Journal of Scientific Computing (2025) 103:32
https://doi.org/10.1007/s10915-025-02842-z

First and Zeroth-Order Implementations of the Regularized Newton Method with Lazy Approximated Hessians

Nikita Doikov[1] · **Geovani Nunes Grapiglia**[2]

Received: 7 September 2023 / Revised: 4 December 2024 / Accepted: 13 February 2025 /

Abstract

In this work, we develop first-order (Hessian-free) and zeroth-order (derivative-free) implementations of the Cubically Regularized Newton Method for solving general non-convex optimization problems. For that, we employ finite difference approximations of the derivatives. We use a special adaptive search procedure in our algorithms, which simultaneously fits both the regularization constant and the parameters of the finite difference approximations. It makes our schemes free from the need to know the actual Lipschitz constants. Additionally, we equip our algorithms with the lazy Hessian update that reuses a previously computed Hessian approximation matrix for several iterations. Specifically, we prove the global complexity bound of $\mathcal{O}(n^{1/2}\epsilon^{-3/2})$ function and gradient evaluations for our new Hessian-free method, and a bound of $\mathcal{O}(n^{3/2}\epsilon^{-3/2})$ function evaluations for the derivative-free method, where n is the dimension of the problem and ϵ is the desired accuracy for the gradient norm. These complexity bounds significantly improve the previously known ones in terms of the joint dependence on n and ϵ, for the first-order and zeroth-order non-convex optimization.

Keywords Newton method · Derivative-free optimization · Zeroth-order optimization · Global convergence · Worst-case complexity

Mathematics Subject Classification 49M15 · 49M37 · 65K05 · 90C30 · 90C56

N. Doikov was supported by the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract number 22.00133. G. N. Grapiglia was partially funded by the FRS-FNRS, Belgium (Grant No. CDR J.0081.23).

✉ Nikita Doikov
nikita.doikov@epfl.ch

Geovani Nunes Grapiglia
geovani.grapiglia@uclouvain.be

[1] Machine Learning and Optimization Laboratory (MLO), École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

[2] Institute of Information and Communication Technologies, Electronics and Applied Mathematics (ICTEAM/INMA), Université catholique de Louvain (UCLouvain), Ottignies-Louvain-la-Neuve, Belgium

1 Introduction

1.1 Motivation

The Newton Method is a powerful algorithm for solving numerical optimization problems. Employing the matrix of second derivatives (the Hessian of the objective), the Newton Method is able to efficiently tackle *ill-conditioned problems*, which can be very difficult for solving by the first-order Gradient Methods.

While the Newton Method has been remaining popular for many decades due to its exceptional practical performance, the study of its *global* complexity bounds is relatively recent. One of the most theoretically established versions of this method is the Cubically Regularized Newton Method, proposed initially in [24] by Griewank, that achieves a global complexity [9, 10, 31] of the order $\mathcal{O}(\epsilon^{-3/2})$ for finding a second-order stationary point for non-convex objective with Lipschitz continuous Hessian, where $\epsilon > 0$ is the desired accuracy for the gradient norm. The corresponding complexity of the Gradient Method [30] for non-convex functions with Lipschitz continuous gradient is $\mathcal{O}(\epsilon^{-2})$, which is significantly worse. Thus, the Cubic Newton Method (CNM) achieves a *provable improvement* of the global complexity, as compared to the first-order methods.

In recent years, many efficient modifications of CNM have been developed, including *adaptive* and *universal* methods [9, 10, 15, 16, 21, 22] that do not require to know the actual Lipschitz constant of the Hessian and can automatically adapt to the best problem class among the functions with Hölder continuous derivatives. Additionally, *accelerated* second-order schemes [8, 22, 26, 27, 29, 30] have been introduced, which offer improved convergence rates for convex functions and match the lower complexity bounds [1, 2, 30].

Clearly, we pay a significant price for the better convergence rates of CNM which is: computation of second derivatives and solving a more difficult subproblem in each step. Note that for some of the most difficult modern applications, our available information about the objective function $f(\cdot)$ can be restricted to the black-box

$$\textit{First-order oracle:} \quad x \;\mapsto\; \{f(x), \nabla f(x)\},$$

or even to

$$\textit{Zeroth-order oracle:} \quad x \;\mapsto\; \{f(x)\},$$

without a direct access to the problem structure and any ability to compute the second derivatives $\nabla^2 f(x)$ exactly. Thus, in this black-box scenarios, we are interested to use optimization schemes which efficiently employ only the information we have an access to.

First-order implementations of CNM were proposed and analysed in [11] and [20]. In both of these works, the methods employ finite-difference Hessian approximations, and complexity bounds of $\mathcal{O}(n\epsilon^{-3/2})$ calls of the oracle were proved, where n is the dimension of the problem. In [11], a zeroth-order implementation of CNM was also proposed, for which the authors showed a complexity bound of $\mathcal{O}(n^2\epsilon^{-3/2})$ calls of the oracle. At each iteration, methods in [11] and [20] require the computation of one or more Hessian approximations. Recently, in [14], a second-order variant of CNM with *lazy Hessians* was proposed, in which the same Hessian matrix is reused during $m \geq 1$ consecutive iterations (as in [34]). Remarkably, the method with lazy Hessians retains the iteration complexity bound of $\mathcal{O}(\sqrt{m}\epsilon^{-3/2})$ for nonconvex problems. Moreover, when $m = n$, it requires in the worst-case a number of Hessian evaluations smaller by a factor of $\sqrt{n}$ in comparison with the standard CNM.

In this paper, we efficiently combine the use of finite-differences with the reuse of previously computed Hessian approximations to obtain new first and zeroth-order implementations of the CNM. Specifically, our algorithms employ adaptive searches by which the regularization parameters in the models and the finite-difference intervals are simultaneously adjusted (as in [20]). Additionally, to improve the total oracle complexity of our schemes, we employ the lazy Hessian updates [14], reusing each Hessian approximation for several consecutive steps. As the result, we obtain purely first-order (Hessian-free) and zeroth-order (derivative-free) implementations of CNM that are adaptive and need, respectively, at most $\mathcal{O}(n^{1/2}\epsilon^{-3/2})$ and $\mathcal{O}(n^{3/2}\epsilon^{-3/2})$ calls of the oracle to find an ϵ-approximate second-order stationary point of the objective function. These complexity bounds significantly improve the corresponding bounds in [11] and [20] in terms of the dependence on n. Note that our new methods also support *composite problem* formulation (as, e.g. in [22]), which include both unconstrained minimization and minimization with respect to simple convex constraints or additive regularization. In its turn, the smooth (and the difficult) part of the problem can be non-convex. Finally, we report the result of preliminary numerical experiments that illustrate the practical efficiency of the proposed methods .

1.2 Contents

In Sect. 2 we introduce the inexact step of CNM, which is the main primitive of all our algorithmic schemes. Section 3 is devoted to the finite difference approximations of the second- and first-order derivatives of a smooth functions. In Sect. 4, we present first-order (*Hessian-free*) implementation of CNM and establish its global complexity bounds. Section 5 contains zeroth-order (*derivative-free*) implementation of CNM. In Sect. 6, we establish local superlinear convergence for our schemes. Section 7 presents illustrative numerical experiments. In Sect. 8, we discuss our results.

1.3 Notation and Assumptions

By $\|\cdot\|$ we denote the standard Euclidean norm for vectors and the spectral norm for matrices, while notation $\|\cdot\|_F$ is reserved for the matrix Frobenius norm. We denote by $e_1, \ldots, e_n$ the standard basis vectors in $\mathbb{R}^n$.

We want to solve the following minimization problem

$$\min_{x \in Q} \Big\{ F(x) \overset{\text{def}}{=} f(x) + \psi(x) \Big\}, \tag{1}$$

where $Q \overset{\text{def}}{=} \operatorname{dom} \psi \subseteq \mathbb{R}^n$. Function $f : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable, potentially *non-convex*, while the composite part $\psi : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is *simple* proper, closed, and convex, but possibly non-differentiable (e.g. indicator of a given closed convex set Q).

Therefore, our goal is to find a point $\bar{x} \in Q$ with a small (sub)gradient norm:

$$\|\nabla f(\bar{x}) + \psi'(\bar{x})\| \leq \epsilon, \tag{2}$$

where $\psi'(\bar{x}) \in \partial \psi(\bar{x})$ and $\epsilon > 0$ is a desired tolerance. We are aiming to find a point satisfying (2), using only *first-order* or *zeroth-order* black-box oracle calls for f. At the same time, the composite component ψ is assumed to be simple enough, such that the corresponding auxiliary minimization problems that involve ψ can be efficiently solved (we present the form of the subproblem that we require to solve explicitly in the next section).

We assume that F is bounded from below on Q and denote

$$F^\star \overset{\text{def}}{=} \inf_{x \in Q} F(x) \quad > \quad -\infty.$$

To characterize the smoothness of the differentiable part of the objective, we assume the following:

A1 The Hessian of f is Lipschitz continuous, i.e.,

$$\|\nabla^2 f(y) - \nabla^2 f(x)\| \le L\|y - x\|, \qquad \forall x, y \in \mathbb{R}^n, \tag{3}$$

where $L \ge 0$ is the Lipschitz constant. Note that in all our methods, we do not need to know the exact value of L, estimating it *automatically* with an adaptive procedure.

2 Inexact Cubic Newton Step

In this section, we analyze one step of the Cubically Regularized Newton Method with an *approximate* second-order and first-order information. We also assume that the step of the method is computed *inexactly*, which would allow to apply our methods in the large scale setting.

Given $x \in Q$ and $\sigma > 0$, let us define the models for $f(y)$ around x, exact second-order model with cubic regularization:

$$\Omega_{x,\sigma}(y) \overset{\text{def}}{=} f(x) + \langle \nabla f(x), y - x\rangle + \tfrac{1}{2}\langle \nabla^2 f(x)(y - x), y - x\rangle + \tfrac{\sigma\|y-x\|^3}{6}, \tag{4}$$

and an *approximate model*:

$$M_{x,\sigma}(y) \overset{\text{def}}{=} f(x) + \langle g, y - x\rangle + \tfrac{1}{2}\langle B(y - x), y - x\rangle + \tfrac{\sigma\|y-x\|^3}{6}, \tag{5}$$

where $g \in \mathbb{R}^n$ is an approximation to $\nabla f(x)$ and $B \in \mathbb{R}^{n\times n}$ is an approximation to $\nabla^2 f(z)$, with some previous point $z \in \mathbb{R}^n$ from the past. In the simplest case, we can set $z := x$. However, to reduce the iteration cost of our methods, we will use the same anchor point z for several iterations (that we call *lazy Hessian updates*).

Note that due to the cubic regularizer, we can minimize model (5) globally even when the quadratic part is non-convex. Efficient techniques for solving such subproblems by using Linear Algebra tools or gradient-based solvers were extensively developed in the context of trust-region methods [12] and for the Cubically Regularized Newton methods [7, 9, 10, 31].

Let us consider a minimizer for our approximate model (5) augmented by the composite component:

$$\boxed{x^+ \approx \operatorname{argmin}_{y\in Q}\Big\{\bar{M}(y) \overset{\text{def}}{=} M_{x,\sigma}(y) + \psi(y)\Big\}} \tag{6}$$

Note that the subproblem in (6) can be non-convex and have several global minima. We denote an arbitrary global minimum by argmin; the choice of a specific global minimum does not affect the convergence rates of our methods. We will use such point x^+ as the main iteration step in all our methods.

Note that if x^+ is an *exact solution* to (6), then the following first-order optimality condition holds, for all $y \in Q$:

$$\langle g + B(x^+ - x) + \tfrac{\sigma}{2}\|x^+ - x\|(x^+ - x), y - x^+\rangle + \psi(y) \ge \psi(x^+). \tag{7}$$

Indeed, by the definition of directional derivative, we have for any $y \in Q$ and for a sufficiently small $\alpha > 0$ that

$$\begin{aligned}
&\langle g + B(x^+ - x) + \tfrac{\sigma}{2}\|x^+ - x\|(x^+ - x), y - x^+\rangle + \psi(y) - \psi(x^+) \\
&= \langle \nabla M_{x,\sigma}(x^+), y - x^+\rangle + \psi(y) - \psi(x^+) \\
&= \tfrac{1}{\alpha}\Big[M_{x,\sigma}(x^+ + \alpha(y - x^+)) - M_{x,\sigma}(x^+)\Big] + \psi(y) - \psi(x^+) + o(1) \\
&= \tfrac{1}{\alpha}\Big[\bar{M}(x^+ + \alpha(y - x^+)) - \bar{M}(x^+)\Big] \\
&\quad + \tfrac{1}{\alpha}\Big[\alpha\psi(y) + (1-\alpha)\psi(x^+) - \psi(\alpha y + (1-\alpha)x^+)\Big] + o(1),
\end{aligned}$$

where $o(1)$ is a function that goes to 0 with $\alpha \to 0$. Note that the first term in the right hand side of the last expression is nonnegative because x^+ is a minimizer of the model $\bar{M}(\cdot)$, and the second term is nonnegative due to convexity of ψ. Hence, taking the limit $\alpha \to 0$, (7) holds and we have an explicit expression for a specific subgradient of ψ at new point:

$$-g - B(x^+ - x) - \tfrac{\sigma}{2}\|x^+ - x\|(x^+ - x) \overset{(7)}{\in} \partial\psi(x^+).$$

Thus, usually for any solver of (6), along with x^+ we are able to compute the corresponding subgradient vector as well.

In what follows, we will consider *inexact minimizers* of our model. First, we provide the bound for the new gradient norm.

Lemma 1 *Let x^+ be an inexact minimizer of subproblem (6) satisfying the following condition, for some $\theta \geq 0$:*

$$\|\nabla M_{x,\sigma}(x^+) + \psi'(x^+)\| \leq \theta\|x^+ - x\|^2, \tag{8}$$

for a certain $\psi'(x^+) \in \partial\psi(x^+)$. Let, for some $\delta_g, \delta_B \geq 0$, it hold that

$$\begin{aligned}
\|g - \nabla f(x)\| &\leq \delta_g, \\
\|B - \nabla^2 f(z)\| &\leq \delta_B.
\end{aligned} \tag{9}$$

Then, we have

$$\|\nabla f(x^+) + \psi'(x^+)\| \leq \big(\theta + \tfrac{\sigma + L}{2}\big)r^2 + \big(\delta_B + L\|x - z\|\big)r + \delta_g, \tag{10}$$

where $r := \|x^+ - x\|$.

Proof Indeed,

$$\begin{aligned}
&\|\nabla f(x^+) + \psi'(x^+)\| \\
&\quad \leq \quad \|\nabla f(x^+) - \nabla \Omega_{x,\sigma}(x^+)\| + \|\nabla \Omega_{x,\sigma}(x^+) - \nabla M_{x,\delta}(x^+)\| \\
&\qquad\quad + \|\nabla M_{x,\sigma}(x^+) + \psi'(x^+)\| \\
&\quad = \quad \big\|\nabla f(x^+) - \nabla f(x) - \nabla^2 f(x)(x^+ - x) - \tfrac{\sigma}{2} r(x^+ - x)\big\| \\
&\qquad\quad + \|\nabla f(x) - g + (\nabla^2 f(x) - B)(x^+ - x)\| + \|\nabla M_{x,\sigma}(x^+) + \psi'(x^+)\| \\
&\quad \overset{(3),(8)}{\leq} \quad \big(\theta + \tfrac{\sigma + L}{2}\big) r^2 \, + \, \|\nabla^2 f(x) - B\| r \, + \, \|\nabla f(x) - g\| \\
&\quad \overset{(3),(9)}{\leq} \quad \big(\theta + \tfrac{\sigma + L}{2}\big) r^2 \, + \, \big(\delta_B + L\|x - z\|\big) r \, + \, \delta_g.
\end{aligned}$$

□

Now, we can express the progress of one step in terms of the objective function value.

Lemma 2 *Let x^+ satisfy the following condition:*

$$M_{x,\sigma}(x^+) + \psi(x^+) \leq F(x), \tag{11}$$

and let g and B satisfy (9) for some $\delta_g, \delta_B \geq 0$. Then, we have

$$F(x) - F(x^+) \geq \tfrac{\sigma - L}{6} r^3 - \tfrac{1}{2}(\delta_B + L\|x - z\|) r^2 - \delta_g r, \tag{12}$$

where $r := \|x^+ - x\|$.

Proof Indeed, we have

$$\begin{aligned}
F(x^+) \quad &\overset{(3)}{\leq} \quad \Omega_{x,L}(x^+) + \psi(x^+) \\
= \qquad & f(x) + \langle \nabla f(x), x^+ - x\rangle + \tfrac{1}{2}\langle \nabla^2 f(x)(x^+ - x), x^+ - x\rangle \\
&\quad + \tfrac{L}{6}\|x^+ - x\|^3 + \psi(x^+) \\
= \qquad & M_{x,\sigma}(x^+) + \langle \nabla f(x) - g, x^+ - x\rangle \\
&\quad + \tfrac{1}{2}\langle (\nabla^2 f(x) - B)(x^+ - x), x^+ - x\rangle + \tfrac{L - \sigma}{6}\|x^+ - x\|^3 + \psi(x^+) \\
\overset{(11),(9),(3)}{\leq} \quad & F(x) + \delta_g r + \tfrac{1}{2}(\delta_B + L\|x - z\|) r^2 + \tfrac{L - \sigma}{6} r^3,
\end{aligned}$$

and this is (12). □

Finally, we analyze the smallest eigenvalues for the Hessian of our problem. Let us consider the case when the composite part ψ is *twice differentiable*, so the Hessian of the full objective in (1) is well-defined. Then, we denote

$$\xi(y) \overset{\text{def}}{=} \max\Big\{ -\lambda_{\min}(\nabla^2 F(y)), 0 \Big\}, \qquad y \in Q. \tag{13}$$

Thus, the value of $\xi(y) \geq 0$ indicates how large the negative part of the smallest eigenvalue of the Hessian at point y is.

Note that if x^+ is an *exact solution* to our subproblem (6), we can use the following second-order optimality condition (see, e.g. Theorem 1.2.2 in [30]):

$$\begin{aligned} 0 &\preceq B + \tfrac{\sigma}{2}\|x^+ - x\|I + \tfrac{\sigma}{2r}(x^+ - x)(x^+ - x)^\top + \nabla^2\psi(x^+) \\ &\preceq B + \sigma\|x^+ - x\|I + \nabla^2\psi(x^+), \end{aligned} \tag{14}$$

where I is identity matrix. In order to provide the guarantee for $\xi(x^+)$, we can use the relaxed version of (14) with an inexact minimizer x^+ of our model and an arbitrary tolerance parameter $\theta \geq 0$.

Lemma 3 *Let ψ be twice differentiable. Let x^+ satisfy the following condition, for some $\theta \geq 0$:*

$$B + \theta\|x^+ - x\|I + \nabla^2\psi(x^+) \succeq 0. \tag{15}$$

Let, for some $\delta_B \geq 0$, it hold that

$$\|B - \nabla^2 f(z)\| \leq \delta_B. \tag{16}$$

Then, we have

$$\xi(x^+) \leq (L + \theta)r + L\|x - z\| + \delta_B. \tag{17}$$

where $r := \|x^+ - x\|$.

Proof Using Lipschitzness of the Hessian of f (3), we have

$$\begin{aligned} \nabla^2 F(x^+) &\succeq \nabla^2 f(x) + \nabla^2\psi(x^+) - LrI \\ &\succeq \nabla^2 f(z) + \nabla^2\psi(x^+) - (Lr + L\|x - z\|)I \\ &\overset{(16)}{\succeq} B + \nabla^2\psi(x^+) - (Lr + L\|x - z\| + \delta_B)I \\ &\overset{(15)}{\succeq} -(Lr + \theta r + L\|x - z\| + \delta_B)I, \end{aligned}$$

which leads to (17). □

Let us combine all our lemmas together. We justify the following bound for the progress of one step for our inexact composite Cubic Newton Method (CNM):

Theorem 1 *Let $\sigma \geq 2L$. Let x^+ be an inexact minimizer of model (5) satisfying the following two conditions, for a certain $\psi'(x^+) \in \partial\psi(x^+)$:*

$$\begin{aligned} \|\nabla M_{x,\sigma}(x^+) + \psi'(x^+)\| &\leq \tfrac{\sigma}{4}\|x^+ - x\|^2, \\ M_{x,\sigma}(x^+) + \psi(x^+) &\leq F(x), \end{aligned} \tag{18}$$

where g and B satisfy (9) for some $\delta_g, \delta_B \geq 0$. Then, we have

$$F(x) - F(x^+) \geq \tfrac{1}{3\cdot 2^6 \sigma^{1/2}}\|\nabla f(x^+) + \psi'(x^+)\|^{3/2} + \mathcal{E}, \tag{19}$$

where

$$\mathcal{E} \overset{\text{def}}{=} \tfrac{\sigma}{48}\|x^+ - x\|^3 - \tfrac{171}{\sigma^2}\Big[\delta_B^3 + L^3\|x - z\|^3\Big] - \tfrac{3}{\sigma^{1/2}}\delta_g^{3/2}.$$

Assume additionally that ψ *is twice differentiable, and* x^+ *satisfies the following extra condition:*

$$B + \sigma\|x^+ - x\| I + \nabla^2\psi(x^+) \succeq 0. \tag{20}$$

Then, we can improve (19) *, as follows:*

$$\begin{aligned} &F(x) - F(x^+) \\ &\geq \ \max\Big\{ \tfrac{1}{3\cdot 2^6 \sigma^{1/2}} \|\nabla f(x^+) + \psi'(x^+)\|^{3/2}, \ \tfrac{1}{2\cdot 3^6 \sigma^2}\big[\xi(x^+)\big]^3 \Big\} + \mathcal{E}. \end{aligned} \tag{21}$$

Proof We denote $r := \|x^+ - x\|$. Firstly, we bound the negative terms from (12), by using Young's inequality: $ab \leq \frac{a^3}{3} + \frac{2b^{3/2}}{3}$, $a, b \geq 0$. We have

$$\begin{aligned} \frac{1}{2}(\delta_B + L\|x-z\|)r^2 \ &= \ \Big[\frac{\sigma^{2/3} r^2}{2^{10/3}}\Big] \cdot \Big[\frac{2^{10/3}}{2\sigma^{2/3}} \cdot \big(\delta_B + L\|x-z\|\big)\Big] \\ &\leq \ \frac{2}{3}\Big[\frac{\sigma^{2/3} r^2}{2^{10/3}}\Big]^{3/2} + \frac{1}{3}\Big[\frac{2^{10/3}}{2\sigma^{2/3}} \cdot \big(\delta_B + L\|x-z\|\big)\Big]^3 \\ &= \ \frac{\sigma r^3}{3\cdot 2^4} + \frac{2^7}{3\sigma^2}\big(\delta_B + L\|x-z\|\big)^3 \ \leq \ \frac{\sigma r^3}{3\cdot 2^4} + \frac{2^9}{3\sigma^2}\big(\delta_B^3 + L^3\|x-z\|^3\big), \end{aligned} \tag{22}$$

and

$$\delta_g r = \Big[\tfrac{\sigma^{1/3} r}{2^{4/3}}\Big] \cdot \Big[\tfrac{2^{4/3}\delta_g}{\sigma^{1/3}}\Big] \ \leq \ \tfrac{\sigma r^3}{48} + \tfrac{2^3 \delta_g^{3/2}}{3\sigma^{1/2}}. \tag{23}$$

Therefore, for the functional progress, we obtain

$$F(x) - F(x^+) \overset{(12),(22),(23)}{\geq} 2\cdot\tfrac{\sigma}{48} r^3 \ - \ \tfrac{2^9}{3\sigma^2}\big(\delta_B^3 + L^3\|x-z\|^3\big) \ - \ \tfrac{2^3\delta_g^{3/2}}{3\sigma^{1/2}}. \tag{24}$$

Secondly, we can relate r and the new gradient norm by using (10). We get

$$\begin{aligned} \|\nabla f(x^+) + \psi'(x^+)\|^{3/2} \ &\overset{(10)}{\leq} \ \Big(\sigma r^2 + \delta_B r + L\|x-z\| r + \delta_g\Big)^{3/2} \\ &\overset{(*)}{\leq} \ 2\sigma^{1/2}\Big(\sigma r^3 \ + \ \frac{\delta_B^{3/2} r^{3/2}}{\sigma^{1/2}} \ + \ \frac{L^{3/2}\|x-z\|^{3/2} r^{3/2}}{\sigma^{1/2}} \ + \ \frac{\delta_g^{3/2}}{\sigma^{1/2}}\Big) \\ &\overset{(**)}{\leq} \ 2\sigma^{1/2}\Big(2\sigma r^3 + \frac{\delta_B^3}{2\sigma^2} + \frac{L^3\|x-z\|^3}{2\sigma^2} + \frac{\delta_g^{3/2}}{\sigma^{1/2}}\Big), \end{aligned} \tag{25}$$

where we used in $(*)$ Jensen's inequality: $(\sum_{i=1}^4 a_i)^{3/2} \leq 2\sum_{i=1}^4 a_i^{3/2}$ for non-negative numbers $\{a_i\}_{i=1}^4$, and in $(**)$ Young's inequality: $ab \leq \frac{a^2}{2} + \frac{b^2}{2}$, valid for any $a, b \geq 0$. Rearranging the terms, we obtain

$$\sigma r^3 \overset{(25)}{\geq} \tfrac{1}{4\sigma^{1/2}}\|\nabla f(x^+) + \psi'(x^+)\|^{3/2} - \tfrac{1}{4\sigma^2}\big(\delta_B^3 + L^3\|x-z\|^3\big) - \tfrac{\delta_g^{3/2}}{2\sigma^{1/2}}. \tag{26}$$

Combining (24) and (26) gives (19).

Finally, assuming twice differentiability of the composite part and using Lemma 3 for the extra condition (20) on x^+, we get

$$\begin{aligned} \big[\xi(x^+)\big]^3 \ &\overset{(17)}{\leq} \ \Big[\tfrac{3}{2}\sigma r + L\|x-z\| + \delta_B\Big]^3 \\ &\overset{(*)}{\leq} \ \tfrac{3^5}{2^3}\sigma^3 r^3 + 3^2 L^3\|x-z\|^3 + 3^2\delta_B^3, \end{aligned} \tag{27}$$

where we used in $(*)$ Jensen's inequality: $(\sum_{i=1}^{3} a_i)^3 \leq 3^2 \sum_{i=1}^{3} a_i$ for non-negative numbers $\{a_i\}_{i=1}^{3}$. Hence, rearranging the terms, we obtain

$$\sigma r^3 \overset{(27)}{\geq} \tfrac{2^3}{3^5 \sigma^2}\big[\xi(x^+)\big]^3 - \big(\tfrac{2}{3}\big)^3 \tfrac{1}{\sigma^2}\big(\delta_B^3 + L^3\|x - z\|^3\big).$$

Combining it with (24) justifies the improved bound (21). □

3 Finite Difference Approximations

In this section, we recall important bounds on finite difference approximations for the Hessian and for the gradient of our objective.

Let us start with the first-order approximation of the Hessian, that will lead us to the first-order (Hessian-free) implementation of the Cubic Newton Method. See, e.g., Lemma 3 in [20].

Lemma 4 *Suppose that A1 holds. Given $\bar{x} \in \mathbb{R}^n$ and $h > 0$, let $A \in \mathbb{R}^{n \times n}$ be defined by*

$$A = \left[\frac{\nabla f(\bar{x} + he_1) - \nabla f(\bar{x})}{h}, \ldots, \frac{\nabla f(\bar{x} + he_n) - \nabla f(\bar{x})}{h}\right]. \tag{28}$$

Then, the matrix

$$B = \tfrac{1}{2}\left(A + A^\top\right) \tag{29}$$

satisfies

$$\|B - \nabla^2 f(\bar{x})\| \leq \tfrac{\sqrt{n}L}{2} h. \tag{30}$$

Now, let us consider zeroth-order approximations of the derivatives, that requires computing only the objective function value (see, e.g., Section 7.1 in [33]). We establish explicit bounds necessary for the analysis of our methods and provide their proofs to ensure completeness of our presentation. The following lemma gives a zeroth-order approximation guarantee for the gradient.

Lemma 5 *Suppose that A1 holds. Given $\bar{x} \in \mathbb{R}^n$ and $h > 0$, let $g \in \mathbb{R}^n$ be defined by*

$$g_i = \frac{f(\bar{x} + he_i) - f(\bar{x} - he_i)}{2h}, \quad i = 1, \ldots, n. \tag{31}$$

Then,

$$\|g - \nabla f(\bar{x})\| \leq \tfrac{\sqrt{n}L}{6} h^2. \tag{32}$$

Proof By A1 we have

$$\left| f(\bar{x} + he_i) - f(\bar{x}) - h\langle \nabla f(\bar{x}), e_i\rangle - \frac{h^2}{2}\langle \nabla^2 f(\bar{x}) e_i, e_i\rangle \right| \leq \tfrac{Lh^3}{6} \tag{33}$$

and

$$\left| f(\bar{x}) - h\langle \nabla f(\bar{x}), e_i\rangle + \frac{h^2}{2}\langle \nabla^2 f(\bar{x}) e_i, e_i\rangle - f(\bar{x} - he_i) \right| \leq \tfrac{Lh^3}{6}. \tag{34}$$

Summing (33) and (34) and using the triangle inequality, we get

$$\left| f(\bar{x} + he_i) - f(\bar{x} - he_i) - 2h\,[\nabla f(\bar{x})]_i \right| \leq \tfrac{Lh^3}{3} \tag{35}$$

Therefore,

$$|g_i - [\nabla f(\bar{x})]_i| = \left|\tfrac{f(\bar{x}+he_i)-f(\bar{x}-he_i)}{2h} - [\nabla f(\bar{x})]_i\right| \overset{(35)}{\leq} \tfrac{Lh^2}{6}.$$

Thus, we conclude

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

□

Finally, we provide a zeroth-order approximation guarantee for the Hessian. A similar error bound was established recently in [6] (Proposition 2.7) for trust-region methods with noisy oracles.

Lemma 6 *Suppose that A1 holds. Given $\bar{x} \in \mathbb{R}^n$ and $h > 0$, let $A \in \mathbb{R}^{n\times n}$ be defined by*

$$A_{ij} = \frac{f(\bar{x}+he_i+he_j) - f(\bar{x}+he_i) - f(\bar{x}+he_j) + f(\bar{x})}{h^2}, \tag{36}$$

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

$$B = \tfrac{1}{2}\left(A + A^{\top}\right) \tag{37}$$

satisfies

$$\|B - \nabla^2 f(\bar{x})\| \leq \tfrac{(\sqrt{2}+1)nL}{3}h. \tag{38}$$

Proof By A1 we have the following inequalities:

$$\begin{aligned} &\Big|f(\bar{x}+he_i+he_j) - f(\bar{x}) - h\langle\nabla f(\bar{x}), e_i\rangle - h\langle\nabla f(\bar{x}), e_j\rangle \\ &-\tfrac{h^2}{2}\langle\nabla^2 f(\bar{x})e_i, e_i\rangle - h^2\langle\nabla^2 f(\bar{x})e_i, e_j\rangle - \tfrac{h^2}{2}\langle\nabla^2 f(\bar{x})e_j, e_j\rangle\Big| \leq \tfrac{\sqrt{2}Lh^3}{3}, \end{aligned} \tag{39}$$

$$\left|f(\bar{x}) + h\langle\nabla f(\bar{x}), e_i\rangle + \tfrac{h^2}{2}\langle\nabla^2 f(\bar{x})e_i, e_i\rangle - f(\bar{x}+he_i)\right| \leq \tfrac{Lh^3}{6}, \tag{40}$$

and

$$\left|f(\bar{x}) + h\langle\nabla f(\bar{x}), e_j\rangle + \tfrac{h^2}{2}\langle\nabla^2 f(\bar{x})e_j, e_j\rangle - f(\bar{x}+he_j)\right| \leq \tfrac{Lh^3}{6}. \tag{41}$$

Summing (39)–(41), and using the triangle inequality, we get

$$\begin{aligned} &\left|f(\bar{x}+he_i+he_j) - f(\bar{x}+he_i) - f(\bar{x}+he_j) + f(\bar{x}) - h^2\langle\nabla^2 f(\bar{x})e_i, e_j\rangle\right| \\ &\leq \tfrac{(\sqrt{2}+1)Lh^3}{3}. \end{aligned}$$

Hence,

$$h^2\left|\tfrac{f(\bar{x}+he_i+he_j)-f(\bar{x}+he_i)-f(\bar{x}+he_j)+f(\bar{x})}{h^2} - \left[\nabla^2 f(\bar{x})\right]_{ij}\right| \leq \tfrac{(\sqrt{2}+1)Lh^3}{3}$$

and, consequently,

$$\left|A_{ij} - \left[\nabla^2 f(\bar{x})\right]_{ij}\right| \leq \tfrac{(\sqrt{2}+1)L}{3}h.$$

Thus, we finally obtain

$$\|B - \nabla^2 f(\bar{x})\| \leq \|A - \nabla^2 f(\bar{x})\| \leq n\|A - \nabla^2 f(\bar{x})\|_{\max} \leq \tfrac{(\sqrt{2}+1)nL}{3}h,$$

which is the required bound. □

4 Hessian-Free CNM with Lazy Hessians

Let us present our first algorithm, which is the *Hessian-free* implementation of the Cubic Newton Method (CNM) [31]. In each iteration of our algorithm, we use an adaptive search to fit simultaneously the regularization constant σ and the parameter h of finite difference approximation of the Hessian (see Lemma 4). Therefore, our algorithm does not need to fix these parameters in advance, adjusting them automatically.

After the new approximation $B_{k,\ell} \approx \nabla^2 f(x_k)$ of the Hessian is computed, where $k \geq 0$ is the current iteration and ℓ is the adaptive search index, we keep using the same matrix $B_{k,\ell}$ for the next m Cubic Newton steps (6), where $m \geq 1$ is our global key parameter.

If we set $m := 1$, it means we update the Hessian approximation each Cubic Newton step, which can be costly from the computational point of view. Instead, we can use $m > 1$ (lazy Hessian updates [14]), that reuses the same Hessian approximation for several steps and thus reduces the arithmetic complexity.

Let us denote by $(\hat{x}, \alpha) = \texttt{CubicSteps}\ (x, B, \sigma, m, \epsilon)$ an auxiliary procedure that performs m inexact Cubic Newton steps (6), starting from point $x \in Q$ and using the same given matrix $B = B^\top$ and regularization constant $\sigma > 0$ for all steps, while recomputing the gradients. Parameter $\epsilon > 0$ is used for validating a certain stopping condition. We can write this procedure in the algorithmic form, as follows.

```
Algorithm 1: CubicSteps($x, B, \sigma, m, \epsilon$)
    Step 0. Set $x_0 := x$ and $t := 0$.
    Step 1. If $t = m$ then stop and return $(x_t,\ \texttt{success})$.
    Step 2. Compute $x_{t+1}$ as an approximate solution to the subproblem
        $\min_{y \in Q} \Big\{ M_{x_t,\sigma}(y) + \psi(y) \Big\}$,   where
        $M_{x_t,\sigma}(y) \equiv f(x_t) + \langle \nabla f(x_t), y - x_t \rangle + \frac{1}{2}\langle B(y - x_t), y - x_t \rangle + \frac{\sigma}{6}\|y - x_t\|^3$
    such that
        $M_{x_t,\sigma}(x_{t+1}) + \psi(x_{t+1}) \leq F(x_t)$   and
        $\|\nabla M_{x_t,\sigma}(x_{t+1}) + \psi'(x_{t+1})\| \leq \frac{\sigma}{4}\|x_{t+1} - x_t\|^2$   for   $\psi'(x_{t+1}) \in \partial\psi(x_{t+1})$,      (42)
    and (optionally, if $\psi$ is twice differentiable) such that
        $B + \sigma\|x_{t+1} - x_t\| I + \nabla^2\psi(x_{t+1}) \succeq 0$.      (43)
    Step 3. If $\|\nabla f(x_{t+1}) + \psi'(x_{t+1})\| \leq \epsilon$ then stop and return $(x_{t+1},\ \texttt{solution})$.
    Step 4. If $F(x_0) - F(x_{t+1}) \geq \frac{\epsilon^{3/2}}{384\sigma^{1/2}}(t+1)$ holds then set $t := t + 1$ and go to Step 1. Otherwise,
        stop and return $(x_{t+1},\ \texttt{halt})$.
```

This procedure returns the resulting point $\hat{x} \in Q$ and a status variable

$$\alpha \in \{\ \texttt{success},\ \ \texttt{solution},\ \ \texttt{halt}\}$$

that corresponds respectfully to finishing all the steps *successfully*, finding a point with *small gradient norm*, and *halting* the procedure due to insufficient progress in terms of the objective function. In the last case, we will need to update our estimates σ and B adaptively and restart this procedure with new parameters.

The next lemma shows that for sufficiently big value of σ and small enough h (the parameter of finite difference approximation of the Hessian), the result of Algorithm 1 always belongs to { `success`, `solution`}, that it either makes a significant progress in the function value, or solves the initial problem (1).

Lemma 7 *Suppose that A1 holds. Given* $x \in Q$, $\epsilon > 0$, $\sigma > 0$, *and* $m \in \mathbb{N} \setminus \{0\}$, *let* $(\hat{x}, \alpha)$ *be the corresponding output of Algorithm 1 with* $B = \frac{1}{2}(A + A^{\top})$, *where*

$$A = \left[\frac{\nabla f(x+he_1) - \nabla f(x)}{h}, \ldots, \frac{\nabla f(x+he_n) - \nabla f(x)}{h} \right] \tag{44}$$

for some $h > 0$. *If*

$$\sigma \geq \left(2^4 \cdot 3^2 \cdot 19\right)^{\frac{1}{3}} mL \quad \text{and} \quad h \leq \left[\frac{\sigma^{3/2}\epsilon^{3/2}}{(2^4 \cdot 3^3 \cdot 19) n^{3/2} L^3} \right]^{\frac{1}{3}}, \tag{45}$$

then either $\alpha =$ `solution` *(and thus* $\|\nabla f(\hat{x}) + \psi'(\hat{x})\| \leq \epsilon$*), or* $\alpha =$ `success` *and so we have*

$$F(x) - F(\hat{x}) \geq \frac{\epsilon^{3/2}}{2 \cdot (192) \sigma^{1/2}} m. \tag{46}$$

Proof Suppose that

$$\|\nabla f(\hat{x}) + \psi'(\hat{x})\| > \epsilon. \tag{47}$$

Hence, $\alpha \neq$ `solution`.

Let us denote by $t^\star$ the last value of t checked in Step 1. Clearly, $t^\star \leq m$ and we need just to prove that $t^\star = m$. Suppose that $t^\star < m$, and hence inequality in Step 4 of the algorithm does not hold for $t := t^\star$.

It follows from (44) and Lemma 4 that

$$\|B - \nabla^2 f(x)\| \leq \delta_B \tag{48}$$

for

$$\delta_B = \frac{\sqrt{n}L}{2} h.$$

Then, by the second inequality in (45) we get

$$\frac{171}{\sigma^2}\delta_B^3 = \frac{3^2 \cdot 19}{2^3} \cdot \frac{n^{3/2}L^3}{\sigma^2} \cdot h^3 \leq \frac{\epsilon^{3/2}}{2^7 \cdot 3\sigma^{1/2}} = \frac{\epsilon^{3/2}}{2 \cdot (192)\sigma^{1/2}}. \tag{49}$$

Hence, in view of (42) and (48), Theorem 1 with $\delta_g := 0$ and $z := x = x_0$ gives

$$\begin{aligned}
& F(x_t) - F(x_{t+1}) \\
& \geq \frac{\sigma}{48}\|x_{t+1} - x_t\|^3 + \frac{1}{192\sigma^{1/2}}\|\nabla f(x_{t+1}) + \psi'(x_{t+1})\|^{3/2} \\
& \quad - \frac{171}{\sigma^2}\delta_B^3 - \frac{171 L^3}{\sigma^2}\|x_t - x_0\|^3 \\
& \overset{(49)}{\geq} \frac{\sigma}{48}\|x_{t+1} - x_t\|^3 + \frac{1}{192\sigma^{1/2}}\|\nabla f(x_{t+1}) + \psi'(x_{t+1})\|^{3/2} \\
& \quad - \frac{\epsilon^{3/2}}{2(192)\sigma^{1/2}} - \frac{171 L^3}{\sigma^2}\|x_t - x_0\|^3 \\
& \overset{(47)}{\geq} \frac{\epsilon^{3/2}}{2(192)\sigma^{1/2}} + \frac{\sigma}{48}\|x_{t+1} - x_t\|^3 - \frac{171 L^3}{\sigma^2}\|x_t - x_0\|^3,
\end{aligned} \tag{50}$$

for any $0 \leq t \leq t^\star$. Finally, summing up these inequalities, and using the triangle inequality, we obtain

$$F(x_0) - F(x_{t^\star+1}) \geq \frac{\epsilon^{3/2}}{2(192)\sigma^{1/2}}(t^\star + 1) + \frac{\sigma}{48}\sum_{i=1}^{t^\star+1} r_i^3 - \frac{171L^3}{\sigma^2}\sum_{i=1}^{t^\star}\Big(\sum_{j=1}^{i} r_j\Big)^3,$$

where $r_i := \|x_i - x_{i-1}\|$. Using Lemma B.1 from [14] and our choice of σ (45) we conclude that

$$F(x_0) - F(x_{t^\star+1}) \geq \frac{\epsilon^{3/2}}{2(192)\sigma^{1/2}}(t^\star + 1),$$

which contradicts that inequality in Step 4 does not hold. Hence, $t^\star = m$ and $\alpha =$ `success`. □

Remark 1 Note that, using the same reasoning as in the proof of Lemma 7, one can show that unless an ϵ-solution was found, for all iterations $\{x_t\}_{1 \leq t \leq m}$ generated by Algorithm 1 it holds that

$$F(x_0) - F(x_t) \geq \frac{\epsilon^{3/2}}{384\sigma^{1/2}}t,$$

under assumptions (45).

For establishing the global convergence to a *second-order stationary point*, we can use our procedure with a stronger guarantee on the solution to the subrpoblem (43). This is optional. In case we use extra guarantee (43), the procedure should not be stopped in Step 3 anymore, since then we are interested in points with *both* small norm of the gradient and bounded smallest eigenvalue.

We can justify the following analogue of Lemma 7 when using condition (43):

Lemma 8 *Consider the sequence $\{x_t\}_{t=1}^m$ generated by Algorithm 1 with extra condition (43) on the inexact solution to the subproblem and without stop*[1] *in Step 3. Then, under the conditions of Lemma 7, we have either*

$$\min_{1 \leq t \leq m}\Bigg[\Delta_t \overset{\text{def}}{=} \max\Big\{\|\nabla f(x_t) + \psi'(x_t)\|,\ \frac{1}{\sigma}\Big(\frac{2}{3}\Big)^{\frac{10}{3}}\big[\xi(x_t)\big]^2\Big\}\Bigg] \leq \epsilon, \tag{51}$$

or

$$F(x) - F(\hat{x}) \geq \frac{\epsilon^{3/2}}{2\cdot(192)\sigma^{1/2}}m. \tag{52}$$

Proof Suppose that (51) does not hold, hence

$$\Delta_t \geq \epsilon, \qquad 1 \leq t \leq m. \tag{53}$$

In view of extra inexact condition (43), from Theorem 1 with $\delta_g := 0$ and $z := x = x_0$ we obtain the following guarantee for one step:

$$\begin{aligned} &F(x_t) - F(x_{t+1}) \\ &\overset{(21)}{\geq} \frac{\sigma}{48}\|x_{t+1} - x_t\|^3 + \frac{1}{192\sigma^{1/2}}\Delta_{t+1}^{3/2} - \frac{171}{\sigma^2}\delta_B^3 - \frac{171L^3}{\sigma^2}\|x_t - x_0\|^3 \\ &\overset{(49),(53)}{\geq} \frac{\epsilon^{3/2}}{2(192)\sigma^{1/2}} + \frac{\sigma}{48}\|x_{t+1} - x_t\|^3 - \frac{171L^3}{\sigma^2}\|x_t - x_0\|^3. \end{aligned}$$

It remains to sum up these inequalities for all $0 \leq t \leq m - 1$ and apply the same reasoning as in Lemma 7 to get (52). □

---

[1] Thus, α can be either `success` or `halt` in this case.

We are ready to present our whole algorithm, which is first-order implementation of CNM. It uses procedure `CubicSteps` as the basic subroutine.

Algorithm 2: First-Order CNM

```
Step 0. Given $x_0 \in Q$, $\tau_0 > 0$, $\epsilon > 0$, $m \in \mathbb{N} \setminus \{0\}$, set $k := 0$.
Step 1. Set $\ell := 0$.
Step 1.1. Using
    $\sigma_{k,\ell} = (2^4 \cdot 3^2 \cdot 19)^{1/3} (2^{\ell} \tau_k) m$    (54)
and
    $h_{k,\ell} = \Big[ \frac{\sigma_{k,\ell}^{3/2} \epsilon^{3/2}}{(2^4 \cdot 3^3 \cdot 19) n^{3/2} (2^{\ell} \tau_k)^3} \Big]^{1/3}$    (55)
compute $B_{k,\ell} = \frac{1}{2}\big(A_{k,\ell} + A_{k,\ell}^{\top}\big)$ with
    $A_{k,\ell} = \Big[ \frac{\nabla f(x_k + h_{k,\ell} e_1) - \nabla f(x_k)}{h_{k,\ell}}, \ \ldots, \ \frac{\nabla f(x_k + h_{k,\ell} e_n) - \nabla f(x_k)}{h_{k,\ell}} \Big].$    (56)
Step 1.2. Perform $m$ inexact Cubic steps using the same Hessian approximation:
    $(\hat{x}_{k,\ell}, \alpha_{k,\ell}) :=$ CubicSteps($x_k$, $B_{k,\ell}$, $\sigma_{k,\ell}$, $m$, $\epsilon$).
Step 2. If $\alpha_{k,\ell} =$ halt, then set $\ell := \ell + 1$ and go to Step 1.1.
Step 3. Set $x_{k+1} = \hat{x}_{k,\ell}$.
Step 4. If $\alpha_{k,\ell} =$ success, then set $\ell_k = \ell$, $\tau_{k+1} = \max\{\tau_0, \ 2^{\ell_k - 1} \tau_k\}$, $k := k + 1$, and go to Step 1. Stop otherwise.
```

Due to Lemmas 7 and 8, this algorithm is well-defined and its inner loop of the adaptive search (Steps 1-2) always quits with a sufficiently big finite value of ℓ and the method continues to Step 3.

In the following lemmas, we show how to bound the maximal value for the regularization parameter and the total number of inner loop steps in our algorithm.

Lemma 9 *Suppose that A1 holds and let $\{\tau_k\}_{k \geq 0}$ be generated by Algorithm 2. Then*

$$\tau_k \leq \max\big\{\tau_0, L\big\}, \qquad \forall k \geq 0. \tag{57}$$

Proof Clearly, (57) is true for $k = 0$. Suppose that it is also true for some $k \geq 0$. If $\ell_k = 0$, then it follows from the definition of τ_{k+1} and from the induction assumption that

$$\tau_{k+1} = \tfrac{1}{2}\tau_k \ < \ \tau_k \ \leq \ \max\big\{\tau_0, L\big\},$$

and so (57) is true for $k + 1$. Now, suppose that $\ell_k > 0$. In this case, we also must have

$$\tau_{k+1} \leq \max\big\{\tau_0, L\big\},$$

since otherwise we would have

$$2^{\ell_k - 1} \tau_k > L$$

and by (54), (55), (56) and Lemma 7, the inner procedure `CubicSteps` would return $\alpha_{k,\ell} \in \{$ `success`, `solution`$\}$ for some $\ell \leq \ell_k - 1$, contradicting the definition of ℓ_k. Thus, (57) is also true for $k + 1$ in this case. □

Lemma 10 *Suppose that A1 holds and let FO_T be the total number of function and gradient evaluations of $f(\cdot)$ performed by Algorithm 2 during the first T iteration. Then*

$$FO_T \leq \left[5 + 2(n+m)\right] \cdot T \; + \; \left[2 + n + m\right] \cdot \log_2 \tfrac{\max\{\tau_0, L\}}{\tau_0}. \tag{58}$$

Proof The total number of function and gradient evaluations performed at the $k$the iteration of Algorithm 2 is bounded from above by

$$1 + \left[(n+1) + (m+1)\right] \cdot (\ell_k + 1).$$

Since $\tau_{k+1} = \max\{\tau_0, 2^{\ell_k - 1}\tau_k\}$, we have $2^{\ell_k - 1} \leq \tau_{k+1}/\tau_k$, and so

$$\ell_k - 1 \leq \log_2 \tau_{k+1} - \log_2 \tau_k.$$

Therefore,

$$1 + \left[(n+1) + (m+1)\right] \cdot (\ell_k + 1) \leq 1 + \left[2 + n + m\right] \cdot (2 + \log_2 \tau_{k+1} - \log_2 \tau_k).$$

Thus,

$$\begin{aligned} \mathrm{FO}_T &\leq \sum_{k=0}^{T-1} 1 + \left[2 + n + m\right] \cdot (2 + \log_2 \tau_{k+1} - \log_2 \tau_k) \\ &= T + \left[2 + n + m\right] \cdot 2T + \left[2 + n + m\right] \cdot \log_2 \tfrac{\tau_T}{\tau_0} \\ &\leq \left[5 + 2(n+m)\right] \cdot T + \left[2 + n + m\right] \cdot \log_2 \tfrac{\max\{\tau_0, L\}}{\tau_0}, \end{aligned}$$

where the last inequality follows from Lemma 9. □

We are ready to establish the global complexity bound for our Hessian-free CNM.

Theorem 2 *Suppose that A1 holds and let $\{x_k\}_{k \geq 1}$ be generated by Algorithm 2. Let $T(\epsilon) \leq +\infty$ be the first iteration index such that $\|\nabla f(x_{T(\epsilon)}) + \psi'(x_{T(\epsilon)})\| \leq \epsilon$, for a certain $\psi'(x_{T(\epsilon)}) \in \partial\psi(x_{T(\epsilon)})$. We have*

$$T(\epsilon) \leq \tfrac{384 \cdot (2^7 \cdot 3^2 \cdot 19)^{1/6} \max\{\tau_0, L\}^{1/2} (F(x_0) - F^\star)}{\sqrt{m}} \cdot \epsilon^{-3/2} \tag{59}$$

and, consequently, the total number of the function and gradient evaluations $FO_{T(\epsilon)}$ is bounded as

$$\begin{aligned} &FO_{T(\epsilon)} \\ &\leq \quad 384(2^7 \cdot 3^2 \cdot 19)^{1/6} \tfrac{[5 + 2(n+m)]}{\sqrt{m}} \max\{\tau_0, L\}^{1/2} (F(x_0) - F^\star) \cdot \epsilon^{-3/2} \qquad (60) \\ &\quad + \; [2 + n + m] \log_2 \tfrac{\max\{\tau_0, L\}}{\tau_0}. \end{aligned}$$

Proof By the definition of $T(\epsilon)$, we have

$$\|\nabla f(x_k) + \psi'(x_k)\| \geq \epsilon, \quad \text{for} \quad k = 0, \ldots, T(\epsilon) - 1, \quad \text{and} \quad \forall \psi'(x_k) \in \partial\psi(x_k).$$

Consequently, by the combination of Algorithm 1 (Step 4) and Algorithm 2 (Step 2), and using Lemma 7 we have

$$F(x_k) - F(x_{k+1}) \geq \frac{\epsilon^{3/2} m}{(384)\sigma_{k,\ell_k}^{1/2}} \quad \text{for} \quad k = 0, \dots, T(\epsilon) - 1. \tag{61}$$

Moreover, by Lemma 9 we also have

$$\begin{aligned} \sigma_{k,\ell_k} &= (2^4 \cdot 3^2 \cdot 19)^{1/3} m (2^{\ell_k} \tau_k) \;\leq\; (2^4 \cdot 3^2 \cdot 19)^{1/3} m (2\tau_{k+1}) \\ &\leq (2^7 \cdot 3^2 \cdot 19)^{1/3} m \cdot \max\{\tau_0, L\}. \end{aligned} \tag{62}$$

Combining (61) and (62), it follows that

$$F(x_k) - F(x_{k+1}) \geq \frac{\epsilon^{3/2}\sqrt{m}}{384 \cdot (2^7 \cdot 3^2 \cdot 19)^{1/6} \max\{\tau_0, L\}^{1/2}}, \quad \text{for} \quad k = 0, \dots, T(\epsilon) - 1.$$

Summing up these inequalities and using the lower bound $F^\star$ on $F(\cdot)$, we get

$$\begin{aligned} F(x_0) - F^\star &\geq F(x_0) - F(x_{T(\epsilon)}) \\ &= \sum_{k=0}^{T(\epsilon)-1} F(x_k) - F(x_{k+1}) \\ &\geq \frac{\epsilon^{3/2}\sqrt{m}}{384 \cdot (2^7 \cdot 3^2 \cdot 19)^{1/6} \max\{\tau_0, L\}^{1/2}} T(\epsilon) \end{aligned}$$

which inplies (59). Finally, combining (59) and Lemma 10 we obtain (60). □

Corollary 1 *By taking $m := n$, it follows from Theorem 2 that Algorithm 2 needs at most*

$$\mathcal{O}\left(n^{1/2}\epsilon^{-3/2} + n\right)$$

total function and gradient evaluations of $f(\cdot)$ to generate x_k such that $\|\nabla f(x_k) + \psi'(x_k)\| \leq \epsilon$.

Let us establish a similar complexity result for reaching the *second-order* stationary points by Algorithm 2, providing the guarantee on the values of $\xi(\cdot)$ (see definition (13)).

Theorem 3 *Suppose that A1 holds. Let $x_{k,\ell}(t)$ be the t-th iterate of Algorithm 1 with extra condition (43) and without stop in Step 3, applied at the k-th iteration of Algorithm 2. Let $T(\epsilon) \leq +\infty$ be the first iteration index such that*

$$\max\left\{ \begin{array}{l} \|\nabla f(x_{T(\epsilon),\ell}(t)) + \psi'(x_{T(\epsilon),\ell}(t))\|, \\ \frac{1}{2^2 3^3 \cdot m \cdot \max\{\tau_0, L\}} \left[\xi(x_{T(\epsilon),\ell}(t))\right]^2 \end{array} \right\} \;\leq\; \epsilon, \tag{63}$$

for some $\ell \geq 0$ and $t \in \{0, \dots, m\}$. Then, bounds (59) and (60) hold.

Proof The proof is similar to those one of Theorem 2, using Lemma 8 instead of Lemma 7. □

Therefore, we conclude that our Hessian-free scheme achieves the second-order stationary guarantee (63), even though the method does not need to compute directly *any second-order information*, using solely the first-order oracle for $f(\cdot)$.

5 Zeroth-Order CNM

In this section, we present the *zeroth-order* implementation of the Cubic Newton Method, which uses only the *function evaluations* for $f(\cdot)$ to solve our optimization problem (1). Hence, we will use finite difference approximations *both* for the Hessian and for the gradients.

Note that approximating the Hessian matrix (37) remains to be *n times more expensive* than the gradient vector (31). Therefore, we keep using each approximation $B_{k,\ell} \approx \nabla^2 f(x_k)$ for consecutive $m \geq 1$ inexact cubic steps, while updating the gradient estimates each step. In what follows, we show that the optimal schedule is $\boxed{m := n}$, which gives the best zeroth-order oracle complexity for our scheme.

Let us denote by $(\hat{x}, \alpha) = \texttt{ZerothOrderCubicSteps}(x, B, \tau, m, \epsilon)$ an auxiliary procedure that performs m inexact Cubic Newton steps (6), starting from a point $x \in Q$, using the same given matrix $B = B^\top$, and estimating the new gradients with finite differences. We use $\sigma > 0$ as a regularization parameter, and $\epsilon > 0$ is the target accuracy (2). The procedure returns the last computed iterate $\hat{x}$ and a status variable

$$\alpha \in \{\texttt{success}, \texttt{halt}\},$$

which indicates whether the progress condition was satisfied for all steps or not. We define this procedure formally as Algorithm 3.

```
Algorithm 3: ZerothOrderCubicSteps($x, B, \sigma, m, \epsilon$)
    Step 0. Set $x_0 := x$ and $t := 0$.
    Step 1. If $t = m$ then stop and return $(x_t,$ success$)$.
    Step 2. For
        $h_g = \frac{1}{3^{1/3}} \left[ \frac{\epsilon m}{\sigma n^{1/2}} \right]^{1/2}$        (64)
    compute $g_t \in \mathbb{R}^n$ by
        $[g_t]^{(i)} = \frac{f(x_t + h_g e_i) - f(x_t - h_g e_i)}{2h_g}, \quad i = 1, \ldots, n.$        (65)
    Step 3. Compute $x_{t+1}$ as an approximate solution to the subproblem
        $\min_{y \in Q} \Big\{ M_{x_t,\sigma}(y) + \psi(y) \Big\}, \qquad \text{where}$
        $M_{x_t,\sigma}(y) \equiv f(x_t) + \langle g_t, y - x_t \rangle + \frac{1}{2}\langle B(y - x_t), y - x_t \rangle + \frac{\sigma}{6}\|y - x_t\|^3$
    such that
        $M_{x_t,\sigma}(x_{t+1}) + \psi(x_{t+1}) \leq F(x_t) \quad \text{and}$
        $\|\nabla M_{x_t,\sigma}(x_{t+1}) + \psi'(x_{t+1})\| \leq \frac{\sigma}{4}\|x_{t+1} - x_t\|^2 \ \text{ for } \ \psi'(x_{t+1}) \in \partial\psi(x_{t+1}),$        (66)
    and (optionally, if $\psi$ is twice differentiable) such that
        $B + \sigma\|x_{t+1} - x_t\| I + \nabla^2\psi(x_{t+1}) \succeq 0.$        (67)
    Step 4. If $F(x_0) - F(x_{t+1}) \geq \frac{\epsilon^{3/2}}{384\sigma^{1/2}}(t + 1)$ holds then set $t := t + 1$ and go to Step 1. Otherwise,
    stop and return $(x_{t+1},$ halt$)$.
```

We can prove the following main result about this procedure.

Lemma 11 *Suppose that A1 holds. Given $x \in Q$, $\epsilon > 0$, $\sigma > 0$, and $m \in \mathbb{N} \setminus \{0\}$, let $(\hat{x}, \alpha)$ be the corresponding output of Algorithm 3 with $B = \frac{1}{2}(A + A^\top)$, where*

$$A^{(i,j)} = \tfrac{f(x+he_i+he_j)-f(x+he_i)-f(x+he_j)+f(x)}{h^2}, \quad i, j = 1, \ldots, n, \tag{68}$$

for some $h > 0$. If

$$\sigma \geq (2^4 \cdot 3^2 \cdot 19)^{\frac{1}{3}} mL \quad \text{and} \quad h \leq \tfrac{1}{\sqrt{2}+1}\left[\tfrac{\sigma^{3/2}\epsilon^{3/2}}{(2^8 \cdot 19)n^3L^3}\right]^{\frac{1}{3}}, \tag{69}$$

then, for the iterations $\{x_t\}_{t=1}^m$ of Algorithm 3, we have either

$$\min_{t=1,\ldots,m} \|\nabla f(x_t) + \psi'(x_t)\| \leq \epsilon, \tag{70}$$

or

$$F(x) - F(\hat{x}) \geq \tfrac{\epsilon^{3/2}}{2(192)\sigma^{1/2}} m. \tag{71}$$

Proof By (65) and Lemma 5 we have

$$\|g_t - \nabla f(x_t)\| \leq \delta_g \tag{72}$$

for

$$\delta_g = \tfrac{\sqrt{n}L}{6} h_g^2. \tag{73}$$

In view of (64) and the assumption (69) it follows that

$$\begin{aligned} \tfrac{3}{\sigma^{1/2}} \cdot \delta_g^{3/2} &\overset{(73)}{=} \tfrac{3}{\sigma^{1/2}} \cdot \tfrac{n^{3/4}L^{3/2}}{6^{3/2}} h_g^3 \overset{(64)}{=} \tfrac{\epsilon^{3/2}}{2^8 3\sigma^{1/2}} \cdot \tfrac{1}{\sigma^{3/2}} \cdot \tfrac{2^{13/2}m^{3/2}L^{3/2}}{3^{1/2}} \\ &\overset{(69)}{\leq} \tfrac{\epsilon^{3/2}}{4(192)\sigma^{1/2}}. \end{aligned} \tag{74}$$

On the other hand, by (68) and Lemma 6 we have

$$\|B - \nabla^2 f(x)\| \leq \delta_B \tag{75}$$

for

$$\delta_B = \tfrac{(\sqrt{2}+1)nL}{3} h.$$

Then, in view of (69), it follows that

$$\begin{aligned} \tfrac{171}{\sigma^2} \cdot \delta_B^3 &= \tfrac{171}{\sigma^2} \cdot \tfrac{(\sqrt{2}+1)^3 n^3 L^3}{3^3} \cdot h^3 \leq \tfrac{171}{\sigma^2} \cdot \tfrac{n^3L^3}{3^3} \cdot \left[\tfrac{\sigma^{3/2}\epsilon^{3/2}}{(2^8 \cdot 19)n^3L^3}\right] \\ &= \tfrac{3^2 \cdot 19}{3^3 \cdot 2^8 \cdot 19} \cdot \tfrac{\epsilon^{3/2}}{\sigma^{1/2}} = \tfrac{\epsilon^{3/2}}{4(192)\sigma^{1/2}}. \end{aligned} \tag{76}$$

Combining (74) and (76), we have

$$\tfrac{171}{\sigma^2}\delta_B^3 + \tfrac{3}{\sigma^{1/2}}\delta_g^{3/2} \leq \tfrac{\epsilon^{3/2}}{2(192)\sigma^{1/2}}. \tag{77}$$

Then, by (66), (72), (75), (77) and Theorem 1 with $z = x$, we obtain

$$\begin{aligned} F(x_{t-1}) - F(x_t) &\geq \tfrac{\sigma}{48}\|x_t - x_{t-1}\|^3 + \tfrac{1}{192\sigma^{1/2}}\|\nabla f(x_t) + \psi'(x_t)\|^{3/2} \\ &\quad - \tfrac{1}{2(192)\sigma^{1/2}}\epsilon^{3/2} - \tfrac{171L^3}{\sigma^2}\|x_{t-1} - x_t\|^3, \end{aligned} \tag{78}$$

for $t = 1, \dots, m$. Consequently, if (70) is not true, then

$$F(x_{t-1}) - F(x_t) \geq \frac{\sigma}{48}\|x_t - x_{t-1}\|^3 + \frac{1}{2(192)\sigma^{1/2}}\epsilon^{3/2} - \frac{171L^3}{\sigma^2}\|x_{t-1} - x_0\|^3$$

for $t = 1, \dots, m$. Finally, summing up these inequalities and using Lemma B.1 in [14] for our choice (69) of σ, we conclude that (71) is true. □

Employing a stronger condition (67) on the solution to the subproblem, we can also justify the progress of our procedure in terms of the *second-order stationarity measure*.

Let us formulate our new optimization method for solving problem (1), which is the zeroth-order implementation of CNM (Algorithm 4).

Algorithm 4: Zeroth-Order CNM

```
Step 0. Given $x_0 \in Q$, $\tau_0 > 0$, $\epsilon > 0$, $m \in \mathbb{N} \setminus \{0\}$, set $k := 0$.
Step 1. Set $\ell := 0$.
Step 1.1. Using
    $\sigma_{k,\ell} = (2^4 \cdot 3^2 \cdot 19)^{1/3}(2^{\ell}\tau_k)m$    (79)
  and
    $h_{k,\ell} = \frac{1}{\sqrt{2}+1}\left[\frac{\sigma_{k,\ell}^{3/2}\epsilon^{3/2}}{(2^8 \cdot 19)n^3(2^{\ell}\tau_k)^3}\right]^{1/3}$    (80)
  compute $B_{k,\ell} = \frac{1}{2}\left(A_{k,\ell} + A_{k,\ell}^{\top}\right)$ with
    $\left[A_{k,\ell}\right]^{(i,j)} = \frac{f(x_k + h_{k,\ell}e_i + h_{k,\ell}e_j) - f(x_k + h_{k,\ell}e_i) - f(x_k + h_{k,\ell}e_j) + f(x_k)}{h_{k,\ell}^2}$    (81)
  for $i, j = 1, \dots, n$.
Step 1.2. Perform $m$ inexact zeroth-order Cubic steps with the same Hessian approximation:
    $(\hat{x}_{k,\ell}, \alpha_{k,\ell}) = \texttt{ZerothOrderCubicSteps}(x_k, B_{k,\ell}, \sigma_{k,\ell}, m, \epsilon)$.
Step 2. If $\alpha_{k,\ell} = \texttt{halt}$, then set $\ell := \ell + 1$ and go to Step 1.1.
Step 3. Set $\ell_k = \ell$, $x_{k+1} = \hat{x}_{k,\ell_k}$, $\tau_{k+1} = \max\{\tau_0,\ 2^{\ell_k - 1}\tau_k\}$, $k := k + 1$, and go to Step 1.
```

Lemma 12 *Consider the sequence $\{x_t\}_{t=1}^m$ generated by Algorithm 3 with extra condition (67) on the inexact solution to the subproblem. Then, under the assumptions of Lemma 11, we have either*

$$\min_{1 \leq t \leq m}\left[\Delta_t \overset{\text{def}}{=} \max\left\{\|\nabla f(x_t) + \psi'(x_t)\|,\ \frac{1}{\sigma}\left(\frac{2}{3}\right)^{\frac{10}{3}}\left[\xi(x_t)\right]^2\right\}\right] \leq \epsilon, \tag{82}$$

or

$$F(x) - F(\hat{x}) \geq \frac{\epsilon^{3/2}}{2 \cdot (192)\sigma^{1/2}}m. \tag{83}$$

Proof The proof follows the reasoning of Lemma 11, using the stronger one step guarantee provided by Theorem 1. □

Lemma 13 *Suppose that A1 holds and let* $\{\tau_k\}_{k\geq 0}$ *be generated by Algorithm 4. Then*

$$\tau_k \leq \max\{\tau_0, L\}, \qquad \forall k \geq 0. \tag{84}$$

Proof It follows exactly as in the proof of Lemma 9, using Lemma 11 to conclude that

$$\tau_{k+1} \leq \max\{\tau_0, L\}$$

when $\ell_k > 0$. □

Lemma 14 *Suppose that A1 holds and let* ZO_T *be the total number of function evaluations of* $f(\cdot)$ *performed by Algorithm 4 during the first* T *iterations. Then,*

$$ZO_T \leq \left[4 + 4mn + 6n^2\right] \cdot T \;+\; \left[2 + 2mn + 3n^2\right] \cdot \log_2 \tfrac{\max\{\tau_0, L\}}{\tau_0}.$$

Proof The number of function evaluations performed by Algorithm 4 (including those ones performed by Algorithm 3 in Step 2) is bounded from above by

$$[2 + 2mn + 3n^2] \cdot (\ell_k + 1).$$

Since $\tau_{k+1} = \max\{\tau_0, 2^{\ell_k - 1}\tau_k\}$, we have $2^{\ell_k - 1} \leq \tau_{k+1}/\tau_k$, and so

$$\ell_{k+1} \leq 2 + \log_2 \tau_{k+1} - \log_2 \tau_k.$$

Thus,

$$\begin{aligned} \mathrm{ZO}_T &\leq \sum_{k=0}^{T-1} [2 + 2mn + 3n^2] \cdot (2 + \log_2 \tau_{k+1} - \log_2 \tau_k) \\ &= [2 + 2mn + 3n^2] \cdot (2T + \log_2 \tau_T - \log_2 \tau_0) \\ &\leq [2 + 2mn + 3n^2] \cdot (2T + \log_2 \tfrac{\max\{\tau_0, L\}}{\tau_0}), \end{aligned}$$

where the last inequality follows from Lemma 13. □

We prove the following main result.

Theorem 4 *Suppose that A1. Let* $x_{k,\ell}(t)$ *be the* t*-th iterate of Algorithm 3 applied at the* k*-th iteration of Algorithm 4 in the* ℓ*-th inner loop. Let* $T(\epsilon) \leq +\infty$ *be the first iteration index such that*

$$\|\nabla f(x_{T(\epsilon),\ell}(t)) + \psi'(x_{T(\epsilon),\ell}(t))\| \leq \epsilon$$

for some $\ell \geq 0$ *and* $t \in \{0, \ldots, m\}$. *Then,*

$$T(\epsilon) \leq \tfrac{384 \cdot (2^7 \cdot 3^2 \cdot 19)^{1/6} \max\{\tau_0, L\}^{1/2} (f(x_0) - f^\star)}{\sqrt{m}} \epsilon^{-3/2} \tag{85}$$

and, consequently, the total number of the function evaluations is bounded as

$$\begin{aligned} ZO_{T(\epsilon)} \;\; &\leq \;\; \mathcal{O}\Big(\tfrac{mn + n^2}{\sqrt{m}} \max\{\tau_0, L\}^{1/2} (f(x_0) - f^\star) \cdot \epsilon^{-3/2} \\ &\quad + \; (mn + 3n^2) \log_2 \tfrac{\max\{\tau_0, L\}}{\tau_0} \Big). \end{aligned} \tag{86}$$

Proof Similarly to the proof of Theorem 2, we get (85) from Lemmas 11 and 13. Then, combining (85) with Lemma 14, we get (86). □

Corollary 2 *By taking* $\boxed{m := n}$, *it follows from Theorem 4 that Algorithm 4 needs at most*

$$\mathcal{O}(n^{3/2}\epsilon^{-3/2})$$

function evaluations of $f(\cdot)$ *to find a point* $\bar{x}$ *such that* $\|\nabla f(\bar{x}) + \psi'(\bar{x})\| \leq \epsilon$.

Finally, we can establish the convergence result in terms of the *second-order stationary point*. The proof is identical and it just needs to replace Lemma 11 by Lemma 12.

Theorem 5 *Suppose that A1 holds. Let* $x_{k,\ell}(t)$ *be the* t*-th iterate of Algorithm 3 with extra condition (67) on the inexact solution to the subproblem, applied at the* k*-th iteration of Algorithm 4 in the* ℓ*-th inner loop. Let* $T(\epsilon) \leq +\infty$ *be the first iteration index such that*

$$\max\left\{ \|\nabla f(x_{T(\epsilon),\ell}(t)) + \psi'(x_{T(\epsilon),\ell}(t))\|,\ \tfrac{1}{2^2 3^3\cdot m\cdot\max\{\tau_0, L\}}\left[\xi(x_{T(\epsilon),\ell}(t))\right]^2 \right\} \leq \epsilon$$

for some $\ell \geq 0$ *and* $t \in \{0, \ldots, m\}$. *Then, bounds (85) and (86) hold.*

6 Local Superlinear Convergence

One of the main classical results about Newton's Method is its *local quadratic convergence*, which dates back to the works of Fine [18], Bennett [3], and Kantorovich [25]. It assumes that the iterates of the method are already in a neighbourhood of a non-degenerate solution (a strict local minimum $x^\star$ satisfying $\nabla^2 f(x^\star) \succ 0$), and it shows importantly that under this condition the method converges very fast.

Later on, a local superlinear convergence of the Newton Method that uses the same Hessian for $m \geq 1$ consecutive steps, where m is a parameter, was established by Shamanskii in [34], and recently in [14]. The local quadratic convergence of the CNM with finite difference Hessian approximations was studied in [20].

In this section, we justify local superlinear convergence for our implementations of the inexact composite CNM. To quantify our problem class, we additionally assume the following[2]:

A2 The Hessian of f is below bounded on Q, for some $\mu > 0$:

$$\nabla^2 f(x) \succeq \mu I, \qquad \forall x \in Q. \tag{87}$$

It is well known that bound (87) means that our composite objective $F(\cdot)$ is *strongly convex* on Q with parameter $\mu > 0$. Thus, it has unique minimizer $x^\star \in Q$, and the following standard inequality holds [30]:

$$\|x - x^\star\| \leq \tfrac{1}{\mu}\|F'(x)\|, \qquad \forall x \in Q,\ F'(x) \in \partial F(x). \tag{88}$$

---

[2] Note that for simplicity we assume here strong convexity for the whole feasible set Q, while it can be possible to restrict our analysis to a neighbourhood of a non-degenerate local minimum.

Let us study one iteration $k \geq 0$ of our first-order CNM (Algorithm 2). First, we have the following bounds, for any $\ell \geq 0$:

$$\begin{aligned} \sigma_{k,\ell} &\overset{\textit{Step 1.1}}{=} (2^4 \cdot 3^2 \cdot 19)^{1/3}(2^\ell \tau_k)m \overset{\textit{Step 4, (57)}}{\leq} (2^7 \cdot 3^2 \cdot 19)^{1/3} m \cdot \max\{\tau_0, L\}, \\ h_{k,\ell} &\overset{\textit{Step 1.1}}{=} \left[\frac{\sigma_{k,\ell}^{3/2}\epsilon^{3/2}}{(2^4\cdot 3^3\cdot 19)n^{3/2}(2^\ell\tau_k)}\right]^{1/3} = \left[\frac{2^{3/2}m^{3/2}\epsilon^{3/2}}{(2^4\cdot 3^4\cdot 19)^{1/2}n^{3/2}(2^\ell\tau_k)^{3/2}}\right]^{1/3} \\ &= c \cdot \sqrt{\frac{m\epsilon}{2^\ell n \tau_k}} \leq c \cdot \sqrt{\frac{m\epsilon}{n\tau_k}} \overset{\textit{Step 4}}{\leq} c \cdot \sqrt{\frac{m\epsilon}{n\tau_0}}, \end{aligned} \tag{89}$$

where $c := \frac{1}{(2\cdot 3^4 \cdot 19)^{1/6}}$ is a numerical constant.

Let us consider the following set, for a fixed $\epsilon, \kappa > 0$ and some given selection of subgradients $F'(x) \in \partial F(x)$:

$$\mathcal{Q}_{\epsilon,\kappa} \overset{\text{def}}{=} \left\{ x \in Q \;:\; \epsilon \leq \|F'(x)\| \leq \tfrac{\kappa}{2} \right\}, \tag{90}$$

where κ is the following constant describing the *region of quadratic convergence*:

$$\kappa \overset{\text{def}}{=} \mu^2 \cdot \tfrac{1}{2}\left[3 \cdot 2^6 \left(\tfrac{2}{3}\right)^{1/3} m \cdot \max\{\tau_0, L\} + 8L + \frac{c^2 L^2 m}{\tau_0}\right]^{-1} \sim \frac{\mu^2}{mL}. \tag{91}$$

By (90), we assume the desired accuracy ϵ to be sufficiently small:

$$\epsilon \leq \tfrac{\kappa}{2} \leq \frac{\tau_0 \mu^2}{mL^2c^2}. \tag{92}$$

Then, we have

$$h_{k,\ell} \overset{(89)}{\leq} c \cdot \sqrt{\frac{m\epsilon}{n\tau_0}} \overset{(92)}{\leq} \frac{\mu}{L\sqrt{n}}. \tag{93}$$

Therefore, due to Lemma 4, for all Hessian approximations $B_{k,\ell}$ constructed in Algorithm 2, it holds:

$$\|B_{k,\ell} - \nabla^2 f(x_k)\| \overset{(30)}{\leq} \frac{\sqrt{n}L}{2} h_{k,\ell} \overset{(93)}{\leq} \frac{\mu}{2}. \tag{94}$$

Taking into account our assumption A2, we conclude that our Hessian approximations are *always positive definite*:

$$B_{k,\ell} \succeq \tfrac{\mu}{2} I. \tag{95}$$

In this case, we can easily bound the lenght of one inexact CNM step, as follows.

Lemma 15 *Let x^+ be an inexact minimizer of model (5) satisfying the following condition:*

$$\|\nabla M_{x,\sigma}(x^+) + \psi'(x^+)\| \leq \tfrac{\sigma}{4}\|x^+ - x\|^2, \tag{96}$$

for a certain $\psi'(x^+) \in \partial\psi(x^+)$, where g satisfy (9) for some $\delta_g \geq 0$, and $B \succeq \frac{\mu}{2}I$. Then, we have

$$r := \|x^+ - x\| \leq \tfrac{2}{\mu}\Big(\|F'(x)\| + \delta_g\Big), \qquad \forall F'(x) \in \partial F(x). \tag{97}$$

Proof Indeed, we get that

$$\begin{aligned} \tfrac{\sigma r^3}{4} &\overset{(96)}{\geq} \langle \nabla M_{x,\sigma}(x^+) + \psi'(x^+), x^+ - x \rangle \\ &= \langle g + B(x^+ - x) + \tfrac{\sigma}{2} r(x^+ - x) + \psi'(x^+), x^+ - x \rangle \\ &\geq \langle g + \psi'(x^+), x^+ - x \rangle + \tfrac{\mu r^2}{2} + \tfrac{\sigma r^3}{2}. \end{aligned} \tag{98}$$

Hence, rearranging the terms and using convexity of ψ, we obtain, for any $\psi'(x) \in \partial \psi(x)$,

$$\begin{aligned} \tfrac{\mu r^2}{2} &\overset{(98)}{\leq} \langle g + \psi'(x^+), x - x^+ \rangle - \tfrac{\sigma r^3}{4} \;\; \leq \;\; \langle g + \psi'(x), x - x^+ \rangle - \tfrac{\sigma r^3}{4} \\ &\leq r\Big(\|\nabla f(x) + \psi'(x)\| + \delta_g \Big) \;\; = \;\; r\Big(\|F'(x)\| + \delta_g \Big), \end{aligned}$$

which is (97). □

Now, let us look at the local progress given by one inexact CNM step $x \mapsto x^+$, with anchor point $z := x_k$. Assuming that $x \in \mathcal{Q}_{\epsilon,\kappa}$ and under assumptions of Lemma 1 with $\theta = \frac{\sigma_{k,\ell}}{4}$, $\delta_g = 0$, and $\delta_B \overset{(30)}{=} \frac{\sqrt{n}L}{2} h_{k,\ell}$, we get

$$\begin{aligned} \|F'(x^+)\| &\overset{(10)}{\leq} \Big(\tfrac{3}{4}\sigma_{k,\ell} + \tfrac{L}{2} \Big) r^2 + \Big(\tfrac{\sqrt{n}L}{2} h_{k,\ell} + L\|x - x_k\| \Big) r \\ &\overset{(89)}{\leq} \Big(\tfrac{3}{4}\sigma_{k,\ell} + \tfrac{L}{2} \Big) r^2 + \Big(\tfrac{cL}{2}\sqrt{\tfrac{m}{\tau_0}} \cdot \sqrt{\epsilon} + L\|x - x_k\| \Big) r \\ &\leq \Big(\tfrac{3}{4}\sigma_{k,\ell} + \tfrac{L}{2} + \tfrac{c^2L^2m}{8\tau_0} \Big) r^2 + Lr\|x - x^\star\| + Lr\|x_k - x^\star\| + \tfrac{\epsilon}{2} \\ &\overset{(97),(88)}{\leq} \tfrac{1}{\mu^2}\Big(3\sigma_{k,\ell} + 4L + \tfrac{c^2L^2m}{2\tau_0} \Big) \|F'(x)\|^2 + \tfrac{2L}{\mu^2}\|F'(x)\|\|F'(x_k)\| + \tfrac{\epsilon}{2}. \end{aligned} \tag{99}$$

We see that the first term in the right hand side of (99) is responsible for the local quadratic convergence in terms of the (sub)gradient norm, as in the classical Newton's Method, and the last two terms appear due to the inexactness of our Hessian approximations. It remains to combine all our observations together.

Theorem 6 *Suppose that A1 and A2 hold. Let $x_0 \in \mathcal{Q}_{\epsilon,\kappa}$, given by* (90) *and κ given by* (91)*. Let $\{x_k\}_{k \geq 1}$ be generated by Algorithm* 2 *and denote by $T(\epsilon) \leq +\infty$ be the first iteration index such that $\|\nabla f(x_{T(\epsilon)}) + \psi'(x_{T(\epsilon)})\| \leq \epsilon$, for a certain $\psi'(x_{T(\epsilon)}) \in \partial\psi(x_{T(\epsilon)})$. We have*

$$T(\epsilon) \leq \tfrac{1}{\log_2(1+m)} \log_2 \log_2 \tfrac{\kappa}{\epsilon} + 1. \tag{100}$$

Proof By the definition of $T(\epsilon)$, we have

$$\|F'(x_k)\| \;\; \equiv \;\; \|\nabla f(x_k) + \psi'(x_k)\| \geq \epsilon, \quad \text{for} \quad k = 0, \ldots, T(\epsilon) - 1,$$

and for all iterations generated by Algorithm 1 launched from Algorithm 2. We prove by induction that

$$\tfrac{1}{\kappa}\|F'(x_k)\| \leq \left(\tfrac{1}{2}\right)^{(1+m)^k + 1}, \qquad k = 0, \ldots, T(\epsilon) - 1, \tag{101}$$

which immediately leads to the desired bound.

For $k = 0$, inequality (101) holds due to our assumption: $x_0 \in \mathcal{Q}_{\epsilon,\kappa}$, and this is the base of our induction. Assume that it holds for some $k \geq 0$, and consider one iteration of

Algorithm 2. In *Step 1.2* it runs `CubicSteps` (Algorithm 1) and will do the adaptive search until gets status $\alpha_{k,\ell} = \texttt{success}$ ($\alpha_{k,\ell} = \texttt{solution}$ is impossible by our assumption).

Hence, x_{k+1} will be computed as m inexact Cubic steps performed from the point x_k. Denoting these steps by $x_k^0 \mapsto x_k^1 \mapsto \ldots \mapsto x_k^m$ ($x_k^0 \equiv x_k$ and $x_k^m \equiv x_{k+1}$), we conclude that, for each $0 \le t \le m-1$:

$$\begin{aligned} \|F'(x_k^{t+1})\| &\overset{(99),(89)}{\le} \tfrac{1}{2\kappa}\Big(\|F'(x_k^t)\|^2 + g_k\|F'(x_k^t)\|\Big) + \tfrac{\epsilon}{2} \\ &\le \tfrac{1}{2\kappa}\Big(\|F'(x_k^t)\|^2 + \|F'(x_k^0)\|\|F'(x_k^t)\|\Big) + \tfrac{1}{2}\|F'(x_k^{t+1})\|. \end{aligned} \tag{102}$$

Now, assuming that

$$\tfrac{1}{\kappa}\|F'(x_k^t)\| \le \left(\tfrac{1}{2}\right)^{(1+t)(1+m)^k+1} \tag{103}$$

(which holds for $t=0$ by (101)), we have

$$\begin{aligned} \frac{1}{\kappa}\|F'(x_k^{t+1})\| &\overset{(102)}{\le} \frac{1}{\kappa}\Big(\|F'(x_k^t)\| + \|F'(x_k^0)\|\Big)\cdot\frac{1}{\kappa}\|F'(x_k^t)\| \\ &\overset{(103)}{\le} \Big(\Big(\frac{1}{2}\Big)^{(1+t)(1+m)^k+1} + \Big(\frac{1}{2}\Big)^{(1+m)^k+1}\Big)\cdot\Big(\frac{1}{2}\Big)^{(1+t)(1+m)^k+1} \\ &\le \Big(\frac{1}{2}\Big)^{(1+t+1)(1+m)^k+1}. \end{aligned}$$

Thus, (103) holds for all $0 \le t \le m$, and for $t = m$ it gives (101) for the next iterate. □

Finally, let us discuss the local superlinear convergence for our derivative-free CNM (Algorithm 4), while the analysis remains similar to the Hessian-free version. For the derivative-free method, we have, for a fixed iteration $k \ge 0$ and for any $\ell \ge 0$:

$$\begin{aligned} \sigma_{k,\ell} &\overset{\textit{Step 1.1}}{=} (2^4\cdot 3^2\cdot 19)^{1/3}(2^\ell\tau_k)m \\ &\overset{\textit{Step 3},\,(84)}{\le} (2^7\cdot 3^2\cdot 19)^{1/3} m\cdot\max\{\tau_0, L\}, \qquad \text{and} \\ h_{k,\ell} &\overset{\textit{Step 1.1}}{=} \tfrac{1}{\sqrt{2}+1}\left[\tfrac{\sigma_{k,\ell}^{3/2}\epsilon^{3/2}}{(2^8\cdot 19)n^3(2^\ell\tau_k)^3}\right]^{1/3} \\ &= \tfrac{1}{\sqrt{2}+1}\left[\tfrac{3\epsilon^{3/2}m^{3/2}}{(2^{12}\cdot 19)^{1/2}n^3(2^\ell\tau_k)^{3/2}}\right]^{1/3} \overset{\textit{Step 3}}{\le} \tfrac{c_B\epsilon^{1/2}m^{1/2}}{n\tau_0^{1/2}}, \end{aligned} \tag{104}$$

where $c_B := \frac{3^{1/3}}{(\sqrt{2}+1)\cdot 2^2\cdot 19^{1/6}}$, and in each call of Algorithm 3, the gradient finite difference parameter is

$$h_g \overset{Step 2 in Alg.\ 3}{=} \tfrac{1}{3^{1/3}}\left[\tfrac{\epsilon m}{\sigma_{k,\ell}n^{1/2}}\right]^{1/2} \le c_g\cdot\tfrac{\epsilon^{1/2}}{n^{1/4}\tau_0^{1/2}}, \tag{105}$$

where $c_g := \frac{1}{(2^7\cdot 3^4\cdot 19)^{1/6}}$. Therefore, due to Lemmas 5 and 6, all our gradient and Hessian approximations used in Algorithms 3 and 4 satisfy the following guarantees:

$$\begin{aligned} \|g_t - \nabla f(x_t)\| &\overset{(32)}{\leq} \frac{\sqrt{n}L}{6} h_g^2 \overset{(105)}{\leq} \frac{c_g^2}{6}\cdot\frac{\epsilon L}{\tau_0}, \\ \|B_{k,\ell} - \nabla^2 f(x_k)\| &\overset{(38)}{\leq} \frac{(\sqrt{2}+1)nL}{3} h_{k,\ell} \overset{(104)}{\leq} \frac{(\sqrt{2}+1)c_B}{3}\cdot\frac{\epsilon^{1/2}m^{1/2}L}{\tau_0^{1/2}}. \end{aligned} \tag{106}$$

In particular, assuming that ϵ is sufficiently small (92), we ensure $\|B_{k,\ell} - \nabla^2 f(x_k)\| \leq \frac{\mu}{2}$, and hence our Hessian approximations are positive definite: $B_{k,\ell} \succeq \frac{\mu}{2} I$.

Let us assume that the initial regularization parameter is sufficiently big:

$$\tau_0 \geq \frac{2c_g^2 L}{3}. \tag{107}$$

Using Lemma 15, we can bound one (zeroth-order) inexact CNM step $x \mapsto x^+$ for a point $x \in \mathcal{Q}_{\epsilon,\kappa}$, as follows:

$$\begin{aligned} r \;:=\; \|x^+ - x\| &\overset{(97),(106)}{\leq} \frac{2}{\mu}\Big(\|F'(x)\| + \frac{c_g^2}{6}\cdot\frac{\epsilon L}{\tau_0}\Big) \\ &\overset{(107)}{\leq} \frac{2}{\mu}\Big(\|F'(x)\| + \frac{\epsilon}{4}\Big) \overset{(90)}{\leq} \frac{5}{2\mu}\|F'(x)\|. \end{aligned} \tag{108}$$

It remains to apply Lemma 1 with $\theta = \frac{\sigma_{k,\ell}}{4}$, $\delta_g = \frac{c_g^2 \epsilon L}{6\tau_0} \overset{(107)}{\leq} \frac{\epsilon}{4}$, $\delta_B = \frac{2c_B \epsilon^{1/2} m^{1/2} L}{3\tau_0^{1/2}}$ and anchor point $z := x_k$. We obtain

$$\begin{aligned} \|F'(x^+)\| &\overset{(10)}{\leq} \Big(\frac{3}{4}\sigma_{k,\ell} + \frac{L}{2}\Big) r^2 + \Big(\frac{2c_B}{3}\cdot\frac{\epsilon^{1/2}m^{1/2}L}{\tau_0^{1/2}} r + L\|x - x_k\|\Big) r + \frac{\epsilon}{4} \\ &\leq \Big(\frac{3}{4}\sigma_{k,\ell} + \frac{L}{2} + \frac{4c_B^2 m L^2}{9\tau_0}\Big) r^2 + Lr\|x - x^\star\| + Lr\|x_k - x^\star\| + \frac{\epsilon}{2} \\ &\overset{(108),(88)}{\leq} \frac{1}{\mu^2}\Big(\frac{75}{4}\sigma_{k,\ell} + \frac{45L}{8} + \frac{4c_B^2 m L^2}{9\tau_0}\Big)\|F'(x)\|^2 + \frac{5L}{2\mu^2}\|F'(x)\|\cdot\|F'(x_k)\| + \frac{\epsilon}{2}. \end{aligned}$$

We see that this inequality has the same structure as (99) established for the Hessian-free CNM. Applying bound (104), it is easy to verify that we can use the same local region, given by (90), (91). Therefore, repeating the previous reasoning, we prove the following local superlinear convergence.

Theorem 7 *Suppose that A1 and A2 hold. Let* $x_0 \in \mathcal{Q}_{\epsilon,\kappa}$, *given by* (90) *and* κ *given by* (91). *Let initial regularization parameter* τ_0 *be sufficiently big* (107). *Let* $x_{k,\ell}(t)$ *be the* t*-th iterate of Algorithm* 3 *applied at the* k*-th iteration of Algorithm* 4 *in the* ℓ*-th inner loop. Let* $T(\epsilon) \leq +\infty$ *be the first iteration index such that* $\|\nabla f(x_{T(\epsilon),\ell}) + \psi'(x_{T(\epsilon),\ell})\| \leq \epsilon$, *for some* $\ell \geq 0$ *and* $t \in \{0, \ldots m\}$. *Then,*

$$T(\epsilon) \leq \frac{1}{\log_2(1+m)} \log_2 \log_2 \frac{\kappa}{\epsilon} + 1. \tag{109}$$

Let us compare our results about the local convergence of our methods (Theorems 6 and 7) with the classical results [13] for iterations of general quasi-Newton methods. In [13], the authors establish an asymptotic local *superlinear* rate of convergence under the weak assumption of convergence for the Hessian approximation matrices along the directions of the update. In contrast, for our algorithms, we establish a local *quadratic* convergence rate with respect to every m-th iteration (where $m \geq 1$ is a parameter of the method), and a *linear* rate between these iterations. The price we pay for achieving the local quadratic rate is the need to approximate the Hessian with finite differences once every m steps. However, as we show in our theory, by making an appropriate choice of m, it is possible to reduce the arithmetic cost of the method while preserving a fast global rate.

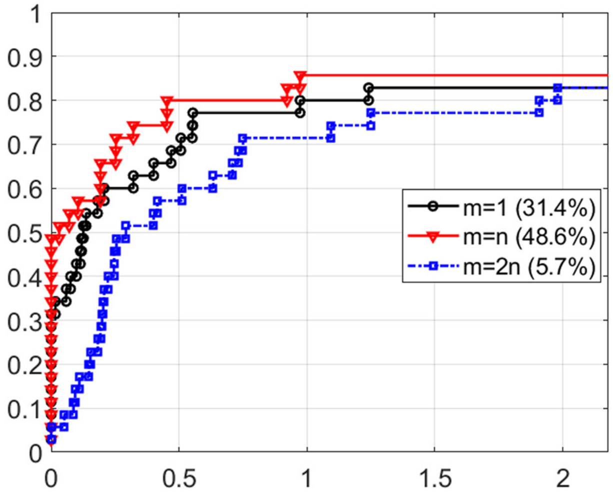

Fig. 1 Performance profiles in $\log_2$ scale for Algorithm 2. For each choice of m, the caption indicates the percentage of problems in which the corresponding code was the best in terms of number of calls of the oracle

7 Illustrative Numerical Experiments

We performed preliminary numerical experiments with Matlab implementations of the proposed methods applied to the set of 35 problems from the Moré-Garbow-Hillstrom collection [28][3] For both algorithms, we considered $\tau_0 = 1$ and $\epsilon = 10^{-4}$, allowing a maximum of 3, 000 calls of the oracle. Moreover, each cubic subproblem was approximately solved by a BFGS method with Armijo line search (using the origin as initial point).

Figure 1 presents the performance profiles [17][4] for Algorithm 2, comparing the variants with $m = 1$, $m = n$ and $m = 2n$ in terms of the number of calls of the oracle required to find the first ϵ-approximate stationary point. For each value x in x-axis, we show in y-axis percentage of the problems for which the corresponding code performs with a factor 2^x of the best performance among all the methods. In accordance with our theory, $m = n$ resulted in the best performance, with the corresponding code requiring less calls of the oracle in 48.6% of the problems.

We performed similar experiments with Algorithm 4, comparing the choices $m = 1$, $m = n$ and $m = 2n$ in terms of the number of function evaluations required to find $\bar{x}$ such that

$$f(\bar{x}) - f_{best} \leq \epsilon \left(f(x_0) - f_{best}\right). \tag{110}$$

For each problem, f_{best} is the smallest value of the objective function obtained by applying the three variants of Algorithm 4 with a budget of 3, 000 function evaluations. Figure 2 presents the corresponding performance profiles. Again, the variant with $m = n$ outperformed the others, requiring less function evaluations in 60.0% of the problems.

In addition, we compared the variants of Algorithm 4 with the MATLAB function **fminunc**, which implements a quasi-Newton method with finite-difference gradients. For each

---

[3] For each problem, n was chosen as in [5], resulting in a set of problems with dimensions ranging from 2 to 40.

[4] The performance profiles were generated using the code **perf.m** freely available in the website https://www.mcs.anl.gov/~more/cops/.

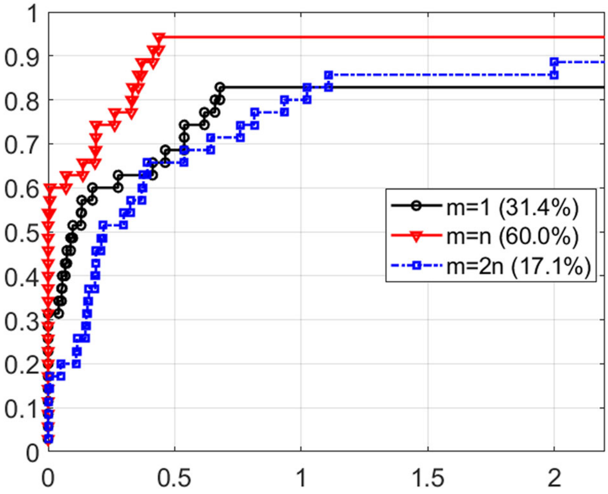

Fig. 2 Performance profiles in $\log_2$ scale for Algorithm 4

Table 1 Percentage of problems in which an ϵ-approximate solution was found

Solver/Precision	$\epsilon = 10^{-1}$	$\epsilon = 10^{-2}$	$\epsilon = 10^{-3}$	$\epsilon = 10^{-4}$
fminunc	100%	97%	94%	94%
Alg. 4 ($m = n$)	86%	71%	66%	54%
Alg. 4 ($m = 1$)	83%	68%	63%	51%
Alg. 4 ($m = 2n$)	83%	68%	63%	46%

problem, we allowed each solver a budget of $100(n + 1)$ function evaluations, where n is the number of variables in the corresponding problem. Table 1 presents the percentage of problems in which each solver found an ϵ-approximate solution in the sense of (110), while Table 2 reports the best function value achieved by each solver within the given budget of function evaluations.

As we can see in Tables 1 and 2, for most of the problems, Algorithm 4 with $m = n$ was able to find approximate solutions with function values comparable to, but slightly worse than, those obtained by **fminunc**. Notable exceptions are problems 4 and 14 in Table 2, where the poor performance of Algorithm 4 could be explained by the fact that the Hessians of the corresponding objective functions are not globally Lipschitz continuous.

8 Discussion

In this paper, we have developed new first-order and zeroth-order implementations of the Cubically Regularized Newton Method, that need, correspondingly, at most $\mathcal{O}(n^{1/2}\epsilon^{-3/2})$ and $\mathcal{O}(n^{3/2}\epsilon^{-3/2})$ calls of the oracle to find an ϵ-approximate second-order stationary point. Along with improved complexity guarantees, one of the main advantages of our schemes is the adaptive search, which makes the algorithms free from the need to fix the actual Lipschitz constant and the finite-difference approximation parameters.

Table 2 Results for the More-Garbow-Hillstrom problems

Problem	n	**fminunc**	$m = 1$	$m = n$	$m = 2n$
1. Rosenbrock	2	0.0000E+00	0.0007E+00	0.0310E+00	0.7791E+00
2. Freudenstein and Roth	2	4.8984E+01	4.8987E+01	4.8984E+01	4.8984E+01
3. Powell badly scaled	2	1.3518E−01	1.5390E−01	1.1340E−01	7.6836E−02
4. Brown badly scaled	2	9.3083E−05	9.7868E+11	9.8940E+11	9.9196E+11
5. Beale	2	3.2516E−14	4.8388E−12	1.1445E−12	2.3686E−10
6. Jenrich-Sampson	2	1.2436E+02	1.2436E+02	1.2436E+02	1.2486E+02
7. Helical Valley	3	5.4759E−15	2.0550E−02	1.3722E+00	1.2960E+00
8. Bard	3	8.2148E−03	9.0025E−03	8.8327E−03	8.5855E−03
9. Gaussian	3	1.1279E−08	1.1279E−08	1.1279E−08	1.1279E−08
10. Meyer	3	1.1212E+05	1.5728E+09	1.5728E+09	1.5399E+09
11. Gulf Res. and Dev	3	3.5662E−08	3.8500E−02	4.1820E−03	4.1983E−03
12. Box 3-Dimensional	3	2.0699E−15	2.3621E−01	2.1647E−01	1.3939E+02
13. Powell Singular	4	1.1644E−11	4.5673E−05	9.0348E−06	3.1404E−05
14. Wood	4	5.7534E−13	7.8371E+00	7.7964E+00	7.8528E+00
15. Kowalik and Osborne	4	3.0750E−04	3.8951E−04	3.8169E−04	3.8761E−04
16. Brown and Dennis	4	8.5822E+04	8.5822E+04	8.5822E+04	8.5822E+04
17. Osborne 1	5	7.7019E−05	4.8416E−01	5.4781E−01	4.4852E−01
18. Biggs EXP6	6	5.6556E−03	2.6781E−01	2.6233E−01	2.6626E−01
19. Osborne 2	11	4.0137E−02	3.0959E−01	3.5320E−01	3.8841E−01
20. Watson	6	2.2876E−03	8.8480E−03	9.8373E−03	1.3510E−02
21. Ext. Rosenbrock	10	1.0017E−10	8.2115E+00	7.7845E+00	9.2702E+00
22. Ext. Powell Singular	12	3.4932E−11	4.2907E−01	2.0729E−02	6.8610E−02
23. Penalty function I	4	2.2499E−05	2.4179E−05	2.4078E−05	2.4254E−05
24. Penalty function II	4	1.0936E−05	9.5536E−06	9.5538E−06	9.5539E−06
25. Variably dim. func	10	3.8700E−15	2.8365E+01	4.9209E+00	36.3277E+00
26. Trigonometric func	10	2.7950E−05	4.2014E−05	2.7957E−05	7.8759E−05
27. Brown almost linear	40	3.6288E−13	3.4055E−05	4.0736E−08	9.1610E−08
28. Disc. boundary val	10	1.5351E−12	6.2199E−05	1.8338E−05	2.3822E−05
29. Disc. integral eq	10	4.1779E−16	8.2681E−16	2.3860E−13	2.9883E−13
30. Broyden tridiagonal	10	8.9420E−14	4.5694E−11	1.1110E−11	1.1435E−11
31. Broyden banded	10	3.0572E+00	7.8009E−10	1.9226E−10	6.1805E−10
32. Linear	10	5.5510E−16	6.1154E−01	2.7622E−03	5.2793E−03
33. Linear-1	10	2.1428E+00	2.1428E+00	2.1428E+00	2.1428E+00
34. Linear-0	10	3.6470E+00	3.6470E+00	3.6470E+00	3.6470E+00
35. Chebyquad	8	3.5168E−03	9.8700E−03	4.5433E−03	4.8938E−03

While in this work we study the general class of non-convex optimization problems, it can be interesting to investigate the global performance of our methods for convex objectives. Indeed, it is well-known that, when the problem is convex, the rate of minimizing the gradient norm can be improved and the methods can be accelerated [22]. Hence, it seems to be an important direction for future research to study the complexities of first-order and zeroth-order regularized Newton schemes in convex case.

Another interesting question is related to comparison of our new schemes with derivative-free implementation of the first-order and direct-search methods [4, 19, 23, 32]. These methods need at most $\mathcal{O}(n\epsilon^{-2})$ function evaluations to find a first-order ϵ-stationary point (in expectation or with high probability for stochastic methods [4, 23, 32], or in terms of the full gradient norm for a deterministic method [19]). We see that bound $\mathcal{O}(n\epsilon^{-2})$ is worse than ours $\mathcal{O}(n^{3/2}\epsilon^{-3/2})$ in terms of dependence on ϵ, but has a better dimension factor. However, note that these complexity bounds are obtained for *different problem classes*, assuming either the first or second derivative to be Lipschitz continuous. Therefore, the development of universal schemes that can automatically achieve the best possible complexity bounds across various problem classes appears to be important, both from practical and theoretical perspectives.

Finally, it would be interesting to compare our methods with recent results on adaptive finite-difference methods [35], which automatically adjust the finite-difference interval to balance truncation error and measurement error, making them suitable for noisy derivative-free optimization. We keep these questions for further research.

Acknowledgements We are very grateful to the associate editor and the two anonymous referees for valuable comments that significantly improved the initial version of this paper.

Funding Open access funding provided by EPFL Lausanne. N. Doikov was supported by the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract number 22.00133. G. N. Grapiglia was partially funded by the FRS-FNRS, Belgium (Grant No. CDR J.0081.23).

Data Availability The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

References

1. Agarwal, N., Hazan, E.: Lower bounds for higher-order convex optimization. In: Conference on Learning Theory, pp. 774–792. PMLR (2018)
2. Arjevani, Y., Shamir, O., Shiff, R.: Oracle complexity of second-order methods for smooth convex optimization. Math. Program. **178**(1–2), 327–360 (2019)
3. Bennett, A.A.: Newton's method in general analysis. Proc. Natl. Acad. Sci. U.S.A. **2**(10), 592 (1916)
4. Bergou, E.H., Gorbunov, E., Richtárik, P.: Stochastic three points method for unconstrained smooth minimization. SIAM J. Optim. **30**(4), 2726–2749 (2020)
5. Birgin, E.G., Gardenghi, J., Martínez, J.M., Santos, S.A.: On the use of third-order models with fourth-order regularization for unconstrained optimization. Optimizat. Lett. **14**(4), 815–838 (2020)
6. Cao, L., Berahas, A.S., Scheinberg, K.: First-and second-order high probability complexity bounds for trust-region methods with noisy oracles. Mathematical Programming pp. 1–52 (2023)
7. Carmon, Y., Duchi, J.: Gradient descent finds the cubic-regularized nonconvex Newton step. SIAM J. Optim. **29**(3), 2146–2178 (2019)
8. Carmon, Y., Hausler, D., Jambulapati, A., Jin, Y., Sidford, A.: Optimal and adaptive Monteiro–Svaiter acceleration. Adv. Neural. Inf. Process. Syst. **35**, 20338–20350 (2022)
9. Cartis, C., Gould, N.I.M., Toint, P.H.L.: Adaptive cubic regularisation methods for unconstrained optimization Part I: Motivation, convergence and numerical results. Math. Programm. **127**(2), 245–295 (2011)

10. Cartis, C., Gould, N.I.M., Toint, P.L.: Adaptive cubic regularisation methods for unconstrained optimization. Part II: worst-case function-and derivative-evaluation complexity. Math. Programm. **130**(2), 295–319 (2011)
11. Cartis, C., Gould, N.I.M., Toint, P.L.: On the oracle complexity of first-order and derivative-free algorithms for smooth nonconvex minimization. SIAM J. Optim. **22**(1), 66–86 (2012)
12. Conn, A.R., Gould, N.I.M., Toint, PL.: Trust region methods. SIAM (2000)
13. Dennis, J.E., Moré, J.J.: A characterization of superlinear convergence and its application to quasi-Newton methods. Math. Comput. **28**(126), 549–560 (1974)
14. Doikov, N., Chayti, E.M., Jaggi, M.: Second-order optimization with lazy Hessians. In: International Conference on Machine Learning. PMLR (2023)
15. Doikov, N., Mishchenko, K., Nesterov, Y.: Super-universal regularized Newton method. SIAM J. Optim. **34**(1), 27–56 (2024)
16. Doikov, N., Nesterov, Y.: Minimizing uniformly convex functions by cubic regularization of Newton method. J. Optimizat. Theory Appl. **189**(1), 1–23 (2021)
17. Dolan, E.D., Moré, J.J.: Benchmarking optimization software with performance profiles. Math. Program. **91**, 201–213 (2002)
18. Fine, H.B.: On Newton's method of approximation. Proc. Natl. Acad. Sci. U.S.A. **2**(9), 546 (1916)
19. Grapiglia, G.N.: Worst-case evaluation complexity of a derivative-free quadratic regularization method. Optimizat. Lett. **18**(1), 195–213 (2023)
20. Grapiglia, G.N., Gonçalves, M.L., Silva, G.: A cubic regularization of Newton's method with finite difference Hessian approximations. Numer. Algorith. **90**, 607–630 (2022)
21. Grapiglia, G.N., Nesterov, Y.: Regularized Newton methods for minimizing functions with Hölder continuous Hessians. SIAM J. Optim. **27**(1), 478–506 (2017)
22. Grapiglia, G.N., Nesterov, Y.: Accelerated regularized Newton methods for minimizing composite convex functions. SIAM J. Optim. **29**(1), 77–99 (2019)
23. Gratton, S., Royer, C.W., Vicente, L.N., Zhang, Z.: Direct search based on probabilistic descent. SIAM J. Optim. **25**(3), 1515–1541 (2015)
24. Griewank, A.: The modification of Newton's method for unconstrained optimization by bounding cubic terms. Tech. rep., Technical report NA/12 (1981)
25. Kantorovich, L.V.: On Newton's method for functional equations. In: Dokl. Akad. Nauk SSSR, vol. 59, pp. 1237–1240 (1948)
26. Kovalev, D., Gasnikov, A.: The first optimal acceleration of high-order methods in smooth convex optimization. Adv. Neural. Inf. Process. Syst. **35**, 35339–35351 (2022)
27. Monteiro, R.D., Svaiter, B.F.: An accelerated hybrid proximal extragradient method for convex optimization and its implications to second-order methods. SIAM J. Optim. **23**(2), 1092–1125 (2013)
28. Moré, J.J., Garbow, B.S., Hillstrom, K.E.: Testing unconstrained optimization software. ACM Trans. Math. Softw. (TOMS) **7**(1), 17–41 (1981)
29. Nesterov, Y.: Accelerating the cubic regularization of Newton's method on convex problems. Math. Program. **112**(1), 159–181 (2008)
30. Nesterov, Y.: Lectures on convex optimization, vol. 137. Springer (2018)
31. Nesterov, Y., Polyak, B.T.: Cubic regularization of Newton's method and its global performance. Math. Program. **108**(1), 177–205 (2006)
32. Nesterov, Y., Spokoiny, V.: Random gradient-free minimization of convex functions. Found. Comput. Math. **17**, 527–566 (2017)
33. Nocedal, J., Wright, S.J.: Numerical optimization. Springer, Berlin (2006)
34. Shamanskii, V.: A modification of Newton's method. Ukr. Math. J. **19**(1), 118–122 (1967)
35. Shi, H.J.M., Xie, Y., Xuan, M.Q., Nocedal, J.: Adaptive finite-difference interval estimation for noisy derivative-free optimization. SIAM J. Sci. Comput. **44**(4), A2302–A2321 (2022)