THEORETICAL STUDY OF AN ADAPTIVE CUBIC REGULARIZATION METHOD WITH DYNAMIC INEXACT HESSIAN INFORMATION

STEFANIA BELLAVIA†, GIANMARCO GURIOLI‡ AND BENEDETTA MORINI†

Abstract. We consider the Adaptive Regularization with Cubics approach for solving nonconvex optimization problems and propose a new variant based on inexact Hessian information chosen dynamically. The theoretical analysis of the proposed procedure is given. The key property of ARC framework, constituted by optimal worst-case function/derivative evaluation bounds for first- and second-order critical point, is guaranteed. Application to large-scale finite-sum minimization based on sub-sampled Hessian is discussed and analyzed in both a deterministic and probabilistic manner.

Key words. Adaptive regularization with cubics; nonconvex optimization; worst-case analysis, finite-sum optimization.

1. Introduction. Numerical methods based on the Adaptive Regularization with Cubics (ARC) constitute an important class of Newton-type procedures for the solution of the unconstrained, possibly nonconvex, optimization problem

\[ \min_{x \in \mathbb{R}^n} f(x), \] (1.1)

where \( f : \mathbb{R}^n \to \mathbb{R} \) is smooth and bounded below. Successively to the seminal works [9, 10], ARC methods have become a very active area of research due to their worst-case iteration and computational complexity bounds for achieving a desired level of accuracy in first-order and second-order optimality conditions. Under reasonable assumptions on \( f \) and a suitable realization of the adaptive cubic regularization method with derivatives of \( f \) up to order 2, Cartis et al. proved that a first- and second-order critical point is found in at most \( O(\max(\epsilon^{-3/2}, \epsilon_H^{-3})) \) function and derivative evaluations where \( \epsilon \) and \( \epsilon_H \) are positive prefixed first-order and second-order optimality tolerances [4, 10, 11, 12]; this complexity result is known to be sharp and optimal with respect to steepest descent, Newton’s method and Newton’s method embedded into a linesearch or a trust-region strategy [8, 11].

Of particular practical interest is the ARC algorithm where exact second-derivatives of \( f \) are not required [9]. Inexact Hessian information is used and suitable approximations of the Hessian make the algorithm convenient for problems where the evaluation of second-derivatives is expensive. Clearly, the agreement between the Hessian and its approximation characterizes complexity and convergence rate behaviour of the procedure; in [9, 10] the well-known Dennis-Moré condition [16] and slightly stronger agreements are considered.

Recently, Newton-type methods with inexact Hessian information, and possibly inexact gradient information, have received large attention see e.g., [1, 2, 5, 6, 13, 14, 18, 22, 23, 24, 25, 26]. The interest in such methods is motivated by problems where the derivative information about \( f \) is computationally expensive, such as large-scale optimization problems arising in machine learning.

†Dipartimento di Ingegneria Industriale, Università di Firenze, viale G.B. Morgagni 40, 50134 Firenze, Italia, stefania.bellavia@unifi.it, benedetta.morini@unifi.it. Members of the INdAM Research Group GNCS.

‡Dipartimento di Matematica e Informatica “Ulisse Dini”, Università di Firenze, viale G.B. Morgagni 67a, 50134 Firenze, Italia, gianmarco.gurioli@unifi.it

*Work partially supported by INdAM-GNCS under Progetti di Ricerca 2018.
and data analysis modeled as

\[
\min_{x \in \mathbb{R}^n} f(x) = \frac{1}{N} \sum_{i=1}^{N} \phi_i(x), \tag{1.2}
\]

with \( N \) being a positive scalar and \( \phi_i : \mathbb{R}^n \rightarrow \mathbb{R} \). Experimental studies have shown that second-order methods can be more efficient on badly-scaled or ill-conditioned problems than first-order methods even though inexact Hessian information is built via random sampling methods, see e.g., [2, 5, 14, 18, 25, 26]. In addition, these methods seem to potentially take advantage of second-order curvature information to escape from saddle points [25]. ARC methods with probabilistic models have been proposed and studied in [13, 14, 18, 24, 25, 26]; much effort has been devoted to weaken the request on the level of resemblance between the Hessian and its approximation though preserving optimal complexity bounds.

This work focuses on a variant of the ARC methods for problem (1.1) with inexact Hessian information and presents a strategy for choosing the Hessian approximation dynamically. We propose a rule for fixing the desired accuracy in the Hessian approximation and incorporate it into the ARC framework; the agreement between the Hessian of \( f \) and its approximation can be loose at the beginning of the iterative process and increases progressively as the norm of stepsize drops below one and a stationary point for (1.1) is approached. The resulting ARC variant is supported by a milder adaptive condition on the inexact Hessian than the proposals in [14, 18, 24], though capable to maintain complexity results. The new algorithm is analyzed theoretically and first- and second-order optimal complexity bounds are proved in a deterministic manner; in particular, we show that the complexity bounds and convergence properties of our scheme match those of the ARC methods mentioned above. Our proposal has been motivated by the pervasiveness of finite-sum minimization problems (1.2) and the significant interest in unconstrained optimization methods with inexact Hessian information. Therefore, we discuss the application of our method to this relevant class of problems and show that it is compatible with sub-sampled Hessian approximations adopted in literature; in this context, we give probabilistic and deterministic results.

The paper is organized as follows. In Section 2 we briefly review the ARC framework, then in Section 3 we introduce our variant based on a dynamic rule for building the inexact Hessian. The first-order iteration complexity bound of the resulting algorithm is studied in Section 4 along with the asymptotic behaviour of the generated sequence; complexity bounds and convergence to second-order points are analyzed in Section 5. The application of our algorithm to the finite-sum optimization problem is discussed in Section 6, while the relevant differences of our proposal from the closely related works in the literature are discussed in Section 7.

**Notations.** The Euclidean vector and matrix norm is denoted as \( \| \cdot \| \). Given the scalar or vector or matrix \( v \), and the non-negative scalar \( \chi \), we write \( v = O(\chi) \) if there is a constant \( g \) such that \( \| v \| \leq g\chi \). Given any set \( S \), \( |S| \) denotes its cardinality.

**2. The adaptive regularization framework.** The ARC approach for unconstrained optimization, firstly proposed in [17, 21, 27], is based on the use of a cubic model for \( f \) and is a globally convergent second-order procedure. If \( f \) is smooth and the Hessian matrix \( \nabla^2 f \) is globally Lipschitz continuous on \( \mathbb{R}^n \) with \( \ell_2 \)-norm Lipschitz constant \( L \), i.e.,

\[
\| \nabla^2 f(x) - \nabla^2 f(y) \| \leq L \| x - y \|, \quad \forall x, y \in \mathbb{R}^n, \quad \exists L > 0,
\]
then the Taylor’s expansion of \( f \) at \( x_k \in \mathbb{R}^n \) with increment \( s \in \mathbb{R}^n \) implies

\[
f(x_k + s) \leq f(x_k) + \nabla f(x_k)^T s + \frac{1}{2} s^T \nabla^2 f(x_k) s + \frac{L}{6} \|s\|^3 \overset{\text{def}}{=} m^C(x_k, s).
\] (2.1)

Consequently, any step \( s \) satisfying \( m^C(x_k, s) < m^C(x_k, 0) = f(x_k) \) provides a reduction of \( f \) at \( x_k + s \) with respect to the current value \( f(x_k) \).

The ARC approach has received growing interest starting from the papers by Cartis et al. [9, 10] where it is not required the knowledge of either exact second-derivatives of \( f \) or the Lipschitz constant \( L \). Specifically, the cubic model used at iteration \( k \) has the form

\[
m(x_k, s, \sigma_k) = f(x_k) + \nabla f(x_k)^T s + \frac{1}{2} s^T B_k s + \frac{\sigma_k}{3} \|s\|^3 \overset{\text{def}}{=} T_2(x_k, s) + \frac{\sigma_k}{3} \|s\|^3,
\] (2.2)

where \( B_k \in \mathbb{R}^{n \times n} \) is a symmetric approximation of \( \nabla^2 f(x_k) \) and \( \sigma_k > 0 \) is the cubic regularization parameter chosen adaptively to ensure the overestimation property as in (2.1). The relevance of such procedure lies on its worst-case evaluation complexity for finding an \( \epsilon \)-approximate first-order critical point, i.e., a point \( \hat{x} \) such that

\[
\|\nabla f(\hat{x})\| \leq \epsilon.
\] (2.3)

In fact, in [10] worst-case iteration complexity of order \( O(\epsilon^{-3/2}) \) complexity bound is proved, provided that: (a) the step \( s_k \) is the global minimizer of \( m(x_k, s, \sigma_k) \) over a subspace of \( \mathbb{R}^n \) including \( \nabla f(x_k) \), see e.g. [3, 7, 9]; (b) the actual objective decrease \( f(x_k) - f(x_k + s_k) \) is a prefixed fraction of the predicted model reduction \( f(x_k) - m(x_k, s_k, \sigma_k) \), i.e.,

\[
\pi_k = \frac{f(x_k) - f(x_k + s_k)}{f(x_k) - m(x_k, s_k, \sigma_k)} \geq \eta_1,
\] (2.4)

for some \( \eta_1 \in (0, 1) \); (c) the agreement between \( \nabla^2 f(x_k) \) and \( B_k \) along \( s_k \) is such that

\[
\|\nabla^2 f(x_k) - B_k s_k\| \leq \chi \|s_k\|^2,
\] (2.5)

for all \( k \geq 0 \) and some constant \( \chi > 0 \).

A main advancement in ARC algorithm was obtained by Birgin et al. in the paper [4] where ARC is generalized to higher order regularized models and significant modifications in the step computation and acceptance criterion are introduced. The Algorithm 2.1 detailed below is proposed in [4] and here restricted to the version based on second order model and cubic regularization; as in [4] \( B_k \) is supposed to be equal to \( \nabla^2 f(x_k) \). Remarkably, global optimization of \( m(x_k, s, \sigma_k) \) over a subspace of \( \mathbb{R}^n \) is no longer required and conditions (2.7)–(2.8) on the step \( s_k \) are quite standard in unconstrained optimization when a model is approximately minimized. A further distinguishing feature is that the denominator in (2.9) involves the second-order Taylor expansion of \( f \) without the regularizing term, whereas the denominator in (2.4) involves the cubic model \( m(x_k, s, \sigma_k) \) itself. Analogously to the algorithm in [10], Algorithm 2.1 finds an \( \epsilon \)-approximation first-order critical point in at most \( O(\epsilon^{-3/2}) \) evaluations of \( f \) and its derivatives \( \nabla f, \nabla^2 f \) ([4]).

In this work, we propose a variant of Algorithm 2.1 employing a model of the form (2.2) and matrix \( B_k \) such that

\[
\|\nabla^2 f(x_k) - B_k\| \leq C_k, \quad C_k \leq C,
\] (2.6)
for all $k \geq 0$ and some positive scalars $C_k$ and $C$. Matrices $B_k$ satisfying (2.6) can be built approximating $\nabla^2 f(x)$ by finite differences or interpolating functions [15] while, in the relevant class of large-scale finite-sum minimization (1.2), condition (2.6) can be enforced in probability via subsampling Hessian approximations, see e.g., [5, 24]. The accuracy $C_k$ on the inexact Hessian information is dynamically chosen and when the norm of the step is smaller than one it depends on the current gradient’s norm. We will show that eventually condition (2.6) implies (2.5). Overall, our dynamic control is less stringent than the proposals made in [14, 18, 24, 26], though preserving the complexity bound $O(\epsilon^{-3/2})$. In the subsequent sections we present and study our variant of the ARC algorithm. We refer to Sections 6 and 7 for a discussion on the application to the finite-sum optimization problem and the comparison with the above mentioned related works in the literature.

### Algorithm 2.1: ARC algorithm [4]

**Step 0: Initialization.** Given an initial point $x_0$, the initial regularizer $\sigma_0 > 0$, the accuracy level $\epsilon$. Given $\theta$, $\eta_1$, $\eta_2$, $\gamma_1$, $\gamma_2$, $\gamma_3$, $\sigma_{\min}$ s.t.

\[
\theta > 0, \quad \sigma_{\min} \in (0, \sigma_0], \quad 0 < \eta_1 \leq \eta_2 < 1, \quad 0 < \gamma_1 < 1 < \gamma_2 < \gamma_3.
\]

Compute $f(x_0)$ and set $k = 0$.

**Step 1: Test for termination.** Evaluate $\nabla f(x_k)$. If $\|\nabla f(x_k)\| \leq \epsilon$, terminate with the approximate solution $\hat{x} = x_k$. Otherwise, compute $B_k = \nabla^2 f(x_k)$.

**Step 2: Step computation.** Compute the step $s_k$ by approximately minimizing the model $m(x_k, s, \sigma_k)$ w.r.t. $s$ so that

\[
\begin{align*}
m(x_k, s_k, \sigma_k) &< m(x_k, 0, \sigma_k), \quad \text{(2.7)} \\
\|\nabla_s m(x_k, s_k, \sigma_k)\| &\leq \theta \|s_k\|^2. \quad \text{(2.8)}
\end{align*}
\]

**Step 3: Acceptance of the trial step.** Compute $f(x_k + s_k)$ and define

\[
\rho_k = \frac{f(x_k) - f(x_k + s_k)}{T_2(x_k, 0) - T_2(x_k, s_k)}. \quad \text{(2.9)}
\]

If $\rho_k \geq \eta_1$, then define $x_{k+1} = x_k + s_k$; otherwise define $x_{k+1} = x_k$.

**Step 4: Regularization parameters update.** Set

\[
\sigma_{k+1} \in \begin{cases} 
\max(\sigma_{\min}, \gamma_1 \sigma_k), & \text{if } \rho_k \geq \eta_2, \\
\sigma_k, & \text{if } \rho_k \in [\eta_1, \eta_2), \\
\gamma_2 \sigma_k, & \text{if } \rho_k < \eta_1.
\end{cases} \quad \text{(2.10)}
\]

Set $k = k + 1$ and go to Step 1 if $\rho_k \geq \eta_1$, or to Step 2 otherwise.

### 3. An adaptive choice of the inexact Hessian.** In this section, we propose and study a variant of Algorithm 2.1 which maintains the complexity bound $O(\epsilon^{-3/2})$. Our algorithm is based on the use of an approximation $B_k$ of $\nabla^2 f(x_k)$ in the construction of the cubic model and a rule for choosing the level of agreement between $B_k$ and $\nabla^2 f(x_k)$. The accuracy requirements in the approximate minimization of $m(x_k, s, \sigma_k)$ consist of (2.7) and a condition on $\|\nabla_s m(x_k, s_k, \sigma_k)\|$ which includes condition (2.8) but it is not limited to it.
Our analysis is carried out under the following Assumptions on the function \( f \) and the matrix \( B_k \) used in the model (2.2).

**Assumption 3.1.** The objective function \( f \) is twice continuously differentiable on \( \mathbb{R}^n \) and its Hessian is Lipschitz continuous on the path of iterates with Lipschitz constant \( L \),

\[
\| \nabla^2 f(x_k + \alpha s_k) - \nabla^2 f(x_k) \| \leq L \alpha \| s_k \|, \quad \forall k \geq 0, \quad \alpha \in [0, 1].
\]

**Assumption 3.2.** For all \( k \geq 0 \) and some \( \kappa_B \geq 0 \), it holds

\[
\| B_k \| \leq \kappa_B.
\]

Further, we suppose that the step \( s_k \) computed has the following properties.

**Assumption 3.3.** For all \( k \geq 0 \) and some \( 0 \leq \theta_k \leq \theta, \theta \in (0, 1) \), \( s_k \) satisfies

\[
m(x_k, s_k, \sigma_k) < m(x_k, 0, \sigma_k), \quad (3.1)
\]

\[
\| \nabla m(x_k, s_k, \sigma_k) \| \leq \theta_k \| \nabla f(x_k) \|, \quad (3.2)
\]

By (2.1) and (2.2) it easily follows

\[
m^C(x_k, s) = T_2(x_k, s) + \frac{1}{2} s^T (\nabla^2 f(x_k) - B_k) s + \frac{L}{6} \| s \|^3. \quad (3.3)
\]

Then, (2.1) yields

\[
f(x_k + s) \leq T_2(x_k, s) + E_k(s), \quad (3.4)
\]

where

\[
E_k(s) = \frac{1}{2} \| \nabla^2 f(x_k) - B_k \| \| s \|^2 + \frac{L}{6} \| s \|^3. \quad (3.5)
\]

Now, we make our key requirement on the agreement between \( B_k \) and \( \nabla^2 f(x_k) \) and analyze its effects on ARC algorithm.

**Assumption 3.4.** Let \( B_k \in \mathbb{R}^{n \times n} \) satisfy

\[
\Delta_k = \nabla^2 f(x_k) - B_k, \quad \| \Delta_k \| \leq C_k, \quad (3.6)
\]

\[
C_k \leq C, \quad (3.7)
\]

\[
C_k \leq \alpha(1 - \theta) \| \nabla f(x_k) \|, \quad \text{if} \quad \| s_k \| < 1, \quad (3.8)
\]

for all \( k \geq 0 \), with \( \alpha, C_k \) and \( C \) positive scalars, \( s_k \in \mathbb{R}^n \) and \( \theta \in (0, 1) \) as in Assumption 3.3.

Bounds on \( \| \Delta_k \| \) and on \( E_k(s_k) \) involving \( \| s_k \| \) are derived below and show that \( E_k(s_k) = O(\| s_k \|^3) \).

**Lemma 3.1.** Let Assumptions 3.1–3.4 hold, and let \( E_k(s_k) \) and \( \Delta_k \) as in (3.5), (3.6). Then

\[
\| \Delta_k \| \leq \begin{cases}
    C_k \| s_k \|, & \text{if} \quad \| s_k \| \geq 1, \\
    \alpha(\kappa_B + \sigma_k) \| s_k \|, & \text{if} \quad \| s_k \| < 1,
\end{cases} \quad (3.9)
\]
and

\[
E_k(s_k) \leq \begin{cases} 
\frac{1}{2} \left( \frac{C_k + L}{3} \right) \|s_k\|^3, & \text{if } \|s_k\| \geq 1, \\
\frac{1}{2} \left( \alpha (\kappa_B + \sigma_k) + \frac{L}{3} \right) \|s_k\|^3, & \text{if } \|s_k\| < 1.
\end{cases}
\]  

(3.10)

Proof. First consider the case \(\|s_k\| \geq 1\). Trivially, the inequality in (3.6) gives (3.9) and

\[E_k(s_k) \leq \frac{1}{2} C_k \|s_k\|^3 + \frac{L}{6} \|s_k\|^3,\]

i.e., the first bound in (3.10).

Suppose now that \(\|s_k\| < 1\). Using (3.2), Assumptions 3.2 and 3.3 we obtain

\[
\theta \|\nabla f(x_k)\| \geq \|\nabla m(x_k, s_k, \sigma_k)\| \\
= \|\nabla f(x_k) + B_k s_k + \sigma_k s_k\| \\
\geq \|\nabla f(x_k)\| - \|B_k\| \|s_k\| - \sigma_k \|s_k\|^2 \\
\geq \|\nabla f(x_k)\| - \kappa_B \|s_k\| - \sigma_k \|s_k\|,
\]

which gives

\[\|s_k\| \geq \frac{(1 - \theta) \|\nabla f(x_k)\|}{\kappa_B + \sigma_k}.
\]

(3.12)

Thus, (3.6) yields

\[\|\Delta_k\| \leq C_k \frac{\|s_k\|}{\|s_k\|} \|s_k\| \leq \frac{C_k (\kappa_B + \sigma_k)}{(1 - \theta) \|\nabla f(x_k)\|} \|s_k\|.
\]

Finally, (3.8) implies (3.9) and this along with (3.5) gives (3.10).

We can now deduce an important upper bound on the regularization parameter \(\sigma_k\).

Lemma 3.2. Let Assumptions 3.1–3.4 hold. Suppose that the scalar \(\alpha\) in Assumption 3.4 is such that \(\alpha \in \left[0, \frac{2}{3}\right]\) and that the constant \(\eta_2\) in Algorithm 2.1 is such that \(\eta_2 \in \left(0, \frac{2 - 3\alpha}{2}\right)\). Then it holds

\[
\sigma_k \leq \sigma_{\text{max}} \overset{\text{def}}{=} \max \left\{ \sigma_0, \frac{3C + L}{2(1 - \eta_2)}, \frac{3\alpha \kappa_B + L}{2 - 3\alpha - 2\eta_2} \right\} \quad \forall k \geq 0,
\]

(3.13)

where \(\gamma_3\) is the constant used in (2.10).

Proof. First, we establish when the overestimation property \(f(x_k + s_k) \leq m(x_k, s_k, \sigma_k)\) is verified. Using (2.2), (3.3), (3.4) we see that if \(E_k(s_k) \leq \sigma_k \|s_k\|^3/3\), then \(m^C(x_k, s_k) \leq m(x_k, s_k, \sigma_k)\) which implies that \(m(x_k, s_k, \sigma_k)\) overestimates \(f(x_k + s)\).

If \(\|s_k\| \geq 1\) and

\[
\frac{1}{2} \left( C + \frac{L}{3} \right) \leq \frac{\sigma_k}{3}, \quad \text{i.e.,} \quad \sigma_k \geq \frac{3C + L}{2},
\]
then (3.10) implies $m^C(x_k, s_k) \leq m(x_k, s_k, \sigma_k)$. Analogously, if $\|s_k\| < 1$ and

$$
\frac{1}{2} \left( \alpha(\kappa_B + \sigma_k) + \frac{L}{3} \right) \leq \frac{\sigma_k}{3}, \quad i.e., \quad \sigma_k \geq \frac{3\alpha\kappa_B + L}{2 - 3\alpha},
$$

then (3.10) implies $m^C(x_k, s_k) \leq m(x_k, s_k, \sigma_k)$. Note that $2 - 3\alpha > 0$ by assumption.

Now, we search for conditions on $\sigma_k$ ensuring $\rho_k \geq \eta_2$. By (2.2) and (3.1) it follows $\|s_k\| \neq 0$ and

$$
0 < m(x_k, 0, \sigma_k) - m(x_k, s_k, \sigma_k) = T_2(x_k, 0) - T_2(x_k, s_k) - \frac{\sigma_k}{3}\|s_k\|^3.
$$

Thus

$$
T_2(x_k, 0) - T_2(x_k, s_k) > \frac{\sigma_k}{3}\|s_k\|^3 > 0,
$$

and by (3.4) and the fact that $E_k(s_k) > 0$

$$
1 - \rho_k = \frac{f(x_k + s_k) - T_2(x_k, s_k)}{\sigma_k} \leq \frac{E_k(s_k)}{T_2(x_k, 0) - T_2(x_k, s_k)} \leq \frac{3E_k(s_k)}{\sigma_k\|s_k\|^3}.
$$

If $\|s_k\| \geq 1$, using (3.7) and (3.10) we obtain

$$
1 - \rho_k < \frac{3}{2\sigma_k} \left( C + \frac{L}{3} \right),
$$

and $\rho_k \geq \eta_2$ is guaranteed when

$$
\sigma_k \geq \frac{3C + L}{2(1 - \eta_2)}.
$$

On the other hand, if $\|s_k\| < 1$ then (3.10) and (3.16) give

$$
1 - \rho_k < \frac{3}{2\sigma_k} \left( \alpha(\kappa_B + \sigma_k) + \frac{L}{3} \right),
$$

and $\rho_k \geq \eta_2$ is guaranteed when

$$
\sigma_k \geq \frac{3\alpha\kappa_B + L}{2 - 3\alpha - 2\eta_2},
$$

noting that the denominator is strictly positive by assumption. Then, the updating rule (2.10) implies $\sigma_{k+1} \leq \sigma_k$ in case $\rho_k \geq \eta_2$ and, more generally, inequality (3.13). [ ]

We observe that the value of $\alpha$ in (3.8) determines the accuracy of $B_k$ as an approximation to $\nabla^2 f(x_k)$ and the admitted maximum value of $\eta_2$. For decreasing values of $\alpha$, the accuracy of the Hessian approximation increases and $\eta_2$ reaches one. On the other hand, if $\alpha$ tends to $\frac{2}{3}$ then the accuracy of the Hessian approximation reduces, $\eta_2$ tends to zero and $\sigma_{\text{max}}$ tends to infinity. [ ]

*Values $\eta_2 = \frac{3}{4}$ and $\eta_2 = \frac{9}{10}$ used in the literature for the trust-region and ARC frameworks are achieved setting $\alpha = \frac{1}{6}$ and $\alpha = \frac{1}{15}$ respectively.
We also note that as long as Lemma 3.1 and 3.2 hold, (3.9) implies (2.5).

On the base of the previous analysis we sketch our version of Algorithm 2.1 denoted as Algorithm 3.1. The main feature is the adaptive rule for choosing matrices $B_k$ which are supposed to satisfy Assumption 3.4. At the beginning of $k$th iteration, the variable flag is equal to either 1 or 0 and determines the value of $C_k$; specifically $C_k = C_0$ if flag = 1, $C_k = \alpha (1 - \theta) \| \nabla f(x_k) \|$ otherwise with $\nabla f(x_k)$ being available (at iteration $k = 0$, flag is set equal to 1). Scalars $C_0$ and $\alpha$ are initialized at Step 0; the choice of $\alpha$ and $\eta_2$ is in accordance to the results presented above. Then $B_k$ is computed at Step 2 and the trial step $s_k$ is computed at Step 3.

**Algorithm 3.1:** ARC algorithm with dynamic Hessian accuracy

**Step 0: Initialization.** Given an initial point $x_0$, the initial regularizer $\sigma_0 > 0$, the accuracy level $\epsilon$. Given $\theta_0$, $\theta$, $\alpha$, $\eta_1$, $\eta_2$, $\gamma_1$, $\gamma_2$, $\gamma_3$, $\sigma_{\text{min}}$, $C_0$ s.t.

$$0 < \theta_0 \leq \theta, \quad \alpha \in \left(0, \frac{2}{3}\right), \quad \sigma_{\text{min}} \in (0, \sigma_0], \quad 0 < \eta_1 \leq \eta_2 < \frac{2 - 3\alpha}{2}, \quad 0 < \gamma_1 < 1 < \gamma_2 < \gamma_3, \quad C_0 > 0$$

Compute $f(x_0)$ and set $k = 0$, flag = 1.

**Step 1: Test for termination.** If $\| \nabla f(x_k) \| \leq \epsilon$, terminate with the current solution $\hat{x} = x_k$.

**Step 2: Hessian approximation.** Compute $B_k$ satisfying (3.6).

**Step 3: Step computation.** Choose $\theta_k \leq \theta$. Compute the step $s_k$ satisfying (3.1) and (3.2).

**Step 4: Check on $\| s_k \|$.**

If $\| s_k \| < 1$ and flag = 1 and $C_0 > \alpha (1 - \theta) \| \nabla f(x_k) \|$ set $x_{k+1} = x_k$, $\sigma_{k+1} = \sigma_k$, (unsuccessful iteration) set $C_{k+1} = \alpha (1 - \theta) \| \nabla f(x_k) \|$, flag = 0, set $k = k + 1$ and go to Step 2.

**Step 5: Acceptance of the trial step and parameters update.**

Compute $f(x_k + s_k)$ and $\rho_k$ in (2.9). If $\rho_k \geq \eta_1$ define $x_{k+1} = x_k + s_k$, set

$$\sigma_{k+1} \in \begin{cases} \max(\sigma_{\text{min}}, \gamma_1 \sigma_k), \sigma_k, \quad \text{if } \rho_k \geq \eta_2, \\ \sigma_k, \gamma_2 \sigma_k, \quad \text{if } \rho_k \in (\eta_1, \eta_2), \end{cases}$$

(very successful iteration)

(successful iteration)

If $\| s_k \| \geq 1$ set $C_{k+1} = C_0$, flag = 1. Otherwise set $C_{k+1} = \alpha (1 - \theta) \| \nabla f(x_{k+1}) \|$, flag = 0. Set $k = k + 1$ and go to Step 1.

else define $x_{k+1} = x_k$, $\sigma_{k+1} \in [\gamma_2 \sigma_k, \gamma_3 \sigma_k]$, unsuccessful iteration

$C_{k+1} = C_k$, $B_{k+1} = B_k$, set $k = k + 1$ and go to Step 3.

Step 4 is devoted to a check on the accordance between $C_k$ and $\| s_k \|$. In fact, (3.8) is required to hold if $\| s_k \| < 1$ whereas $\| s_k \|$ is not available when $B_k$ is formed. Therefore, if $\| s_k \| < 1$, flag = 1 and $C_0 > \alpha (1 - \theta) \| \nabla f(x_k) \|$ then the step is rejected and the iteration is unsuccessful. Variable flag is set equal to 0 and $B_k$ is recomputed at the successive iteration. This unsuccessful iteration is ascribed to the choice of matrix $B_k$, then the regularization parameter is left unchanged. On
the other hand, if the level of accuracy in matrix $B_k$ with respect to $\nabla^2 f(x_k)$ fulfills the requests (3.7)–(3.8), in Step 5 we proceed for acceptance of the trial steps and update of the regularizing parameter as in Algorithm 2.1. Summarizing, by construction, Assumption 3.4 is satisfied at every successful iteration and at any unsuccessful iteration detected in Step 5.

Finally, both flag and $C_k$ are updated in Step 5 as follows. If the iteration is successful, we update flag and $C_k$ following (3.7)–(3.8) and using the norm of the accepted trial step; clearly, this is a prediction as the step $s_{k+1}$ is not available at this stage and such a setting may be rejected at Step 4 of the successive iteration. If the iteration is unsuccessful, then we do not change either $C_k$ or $B_k$.

The classification of successful and unsuccessful iterations of the Algorithm 3.1 between 0 and $k$ can be made introducing the sets

$$S_k = \{ 0 \leq j \leq k \mid j \text{ successful in the sense of Step 5} \} ,$$

$$U_{k,1} = \{ 0 \leq j \leq k \mid j \text{ unsuccessful in the sense of Step 5} \} ,$$

$$U_{k,2} = \{ 0 \leq j \leq k \mid j \text{ unsuccessful in the sense of Step 4} \} .$$

More insight into the settings of $C_k$ and $\sigma_k$ in our algorithm, first note that $C_k$ satisfies

$$C_k = \alpha \omega(s_k)(1 - \theta)\|\nabla f(x_k)\| + (1 - \omega(s_k))C_0,$$

where $\omega : W \rightarrow \{ 0, 1 \}$ denotes the characteristic function of $W = \{ s_k : \| s_k \| < 1 \}$. It follows that if

$$\|\nabla f(x)\| \leq \kappa_g,$$

for all $x$ in an open convex set $X$ containing $\{ x_k \}$, then $C_k \leq C \overset{\text{def}}{=} \max\{ C_0, \alpha(1 - \theta)\kappa_g \}$ and (3.7) holds.

Second, we observe that the update of $\sigma_k$ is not affected by unsuccessful iterations in the sense of Step 4. In fact, we have $\sigma_{k+1} = \sigma_k$ whenever an unsuccessful iteration occurs at Step 4 and the rule for adapting $\sigma_j$, $j \leq k$, has the form

$$\sigma_{j+1} \geq \gamma_1 \sigma_j, \quad j \in S_k ,$$

$$\sigma_{j+1} \geq \gamma_2 \sigma_j, \quad j \in U_{k,1} ,$$

$$\sigma_{j+1} = \sigma_j, \quad j \in U_{k,2} .$$

As a consequence, the upper bound on the scalars $\sigma_k$ established in Lemma 3.2 is still valid.

4. **Complexity analysis.** In this section we study the iteration complexity of Algorithm 3.1 assuming that $f$ is bounded below, i.e., there exists $f_{\text{low}}$, such that

$$f(x) \geq f_{\text{low}}, \forall x \in \mathbb{R}^n .$$

We consider two possible stopping criteria for the approximate minimization of model $m_k$ at Step 3. Given $\theta \in (0, 1)$, the first criterion has the form

$$\|\nabla_m(x_k, s_k, \sigma_k)\| \leq \theta \min \left( \| s_k \|^2, \| \nabla f(x_k) \| \right) ,$$

(4.1)
which amounts to (3.2) with \( \theta_k = \theta \min \left( 1, \frac{\|s_k\|^2}{\|\nabla f(x_k)\|} \right) \). The second criterion is considered in [9, Eqn. (3.28)] and takes the form

\[
\|\nabla_s m(x_k, s_k, \sigma_k)\| \leq \theta \min(1, \|s_k\|) \|\nabla f(x_k)\|. \tag{4.2}
\]

It corresponds to the choice \( \theta_k = \theta \min(1, \|s_k\|) \) in (3.2).

**Lemma 4.1.** Let Assumptions 3.1 and 3.2 hold. Suppose that \( \alpha \in \left[ 0, \frac{2}{3} \right) \) and \( \eta_2 \in \left( 0, \frac{2 - 3\alpha}{2} \right) \) in Algorithm 3.1. Then, at iteration \( k \in S_k \cup U_{k,1} \)

\[
\|s_k\| \geq \sqrt{\zeta \|\nabla f(x_k + s_k)\|},
\]

for some positive \( \zeta \), both when \( s_k \) satisfies (4.1) and when \( s_k \) satisfies (4.2) and the norm of the Hessian is bounded above by a constant \( \kappa_H \) on the path of iterates,

\[
\|\nabla^2 f(x_k + \alpha s_k)\| \leq \kappa_H, \quad \forall k \geq 0, \quad \alpha \in [0, 1].
\]

**Proof.** Taylor expansion of \( f \) gives

\[
f(x_k + s) = f(x_k) + s^T \nabla f(x_k) + \frac{1}{2} s^T \nabla^2 f(x_k) s + \int_0^1 (1 - \tau) s^T (\nabla^2 f(x_k + \tau s) - \nabla^2 f(x_k)) s \, d\tau,
\]

and consequently,

\[
\nabla f(x_k + s) = \nabla f(x_k) + \nabla^2 f(x_k) s + 2 \int_0^1 (1 - \tau) (\nabla^2 f(x_k + \tau s) - \nabla^2 f(x_k)) s \, d\tau. \tag{4.3}
\]

Then, noting that the assumptions of Lemma 3.2 hold at iterations \( k \in S_k \cup U_{k,1} \), using the Lipschitz continuity of \( \nabla^2 f \), (3.7), (3.9) (valid at \( k \in S_k \cup U_{k,1} \)) and (3.13), we derive

\[
\|\nabla f(x_k + s_k) - \nabla_s T_2(x_k, s_k)\| = \|\nabla f(x_k + s_k) - \nabla f(x_k) - B_k s_k\|
\]

\[
\leq \|\Delta_k s_k\| + 2 \int_0^1 (1 - \tau) \| (\nabla^2 f(x_k + \tau s_k) - \nabla^2 f(x_k)) s_k \| \, d\tau
\]

\[
\leq \|\Delta_k\| \|s_k\| + \frac{L}{3} \|s_k\|^2
\]

\[
\leq \left( \max(C, \alpha (\kappa_B + \sigma_{\max})) + \frac{L}{3} \right) \|s_k\|^2. \tag{4.4}
\]

Moreover, by (3.11)

\[
\nabla f(x_k + s_k) = \nabla f(x_k + s_k) - \nabla_s T_2(x_k, s_k) + \nabla_s T_2(x_k, s_k) + \sigma_k \|s_k\| s_k - \sigma_k \|s_k\| s_k
\]

\[
= \nabla f(x_k + s_k) - \nabla_s T_2(x_k, s_k) + \nabla_s m(x_k, s_k, \sigma_k) - \sigma_k \|s_k\| s_k. \tag{4.5}
\]

Now consider the case \( s_k \) satisfying (4.1). Condition (4.1) along with (4.5) and (4.4) yield

\[
\|\nabla f(x_k + s_k)\| \leq \left( \max(C, \alpha (\kappa_B + \sigma_{\max})) + \frac{L}{3} + \theta + \sigma_{\max} \right) \|s_k\|^2,
\]

10
which gives the claim with \( \zeta = 1/(\max (C, \alpha (\kappa B + \sigma_{\max})) + L/3 + \theta + \sigma_{\max}) \)

We turn now the attention to the case \( s_k \) satisfying (4.2). Combining

\[
\nabla f(x_k + s_k) = \nabla f(x_k) + \int_0^1 \nabla^2 f(x_k + ts_k) s_k dt
\]

and the boundness of the Hessian we have

\[
\| \nabla f(x_k) \| \leq \| \nabla f(x_k + s_k) \| + \kappa H \| s_k \|,
\]

for some positive \( L_g \), and by (4.2)

\[
\| \nabla m(x_k, s_k, \sigma_k) \| \leq \theta \min(1, \| s_k \|) \| \nabla f(x_k + s_k) \| + \theta \min(1, \| s_k \|) L_g \| s_k \|
\]

Thus, (4.4) and (4.5) give

\[
(1 - \theta) \| \nabla f(x_k + s_k) \| \leq (\max (C, \alpha (\kappa B + \sigma_{\max})) + L/3 + \theta \kappa H + \sigma_{\max}) \| s_k \|^2,
\]

and the claim follows with \( \zeta = (1 - \theta)/(\max (C, \alpha (\kappa B + \sigma_{\max})) + L/3 + \theta \kappa H + \sigma_{\max}) \).

**Theorem 4.2.** Suppose that \( f \) in (1.1) is lower bounded by \( f_{\text{low}} \) and the assumptions of Lemma 4.1 hold. Then Algorithm 3.1 requires at most

\[
\mathcal{I}_S = \left[ \kappa_s \frac{f(x_0) - f_{\text{low}}}{\varepsilon^{3/2}} \right],
\]

successful iterations and at most

\[
\mathcal{I}_T = \left[ \kappa_s \frac{f(x_0) - f_{\text{low}}}{\varepsilon^{3/2}} \right] \left( 1 + \frac{1}{\log \gamma_2} \log \left( \frac{\sigma_{\max}}{\sigma_0} \right) + |\kappa_u(f(x_0) - f_{\text{low}})| \right),
\]

iterations to produce an iterate \( x_{\tilde{k}} \) satisfying (2.3), with \( \kappa_s = \frac{3}{\eta_1 \sigma_{\min} \zeta^{3/2}} \) and \( \zeta \) as in Lemma 4.1, and \( \kappa_u = \frac{3}{\eta_1 \sigma_{\min}} \).

**Proof.** The mechanism of Algorithm 3.1 for updating \( \sigma_k \) has the form (3.20)–(3.22). An unsuccessful iteration in \( U_k,2 \) does not affect the value of the regularization parameter as \( \sigma_{k+1} = \sigma_k \). Moreover, the assumptions of Lemma 3.2 hold at iterations \( k \in S_k \). Hence, \( \sigma_k \leq \sigma_{\max} \), for all \( k \geq 0 \), due to Lemma 3.2.

The upper bound on the cardinality \( |S_k| \) of \( S_k \) follows from [4, Theorem 2.5]. Then, by using (3.15) and Lemma 4.1, at each successful iteration before termination it holds

\[
f(x_k) - f(x_k + s_k) \geq \eta_1 (T_2(x_k, 0) - T_2(x_k, s_k))
\geq \eta_1 \frac{\sigma_k}{3} \| s_k \|^3
\geq \eta_1 \frac{\sigma_{\min}}{3} \zeta^{3/2} \| \nabla f(x_k + s_k) \|^3/2
\overset{\text{def}}{=} \kappa_{\gamma}^{-1} \| \nabla f(x_k + s_k) \|^{3/2}.
\]

(4.7)
Consequently, before termination (2.3) it holds $f(x_k) - f(x_k + s_k) \geq \kappa^{-1} \epsilon^{3/2}$ which implies

$$f(x_0) - f(x_{k+1}) = \sum_{j \in S_k} (f(x_j) - f(x_j + s_j)) \geq |S_k| \kappa^{-1} \epsilon^{3/2},$$

and (4.6).

The upper bound on $|U_{k,1}|$ follows from [4, Lemma 2.4]. In particular, by (3.20)–(3.22) it holds $\sigma_0 \gamma_1 \frac{|S_k|}{\gamma_2} \leq \sigma_k$ and (3.13) implies

$$|U_{k,1}| \leq |S_k| \left( \frac{\log \gamma_1}{\log \gamma_2} + \frac{1}{\log \gamma_2} \log \left( \frac{\sigma_{\text{max}}}{\sigma_0} \right) \right).$$

As for $|U_{k,2}|$, it is less or equal than the number of successful iterations with $\|s_k\| \geq 1$. By construction, an unsuccessful iteration in $U_{k,2}$ occurs at most once between two successful iterations with the first one such that flag $= 1$, and it can not occur between two successful iterations if flag is null at the first of such iterations. In fact, flag is reassigned only at the end of a successful iteration and can be set to one only in case of successful iteration with $\|s_k\| \geq 1$, see Step 5 of Algorithm 3.1, except for the first iteration. If the case flag $= 1$ and $\|s_k\| < 1$ occurs then flag is set to zero and is not further changed until the subsequent successful iteration.

Noting that, by (4.7),

$$f(x_0) - f(x_{k+1}) = \sum_{j \in S_k} (f(x_j) - f(x_j + s_j)) \geq \eta \sigma_{\text{min}} \sum_{j \in S_k} \|s_k\|^3,$$

we can claim that

$$f(x_0) - f_{\text{low}} \geq \eta \frac{\sigma_{\text{min}}}{3} |U_{k,2}| \overset{\text{def}}{=} \kappa^{-1} u \|s_k\|^3.$$

Then, we obtain $|U_{k,2}| \leq \lfloor \kappa u (f(x_0) - f_{\text{low}}) \rfloor + 1$ (counting the iteration $k = 0$) and the proof is concluded.

The complexity analysis presented above implies

$$\liminf_{k \to \infty} \|\nabla f(x_k)\| = 0.$$
Theorem 4.3. Suppose that \( f \) in (1.1) is lower bounded by \( f_{\text{low}} \), and that the assumptions of Theorem 4.2 hold. Then, the steps \( s_k \) and the iterate \( x_k \) generated by Algorithm 3.1 satisfy

\[
\|s_k\| \to 0, \quad \text{as } k \to \infty, \quad k \in \mathcal{S},
\]

and

\[
\|\nabla f(x_k)\| \to 0, \quad \text{as } k \to \infty.
\]

Moreover, unsuccessful iterations in \( U_k \) do not occur eventually.

Proof. The first claim is proved paralleling [9, Lemma 5.1]. In particular, by (3.14)

\[
f(x_k) - f(x_{k+1}) \geq \eta_1(T_2(x_k,0) - T_2(x_k, s_k)) \geq \eta_1 \frac{\sigma_{\min}}{3} \|s_k\|^3, \quad k \in \mathcal{S}.
\]

Since \( f \) is lower bounded by \( f_{\text{low}} \), one has

\[
f(x_0) - f_{\text{low}} \geq f(x_0) - f(x_{k+1}) = \sum_{j=0}^{k} (f(x_j) - f(x_{j+1})) \geq \eta_1 \frac{\sigma_{\min}}{3} \sum_{j=0, j \in S} \|s_j\|^3, \quad k \geq 0,
\]

which implies convergence of the series \( \sum_{k=0, k \in S} \|s_k\|^3 \) and the first claim as a consequence.

As for \( \|\nabla f(x_k)\| \), Lemma 4.1 provides

\[
\zeta \|\nabla f(x_{k+1})\| \leq \|s_k\|^2 \to 0, \quad \text{as } k \to \infty, \quad k \in \mathcal{S}.
\]

This fact along with \( \nabla f(x_{k+1}) = \nabla f(x_k) \) at unsuccessful iterations provides the convergence of \( \{\|\nabla f(x_k)\|\} \) to zero.

Finally, the behaviour of \( \{\|s_k\|\}_{k \in S} \) implies that eventually all successful iterations are such that \( \|s_k\| < 1 \). Thus, the mechanism of Algorithm 3.1 gives \( \text{flag} = 0 \) for all \( k \) sufficiently large and unsuccessful iterations in the sense of Step 4 cannot occur. \( \Box \)

5. Convergence to second order critical point. In this section we focus on the convergence of the sequence generated by our procedure to second-order critical points \( x^* \):

\[
\nabla f(x^*) = 0 \quad \text{and} \quad \lambda_{\min}(\nabla^2 f(x^*)) \geq 0.
\]

First, we analyze the asymptotic behaviour of \( \{x_k\} \) in the case where the approximate Hessian \( B_k \) becomes positive definite along a converging subsequence of \( \{x_k\} \). In such a context, we show \( q \)-quadratic convergence of \( \{x_k\} \) under an additional mild requirement on the step, namely the Cauchy condition. Second, we consider the case where the model \( B_k \) is not convex and obtain a second order complexity bound in line with the study of Cartis et al. [11].

Theorem 5.1. Suppose that \( f \) in (1.1) is lower bounded by \( f_{\text{low}} \), and that the assumptions of Theorem 4.2 hold. Suppose that \( \{x_k\} \) is a subsequence of successful iterates converging to some \( x^* \) and that \( B_{k_i} \) are positive definite whenever \( x_{k_i} \) is sufficiently close to \( x^* \). Then

i) \( x_k \to x^* \) as \( k \to \infty \) and \( x^* \) is second-order critical.

ii) If \( s_k \) satisfies

\[
m(x_k, s_k, \sigma_k) \leq m(x_k, s_k^C, \sigma_k), \quad \forall k \geq 0,
\]

where \( s_k^C \) is the Cauchy step, i.e.

\[
s_k^C = -\alpha_k^C \nabla f(x_k) \quad \text{and} \quad \alpha_k^C = \arg\min_{\alpha \geq 0} m_k(x_k, -\alpha \nabla f(x_k), \sigma_k),
\]

then all the iterations are eventually successful and \( x_k \to x^* \) \( q \)-quadratically.
Proof. i) From (4.9), (3.9) and (3.13) it follows

$$\|\nabla^2 f(x_k) - B_k\| \leq \alpha (\kappa_B + \sigma_{\text{max}}) \|s_k\| \to 0, \quad k \to \infty, \quad k \in \mathcal{S}. \quad (5.2)$$

As a consequence, standard perturbation results on the eigenvalues of symmetric matrices and the convergence of \(\{x_k\}\) to \(x^*\) give that \(\nabla^2 f(x^*)\) is positive definite. Thus, \(x^*\) is an isolated limit point and the claim i) is completed by using (4.9) and [20, Lemma 4.10].

ii) From the convergence of \(\{x_k\}\) to \(x^*\), (5.2) and the positive definiteness of \(\nabla^2 f(x^*)\) it follows that

$$\lambda_{\text{min}}(B_k) \geq \lambda, \quad \forall k \in \mathcal{S} \text{ sufficiently large.}$$

Moreover, we know that unsuccessful iterations in \(U_2\) do not occur eventually. Then, taking into account that \(B_k\) is not modified along the unsuccessful iterations in \(U_1\), we conclude that

$$\lambda_{\text{min}}(B_k) \geq \lambda, \quad \forall k \text{ sufficiently large.}$$

In order to show that all the iterations are eventually successful, we start using (3.2), (3.11) and obtaining

$$\frac{\|s_k\|}{\|(B_k + \sigma_k s_k \|s_k\| I)^{-1}\|} - \|\nabla f(x_k)\| \leq \theta \|\nabla f(x_k)\|. \quad (5.3)$$

Since

$$\|(B_k + \sigma_k s_k \|s_k\| I)^{-1}\| = \frac{1}{\lambda_{\text{min}}(B_k) + \sigma_k \|s_k\|} \leq \frac{1}{\lambda},$$

we get

$$\|s_k\| \leq \frac{1 + \theta}{\lambda} \|\nabla f(x_k)\|, \quad \forall k \text{ sufficiently large,}$$

and \(\|s_k\| \to 0\) due to (4.10). Moreover, by (3.14), (5.1) and [9, Lemma 2.1]

$$T_2(x_k, 0) - T_2(x_k, s_k) \geq m(x_k, 0, \sigma_k) - m(x_k, s_k, \sigma_k) \geq \frac{\|\nabla f(x_k)\|}{6\sqrt{2}} \min \left( \frac{\|\nabla f(x_k)\|}{1 + \|B_k\|}, \frac{1}{2} \sqrt{\frac{\|\nabla f(x_k)\|}{\sigma_k}} \right), \quad (5.4)$$

and Assumption 3.2 and Lemma 3.2 yield

$$T_2(x_k, 0) - T_2(x_k, s_k) \geq \frac{\|\nabla f(x_k)\|}{6\sqrt{2}} \min \left( \frac{\|\nabla f(x_k)\|}{1 + \kappa_B}, \frac{1}{2} \sqrt{\frac{\|\nabla f(x_k)\|}{\sigma_{\text{max}}}} \right).$$

Thus, eventually (4.10) and (5.3) give

$$T_2(x_k, 0) - T_2(x_k, s_k) \geq \frac{\|\nabla f(x_k)\|^2}{6\sqrt{2}(1 + \kappa_B)} \geq \frac{\lambda^2}{6\sqrt{2}(1 + \kappa_B)(1 + \theta)^2} \|s_k\|^2 \overset{\text{def}}{=} \kappa_c \|s_k\|^2.$$
and by (3.10) and (3.4)

\[ 1 - \rho_k = \frac{f(x_k + s_k) - T_2(x_k, s_k)}{T_2(x_k, 0) - T_2(x_k, s_k)} \leq \frac{E_k(s_k)}{\kappa_c \|s_k\|^2} < \frac{(\alpha (\kappa B + \sigma_{\max}) + L/3) \|s_k\|^3}{2\kappa_c \|s_k\|^2}, \]

i.e., \( \rho_k \to 1 \) and the iterations are very successful eventually.

Finally, (5.3) and Lemma 4.1 provide

\[ \|\nabla f(x_{k+1})\| \leq \frac{\|s_k\|^2}{\zeta} \leq \frac{(1 + \theta)^2}{\zeta 2} \|\nabla f(x_k)\|^2, \quad \forall k \text{ sufficiently large}, \]

and the \( q \)-quadratic convergence of the sequence \( \{x_k\} \) follows in a standard way by means of the Taylor’s expansion. \( \Box \)

Dropping the assumption that \( B_k \) is positive definite, convergence to second order critical points can be studied. Following [11] where a modification of the ARC algorithm in [9] is proposed, we equip Algorithm 3.1 with a further stopping criterion and impose an additional condition on the step. First, Algorithm 3.1 is stopped when

\[ \|\nabla f(x_k)\| \leq \epsilon \quad \text{and} \quad \lambda_{\min}(B_k) \geq -\epsilon_H, \quad \epsilon, \epsilon_H > 0, \quad (5.5) \]

which represents the approximate counterpart of the second-order optimality conditions with the Hessian matrix approximated by \( B_k \). The above criterion does not imply, in general, vicinity to local minima, as well as it does not guarantee the iterates to be distant from saddle points. Then, the possibility of referring to the strict-saddle property [19] may play a significant role; indeed, under its validity, (5.5) implies closeness to a local minimum for sufficiently small values of the tolerances \( \epsilon \) and \( \epsilon_H \).

Second, the trial step \( s_k \) computed in Step 2.2 of Algorithm 3.1 is required to satisfy the following additional condition: if \( B_k \) is not positive semidefinite, then

\[ m(x_k, s_k, \sigma_k) \leq m(x_k, s_k^E, \sigma_k), \quad (5.6) \]

where \( s_k^E \) is defined as

\[ s_k^E = \alpha_k^E u_k \quad \text{and} \quad \alpha_k^E = \arg \min_{\alpha \geq 0} m_k(\alpha u_k), \quad (5.7) \]

and \( u_k \) is an approximation of the eigenvector of \( B_k \) associated with its smallest eigenvalue \( \lambda_{\min}(B_k) \), in the sense that

\[ \nabla f(x_k)^T u_k \leq 0 \quad \text{and} \quad u_k^T B_k u_k \leq \kappa_{\text{snc}} \lambda_{\min}(B_k) \|u_k\|^2, \quad (5.8) \]

for some constant \( \kappa_{\text{snc}} \in (0, 1] \). Note that the minimization in (5.7) is global which implies

\[ \nabla f(x_k)^T s_k^E + (s_k^E)^T B_k s_k^E + \sigma_k \|s_k^E\|^3 = 0, \quad (5.9) \]

\[ (s_k^E)^T B_k s_k^E + \sigma_k \|s_k^E\|^3 \geq 0. \quad (5.10) \]

We refer to the resulting algorithm as ARC Second Order critical point (ARC\_SO). The termination criterion adopted here does not affect the mechanism for updating \( \sigma_k \), then the upper bound \( \sigma_{\max} \) on \( \sigma_k \) given in Lemma 3.2 is still valid.
Let $\tilde{S}_k$ denote the set of indexes of successful iterations of ARC.SO whenever $\|\nabla f(x_k)\| > \epsilon$ and/or $\lambda_{\min}(B_k) < -\epsilon_H$, i.e., the indexes of successful iterations before (5.5) is met. Following [11] we also let $\tilde{S}_k^{(1)}$ be the set of indices of successful iterations where $\|\nabla f(x_k)\| > \epsilon$ and $\tilde{S}_k^{(2)}$ be the set of indexes of successful iterations where $\lambda_{\min}(B_k) < -\epsilon_H$. Let $U_{k,1}$ and $U_{k,2}$ denote the set of unsuccessful iterations of ARC.SO analogously to (3.18) and (3.19). Remarkably, the cardinality of both $\tilde{S}_k^{(1)}$ and $U_{k,2}$ is the same as in Algorithm 3.1, see Theorem 4.2, while proceeding as in Theorem 4.2 the cardinality of $U_{k,1}$ is bounded in terms of the number of successful iterations $\tilde{S}_k$, see also [11, Lemma 2.6]. Hence, it remains to derive the cardinality of $\tilde{S}_k^{(2)}$.

**Lemma 5.2.** Suppose that $f$ in (1.1) is lower bounded by $f_{\text{low}}$ and the assumptions of Theorem 4.2 hold. Suppose that $s_k$ satisfies (5.6). Then, the number of successful iterations of Algorithm ARC.SO with $\lambda_{\min}(B_k) < -\epsilon_H$ is bounded above by

$$\left\lceil \frac{\kappa_e (f(x_0) - f_{\text{low}})}{\epsilon_H^3} \right\rceil,$$

where $\kappa_e = \frac{3\sigma_{\text{max}}^2 \sigma_{\text{min}}^2}{\eta_1 \kappa_{\text{nc}}^3}$.

**Proof.** The proof parallels that of [11, Lemma 2.8]. We have

$$f(x_k) - f(x_k + s_k) \geq \eta_1 (T_2(x_k, 0) - T_2(x_k, s_k))$$

$$= \eta_1 (m(x_k, 0, \sigma_k) - m(x_k, s_k, \sigma_k) + \frac{\sigma_k}{3} \|s_k\|^3)$$

$$\geq \eta_1 (m(x_k, 0, \sigma_k) - m(x_k, s_k^E, \sigma_k))$$

$$\geq \eta_1 \sigma_k \|s_k^E\|^3$$

$$\geq \eta_1 \eta_1 \lambda_{\min}(B_k)^3 \kappa_{\text{nc}}^3$$

$$= \eta_1 \frac{3\sigma_{\text{max}}^2 \sigma_{\text{min}}^2}{6\sigma_{\text{max}}^2}$$

in which we have used (3.14), (5.6), (5.9), (5.10) and (5.8). As a consequence, letting $\kappa_e$ as in the statement of the theorem, before termination it holds

$$f(x_0) - f_{\text{low}} \geq f(x_0) - f(x_{k+1}) \geq \sum_{j \in \tilde{S}_k^{(2)}} (f(x_j) - f(x_j + s_j)) \geq |\tilde{S}_k^{(2)}| \kappa_e^{-1} \epsilon_H^3,$$

and the claim follows. \qed

We thus conclude that Algorithm ARC.SO produces an iterate $x_\tilde{k}$ satisfying (5.5) within at most

$$O\left(\max(\epsilon^{-3/2}, \epsilon_H^{-3})\right),$$

iterations, in accordance with the complexity result in [11].

**6. Finite sum minimization.** Large-scale instances of the finite-sum problem (1.2) can be conveniently solved by subsampled procedures where $\nabla f^2(x_k)$ is approximated by randomly sampling component functions $\phi_i$ [5]. The resulting approximation of $\nabla f^2(x_k)$ takes the form

$$\nabla^2 f_{D_k}(x_k) = \frac{1}{|D_k|} \sum_{i \in D_k} \nabla^2 \phi_i(x_k),$$

(6.1)
with $\mathcal{D}_k \subset \{1, 2, \ldots, N\}$ and $|\mathcal{D}_k|$ being the so-called sample size.

We discuss the application of Algorithm 3.1 to problem (1.2) with

$$B_k = \nabla^2 f_{\mathcal{D}_k}(x_k),$$

(6.2)
giving both deterministic and probabilistic results. The application of Algorithm 3.1 to problem (1.2) with such Hessian approximation is supported by results in the literature which give the sample size required for $B_k$ to satisfy condition (3.6) in probability and will be addressed below.

Let us make the following assumption on the objective function.

**Assumption 6.1.** For each $i \in \{1, 2, \ldots, N\}$ there exists $\kappa_{\phi,i} \geq 0$, such that

$$\sup_{x \in \mathbb{R}^n} \| \nabla^2 \phi_i(x) \| \leq \kappa_{\phi,i}.$$

Denoting $\kappa_\phi = \max_{i \in \{1, 2, \ldots, N\}} \kappa_{\phi,i}$, trivially $\sup_{x \in \mathbb{R}^n} \| \nabla^2 f(x) \| \leq \kappa_\phi$, and Assumption 3.2 is satisfied with $\kappa_B = \kappa_\phi$.

Uniform and non-uniform sampling strategies have been proposed [5, 14, 18, 24, 25]; for instance, the following Lemma provides the size of uniform sampling which probabilistically satisfies (3.6).

**Lemma 6.1.** Assume that Assumption 6.1 holds, $C_k > 0$ is given, the subsample $\mathcal{D}_k$ is chosen randomly and uniformly from $\{1, 2, \ldots, N\}$ and $B_k$ is as in (6.2). Then, given $\bar{\delta} \in (0, 1)$,

$$\Pr(\| \nabla^2 f(x_k) - B_k \| \leq C_k) \geq 1 - \bar{\delta},$$

(6.3)

whenever the cardinality $|\mathcal{D}_k|$ of $\mathcal{D}_k$ satisfies

$$|\mathcal{D}_k| \geq \frac{16 \kappa_\phi^2}{C_k^2} \ln \left( \frac{2n}{\bar{\delta}} \right),$$

(6.4)

**Proof.** See [24, Lemma 4]. □

We first give deterministic results, namely properties which are valid independently from Assumption 3.4 on $B_k$, now guaranteed with probability $1 - \bar{\delta}$ by Lemma 6.1. In the following theorem the only requirement on $B_k$ is the boundness of its norm, i.e. Assumption 3.2; concerning the trial step $s_k$, the Cauchy condition (5.1) is assumed.$^1$

**Theorem 6.2.** Let $f \in C^2(\mathbb{R}^n)$. Suppose that $f$ in (1.1) is lower bounded by $f_{\text{low}}$, Assumption 6.1 and condition (5.1) hold. Then,

i) Given $\epsilon > 0$, Algorithm 3.1 takes at most $O(\epsilon^{-2})$ successful iterations to satisfy $\| \nabla f(x_k) \| < \epsilon$.

ii) $\| \nabla f(x_k) \| \to 0$, as $k \to \infty$ and therefore all the accumulation points of the sequence $\{x_k\}$, if any, are first-order stationary points.

iii) If $\{x_k\}$ is a subsequence of iterates converging to some $x^*$ such that $\nabla f^2(x^*)$ is definite positive, then $x_k \to x^*$ as $k \to \infty$.

---

$^1$This result is valid independently from the specific form of $f$ considered in this section, provided that the norm of the Hessian of $f$ is bounded in an open convex set containing all the sequence $\{x_k\}$ and Assumptions 3.2 holds.
Proof. i) The claim follows from Lemma 3.1–3.3 and Corollary 3.4 in [10]. In fact, despite the acceptance criterion in [10] is (2.4) instead of (2.9), we can rely on the proof of [10, Lemma 3.2] thanks to (5.4) and considering that
\[ f(x_k + s_k) - T_2(x_k, s_k) \leq (\kappa_o + \kappa_B) ||s_k||^2, \quad k \geq 0. \]

ii) The sub-optimal complexity result in Item i) guarantees that \( \liminf_{k \to \infty} ||\nabla f(x_k)|| = 0 \) and that the number of successful iterations is finite. Moreover, \( \lim_{k \to \infty} ||\nabla f(x_k)|| = 0 \) follows by Assumption 6.1 and [9, Corollary 2.6].

iii) Proceeding as in Theorem 4.3 we obtain (4.9). Since \( \nabla^2 f(x^*) \) in positive definite, \( x^* \) is an isolated limit point; consequently, (4.9) and Lemma [20, Lemma 4.10] yield the claim. \( \square \)

Focusing on the optimal complexity result, we observe that Algorithm 3.1 requires at most \( O(\epsilon^{-3/2}) \) iterations to satisfy \( ||\nabla f(x_k)|| \leq \epsilon \) with probability \( 1 - \delta, \delta \in (0, 1) \), provided that the sample size is chosen accordingly to (6.4) and \( \delta \) is suitable chosen. In fact, let \( E_i \) be the event: “the relation \( ||\nabla^2 f(x_i) - B_i|| \leq C_i \) holds at iteration \( i, 1 \leq i \leq k' \), and \( E(k) \) be the event: “the relation \( ||\nabla^2 f(x_i) - B_i|| \leq C_i \) holds for the entire \( k \) iterations”. If the events \( E_i \) are independent, then due to (6.3)

\[ Pr(E(k)) \equiv Pr \left( \bigcap_{i=1}^{k} E_i \right) = (1 - \bar{\delta})^k. \]

Thus, requiring that the event \( E(k) \) occurs with probability \( 1 - \delta \), we obtain

\[ Pr(E(k)) = (1 - \bar{\delta})^k = 1 - \delta, \quad \text{i.e.,} \quad \bar{\delta} = 1 - \sqrt[3]{1 - \delta} = O \left( \frac{\delta}{k} \right). \]

Taking into account the iteration complexity, \( k = O(\epsilon^{-3/2}) \) we deduce the following choice of \( \bar{\delta} \):

\[ \bar{\delta} = O(\delta^{3/2}). \tag{6.5} \]

Summarizing, choosing, at each iteration, \( \bar{\delta} \) according to (6.5) and the sample size according to (6.4), the complexity result in Theorem 4.2 holds with probability \( 1 - \delta \). We underline that the resulting per-iteration failure probability \( \delta \) is not too demanding in what concerns the sample size, because it influences only the logarithmic factor in (6.4), see [24].

Finally, we focus on the subclass of problems where the function \( \phi_i \) are strongly convex and denote with \( x^* \) the unique minimizer of problem (1.2). Problems of this type arise, for instance, in classification procedures. In this case, trivially \( B_k \) is positive definite. From Theorem 5.1, Item ii), we can conclude that, for \( k \) sufficiently large, say \( k \geq \bar{k} \), with probability \( (1 - \bar{\delta})^{\bar{k}} \) there exists \( M > 0 \) such that

\[ ||x_{k+1} - x^*|| \leq M||x_k - x^*||^2, \quad k = \bar{k}, \ldots, \bar{k} + k_o - 1. \]

Specifically, proceeding as in [23, Theorem 2] and denoting with \( E_i \) the event: “the relation \( ||\nabla^2 f(x_i) - B_i|| \leq C_i \) holds at iteration \( i, i \geq \bar{k} \)”, we have that the overall success probability in consecutive \( k_o \) iterations is

\[ Pr \left( \bigcap_{i=\bar{k}}^{\bar{k}+k_o-1} E_i \right) = (1 - \bar{\delta})^{k_o}, \]

which concludes our argument.
7. Related work. Variants of ARC based on suitable approximations of the gradient and/or the Hessian of $f$ have been discussed in a few recent lines of work reviewed in this section. Besides the algorithm in [9, 10, 11], which employs approximations for the Hessian and is suited for a generic nonconvex function $f$, works [14, 18, 24, 26] propose variants of the algorithm given in [9] where the gradient and/or the Hessian approximations can be performed via sub-sampling techniques [2, 5] and are applicable to the relevant class of large-scale finite-sum minimization (1.2) arising in machine learning; probabilistic complexity and convergence analysis is carried out.

Cartis et al. [9, 10, 11] analyze ARC framework under varying assumptions on the Hessian approximation $B_k$ and establish optimal and sub-optimal worst-case iteration bounds for first- and second-order optimality. First-order complexity was shown to be of $O(\epsilon^{-2})$ iterations under Assumption 3.2 and, as mentioned in Section 2, of $O(\epsilon^{-3/2})$ iterations when, in addition, $B_k$ resembles the true Hessian and condition (2.5) is satisfied.

Kohler et al. [18] propose and study a variant of ARC algorithm suited for finite-sum minimization not necessarily convex. A sub-sampling scheme for the gradient and the Hessian of $f$ is applied while maintaining first-order complexity of $O(\epsilon^{-3/2})$ iterations. The sampling scheme provided guarantees that the sub-sampled gradient $g(x_k)$ satisfies
\[ \| \nabla f(x_k) - g(x_k) \| \leq M \| s_k \|^2, \quad \forall k \geq 0, \quad M > 0, \]
with prefixed probability, and the sub-sampled Hessian $B_k$ satisfies the strict condition (2.5) with prefixed probability.

Xu et al. [24] develop and study a version of ARC algorithm where a major relaxation on the level of resemblance between $\nabla^2 f(x_k)$ and $B_k$ is made over (2.5). Matrix $B_k$ is supposed to satisfy Assumption 3.2 and
\[ \| (\nabla^2 f(x_k) - B_k) s_k \| \leq \mu \| s_k \|, \quad \mu \in (0, 1), \]
and the latter condition can be enforced building $B_k$ such that $\| \nabla^2 f(x_k) - B_k \| \leq \mu$. Non convex finite-sum minimization is the motivating application for the proposal, and uniform and non-uniform sampling strategies are provided to construct matrices $B_k$ satisfying $\| \nabla^2 f(x_k) - B_k \| \leq \mu$ with prefixed probability. In particular, unlike the rule in [18], the rule for choosing the sample size at iteration $k$ does not depend on the step $s_k$ which is not available when $B_k$ is built. Worst-case iteration count of order $\epsilon^{-3/2}$ is shown when $\mu = O(\epsilon)$ while suboptimal worst-case iteration count of order $\epsilon^{-2}$ is achieved if $\mu = O(\sqrt{\epsilon})$. Note that the accuracy requirement on $B_k$ is fixed along the iterations and depends on the accuracy requirement on the gradient’s norm, that is on the gradient’s norm at the final iteration. Additionally, the use of approximate gradient via subsampling is addressed in [26].

Chen et al. [14] propose an ARC procedure for convex optimization via random sampling. Function $f$ is convex and defined as finite-sum (1.2) of possibly nonconvex functions. Semidefinite positive sub-sampled approximations $B_k$ satisfying $\| \nabla^2 f(x_k) - B_k \| \leq \mu_k$, $\mu_k \in (0, 1)$, are built with a prefixed probability. Iteration complexity of order $O(\epsilon^{-1/2})$ is proved with respect to the fulfillment of condition $f(x_k) - f(x^*) \leq \epsilon$, $x^*$ being the global minimum of (1.2); the scalar $\mu_k$ is updated as $\mu_{k+1} = O(\min(\mu_k, \| \nabla f(x_k) \|))$, and the model $m(x_k, s, \sigma_k)$ is minimized on a subspace of $\mathbb{R}^n$ imposing the strict condition $\| \nabla m(x_k, s_k, \sigma_k) \| \leq \theta \min(\| \nabla f(x_k) \|, \| \nabla^2 f(x_k) \|^3, \| s_k \|^2), \theta \in (0, 1)$.

Summarizing, our proposal differs from the above works in the following respects. Our scheme is based on the relaxed Assumption 3.4 for $B_k$ over (2.5), unlike [18]; it improves upon [24, 26]
in the construction of $B_k$ as the level of resemblance between $\nabla^2 f(x_k)$ and $B_k$ is not maintained fixed along iterations but adaptively chosen; it improves upon [14] as the prescribed accuracy on $B_k$ (and the sample size) may reduce at some iteration, the ultimate accuracy on $\|\nabla m(x_k, s_k, \sigma_k)\|$ is milder, and our complexity results are optimal for nonconvex problems while the analysis in [14] is limited to convex problems.

8. Conclusions and perspectives. We proposed an ARC algorithm for solving nonconvex optimization problems based on a dynamic rule for building inexact Hessian information. The new algorithm maintains the distinguishing features of ARC framework, i.e., the optimal worst-case iteration bound for first- and second-order critical point. Application to large-scale finite-sum minimization is sketched and analyzed. Our next goal is to implement and test the new algorithm on such class of problems, possibly including inexact gradient information.

REFERENCES


