Inexact Cubic Regularization Method with Adaptive Reuse of Hessian Approximations

Vilmar G. Filho * Max L. N. Gonçalves *

May 14, 2026

Abstract

This work introduces an inexact cubic regularization method with adaptive reuse of Hessian approximations to solve general non-convex optimization problems. In the proposed approach, the gradient is computed inexactly and updated at every iteration, whereas the Hessian approximation is updated at a specific iteration and then reused for m subsequent iterations (a lazy strategy), where the value of m may vary throughout the procedure. The method can be implemented either in a Hessian-free or a derivative-free manner. Implementations that approximate derivative information via finite-difference schemes are discussed. We provide iteration-complexity guarantees showing that the method reaches an approximate critical point. We also establish bounds on the total gradient and function evaluations required, including the case in which only function values are used. Numerical experiments are reported to illustrate the behavior of the proposed method and to compare its performance with existing lazy cubic algorithms.

keywords: Cubic regularization method; Lazy strategies; Iteration-complexity analysis; Non-convex optimization

AMS subject classifications: 49M15, 49M37, 65K05, 90C26.

1 Introduction

In this paper, we focus on the following general non-convex optimization problem:

$$\min_{x\in Q} F(x) := f(x) + \psi(x), \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a twice continuously differentiable function, potentially non-convex, $\psi : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is a proper, closed and convex function that may be non-differentiable, and $Q = \mathrm{Dom}(\psi)$. Our analysis will be conducted under the following assumptions:

(A1) the Hessian of f is L-Lipschitz continuous, i.e.

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq L\|x - y\|, \quad \forall x, y \in \mathbb{R}^n; \tag{2}$$

*IME, Universidade Federal de Goiás, Goiânia, GO 74001-970, Brazil. (E-mails: vilmargehlen@discente.ufg.br and maxlng@ufg.br). The work of these authors was supported in part by CAPES, FAPEG (Grant No. 202510267001610) and FAPESC (Grant No. 2024TR002238).

(A2) there exists a constant $F^\star$ such that $F(x) \geq F^\star$ for all $x \in Q$.

Among the different approaches to solve problem (1), one of the most effective in terms of complexity-iteration guarantees is the Cubic Regularization (CR) method [13,18]. This method, originally proposed for the case $\psi = 0$, modifies the classical Newton method by adding a cubic regularization term to its quadratic model, ensuring global convergence and providing stronger complexity bounds. In the setting $\psi = 0$, at each iteration, the CR method generates a new iterate x_{t+1} by solving the subproblem

$$x_{t+1} \in \arg\min_{y \in \mathbb{R}^n} M_{x_t,\sigma_t}^{g_t,B_t}(y),$$

where $\sigma_t > 0$ is a regularization parameter that controls the influence of the cubic term, and

$$M_{x,\sigma}^{g,B}(y) = f(x) + \langle g, y - x\rangle + \frac{1}{2}\langle B(y-x), y-x\rangle + \frac{\sigma}{6}\|y-x\|^3, \tag{3}$$

with g denoting an approximation to the gradient of f at x, and B a symmetric approximation to its Hessian. Under standard smoothness assumptions on f (see **(A1)**–**(A2)**), Nesterov and Polyak [18] proved that the vanilla CR method (that is, $g = \nabla f$, $B = \nabla^2 f$ and σ_t fixed and proportional to L) requires at most $\mathcal{O}(\varepsilon^{-3/2})$ iterations to generate an ε-approximate stationary point (namely, a point x_t satisfying $\|\nabla f(x_t)\| \leq \varepsilon$), where $\varepsilon > 0$ is a given accuracy tolerance. In contrast, the standard Newton method (without regularization) may require up to $\mathcal{O}(\varepsilon^{-2})$ iterations to reach the same level of accuracy [5]. Since the vanilla CR method relies on exact evaluations of the gradient and Hessian—which can be computationally demanding in many applications—several recent works have proposed inexact variants of CR, in which either or both of these quantities are computed approximately; see, for example, [2–4, 6, 7, 9, 11, 12, 15, 19, 20]. Some of these works also incorporate line-search strategies to estimate or adapt the Lipschitz constant of the Hessian, further enhancing robustness and efficiency in practice.

In this work, we introduce an inexact cubic regularization method with adaptive reuse of Hessian approximations to solve the general problem (1). In the proposed approach, the gradient is computed inexactly and updated at every iteration, whereas the Hessian approximation is updated at a specific iteration and then reused for m subsequent iterations (a lazy strategy), where the value of m may vary throughout the procedure. The method can be implemented either in a Hessian-free or a derivative-free manner. Implementations that approximate derivative information via finite-difference schemes are discussed. The new algorithm jointly adjusts the regularization parameter and the accuracy of the derivative approximations using a nonmonotone line search criterion, without requiring any prior knowledge of the Lipschitz constant L. A key advantage of the lazy strategy is its ability to reduce computational overhead by reusing the same Hessian approximation over multiple iterations, thus avoiding costly second-order updates at every step. This approach can substantially decrease the overall computational cost while retaining the main benefits of second-order information. Moreover, allowing the number m of Hessian approximation reuses to vary adaptively during the iterations adds flexibility and can lead to better overall performance. For example, in the early phase of the optimization, when the iterates are far from a stationary point, using a smaller m allows more frequent Hessian updates and, consequently, more accurate search directions. Conversely, as the algorithm approaches a stationary point, employing a larger m reduces computational effort without deteriorating convergence behavior. This adaptive adjustment of m effectively balances accuracy and efficiency, making the method both robust and computationally efficient across different stages of the optimization process.

From a theoretical perspective, first-order iteration-complexity results are established for the proposed method. Specifically, for a given tolerance $\varepsilon > 0$, it is shown that the algorithm requires at most $\mathcal{O}\left((m_{\max}+1)^{1/2}\varepsilon^{-3/2}\right)$ outer iterations to produce an ε-approximate stationary point of (1) (see Definition (1) below), where $m_{\max}$ denotes the maximum number of times a Hessian approximation is reused within any block of consecutive iterations. Furthermore, when exact gradient information is available and the Hessian approximations are computed from gradient values, the method requires at most

$$\mathcal{O}\left((n+m_{\min})(m_{\min}+1)^{-1}(m_{\max}+1)^{1/2}\varepsilon^{-3/2} + n\log_2(m_{\max}+1)\right)$$

gradient and function evaluations to reach an ε-approximate critical point, where $m_{\min}$ denotes the minimum number of times a Hessian approximation is reused within any block of consecutive iterations, and n is the dimension of the problem. If the algorithm is implemented using only function values, the bound becomes

$$\mathcal{O}\left((n^2+n(m_{\min}+1))(m_{\min}+1)^{-1}(m_{\max}+1)^{1/2}\varepsilon^{-3/2} + n^2\log_2(m_{\max}+1)\right)$$

function evaluations.

From a practical perspective, numerical experiments are conducted to assess the effectiveness of the proposed approach. The method is tested on a set of 35 problems from the Moré–Garbow–Hillstrom collection [17]. The results demonstrate that the adaptive reuse of Hessian approximations can significantly reduce computational cost without compromising accuracy. In particular, the proposed strategy outperforms existing lazy cubic regularization schemes in [9].

Previous related works. We note that CR methods with lazy strategies were recently proposed in [8, 9], representing significant advances in the design of practical cubic regularization algorithms. The work in [8], in particular, introduces a CR method using exact derivative information and lazy Hessian updates to solve problem (1) with $\psi = 0$, and discusses iteration-complexity bounds for this method. The subsequent work [9] extended this framework by developing first- and zeroth-order implementations of CR methods with lazy approximated Hessians for the general composite problem (1). Our proposed algorithm shares some features with these approaches but also introduces some key differences (see Remark 2 below). In short, our method adaptively adjusts the number of Hessian approximation reuses during the iterations, providing additional flexibility that can enhance overall performance. Moreover, it accommodates different finite-difference schemes (forward, backward, and central) for approximating derivatives, whereas the methods in [9] employ only some of these approaches. Another distinction lies in the choice of the finite-difference parameter h: while in [9] this parameter depends explicitly on ε—which may lead to excessively small and potentially unstable steps—our framework allows a more robust and practical selection. Finally, the acceptance criteria for new iterates differ substantially. Our algorithm employs a nonmonotone condition (see (17) and (19)), whereas the methods in [9] rely on conditions tied to the prescribed accuracy ε. From an analytical standpoint, the iteration-complexity analysis of our method is more challenging, as it must account for the possibility of a variable number of Hessian reuses throughout the procedure. Nonetheless, when this number is fixed, our complexity bounds are similar to those obtained in [9] (see Remarks (4) and (6)).

Organization of the paper. Section 2 introduces the main definitions and preliminary results, with particular attention to finite-difference schemes for approximating the gradient and Hessian of f. Section 3 formally describes the proposed algorithm for solving problem (1) and presents its

iteration-complexity results, whose proofs are postponed to Subsection 3.1. Section 4 reports the numerical experiments conducted to evaluate the performance of the method. Finally, concluding remarks are provided in Section 5.

2 Preliminary

In this section, we introduce some basic definitions and preliminary results that will be used throughout the paper. We focus primarily on techniques for approximating the gradient and Hessian of the smooth function f using finite-difference schemes.

We begin by observing that assumption **(A1)** implies (see [18, Lemma 1]) that

$$\|\nabla f(y) - \nabla f(x) - \nabla^2 f(x)(y-x)\| \leq \frac{L}{2}\|y-x\|^2, \quad \forall\, x, y \in \mathbb{R}^n, \tag{4}$$

and

$$\left| f(y) - f(x) - \langle \nabla f(x), y-x\rangle - \frac{1}{2}\langle \nabla^2 f(x)(y-x), y-x\rangle \right| \leq \frac{L}{6}\|y-x\|^3, \quad \forall\, x, y \in \mathbb{R}^n. \tag{5}$$

We next present the concept of approximate first-order stationary points of (1).

Definition 1. *Given $\varepsilon > 0$, we say that $\bar{x} \in \mathbb{R}^n$ is an ε-approximate (first-order) stationary or critical point of (1) if $\|\nabla f(\bar{x}) + \psi'(\bar{x})\| \leq \varepsilon$, for some $\psi'(\bar{x}) \in \partial\psi(\bar{x})$, where $\partial\psi$ denotes the subdifferential of ψ.*

In what follows, we present finite-difference techniques for approximating the gradient and Hessian of f, beginning with the gradient approximations.

Lemma 1. *Suppose that **(A1)** holds. Given $x \in \mathbb{R}^n$ and $h > 0$, define $g(x) \in \mathbb{R}^n$ as*

$$g(x) = \left(\frac{f(x+he_1) - f(x-he_1)}{2h}, \ldots, \frac{f(x+he_n) - f(x-he_n)}{2h} \right). \tag{6}$$

Then,

$$\|g(x) - \nabla f(x)\| \leq \frac{\sqrt{n}L}{6}h^2.$$

Proof. See [9, Lemma 5] or [10, Lemma 3.9]. □

We next introduce an alternative gradient approximation approach, which requires an extra assumption that holds when ∇f is L_1-Lipschitz continuous. The proof of the next lemma can be found, for instance, in [10, Lemma 3.10].

Lemma 2. *Suppose that there exists $L_1 > 0$ such that*

$$\|f(y) - f(x) - \nabla f(x)(y-x)\| \leq \frac{L_1}{2}\|y-x\|^2, \quad \forall\, x, y \in \mathbb{R}^n. \tag{7}$$

Given $x \in \mathbb{R}^n$ and $h > 0$, define $g(x) \in \mathbb{R}^n$ as

$$g(x) = \left(\frac{f(x+he_1) - f(x)}{h}, \ldots, \frac{f(x+he_n) - f(x)}{h} \right),$$

or

$$g(x) = \left(\frac{f(x + he_1) - f(x - he_1)}{2h}, \ldots, \frac{f(x + he_n) - f(x - he_n)}{2h} \right),$$

or

$$g(x) = \left(\frac{f(x) - f(x - he_1)}{h}, \ldots, \frac{f(x) - f(x - he_n)}{h} \right).$$

Then,

$$\|g(x) - \nabla f(x)\| \leq \frac{\sqrt{n} L_1}{2} h.$$

We now turn our attention to approximating the Hessian of f using either function or gradient evaluations.

Lemma 3. *Suppose that* ***(A1)*** *holds. Given $x \in \mathbb{R}^n$ and $h > 0$, define $A(x) \in \mathbb{R}^{n \times n}$ as*

$$A(x) = \left[\frac{\nabla f(x + he_1) - \nabla f(x)}{h}, \ldots, \frac{\nabla f(x + he_n) - \nabla f(x)}{h} \right], \tag{8}$$

or

$$A(x) = \left[\frac{\nabla f(x + he_1) - \nabla f(x - he_1)}{2h}, \ldots, \frac{\nabla f(x + he_n) - \nabla f(x - he_n)}{2h} \right], \tag{9}$$

or

$$A(x) = \left[\frac{\nabla f(x) - \nabla f(x - he_1)}{h}, \ldots, \frac{\nabla f(x) - \nabla f(x - he_n)}{h} \right]. \tag{10}$$

Then, the matrix

$$B(x) = \frac{1}{2} \left(A(x) + A^\top(x) \right), \tag{11}$$

satisfies

$$\|B(x) - \nabla^2 f(x)\| \leq \frac{\sqrt{n} L}{2} h.$$

Proof. See [12, Lemma 3] or [10, Lemma 3.12]. □

We conclude this section by discussing approaches to approximate the Hessian of f using only function evaluations. The proof of the next lemma can be found, for instance, in [9, Lemma 6] and [10, Lemma 3.14]).

Lemma 4. *Suppose that* ***(A1)*** *holds. Given $x \in \mathbb{R}^n$ and $h > 0$, define $A(x) \in \mathbb{R}^{n \times n}$ as*

$$A_{ij}(x) = \frac{f(x + he_i + he_j) - f(x + he_i) - f(x + he_j) + f(x)}{h^2}, \tag{12}$$

or

$$A_{ij}(x) = \frac{f(x + h(e_i + e_j)) - f(x + h(e_i - e_j)) - f(x + h(e_j - e_i)) + f(x - h(e_i + e_j))}{4h^2}, \tag{13}$$

for $i, j = 1, \ldots, n$. Then, the matrix

$$B(x) = \frac{1}{2} \left(A(x) + A^\top(x) \right), \tag{14}$$

satisfies

$$\|B(x) - \nabla^2 f(x)\| \leq \frac{(1 + \sqrt{2}) n L h}{3}.$$

3 Inexact CR method with adaptive Hessian reuse

In this section, we formally present the cubic regularization method with inexact derivative information and adaptive reuse of Hessian approximations for solving (1). We then state its iteration-complexity results, whose proofs are postponed to Subsection 3.1.

Algorithm 1. Inexact CR method with adaptive reuse of Hessian approximations

Step 0. Choose $x_{-1}, x_0 \in Q$, $\sigma_0 > 0$, $\theta \geq 0$, $\bar{\kappa}_g \geq 0$, $\bar{\kappa}_B \geq 0$, and $m_{\min}, m_{\max} \in \mathbb{N}$ such that $0 \leq m_{\min} \leq m_{\max}$. Set $t := 0$, $\tau := 0$, and $m_0 := 0$.

Step 1. Set $\tau := \tau + 1$ and choose $m_\tau \in [m_{\min}, m_{\max}]$. Find the smallest integer $i \geq 0$ such that $2^{i-1}\sigma_t \geq \sigma_0(m_{\max}+1)$.

Step 1.1. Construct a vector $g_{t,i}$ and a symmetric matrix $B_{t,i}$ such that

$$\|g_{t,i} - \nabla f(x_t)\| \leq \frac{\bar{\kappa}_g}{2^{i-1}}\|x_t - x_{t-1}\|^2, \qquad \|B_{t,i} - \nabla^2 f(x_t)\| \leq \frac{\bar{\kappa}_B}{2^{i-1}}\|x_t - x_{t-1}\|. \tag{15}$$

Step 1.2. Compute $x_{t,i}^+$ such that

$$M_{x_t,2^i\sigma_t}^{g_{t,i},B_{t,i}}(x_{t,i}^+) + \psi(x_{t,i}^+) \leq F(x_t), \quad \left\|\nabla M_{x_t,2^i\sigma_t}^{g_{t,i},B_{t,i}}(x_{t,i}^+) + \psi'(x_{t,i}^+)\right\| \leq \theta\|x_{t,i}^+ - x_t\|^2, \tag{16}$$

for some $\psi'(x_{t,i}^+) \in \partial\psi(x_{t,i}^+)$, where $M_{x,\sigma}^{g,B}(\cdot)$ is as in (3).

Step 1.3. Set $\gamma_1 := m_\tau + 1$ and $\gamma_2 := 1$ if $\tau = 1$, and $\gamma_1 := 12$ and $\gamma_2 := m_\tau + 1$ otherwise. If

$$F(x_t) - F(x_{t,i}^+) \geq \frac{2^i\sigma_t}{12}\|x_{t,i}^+ - x_t\|^3 - \frac{\sigma_0(\gamma_1\gamma_2+1)}{8\gamma_1}\|x_t - x_{t-1}\|^3, \tag{17}$$

set $i_t := i$ and go to Step 2. Otherwise, set $i := i + 1$ and go to Step 1.1.

Step 2. Set $x_{t+1} := x_{t,i_t}^+$, $\sigma_{t+1} := 2^{i_t-1}\sigma_t$, $B_t := B_{t,i_t}$, and $t := t + 1$.

Step 3. If $m_\tau > 0$, let $\hat{t} := t$, $B := B_{\hat{t}-1}$ and go to Step 4. Otherwise, return to Step 1.

Step 4. Find the smallest integer $j \geq 0$ such that $2^{j-1}\sigma_t \geq \sigma_0(m_{\max}+1)$.

Step 4.1 Construct $g_{t,j}$ such that

$$\|g_{t,j} - \nabla f(x_t)\| \leq \frac{\bar{\kappa}_g}{2^{j-1}}\|x_t - x_{t-1}\|^2. \tag{18}$$

Step 4.2. Compute $x_{t,j}^+$ such that

$$M_{x_t,2^j\sigma_t}^{g_{t,j},B}(x_{t,j}^+) + \psi(x_{t,j}^+) \leq F(x_t), \quad \left\|\nabla M_{x_t,2^j\sigma_t}^{g_{t,j},B}(x_{t,j}^+) + \psi'(x_{t,j}^+)\right\| \leq \theta\|x_{t,j}^+ - x_t\|^2,$$

for some $\psi'(x_{t,j}^+) \in \partial\psi(x_{t,j}^+)$, where $M_{x,\sigma}^{g,B}(\cdot)$ is as in (3).

Step 4.3. If

$$\begin{aligned} F(x_t) - F(x_{t,j}^+) \geq{}& \frac{2^j\sigma_t}{12}\|x_{t,j}^+ - x_t\|^3 - \frac{\sigma_0}{32(m_\tau+1)m_\tau}\|x_t - x_{\hat{t}-1}\|^3 - \frac{\sigma_0}{8\gamma_1}\|x_{\hat{t}-1} - x_{\hat{t}-2}\|^3 \\ &- \frac{\sigma_0\gamma_2}{8}\|x_t - x_{t-1}\|^3, \end{aligned} \tag{19}$$

set $j_t := j$ and go to Step 5. Otherwise, set $j := j + 1$ and go to Step 4.1.

Step 5. Set $x_{t+1} := x_{t,j_t}^+$, $\sigma_{t+1} := 2^{j_t-1}\sigma_t$, and $t := t + 1$.

Step 6. If $t < \hat{t} + m_\tau$, return to Step 4. Otherwise, return to Step 1.

Remark 1. *(i) Note that τ denotes a block of consecutive iterations that share the same Hessian approximation, and m_τ indicates the number of times this Hessian approximation is reused within block τ. For instance, if $\tau = 1$ (the first block), then B_0 is an approximation of $\nabla^2 f(x_0)$, and this same approximation is used for the subsequent m_1 iterations. In this case, the sequence $\{x_t\}_{t=1}^{m_1+1}$ corresponds to the first block. Similarly, the sequence $\{x_t\}_{t=m_1+2}^{m_1+m_2+2}$ corresponds to the second block, with a fixed Hessian approximation B_{m_1+1} of $\nabla^2 f(x_{m_1+1})$. This process continues block by block, with the Hessian approximation updated at the beginning of each new block.*

(ii) Implementations of Algorithm 1, in which the derivatives are computed using finite-difference schemes satisfying (15)*, will be discussed in details in Corollaries* (6) *and* (7)*, and Remarks 3 and 5.*

(iii) Note that $x_{t,i}^+$ as in Step 1.2 *is an inexact solution of the problem*

$$\min_{y\in Q}\left(M_{x_t,2^i\sigma_t}^{g_{t,i},B_{t,i}}(y)+\psi(y)=f(x_t)+\langle g_{t,i},y-x_t\rangle+\frac{1}{2}\langle B_{t,i}(y-x_t),y-x_t\rangle+\frac{2^i\sigma_t}{6}\|y-x_t\|^3+\psi(y)\right).$$

Conditions in (16) *imply that $x_{t,i}^+$ yields a decrease in the cubic regularized model plus the function ψ, and that it is an approximate first-order stationary point of the above problem.*

(iv) Note that the sequence of parameters $\{\sigma_t\}$ can be nonmonotone. Indeed, if $i_t = 0$, then $\sigma_{t+1} = 2^{i_t-1}\sigma_t = \sigma_t/2 < \sigma_t$. Moreover, both conditions (17) *and* (19) *allow the acceptance of a trial point $x_{t,i}^+$ satisfying $F(x_{t,i}^+) > F(x_t)$. Consequently, the sequence $\{F(x_t)\}_{t\geq 0}$ may also be nonmonotone.*

Remark 2. *As discussed in the introduction, our algorithm shares certain similarities with the approaches proposed in the recent work [9], but also exhibits some key differences. First, our method adaptively adjusts the number of Hessian-approximation reuses m_τ within each block of iterations, providing additional flexibility that can enhance overall performance. Second, the inequalities in* (15) *can accommodate different finite-difference schemes (forward, backward, and central) for approximating derivatives, whereas the methods in [9] employ only* (8) *for the first-order method, and* (6) *and* (12) *for the zero-order method. Third, the choice of the finite-difference parameter h in [9] depends explicitly on the target accuracy ε, while in our implementations of the algorithm this parameter does not depend on ε and can instead be controlled by freely chosen constants (κ_g and κ_B); see Corollaries 6 and 7, and Remarks 3 and 5. Finally, our acceptance criteria for new iterates (conditions* (17) *and* (19)*) are based on a nonmonotone strategy, whereas the methods in [9] rely on conditions explicitly tied to the prescribed accuracy ε. We also note that, in addition to allowing for different finite-difference Hessian approximations, the proposed algorithm with $\bar\kappa_g = 0$, $m_{\max} = m_{\min} = 0$ and $\psi = 0$ closely resembles the CR method with finite-difference Hessian approximations proposed in [12].*

We next discuss the iteration-complexity bounds of Algorithm 1, with the corresponding proofs presented in the next section. From now on, let $\{x_t\}_{t=1}^T$ be the sequence generated by Algorithm 1. We begin by establishing a bound in terms of outer iterations.

Theorem 5. *Suppose that **(A1)** and **(A2)** hold. Then,*

$$\sum_{t=0}^{T-1}\|\nabla f(x_{t+1})+\psi'(x_{t+1})\|^{\frac{3}{2}}\leq\frac{48\bar\lambda(F(x_0)-F^\star)}{\sigma_0}+6\left(2\bar\lambda+\bar\kappa_B^{\frac{3}{2}}+\bar\kappa_g^{\frac{3}{2}}\right)\|x_0-x_{-1}\|^3, \tag{20}$$

where

$$\bar\lambda:=\frac{2(m_{\max}+1)^{\frac{1}{2}}\left[10L^{\frac{3}{2}}+16\bar\kappa_B^{\frac{3}{2}}+\sigma_0^{\frac{3}{2}}+(\theta+4L+3\bar\kappa_B)\sigma_0^{\frac{1}{2}}+114\bar\kappa_g^3\sigma_0^{-\frac{3}{2}}\right]^{\frac{3}{2}}}{\sigma_0^{\frac{3}{4}}}+\frac{4\sqrt{2}\bar\kappa_g^{\frac{3}{2}}}{(m_{\max}+1)}. \tag{21}$$

As a consequence, given $\varepsilon > 0$, Algorithm 1 needs at most $\mathcal{O}\left((m_{\max}+1)^{\frac{1}{2}}\varepsilon^{-\frac{3}{2}}\right)$ outer iterations to generate an $\varepsilon-$approximate critical point for problem (1)*.*

We now discuss implementations of Algorithm 1 and their iteration-complexity bounds in which exact gradient information is available and the Hessian approximations are computed from gradient values using the finite-difference expressions.

Corollary 6. *Suppose that **(A1)** and **(A2)** hold. In Steps 1.1 and 4.1, assume that $g_{t,i} = g_{t,j} := \nabla f(x_t)$ and $B_{t,i} := B(x_t)$, where $B(x)$ is as in (11) and $A(x)$ is specified in either (8) or (10), with*

$$h := \frac{2\kappa_B}{2^{i-1}\sqrt{n}}\|x_t - x_{t-1}\|, \tag{22}$$

for some constant $\kappa_B > 0$. Then, the inequalities in (15) and (18) hold trivially with $\bar{\kappa}_g := 0$ and $\bar{\kappa}_B := L\kappa_B$. Moreover, the total number of function and gradient evaluations up to the T-th iteration, denoted by $FGE(T)$, is bounded as follows:

$$FGE(T) \leq \frac{T(n+1)}{m_{\min}+1}\left[\log_2 \hat{\lambda} + 2\right] + (n+2)\log_2\left((m_{\max}+1)\hat{\lambda}\right) + 3(T+n+1), \tag{23}$$

where $\hat{\lambda} := [2L\sigma_0^{\frac{1}{2}} + 10L^{\frac{3}{2}} + 16L^{\frac{3}{2}}\kappa_B^{\frac{3}{2}} + \sigma_0^{\frac{3}{2}}]/\sigma_0^{\frac{3}{2}}$. As a consequence, given $\varepsilon > 0$, the total number of FGE required to obtain an ε-approximate critical point is

$$\mathcal{O}\left((n+m_{\min})(m_{\min}+1)^{-1}(m_{\max}+1)^{1/2}\varepsilon^{-3/2} + n\log_2(m_{\max}+1)\right).$$

Remark 3. *In particular, if in Corollary 6 the matrix $A(x)$ is defined as in (9), with h as in (22), then the inequalities in (15) hold trivially with $\bar{\kappa}_g := 0$ and $\bar{\kappa}_B := L\kappa_B$. Consequently, the bound in (23) becomes*

$$FGE(T) \leq \frac{T(2n+1)}{m_{\min}+1}\left[\log_2 \hat{\lambda} + 2\right] + 2(n+1)\log_2\left((m_{\max}+1)\hat{\lambda}\right) + 3(T+2n+1).$$

Remark 4. *Considering the first-order (Hessian-free) implementation of Algorithm 1 described in Corollary 6, and taking $m_{\min} = m_{\max} = m$, our iteration-complexity bounds reduce to*

$$\mathcal{O}\Big((m+1)^{1/2}\varepsilon^{-3/2}\Big) \quad \text{and} \quad \mathcal{O}\Big((n+m)(m+1)^{-1/2}\varepsilon^{-3/2} + n\log_2(m+1)\Big)$$

in terms of outer iterations and FGE, respectively. These bounds are comparable, with respect to the dependence on ε, n and m, to those established in [9, Theorem 2] for the first-order CN method [9, Algorithm 2]. On the other hand, when $m_{\min} = m_{\max} = 0$ and $\psi = 0$, our bounds are consistent with those in [12, Theorem 2 and Corollary 1], derived for the CR method with finite-difference Hessian approximations. As noted in [9], choosing $m = n$ yields an improved FGE bound by a factor of $\sqrt{n}$ compared with the case $m_{\min} = m_{\max} = 0$.

We next discuss implementations of Algorithm 1, along with its iteration-complexity bounds, in which derivative approximations are computed from function values using finite-difference schemes.

Corollary 7. *Suppose that **(A1)** and **(A2)** hold. In Steps 1.1 and 4.1, assume that $g_{t,i} = g_{t,j} = g(x_t)$ and $B_{t,i} := B(x_t)$, where $g(x)$ is as in (6) and $B(x)$ is as in (14) and $A(x)$ is specified in either (12) or (13), with*

$$h := \min\left\{\left(\frac{6\kappa_g}{\sqrt{n}2^{i-1}}\right)^{\frac{1}{2}}, \frac{3\kappa_B}{(1+\sqrt{2})n2^{i-1}}\right\}\|x_t - x_{t-1}\|,$$

for some constants $\kappa_g, \kappa_B > 0$. Then, the inequalities in (15) and (18) (for the latter, taking $h := (6\kappa_g\|x_t - x_{t-1}\|^2/(\sqrt{n}2^{j-1}))^{1/2}$) hold trivially with $\bar{\kappa}_g := L\kappa_g$ and $\bar{\kappa}_B := L\kappa_B$. Moreover, the total number of function evaluations up to the T-th iteration, denoted by $FE(T)$, is bounded as follows:

$$FE(T) \leq \frac{4n^2T}{m_{\min}+1}\left[\log_2 \lambda + 2\right] + 2(2n^2+n+1)\log_2\left((m_{\max}+1)\lambda\right) + 4((n+1)T + 2n^2), \tag{24}$$

where $\lambda := [2L\sigma_0^{\frac{1}{2}} + 10L^{\frac{3}{2}} + 16L^{\frac{3}{2}}\kappa_B^{\frac{3}{2}} + \sigma_0^{\frac{3}{2}} + 114L^3\kappa_g^3\sigma_0^{-\frac{3}{2}}]/\sigma_0^{\frac{3}{2}}$. *As a consequence, given* $\varepsilon > 0$, *the total number of FE required to obtain an* ε*-approximate critical point is*

$$\mathcal{O}\Big((n^2 + n(m_{\min}+1))(m_{\min}+1)^{-1}(m_{\max}+1)^{1/2}\varepsilon^{-3/2} + n^2\log_2(m_{\max}+1)\Big).$$

Remark 5. *If* f *also satisfies* (7) *and* $g(x)$ *in Corollary 7 is defined according to one of the three options in Lemma 2, with*

$$h := \min\left\{\frac{2\kappa_g\|x_t - x_{t-1}\|}{\sqrt{n}\,2^{i-1}}, \frac{3\kappa_B}{(1+\sqrt{2})\,n\,2^{i-1}}\right\}\|x_t - x_{t-1}\|,$$

then the inequalities in (15) *hold trivially with* $\bar{\kappa}_g := L_1\kappa_g$ *and* $\bar{\kappa}_B := L\kappa_B$. *Consequently, the iteration-complexity results in Corollary 7 remain valid.*

Remark 6. *For the zero-order implementation of Algorithm 1 described in Corollary 7, and setting* $m_{\min} = m_{\max} = m$, *the iteration-complexity bounds simplify to*

$$\mathcal{O}\Big((m+1)^{1/2}\varepsilon^{-3/2}\Big) \quad \text{and} \quad \mathcal{O}\Big((n^2+mn)(m+1)^{-1/2}\varepsilon^{-3/2} + n^2\log_2(m+1)\Big)$$

in terms of outer iterations and function evaluations (FE), respectively. These bounds are comparable, in their dependence on ε, n, *and* m, *to those established in [9, Theorem 4] for the zero-order CN method [9, Algorithm 4].*

3.1 Proofs of Theorem 5 and Corollaries 6 and 7

We now proceed with the proofs of Theorems 5, and Corollaries 6 and 7. To this end, we first recall the Young's inequality (see [16]): given positive numbers a, b, p, q satisfying $p, q > 1$ and $1/p + 1/q = 1$, we have

$$ab \le \frac{a^p}{p} + \frac{b^q}{q}. \tag{25}$$

For clarity, in the proofs of the next two results, we omit the index t from the iterates generated by Algorithm 1. The following proposition is crucial in establishing that the inner procedures of the algorithm terminate in a finite number of iterations.

Proposition 8. *Let* $\hat{\kappa}_g, \hat{\kappa}_B \ge 0$, $\sigma > 0$ *and* $x, z, \hat{x}, \hat{z} \in \mathbb{R}^n$ *be given. Assume that* $x^+, g \in \mathbb{R}^n$ *and* $B \in \mathbb{R}^{n\times n}$ *satisfy*

$$\|g - \nabla f(x)\| \le \hat{\kappa}_g\|x - \hat{x}\|^2, \qquad \|B - \nabla^2 f(z)\| \le \hat{\kappa}_B\|z - \hat{z}\| \tag{26}$$

and

$$M_{x,\sigma}^{g,B}(x^+) + \psi(x^+) \le F(x). \tag{27}$$

If

$$\sigma \ge 2(L + \sqrt{2}L^{\frac{3}{2}}\hat{\rho}^{\frac{1}{2}} + \sqrt{2}\hat{\kappa}_B^{\frac{3}{2}}\bar{\rho}^{\frac{1}{2}} + 2\hat{\kappa}_g^3\tilde{\rho}^2), \tag{28}$$

for some $\hat{\rho}, \bar{\rho}, \tilde{\rho} > 0$, *then*

$$F(x) - F(x^+) \ge \frac{\sigma}{12}\|x^+ - x\|^3 - \frac{1}{3\hat{\rho}}\|x - z\|^3 - \frac{1}{3\bar{\rho}}\|z - \hat{z}\|^3 - \frac{2}{3\tilde{\rho}}\|x - \hat{x}\|^3.$$

Proof. From (5) and definition of $M_{x,\sigma}^{g,B}(\cdot)$ in (3), we have

$$\begin{aligned} f(x^+) &\le f(x) + \langle\nabla f(x), x^+ - x\rangle + \frac{1}{2}\langle\nabla^2 f(x)(x^+ - x), x^+ - x\rangle + \frac{L}{6}\|x^+ - x\|^3 \\ &= M_{x,\sigma}^{g,B}(x^+) + \langle\nabla f(x) - g, x^+ - x\rangle + \frac{1}{2}\langle(\nabla^2 f(x) - B)(x^+ - x), x^+ - x\rangle + \frac{L-\sigma}{6}\|x^+ - x\|^3. \end{aligned}$$

The last inequality, combined with (2), (26), (27), $F = f + \psi$ and the Cauchy-Schwarz inequality, yields

$$\begin{aligned} F(x^+) &\leq F(x) + \|\nabla f(x) - g\|\|x^+ - x\| + \frac{1}{2}\|\nabla^2 f(x) - \nabla^2 f(z)\|\|x^+ - x\|^2 + \frac{1}{2}\|B - \nabla^2 f(z)\|\|x^+ - x\|^2 \\ &\quad + \frac{L-\sigma}{6}\|x^+ - x\|^3 \\ &\leq F(x) + \hat{\kappa}_g\|x - \hat{x}\|^2\|x^+ - x\| + \frac{L}{2}\|x - z\|\|x^+ - x\|^2 + \frac{\hat{\kappa}_B}{2}\|z - \hat{z}\|\|x^+ - x\|^2 + \frac{L-\sigma}{6}\|x^+ - x\|^3. \end{aligned}$$

From the inequality in 25 with $p = 3/2$, $q = 3$, $a = \|x - \hat{x}\|^2/\tilde{\rho}^{2/3}$ and $b = \hat{\kappa}_g\tilde{\rho}^{2/3}\|x^+ - x\|$, we get

$$\hat{\kappa}_g\|x - \hat{x}\|^2\|x^+ - x\| = \frac{\|x - \hat{x}\|^2}{\tilde{\rho}^{\frac{2}{3}}}\hat{\kappa}_g\tilde{\rho}^{\frac{2}{3}}\|x^+ - x\| \leq \frac{2}{3\tilde{\rho}}\|x - \hat{x}\|^3 + \frac{\hat{\kappa}_g^3\tilde{\rho}^2}{3}\|x^+ - x\|^3.$$

Again, from the inequality in (25) with $p = 3, q = 3/2,\ a = \|x - z\|/\hat{\rho}^{\frac{1}{3}}$ and $b = L\hat{\rho}^{\frac{1}{3}}\|x^+ - x\|^2/2$, we obtain

$$\frac{L}{2}\|x - z\|\|x^+ - x\|^2 = \frac{\|x - z\|}{\hat{\rho}^{\frac{1}{3}}}\frac{L\hat{\rho}^{\frac{1}{3}}}{2}\|x^+ - x\|^2 \leq \frac{1}{3\hat{\rho}}\|x - z\|^3 + \frac{\sqrt{2}L^{\frac{3}{2}}\hat{\rho}^{\frac{1}{2}}}{6}\|x^+ - x\|^3.$$

Again, from the inequality in 25 with $p = 3$, $q = 3/2$, $a = \|z - \hat{z}\|/\bar{\rho}^{1/3}$ and $b = \hat{\kappa}_B\bar{\rho}^{1/3}\|x^+ - x\|^2/2$, we have

$$\frac{\hat{\kappa}_B}{2}\|z - \hat{z}\|\|x^+ - x\|^2 = \frac{\|z - \hat{z}\|}{\bar{\rho}^{\frac{1}{3}}}\frac{\hat{\kappa}_B}{2}\bar{\rho}^{\frac{1}{3}}\|x^+ - x\|^2 \leq \frac{1}{3\bar{\rho}}\|z - \hat{z}\|^3 + \frac{\sqrt{2}\hat{\kappa}_B^{\frac{3}{2}}\bar{\rho}^{\frac{1}{2}}}{6}\|x^+ - x\|^3.$$

It follows from the last four inequalities that

$$F(x^+) \leq F(x) + \left(\frac{L + \sqrt{2}L^{\frac{3}{2}}\hat{\rho}^{\frac{1}{2}} + \sqrt{2}\hat{\kappa}_B^{\frac{3}{2}}\bar{\rho}^{\frac{1}{2}} + 2\hat{\kappa}_g^3\tilde{\rho}^2 - \sigma}{6}\right)\|x^+ - x\|^3 + \frac{\|x - z\|^3}{3\hat{\rho}} + \frac{\|z - \hat{z}\|^3}{3\bar{\rho}} + \frac{2\|x - \hat{x}\|^3}{3\tilde{\rho}},$$

which, combined with (28), implies the desired inequality. □

The next proposition establishes a bound on $\|\nabla f(x^+) + \psi'(x^+)\|$ under certain conditions on x^+, g and B.

Proposition 9. *Assume that $g \in \mathbb{R}^n$ and $B \in \mathbb{R}^{n\times n}$ satisfy (26) for some $\hat{\kappa}_g, \hat{\kappa}_B \geq 0$ and $x, z, \hat{x}, \hat{z} \in \mathbb{R}^n$. Moreover, suppose that $x^+ \in \mathbb{R}^n$ satisfies*

$$\|\nabla M_{x,\sigma}^{g,B}(x^+) + \psi'(x^+)\| \leq \theta\|x^+ - x\|^2, \tag{29}$$

for some $\psi'(x^+) \in \partial\psi(x^+)$, $\theta \geq 0$ and $\sigma > 0$. If $\rho, \rho^\star > 0$, then

$$\|\nabla f(x^+) + \psi'(x^+)\|^{\frac{3}{2}} \leq \frac{1}{\sqrt{2}}\left[\eta^{\frac{3}{2}}\|x^+ - x\|^3 + (2\hat{\kappa}_g)^{\frac{3}{2}}\|x - \hat{x}\|^3 + \frac{L^{\frac{3}{2}}\|x - z\|^3}{\rho^{\frac{3}{2}}} + \frac{\hat{\kappa}_B^{\frac{3}{2}}\|z - \hat{z}\|^3}{\rho^{\star\frac{3}{2}}}\right], \tag{30}$$

where $\eta := \sigma + L + L\rho + \hat{\kappa}_B\rho^\star + 2\theta$.

Proof. From the definition of $M_{x,\sigma}^{g,B}(\cdot)$ in (3), we have

$$\nabla M_{x,\sigma}^{g,B}(y) = g + B(y - x) + \frac{\sigma}{2}\|y - x\|(y - x).$$

Hence, using (29) and the triangle inequality, we obtain

$$\begin{aligned} \|\nabla f(x^+) + \psi'(x^+)\| &\leq \|\nabla f(x^+) - \nabla M_{x,\sigma}^{g,B}(x^+)\| + \|\nabla M_{x,\sigma}^{g,B}(x^+) + \psi'(x^+)\| \\ &\leq \|\nabla f(x^+) - g - B(x^+ - x)\| + \frac{\sigma}{2}\|x^+ - x\|^2 + \theta\|x^+ - x\|^2 \\ &\leq \|\nabla f(x^+) - \nabla f(x) - \nabla^2 f(x)(x^+ - x)\| + \|(\nabla^2 f(x) - \nabla^2 f(z))(x^+ - x)\| \\ &\quad + \|(\nabla^2 f(z) - B)(x^+ - x)\| + \|\nabla f(x) - g\| + \left(\frac{\sigma}{2} + \theta\right)\|x^+ - x\|^2. \end{aligned}$$

From last inequality, (2), (4) and (26), we get

$$\|\nabla f(x^+) + \psi'(x^+)\| \leq \left(\frac{L+\sigma}{2} + \theta\right)\|x^+ - x\|^2 + L\|x - z\|\|x^+ - x\| + \hat{\kappa}_B\|z - \hat{z}\|\|x^+ - x\| + \hat{\kappa}_g\|x - \hat{x}\|^2.$$

On the other hand, it follows from the inequality in (25) with $p = q = 2$, $a = \|x - z\|/\rho^{1/2}$ and $b = \rho^{1/2}\|x^+ - x\|$ that

$$\|x - z\|\|x^+ - x\| \leq \frac{\|x - z\|^2}{2\rho} + \frac{\rho\|x^+ - x\|^2}{2}.$$

Again, by the same inequality with $p = q = 2$, $a = \|z - \hat{z}\|/\rho^{\star\frac{1}{2}}$ and $b = \rho^{\star\frac{1}{2}}\|x^+ - x\|$, we have

$$\|z - \hat{z}\|\|x^+ - x\| \leq \frac{\|z - \hat{z}\|^2}{2\rho^\star} + \frac{\rho^\star\|x^+ - x\|^2}{2}.$$

Combining the last three inequalities, we find that

$$\|\nabla f(x^+) + \psi'(x^+)\|^{\frac{3}{2}} \leq 2^{\frac{3}{2}}\left[\frac{\eta}{4}\|x^+ - x\|^2 + \frac{2\hat{\kappa}_g}{4}\|x - \hat{x}\|^2 + \frac{L}{4\rho}\|x - z\|^2 + \frac{\hat{\kappa}_B}{4\rho^\star}\|z - \hat{z}\|^2\right]^{\frac{3}{2}}.$$

where $\eta = \sigma + L + L\rho + \hat{\kappa}_B\rho^\star + 2\theta$. Since the function $t \mapsto t^{3/2}$ is convex for $t \geq 0$, it follows from the Jensen's inequality that $((t_1 + t_2 + t_3 + t_4)/4)^{3/2} \leq (t_1^{3/2} + t_2^{3/2} + t_3^{3/2} + t_4^{3/2})/4$ for all $t_1, t_2, t_3, t_4 \geq 0$. Hence,

$$\|\nabla f(x^+) + \psi'(x^+)\|^{\frac{3}{2}} \leq \frac{1}{\sqrt{2}}\left[\eta^{\frac{3}{2}}\|x^+ - x\|^3 + 2^{\frac{3}{2}}\hat{\kappa}_g^{\frac{3}{2}}\|x - \hat{x}\|^3 + \frac{L^{\frac{3}{2}}\|x - z\|^3}{\rho^{\frac{3}{2}}} + \frac{\hat{\kappa}_B^{\frac{3}{2}}\|z - \hat{z}\|^3}{\rho^{\star\frac{3}{2}}}\right],$$

which implies (30). □

We next prove that the sequence of parameters $\{\sigma_t\}$ is bounded from above. In particular, we show that the inner procedures in Steps 1 and 4 of the algorithm end in a finite number of trials.

Lemma 10. *The regularization parameters σ_t in Algorithm 1 satisfies*

$$\sigma_0(m_{max} + 1) \leq \sigma_t \leq \sigma_{\max} := 2L + \frac{(m_{\max} + 1)^{\frac{1}{2}}[10L^{\frac{3}{2}}m_{\max}^{\frac{1}{2}} + 16\bar{\kappa}_B^{\frac{3}{2}} + \sigma_0^{\frac{3}{2}}(m_{\max} + 1)^{\frac{1}{2}}]}{\sigma_0^{\frac{1}{2}}} + \frac{114\bar{\kappa}_g^3}{\sigma_0^2}, \quad (31)$$

for all $t \geq 1$. As a consequence, the inner procedures in Steps 1 and 4 end in a finite number of iterations.

Proof. Let us prove by induction on t that (31) holds. For $t = 1$, by Step 1, we obtain $\sigma_0(m_{\max} + 1) \leq 2^{i_0 - 1}\sigma_0 = \sigma_1$. Now, assume by contradiction that $2^{i_0-1}\sigma_0 = \sigma_1 > \sigma_{\max}$. Hence, since $m_{\max} \geq m_1$, we have

$$\begin{aligned}
2^{i_0-1}\sigma_0 &> \sigma_0(m_{\max} + 1) + 2L + \frac{10L^{\frac{3}{2}}(m_{\max} + 1)^{\frac{1}{2}}m_{\max}^{\frac{1}{2}}}{\sigma_0^{\frac{1}{2}}} + \frac{16\bar{\kappa}_B^{\frac{3}{2}}(m_{\max} + 1)^{\frac{1}{2}}}{\sigma_0^{\frac{1}{2}}} + \frac{114\bar{\kappa}_g^3}{\sigma_0^2} \\
&> 2\left(L + \sqrt{2}L^{\frac{3}{2}}\left(\frac{32m_{\max}(m_{\max} + 1)}{3\sigma_0}\right)^{\frac{1}{2}} + \sqrt{2}\bar{\kappa}_B^{\frac{3}{2}}\left(\frac{96(m_{\max} + 1)}{3\sigma_0}\right)^{\frac{1}{2}} + 2\bar{\kappa}_g^3\left(\frac{16}{3\sigma_0}\right)^2\right) \\
&\geq 2\left(L + \sqrt{2}L^{\frac{3}{2}}\left(\frac{32m_1(m_1 + 1)}{3\sigma_0}\right)^{\frac{1}{2}} + \sqrt{2}\left(\frac{\bar{\kappa}_B}{2^{i_0-2}}\right)^{\frac{3}{2}}\left(\frac{8(m_1 + 1)}{3\sigma_0}\right)^{\frac{1}{2}} + 2\left(\frac{\bar{\kappa}_g}{2^{i_0-2}}\right)^3\left(\frac{16}{3\sigma_0}\right)^2\right),
\end{aligned} \quad (32)$$

where we used the fact that $1 \geq 1/(2^{i_0-2})^{3/2}$ and $i_0 \geq 2$ in the last inequality. Then, by inequality (32) and Proposition 8 with $\sigma = 2^{i_0-1}\sigma_0$, $\hat{\kappa}_g = \bar{\kappa}_g/2^{i_0-2}$, $\hat{\kappa}_B = \bar{\kappa}_B/2^{i_0-2}$, $x^+ = x^+_{0,i_0-1}$, $z = x = x_0$, $\hat{z} = x_{-1}$, $\hat{\rho} = 32m_1(m_1+1)/(3\sigma_0)$, $\bar{\rho} = 8(m_1+1)/(3\sigma_0)$ and $\tilde{\rho} = 16/(3\sigma_0)$ it follows that

$$F(x_0) - F(x^+_{0,i_0-1}) \geq \frac{2^{i_0-1}\sigma_0}{12}\|x^+_{0,i_0-1} - x_0\|^3 - \frac{\sigma_0}{8(m_1+1)}\|x_0 - x_{-1}\|^3 - \frac{\sigma_0}{8}\|x_0 - x_{-1}\|^3.$$

Therefore, (17) is satisfied for $i = i_0 - 1$, contradicting the minimality of i_0, which proves the inequality in (31) for $t = 1$. Now, suppose that (31) holds for some natural number $t > 1$, that is, $\sigma_0(m_{\max}+1) \leq \sigma_t \leq \sigma_{\max}$. Let us consider the case that σ_{t+1} is given in Step 5 (the proof for the case where σ_{t+1} is given in Step 2 follows with similar arguments). We divide the proof into two cases:
Case ($j_t \leq 1$): From Step 4, we obtain

$$\sigma_0(m_{\max}+1) \leq \sigma_{t+1} = 2^{j_t-1}\sigma_t \leq 2^{1-1}\sigma_t = \sigma_t \leq \sigma_{\max}.$$

Case ($j_t \geq 2$): From Step 4, we have $\sigma_{t+1} = 2^{j_t-1}\sigma_t \geq \sigma_0(m_{\max}+1)$. Now, assume by contradiction that $\sigma_{t+1} = 2^{j_t-1}\sigma_t > \sigma_{\max}$. Hence, since $m_{\max} \geq m_\tau$, $12(m_1+1) \geq \gamma_1$ and $1 \geq 1/\gamma_2$, we have

$$\begin{aligned}
2^{j_t-1}\sigma_t &> \sigma_0(m_{\max}+1) + 2L + \frac{10L^{\frac{3}{2}}(m_{\max}+1)^{\frac{1}{2}}m_{\max}^{\frac{1}{2}}}{\sigma_0^{\frac{1}{2}}} + \frac{16\bar{\kappa}_B^{\frac{3}{2}}(m_{\max}+1)^{\frac{1}{2}}}{\sigma_0^{\frac{1}{2}}} + \frac{114\bar{\kappa}_g^3}{\sigma_0^2} \\
&> 2\left(L + \sqrt{2}L^{\frac{3}{2}}\left(\frac{32m_{\max}(m_{\max}+1)}{3\sigma_0}\right)^{\frac{1}{2}} + \sqrt{2}\bar{\kappa}_B^{\frac{3}{2}}\left(\frac{96(m_1+1)}{3\sigma_0}\right)^{\frac{1}{2}} + 2\bar{\kappa}_g^3\left(\frac{16}{3\sigma_0}\right)^2\right) \\
&\geq 2\left(L + \sqrt{2}L^{\frac{3}{2}}\left(\frac{32m_\tau(m_\tau+1)}{3\sigma_0}\right)^{\frac{1}{2}} + \sqrt{2}\left(\frac{\bar{\kappa}_B}{2^{j_t-2}}\right)^{\frac{3}{2}}\left(\frac{8\gamma_1}{3\sigma_0}\right)^{\frac{1}{2}} + 2\left(\frac{\bar{\kappa}_g}{2^{j_t-2}}\right)^3\left(\frac{16}{3\sigma_0\gamma_2}\right)^2\right), \quad (33)
\end{aligned}$$

where we used the fact that $1 \geq 1/(2^{j_t-2})^{3/2}$ and $j_t \geq 2$ in the second inequality. It follows from inequality (33) and Proposition 8 with $\sigma = 2^{j_t-1}\sigma_t$, $\hat{\kappa}_g = \bar{\kappa}_g/2^{j_t-2}$, $\hat{\kappa}_B = \bar{\kappa}_B/2^{j_t-2}$, $x^+ = x^+_{t,j_t-1}$, $x = x_t$, $z = x_{\hat{t}-1}$, $\hat{z} = x_{\hat{t}-2}$, $\hat{\rho} = 32m_\tau(m_\tau+1)/(3\sigma_0)$, $\bar{\rho} = 8\gamma_1/(3\sigma_0)$ and $\tilde{\rho} = 16/(3\sigma_0\gamma_2)$ that

$$\begin{aligned}
F(x_t) - F(x^+_{t,j_t-1}) \geq{}& \frac{2^{j_t-1}\sigma_t}{12}\|x^+_{t,j_t-1} - x_t\|^3 - \frac{\sigma_0}{32m_\tau(m_\tau+1)}\|x_t - x_{\hat{t}-1}\|^3 - \frac{\sigma_0}{8\gamma_1}\|x_{\hat{t}-1} - x_{\hat{t}-2}\|^3 \\
&- \frac{\sigma_0\gamma_2}{8}\|x_t - x_{t-1}\|^3
\end{aligned}$$

Therefore, (19) is satisfied for $j = j_t - 1$, contradicting the minimality of j_t. So, $\sigma_{t+1} \leq \sigma_{\max}$, which concludes the proof of the inequality in (19). □

We now present a recursive inequality that holds in each block of Algorithm 1.

Lemma 11. *Let $\tau_T \in \mathbb{N} - \{0\}$ be the block number associated with the T-th iteration of Algorithm 1, that is,*

$$T = m_0 + m_1 + m_2 + \cdots + m_{\tau_T-1} + \tau_T - 1 + \ell_T, \quad \textit{with} \quad \ell_T \in \mathbb{N}, \quad 1 \leq \ell_T \leq m_{\tau_T} + 1.$$

Then,

$$\begin{aligned}
F(x_T) \leq{}& F(x_{a_T}) - \frac{15\sigma_0(m_{\max}+1)}{96}\sum_{t=a_T}^{T-1}\|x_{t+1} - x_t\|^3 + \frac{\sigma_0(m_{\tau_T}+1)}{8\gamma_1}\|x_{a_T} - x_{a_T-1}\|^3 \\
&+ \frac{\sigma_0\gamma_2}{8}\|x_{a_T} - x_{a_T-1}\|^3 + \frac{\sigma_0\gamma_2}{8}\sum_{t=a_T}^{T-2}\|x_{t+1} - x_t\|^3
\end{aligned}$$

where $a_T := m_0 + m_1 + m_2 + \cdots + m_{\tau_T-1} + \tau_T - 1$, $\gamma_1 := m_{\tau_T} + 1$, $\gamma_2 := 1$ if $\tau_T = 1$ and $\gamma_1 := 12$, $\gamma_2 := m_{\tau_T} + 1$ otherwise.

Proof. Since a_T+1 is the first iteration of the block τ_T, it follows from (17) and the fact that $2\sigma_{t+1} := 2^{i_t}\sigma_t$ for all $t \geq 0$, that

$$F(x_{a_T+1}) \leq F(x_{a_T}) - \frac{\sigma_{a_T+1}}{6}\|x_{a_T+1} - x_{a_T}\|^3 + \frac{\sigma_0}{8\gamma_1}\|x_{a_T} - x_{a_T-1}\|^3 + \frac{\sigma_0\gamma_2}{8}\|x_{a_T} - x_{a_T-1}\|^3.$$

For the other iterations in the block τ_T (if $m_{\tau_T} > 0$), if follows from (19) and the fact that $2\sigma_{t+1} := 2^{j_t}\sigma_t$ for all $t \geq 0$, that

$$\begin{aligned} F(x_{a_T+2}) \leq F(x_{a_T+1}) &- \frac{\sigma_{a_T+2}}{6}\|x_{a_T+2} - x_{a_T+1}\|^3 + \frac{\sigma_0}{32(m_{\tau_T}+1)m_{\tau_T}}\|x_{a_T+1} - x_{a_T}\|^3 \\ &+ \frac{\sigma_0}{8\gamma_1}\|x_{a_T} - x_{a_T-1}\|^3 + \frac{\sigma_0\gamma_2}{8}\|x_{a_T+1} - x_{a_T}\|^3, \end{aligned}$$

$$\vdots$$

$$\begin{aligned} F(x_{a_T+\ell_T}) \leq F(x_{a_T+\ell_T-1}) &- \frac{\sigma_{a_T+\ell_T}}{6}\|x_{a_T+\ell_T} - x_{a_T+\ell_T-1}\|^3 + \frac{\sigma_0}{32(m_{\tau_T}+1)m_{\tau_T}}\|x_{a_T+\ell_T-1} - x_{a_T}\|^3 \\ &+ \frac{\sigma_0}{8\gamma_1}\|x_{a_T} - x_{a_T-1}\|^3 + \frac{\sigma_0\gamma_2}{8}\|x_{a_T+\ell_T-1} - x_{a_T+\ell_T-2}\|^3. \end{aligned}$$

Summing up the above inequalities and using the first inequality in (31), we find that

$$\begin{aligned} F(x_{a_T+\ell_T}) \leq F(x_{a_T}) &- \frac{\sigma_0(m_{\max}+1)}{6}\sum_{t=a_T}^{a_T+\ell_T-1}\|x_{t+1} - x_t\|^3 + \frac{\sigma_0}{32(m_{\tau_T}+1)m_{\tau_T}}\sum_{t=a_T}^{a_T+\ell_T-1}\|x_t - x_{a_T}\|^3 \\ &+ \frac{\sigma_0\ell_T\|x_{a_T} - x_{a_T-1}\|^3}{8\gamma_1} + \frac{\sigma_0\gamma_2}{8}\|x_{a_T} - x_{a_T-1}\|^3 + \frac{\sigma_0\gamma_2}{8}\sum_{t=a_T}^{a_T+\ell_T-2}\|x_{t+1} - x_t\|^3. \end{aligned} \tag{34}$$

Now, since the function $x \mapsto x^3$ is convex for $x \geq 0$, it follows from the Jensen's inequality that, for every $s \geq 2$,

$$\begin{aligned} \left(\frac{\|x_{a_T} - x_{a_T+s}\|}{s}\right)^3 &\leq \left(\frac{\|x_{a_T} - x_{a_T+1}\| + \|x_{a_T+1} - x_{a_T+2}\| + \ldots + \|x_{a_T+s-1} - x_{a_T+s}\|}{s}\right)^3 \\ &\leq \frac{1}{s}\sum_{t=a_T+1}^{a_T+s}\|x_t - x_{t-1}\|^3. \end{aligned}$$

Hence, applying the last inequality for different values of s, we find, for all $\ell_T > 1$, that

$$\begin{aligned} \sum_{t=a_T}^{a_T+\ell_T-1}\|x_t - x_{a_T}\|^3 &= \|x_{a_T+1} - x_{a_T}\|^3 + 2^3\left(\frac{\|x_{a_T+2} - x_{a_T}\|}{2}\right)^3 + 3^3\left(\frac{\|x_{a_T+3} - x_{a_T}\|}{3}\right)^3 \\ &\quad + \ldots + (\ell_T-1)^3\left(\frac{\|x_{a_T+\ell_T-1} - x_{a_T}\|}{\ell_T-1}\right)^3 \\ &\leq \|x_{a_T+1} - x_{a_T}\|^3 + \frac{2^3}{2}\sum_{t=a_T+1}^{a_T+2}\|x_t - x_{t-1}\|^3 + \frac{3^3}{3}\sum_{t=a_T+1}^{a_T+3}\|x_t - x_{t-1}\|^3 \\ &\quad + \cdots + \frac{(\ell_T-1)^3}{\ell_T-1}\sum_{t=a_T+1}^{a_T+\ell_T-1}\|x_t - x_{t-1}\|^3 \\ &\leq \left(1 + 2^2 + 3^2 + \cdots + (\ell_T-1)^2\right)\sum_{t=a_T+1}^{a_T+\ell_T-1}\|x_t - x_{t-1}\|^3 \\ &= \frac{\ell_T(\ell_T-1)(2\ell_T-1)}{6}\sum_{t=a_T+1}^{a_T+\ell_T-1}\|x_t - x_{t-1}\|^3 \leq \frac{(m_{\tau_T}+1)^2 m_{\tau_T}}{3}\sum_{t=a_T}^{a_T+\ell_T-1}\|x_{t+1} - x_t\|^3, \end{aligned}$$

where the last equality follows from the formula $\sum_{i=1}^{s} i^2 = s(s+1)(2s+1)/6$, whereas the last inequality holds since $1 < \ell_T \leq m_{\tau_T} + 1$. Note that the last inequality also holds when $\ell_T = 1$. Hence,

$$\sum_{t=a_T}^{a_T+\ell_T-1} \|x_t - x_{a_T}\|^3 \leq \frac{(m_{\tau_T}+1)^2 m_{\tau_T}}{3} \sum_{t=a_T}^{a_T+\ell_T-1} \|x_{t+1} - x_t\|^3, \quad \forall\, 1 \leq \ell_T \leq m_{\tau_T} + 1. \tag{35}$$

Therefore, combining the last inequality with (34), we get

$$\begin{aligned} F(x_{a_T+\ell_T}) \leq F(x_{a_T}) &- \frac{\sigma_0(m_{\max}+1)}{6} \sum_{t=a_T}^{a_T+\ell_T-1} \|x_{t+1} - x_t\|^3 + \frac{\sigma_0(m_{\tau_T}+1)}{96} \sum_{t=a_T}^{a_T+\ell_T-1} \|x_{t+1} - x_t\|^3 \\ &+ \frac{\sigma_0 \ell_T \|x_{a_T} - x_{a_T-1}\|^3}{8\gamma_1} + \frac{\sigma_0\gamma_2}{8}\|x_{a_T} - x_{a_T-1}\|^3 + \frac{\sigma_0\gamma_2}{8} \sum_{t=a_T}^{a_T+\ell_T-2} \|x_{t+1} - x_t\|^3, \end{aligned}$$

which, combined with $T = a_T + \ell_T$ and $\ell_T \leq m_{\tau_T} + 1 \leq m_{\max} + 1$, implies the desired inequality. □

As a consequence of the previous lemma, we obtain the following bound for the sum of the sequence $\{\|x_{t+1} - x_t\|^3\}$.

Lemma 12. *The following inequality holds:*

$$\sum_{t=0}^{T-1} \|x_{t+1} - x_t\|^3 \leq \frac{1}{m_{\max}+1} \left(\frac{48(F(x_0) - F^\star)}{\sigma_0} + 12\|x_0 - x_{-1}\|^3 \right).$$

Proof. Applying Lemma 11 for the first block of the Algorithm 1 and using that $a_T = m_0 = 0$, $\gamma_1 = m_1 + 1$ and $\gamma_2 = 1$, we obtain

$$\begin{aligned} F(x_{m_1+1}) \leq F(x_0) &- \frac{15\sigma_0(m_{\max}+1)}{96} \sum_{t=0}^{m_1} \|x_{t+1} - x_t\|^3 + \frac{\sigma_0(m_1+1)}{8(m_1+1)}\|x_0 - x_{-1}\|^3 \\ &+ \frac{\sigma_0}{8}\|x_0 - x_{-1}\|^3 + \frac{\sigma_0}{8} \sum_{t=0}^{m_1-1} \|x_{t+1} - x_t\|^3. \end{aligned}$$

Again, applying Lemma 11 for other blocks and using that $\gamma_1 = 12$ and $\gamma_2 = m_{\tau_T} + 1$ we have

$$\begin{aligned} F(x_{m_1+m_2+2}) \leq F(x_{m_1+1}) &- \frac{15\sigma_0(m_{\max}+1)}{96} \sum_{t=m_1+1}^{m_1+m_2+1} \|x_{t+1} - x_t\|^3 + \frac{\sigma_0(m_2+1)}{96}\|x_{m_1+1} - x_{m_1}\|^3 \\ &+ \frac{\sigma_0(m_2+1)}{8}\|x_{m_1+1} - x_{m_1}\|^3 + \frac{\sigma_0(m_2+1)}{8} \sum_{t=m_1+1}^{m_1+m_2} \|x_{t+1} - x_t\|^3 \\ &\vdots \\ F(x_T) \leq F(x_{a_T}) &- \frac{15\sigma_0(m_{\max}+1)}{96} \sum_{t=a_T}^{T-1} \|x_{t+1} - x_t\|^3 + \frac{\sigma_0(m_{\tau_T}+1)}{96}\|x_{a_T} - x_{a_T-1}\|^3 \\ &+ \frac{\sigma_0(m_{\tau_T}+1)}{8}\|x_{a_T} - x_{a_T-1}\|^3 + \frac{\sigma_0(m_{\tau_T}+1)}{8} \sum_{t=a_T}^{T-2} \|x_{t+1} - x_t\|^3, \end{aligned}$$

where $a_T = m_1 + m_2 + \cdots + m_{\tau_T - 1} + \tau_T - 1$. It follows from the above inequalities that

$$\begin{aligned} F(x_T) \leq F(x_0) &- \frac{15\sigma_0(m_{\max}+1)}{96} \sum_{t=0}^{T-1} \|x_{t+1} - x_t\|^3 + \frac{\sigma_0(m_{\max}+1)}{8} \sum_{t=0}^{T-2} \|x_{t+1} - x_t\|^3 \\ &+ \frac{\sigma_0\|x_0 - x_{-1}\|^3}{4} + \frac{\sigma_0(m_2+1)}{96}\|x_{m_1+1} - x_{m_1}\|^3 + \cdots + \frac{\sigma_0(m_{\tau_T}+1)}{96}\|x_{a_T} - x_{a_T-1}\|^3, \end{aligned}$$

which implies that

$$F(x_T) \leq F(x_0) - \frac{\sigma_0(m_{\max}+1)}{48}\sum_{t=0}^{T-1}\|x_{t+1}-x_t\|^3 + \frac{\sigma_0\|x_0-x_{-1}\|^3}{4}.$$

Now, using that $F(x_T) \geq F^\star$, we obtain

$$\frac{\sigma_0(m_{\max}+1)}{48}\sum_{t=0}^{T-1}\|x_{t+1}-x_t\|^3 \leq F(x_0) - F^\star + \frac{\sigma_0}{4}\|x_0-x_{-1}\|^3.$$

Therefore, the desired inequality now follows from the last one. □

The following lemma establishes a relationship between the sums of the sequences $\{\|\nabla f(x_{t+1}) + \psi'(x_{t+1})\|^{\frac{3}{2}}\}$ and $\{\|x_{t+1}-x_t\|^3\}$.

Lemma 13. *Let $\tau_T \in \mathbb{N} - \{0\}$ be the block number associated with the T-th iteration as defined in Lemma 11. Then,*

$$\sqrt{2}\sum_{t=a_T}^{a_T+\ell_T-1}\|\nabla f(x_{t+1}) + \psi'(x_{t+1})\|^{\frac{3}{2}} \leq \left(\tilde{\lambda}^{\frac{3}{2}} + \frac{L^{\frac{3}{2}}(m_{\max}+1)^{\frac{3}{2}}}{3}\right)\sum_{t=a_T}^{a_T+\ell_T-1}\|x_{t+1}-x_t\|^3 + 8\bar{\kappa}_g^{\frac{3}{2}}\sum_{t=a_T-1}^{a_T+\ell_T-2}\|x_{t+1}-x_t\|^3 + 8\bar{\kappa}_B^{\frac{3}{2}}\|x_{a_T}-x_{a_T-1}\|^3. \quad (36)$$

where $a_T = m_0 + m_1 + m_2 + \cdots + m_{\tau_T-1} + \tau_T - 1$, $\ell_T \in \mathbb{N}$ with $1 \leq \ell_T \leq m_{\tau_T}+1$, and $\tilde{\lambda} := 2(\sigma_{\max}+\theta) + L + L(m_{\max}+1) + 2(m_{\max}+1)^{\frac{2}{3}}\bar{\kappa}_B$.

Proof. Since a_T+1 is the first iteration of the block τ_T, it follows from Proposition 9 with $\sigma = 2^{i_{a_T}}\sigma_{a_T}$, $x^+ = x_{a_T+1}$, $x = z = x_{a_T}$, $\hat{x} = \hat{z} = x_{a_T-1}$, $\hat{\kappa}_g = \bar{\kappa}_g/2^{i_{a_T}-1}$, $\hat{\kappa}_B = \bar{\kappa}_B/2^{i_{a_T}-1}$, $\rho = 1$ and $\rho^\star = (m_{\tau_T}+1)^{\frac{2}{3}}$ that

$$\sqrt{2}\|\nabla f(x_{a_T+1}) + \psi'(x_{a_T+1})\|^{\frac{3}{2}} \leq \left(2^{i_{a_T}}\sigma_{a_T} + L + L + (m_{\tau_T}+1)^{\frac{2}{3}}\left(\frac{\bar{\kappa}_B}{2^{i_{a_T}-1}}\right) + 2\theta\right)^{\frac{3}{2}} \times \|x_{a_T+1}-x_{a_T}\|^3 + \left(\frac{2\bar{\kappa}_g}{2^{i_{a_T}-1}}\right)^{\frac{3}{2}}\|x_{a_T}-x_{a_T-1}\|^3 + \left(\frac{2\bar{\kappa}_B}{2^{i_{a_T}-1}}\right)^{\frac{3}{2}}\frac{\|x_{a_T}-x_{a_T-1}\|^3}{(m_{\tau_T}+1)},$$

which, combined with $2^{i_{a_T}-1}\sigma_{a_T} = \sigma_{a_T+1} \leq \sigma_{\max}$ (see Step 2 of Algorithm 1 and (31)), $m_{\tau_T} \leq m_{\max}$, the definition of $\tilde{\lambda}$ and $i_{a_T} \geq 0$, yields

$$\sqrt{2}\|\nabla f(x_{a_T+1}) + \psi'(x_{a_T+1})\|^{\frac{3}{2}} \leq \tilde{\lambda}^{\frac{3}{2}}\|x_{a_T+1}-x_{a_T}\|^3 + 8\bar{\kappa}_g^{\frac{3}{2}}\|x_{a_T}-x_{a_T-1}\|^3 + \frac{8\bar{\kappa}_B^{\frac{3}{2}}\|x_{a_T}-x_{a_T-1}\|^3}{(m_{\tau_T}+1)}. \quad (37)$$

For the other iterations in the block τ_T (if $m_{\tau_T} > 0$), it follows from Proposition 9 with $\sigma = 2^{j_{a_T+\ell_T-1}}\sigma_{a_T+\ell_T-1}$, $x^+ = x_{a_T+\ell_T}$, $x = x_{a_T+\ell_T-1}$, $\hat{x} = x_{a_T+\ell_T-2}$, $z = x_{a_T}$, $\hat{z} = x_{a_T-1}$, $\hat{\kappa}_g = \bar{\kappa}_g/2^{j_{a_T+\ell_T-1}-1}$, $\hat{\kappa}_B = \bar{\kappa}_B/2^{j_{a_T+\ell_T-1}-1}$, $\rho = (m_{\tau_T}+1)^{1/3}m_{\tau_T}^{2/3}$ and $\rho^\star = (m_{\tau_T}+1)^{2/3}$ that

$$\sqrt{2}\|\nabla f(x_{a_T+\ell_T}) + \psi'(x_{a_T+\ell_T})\|^{\frac{3}{2}} \leq \left(\frac{2\bar{\kappa}_g}{2^{j_{a_T+\ell_T-1}-1}}\right)^{\frac{3}{2}}\|x_{a_T+\ell_T-1}-x_{a_T+\ell_T-2}\|^3 + \frac{L^{\frac{3}{2}}\|x_{a_T+\ell_T-1}-x_{a_T}\|^3}{(m_{\tau_T}+1)^{\frac{1}{2}}m_{\tau_T}} + \left(2^{j_{a_T+\ell_T-1}}\sigma_{a_T+\ell_T-1} + L + L(m_{\tau_T}+1)^{1/3}m_{\tau_T}^{2/3} + (m_{\tau_T}+1)^{\frac{2}{3}}\left(\frac{\bar{\kappa}_B}{2^{j_{a_T+\ell_T-1}-1}}\right) + 2\theta\right)^{\frac{3}{2}} \times \|x_{a_T+\ell_T}-x_{a_T+\ell_T-1}\|^3 + \left(\frac{2\bar{\kappa}_B}{2^{j_{a_T+\ell_T-1}-1}}\right)^{\frac{3}{2}}\frac{\|x_{a_T}-x_{a_T-1}\|^3}{(m_{\tau_T}+1)},$$

for every $2 \leq \ell_T \leq m_{\tau_T}+1$. Combining the last inequality with $2^{j_{a_T+\ell_T}-1}\sigma_{a_T+\ell_T} = \sigma_{a_T+\ell_T+1} \leq \sigma_{\max}$ (see Step 5 of Algorithm 1 and (31)), $m_{\tau_T} \leq m_{\max}$, the definition of $\tilde{\lambda}$ and $j_{a_T+\ell_T} \geq 0$, we get

$$\sqrt{2}\|\nabla f(x_{a_T+\ell_T}) + \psi'(x_{a_T+\ell_T})\|^{\frac{3}{2}} \leq \tilde{\lambda}^{\frac{3}{2}}\|x_{a_T+\ell_T} - x_{a_T+\ell_T-1}\|^3 + 8\bar{\kappa}_g^{\frac{3}{2}}\|x_{a_T+\ell_T-1} - x_{a_T+\ell_T-2}\|^3 + \frac{L^{\frac{3}{2}}\|x_{a_T+\ell_T-1} - x_{a_T}\|^3}{(m_{\tau_T}+1)^{\frac{1}{2}}m_{\tau_T}} + \frac{8\bar{\kappa}_B^{\frac{3}{2}}\|x_{a_T} - x_{a_T-1}\|^3}{(m_{\tau_T}+1)},$$

for every $2 \leq \ell_T \leq m_{\tau_T}+1$. Combining the last inequalities with (37), we have

$$\sqrt{2}\sum_{t=a_T}^{a_T+\ell_T-1}\|\nabla f(x_{t+1}) + \psi'(x_{t+1})\|^{\frac{3}{2}} \leq \tilde{\lambda}^{\frac{3}{2}}\sum_{t=a_T}^{a_T+\ell_T-1}\|x_{t+1} - x_t\|^3 + 8\bar{\kappa}_g^{\frac{3}{2}}\sum_{t=a_T-1}^{a_T+\ell_T-2}\|x_{t+1} - x_t\|^3 + \frac{L^{\frac{3}{2}}}{(m_{\tau_T}+1)^{\frac{1}{2}}m_{\tau_T}}\sum_{t=a_T}^{a_T+\ell_T-1}\|x_t - x_{a_T}\|^3 + \frac{8\bar{\kappa}_B^{\frac{3}{2}}\ell_T}{(m_{\tau_T}+1)}\|x_{a_T} - x_{a_T-1}\|^3,$$

which, combined with $\ell_T \leq m_{\tau_T}+1$ and (35), yields

$$\sqrt{2}\sum_{t=a_T}^{a_T+\ell_T-1}\|\nabla f(x_{t+1}) + \psi'(x_{t+1})\|^{\frac{3}{2}} \leq \left(\tilde{\lambda}^{\frac{3}{2}} + \frac{L^{\frac{3}{2}}(m_{\tau_T}+1)^{\frac{3}{2}}}{3}\right)\sum_{t=a_T}^{a_T+\ell_T-1}\|x_{t+1} - x_t\|^3 + 8\bar{\kappa}_g^{\frac{3}{2}}\sum_{t=a_T-1}^{a_T+\ell_T-2}\|x_{t+1} - x_t\|^3 + 8\bar{\kappa}_B^{\frac{3}{2}}\|x_{a_T} - x_{a_T-1}\|^3.$$

Therefore, the desired inequality now follows from the fact that $m_{\tau_T} \leq m_{\max}$. □

We are now ready to prove Theorem 5.

Proof of Theorem 5. Applying the inequality in (36) to multiple blocks ($\tau = 1, \ldots, \tau_T$), we obtain

$$\sqrt{2}\sum_{t=0}^{m_1}\|\nabla f(x_{t+1}) + \psi'(x_{t+1})\|^{\frac{3}{2}} \leq \left(\tilde{\lambda}^{\frac{3}{2}} + \frac{L^{\frac{3}{2}}(m_{\max}+1)^{\frac{3}{2}}}{3}\right)\sum_{t=0}^{m_1}\|x_{t+1} - x_t\|^3 + 8\bar{\kappa}_g^{\frac{3}{2}}\sum_{t=-1}^{m_1-1}\|x_{t+1} - x_t\|^3 + 8\bar{\kappa}_B^{\frac{3}{2}}\|x_0 - x_{-1}\|^3,$$

$$\sqrt{2}\sum_{t=m_1+1}^{m_1+m_2+1}\|\nabla f(x_{t+1}) + \psi'(x_{t+1})\|^{\frac{3}{2}} \leq \left(\tilde{\lambda}^{\frac{3}{2}} + \frac{L^{\frac{3}{2}}(m_{\max}+1)^{\frac{3}{2}}}{3}\right)\sum_{t=m_1+1}^{m_1+m_2+1}\|x_{t+1} - x_t\|^3 + 8\bar{\kappa}_g^{\frac{3}{2}}\sum_{t=m_1}^{m_1+m_2}\|x_{t+1} - x_t\|^3 + 8\bar{\kappa}_B^{\frac{3}{2}}\|x_{m_1+1} - x_{m_1}\|^3,$$

$$\vdots$$

$$\sqrt{2}\sum_{t=a_T}^{T-1}\|\nabla f(x_{t+1}) + \psi'(x_{t+1})\|^{\frac{3}{2}} \leq \left(\tilde{\lambda}^{\frac{3}{2}} + \frac{L^{\frac{3}{2}}(m_{\max}+1)^{\frac{3}{2}}}{3}\right)\sum_{t=a_T}^{T-1}\|x_{t+1} - x_t\|^3 + 8\bar{\kappa}_g^{\frac{3}{2}}\sum_{t=a_T-1}^{T-2}\|x_{t+1} - x_t\|^3 + 8\bar{\kappa}_B^{\frac{3}{2}}\|x_{a_T} - x_{a_T-1}\|^3.$$

where $T = a_T + \ell_T$ and $a_T = m_0 + m_1 + m_2 + \cdots + m_{\tau_T-1} + \tau_T - 1$. It follows from the above inequalities

that

$$\sqrt{2}\sum_{t=0}^{T-1}\|\nabla f(x_{t+1})+\psi'(x_{t+1})\|^{\frac{3}{2}} \leq \left(\tilde{\lambda}^{\frac{3}{2}}+\frac{L^{\frac{3}{2}}(m_{\max}+1)^{\frac{3}{2}}}{3}+8\bar{\kappa}_g^{\frac{3}{2}}\right)\sum_{t=0}^{T-1}\|x_{t+1}-x_t\|^3$$
$$+8(\bar{\kappa}_B^{\frac{3}{2}}+\bar{\kappa}_g^{\frac{3}{2}})\|x_0-x_{-1}\|^3+8\bar{\kappa}_B^{\frac{3}{2}}\|x_{m_1+1}-x_{m_1}\|^3+\ldots+8\bar{\kappa}_B^{\frac{3}{2}}\|x_{a_T}-x_{a_T-1}\|^3$$
$$\leq \left(\tilde{\lambda}^{\frac{3}{2}}+\frac{L^{\frac{3}{2}}(m_{\max}+1)^{\frac{3}{2}}}{3}+8(\bar{\kappa}_B^{\frac{3}{2}}+\bar{\kappa}_g^{\frac{3}{2}})\right)\sum_{t=0}^{T-1}\|x_{t+1}-x_t\|^3+8(\bar{\kappa}_B^{\frac{3}{2}}+\bar{\kappa}_g^{\frac{3}{2}})\|x_0-x_{-1}\|^3,$$

which, combined with Lemma 12, yields

$$\sum_{t=0}^{T-1}\|\nabla f(x_{t+1})+\psi'(x_{t+1})\|^{\frac{3}{2}} \leq \frac{1}{\sqrt{2}(m_{\max}+1)}\left(\tilde{\lambda}^{\frac{3}{2}}+\frac{L^{\frac{3}{2}}(m_{\max}+1)^{\frac{3}{2}}}{3}+8(\bar{\kappa}_B^{\frac{3}{2}}+\bar{\kappa}_g^{\frac{3}{2}})\right)$$
$$\times\left(\frac{48(F(x_0)-F^\star)}{\sigma_0}+12\|x_0-x_{-1}\|^3\right)+\frac{8(\bar{\kappa}_B^{\frac{3}{2}}+\bar{\kappa}_g^{\frac{3}{2}})}{\sqrt{2}}\|x_0-x_{-1}\|^3.$$

On the other hand, it follows from the definitions of $\tilde{\lambda}$ and $\sigma_{\max}$ given in Lemma 13 and inequality (31), respectively, that

$$\tilde{\lambda} \leq \frac{2(m_{\max}+1)[10L^{\frac{3}{2}}+16\bar{\kappa}_B^{\frac{3}{2}}+\sigma_0^{\frac{3}{2}}+(\theta+3L+\bar{\kappa}_B)\sigma_0^{\frac{1}{2}}+114\bar{\kappa}_g^3\sigma_0^{-\frac{3}{2}}]}{\sigma_0^{\frac{1}{2}}}.$$

Therefore, the inequality in (20) follows now from the last two inequalities, some algebraic manipulations, and the definition of $\bar{\lambda}$ in (21). The second part of the theorem follows directly from (20). □

We next prove Corollary 6.

Proof of Corollary 6. Let $\tau_T \in \mathbb{N}-\{0\}$ be the block number associated with the T-th iteration as defined in Lemma 11, that is,

$$T = a_T+\ell_T, \quad \text{with} \quad a_T = m_0+m_1+m_2+\cdots+m_{\tau_T-1}+\tau_T-1, \quad \ell_T\in\mathbb{N}, \quad 1\leq\ell_T\leq m_{\tau_T}+1. \tag{38}$$

Hence, taking into account that the Hessian is updated only in the first iteration (i.e., $(a_T+1)-$iteration) of the block, the number of function and gradient evaluations is bounded by $(i_t+1)(n+2)$ if $t=a_T$, and by (j_t+2) if $t\in\{a_T+1,\ldots,a_T+\ell_T-1\}$. Now, since $\sigma_{a_T+1}=2^{i_{a_T}-1}\sigma_{a_T}$ and $\sigma_{t+1}=2^{j_t-1}\sigma_t$, we have

$$i_{a_T}+1=\log_2\sigma_{a_T+1}-\log_2\sigma_{a_T}+2, \quad j_t+1=\log_2\sigma_{t+1}-\log_2\sigma_t+2. \tag{39}$$

Now,

$$(i_{a_T}+1)(n+2)+\sum_{t=a_T+1}^{a_T+\ell_T-1}(j_t+2)=(i_{a_T}+1)(n+1)+\sum_{t=a_T}^{a_T+\ell_T-1}(j_t+1)+\ell_T-1.$$

Applying the last equality to multiple blocks ($\tau=1,\ldots,\tau_T$) and using (39), we obtain

$$(i_0+1)(n+2)+\sum_{t=1}^{m_1}(j_t+1)=(i_0+1)(n+1)+\sum_{t=0}^{m_1}(\log_2\sigma_{t+1}-\log_2\sigma_t+2)+m_1,$$

$$(i_{m_1+1}+1)(n+2)+\sum_{t=m_1+2}^{m_1+m_2+1}(j_t+1)=(i_{m_1+1}+1)(n+1)+\sum_{t=m_1+1}^{m_1+m_2+1}(\log_2\sigma_{t+1}-\log_2\sigma_t+2)+m_2,$$

$$\vdots$$

$$(i_{a_T}+1)(n+2)+\sum_{t=a_T+1}^{T-1}(j_t+1)\leq (i_{a_T}+1)(n+1)+\sum_{t=a_T}^{a_T+\ell_T-1}(\log_2\sigma_{t+1}-\log_2\sigma_t+2)+m_{\tau_T}.$$

Summing up the last inequalities and using (31) and (39), we obtain

$$\begin{aligned}
FGE(T) &\leq (n+1)[(i_0+1)+\ldots+(i_{a_T}+1)]+\sum_{t=0}^{a_T+\ell_T-1}(\log_2\sigma_{t+1}-\log_2\sigma_t+2)+T\\
&= (n+1)[(\log_2\sigma_1-\log_2\sigma_0+2)+\ldots+(\log_2\sigma_{a_T+1}-\log_2\sigma_{a_T}+2)]+\log_2\frac{\sigma_T}{\sigma_0}+3T\\
&\leq (n+1)\left[(\log_2\sigma_{m_1+2}-\log_2\sigma_{m_1+1}+2)+\ldots+(\log_2\sigma_{a_T+1}-\log_2\sigma_{a_T}+2)\right]\\
&+(n+2)\log_2\frac{\sigma_{\max}}{\sigma_0}+3(T+n+1)\\
&\leq (n+1)(\tau_T-1)\left[\log_2\frac{\sigma_{\max}}{\sigma_0(m_{\max}+1)}+2\right]+(n+2)\log_2\frac{\sigma_{\max}}{\sigma_0}+3(T+n+1). \qquad (40)
\end{aligned}$$

Now, from (38), $m_\tau\geq m_{\min}$ and $\ell_T\geq 0$, we find that

$$\begin{aligned}
T=a_T+\ell_T &= m_0+m_1+m_2+\cdots+m_{\tau_T-1}+\tau_T-1+\ell_T\\
&\geq (\tau_T-1)m_{\min}+\tau_T-1,
\end{aligned}$$

which implies that $\tau_T-1\leq T/(m_{\min}+1)$. Therefore, combining the last two inequalities, we find that

$$FGE(T)\leq\frac{T(n+1)}{m_{\min}+1}\left[\log_2\frac{\sigma_{\max}}{\sigma_0(m_{\max}+1)}+2\right]+(n+2)\log_2\frac{\sigma_{\max}}{\sigma_0}+3(T+n+1),$$

which, combined with the definition of $\sigma_{\max}$ (see (31) with $\bar\kappa_g:=0$ and $\bar\kappa_B:=L\kappa_B$), implies (23). The second statement of the lemma follows from (23) and Theorem 5. □

We next prove Corollary 7.

Proof of Corollary 7. Note first that the Hessian is updated only in the first iteration of each block (that is, when $t=a_T$), while the gradient is updated at every iteration. Hence, the number of function evaluations is bounded by $(i_t+1)(4n^2+2n+2)$ if $t=a_T$, and by $(j_t+1)(2n+2)$ if $t\in a_T+1,\ldots,a_T+\ell_T-1$. Following the same idea as in the proof of Corollary 6, we obtain the following inequality, analogous to (40):

$$FE(T)\leq 4n^2(\tau_T-1)\left[\log_2\frac{\sigma_{\max}}{\sigma_0(m_{\max}+1)}+2\right]+2(2n^2+n+1)\log_2\frac{\sigma_{\max}}{\sigma_0}+4((n+1)T+2n^2).$$

Now, using $\tau_T-1\leq T/(m_{\min}+1)$ and the definition of $\sigma_{\max}$ (see (31) with $\bar\kappa_g:=L\kappa_g$ and $\bar\kappa_B:=L\kappa_B$), we obtain (24). The second statement of the corollary then follows from (24) and Theorem 5. □

4 Numerical Experiments

We illustrate the practical performance of Algorithm 1 on the set of 35 problems from the Moré–Garbow–Hillstrom collection [17]. Our experimental setup closely follows that of [9], allowing for a direct comparison with their cubic regularization methods with lazy Hessian updates. In particular, we adopt the same test problems (see Table 2 therein) and the same performance measures (number of function/gradient or function evaluations). We emphasize that the goal of these experiments is not to provide a comprehensive computational study, but rather to assess the practical behavior of the proposed method and, in particular, to highlight the effect of adaptively reusing Hessian approximations within the cubic regularization framework.

For all algorithms, each cubic subproblem was approximately solved using the Barzilai–Borwein gradient (BBG) method [1], combined with the nonmonotone line search of [14], using the origin as the initial point. All algorithms were implemented in Python and executed on a machine equipped with a 3.5 GHz dual-core Intel Core i5 processor and 16 GB of 2400 MHz DDR4 memory.

In all instances of Algorithm 1, we set $x_{-1}=0$, x_0 as provided by the test library, $\theta=0.5$ and $\sigma_0=0.1$.

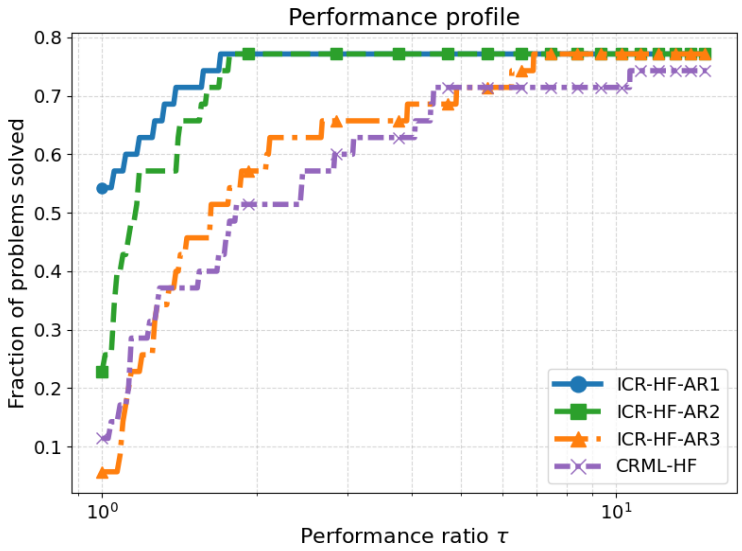

Figure 1: Performance profiles comparing the Hessian-free algorithms in terms of function and gradient evaluations.

4.1 Hessian-free implementations

We first consider Hessian-free implementations of the proposed method, in which exact gradient information is available and Hessian approximations are constructed using gradient values. Specifically, in Steps 1.1 and 4.1 we set $g_{t,i} = g_{t,j} := \nabla f(x_t)$ (that is, $\bar{\kappa}_g = 0$), and define $B_{t,i} := B(x_t)$, where $B(x)$ is given in (11) and $A(x)$ is defined in (8), with $h := 0.2\|x_t - x_{t-1}\|/(2^{i-1}\sqrt{n})$.

We consider the following choices for the reuse parameter m_τ:

- *ICR-HF-AR1*: Algorithm 1 with $m_{\min} = 0$, $m_{\max} = \lfloor 4n/3 \rfloor$, and m_τ chosen as

$$m_\tau = \begin{cases} \lfloor 2n/3 \rfloor, & \text{if } \|x_t - x_{t-1}\| \geq 10^{-2}, \\ \min(m_{\tau-1} + 2,\ m_{\max}), & \text{otherwise .} \end{cases}$$

- *ICR-HF-AR2*: Algorithm 1 with $m_\tau = m_{\min} = m_{\max} = n$;
- *ICR-HF-AR3*: Algorithm 1 with $m_\tau = m_{\min} = m_{\max} = 0$.

For comparison, we also consider the Hessian-free version of the method proposed in [9], which achieved the best performance in their experiments:

- *CRML-HF*: [9, Algorithm 2] with $\tau_0 = 1$, $\varepsilon = \|\nabla f(x_0)\|10^{-8}$, and $m = n$.

All methods are terminated when the stopping criterion $\|\nabla f(x_t)\|/\|\nabla f(x_0)\| \leq 10^{-8}$ is satisfied, or when a maximum of 50,000 function and gradient evaluations is reached.

The performance profile, in terms of function and gradient evaluations, in Figure 1 indicates that the proposed adaptive strategy *ICR-HF-AR1* achieves the best overall performance among all tested methods. In particular, it attains the highest efficiency (54.29%) while maintaining full robustness (77.14%), clearly outperforming the other approaches in terms of efficiency. This behavior highlights the advantage of adaptively selecting the reuse parameter m_τ. In contrast, the variants with a fixed reuse parameter (*ICR-HF-AR2* and *CRML-HF*) exhibit weaker performance compared to the adaptive strategy. Nevertheless, they remain more efficient than the variant without Hessian approximation reuse, *ICR-HF-AR3*. This indicates that incorporating Hessian reuse, even in a fixed manner, is beneficial, while adaptivity further enhances performance.

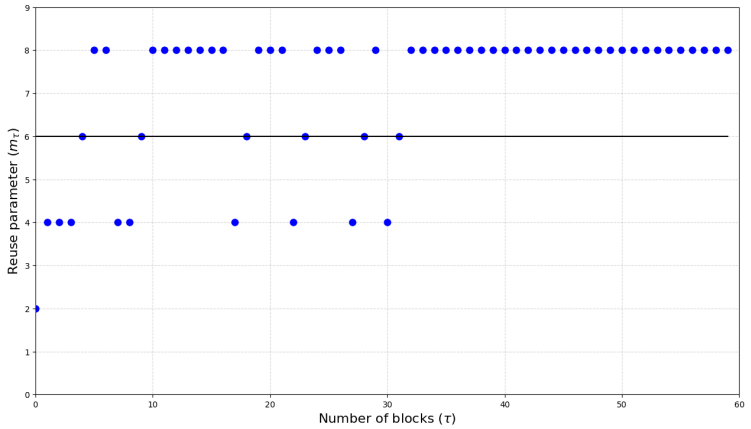

Figure 2: Evolution of the reuse parameter m_τ for *ICR-HF-AR1* applied to the Watson function ($n = 6$).

Figure 2 illustrates the evolution of the reuse parameter m_τ for *ICR-HF-AR1* along the iterations (blocks) for the Watson function with dimension $n = 6$. In this example, the adaptive strategy initially selects smaller values of m_τ, promoting more frequent updates of the Hessian approximation. As the iterations progress, larger values are increasingly chosen, with m_τ often approaching the reference value n.

4.2 Derivative-free implementations

We now consider derivative-free implementations of the proposed method, in which only function values are available and both gradient and Hessian approximations are constructed from these values. Specifically, in Steps 1.1 and 4.1, we set

$$g_{t,i} = g_{t,j} := g(x_t), \qquad B_{t,i} := B(x_t),$$

where $g(x)$ is defined in (6), $B(x)$ is given in (14), and $A(x)$ is specified in (12), with

$$h := \min\left\{\left(\frac{6\times 10^{-2}}{\sqrt{n}\,2^{i-1}}\right)^{\frac{1}{2}}, \frac{3\times 0.75}{(1+\sqrt{2})\,n\,2^{i-1}}\right\}\|x_t - x_{t-1}\|.$$

We consider the same choices of the reuse parameter m_τ as in Section 4.1, and denote the corresponding variants by *ICR-DF-AR1*, *ICR-DF-AR2*, and *ICR-DF-AR3*. For comparison, we also consider the derivative-free method proposed in [9, Algorithm 4], which we denote by *CRML-DF*, with $\tau_0 = 0.06$, $\varepsilon = 10^{-6}\big(f(x_0) - f^*\big)$, and $m = n$.

All methods are terminated when the stopping criterion

$$f(x_t) - f^* \le 10^{-6}\big(f(x_0) - f^*\big)$$

is satisfied, where f^* is provided by the test library, or when a maximum of 50,000 function evaluations is reached.

The performance profile, in terms of function evaluations, in Figure 3 shows that the overall behavior of the derivative-based variants closely mirrors that observed in the previous section. In particular, the adaptive strategy *ICR-DF-AR1* again delivers the best overall performance, achieving the highest efficiency while maintaining strong robustness across the tested problems. This confirms that the benefits of adaptively selecting the reuse parameter m_τ persist in the derivative-based setting. The fixed strategies *ICR-DF-AR2* and *CRML-DF* exhibit competitive but consistently weaker performance compared to the

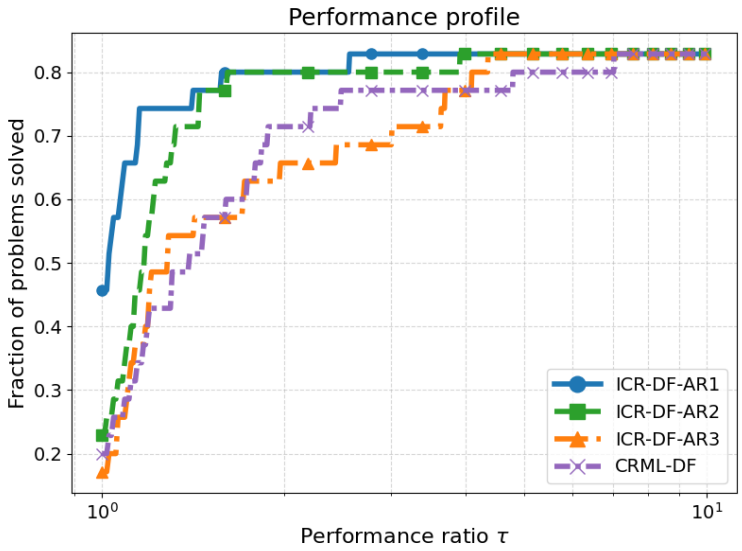

Figure 3: Performance profiles comparing the derivative-free algorithms in terms of function evaluations.

adaptive variant. Although they eventually reach similar robustness levels for larger values of τ. In contrast, *ICR-DF-AR3* shows slower progress in terms of efficiency, despite attaining comparable robustness as τ increases. This reinforces the observation that incorporating Hessian approximation reuse is beneficial, and that adaptivity further enhances performance.

5 Conclusion

In this work, we proposed an inexact cubic regularization method with adaptive reuse of Hessian approximations for solving general non-convex optimization problems. The method combines inexact gradient information with a flexible lazy strategy, allowing the reuse parameter to vary along the iterations. We established iteration-complexity guarantees ensuring convergence to approximate critical points, along with bounds on the total number of gradient and function evaluations. Numerical results demonstrated that the proposed adaptive strategy consistently improves efficiency over fixed reuse schemes and existing lazy cubic regularization methods, while maintaining robustness.

References

[1] J. Barzilai and J. M. Borwein. Two-Point Step Size Gradient Methods. *IMA J. Numer. Anal.*, 8(1):141–148, 1988.

[2] S. Bellavia and G. G. Stochastic analysis of an adaptive cubic regularization method under inexact gradient evaluations and dynamic hessian accuracy. *Optimization*, 71(1):227–261, 2022.

[3] S. Bellavia, G. Gurioli, and B. Morini. Adaptive cubic regularization methods with dynamic inexact Hessian information and applications to finite-sum minimization. *IMA J. Numer. Anal.*, 41(1):764–799, 04 2020.

[4] S. Bellavia, G. Gurioli, B. Morini, and P. L. Toint. Quadratic and cubic regularisation methods with inexact function and random derivatives for finite-sum minimisation. In *In: Proceedings of the 21st International Conference on Computational Science and Its Applications*, page 258–267, 2021.

[5] C. Cartis, N. I. Gould, and P. L. Toint. On the complexity of steepest descent, Newton's and regularized Newton's methods for nonconvex unconstrained optimization problems. *SIAM J. Optim.*, 20(6):2833–2852, 2010.

[6] C. Cartis, N. I. Gould, and P. L. Toint. On the oracle complexity of first-order and derivative-free algorithms for smooth nonconvex minimization. *SIAM J. Optim.*, 22(1):66–86, 2012.

[7] X. Chen, B. Jiang, T. Lin, and S. Zhang. Accelerating adaptive cubic regularization of Newton's method via random sampling. *J. Mach. Learn. Res.*, 23(90):1–38, 2022.

[8] N. Doikov, E. M. Chayti, and M. Jaggi. Second-order optimization with lazy hessians. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 8138–8161. PMLR, 23–29 Jul 2023.

[9] N. Doikov and G. N. Grapiglia. First and zeroth-order implementations of the regularized Newton method with lazy approximated hessians. *J Sci Comput*, 103(32):1–36, 2025.

[10] D. S. Gonçalves, M. L. N. Gonçalves, and J. G. Melo. A cubic regularization method for multiobjective optimization. *arXiv:2506.08181*, 2025.

[11] M. L. N. Gonçalves. Subsampled cubic regularization method for finite-sum minimization. *Optimization*, 74(7):1591–1614, 2025.

[12] G. N. Grapiglia, M. L. N. Gonçalves, and G. N. Silva. A cubic regularization of Newton's method with finite difference hessian approximations. *Numer. Algorithms*, 90, 2022.

[13] A. Griewank. The modification of newton's method for unconstrained optimization by bounding cubic terms. Technical report, Technical report NA/12, 1981.

[14] L. Grippo, F. Lampariello, and S. Lucidi. A nonmonotone line search technique for Newton's method. *SIAM J. Numer. Anal.*, 23(4):707–716, 1986.

[15] J. Kohler and A. Lucchi. Sub-sampled cubic regularization for non-convex optimization. In *In Proceedings of the 34th International Conference on Machine Learning (ICML 2017)*, volume 70, pages 1895 – 1904, 2017.

[16] E. Kreyszig. *Introductory functional analysis with applications*. John Wiley & Sons, 1991.

[17] J. J. Moré, B. S. Garbow, and K. E. Hillstrom. Testing unconstrained optimization software. *ACM Trans. Math. Softw.*, 7(1):17–41, Mar. 1981.

[18] Y. Nesterov and B. T. Polyak. Cubic regularization of newton method and its global performance. *Math. Program.*, 108(1):177–205, 2006.

[19] S. Park, S. H. Jung, and P. M. Pardalos. Combining stochastic adaptive cubic regularization with negative curvature for nonconvex optimization. *J. Optim. Theory Appl.*, 184(3):953–971, 2020.

[20] Z. Wang, Y. Zhou, Y. Liang, and G. Lan. Cubic regularization with momentum for nonconvex optimization. In *Uncertainty in Artificial Intelligence*, pages 313–322. PMLR, 2020.