta_H C_R + 1)\|\eta_k\|^{1+\mu\wedge\nu\wedge\theta} + \|\operatorname{grad} \varphi(\eta_k; \sigma_k)\|.$$

2) *For η_k satisfying (TC.2), we have*

$$-\lambda_{\min}(H_k) \leqslant \|\eta_k\|^{\theta} + \lambda_{\max}(\operatorname{Hess} \varphi(\eta_k; \sigma_k)).$$

Readers may notice that the equivalence of μ, ν, and θ has emerged in the first inequality of Proposition 2.

While Algorithm 1 provides a high-level framework, we will specify the subproblem solver, additional termination criteria, regularization parameter adaptation, and acceptance criteria in dedicated sections for trust region and adaptive regularization methods.

3.1 Iteration Complexity Analysis Framework

We now give a unified analysis framework for the iteration complexity of RARN. A key observation is that, in RARN, the adaptation of the regularization parameter is closely related to the acceptance of a candidate step. We categorize the kth iteration as either "successful" or "unsuccessful" based on whether $R_{x_k}(\eta_k)$ is accepted as x_{k+1}. Formally, we define the following index sets:

$$\mathcal{K} \coloneqq \{k \in \mathbb{N} : \text{the algorithm is not terminated within } k\text{th iteration}\},$$
$$\mathcal{S} \coloneqq \{k \in \mathcal{K} : k\text{th iteration is successful}\}, \quad \text{and} \quad \mathcal{U} \coloneqq \{k \in \mathcal{K} : k\text{th iteration is unsuccessful}\}.$$

Because $|\mathcal{K}| = |\mathcal{S} \sqcup \mathcal{U}| = |\mathcal{S}| + |\mathcal{U}|$, we can establish a connection between the total number of iterations and the range of values for the regularization parameter. We validate this observation in the following lemma, an adaptation from [11, Theorem 2.1].

Lemma 1 (Decomposition of total number of iterations)**.** *Suppose that Algorithm 1 prescribes the following conditions for the regularization parameter: σ_k only shrinks for $k \in \mathcal{S}$, expands for any $k \in \mathcal{U}$, and the shrinkage is explicitly lower bounded by $\sigma_{k+1} \geqslant \max\{\underline{\kappa}\sigma_k, \underline{\sigma}\}$ for some $\underline{\kappa} < 1$, while the expansion is upper bounded by $\sigma_{k+1} \leqslant \bar{\kappa}\sigma_k$ for some $\bar{\kappa} > 1$. Additionally, if* 1) *$\mathcal{S}$ is finite and* 2) *a upper bound $\bar{\sigma} < +\infty$ of σ_k exists, then*

$$|\mathcal{K}| \leqslant \left(1 - \frac{\log \underline{\kappa}}{\log \bar{\kappa}}\right) |\mathcal{S}| + \log_{\bar{\kappa}} \frac{\bar{\sigma}}{\underline{\sigma}}.$$

Similarly, suppose Algorithm 1 specifies that σ_k only expands for $k \in \mathcal{S}$, shrinks for any $k \in \mathcal{U}$, and the expansion is explicitly upper bounded by $\sigma_{k+1} \leqslant \min\{\bar{\kappa}\sigma_k, \bar{\sigma}\}$ for some $\bar{\kappa} > 1$, while the shrinkage is lower bounded by $\sigma_{k+1} \geqslant \underline{\kappa}\sigma_k$ for some $\underline{\kappa} < 1$. Additionally, if 1) *$\mathcal{S}$ is finite and* 2) *a lower bound $\underline{\sigma} > 0$ of σ_k exists, then*

$$|\mathcal{K}| \leqslant \left(1 - \frac{\log \bar{\kappa}}{\log \underline{\kappa}}\right) |\mathcal{S}| + \log_{\underline{\kappa}^{-1}} \frac{\bar{\sigma}}{\underline{\sigma}}.$$

Proof. We prove the first bound; the second bound can be derived similarly. For any $k \in \mathbb{N}$, by the assumptions, we have

$$\bar{\sigma} \geqslant \sigma_k \geqslant \sigma_0 \cdot \underline{\kappa}^{|\mathcal{S} \cap [k]|} \cdot \bar{\kappa}^{|\mathcal{U} \cap [k]|} \geqslant \underline{\sigma} \cdot \underline{\kappa}^{|\mathcal{S} \cap [k]|} \cdot \bar{\kappa}^{|\mathcal{U} \cap [k]|}.$$

Taking the logarithm of both sides and rearranging the above inequality gives

$$|\mathcal{U} \cap [k]| \leqslant -\frac{\log \underline{\kappa}}{\log \bar{\kappa}} \cdot |\mathcal{S} \cap [k]| + \log_{\bar{\kappa}} \frac{\bar{\sigma}}{\underline{\sigma}} \leqslant -\frac{\log \underline{\kappa}}{\log \bar{\kappa}} \cdot |\mathcal{S}| + \log_{\bar{\kappa}} \frac{\bar{\sigma}}{\underline{\sigma}},$$

We get the finite bound of $|\mathcal{U}|$ by letting $k \to \infty$, and then the result follows. □

Remark 1. Lemma 1 directly gives a nonasymptotic bound: $|\mathcal{K}| = O\left(|\mathcal{S}| + \log(\bar{\sigma}/\underline{\sigma})\right)$.

To explicitly bound $|\mathcal{K}|$, we still need to validate the two assumptions in Lemma 1: providing a bound of $|\mathcal{S}|$ and an upper/lower bound of σ_k. For the bound of $|\mathcal{S}|$, if we can establish a minimal decrease in the objective function during successful iterations, then since f is bounded below, the number of successful iterations can be bounded.

To this end, we further partition $\mathcal{S}$ into the following index sets

$$\mathcal{S}_L \coloneqq \{k \in \mathcal{S} : \|\operatorname{grad} f(x_k)\| \leqslant \varepsilon_g\}, \quad \mathcal{S}_G \coloneqq \{k \in \mathcal{S} : \|\operatorname{grad} f(x_k)\| > \varepsilon_g\},$$
$$\mathcal{S}_{GL} \coloneqq \{k \in \mathcal{S} : \|\operatorname{grad} f(x_k)\| > \varepsilon_g \text{ and } \|\operatorname{grad} f(x_{k+1})\| \leqslant \varepsilon_g\}, \quad \text{and}$$
$$\mathcal{S}_{GG} \coloneqq \{k \in \mathcal{S} : \|\operatorname{grad} f(x_k)\| > \varepsilon_g \text{ and } \|\operatorname{grad} f(x_{k+1})\| > \varepsilon_g\}.$$

We list some observations concerning the above index sets. First, $\mathcal{S} = \mathcal{S}_L \sqcup \mathcal{S}_G$ and $\mathcal{S}_G = \mathcal{S}_{GL} \sqcup \mathcal{S}_{GG}$. Moreover, if only the first-order stationarity is targeted, $|\mathcal{S}| = |\mathcal{S}_{GG}| + 1$. Also, for $k \in \mathcal{S}_{GL}$, any immediate successful iteration following k (if exists) must fall into the index set $\mathcal{S}_L$, implying that $|\mathcal{S}_{GL}| \leqslant |\mathcal{S}_L| + 1$. Furthermore, η_k is returned by Line 5 of Algorithm 1 if $k \in \mathcal{S}_G$ and by Line 7 if $k \in \mathcal{S}_L$.

Lemma 2 (Number of successful iterations)**.** *For $k \in \mathcal{S}$, if Lines 5 and 7 in Algorithm 1 produce a minimal decrease of ε_1 and ε_2 in f respectively, we have*

$$|\mathcal{S}| \leqslant 1 + \begin{cases} (f_0 - f_{\min}) \cdot \varepsilon_1^{-1}, & \text{for first-order stationarity,} \\ 2(f_0 - f_{\min}) \cdot (\varepsilon_1 \wedge \varepsilon_2)^{-1}, & \text{for second-order stationarity.} \end{cases}$$

Proof. Since the objective function value only changes when the iteration is successful, we have

$$\begin{aligned} f_0 - f_{\min} &\geqslant \sum_{k\in\mathcal{S}} (f_k - f_{k+1}) \geqslant \sum_{k\in\mathcal{S}_{GG}} (f_k - f_{k+1}) + \sum_{k\in\mathcal{S}_L} (f_k - f_{k+1}) \\ &\geqslant |\mathcal{S}_{GG}| \cdot \varepsilon_1 + |\mathcal{S}_L| \cdot \varepsilon_2 \geqslant (|\mathcal{S}_{GG}| + |\mathcal{S}_L|) \cdot (\varepsilon_1 \wedge \varepsilon_2). \end{aligned}$$

For first-order stationarity, $|\mathcal{S}_L| = 0$. Thus, we have

$$|\mathcal{S}| \leqslant |\mathcal{S}_{GG}| + 1 \leqslant 1 + (f_0 - f_{\min}) \cdot \varepsilon_1^{-1}.$$

For second-order stationarity, any k in $\mathcal{S}_{GL}$ is followed by a successful iteration in $\mathcal{S}_L$. Thus, $|\mathcal{S}_{GL}| \leqslant |\mathcal{S}_L|$. Then, we have

$$|\mathcal{S}| \leqslant |\mathcal{S}_{GG}| + 2|\mathcal{S}_L| + 1 \leqslant 1 + 2(f_0 - f_{\min}) \cdot (\varepsilon_1 \wedge \varepsilon_2)^{-1}.$$

□

By Lemmas 1 and 2, we get an iteration complexity guarantee once we determine the minimal successful decreases ε_1 and ε_2, as well as the regularization parameter bound $\bar{\sigma}/\underline{\sigma}$ for specific algorithms. In the following sections, our objective is to concretize these values.

Corollary 1 (Total number of iterations)**.** *If the conditions in Lemmas 1 and 2 are satisfied, we have*

$$|\mathcal{K}| \leqslant \begin{cases} O(\varepsilon_1^{-1} + \log \bar{\sigma}/\underline{\sigma}), & \text{for first-order stationarity,} \\ O((\varepsilon_1 \wedge \varepsilon_2)^{-1} + \log \bar{\sigma}/\underline{\sigma}), & \text{for second-order stationarity.} \end{cases}$$

4 Riemannian Adaptive 2+α Regularization

Adaptive regularization with cubics (ARC) [12, 4] adds a cubic regularization term to the vanilla quadratic model function (1). However, as we will shortly elucidate, for $C^{2,\mu}$ objective functions and $C^{1,\nu}$ retractions, cubics *under-regularize*. Therefore, for the time being, we first consider a general Riemannian adaptive $2{+}\omega$ regularization (RAR), $\omega \in (0, 1]$. Specifically, the Riemannian $2{+}\omega$ regularized model problem at $x_k \in \mathcal{M}$ reads:

$$\min_{\eta\in T_{x_k}\mathcal{M}} \quad \bar{m}^{\text{ar}}_{x_k}(\eta; \sigma_k) \coloneqq f(x_k) + \langle \eta, \operatorname{grad} f(x_k)\rangle + \frac{1}{2}\langle \eta, \operatorname{Hess} f(x_k)[\eta]\rangle + \frac{\sigma_k}{2+\omega}\|\eta\|^{2+\omega}. \tag{5}$$

We inherit the method framework in Algorithm 1 and propose employing a Lanczos-based Krylov subspace method [12, 4, 8] as the subproblem solver with termination criteria (TC.C), (TC.1), and (TC.2). The method is presented in Algorithm 2. Krylov subspace methods enjoy the following property.

Proposition 3 (Gradient bound by step size)**.** *η_k returned by a Krylov subspace method satisfies*

$$\|g_k\| \leqslant \beta_H\|\eta_k\| + \|\operatorname{grad} \varphi(\eta_k; \sigma_k)\|.$$

Proposition 3 also applies to trust region methods, where φ is the trust region regularization, when η_k resides within the interior of the trust region, i.e., $\|\eta_k\|$ is strictly smaller than the trust region radius.

Algorithm 2: Riemannian Adaptive 2+ω Regularization

```
parameters tolorance $\varepsilon_g, \varepsilon_H \in (0,1)$, regularization adaptation parameters
    $\kappa_3 > \kappa_2 \geqslant 1 > \kappa_1 > 0$, $\underline{\sigma} > 0, \sigma_0 \in [\underline{\sigma}, \infty)$; control parameters $1 > \varrho_2 \geqslant \varrho_1 > 0$
input initial point $x_0 \in \mathcal{M}$.
for $k = 0, 1, \ldots$ do
    if $\|g_k\| > \varepsilon_g$ then
        obtain $\eta_k$ by solving problem (5) with termination criteria (TC.1&C)
    else if $H_k \not\succeq -\varepsilon_H I$ then
        obtain $\eta_k$ by solving problem (5) with termination criteria (TC.2&C)
    else
        return $x_k$
    end
    compute $\rho_k$ using (3)
    set $\sigma_{k+1} = \begin{cases} [\max\{\underline{\sigma}, \kappa_1\sigma_k\}, \sigma_k], & \rho_k > \varrho_2, & \text{// very successful} \\ [\sigma_k, \kappa_2\sigma_k], & \varrho_1 \leqslant \rho \leqslant \varrho_2, & \text{// successful} \\ [\kappa_2\sigma_k, \kappa_3\sigma_k], & \text{o.w.} & \text{// unsuccessful} \end{cases}$
    set $x_{k+1} = \begin{cases} R_{x_k}(\eta_k), & \rho_k \geqslant \varrho_1, \\ x_k & \text{o.w.} \end{cases}$
end
```

4.1 Iteration Complexity of RAR

As discussed in Section 3.1, obtaining an iteration complexity bound of RAR only requires establishing an upper bound of σ_k and the minimal decrease in f for $k \in \mathcal{S}$. When using a second-order retraction and $\omega = \mu$, one can readily derive an upper bound of σ_k, akin to the Euclidean case, as demonstrated in [5, Lemma 2.2] and [12, Lemma 5.2]. However, a general retraction introduces a gradient dependency in the model-actual difference (see Proposition 1 or [4, Assumption 4]). Notably, [4] introduced an additional smoothness assumption on the composition $f \circ R$ ([4, Assumption 2]) to bypass this challenge. We will show that this additional assumption is unnecessary. Additionally, the general order ω introduces a step-size dependency in σ_k, which vanishes when $\omega = \mu \wedge \nu$ (see (9)). These two dependencies cause σ_k much more challenging to bound for RAR with 2+ω regularization. To address the step-size dependency, we will utilize the following proposition.

Proposition 4 (Step-size bound)**.** *If the Cauchy condition (TC.C) is satisfied, we have*

$$\|\eta_k\| \leqslant \left(\frac{3(\beta_H + 1)}{\sigma_k}\right)^{1/\omega} \vee \left(\|g_k\| \wedge \left(\frac{6\|g_k\|}{\sigma_k}\right)^{1/(1+\omega)}\right).$$

The proposition is a direct consequence of the Cauchy condition, and we defer its proof to Appendix A.4. Furthermore, we will show that the step-size has a lower bound determined by ε_g and ε_H. To address the gradient dependency, in addition to Proposition 3, we also need the following proposition.

Proposition 5 (Gradient bound)**.** *The gradient sequence generated by Algorithm 2 is bounded, i.e., there exists $\beta_g > 0$ such that $\|g_k\| \leqslant \beta_g$ for any $k \in \mathcal{K}$.*

In practice, there are various means to bound the gradient sequence. For instance, if the level set $\{x \in \mathcal{M} : f(x) \leqslant f_0\}$ is compact or if the algorithm converges, $\{g_k\}$ is bounded. We provide a proof of Proposition 5 adapted from [12, Theorem 2.5] in Appendix A.5 without any additional assumptions. We are now ready to provide an upper bound of the regularization parameter.

Lemma 3 (Regularization parameter upper bound)**.** *The regularization parameter has an upper bound*

$$\bar{\sigma} \coloneqq C_\sigma \left(\varepsilon_g \wedge \varepsilon_H^{1/\alpha}\right)^{-[\omega-\alpha]_+},$$

where C_σ is a positive constant satisfying $C_\sigma \geqslant 1$ and $[x]_+ \coloneqq \max\{0, x\}$.

Proof. In this proof, we omit the superscript of $\bar{m}_{x_k}^{\mathrm{ar}}$ and the subscript x_k of $\bar{m}_{x_k}$, m_{x_k}, and R_{x_k}. According to the update rule in Algorithm 2, the regularization parameter only expands when $\rho_k \leqslant \varrho_2$. Therefore, we only need to derive the upper bound of σ_k when $\rho_k \leqslant \varrho_2$, in which case (3) gives

$$\begin{aligned}
\varrho_2\left(m(0) - m\left(\eta_k\right)\right) &\geqslant f\left(x_k\right) - f\left(R\left(\eta_k\right)\right) \\
&= m(0) - m\left(\eta_k\right) + m\left(\eta_k\right) - f\left(\exp\left(\eta_k\right)\right) + f\left(\exp(\eta_k)\right) - f\left(R\left(\eta_k\right)\right) \\
&\geqslant m(0) - m\left(\eta_k\right) - |m(\eta_k) - f(\exp(\eta_k))| - |f(\exp(\eta_k)) - f(R(\eta_k))|,
\end{aligned}$$

which can be reformulated as

$$\left(1 - \varrho_2\right)\left(\bar{m}(0) - \bar{m}\left(\eta_k\right) + \varphi(\eta_k; \sigma_k)\right) \leqslant |f(R(\eta_k)) - f(\exp(\eta_k))| + |f(\exp(\eta_k)) - m(\eta_k)|. \tag{6}$$

The first term in the right hand side of (6) can be bounded by Proposition 1. For the second term, by the Taylor expansion on manifolds (see, e.g., [35, Lemma 3]), there exists $\tau \in [0, 1]$ such that

$$|f(\exp(\eta_k)) - m(\eta_k)| = \left|\frac{1}{2}\left\langle \eta_k, (H_k - P_\gamma^{\tau\to 0} H_\tau P_\gamma^{0\to\tau})\eta_k \right\rangle\right| \leqslant \frac{1}{2}\|H_k - P_\gamma^{\tau\to 0} H_\tau P_\gamma^{0\to\tau}\|\|\eta_k\|^2.$$

where γ is the geodesic from x_k to $\exp(\eta_k)$. Then by the Hölder continuity of $\operatorname{Hess} f$, we get

$$|f(\exp(\eta_k)) - m(\eta_k)| \leqslant \frac{1}{2} C_H \tau \|\eta_k\|^{2+\mu} \leqslant C_H \|\eta_k\|^{2+\mu}. \tag{7}$$

Combining (6), (7), the Cauchy condition (4), and Proposition 1 gives

$$(1 - \varrho_2)\varphi(\eta_k; \sigma_k) \leqslant C_H \left\|\eta_k\right\|^{2+\mu} + (1 + C_R \|\eta_k\|^\nu) \cdot \beta_H C_R \left\|\eta_k\right\|^{2+\nu} + C_R \|g_k\|\|\eta_k\|^{1+\nu}. \tag{8}$$

A general $2+\omega$ regularization necessitates the consideration of two cases: small and large step-sizes.
(C.I) When $\|\eta_k\| \leqslant C_{\sigma,1} \coloneqq 1 \wedge (\frac{1-\varrho_2}{2(2+\omega)C_R})^{1/\nu}$, by Proposition 3, we have

$$\frac{1 - \varrho_2}{2 + \omega} \sigma_k \|\eta_k\|^{2+\omega} \leqslant (C_H + (1 + C_R)\beta_H C_R)\|\eta_k\|^{2+\mu\wedge\nu} + C_R(\beta_H + \sigma_k\|\eta_k\|^\omega)\|\eta_k\|^{2+\nu},$$

which gives

$$\frac{1 - \varrho_2}{2(2 + \omega)} \sigma_k \|\eta_k\|^{2+\omega} \leqslant (C_H + (2 + C_R)\beta_H C_R)\|\eta_k\|^{2+\mu\wedge\nu}.$$

Let $C_{\sigma,2} \coloneqq 2(2 + \omega)(C_H + (2 + C_R)\beta_H C_R)/(1 - \varrho_2)$. The above inequality is equivalent to

$$\sigma_k \|\eta_k\|^\omega \leqslant C_{\sigma,2} \|\eta_k\|^{\mu\wedge\nu}. \tag{9}$$

(C.I.I) If $\|g_k\| > \varepsilon_g$, by Proposition 3 and (9), we get

$$\varepsilon_g \leqslant (\beta_H + \sigma_k \|\eta_k\|^\omega)\|\eta_k\| \leqslant (\beta_H + C_{\sigma,2}\|\eta_k\|^{\mu\wedge\nu})\|\eta_k\| \leqslant (\beta_H + C_{\sigma,2})\|\eta_k\|,$$

which gives $\|\eta_k\| \geqslant \varepsilon_g/(\beta_H + C_{\sigma,2})$.
(C.I.II) If $\|g_k\| \leqslant \varepsilon_g$, termination condition (TC.2) is used. Then if the algorithm dose not terminate, by Proposition 2 and (9), we get

$$\begin{aligned}
\varepsilon_H \leqslant& \|\eta_k\|^\theta + \lambda_{\max}\left(\operatorname{Hess}\varphi(\eta_k; \sigma_k)\right) \\
=& \|\eta_k\|^\theta + \lambda_{\max}\left(\sigma_k \|\eta_k\|^\omega \left(\frac{\eta_k \eta_k^T}{\omega\|\eta_k\|^2} + I\right)\right) \\
\leqslant& \|\eta_k\|^\theta + \frac{1 + \omega}{\omega} \sigma_k \|\eta_k\|^\omega \\
\leqslant& (1 + C_{\sigma,2}(1 + \omega)/\omega)\|\eta_k\|^{\mu\wedge\nu\wedge\theta},
\end{aligned}$$

which gives $\|\eta_k\| \geqslant (\frac{\varepsilon_H}{1+C_{\sigma,2}(1+\omega)/\omega})^{1/\alpha}$. Let $C_{\sigma,3} \coloneqq (\beta_H + C_{\sigma,2}) \vee (1 + C_{\sigma,2}(1+\omega)/\omega)^{1/\alpha}$. For (C.I), we have $\|\eta_k\| \geqslant C_{\sigma,3}^{-1}(\varepsilon_g \wedge \varepsilon_H^{1/\alpha})$. Plugging it back into (9) gives

$$\sigma_k \leqslant C_{\sigma,2}\|\eta_k\|^{\mu\wedge\nu-\omega} \leqslant C_{\sigma,2}\|\eta_k\|^{\alpha-\omega} \leqslant C_{\sigma,2}\|\eta_k\|^{-[\omega-\alpha]_+} \leqslant C_{\sigma,4}(\varepsilon_g \wedge \varepsilon_H^{1/\alpha})^{-[\omega-\alpha]_+},$$

where $C_{\sigma,4} \coloneqq C_{\sigma,2}C_{\sigma,3}^{[\omega-\alpha]_+}$.

(C.II) When $\|\eta_k\| > C_{\sigma,1}$, Proposition 4 introduces another two cases.
(C.II.I) When the former term in Proposition 4 is active, we have

$$\sigma_k \leqslant \|\eta_k\|^{-\omega} \cdot 3(\beta_H + 1) \leqslant C_{\sigma,5} \coloneqq 3C_{\sigma,1}^{-\omega}(\beta_H + 1).$$

(C.II.II) When the latter term is active in Proposition 4, (8) and Proposition 5 give

$$\frac{1-\varrho_2}{2+\omega}\sigma_k\|\eta_k\|^{2+\omega} \leqslant C_H\beta_g^{2+\mu} + (1 + C_R\beta_g^{\nu}) \cdot \beta_H C_R \beta_g^{2+\nu} + C_R\beta_g\beta_g^{1+\nu},$$

which further gives

$$\sigma_k \leqslant C_{\sigma,6} \coloneqq \frac{2+\omega}{(1-\varrho_2)C_{\sigma,1}^{2+\omega}}\left(C_H\beta_g^{2+\mu} + C_R\beta_g^{2+\nu} + (1+C_R\beta_g^{\nu})C_R\beta_H\beta_g^{2+\nu}\right).$$

Therefore, for (C.II), we have $\sigma_k \leqslant C_{\sigma,5} \vee C_{\sigma,6}$. Combining (C.I) and (C.II) gives

$$\sigma_k \leqslant (C_{\sigma,4} \vee C_{\sigma,5} \vee C_{\sigma,6}) \cdot (\varepsilon_g \wedge \varepsilon_H^{1/\alpha})^{-[\omega-\alpha]_+}.$$

Let $C_\sigma \coloneqq \kappa_3(C_{\sigma,4} \vee C_{\sigma,5} \vee C_{\sigma,6})$. We conclude σ_k has an upper bound: $\bar{\sigma} \coloneqq C_\sigma(\varepsilon_g \wedge \varepsilon_H^{1/\alpha})^{-[\omega-\alpha]_+}$. □

As we can see, when $\omega > \alpha$, the algorithm *under-regularizes*, meaning it provides inadequate regularization compared to methods with smaller ω. This can result in potentially larger regularization parameters, especially when the tolerance ε_g or ε_H is set to a small value. The underlying reason for this behavior is that in cases where the problem exhibits limited smoothness, the quadratic model function m_x may fail to deliver a sufficiently accurate approximation of the objective function, particularly at points farther away from x. As a remedy, the regularization parameter is increased to constrain the step size.

Lemma 4 (Minimal successful decrease of RAR). *For $k \in \mathcal{S}$, we have*

$$f(x_k) - f(x_{k+1}) \geqslant \begin{cases} C_s(\varepsilon_g/\bar{\sigma})^{\frac{2+\omega}{1+\omega}}, & k \in \mathcal{S}_{GG}, \\ C_s(\varepsilon_H/\bar{\sigma})^{\frac{2+\omega}{\omega}}, & k \in \mathcal{S}_L, \end{cases}$$

where C_s is a positive constant satisfying $C_s \geqslant 1$ and $\bar{\sigma}$ is specified in Lemma 3.

Proof. By the Cauchy condition (4), for $k \in \mathcal{S}$, we have

$$f(x_k) - f(x_{k+1}) \geqslant \varrho_1(m_{x_k}(0) - m_{x_k}(\eta_k)) \geqslant \frac{\varrho_1\underline{\sigma}}{2+\omega}\|\eta_k\|^{2+\omega}. \tag{10}$$

When $\|\eta_k\| \geqslant 1$, (10) directly satisfies the requirement because

$$\frac{\varrho_1\underline{\sigma}}{(2+\omega)}\|\eta_k\|^{2+\omega} \geqslant \frac{\varrho_1\underline{\sigma}}{(2+\omega)} \gtrsim 1 \gtrsim (\varepsilon_g/\bar{\sigma})^{1/(1+\omega)} \vee (\varepsilon_H/\bar{\sigma})^{1/\omega}.$$

Thus, we only consider the case where $\|\eta_k\| < 1$. For $k \in \mathcal{S}_{GG}$, by Proposition 2, we get

$$\varepsilon_g \leqslant C_{s,1}\|\eta_k\|^{1+\alpha} + \bar{\sigma}\|\eta_k\|^{1+\omega}, \tag{11}$$

where $C_{s,1} := 2C_H(C_R+1)+\beta_H C_R + 1$ and $\bar{\sigma}$ is specified in Lemma 3. We claim that the right hand side of (11) is always dominated by $\bar{\sigma}\|\eta_k\|^{1+\omega}$, i.e., $\|\eta_k\|^{1+\alpha} \lesssim \bar{\sigma}\|\eta_k\|^{1+\omega}$. Since $\|\eta_k\| < 1$, this is obviously true when $\alpha \geqslant \omega$. Now suppose $\alpha < \omega$ and the right hand side of (11) is dominated by $C_{s,1}\|\eta_k\|^{1+\alpha}$. Then, (11) gives

$$\|\eta_k\| \gtrsim \varepsilon_g^{1/(1+\alpha)} \geqslant \varepsilon_g \geqslant \varepsilon_g \wedge \varepsilon_H^{1/\alpha}.$$

Since $\alpha < \omega$, by Lemma 3, we get

$$\|\eta_k\|^{\alpha-\omega} \lesssim \left(\varepsilon_g \wedge \varepsilon_H^{1/\alpha}\right)^{\alpha-\omega} \lesssim \bar{\sigma},$$

which indicates $\|\eta_k\|^{\alpha} \lesssim \bar{\sigma}\|\eta_k\|^{\omega}$, affirming the dominance of $\bar{\sigma}\|\eta_k\|^{\omega}$. Therefore, we conclude that there exists $C_{s,2} > 0$ such that (11) is equivalent to

$$\varepsilon_g \leqslant C_{s,2}\bar{\sigma}\|\eta_k\|^{1+\omega},$$

which further gives $\|\eta_k\| \geqslant (\varepsilon_g/(\bar{\sigma}C_{s,2}))^{1/(1+\omega)}$.

For $k \in \mathcal{S}_L$, by Proposition 2, we have

$$\varepsilon_H \leqslant \|\eta_k\|^{\theta} + \bar{\sigma}\|\eta_k\|^{\omega}.$$

Similarly, we claim that the right hand side of the above inequality is dominated by $\bar{\sigma}\|\eta_k\|^{\omega}$. Therefore, there exists $C_{s,3} > 0$ such that

$$\varepsilon_H \leqslant C_{s,3}\bar{\sigma}\|\eta_k\|^{\omega},$$

which gives $\|\eta_k\| \geqslant (\varepsilon_H/(\bar{\sigma}C_{s,3}))^{1/\omega}$. Let $C_s = (C_{s,2}^{-(2+\omega)/(1+\omega)} \wedge C_{s,3}^{-(2+\omega)/\omega}) \cdot \varrho_1 \underline{\sigma}/(2+\omega)$. Then we get the result combining (10). □

Plugging Lemmas 3 and 4 into Corollary 1 gives the iteration complexity of RAR with $2+\omega$ regularization.

Theorem 1 (Iteration complexity of RAR)**.** *Under Assumption 1, Algorithm 2 finds an $(\varepsilon_g, \varepsilon_H)$-approximate second-order stationary point with the following worst-case iteration complexity:*

$$O\left(\max\left\{\varepsilon_g^{-(1+[\omega-\alpha]_+)\cdot\frac{2+\omega}{1+\omega}},\ \varepsilon_g^{-\frac{(2+\omega)[\omega-\alpha]_+}{\omega}}\varepsilon_H^{-\frac{2+\omega}{\omega}},\ \varepsilon_H^{-\left(1+\frac{[\omega-\alpha]_+}{\alpha}\right)\frac{2+\omega}{\omega}}\right\}\right),$$

where $\alpha = \mu \wedge \nu \wedge \theta$.

Proof. By Lemma 3, the second logarithmic term in Corollary 1 is suppressed by the first term. When $\varepsilon_H \leqslant \varepsilon_g^{\alpha}$, Lemma 3 states that $\bar{\sigma} = C_\sigma \varepsilon_H^{-[\omega-\alpha]_+/\alpha}$. We calculate the ratio of two minimal decreases for cases of $k \in \mathcal{S}_{GG}$ and $k \in \mathcal{S}_L$ in Lemma 4:

$$\begin{aligned}\frac{(\varepsilon_g/\bar{\sigma})^{(2+\omega)/(1+\omega)}}{(\varepsilon_H/\bar{\sigma})^{(2+\omega)/\omega}} =& C_\sigma^{(2+\omega)/(\omega(1+\omega))}\frac{(\varepsilon_g\varepsilon_H^{-[\omega-\alpha]_+/\alpha})^{(2+\omega)/(1+\omega)}}{(\varepsilon_H^{1-[\omega-\alpha]_+/\alpha})^{(2+\omega)/\omega}} \\ \geqslant& \frac{(\varepsilon_H^{1/\alpha-[\omega-\alpha]_+/\alpha})^{(2+\omega)/(1+\omega)}}{(\varepsilon_H^{1-[\omega-\alpha]_+/\alpha})^{(2+\omega)/\omega}} \\ =& \left(\varepsilon_H^{\omega-\alpha-[\omega-\alpha]_+-\omega\alpha}\right)^{(2+\omega)/(\alpha\omega(1+\omega))},\end{aligned}$$

where the inequality uses $C_\sigma \geqslant 1$ and $\varepsilon_H \leqslant \varepsilon_g^{\alpha}$. Since $\omega-\alpha-[\omega-\alpha]_+-\omega\alpha < 0$ (which is easy to verify), the above ratio is greater than 1, making the minimal decrease for $k \in \mathcal{S}_L$ a lower bound for Lemma 4:

$$f(x_k) - f(x_{k+1}) \geqslant C_s(\varepsilon_H/\bar{\sigma})^{(2+\omega)/\omega} = C_s C_\sigma^{-\frac{2+\omega}{\omega}} \varepsilon_H^{\left(1+\frac{[\omega-\alpha]_+}{\alpha}\right)\frac{2+\omega}{\omega}}.$$

When $\varepsilon_H > \varepsilon_g^{\alpha}$, Lemma 3 states that $\bar{\sigma} = C_\sigma \varepsilon_g^{-[\omega-\alpha]_+}$, and Lemma 4 gives

$$f(x_k) - f(x_{k+1}) \geqslant C_s \min\left\{ C_\sigma^{-\frac{2+\omega}{1+\omega}} \varepsilon_g^{(1+[\omega-\alpha]_+)\frac{2+\omega}{1+\omega}}, C_\sigma^{-\frac{2+\omega}{\omega}} \varepsilon_g^{\frac{(2+\omega)[\omega-\alpha]_+}{\omega}} \varepsilon_H^{\frac{2+\omega}{\omega}} \right\}.$$

Combining two cases gives the result. □

Corollary 2 (Optimal iteration complexity of RAR)**.** *When $\omega = \alpha$, Theorem 1 achieves the optimal iteration complexity:*

$$O\left(\max\left\{ \varepsilon_g^{-\frac{2+\alpha}{1+\alpha}}, \varepsilon_H^{-\frac{2+\alpha}{\alpha}} \right\} \right).$$

Proof. When $\alpha \geqslant \omega$, Algorithm 2 becomes

$$O\left(\max\left\{ \varepsilon_g^{-\frac{2+\omega}{1+\omega}}, \varepsilon_H^{-\frac{2+\omega}{\omega}} \right\} \right),$$

which monotonically decreases as ω increases. When $\alpha \leqslant \omega$, Algorithm 2 becomes

$$O\left(\max\left\{ \varepsilon_g^{-\frac{(2+\omega)(1+\omega-\alpha)}{1+\omega}}, \varepsilon_g^{-\frac{(2+\omega)(\omega-\alpha)}{\omega}} \varepsilon_H^{-\frac{2+\omega}{\omega}}, \varepsilon_H^{-\frac{2+\omega}{\alpha}} \right\} \right) \geqslant O\left(\max\left\{ \varepsilon_g^{-\frac{(2+\omega)(1+\omega-\alpha)}{1+\omega}}, \varepsilon_H^{-\frac{2+\omega}{\alpha}} \right\} \right),$$

where the right hand side monotonically increases with ω. Therefore, the complexity in Theorem 1 is optimal when $\alpha = \omega$. □

Corollary 2 agrees with Lemma 3 that $\omega > \alpha$ under-regularizes. It also suggests that $\omega < \alpha$ *over-regularizes*, yielding a suboptimal complexity bound.

5 Riemannian Trust Region Methods

Riemannian trust region (RTR) methods [35, 1] add a trust region constraint to the vanilla quadratic model function (1). To facilitate the iteration complexity analysis, we additionally introduce a small *non-adaptive* quadratic regularization term, transforming the Riemannian trust region model problem at $x_k \in \mathcal{M}$ into:

$$\min_{\eta \in T_{x_k}\mathcal{M},\ \|\eta\| \leqslant \Delta_k} \quad \bar{m}_{x_k}^{\mathrm{tr}}(\eta; \Delta_k) = f(x_k) + \langle \eta, \operatorname{grad} f(x_k) \rangle + \frac{1}{2} \langle \eta, \operatorname{Hess} f(x_k)[\eta] \rangle + \frac{1}{4} \varepsilon_H \|\eta\|^2. \tag{12}$$

It is worth noting that we incorporate the trust region as a problem constraint, and thus the regularization term in the model function (2) is $\varphi^{\mathrm{tr}}(\eta) \coloneqq \varepsilon_H \|\eta\|^2/4$. We inherit the method framework in Algorithm 1 and continue to use a Lanczos-based Krylov subspace method as the subproblem solver [24, 19]. When the solution to (12) is situated at the trust region boundary, (TC.1) may not be satisfied. Therefore, we introduce a truncation termination criterion: a Krylov subspace solution is returned as soon as it reaches the trust region boundary, which gives

$$\|\eta_k\| = \Delta_k. \tag{TC.T}$$

Otherwise, we utilize the residual termination criteria (TC.1) or (TC.2). In all cases, (TC.C) must also be satisfied. We define two additional index sets:

$$\mathcal{B} \coloneqq \{k \in \mathcal{K} : \|\eta_k\| = \Delta_k\}, \quad \text{and} \quad \mathcal{I} \coloneqq \{k \in \mathcal{K} : \|\eta_k\| < \Delta_k\}.$$

A more practical truncated conjugate gradient (TCG) method [30, 31] can also be used to solve the subproblem. We defer the detailed description of these subproblem solvers to the next section. The only difference from a standard trust region subproblem solver is that here we pass a regularized Hessian $\bar{H}_k \coloneqq H_k + \frac{1}{2}\varepsilon_H I$ to it. The method is presented in Algorithm 3. When we set $\varepsilon_H = 0$ and omit the second-order certification, our algorithm reduces to the classical one in [1].

Algorithm 3: Riemannian Trust Region Method

```
parameters tolorance $\varepsilon_g, \varepsilon_H \in (0,1)$, trust region adaptation parameters
  $\kappa_2 \geqslant 1 > \kappa_1 > 0, \bar{\Delta} > 0, \Delta_0 \in (0, \bar{\Delta}]$; step acceptance parameter $\varrho \in [0, 1/4)$.
input initial point $x_0 \in \mathcal{M}$.
for $k = 0, 1, \ldots$ do
    if $\|g_k\| > \varepsilon_g$ then
        obtain $\eta_k$ by solving problem (12) with termination criteria (TC.(1/T)&C)
    else if $H_k \nsucceq -\varepsilon_H I$ then
        obtain $\eta_k$ by solving problem (12) with termination criteria (TC.(2/T)&C)
    else
        return $x_k$
    end
    compute $\rho_k$ using (3)
    set $\Delta_{k+1} = \begin{cases} \min\{\bar{\Delta}, \kappa_2\Delta_k\}, & \rho_k > 3/4 \text{ and } \|\eta_k\| = \Delta_k, & \text{// very successful} \\ \kappa_1\Delta_k, & \rho_k < 1/4, & \text{// unsuccessful} \\ \Delta_k, & \text{o.w.} & \text{// successful} \end{cases}$
    set $x_{k+1} = \begin{cases} R_{x_k}(\eta_k), & \rho_k \geqslant \varrho, \\ x_k & \text{o.w.} \end{cases}$
end
```

5.1 Iteration Complexity of RTR

As discussed in Section 3.1, obtaining an iteration complexity bound of RTR only requires establishing a lower bound of Δ_k and the minimal decrease in f for $k \in \mathcal{S}$. Similar to the case of RAR (see Section 4), we need to carefully discuss the gradient dependency introduced by a general retraction in order to establish a lower bound of Δ_k. Moreover, given that our problem does not possess $C^{2,1}$ smoothness, we employ certain strategies to derive the minimal successful decrease, which are essential for achieving the optimal complexity bound as shown in [13].

Lemma 5 (Trust region radius lower bound)**.** *The trust region radius has a lower bound*

$$\underline{\Delta} := C_{\Delta}\varepsilon_H^{\frac{1}{\mu\wedge\nu}},$$

where C_{Δ} is a positive constant.

Proof. In this proof, we omit the superscript of $\bar{m}^{\text{tr}}$ and the subscript x_k of $\bar{m}_{x_k}$, m_{x_k}, and R_{x_k}. According to the update rule in Algorithm 3, the trust region radius only shrinks when $\rho_k < 1/4$. Therefore, we only need to establish the lower bound of Δ_k when $\rho_k < 1/4$. As we are considering the lower bound of the trust region radius, we only need to consider the case of $\|\eta_k\| \leqslant 1$. Similar to (8), when $\rho_k < 1/4$, we have

$$\frac{3}{4}(\bar{m}(0) - \bar{m}(\eta_k)) + \frac{3}{16}\varepsilon_H\|\eta_k\|^2 \leqslant C_d\,\|\eta_k\|^{2+\mu\wedge\nu} + C_R\|g_k\|\|\eta_k\|^{1+\nu}, \tag{13}$$

where $C_d := C_H + (1 + C_R\bar{\Delta}^{\nu})\beta_H C_R$.

(C.I) If $k \in \mathcal{I}$, (13) and Proposition 3 gives

$$\frac{3}{16}\varepsilon_H \leqslant C_d\|\eta_k\|^{\mu\wedge\nu} + C_R(\beta_H + \varepsilon_H/2)\|\eta_k\|^{\nu},$$

which further gives

$$\Delta_k \geqslant \|\eta_k\| \geqslant \left(\frac{3\varepsilon_H}{16(C_d + C_R(\beta_H + \varepsilon_H/2))}\right)^{\frac{1}{\mu\wedge\nu}}. \tag{14}$$

(C.II) If $k \in \mathcal{B}$, by [26, Lemma 4.3], the Cauchy condition (TC.C) and (13) further give

$$\frac{3}{8}\|g_k\| \cdot \min\left\{\Delta_k, \frac{\|g_k\|}{\beta_H + \varepsilon_H/2}\right\} + \frac{3}{16}\varepsilon_H \Delta_k^2 \leqslant C_d \Delta_k^{2+\mu\wedge\nu} + C_R\|g_k\|\Delta_k^{1+\nu}. \tag{15}$$

(C.II.I) If $\Delta_k \geqslant \|g_k\|/(\beta_H + \varepsilon_H/2)$, since $\Delta_k = \|\eta_k\| \leqslant 1$, (15) gives

$$\frac{3}{16}\varepsilon_H \Delta_k^2 \leqslant C_d \Delta_k^{2+\mu\wedge\nu} + C_R\|g_k\|\Delta_k^{1+\nu} \leqslant (C_d + C_R(\beta_H + \varepsilon_H/2))\Delta_k^{2+\mu\wedge\nu},$$

which further gives the result in (14).

(C.II.II) If $\Delta_k < \|g_k\|/(\beta_H + \varepsilon_H/2)$, (15) can be reformulated as

$$\|g_k\|\Delta_k\left(\frac{3}{8} - C_R\Delta_k^{\nu}\right) \leqslant C_d\Delta_k^{2+\nu\wedge\mu} - \frac{3}{16}\varepsilon_H\Delta_k^2.$$

(C.II.II.I) If $\Delta_k \leqslant (\frac{3}{8C_R})^{1/\nu}$, we have

$$C_d\|\Delta_k\|^{2+\nu\wedge\mu} - \frac{3}{16}\varepsilon_H\|\Delta_k\|^2 \geqslant 0$$

which gives $\Delta_k \geqslant (\frac{3\varepsilon_H}{16C_d})^{\frac{1}{\mu\wedge\nu}}$. (C.II.II.II) The other case is just $\Delta_k > (\frac{3}{8C_R})^{1/\nu}$. Therefore, for (C.II.II), we have

$$\Delta_k \geqslant \min\left\{\left(\frac{3}{8C_R}\right)^{\frac{1}{\nu}}, \left(\frac{3\varepsilon_H}{16C_d}\right)^{\frac{1}{\mu\wedge\nu}}\right\}. \tag{16}$$

Combining (14) and (16), and the initial assumption $\|\eta_k\| \leqslant 1$ gives

$$\Delta_k \geqslant \min\left\{1, \left(\frac{3}{8C_R}\right)^{\frac{1}{\nu}}, \left(\frac{3\varepsilon_H}{16C_d}\right)^{\frac{1}{\mu\wedge\nu}}, \left(\frac{3\varepsilon_H}{16(C_d + C_R(\beta_H + \varepsilon_H/2))}\right)^{\frac{1}{\mu\wedge\nu}}\right\}.$$

Combining the above cases, we complete the proof with C_Δ defined accordingly. □

Lemma 6 (Minimal successful decrease of RTR)**.** *There exists a positive constant C_g such that for any $k \in \mathcal{S}$, we have*

$$f(x_k) - f(x_{k+1}) \geqslant \frac{\varrho\varepsilon_H}{4}\begin{cases} \Delta_k^2, & k \in \mathcal{B} \cap \mathcal{S}, \\ C_g^2 \min\{\|g_{k+1}\|^{4/(1+\alpha)}\varepsilon_H^{-2/\alpha}, \varepsilon_H^{2/\alpha}\}, & k \in \mathcal{I} \cap \mathcal{S}_G, \\ (\varepsilon_H/2)^{2/\theta}, & k \in \mathcal{I} \cap \mathcal{S}_L. \end{cases}$$

Proof. First of all, since the Cauchy termination criterion (TC.C) is enforced for all cases, for any $k \in \mathcal{S}$, we have

$$f(x_k) - f(x_{k+1}) \geqslant \varrho(m_{x_k}(0) - m_{x_k}(\eta_k)) = \varrho(\bar{m}_{x_k}(0) - \bar{m}_{x_k}(\eta_k) + \varepsilon_H\|\eta_k\|^2/4) \geqslant \frac{\varrho\varepsilon_H}{4}\|\eta_k\|^2. \tag{17}$$

Therefore, we only need to figure out a lower bound of $\|\eta_k\|$.

(C.I) For $k \in \mathcal{B} \cap \mathcal{S}$, we have $\|\eta_k\| = \Delta_k$.

(C.II) For $k \in \mathcal{I} \cap \mathcal{S}_G$, residual termination condition (TC.1) is used.

(C.II.I) If $\varepsilon_H\|\eta_k\| \leqslant \varepsilon_H^{\frac{1+\alpha}{2\alpha}}\|\eta_k\|^{\frac{1+\alpha}{2}}$, we have $\|\eta_k\| \leqslant 1$, which together with Propositions 2 and 3 gives

$$C_g'\|\eta_k\|^{1+\alpha} + \frac{1}{2}\varepsilon_H^{\frac{1+\alpha}{2\alpha}}\|\eta_k\|^{\frac{1+\alpha}{2}} - \|g_{k+1}\| \geqslant 0, \tag{18}$$

where $C_g' \coloneqq 2C_H(1+C_R) + \beta_H C_R + (\beta_H + \varepsilon_H/2)^{1+\theta}$. By [28, Lemma 17], the solution of this inequality indicates

$$\begin{aligned} \|\eta_k\| &\geqslant \left(\frac{-1+\sqrt{1+16C_g'\|g_{k+1}\|\varepsilon_H^{-\frac{1+\alpha}{\alpha}}}}{4C_g'} \varepsilon_H^{\frac{1+\alpha}{2\alpha}} \right)^{\frac{2}{1+\alpha}} \\ &\geqslant \left(\frac{-1+\sqrt{1+16C_g'}}{4C_g'} \min\left\{ \|g_{k+1}\|\varepsilon_H^{-\frac{1+\alpha}{2\alpha}}, \varepsilon_H^{\frac{1+\alpha}{2\alpha}} \right\} \right)^{\frac{2}{1+\alpha}} \\ &\eqqcolon C_g \min\{\|g_{k+1}\|^{2/(1+\alpha)}\varepsilon_H^{-1/\alpha}, \varepsilon_H^{1/\alpha}\}. \end{aligned} \tag{19}$$

(C.II.II) If $\varepsilon_H\|\eta_k\| > \varepsilon_H^{\frac{1+\alpha}{2\alpha}} \|\eta\|^{\frac{1+\alpha}{2}}$, we have

$$\|\eta_k\|^{1-\frac{1+\alpha}{2}} > \varepsilon_H^{\frac{1+\alpha}{2\alpha}-1} \quad \Longrightarrow \quad \|\eta_k\| > \varepsilon_H^{1/\alpha}.$$

Set $C_g \coloneqq C_g \wedge 1$. Then (C.II.II) also satisfies (19).

(C.III) For $k \in \mathcal{I} \cap \mathcal{S}_L$, since the algorithm does not terminate, residual termination condition (TC.2) is used and Proposition 2 gives

$$\varepsilon_H < -\lambda_{\min}(H_k) \leqslant \|\eta_k\|^\theta + \frac{\varepsilon_H}{2},$$

which further gives $\|\eta_k\| \geqslant (\varepsilon_H/2)^{1/\theta}$. Combining (17) and (C.I-III) gives the result. □

We should note that the termination criterion (TC.2) might appear less practical in the context of RTR because of the fixed nature of $\operatorname{Hess}\bar{m}^{\mathrm{tr}}$. However, when this criterion is applied, i.e., when $H_k \not\succeq -\varepsilon_H I$, we anticipate that the Krylov subspace method will be capable of identifying the negative curvature of the model problem and return a solution located on the trust region boundary. Therefore, we can omit (TC.2) and rely solely on (TC.T). To be consistent with RAR, here we retain (TC.2) as an early stopping criterion, and its inexactness can potentially expedite the process. We will give a concrete subproblem solver and an alternative termination criterion to replace (TC.2) for RTR in Section 6.4.

Plugging Lemmas 5 and 6 into Corollary 1 gives the iteration complexity of RTR.

Theorem 2 (Iteration complexity of RTR)**.** *Under Assumption 1, Algorithm 3 finds an $(\varepsilon_g, \varepsilon_H)$-approximate second-order stationary point with the following worst-case iteration complexity:*

$$O\left(\max\left\{ \varepsilon_H^{-1-\frac{2}{\alpha}}, \varepsilon_g^{-\frac{4}{1+\alpha}} \varepsilon_H^{-1+\frac{2}{\alpha}} \right\} \right),$$

where $\alpha = \mu \wedge \nu \wedge \theta$.

Proof. By Lemma 5, the second logarithmic term in Corollary 1 is suppressed by the first term. For $k \in \mathcal{S}_{GG}$, by Lemmas 5 and 6, we have

$$f(x_k) - f(x_{k+1}) \geqslant \frac{\varrho}{4} \min\left\{ \varepsilon_H \underline{\Delta}^2, C_g^2 \varepsilon_g^{4/(1+\alpha)} \varepsilon_H^{1-2/\alpha}, C_g^2 \varepsilon_H^{1+2/\alpha} \right\} \gtrsim \min\left\{ \varepsilon_H^{1+2/\alpha}, \varepsilon_g^{4/(1+\alpha)} \varepsilon_H^{1-2/\alpha} \right\}.$$

For $k \in \mathcal{S}_L$, by Lemmas 5 and 6, we have

$$f(x_k) - f(x_{k+1}) \geqslant \frac{\varrho}{4} \min\left\{ \varepsilon_H \underline{\Delta}^2, 2^{-2/\theta} \varepsilon_H^{1+2/\theta} \right\} \gtrsim \varepsilon_H^{1+2/\alpha}.$$

Combining two cases with Corollary 1 gives the result. □

Corollary 3 (Optimal iteration complexity of RTR)**.** *By choosing $\varepsilon_H = \varepsilon_g^{\alpha/(1+\alpha)}$, Theorem 1 achieves the optimal (first-order) iteration complexity:*

$$O\left(\varepsilon_g^{-\frac{2+\alpha}{1+\alpha}}\right).$$

Our analysis, especially in light of Corollaries 2 and 3, reveals that the parameters μ, ν, and θ are *equivalently* responsible for governing the iteration complexity of RARN, and this control is encapsulated by the single parameter $\alpha = \mu \wedge \nu \wedge \theta$. In essence, elevating any of these parameters in isolation will not improve the worst-case iteration complexity. Furthermore, this observation can serve as a valuable guideline for algorithm design: when μ, a value determined by the problem, is relatively large, we need to increase ν and θ to match the accuracy of the model function, which approximates the objective function, to achieve the optimal iteration complexity. Conversely, a smaller μ provides the flexibility to relax the smoothness requirement on the retraction and the precision requirement on the subproblem solver, thus effectively curtailing computational costs.

6 Subproblem Solvers and Operation Complexity

In this section, we comprehensively analyze the subproblem solvers employed in RAR and RTR. Our focus centers on introducing Lanczos-based Krylov subspace methods as the central subproblem solver, complemented by integrating minimal eigenvalue oracles (MEO) to assess second-order stationarity. Building upon these subproblem solvers, we provide insights into our algorithms' operation complexity, quantified in terms of the number of Hessian-vector products.

6.1 Lanczos-Based Krylov Subspace Methods

We begin by restating the RARN subproblem:

$$\min_{\eta \in T_x\mathcal{M}} \bar{m}(\eta) \coloneqq \langle \eta, g \rangle + \frac{1}{2}\langle \eta, H\eta \rangle + \varphi(\|\eta\|, \sigma). \tag{20}$$

In this formulation, we omit the reference to a specific point $x \in \mathcal{M}$ and the constant term $f(x)$. Additionally, we simplify the regularization function to only depend on the norm of η. Krylov subspace methods aim to approximately minimize $\bar{m}$ by identifying solutions within specific subspaces known as the Krylov subspaces. In this paper, for the automatic fulfillment of the Cauchy condition (TC.C), we construct the Krylov subspaces based on (H, g). Specifically, the order-j Krylov subspace is defined as $\operatorname{span}\{g, Hg, \ldots, H^{j-1}g\}$. We denote ξ_j as the order-j Krylov subspace solution to (20). Consequently, the Cauchy point $\eta^{\mathrm{C}} = \xi_1$ is the first-order Krylov subspace solution. To efficiently compute the Krylov subspace solutions, we use the Lanczos method to construct an orthogonal Krylov subspace basis $Q_j = (q_0, \ldots, q_{j-1})$, transforming the jth subproblem iteration on the tangent space into the following one in $\mathbb{R}^j$:

$$\min_{u \in \mathbb{R}^j} \|g\| \cdot (u)_1 + \frac{1}{2} u^* Q_j^* H Q_j u + \varphi(\|u\|, \sigma),$$

where $(u)_1 \in \mathbb{R}$ is the first component of $u \in \mathbb{R}^j$, u^* is the transpose of u, and $Q_j^* : T_x\mathcal{M} \to \mathbb{R}^j$ is the adjoint of Q_j, and we use the fact that $Q_j^* g = \|g\| e_1$ and $\|Q_j u\| = \|u\|$. The tridiagonal structure of $Q_j^* H Q_j$ streamlines the efficient solution of the above problem, requiring only $O(1)$ Hessian-vector products. For a more detailed explanation of Lanczos-based Krylov subspace methods, please refer to, e.g., [4, Section 8], [18, Chapter 10], and [32, Lecture 36]. We present several conditioned operation complexities of Krylov subspace methods.

Proposition 6 (Operation complexity of Krylov subspace methods)**.** *Suppose η_k is returned by a Krylov subspace method, and η^* is the exact solution to (20). We have*

1. *for the adaptive regularization subproblem, η_k is an ε-optimal solution to (20) with at most $O(\varepsilon^{-1/2}\|\eta^*\|)$ Hessian-vector product operations [8, Corollary 4.2][2];*
2. *for the adaptive regularization subproblem, if $H \succeq \varepsilon_H I$, η_k is an ε-optimal solution to (20) with at most $O(\lambda_{\min}(H)^{-1/2}\log(1/\varepsilon))$ Hessian-vector product operations [8, Corollary 4.2];*
3. *for the* non-regularized *trust region subproblem, if $H \succeq \varepsilon_H I$, η_k is an ε-optimal solution to (20) with at most $O(\lambda_{\min}(H)^{-1/2}\log(1/\varepsilon))$ Hessian-vector product operations [8, Theorem 4.1].*

In Proposition 6, we omit the operation complexity's dependency on $\bar{m}_{x_k}(0) - \min_\eta \bar{m}_{x_k}(\eta)$, which can be bounded by $\|g_k\|$. As our analysis shows that both RAR and RTR converge to an approximate stationary point, this dependence becomes a constant factor and can thus be neglected. We elaborate further on this aspect in Appendix A.6.

6.2 Minimal Eigenvalue Oracle

We have not yet specified how we test for second-order stationarity and what to do when $H \not\succeq \varepsilon_H I$ in Items 2 and 3 of Proposition 6. To address this, we assume our algorithm can access a minimal eigenvalue oracle (MEO). Following [17], we require the MEO to indicate whether $H \succeq -\varepsilon_H I$ or, if not, to return a unit vector η^{E} such that

$$\langle \eta^{\mathrm{E}}, g\rangle \leqslant 0 \quad \text{and} \quad \langle \eta^{\mathrm{E}}, H\eta^{\mathrm{E}}\rangle \leqslant -\frac{1}{2}\varepsilon_H\|\eta^{\mathrm{E}}\|^2, \tag{TC.E}$$

with at most $O(n \wedge \varepsilon_H^{-1/2}\log(1/\delta))$ Hessian-vector products, where n is the manifold dimension and δ is the probability that the MEO incorrectly claims $H \succeq -\varepsilon_H$. Actually, Krylov subspace methods, with a Lanczos-based orthogonal basis or an H-conjugate basis (resulting in conjugate gradient methods), satisfy these requirements with an operation complexity of $O(n \wedge \varepsilon_H^{-1/2}\log(n/\delta^2))$ [29, Appendix B.2].

6.3 Operation Complexity of RAR

In this subsection, we provide a concrete algorithm of adaptive regularization methods with a strong operation complexity guarantee. To simplify the analysis, we set $\alpha := \mu = \nu = \theta = \omega$, the optimal parameter profile discussed in previous sections. As for the subproblem solver, we substitute Lines 4 to 10 of Algorithm 2 with Algorithm 4. Based on Item 1 of Proposition 6, we limit the number of subproblem iterations to $K_{\mathrm{sub}} = O(\varepsilon_H^{-1/2})$, and propose a new termination criterion:

$$m(0) - m(\eta) \geqslant \frac{\alpha\varepsilon_H^{(2+\alpha)/\alpha}}{12\bar{\sigma}^{2/\alpha}}. \tag{TC.D}$$

which directly fulfills the requirement for minimal successful decrease (Lemma 4, albeit potentially with a different constant), rendering (TC.2) obsolete. We have the following proposition.

Proposition 7. *If $H_k \not\succeq -\varepsilon_H I$, then a Krylov subspace method will find a solution satisfying (TC.D) within at most $O(\varepsilon_H^{-1/2})$ iterations.*

We first try solving the subproblem using a Lanczos-based Krylov subspace method (Lines 2 to 8). By Proposition 7, if `max_flag` remains `true` after K_{sub} iterations, we know that $H_k \succeq -\varepsilon_H I$ (Lines 9 to 14). Subsequently, if $\|g_k\| \leqslant \varepsilon_g$, it signifies that x_k is an $(\varepsilon_g, \varepsilon_H)$-approximate second-order stationary point and we can terminate the algorithm (Line 11). If $\|g_k\| > \varepsilon_g$, because the Krylov subspace method fails to meet an appropriate termination criterion within the specified number of

[2] [8] considers cubic regularization with fixed regularization parameter, but it still applies here because the implied complexity does not involve the regularization order or parameter.

iterations, we still need to efficiently find an iteration step that aims for the first-order stationarity. Since now $H_k \succeq -\varepsilon_H I$, we solve (20) again with a regularized Hessian $\bar{H}_k = H_k + 2\varepsilon_H I$ to obtain η_k (Line 13), which is guaranteed to satisfy the residual termination criterion (TC.1) with at most $\widetilde{O}(\varepsilon_H^{-1/2})$ iterations by Item 2 of Proposition 6.

On the other hand, even if `max_flag = false`, we cannot assert that $H_k \not\succeq -\varepsilon_H I$ (Lines 15 to 24). Therefore, when `max_flag = false` and the first-order stationarity is achieved ($\|g_k\| \leqslant \varepsilon_g$), we still need to call an MEO to test the second-order stationarity. If the MEO claims the presence of the second-order stationarity, we terminate the algorithm (Line 19); otherwise, we continue the algorithm with the returned subproblem solution $\eta_k = \xi_j$ (Line 22).

Algorithm 4: Subproblem solver for RAR

```
1  set max_flag = true
2  for $j = 1, ..., K_{\text{sub}}$ do
3      get order-$j$ Krylov subspace solution $\xi_j$
4      if ($\|g_k\| > \varepsilon_g$ and $\xi_j$ satisfies (TC.1)) or $\xi_j$ satisfies (TC.D) then
5          set max_flag = false
6          break for
7      end
8  end
9  if max_flag = true then
10     if $\|g_k\| \leqslant \varepsilon_g$ then
11         return $x_k$ and terminate the outer algorithm
12     else
13         return $\eta_k$ by solving problem (5) with $H_k$ replace with $\bar{H}_k = H_k + 2\varepsilon_H I$ and
            termination criterion (TC.1)
14     end
15 else // max_flag = false
16     if $\|g_k\| \leqslant \varepsilon_g$ then
17         call an MEO to test the second-order stationarity
18         if the MEO indicates that $H_k \succeq -\varepsilon_H I$ then
19             return $x_k$ and terminate the outer algorithm
20         end
21     else
22         return $\eta_k = \xi_j$
23     end
24 end
```

Due to the utilization of a regularized Hessian (Line 13), a minor adjustment is required for the iteration complexity of RAR (Corollary 2). When combined with the operation complexity of Lanczos-based Krylov subspace methods and MEO, we obtain an operation complexity guarantee of RAR.

Corollary 4 (Operation complexity of RAR)**.** *Algorithms 2 and 4 finds an $(\varepsilon_g, \varepsilon_H)$-approximate second-order stationary point with the following worst-case operation complexity:*

$$\widetilde{O}\left(\max\left\{\varepsilon_g^{-\frac{2(2+\alpha)}{1+\alpha}}\varepsilon_H^{\frac{2+\alpha}{\alpha}}, \varepsilon_H^{-\frac{2+\alpha}{\alpha}}\right\}\cdot\varepsilon_H^{-1/2}\right),$$

where $\widetilde{O}$ suppresses the logarithmic dependency on ε_H. When $\varepsilon_H = \varepsilon_g^{\alpha/(1+\alpha)}$, the complexity becomes

$$\widetilde{O}\left(\varepsilon_g^{-\frac{4+3\alpha}{2(1+\alpha)}}\right),$$

which further becomes $\widetilde{O}(\varepsilon_g^{-7/4})$ when $\alpha = 1$.

We thoroughly examine in Appendix B the applicability of Lemmas 3 and 4, and thus Corollary 2, when employing Algorithm 4, making this operation complexity valid.

6.4 Operation Complexity of RTR

Similarly, we adopt the parameter profile $\alpha := \mu = \nu = \theta$ in this subsection. For RTR, we substitute Lines 4 to 10 of Algorithm 3 with Algorithm 5. Additionally, to ensure $\bar{H}_k \succeq \varepsilon_H I$ indicates $H_k \succeq -\varepsilon_H I$, we increase the regularization strength by setting the regularized Hessian as $\bar{H}_k = H_k + 2\varepsilon_H I$, which aligns with its usage in Algorithm 4. This modification deprecates termination criterion (TC.2). Based on Item 3 of Proposition 6 applied to the new regularized Hessian $\bar{H}_k$, we limit the number of subproblem iterations to $K_{\text{sub}} = \widetilde{O}(\varepsilon_H^{-1/2})$, which is sufficient to meet any residual termination criterion if $\bar{H}_k \succeq \varepsilon_H I$. Therefore, we first try solving the subproblem using a Lanczos-based Krylov subspace method (Lines 2 to 8). If `max_flag` remains `true` after K_{sub} iterations, we know that $H_k \not\succeq -\varepsilon_H I$, and we use an MEO to find a suitable iteration step corresponding to its minimal eigenvalue (Lines 9 to 11).

On the other hand, if `max_flag = false`, we cannot claim that $H_k \succeq -\varepsilon_H I$ (Lines 12 to 23). If we are confident that x_k is not a second-order stationary point, we proceed with the returned subproblem solution $\eta_k = \xi_j$ (Line 14). Otherwise, we once again call an MEO to test the second-order stationarity. When making use of the return of an MEO, we set $\eta_k = \Delta_k \eta^{\mathrm{E}}$ to satisfy (TC.T) and thus replace (TC.2).

We remark that in the subproblem solver for RAR (Algorithm 4), we exclusively employ an MEO to assess second-order stationarity and do not rely on it to provide an iteration step. This is because for the RAR subproblem, a Krylov subspace method performs well (efficiently finds a solution that offers sufficient model decrease) when $H_k \not\succeq -\varepsilon_H I$ (see Proposition 7). However, Krylov subspace methods for the RTR subproblem do not enjoy this nice property when $H_k \not\succeq -\varepsilon_H I$, especially in *hard cases* (see [26, Chapter 4] or [16, Chapter 7]). Therefore, we also incorporate an MEO into our subproblem solver for RTR to determine the iteration steps. This approach is characterized by its simplicity and efficiency (recall that a Krylov subspace method is an MEO), aligning with established practices in the field [17, 33].

By Corollary 3 and the operation complexity of Lanczos-based Krylov subspace methods and MEO, we obtain an operation complexity guarantee of RTR.

Corollary 5 (Operation complexity of RTR)**.** *Algorithms 3 and 5 finds an $(\varepsilon_g, \varepsilon_H)$-approximate second-order stationary point with the following worst-case operation complexity:*

$$\widetilde{O}\left(\max\left\{\varepsilon_H^{-1-\frac{2}{\alpha}}, \varepsilon_g^{-\frac{4}{1+\alpha}}\varepsilon_H^{-1+\frac{2}{\alpha}}\right\} \cdot \varepsilon_H^{-1/2}\right).$$

where $\widetilde{O}$ suppresses the logarithmic dependency on ε_H. Let $\varepsilon_H = \varepsilon_g^{\alpha/(1+\alpha)}$, the complexity becomes

$$\widetilde{O}\left(\varepsilon_g^{-\frac{4+3\alpha}{2(1+\alpha)}}\right),$$

which further becomes $\widetilde{O}(\varepsilon_g^{-7/4})$ when $\alpha = 1$.

Algorithm 5: Subproblem solver for RTR

```
set max_flag = true
for $j = 1, ..., K_{\text{sub}}$ do
    get order-$j$ Krylov subspace solution $\xi_j$ with the regularized Hessian $\bar{H}_k$
    if $\xi_j$ satisfies (TC.1/T) then
        set max_flag = false
        break for
    end
end
if max_flag = true then
    call an MEO to return $\eta^{\mathrm{E}}$
    return $\eta_k = \Delta_k \eta^{\mathrm{E}}$
else // max_flag = false
    if $\|g_k\| > \varepsilon_g$ or $\|\xi_j\| = \Delta_k$ then
        return $\eta_k = \xi_j$
    else
        call an MEO to test the second-order stationarity
        if the MEO indicates that $H_k \succeq -\varepsilon_H I$ then
            return $x_k$ and terminate the outer algorithm
        else
            return $\eta_k = \Delta_k \eta^{\mathrm{E}}$, where $\eta^{\mathrm{E}}$ is obtained through the MEO
        end
    end
end
```

We thoroughly examine in Appendix C the applicability of Lemmas 5 and 6, and thus Corollary 3, when employing Algorithm 5, making this operation complexity valid.

Remark 2 (TCG as RTR's subproblem sovler). Truncated conjugate gradient (TCG) methods [30, 31] constitute another practical choice for the subproblem solver in RTR [35, 1]. For trust region methods designed for C^2 problems in Euclidean spaces with TCG as the subproblem solver, [17] provides an iteration complexity of $\widetilde{O}(\varepsilon_g^{-3/2})$ and an operation complexity of $\widetilde{O}(\varepsilon_g^{-7/4})$. Remarkably, with minimal adjustments, our analysis seamlessly extends to the context of RTR employing TCG as the subproblem solver, yielding an iteration complexity given in Corollary 3 and an operation complexity given in Corollary 5. This adaptability arises due to the commensurate operation complexity of TCG and Lanczos-based Krylov subspace methods. Additionally, when TCG terminates within the trust region ($\|\eta_k\| < \Delta_k$), it equates to a Krylov subspace method (with a H_k-conjugate basis). When $\|\eta_k\| = \Delta_k$, we only necessitate the Cauchy condition (TC.C), a criterion also satisfied by TCG. Thus, our analysis aptly extends to encompass TCG as well.

We conclude with some additional insights into the unified view of the subproblem procedures for RAR and RTR. When `max_flag = true`, indicating that the subproblem solver falls short of meeting the termination criteria within $\widetilde{O}(\varepsilon_H^{-1/2})$ iterations, both methods gain the knowledge about the approximate positive definiteness of H_k: RAR can deduce that $H_k \succeq -\varepsilon_H I$, while RTR can assert that $H_k \not\succeq -\varepsilon_H I$. However, despite this information, obtaining a suitable iteration step remains elusive at this point for both methods. Consequently, an additional process is required by both RAR and RTR to determine η_k: RAR employs a regularization on the Hessian, and RTR can leverage an MEO. Conversely, when `max_flag = false` and $\|g_k\| \leqslant \varepsilon_g$, both procedures still lack information regarding the approximate positive definiteness of H_k; thus, both depend on an MEO to test the second-order stationarity. However, since RAR can find a suitable iteration point satisfying the second-order termination criterion (TC.2) or (TC.D), it does not rely on the MEO to return η_k.

References

[1] P.-A. Absil, C. G. Baker, and K. A. Gallivan. Trust-region methods on Riemannian manifolds. *Foundations of Computational Mathematics*, 7:303–330, 2007.

[2] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, Princeton, N.J. ; Woodstock, 2008. ISBN 978-0-691-13298-3.

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

[4] N. Agarwal, N. Boumal, B. Bullins, and C. Cartis. Adaptive regularization with cubics on manifolds. *Mathematical Programming*, 188:85–134, 2021.

[5] E. G. Birgin, J. Gardenghi, J. M. Martínez, S. A. Santos, and P. L. Toint. Worst-case evaluation complexity for unconstrained nonlinear optimization using high-order regularized models. *Mathematical Programming*, 163(1-2):359–368, 2017. ISSN 0025-5610, 1436-4646. doi: 10.1007/s10107-016-1065-8.

[6] N. Boumal. *An Introduction to Optimization on Smooth Manifolds*. Cambridge University Press, 2023.

[7] N. Boumal, P.-A. Absil, and C. Cartis. Global rates of convergence for nonconvex optimization on manifolds. *IMA Journal of Numerical Analysis*, 39(1):1–33, 2019.

[8] Y. Carmon and J. C. Duchi. First-order methods for nonconvex quadratic minimization. *SIAM Review*, 62(2):395–436, 2020. ISSN 0036-1445, 1095-7200. doi: 10.1137/20M1321759.

[9] Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford. Accelerated methods for nonconvex optimization. *SIAM Journal on Optimization*, 28(2):1751–1772, 2018.

[10] Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford. Lower bounds for finding stationary points I. *Mathematical Programming*, 184(1-2):71–120, 2020.

[11] C. Cartis, N. I. M. Gould, and P. L. Toint. Adaptive cubic regularisation methods for unconstrained optimization. Part II: Worst-case function- and derivative-evaluation complexity. *Mathematical Programming*, 130(2):295–319, 2011. ISSN 0025-5610, 1436-4646. doi: 10.1007/s10107-009-0337-y.

[12] C. Cartis, N. I. M. Gould, and P. L. Toint. Adaptive cubic regularisation methods for unconstrained optimization. Part I: Motivation, convergence and numerical results. *Mathematical Programming*, 127(2):245–295, 2011. ISSN 0025-5610, 1436-4646. doi: 10.1007/s10107-009-0286-5.

[13] C. Cartis, N. I. M. Gould, and P. L. Toint. Optimal Newton-type methods for nonconvex smooth optimization problems. *Technical Report ERGO Technical Report 11–009*, 2011.

[14] C. Cartis, N. I. M. Gould, and M. Lange. On monotonic estimates of the norm of the minimizers of regularized quadratic functions in Krylov spaces. *BIT Numerical Mathematics*, 60(3):583–589, 2020.

[15] C. Cartis, N. I. M. Gould, and P. L. Toint. Sharp worst-case evaluation complexity bounds for arbitrary-order nonconvex optimization with inexpensive constraints. *SIAM Journal on Optimization*, 30(1):513–541, 2020.

[16] A. R. Conn, N. I. M. Gould, and P. L. Toint. *Trust Region Methods*. SIAM, 2000.

[17] F. E. Curtis, D. P. Robinson, C. W. Royer, and S. J. Wright. Trust-region Newton-CG with strong second-order complexity guarantees for nonconvex optimization. *SIAM Journal on Optimization*, 31(1):518–544, 2021.

[18] G. H. Golub and C. F. Van Loan. *Matrix Computations*. JHU press, 2013.

[19] N. I. M. Gould, S. Lucidi, M. Roma, and P. L. Toint. Solving the trust-region subproblem using the Lanczos method. *SIAM Journal on Optimization*, 9(2):504–525, 1999.

[20] G. N. Grapiglia and Y. Nesterov. Regularized Newton methods for minimizing functions with Hölder continuous hessians. *SIAM Journal on Optimization*, 27(1):478–506, 2017.

[21] Y. Hsia, R.-L. Sheu, and Y.-x. Yuan. Theory and application of p-regularized subproblems for p$>$ 2. *Optimization Methods and Software*, 32(5):1059–1077, 2017.

[22] R.-J. Jiang, Z.-S. Zhou, and Z.-R. Zhou. Cubic regularization methods with second-order complexity guarantee based on a new subproblem reformulation. *Journal of the Operations Research Society of China*, 10(3):471–506, 2022.

[23] H. Li and Z. Lin. Restarted nonconvex accelerated gradient descent: No more polylogarithmic factor in the $O(\varepsilon^{-7/4})$ complexity. In *International Conference on Machine Learning*, pages 12901–12916. PMLR, 2022.

[24] L. Lukšan, C. Matonoha, and J. Vlček. On Lagrange multipliers of trust-region subproblems. *BIT Numerical Mathematics*, 48(4):763–768, 2008.

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

[26] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer Series in Operations Research. Springer, New York, 2nd ed edition, 2006. ISBN 978-0-387-30303-1.

[27] C. Qi. *Numerical Optimization Methods on Riemannian Manifolds*. PhD thesis, Florida State University, 2011.

[28] C. W. Royer and S. J. Wright. Complexity analysis of second-order line-search algorithms for smooth nonconvex optimization. *SIAM Journal on Optimization*, 28(2):1448–1477, 2018.

[29] C. W. Royer, M. O'Neill, and S. J. Wright. A Newton-CG algorithm with complexity guarantees for smooth unconstrained optimization. *Mathematical Programming*, 180:451–488, 2020.

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

[31] P. Toint. Towards an efficient sparsity exploiting Newton method for minimization. In *Sparse Matrices and Their Uses*, pages 57–88. Academic press, 1981.

[32] L. N. Trefethen and D. Bau. *Numerical Linear Algebra*, volume 181. SIAM, 2022.

[33] P. Xu, F. Roosta, and M. W. Mahoney. Newton-type methods for non-convex optimization under inexact Hessian information. *Mathematical Programming*, 184(1-2):35–70, 2020.

[34] Z. Yao, P. Xu, F. Roosta, S. J. Wright, and M. W. Mahoney. Inexact Newton-CG algorithms with complexity guarantees. *IMA Journal of Numerical Analysis*, 43(3):1855–1897, 2023.

[35] C. Zhang, R. Xiao, W. Huang, and R. Jiang. Riemannian trust region methods for SC^1 minimization. *arXiv preprint arXiv:2307.00490*, 2023.

[36] F. Zhang. *Matrix Theory: Basic Results and Techniques.* Springer, 2011.

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

Appendix

A Proof of Propositions

A.1 Proof of Proposition 1

Proof. The first inequality is given by [35, Proposition 3] and the second is by [35, Corollary 5.1]. In this proof, we omit the subscript x_k of R_{x_k} and $\exp_{x_k}$. By the Taylor expansion on manifolds ([6, Section 4.1]), there exists $\tau \in [0,1]$ such that

$$f(R(\eta_k)) = f(\exp(\eta_k)) + \left\langle \operatorname{grad} f(\gamma(\tau)), \gamma'(\tau) \right\rangle,$$

where γ is the geodesic from $\exp(\eta_k)$ to $R(\eta_k)$. Then, we have

$$\begin{align*}
& |f(R(\eta_k)) - f(\exp(\eta_k))| \\
=& \left| \left\langle \operatorname{grad} f(\gamma(\tau)), \gamma'(\tau) \right\rangle \right| \\
=& \left| \left\langle P_\gamma^{\tau \to 0} \operatorname{grad} f(\gamma(\tau)), \gamma'(0) \right\rangle \right| \tag{21} \\
\leqslant& \| P_\gamma^{\tau \to 0} \operatorname{grad} f(\gamma(\tau)) \| \| \gamma'(0) \| \\
=& \| \operatorname{grad} f(x_k) - \operatorname{grad} f(x_k) + P_\gamma^{\tau \to 0} \operatorname{grad} f(\gamma(\tau)) \| \cdot \operatorname{dist}(\exp(\eta_k), R(\eta_k)) \tag{22} \\
\leqslant& (\| \operatorname{grad} f(x_k) \| + \beta_H \operatorname{dist}(x_k, \gamma(\tau))) \cdot \operatorname{dist}(\exp(\eta_k), R(\eta_k)), \tag{23}
\end{align*}$$

where (21) use the fact that the parallel transport $P_\gamma^{\tau \to 0}$ preserves inner product and $P_\gamma^{\tau \to 0} \gamma'(\tau) = \gamma(0)$ becomes γ' is parallel along γ; (22) is by the definition of a geodesic in Section 2; and in (23), β_H is the uniform operator norm bound of $\operatorname{Hess} f$, and then $\operatorname{grad} f$ is β_H-Lipschitz continuous. Then by the triangle inequality, we have

$$\operatorname{dist}(x_k, \gamma(\tau)) \leqslant \operatorname{dist}(x_k, \exp(\eta_k)) + \operatorname{dist}(\exp(\eta_k), \gamma(\tau)) \leqslant \|\eta_k\| + \operatorname{dist}(\exp(\eta_k), R(\eta_k)). \tag{24}$$

Combining (23), (24), and the first inequality of Proposition 1 gives the desired result. □

A.2 Proof of Proposition 2

Proof. We denote $r_k \coloneqq \operatorname{grad} \bar{m}_{x_k}(\eta_k)$ and $s_k \coloneqq \operatorname{grad} \varphi(\eta_k; \sigma_k)$. Then, $r_k = g_k + H_k \eta_k + s_k$. For $k \in \mathcal{S}$, we have

$$\begin{aligned}
\|g_{k+1}\| &= \|g_{k+1} + P_{k,k+1}(r_k - g_k - H_k \eta_k - s_k)\| \\
&\leqslant \|g_{k+1} - P_{k,k+1}(g_k + H_k \eta_k)\| + \|r_k\| + \|s_k\| \\
&\leqslant \underbrace{\|g_{k+1} - P_{k,k+1}(g_k + H_k \exp_{x_k}^{-1}(x_{k+1}))\|}_{S_1} + \underbrace{\|H_k(\exp_{x_k}^{-1}(x_{k+1}) - \eta_k)\|}_{S_2} + \|\eta_k\|^{1+\theta} + \|s_k\|,
\end{aligned}$$

where we write $P_{k,k+1} \coloneqq P_{x_k,x_{k+1}}$, which is norm preserving, for notational simplicity and use (TC.1): $\|r_k\| \leqslant \|\eta_k\|^{1+\theta}$. Since $f \in C^{2,\mu}$, by the Taylor expansion of $\operatorname{grad} f(x_k)$ ([6, Section 4.1]), we have

$$\begin{aligned} S_1 \leqslant & C_H \operatorname{dist}(x_k, x_{k+1})^{1+\mu} \\ \leqslant & C_H (\operatorname{dist}(x_k, \exp_{x_k}(\eta_k)) + \operatorname{dist}(\exp_{x_k}(\eta_k), R_{x_k}(\eta_k)))^{1+\mu} \\ \leqslant & 2C_H \operatorname{dist}(x_k, \exp_{x_k}(\eta_k))^{1+\mu} + 2C_H \operatorname{dist}(\exp_{x_k}(\eta_k), R_{x_k}(\eta_k))^{1+\mu} \qquad (25) \\ \leqslant & 2C_H \|\eta_k\|^{1+\mu} + 2C_H C_R^{1+\mu} \|\eta_k\|^{(1+\mu)(1+\nu)}, \qquad (26) \end{aligned}$$

where (25) is by the convexity of function $z \mapsto z^{1+\mu}$ and (26) is by Proposition 1, which also gives

$$S_2 \leqslant \beta_H C_R \|\eta_k\|^{1+\nu}.$$

Substituting S_1 and S_2 with the above bounds gives the first result of Proposition 2.

For η_k satisfying termination criterion (TC.2), we have

$$\begin{aligned} -\lambda_{\min}(H_k) =& \lambda_{\max}(-H_k) = \lambda_{\max}(-\operatorname{Hess} \bar{m}(\eta_k) + \operatorname{Hess} \varphi(\eta_k; \sigma_k)) \\ \leqslant & -\lambda_{\min}(\operatorname{Hess} \bar{m}(\eta_k)) + \lambda_{\max}(\operatorname{Hess} \varphi(\eta_k; \sigma_k)) \\ \leqslant & \|\eta_k\|^{\theta} + \lambda_{\max}(\operatorname{Hess} \varphi(\eta_k; \sigma_k)), \end{aligned}$$

where the first inequality is by [36, Theorem 10.21] and uses the fact that $\operatorname{Hess} \bar{m}$ and $\operatorname{Hess} \varphi$ are Hermitian. □

A.3 Proof of Proposition 3

Proof. Let ξ_1 be the minimizer of (5) restricted in the first Krylov subspace $\operatorname{span}\{g_k\}$. Then we have $\operatorname{grad}_\tau \bar{m}(\tau g_k) = \tau \langle g_k, \operatorname{grad} \bar{m}(\xi_1) \rangle = 0$, where $\xi_1 = \tau g_k$. Calculating the gradient of $\bar{m}$ gives

$$\langle g_k, g_k + H_k \xi_1 + \operatorname{grad} \varphi(\xi_1) \rangle = 0, \qquad (27)$$

which further gives

$$\|g_k\|^2 = -\langle g_k, H_k \xi_1 \rangle - \langle g_k, \operatorname{grad} \varphi(\xi_1) \rangle \leqslant \|g_k\| \|H_k\| \|\xi_1\| + \|g_k\| \|\operatorname{grad} \varphi(\xi_1)\|.$$

Then for RAR, since $\|\operatorname{grad} \varphi(\xi)\| = \sigma_k \|\xi\|^{1+\omega}$ increases with the magnitude of ξ and $\|\xi_1\| \leqslant \|\eta_k\|$ [14, Theorem 1], we have

$$\|g_k\| \leqslant \beta_H \|\xi_1\| + \|\operatorname{grad} \varphi(\xi_1)\| \leqslant \beta_H \|\eta_k\| + \|\operatorname{grad} \varphi(\eta_k)\|. \qquad (28)$$

For RTR, $\|\xi_1\| \leqslant \|\eta_k\|$ also holds [30, 24]. Therefore, if $\|\eta_k\| < \Delta_k$, we know $\|\xi_1\| < \Delta_k$, and then (27) still holds. Since $\|\operatorname{grad} \varphi(\xi)\|$ also increases with the magnitude of ξ for RTR, (28) still holds. □

A.4 Proof of Proposition 4

Proof. In this proof, we omit the superscript and subscript of $\bar{m}^{\mathrm{ar}}_{x_k}$. By the Cauchy termination criterion (TC.C), we have

$$0 \leqslant \bar{m}(0) - \bar{m}(\eta_k) \leqslant \|g_k\| \|\eta_k\| + \frac{1}{2} \beta_H \|\eta_k\|^2 - \frac{\sigma_k}{2+\omega} \|\eta_k\|^{2+\omega}. \qquad (29)$$

If $\|\eta_k\| \geqslant \|g_k\|$, (29) gives

$$\|\eta_k\|^2 \cdot \left(1 + \frac{\beta_H}{2} - \frac{\sigma_k}{2+\omega} \|\eta_k\|^{\omega}\right) \geqslant 0.$$

Therefore, we have

$$\|\eta_k\| \leqslant \left(\frac{(2+\beta_H)(2+\omega)}{2\sigma_k}\right)^{1/\omega} \vee \|g_k\|. \tag{30}$$

Another decomposition of the right hand side (29) gives

$$0 \leqslant \|\eta_k\| \cdot \left(\|g_k\| - \frac{\sigma_k}{2(2+\omega)}\|\eta_k\|^{1+\omega}\right) + \|\eta_k\|^2 \cdot \left(\frac{\beta_H}{2} - \frac{\sigma_k}{2(2+\omega)}\|\eta_k\|^{\omega}\right).$$

Hence the two terms on the right hand side of the above inequality cannot both be negative. This gives

$$\|\eta_k\| \leqslant \left(\frac{2\beta_H(2+\omega)}{2\sigma_k}\right)^{1/\omega} \vee \left(\frac{2(2+\omega)\|g_k\|}{\sigma_k}\right)^{1/(1+\omega)}. \tag{31}$$

Utilizing $\omega \in (0,1]$ and the fact that $(a \vee b) \wedge (a \vee c) = a \vee (b \wedge c)$, (30) and (31) give

$$\|\eta_k\| \leqslant \left(\frac{3(\beta_H+1)}{\sigma_k}\right)^{1/\omega} \vee \left(\|g_k\| \wedge \left(\frac{6\|g_k\|}{\sigma_k}\right)^{1/(1+\omega)}\right).$$

□

A.5 Proof of Proposition 5

Before proving Proposition 5, we provide a lemma on the lower bound of the model decrease.

Lemma 7 (Cauchy decrease)**.** *If η_k satisfies the Cauchy termination criterion (TC.C), then we have*

$$\bar{m}_{x_k}^{\mathrm{ar}}(0) - \bar{m}_{x_k}^{\mathrm{ar}}(\eta_k) \geqslant \frac{\|g_k\|^2}{4\left(\beta_H \vee \left(\sigma_k^{1/(1+\omega)}\|g_k\|^{\omega/(1+\omega)}\right)\right)}.$$

Proof. In this proof, we omit the superscript and subscript of $\bar{m}_{x_k}^{\mathrm{ar}}$. Let η^{C} be the Cauchy point. Then for any $\tau \in \mathbb{R}$, we have $\bar{m}(\eta^{\mathrm{C}}) \leqslant \bar{m}(-\tau g_k)$. Therefore, we have

$$\begin{aligned}\bar{m}(0) - \bar{m}(\eta^{\mathrm{C}}) &\geqslant \bar{m}(0) - \bar{m}\left(-\tau g_k\right)\\ &= \tau\left\|g_k\right\|^2 - \frac{1}{2}\tau^2\left\langle g_k, H_k g_k\right\rangle - \frac{\sigma_k}{2+\omega}\tau^{2+\omega}\left\|g_k\right\|^{2+\omega}\\ &\geqslant \tau\left\|g_k\right\|^2\left(1 - \frac{\tau\beta_H}{2} - \frac{\sigma_k\tau^{1+\omega}\left\|g_k\right\|^{\omega}}{2+\omega}\right).\end{aligned}$$

The above inequality holds for any τ. Let $\tau = \frac{1}{2\beta_H} \vee \left(\frac{2+\omega}{4\sigma_k\|g_k\|^{\omega}}\right)^{1/(1+\omega)}$. We get

$$\begin{aligned}\bar{m}(0) - \bar{m}\left(\eta^{\mathrm{C}}\right) &\geqslant \tau\left\|g_k\right\|^2\left(\left(\frac{1}{2} - \frac{1}{2\beta_H}\cdot\frac{\beta_H}{2}\right) + \left(\frac{1}{2} - \frac{2+\omega}{4\sigma_k\left\|g_k\right\|^{\omega}}\cdot\frac{\sigma_k\left\|g_k\right\|^{\omega}}{2+\omega}\right)\right)\\ &= \frac{\tau\left\|g_k\right\|^2}{2}\\ &\geqslant \frac{\left\|g_k\right\|^2}{4\left(\beta_H \vee \left(\sigma_k^{1/(1+\omega)}\left\|g_k\right\|^{\omega/(1+\omega)}\right)\right)}.\end{aligned}$$

Notice that the above inequality also holds for $g_k = 0$, in which case we let $\tau = 1/(2\beta_H)$. □

Proof of Proposition 5. Our proof follows the same logic of [12, Theorem 2.5 and Corollary 2.6] which proves that $\lim_k \|g_k\| = 0$. Here we aim to show $\limsup_k \|g_k\| < +\infty$. The result automatically holds if $\mathcal{S}$ is finite. Thus, we assume $|\mathcal{S}| = +\infty$ in the rest of the proof. We first claim that $\liminf_k \|g_k\| < +\infty$. If not, for any $C > 0$, there exists $K_1 \in \mathbb{N}$ such that for any $k \geqslant K_1$, it holds that $\|g_k\| > C$. By Lemma 7, we have

$$\sum_{k\in\mathcal{S}} f(x_k) - f(x_{k+1}) \geqslant \sum_{k\in\mathcal{S},\ k\geqslant K_1} \frac{\varrho_1\|g_k\|^2}{4\left(\beta_H \vee \left(\sigma_k^{1/(1+\omega)}\|g_k\|^{\omega/(1+\omega)}\right)\right)}.$$

Since f is bounded below, we know the summand sequence of the right hand side of the above inequality converges to zero. Then since $\|g_k\| > C$, we know that β_H in the denominator must be inactive when $k \geqslant K_2$ for some $K_2 \geqslant K_1$. Therefore, we get

$$+\infty > \sum_{k\in\mathcal{S}} f_k - f_{k+1} \geqslant \sum_{k\in\mathcal{S},\ k\geqslant K_2} \frac{\varrho_1\|g_k\|^{2-\omega/(1+\omega)}}{4\sigma_k^{1/(1+\omega)}} \geqslant \frac{\varrho_1 C}{4} \sum_{k\in\mathcal{S},\ k\geqslant K_2} \left(\frac{\|g_k\|}{\sigma_k}\right)^{1/(1+\omega)},$$

which gives

$$\lim_{k_1\to\infty} \sum_{k\in\mathcal{S},\ k\geqslant k_1} \left(\frac{\|g_k\|}{\sigma_k}\right)^{1/(1+\omega)} = 0, \tag{32}$$

which further gives $\lim_{\mathcal{S}\ni k\to\infty} \|g_k\|/\sigma_k = 0$ and thus $\lim_{\mathcal{S}\ni k\to\infty} \sigma_k = +\infty$ due to $\|g_k\| > C$. Let $C \geqslant \beta_H + 1$. Then, by Proposition 4, there exists $K_3 \geqslant K_2$ such that for any $k \in \mathcal{S}$ and $k \geqslant K_3$, we have

$$\|\eta_k\| \leqslant \left(\frac{6\|g_k\|}{\sigma_k}\right)^{1/(1+\omega)} \leqslant 1. \tag{33}$$

Then, by the triangle inequality and Proposition 1, for any $k_2 \geqslant k_1 \geqslant K_3$, we have

$$\begin{aligned}
\operatorname{dist}(x_{k_2}, x_{k_1}) &\leqslant \sum_{\substack{k\in\mathcal{S}\\ k_1\leqslant k<k_2}} \operatorname{dist}(x_k, x_{k+1}) \\
&= \sum_{\substack{k\in\mathcal{S}\\ k_1\leqslant k<k_2}} \|\exp_{x_k}^{-1}(R_{x_k}(\eta_k))\| \\
&\overset{\text{Prop. 1}}{\leqslant} \sum_{\substack{k\in\mathcal{S}\\ k_1\leqslant k<k_2}} \|\eta_k\| \cdot (1 + C_R\|\eta_k\|^\nu) \\
&\overset{(33)}{\leqslant} (1 + C_R) \cdot \sum_{\substack{k\in\mathcal{S}\\ k_1\leqslant k<k_2}} \|\eta_k\| \\
&\overset{(33)}{\leqslant} 6^{1/(1+\omega)}(1 + C_R) \cdot \sum_{\substack{k\in\mathcal{S}\\ k_1\leqslant k<k_2}} \left(\frac{\|g_k\|}{\sigma_k}\right)^{1/(1+\omega)} \\
&\leqslant 6^{1/(1+\omega)}(1 + C_R) \cdot \sum_{\substack{k\in\mathcal{S}\\ k\geqslant k_1}} \left(\frac{\|g_k\|}{\sigma_k}\right)^{1/(1+\omega)} \\
&\overset{(32)}{\to} 0 \quad \text{as} \quad k_1 \to +\infty.
\end{aligned} \tag{34}$$

Therefore, $\{x_k\}$ is a Cauchy sequence. Since $\mathcal{M}$ is complete, we know $\{x_k\}$ converges, contradicting the hypothesis $\liminf_k \|g_k\| = +\infty$.

Now suppose $\liminf_k \|g_k\| \leqslant C$. If $\limsup_k \|g_k\| = +\infty$, then there exists $\mathcal{S}_C \subset \mathcal{S}$ such that $|\mathcal{S}_C| = +\infty$ and $\|g_k\| \geqslant C+1$ for any $k \in \mathcal{S}_C$. For any $k_1 \in \mathcal{S}_C$, let k_2 be the smallest integer such that $k_1 \leqslant k_2 \in \mathcal{S}$ and $\|g_{k_2}\| \leqslant C$. Consider the infinite index set formed by these consecutive index sequences (say $\mathcal{J}$), whose definition implies $\|g_k\| > C$ for all $k \in \mathcal{J}$. We observe a similar pattern as in (32–34), which only relies on the condition that $\|g_k\| > C$ within the specified index set. This leads us to the conclusion that

$$\operatorname{dist}(x_{k_2}, x_{k_1}) \to 0 \quad \text{as} \quad k_1 \to +\infty.$$

By the Lipschitzness of $\operatorname{grad} f$, we get

$$\|g_{k_1}\| \leqslant \|g_{k_2}\| + \|g_{k_1} - g_{k_2}\| \to \|g_{k_2}\| \leqslant C \quad \text{as} \quad k_1 \to +\infty,$$

contradicting to the assumption that $\|g_{k_1}\| \geqslant C+1$. Therefore, we conclude $\limsup_k \|g_k\| < +\infty$ and thus $\{g_k\}$ is bounded. □

A.6 Remark on Proposition 6

Item 1 in Proposition 6 is directly given by [8, Corollary 4.2], our construction of the Krylov subspace basis, and the fact that $|\lambda_{\max}(H_k)| \vee |\lambda_{\min}(H_k)| \leqslant \beta_H$. Similarly, for Items 2 and 3, we simply need to obtain an upper bound on $\bar{m}_{x_k}(0) - \bar{m}_{x_k}(\eta^*)$, where $\eta^* := \operatorname{argmin}_\eta \bar{m}_{x_k}(\eta)$. Recall that we set $\alpha = \mu = \nu = \theta = \omega$.

For RAR (Item 2), by Proposition 4 (which remains valid when using Algorithm 4 as the subproblem solver, as discussed in Appendix B), we have

$$\|\eta^*\| \leqslant \left(\frac{3(\beta_H + 1)}{\underline{\sigma}}\right)^{1/\alpha} \vee \|g_k\| = O(\|g_k\| \vee 1).$$

Then, we get

$$\begin{aligned} \bar{m}^{\text{ar}}_{x_k}(0) - \bar{m}^{\text{ar}}_{x_k}(\eta^*) \leqslant & \|g_k\|\|\eta^*\| + \frac{1}{2}\|H_k\|\|\eta^*\|^2 + \frac{\sigma_k}{2+\alpha}\|\eta^*\|^{2+\alpha} \\ \leqslant & \|g_k\|\|\eta^*\| + \frac{\beta_H}{2}\|\eta^*\|^2 + \frac{\bar{\sigma}}{2+\alpha}\|\eta^*\|^{2+\alpha} \\ = & O(\|g_k\|^{2+\alpha} \vee 1), \end{aligned}$$

where we use the fact that $\bar{\sigma} = O(1)$ when $\alpha = \mu = \nu = \theta = \omega$, which is regardless of the subproblem solver (see Appendix B).

For RTR (Item 3), we have

$$\bar{m}^{\text{tr}}_{x_k}(0) - \bar{m}^{\text{tr}}_{x_k}(\eta^*) \leqslant \|g_k\|\|\eta^*\| + \frac{1}{2}(\|H_k\| + 2\varepsilon_H)\|\eta^*\|^2 \leqslant \|g_k\|\bar{\Delta} + \frac{\beta_H + 2\varepsilon_H}{2}\bar{\Delta}^2 = O(\|g_k\| \vee 1).$$

Then, we can proceed the analysis with the maximum number of subproblem iterations explicitly dependent on $\|g_k\|$, i.e., $K_{\text{sub}} := \widetilde{O}((\|g_k\|^{2+\alpha} \vee 1) \cdot \varepsilon_H^{-1/2})$. Notably, both RAR and RTR converge to a *small gradient region* characterized by $\|g_k\| \leqslant \varepsilon_g$. Consequently, we can assert the existence of a positive constant β_g such that $\|g_k\| \leqslant \beta_g$ for all $k \in \mathcal{K}$. Hence, the gradient norm dependency becomes a constant factor and can be disregarded.

A.7 Proof of Proposition 7

Proof. In this proof, we omit the subscript and superscript of $\bar{m}^{\text{ar}}_{x_k}$. By the optimality condition of $\min \bar{m}(\eta)$ [21, Theorem 1.1], we know

$$\begin{cases} g_k + H_k\eta^* + \sigma_k\|\eta^*\|^\alpha\eta^* = \operatorname{grad} \bar{m}(\eta^*) = 0, & \text{first-order optimality,} \\ \lambda_{\min}(H_k) + \sigma_k\|\eta^*\|^\alpha \geqslant \lambda_{\min}(\operatorname{Hess} \bar{m}(\eta^*)) \geqslant 0, & \text{second-order optimality,} \end{cases}$$

where $\eta^* = \operatorname{argmin}_\eta \bar{m}^{\mathrm{ar}}_{x_k}(\eta)$. Since $H_k \not\succeq -\varepsilon_H I$, by the second-order optimality condition, we get $\|\eta^*\| \geqslant (\varepsilon_H/\sigma_k)^{1/\alpha}$. The first-order optimality condition gives

$$\begin{aligned}\bar{m}(0) - \bar{m}(\eta^*) &= -\langle g_k, \eta^*\rangle - \frac{1}{2}\langle \eta^*, H_k\eta^*\rangle - \frac{\sigma_k}{2+\alpha}\|\eta^*\|^{2+\alpha} \\ &= \langle \eta^*, H_k\eta^*\rangle + \sigma_k\|\eta^*\|^{2+\alpha} - \frac{1}{2}\langle \eta^*, H_k\eta^*\rangle - \frac{\sigma_k}{2+\alpha}\|\eta^*\|^{2+\alpha} \\ &= \frac{1}{2}\langle \eta^*, H_k\eta^*\rangle + \frac{1+\alpha}{2+\alpha}\sigma_k\|\eta^*\|^{2+\alpha}.\end{aligned}$$

Combined with the second-order optimality condition, we get

$$\bar{m}(0) - \bar{m}(\eta^*) \geqslant \frac{1}{2}(-\sigma_k\|\eta^*\|^{2+\alpha}) + \frac{1+\alpha}{2+\alpha}\sigma_k\|\eta^*\|^{2+\alpha} = \frac{\alpha}{2(2+\alpha)}\sigma_k\|\eta^*\|^{2+\alpha}.$$

Then if $\bar{m}(\eta) - \bar{m}(\eta^*) \leqslant \alpha\sigma_k\|\eta^*\|^{2+\alpha}/12$, we have

$$\begin{aligned}m(0) - m(\eta) \geqslant& \bar{m}(0) - \bar{m}(\eta) = \bar{m}(0) - \bar{m}(\eta^*) - (\bar{m}(\eta) - \bar{m}(\eta^*)) \geqslant \frac{\alpha\sigma_k}{2(2+\alpha)}\|\eta^*\|^{2+\alpha} - \frac{\alpha\sigma_k}{12}\|\eta^*\|^{2+\alpha} \\ \geqslant& \frac{\alpha\sigma_k}{12}\|\eta^*\|^{2+\alpha} \geqslant \frac{\alpha\varepsilon_H^{(2+\alpha)/\alpha}}{12\bar{\sigma}^{2/\alpha}}.\end{aligned}$$

By Item 1 of Proposition 6, to make $\bar{m}(\eta) - \bar{m}(\eta^*) \leqslant \alpha\sigma_k\|\eta^*\|^{2+\alpha}/12$, we need at most

$$O((\sigma_k\|\eta^*\|^{2+\alpha})^{-1/2}\|\eta^*\|) = O(\|\eta^*\|^{-\alpha/2}) \leqslant O((\varepsilon_H^{1/\alpha})^{-\alpha/2}) = O(\varepsilon_H^{-1/2})$$

iterations. □

B Validating Iteration Complexity of RAR

In this subsection, we revisit Section 4 while using Algorithm 4 as the RAR subproblem process. We highlight the only differences between Algorithm 4 and a pure Krylov subspace method with (TC.1) and (TC.2): a new termination criterion (TC.D) replacing (TC.2) (Line 4) and occasional recalculations of the iteration step using a regularized Hessian $\bar{H}_k = H_k + 2\varepsilon_H I$ (Line 13). The MEO does not impact the iteration complexity because we only use it to test the second-order stationarity for RAR. Intuitively, the regularized Hessian will merely introduce a slightly larger Hessian operator norm bound $\beta_H + 2\varepsilon_H$; and new termination criterion (TC.D), which replaces (TC.2), directly applies a condition on the model decrease, which will not only fulfill all previous discussions but also make their establishment more straightforward.

We begin by examining Propositions 3 to 5, which only rely on the Cauchy condition (TC.C). One can observe that in these propositions, as well as in Lemma 7, any reliance on H_k gets transformed into its corresponding norm bound β_H. Therefore, the regularized Hessian merely leads to a slightly larger norm bound: $\beta_H + 2\varepsilon_H$. Without loss of generality, we can substitute all instances of β_H with $\beta_H + 2\varepsilon_H$. Furthermore, the corollary inequality (4) of the Cauchy condition remains valid because

$$m_{x_k}(0) - m_{x_k}(\eta_k) = \bar{m}_{x_k}(0) - \bar{m}_{x_k}(\eta_k) + \varepsilon_H\|\eta_k\|^2 + \varphi(\eta_k;\sigma_k) \geqslant \varepsilon_H\|\eta_k\|^2 + \varphi(\eta_k;\sigma_k) \geqslant \varphi(\eta_k;\sigma_k).$$

We then examine Lemma 3. Recall that we have set $\omega = \mu = \nu$. Then, for (C.I) in Lemma 3, (9) directly gives a bound $\sigma_k \leqslant C_{\sigma,2}$. Thus, we do not need to consider (C.I.II) where (TC.2) comes into play. The bound of (C.II) remains unchanged as it does not involve any termination criterion. Therefore, using Algorithm 4, the upper bound of σ_k reduces to

$$\bar{\sigma} = \kappa_3(C_{\sigma,2} \vee C_{\sigma,5} \vee C_{\sigma,6}).$$

For Lemma 4, a slight modification is required to accommodate the utilization of a regularized Hessian. We opt for a single lower bound on the model decrease for any $k \in \mathcal{S}_{GG} \sqcup \mathcal{S}_L$:

$$m_{x_k}(0) - m_{x_k}(\eta_k) \gtrsim \min\left\{ \varepsilon_g^{\frac{2(2+\alpha)}{1+\alpha}} \varepsilon_H^{-\frac{2+\alpha}{\alpha}}, \varepsilon_H^{\frac{2+\alpha}{\alpha}} \right\}. \tag{35}$$

It is evident that η_k returned using (TC.D) directly satisfies this lower bound. For η_k returned using (TC.1) and an unregularized Hessian, since Algorithm 4 also necessitates $\|g_k\| > \varepsilon_g$ in this case, $k \in \mathcal{S}_{GG} \sqcup \mathcal{S}_L$ implies $k \in \mathcal{S}_{GG}$. Lemma 4 for $k \in \mathcal{S}_{GG}$ directly gives (35) without adaptation in this case, because

$$\underbrace{\min\left\{ \varepsilon_g^{\frac{2(2+\alpha)}{1+\alpha}} \varepsilon_H^{-\frac{2+\alpha}{\alpha}}, \varepsilon_H^{\frac{2+\alpha}{\alpha}} \right\}}_{(35)} \Big/ \underbrace{\varepsilon_g^{\frac{2+\alpha}{1+\alpha}}}_{\text{Lem. } 4} = \min\left\{ \varepsilon_g^{\frac{2+\alpha}{1+\alpha}} \varepsilon_H^{-\frac{2+\alpha}{\alpha}}, \left(\varepsilon_g^{\frac{2+\alpha}{1+\alpha}} \varepsilon_H^{-\frac{2+\alpha}{\alpha}} \right)^{-1} \right\} \leqslant 1.$$

Finally, if η_k is returned using a regularized Hessian, (11) becomes

$$\varepsilon_g \leqslant (C_{s,1} + \bar{\sigma}) \|\eta_k\|^{1+\alpha} + 2\varepsilon_H \|\eta_k\|,$$

which resembles (18). Thus, similar to (C.II.I) and (C.II.II), we get $\|\eta_k\| \gtrsim \varepsilon_g^{1/(1+\alpha)}$ when $\|\eta_k\|^\alpha \gtrsim \varepsilon_H$; and when $\|\eta_k\|^\alpha \not\gtrsim \varepsilon_H$, we have

$$\begin{aligned}
\|\eta_k\| &\geqslant \left(\frac{-1 + \sqrt{1 + (C_{s,1} + \bar{\sigma}) \varepsilon_g \varepsilon_H^{-\frac{1+\alpha}{\alpha}}}}{C_{s,1} + \bar{\sigma}} \varepsilon_H^{\frac{1+\alpha}{2\alpha}} \right)^{\frac{2}{1+\alpha}} \\
&\geqslant \left(\frac{-1 + \sqrt{1 + C_{s,1} + \bar{\sigma}}}{C_{s,1} + \bar{\sigma}} \min\left\{ \varepsilon_g \varepsilon_H^{-\frac{1+\alpha}{2\alpha}}, \varepsilon_H^{\frac{1+\alpha}{2\alpha}} \right\} \right)^{\frac{2}{1+\alpha}} \\
&\gtrsim \min\left\{ \varepsilon_g^{\frac{2}{1+\alpha}} \varepsilon_H^{-\frac{1}{\alpha}}, \varepsilon_H^{\frac{1}{\alpha}} \right\}.
\end{aligned}$$

Then by the Cauchy condition (4), (35) holds. To recap, (35) holds for all cases. Therefore, the iteration complexity of RAR in Corollary 2 turns into

$$O\left(\max\left\{ \varepsilon_g^{-\frac{2(2+\alpha)}{1+\alpha}} \varepsilon_H^{\frac{2+\alpha}{\alpha}}, \varepsilon_H^{-\frac{2+\alpha}{\alpha}} \right\} \right).$$

Combined with the operation operation complexity of Lanczos-based Krylov subspace methods and MEO, we get Corollary 4.

C Validating Iteration Complexity of RTR

In this subsection, we revisit Section 5 while using Algorithm 5 as the RTR subproblem process. We highlight the only difference between Algorithm 5 and a pure Krylov subspace method with (TC.1), (TC.2) and (TC.T): we will occasionally use η^{E} returned by an MEO with (TC.E) to construct the iteration step, on which we do not impose any other conditions such as the Cauchy condition (TC.C). Intuitively, we only invoke an MEO if there is uncertainty regarding the second-order stationarity, and it will return an eigenvector associated with $\lambda_{\min}(H_k)$ when $H_k \not\succeq -\varepsilon_H I$. Similar to (TC.D), this approach will not only fulfill our previous discussions but also make their establishment more straightforward.

We only need to examine Lemmas 5 and 6. According to Algorithm 5, an MEO is invoked only if `max_flag = true` or $\|\xi_j\| < \Delta_k$. In both scenarios, we know that the first-order Krylov subspace solution resides within the interior of the trust region, and thereby $\|g_k\| \leqslant (\beta_H + 2\varepsilon_H)\|\xi_1\|$ (see Appendix A.3). Therefore, when using an MEO, (13) in Lemma 5 should be reformulated as

$$\frac{3}{4}(m_{x_k}(0) - m_{x_k}(\eta_k)) \leqslant C_d\|\eta_k\|^{2+\alpha} + C_R(\beta_H + 2\varepsilon_H)\|\xi_1\|\|\eta_k\|^{1+\alpha},$$

where we set $\alpha = \mu = \nu$. By (TC.E) and $\|\eta_k\| = \Delta_k$, we have

$$\frac{3}{8}\varepsilon_H \Delta_k^2 \leqslant (C_d + C_R(\beta_H + 2\varepsilon_H))\Delta_k^{2+\alpha},$$

which gives $\Delta_k \gtrsim \varepsilon_H^{1/\alpha}$. Therefore, Lemma 5 still holds using an MEO, perhaps with different constants.

For Lemma 6, if η_k is returned by an MEO, by (TC.E), we directly have

$$m_{x_k}(0) - m_{x_k}(\eta_k) = -\Delta_k \left\langle g_k, \eta^{\mathrm{E}} \right\rangle - \frac{\Delta_k^2}{2} \left\langle \eta^{\mathrm{E}}, H_k \eta^{\mathrm{E}} \right\rangle \geqslant \frac{1}{4}\varepsilon_H \Delta_k^2.$$

Therefore, Lemma 6 also holds for MEO. In summary, the iteration complexity of RTR in Corollary 3 remains valid when using Algorithm 5, as does the operation complexity in Corollary 5.