Optimization Methods & Software
iFirst, 2011, 1–23

Taylor & Francis
Taylor & Francis Group

Evaluation complexity of adaptive cubic regularization methods for convex unconstrained optimization

Coralia Cartis[a]*, Nicholas I.M. Gould[b] and Philippe L. Toint[c]

[a]*School of Mathematics, University of Edinburgh, The King's Buildings, Edinburgh EH9 3JZ, UK;* [b]*Computational Science and Engineering Department, Rutherford Appleton Laboratory, Chilton, Oxfordshire OX11 0QX, UK;* [c]*Department of Mathematics, FUNDP – University of Namur, 61, rue de Bruxelles, B-5000 Namur, Belgium*

(*Received 6 November 2010; final version received 27 June 2011*)

The adaptive cubic regularization algorithms described in Cartis, Gould and Toint [Adaptive cubic regularisation methods for unconstrained optimization *Part II: Worst-case function- and derivative-evaluation complexity*, Math. Program. (2010), doi:10.1007/s10107-009-0337-y (online)]; [*Part I: Motivation, convergence and numerical results*, Math. Program. 127(2) (2011), pp. 245–295] for unconstrained (nonconvex) optimization are shown to have improved worst-case efficiency in terms of the function- and gradient-evaluation count when applied to convex and strongly convex objectives. In particular, our complexity upper bounds match in order (as a function of the accuracy of approximation), and sometimes even improve, those obtained by Nesterov [*Introductory Lectures on Convex Optimization*, Kluwer Academic Publishers, Dordrecht, 2004; *Accelerating the cubic regularization of Newton's method on convex problems*, Math. Program. 112(1) (2008), pp. 159–181] and Nesterov and Polyak [*Cubic regularization of Newton's method and its global performance*, Math. Program. 108(1) (2006), pp. 177–205] for these same problem classes, without requiring exact Hessians or exact or global solution of the subproblem. An additional outcome of our approximate approach is that our complexity results can naturally capture the advantages of both first- and second-order methods.

Keywords: Newton's method, cubic regularization, nonlinear optimization

1. Introduction

State-of-the-art methods for unconstrained smooth optimization typically depend on trust-region [6] or line-search [7] techniques to globalize Newton-like iterations. Of late, a third alternative, in which a local cubic overestimator of the objective is used as the basis for a regularization strategy for the step computation, has been proposed [8,11,12] (see [5, §1] for a detailed description of these contributions). Such ideas have been refined so that they are now well suited to large-scale computation for a wide class of nonlinear nonconvex objectives; rigorous convergence and complexity analyses under weak assumptions, together with promising numerical experience with these techniques, are available [2,5]. Our objective in this paper is to show that the complexity

*Corresponding author. Email: coralia.cartis@ed.ac.uk

ISSN 1055-6788 print/ISSN 1029-4937 online

http://dx.doi.org/10.1080/10556788.2011.602076
http://www.tandfonline.com

bounds for this type of algorithms significantly improve in the presence of convexity or strong convexity.

Specifically, at each iteration of what we call an Adaptive Regularization with Cubics (ARC) framework, a possibly nonconvex model

$$m_k(s) \overset{\text{def}}{=} f(x_k) + s^{\mathrm{T}} g_k + \tfrac{1}{2} s^{\mathrm{T}} B_k s + \tfrac{1}{3}\sigma_k \|s\|^3 \tag{1}$$

is employed as an approximation to the smooth objective $f(x_k + s)$ we wish to minimize. Here, $\sigma_k > 0$ is a regularization weight, we have written $\nabla f(x_k) = g(x_k) = g_k$ and here and hereafter we choose the Euclidean norm $\|\cdot\| = \|\cdot\|_2$. To compute the change s_k to x_k, the model m_k is globally minimized, either exactly or approximately, with respect to $s \in \mathbb{R}^n$. Note that if B_k is taken to be the Hessian $H(x)$ of f, and the latter is globally Lipschitz continuous with Lipschitz constant $\sigma_k/2$, we have the overestimation property $f(x_k + s) \le m_k(s)$ for all $s \in \mathbb{R}^n$ [5, §1]. Thus in this case, minimizing m_k with respect to s forces a decrease in f from the value $f(x_k)$, since $f(x_k) = m_k(0)$. In the general ARC algorithmic framework, H need not be Lipschitz, nor need B_k be $H(x_k)$, but in this case, σ_k must be adjusted as the computation proceeds to ensure convergence [2,5, 2.1]. The generic ARC framework [2,5, §2.1] may be summarized as follows.

ALGORITHM 1.1 (ARC [2,5]) *Given* x_0, $\gamma_2 \ge \gamma_1 > 1$, $1 > \eta_2 \ge \eta_1 > 0$, *and* $\sigma_0 > 0$, *for* $k = 0, 1, \ldots$ *until convergence*,

1. *compute a step* s_k *for which*

$$m_k(s_k) \le m_k(s_k^{\mathrm{C}}), \tag{2}$$

where the Cauchy point

$$s_k^{\mathrm{C}} = -\alpha_k^{\mathrm{C}} g_k \quad \textit{and} \quad \alpha_k^{\mathrm{C}} = \arg\min_{\alpha \in \mathbb{R}_+} m_k(-\alpha g_k); \tag{3}$$

2. *compute* $f(x_k + s_k)$ *and*

$$\rho_k = \frac{f(x_k) - f(x_k + s_k)}{f(x_k) - m_k(s_k)}; \tag{4}$$

3. *set*

$$x_{k+1} = \begin{cases} x_k + s_k & \textit{if } \rho_k \ge \eta_1, \\ x_k & \textit{otherwise}; \end{cases} \tag{5}$$

4. *set*

$$\sigma_{k+1} \in \begin{cases} (0, \sigma_k] & \textit{if } \rho_k > \eta_2 \ [\textit{very successful iteration}], \\ [\sigma_k, \gamma_1\sigma_k] & \textit{if } \eta_1 \le \rho_k \le \eta_2 \ [\textit{successful iteration}], \\ [\gamma_1\sigma_k, \gamma_2\sigma_k] & \textit{otherwise} \ [\textit{unsuccessful iteration}]. \end{cases} \tag{6}$$

For a detailed description of the algorithm construction, including a justification that (2)–(4) are well defined until termination, see [5]. The above ARC algorithm is a very general first-order framework that due to the Cauchy condition (2) ensures at least a steepest-descent-like decrease in each (successful) iteration. This is sufficient to ensure global convergence of ARC both to first-order critical points [5, §2.1] and with steepest-descent-like function-evaluation complexity bounds of order ϵ^{-2} [2, §3] to guarantee

$$\|g_k\| \le \epsilon. \tag{7}$$

These results require that $g(x)$ is uniformly and Lipschitz continuous (respectively) and that $\{B_k\}$ is uniformly bounded above. Clearly, the Cauchy point s_k^{C} achieves (2) in a computationally inexpensive way [5, §2.1]; the choice of interest, however, is when s_k is an (approximate global) minimizer of $m_k(s)$ and B_k, a nontrivial approximation to the Hessian $H(x_k)$ (see Section 3).

Although m_k might be nonconvex, its global minimizer over $\mathbb{R}^n$ is always well defined and can be characterized in a computationally viable way [5, Theorem 3.1; 8,11]. This characterization is best suited for exact computation when B_k is sparse or of modest size. For large problems, a suitable alternative is to improve upon the Cauchy point by globally minimizing m_k over (nested and increasing) subspaces that include g_k – which ensures (2) remains satisfied – until a suitable termination condition is achieved. (For instance, in our ARC implementation [5], the successive subspaces that the model is minimized over are generated using Lanczos method.) These ARC variants are summarized in Algorithm 1.2, where $h_k(\|s_k\|, \|g_k\|)$ is some generic function of $\|s_k\|$ and $\|g_k\|$, with specific examples of suitable choices given in (10) and (11) below.

ALGORITHM 1.2 (ARC$_{(\text{h})}$ [2,5]) *In each iteration k of Algorithm* 1.1, *compute s_k in Step* 1 *as the global minimizer of*

$$\min_{s\in\mathbb{R}^n} \quad m_k(s) \text{ subject to } s \in \mathcal{L}_k, \tag{8}$$

where $\mathcal{L}_k$ is a subspace of $\mathbb{R}^n$ containing g_k, and such that the termination condition

$$\textbf{TC.h} \quad \|\nabla_s m_k(s_k)\| \leq \theta_k \|g_k\|, \quad \text{where } \theta_k \overset{\text{def}}{=} \kappa_\theta \min(1, h_k) \quad \text{and} \quad h_k \overset{\text{def}}{=} h_k(\|s_k\|, \|g_k\|) > 0, \tag{9}$$

is satisfied, for some constant $\kappa_\theta \in (0,1)$ chosen at the start of the algorithm.

Clearly, TC.h is satisfied when s_k is the global minimizer of m_k over the whole space, but one hopes that termination of the subspace minimization will occur well before this inevitable outcome, at least in early stages of the iteration. Note that, in fact, TC.h only requires an approximate critical point of the model, and as such the global subspace minimization in (8) may only need to hold along the one-dimensional subspace determined by s_k [5, (3.11), (3.12)], provided (2) holds.

For ARC$_{(\text{h})}$ to be a 'proper' second-order method, a careful choice of h_k needs to be made, such as $h_k = \|s_k\|$ or $h_k = \|g_k\|^2$, yielding the termination criteria

$$\textbf{TC.s} \quad \|\nabla_s m_k(s_k)\| \leq \theta_k \|g_k\|, \quad \text{where } \theta_k = \kappa_\theta \min(1, \|s_k\|) \tag{10}$$

and

$$\textbf{TC.g2} \quad \|\nabla_s m_k(s_k)\| \leq \theta_k \|g_k\|, \quad \text{where } \theta_k = \kappa_\theta \min(1, \|g_k\|^2). \tag{11}$$

Forthwith, we refer to ARC$_{(\text{h})}$ with TC.s and with TC.g2 as ARC$_{(\text{S})}$ and ARC$_{(\text{g2})}$, respectively. The benefit of requiring the more stringent conditions (8), and (10) or (11), in the above ARC variants is that ARC$_{(\text{S})}$ and ARC$_{(\text{g2})}$ are also guaranteed to converge locally Q-quadratically and globally to second-order critical points [5, §4.2, §5] and to have improved function-evaluation complexity of order $\epsilon^{-3/2}$ to ensure (7) [2, §5], provided $H(x)$ is globally Lipschitz continuous along the path of the iterates and there is sufficiently good agreement between the $H(x_k)$ and its approximation B_k.

In this paper, we investigate the worst-case function-evaluation complexity of the basic ARC framework and its second-order variants ARC$_{(\text{S})}$ and/or ARC$_{(\text{g2})}$ when applied to the minimization of special classes of objectives, namely convex and strongly convex ones. In particular, we show that as expected, these algorithms satisfy improved bounds compared with the nonconvex case. Specifically, generic ARC (Algorithm 1.1) takes at most $\mathcal{O}(\epsilon^{-1})$ and $\mathcal{O}(\log \epsilon^{-1})$ function evaluations to reach the neighbourhood

$$f(x_k) - f_* \leq \epsilon \tag{12}$$

of the (global) minimum f_* of convex and strongly convex objectives, respectively, with Lipschitz continuous gradients, where the dependence of these bounds on problem conditioning is carefully considered (see Section 2.4.1). Unsurprisingly, due to the simple Cauchy decrease condition (2) required on the step, these bounds match in order those for standard steepest descent methods on the same classes of objectives [9].

When applied to convex objectives with bounded level sets and globally Lipschitz continuous Hessian, ARC$_{(g2)}$ with $B_k = H(x_k)$ will reach approximate optimality in the (12) sense in at most $\mathcal{O}(\epsilon^{-1/2})$ function evaluations; this matches in order the bound obtained in [10,11] for cubic regularization on the same problem class when the exact subproblem solution is computed in each iteration. Note that asymptotically, in ARC$_{(g2)}$, the subproblem is solved to higher accuracy than in ARC$_{(S)}$, which seems to be crucial when deriving the improved bound compared with the first-order basic ARC. We also present an illustration on a common convex objective that indicates that despite being worst case, the bounds presented here may be tight.

If the objective is strongly convex, then ARC$_{(S)}$ and ARC$_{(g2)}$ (with approximate Hessians as B_k) require at most $\mathcal{O}(|\log\kappa| + |\log\log\epsilon|)$ function evaluations to satisfy (12), where κ is a problem-dependent constant and where the double logarithm term expresses the local Q-quadratic rate of convergence of these variants. The strongly convex-case bound improves that obtained in [10,11] for cubic regularization with exact subproblem solution in that the former has a logarithmic dependence on κ while the latter only includes a polynomial dependence on problem condition numbers. Our result is a direct consequence of using increasing accuracy in the subproblem solution with first-order-like behaviour, and hence complexity early on, and second-order characteristics asymptotically.

Note that the assumption labelling used throughout the paper was chosen to maintain consistency with notation introduced in [2,5]. The structure of the paper is as follows. Section 2 analyses the complexity of basic ARC, while Section 3 that of the second-order variants ARC$_{(S)}$ and ARC$_{(g2)}$, in the convex and strongly convex cases. Section 3.3 presents a convex example of inefficient ARC behaviour with $\mathcal{O}(\epsilon^{-1/2})$ complexity, and Section 4 draws some conclusions and open questions.

2. The complexity of the basic ARC framework

This section addresses the basic ARC algorithm, Algorithm 1.1. We assume that

$$\textbf{AF.1} \qquad f \in C^1(\mathbb{R}^n), \tag{13}$$

and that the gradient g is Lipschitz continuous on an open convex set X containing all the iterates $\{x_k\}$,

$$\textbf{AF.4} \qquad \|g(x) - g(y)\| \le \kappa_{\rm H}\|x-y\|, \quad \text{for all } x, y \in X \text{ and some } \kappa_{\rm H} \ge 1. \tag{14}$$

If $f \in \mathcal{C}^2(\mathbb{R}^n)$, then AF.4 is satisfied if the Hessian $H(x)$ is bounded above on X. Note, however, that for now, we only assume AF.1. In particular, no Lipschitz continuity of $H(x)$ will be required in this section.

The model m_k is assumed to achieve

$$\textbf{AM.1} \qquad \|B_k\| \le \kappa_{\rm B}, \quad \text{for all } k \ge 0 \text{ and some } \kappa_{\rm B} \ge 1. \tag{15}$$

In the case when $f \in \mathcal{C}^2(\mathbb{R}^n)$ and $B_k = H(x_k)$ for all k, then AF.4 implies AM.1 with $\kappa_{\rm B} = \kappa_{\rm H}$.

Naturally, we assume f is bounded below, letting $f_* > -\infty$ be the (global) minimum of f and

$$\Delta_k \overset{\rm def}{=} f(x_k) - f_*, \quad \text{for all } k \ge 0. \tag{16}$$

2.1 *Relating successful and total iteration counts*

Note that the total number of ARC iterations is the same as the number of function evaluations (as we also need to evaluate f on unsuccessful iterations in order to be able to compute ρ_k in (4)), while the number of successful ARC iterations is the same as that of gradient evaluations.

Let us introduce some useful notation. Throughout, denote the index set

$$\mathcal{S} \overset{\text{def}}{=} \{k \geq 0 : k \text{ successful or very successful in the sense of (6)}\}, \tag{17}$$

and, given any $j \geq 0$, let

$$\mathcal{S}_j \overset{\text{def}}{=} \{k \leq j : k \in \mathcal{S}\}, \tag{18}$$

with $|\mathcal{S}_j|$ denoting the cardinality of the latter.

Concerning σ_k, we may require that on each very successful iteration $k \in \mathcal{S}_j$, σ_{k+1} is chosen such that

$$\sigma_{k+1} \geq \gamma_3 \sigma_k, \quad \text{for some } \gamma_3 \in (0, 1]. \tag{19}$$

Note that (19) allows $\{\sigma_k\}$ to converge to zero on very successful iterations (but no faster than $\{\gamma_3^k\}$). A stronger condition on σ_k is

$$\sigma_k \geq \sigma_{\min}, \quad k \geq 0, \tag{20}$$

for some $\sigma_{\min} > 0$. These conditions on σ_k and the construction of ARC's Steps 2–4 allow us to quantify the total iteration count as a function of the successful ones.

THEOREM 2.1 *For any fixed $j \geq 0$, let $\mathcal{S}_j$ be defined in* (18). *Assume that* (19) *holds and let $\overline{\sigma} > 0$ be such that*

$$\sigma_k \leq \overline{\sigma}, \quad \textit{for all } k \leq j. \tag{21}$$

Then

$$j \leq \left\lceil 1 - \frac{\log \gamma_3}{\log \gamma_1} \right\rceil \cdot |\mathcal{S}_j| + \left\lceil \frac{1}{\log \gamma_1} \log\left(\frac{\overline{\sigma}}{\sigma_0}\right) \right\rceil. \tag{22}$$

In particular, if σ_k satisfies (20), *then it also achieves* (19) *with $\gamma_3 = \sigma_{\min}/\overline{\sigma}$, and we have that*

$$j \leq \left\lceil 1 + \frac{2}{\log \gamma_1} \log\left(\frac{\overline{\sigma}}{\sigma_{\min}}\right) \right\rceil \cdot |\mathcal{S}_j|. \tag{23}$$

Proof Apply [2, Theorem 2.1] and the fact that the unsuccessful iterations up to j together with $\mathcal{S}_j$ form a partition of $\{0, \ldots, j\}$. ■

Values for $\overline{\sigma}$ in (21) are provided in (28), and under stronger assumptions, in (57). (Note that due to Lemmas 2.4 and 2.6, the condition required for (28) is achieved by the gradient of convex and strongly convex functions, with appropriate values of ϵ, whenever $\Delta_k > \epsilon$.) Thus, based on the above theorem, we are left with bounding the successful iteration count $|\mathcal{S}_j|$ until iteration j that is within ϵ of the optimum, which we focus on for the remainder of the paper and which has the outcome that the total iteration count up to j is of the same order in ϵ as $|\mathcal{S}_j|$.

2.2 *Some useful properties*

The next lemma summarizes some useful properties of the basic ARC iteration.

Lemma 2.2 *Suppose that the step s_k satisfies* (2).

(i) [5, Lemma 2.1] *Let* AM.1 *hold. Then, for $k \geq 0$, we have that*

$$f(x_k) - m_k(s_k) \geq \frac{\|g_k\|}{6\sqrt{2}} \min\left[\frac{\|g_k\|}{\kappa_B}, \frac{1}{2}\sqrt{\frac{\|g_k\|}{\sigma_k}}\right], \tag{24}$$

and so Δ_k in (16) *is monotonically decreasing*,

$$\Delta_{k+1} \leq \Delta_k, \quad k \geq 0. \tag{25}$$

(ii) [2, Lemma 3.2] *Let* AF.1, AF.4 *and* AM.1 *hold. Also, assume that*

$$\sqrt{\sigma_k \|g_k\|} > \frac{108\sqrt{2}}{1-\eta_2}(\kappa_H + \kappa_B) \overset{\text{def}}{=} \kappa_{HB}. \tag{26}$$

Then iteration k is very successful and

$$\sigma_{k+1} \leq \sigma_k. \tag{27}$$

(iii) [2, Lemma 3.3] *Let* AF.1, AF.4 *and* AM.1 *hold. For any $\epsilon > 0$ and $j \geq 0$ such that $\|g_k\| > \epsilon$ for all $k \in \{0, \ldots, j\}$, we have*

$$\sigma_k \leq \max\left(\sigma_0, \frac{\gamma_2 \kappa_{HB}^2}{\epsilon}\right), \quad 0 \leq k \leq j. \tag{28}$$

A generic property follows.

Lemma 2.3 *Assume* AF.1, AF.4 *and* AM.1 *hold, and that when applying ARC to minimizing f,*

$$\Delta_k \leq \kappa_c \|g_k\|^p, \quad \textit{for all } k \geq 0, \tag{29}$$

for some $\kappa_c > 0$ and $p > 0$, with Δ_k defined in (16). *Then*

$$f(x_k) - m_k(s_k) \geq \kappa_m \Delta_k^{2/p}, \quad \textit{for all } k \geq 0, \tag{30}$$

where κ_{HB} is defined in (26) *and*

$$\kappa_m \overset{\text{def}}{=} \frac{1}{12\sqrt{2}\kappa_c^{2/p}} \min\left(\sqrt{\frac{\kappa_c^{1/p}}{\sigma_0 \Delta_0^{1/p}}}, \frac{1}{\sqrt{\gamma_2}\kappa_{HB}}\right). \tag{31}$$

Proof We first show that

$$\sigma_k \Delta_k^{1/p} \leq \max(\sigma_0 \Delta_0^{1/p}, \gamma_2 \kappa_c^{1/p} \kappa_{HB}^2), \quad \text{for all } k \geq 0. \tag{32}$$

For this, we use the implication

$$\sigma_k \Delta_k^{1/p} > \kappa_c^{1/p} \kappa_{HB}^2 \Longrightarrow \sigma_{k+1} \Delta_{k+1}^{1/p} \leq \sigma_k \Delta_k^{1/p}, \tag{33}$$

which follows from (27) in Lemma 2.2(ii), (29) and (25). Thus, when $\sigma_0 \Delta_0^{1/p} \leq \gamma_2 \kappa_c^{1/p} \kappa_{HB}^2$, (33) implies $\sigma_k \Delta_k^{1/p} \leq \gamma_2 \kappa_c^{1/p} \kappa_{HB}^2$, where the factor γ_2 is introduced for the case when $\sigma_k \Delta_k^{1/p}$ is less than

$\kappa_{\text{c}}^{1/p}\kappa_{\text{HB}}^2$ and the iteration k is not very successful. Letting $k = 0$ in (33) gives the first inequality in (32) when $\sigma_0\Delta_0^{1/p} \geq \gamma_2\kappa_{\text{c}}^{1/p}\kappa_{\text{HB}}^2$, since $\gamma_2 > 1$. Next, we deduce from (24) and (29) that

$$f(x_k) - m_k(s_k) \geq \frac{\Delta_k^{2/p}}{6\sqrt{2}\kappa_{\text{c}}^{1/p}} \min\left(\frac{1}{\kappa_{\text{c}}^{1/p}\kappa_{\text{B}}}, \frac{1}{2\kappa_{\text{c}}^{1/(2p)}\sqrt{\sigma_k\Delta_k^{1/p}}}\right),$$

which together with (32) and the definition of κ_{HB} gives (30) and (31). ■

In the next two sections, we show that when applied to convex and strongly convex functions with globally Lipschitz continuous gradients, the basic ARC algorithm, with only the Cauchy condition for the step computation, satisfies the same upper iteration complexity bounds – namely $\mathcal{O}(\epsilon^{-1})$ and $\mathcal{O}(|\log \epsilon|)$, respectively – as steepest descent when applied to these problem classes [9, Theorems 2.1.14, 2.1.15].

2.3 *Basic ARC complexity on convex objectives*

Let us now assume that

$$\textbf{AF.7} \qquad f \text{ is convex} \tag{34}$$

and also that the level sets of f are bounded, namely

$$\textbf{AF.8} \qquad \|x - x_*\| \leq D, \quad \text{for all } x \text{ such that } f(x) \leq f(x_0), \tag{35}$$

where x_* is any global minimizer of f and $D \geq 1$. The following property specifies the values of p and κ_{c} for which (29) holds in the convex case.

LEMMA 2.4 *Assume* AF.1 *and* AF.7–AF.8 *hold, and let* $f_* = f(x_*)$ *be the (global) minimum of* f. *When applying ARC to minimizing* f, *we have for* (16),

$$\Delta_k \leq D\|g_k\|, \quad \textit{for all } k \geq 0. \tag{36}$$

Proof AF.7 implies $f(x) - f(y) \geq g(y)^{\text{T}}(x - y)$, for all $x, y \in \mathbb{R}^n$. This with $x = x_*$ and $y = x_k$, the Cauchy–Schwarz inequality, $f(x_k) \leq f(x_0)$, and AF.8 give (36). ■

An $\mathcal{O}(\epsilon^{-1})$ upper bound on the ARC iteration count for reaching within ϵ optimality of the objective value is given next.

THEOREM 2.5 *Assume* AF.1, AF.4, AF.7–AF.8 *and* AM.1 *hold, and let* $f_* = f(x_*)$ *be the (global) minimum of* f. *Then, when applying ARC to minimizing* f, *we have*

$$\Delta_j = f(x_j) - f_* \leq \frac{1}{|\mathcal{S}_j|\eta_1\kappa_{\text{m}}^c}, \quad j \geq 0, \tag{37}$$

where $\mathcal{S}_j$ *is defined in* (18), *and* κ_{m}^c *has the expression*

$$\kappa_{\text{m}}^c \overset{\text{def}}{=} \frac{1}{12\sqrt{2}D^2} \min\left(\sqrt{\frac{D}{\sigma_0\Delta_0}}, \frac{1}{\sqrt{\gamma_2}\kappa_{\text{HB}}}\right). \tag{38}$$

Thus, given any $\epsilon > 0$, *ARC takes at most*

$$\left\lceil \frac{\kappa_{\text{s}}^c}{\epsilon} \right\rceil \tag{39}$$

successful iterations and gradient evaluations to generate $f(x_j) - f_* \leq \epsilon$, *where* $\kappa_{\text{s}}^c \overset{\text{def}}{=} (\eta_1\kappa_{\text{m}}^c)^{-1}$.

Proof From (4) and (5), we have

$$f(x_k) - f(x_{k+1}) \geq \eta_1(f(x_k) - m_k(s_k)), \quad k \in \mathcal{S}. \tag{40}$$

Lemma 2.4 implies that the conditions of Lemma 2.3 are satisfied with $p = 1$ and $\kappa_{\mathrm{c}} = D$, and so (30) and (40) imply

$$f(x_k) - f(x_{k+1}) \geq \eta_1 \kappa_{\mathrm{m}}^c \Delta_k^2,$$

where κ_{m}^c is defined in (38). Thus, recalling (16), we have

$$\Delta_k - \Delta_{k+1} \geq \eta_1 \kappa_{\mathrm{m}}^c \Delta_k^2, \quad k \in \mathcal{S},$$

or equivalently,

$$\frac{1}{\Delta_{k+1}} - \frac{1}{\Delta_k} = \frac{\Delta_k - \Delta_{k+1}}{\Delta_k \Delta_{k+1}} \geq \eta_1 \kappa_{\mathrm{m}}^c \frac{\Delta_k}{\Delta_{k+1}} \geq \eta_1 \kappa_{\mathrm{m}}^c, \quad k \in \mathcal{S},$$

where in the last inequality, we used (25). Since $\Delta_k = \Delta_{k+1}$ for any $k \notin \mathcal{S}$, summing up the above inequalities up to j gives

$$\frac{1}{\Delta_j} \geq \frac{1}{\Delta_0} + |\mathcal{S}_j| \eta_1 \kappa_{\mathrm{m}}^c \geq |\mathcal{S}_j| \eta_1 \kappa_{\mathrm{m}}^c, \quad j \geq 0,$$

which gives (37), and hence, also (39). ■

2.4 *Basic ARC complexity on strongly convex objectives*

When we know even more information about f, namely, that f is strongly convex, a global linear rate of convergence, and hence, an improved iteration complexity of at most $\mathcal{O}(\log \epsilon^{-1})$, can be proved for the ARC basic framework, as we show next. This represents, as expected, a marked improvement over the global sublinear rate of convergence obtained in the nonconvex and convex cases, and the corresponding iteration complexity bounds.

Let us assume that f is strongly convex, namely, there exists a constant $\mu > 0$ such that

$$\textbf{AF.9} \qquad f(y) \geq f(x) + g(x)^{\mathrm{T}}(y - x) + \frac{\mu}{2}\|y - x\|^2, \quad \forall\, x, y \in \mathbb{R}^n. \tag{41}$$

When AF.9 holds, f has a unique minimizer, say x_*.

The next property specifies the values of p and κ_{c} for which (29) holds in the strongly convex case.

LEMMA 2.6 *Assume* AF.1 *and* AF.9 *hold, and let x_* be the global minimizer of f. When applying ARC to minimizing f, we have*

$$\Delta_k \leq \frac{1}{2\mu}\|g_k\|^2, \quad \textit{for all } k \geq 0. \tag{42}$$

Proof AF.9 implies $f(y) \leq f(x) + g(x)^{\mathrm{T}}(y - x) + (1/2\mu)\|g(x) - g(y)\|^2$, for all $x, y \in \mathbb{R}^n$ (see [9, Theorem 2.1.10] and its proof). Letting $x = x_*$ and $y = x_k$ in the latter gives (42). ■

An $\mathcal{O}(\log \epsilon^{-1})$ upper bound on the ARC iteration count for reaching within ϵ optimality of the objective value is given next.

THEOREM 2.7 *Assume* AF.1, AF.4, AF.9 *and* AM.1 *hold, and let x_* be the global minimizer of f. Then, when applying ARC to minimizing f, we have*

$$\Delta_j = f(x_j) - f_* \leq (1 - \eta_1 \kappa_{\mathrm{m}}^{sc})^{|\mathcal{S}_j|} \Delta_0, \quad j \geq 0, \tag{43}$$

where $\mathcal{S}_j$ is defined in (18), *and κ_{m}^{sc} has the expression*

$$\kappa_{\mathrm{m}}^{sc} \overset{\text{def}}{=} \frac{\mu}{6\sqrt{2}} \min\left(\frac{1}{\sqrt{\sigma_0\sqrt{2\mu\Delta_0}}}, \frac{1}{\sqrt{\gamma_2\kappa_{\mathrm{HB}}}} \right) \in (0,1). \tag{44}$$

Thus, given any $\epsilon > 0$, ARC takes at most

$$\left\lceil \kappa_{\mathrm{s}}^{sc} \log \frac{\Delta_0}{\epsilon} \right\rceil \tag{45}$$

successful iterations and gradient evaluations to generate $f(x_j) - f_* \leq \epsilon$, where $\kappa_{\mathrm{s}}^{sc} \overset{\text{def}}{=} (\eta_1\kappa_{\mathrm{m}}^{sc})^{-1}$.

Proof Lemma 2.6 implies that (29) holds with $p = 2$ and $\kappa_{\mathrm{c}} = 1/(2\mu)$, and so the conditions of Lemma 2.3 are satisfied and it follows immediately from (30), (31) and (40) and the above choices of p and κ_{c} that

$$\Delta_k - \Delta_{k+1} = f(x_k) - f(x_{k+1}) \geq \eta_1 \kappa_{\mathrm{m}}^{sc} \Delta_k,$$

where κ_{m}^{sc} is defined in (44), which immediately gives (43) since $\Delta_k = \Delta_{k+1}$ for any $k \notin \mathcal{S}$. To show that $\kappa_{\mathrm{m}}^{sc} < 1$, use $\gamma_2 \geq 1$, $\kappa_{\mathrm{HB}} > \kappa_{\mathrm{H}}$ and $\kappa_{\mathrm{H}}/\mu \geq 1$; the latter inequality follows from (42) and from (49) with $x = x_k$. The bound (43) and the inequality $(1 - \eta_1\kappa_{\mathrm{m}}^{sc})^{|\mathcal{S}_j|} \leq \mathrm{e}^{-\eta_1\kappa_{\mathrm{m}}^{sc}|\mathcal{S}_j|}$ imply that $\Delta_j \leq \epsilon$ provided $\mathrm{e}^{-\eta_1\kappa_{\mathrm{m}}^{sc}|\mathcal{S}_j|}\Delta_0 \leq \epsilon$, which then gives (45) by applying the logarithm. ■

2.4.1 *Some remarks on basic ARC's complexity for convex and strongly convex objectives*

Let us comment on the results in Theorems 2.5 and 2.7. Note that, despite AF.7 or AF.9, no convexity assumption was made on m_k, confirming the basic ARC framework to be a steepest-descent-like method. The only model assumption is AM.1. Our results match in order, as a function of the accuracy ϵ, the (nonoptimal) complexity bounds for steepest descent applied to convex and strongly convex objectives with Lipschitz continuous gradients given in [9, Corollary 2.1.2, Theorem 2.1.15].

Let us now discuss the condition numbers that occur in our bounds and their connection to standard measures of conditioning. Consider first the convex-case bound in Theorem 2.5. Assume that the initial regularization parameter σ_0 is chosen small enough, namely $\sigma_0 \leq 1/\|g_0\|$. Then (36) implies that $D/(\sigma_0\Delta_0) \geq 1$ and so (38) becomes $\kappa_{\mathrm{m}}^{c} = (12\sqrt{2\gamma_2}\kappa_{\mathrm{HB}}D^2)^{-1}$, where we also used that $\gamma_2, \kappa_{\mathrm{HB}} \geq 1$. Recalling (26) and that γ_2, η_1 and η_2 are user-chosen constants, we deduce that the bound (39) is a problem-independent constant multiple of

$$\frac{\max(\kappa_{\mathrm{B}}, \kappa_{\mathrm{H}})D^2}{\epsilon},$$

where D measures the size of the $f(x_0)$-level set, and κ_{H} and κ_{B} are the exact and approximate Lipschitz constants of the gradient, respectively. The displayed expression coincides with the bound in [9, Corollary 2.1.2] when the exact Hessian is used in place of B_k so that $\kappa_{\mathrm{B}} = \kappa_{\mathrm{H}}$ and all iterations are successful.

Consider now the strongly convex case and Theorem 2.7. Choosing again $\sigma_0 \leq 1/\|g_0\|$, (42) provides that $\sigma_0\sqrt{2\mu\Delta_0} \leq 1$. Using this, $\gamma_2 \geq 1$ and $\kappa_{\rm HB} \geq 1$, (44) becomes $\kappa_{\rm m}^{sc} = (6\sqrt{2\gamma_2}\kappa_{\rm HB}/\mu)^{-1}$. Employing (26) for the expression of $\kappa_{\rm HB}$, (43) now becomes

$$\Delta_j = f(x_j) - f_* \leq \left(1 - \frac{\overline{\eta}}{c(H)}\right)^{|\mathcal{S}_j|} \Delta_0, \tag{46}$$

where $\overline{\eta} \overset{\rm def}{=} \eta_1(1-\eta_2)/(2592\sqrt{\gamma_2}) \in (0,1)$ and

$$c(H) \overset{\rm def}{=} \frac{\max(\kappa_{\rm H}, \kappa_{\rm B})}{\mu}. \tag{47}$$

Note that $c(H)$ is a uniform upper bound on Hessian's condition number, which equals the common measure $\kappa_{\rm H}/\mu$ when exact Hessians are employed in place of B_k. Recalling that $\eta_{1,2}$ and γ_2 are user-chosen parameters, we deduce that, whenever $\sigma_0 \leq 1/\|g_0\|$, (45) is a problem-independent constant multiple of

$$c(H)\log\frac{\Delta_0}{\epsilon}, \tag{48}$$

where $c(H)$ is defined in (47). When $B_k = H(x_k)$, the function-decrease bound for steepest descent method in [9, Theorem 2.1.15] has a similar form to the simplified bound (46) with the term $1 - \overline{\eta}/c(H)$ replaced by the slightly smaller expression $(c(H)-1)^2/(c(H)+1)^2$.

Note that both (39) and (45) are worse than the complexity bounds of the optimal gradient method [9]. The latter enjoys a worst-case bound of order $\mathcal{O}(1/\sqrt{\epsilon})$ when applied to convex objectives [9, Theorems 2.1.7, §2.2.1], and of order $\mathcal{O}((\sqrt{c(H)}-1)^2/(\sqrt{c(H)}+1)^2\log\epsilon^{-1})$ for strongly convex functions. These two upper bounds match the lower complexity bounds for the minimization of convex and strongly convex functions with Lipschitz continuous gradient by means of gradient methods [9], and hence they are optimal from a worst-case complexity point of view.

2.5 *Complexity of basic ARC generating approximately-optimal gradients*

Let us address the implication of the above results on the ARC's complexity for achieving (7). This issue is important as the latter can be used as a termination condition for ARC, while Δ_k in (16), whose complexity was estimated above, cannot be computed in practice since f_* and x_* are unknown. The following generic property is useful in this and other contexts.

Lemma 2.8 *Let* AF.1 *and* AF.4 *hold, and assume* f *is bounded below by* f_*. *Then*

$$f(x) - f_* \geq f(x) - f(x - \alpha g(x)) \geq \frac{1}{2\kappa_{\rm H}}\|g(x)\|^2, \quad \textit{for all } \alpha \geq 0 \textit{ and } x \in \mathbb{R}^n. \tag{49}$$

Thus, when ARC is applied to minimizing f, *we have*

$$\Delta_k \geq \frac{1}{2\kappa_{\rm H}}\|g_k\|^2, \quad k \geq 0, \tag{50}$$

and so, for any $\epsilon > 0$, $\|g_j\| \leq \epsilon$ *holds whenever*

$$f(x_j) - f_* \leq \frac{\epsilon^2}{2\kappa_{\rm H}}. \tag{51}$$

Proof First-order Taylor expansion and AF.4 give the overestimation property

$$f(x+s) = f(x) + g(x)^{\mathrm{T}}s + \int_0^1 (g(x+ts) - g(x))\,\mathrm{d}t \le f(x) + g(x)^{\mathrm{T}}s + \frac{\kappa_{\mathrm{H}}}{2}\|s\|^2, \quad \text{for all } x, s \in \mathbb{R}^n.$$

Thus, letting $s = -\alpha g(x)$, we obtain

$$f(x) - f(x - \alpha g(x)) \ge \left(\alpha - \frac{\kappa_{\mathrm{H}}}{2}\alpha^2\right)\|g(x)\|^2, \quad \text{for all } \alpha \ge 0.$$

The minimum of the right-hand side of the above inequality is attained at $\alpha_* = 1/\kappa_{\mathrm{H}}$, giving (49). ■

Under the conditions of Theorem 2.5, ARC will take at most $\mathcal{O}(\epsilon^{-2})$ successful iterations to ensure (51) when applied to convex objectives. For strongly convex functions, Theorem 2.7 implies the same order of complexity of $|\log\epsilon|$ for $\|g_j\| \le \epsilon$. (Note that the term $f(x_0) - f_*$ in (37) and (43) can be replaced by $D\|g_0\|$ and $\|g_0\|^2/(2\mu)$, respectively.)

Now recall [2, Corollary 3.4], which states that, when applied to nonconvex objectives, the basic ARC scheme takes at most $\mathcal{O}(\epsilon^{-2})$ iterations to generate a *first* iterate k with $\|g_j\| \le \epsilon$. Hence, we see that the difference between the convex and nonconvex cases is not so great, and the bound improvement (for g_j) is somewhat slight. Namely, as the bound on g_j in the convex case was obtained from that on the function values $f(x_j)$ which decrease monotonically, it follows from (50) that once $\|g_k\| \le \epsilon$, it will remain as such for all subsequent iterations, and so the $\mathcal{O}(\epsilon^{-2})$ iteration bound represents the maximum *total* number of (successful) iterations with $\|g_k\| > \epsilon$ that may occur. Clearly, there is a marked improvement in ARC's worst-case complexity for the strongly convex case.

3. The complexity of second-order ARC variants

Let us now consider the complexity of Algorithm 1.2 with inner iteration termination criteria (10) and (11), namely of the $\mathrm{ARC}_{(\mathrm{S})}$ and $\mathrm{ARC}_{(\mathrm{g2})}$ variants. For the remainder of the paper, we assume that

AF.3 $$f \in C^2(\mathbb{R}^n). \tag{52}$$

While no assumption on the Hessian of f being globally or locally Lipschitz continuous has been imposed in the complexity results of Section 2.2, we now require that the objective's Hessian is globally Lipschitz continuous on the path of the iterates, namely, there exists a constant $L > 0$ independent of k such that

AF.6 $$\|H(x) - H(x_k)\| \le L\|x - x_k\|, \quad \text{for all } x \in [x_k, x_k + s_k] \text{ and all } k \ge 0, \tag{53}$$

and that B_k and $H(x_k)$ agree along s_k in the sense that

AM.4 $$\|(H(x_k) - B_k)s_k\| \le C\|s_k\|^2, \quad \text{for all } k \ge 0 \text{ and some constant } C > 0. \tag{54}$$

By using finite differences on the gradient for computing B_k, we showed in [4] that AM.4 can be achieved in $\mathcal{O}(n|\log\epsilon|)$ additional iterations and gradient evaluations (for any user-chosen constant C).

Next, we recall some results for ARC$_{(\mathrm{h})}$, in particular, necessary conditions for the global subproblem solution (8) and expressions for the model decrease (see Lemma 3.1(i)); also, some general properties that hold for a large class of (nonconvex) functions (see Lemma 3.1(ii) and (iii)).

Lemma 3.1 (i) [5, Lemmas 3.2, 3.3] *Let s_k be the global minimizer of* (8) *for any $k \geq 0$. Then,*

$$g_k^\top s_k + s_k^\top B_k s_k + \sigma_k \|s_k\|^3 = 0, \tag{55}$$

and

$$f(x_k) - m_k(s_k) = \tfrac{1}{2} s_k^{\mathrm{T}} B_k s_k + \tfrac{2}{3}\sigma_k \|s_k\|^3. \tag{56}$$

(ii) [5, Lemma 5.2] *Let* AF.3, AF.6 *and* AM.4 *hold. Then,*

$$\sigma_k \leq \max(\sigma_0, \tfrac{3}{2}\gamma_2(C+L)) \overset{\text{def}}{=} L_0, \quad \textit{for all } k \geq 0. \tag{57}$$

(iii) [2, Lemma 5.2] *Let* AF.3–AF.4, AF.6, AM.4 *and* TC.s *hold. Then, s_k satisfies*

$$\|s_k\| \geq \kappa_g \sqrt{\|g_{k+1}\|} \quad \textit{for all successful iterations } k, \tag{58}$$

where κ_g is the positive constant

$$\kappa_g \overset{\text{def}}{=} \sqrt{\frac{1-\kappa_\theta}{L + C + L_0 + \kappa_\theta \kappa_{\mathrm{H}}}}. \tag{59}$$

Note that in our second-order ARC variants in [2,5], we employ the more general condition (55) and an approximate nonnegative curvature requirement [5, (3.12)] for defining the choice of s_k, which may hold at other points (of local minimum) than the global minimizer over $\mathcal{L}_k$ as prescribed by (8). When the model is convex, as it is often the case here, such situations do not arise.

The bound (58) ensures that the step s_k does not become too small compared with the size of the gradient, and it is a crucial ingredient for obtaining, as shown in [2, Corollary 5.3], an $\mathcal{O}(\epsilon^{-3/2})$ upper bound on the iteration count of ARC$_{(\mathrm{S})}$ to generate $\|g_k\| \leq \epsilon$ for general nonconvex functions. Next, we improve the order of this bound for convex and strongly convex objectives.

Despite solving the subproblem to higher accuracy than the generic ARC framework, the second-order ARC variants still only evaluate the objective function and its gradient once in each (major) iteration and each successful iteration, respectively; hence, the correspondence between (successful) iteration count and the number of (gradient) function evaluations continues to hold. Recall also Theorem 2.1 that relates the total number of iterations to that of successful ones.

3.1 *ARC$_{(\mathrm{g2})}$ complexity on convex objectives*

Here, we prove an $\mathcal{O}(1/\sqrt{\epsilon})$ iteration upper bound for ARC$_{(\mathrm{g2})}$ to achieve (12), which improves the steepest-descent-like bound of order $1/\epsilon$ for basic ARC in Theorem 2.5.

A stronger requirement than AF.6 is required in this section, namely, that the Hessian is globally Lipschitz continuous

$$\textbf{AF.6}' \qquad \|H(x) - H(y)\| \leq L\|x-y\|, \quad \text{for all } x, y \in \mathbb{R}^n. \tag{60}$$

Note that AF.6′ and AF.8 imply AF.4 on the $f(x_0)$-level set of f, which is the required domain of gradient Lipschitz continuity for the results in this section.

We also employ the true Hessian values for B_k, namely, we make the following choice in $\mathrm{ARC}_{(\mathrm{g}2)}$,

$$B_k = H(x_k), \quad \text{for all } k \geq 0. \tag{61}$$

Thus, AM.4 holds in this case with $C = 0$, and AF.4 (or AF.6′ and AF.8) implies AM.1.

A useful lemma is given first.

Lemma 3.2 *Let* AF.3, AF.6′ *and* AF.7–AF.8 *hold. Let* $f_* = f(x_*)$ *be the (global) minimum of* f. *Consider the subproblem* (8) *with* $B_k = H(x_k)$ *and for a(ny) subspace* $\mathcal{L}_k$ *of* $\mathbb{R}^n$ *with* $g \in \mathcal{L}_k$. *Then,*

$$\min_{s \in \mathcal{L}_k} m_k(s) \leq f(x_k) - 2\kappa^c_{\mathrm{m(g2)}}[f(x_k) - f(x_k + s_k^*)]^{3/2}, \tag{62}$$

where s_k^* *is a (global) minimizer of* $f(x_k + s)$ *over* $s \in \mathcal{L}_k$, *and where*

$$\kappa^c_{\mathrm{m(g2)}} \stackrel{\mathrm{def}}{=} (6D\sqrt{6DL_1})^{-1} \quad \textit{and} \quad L_1 \stackrel{\mathrm{def}}{=} \max(\sigma_0, \gamma_2 L, \kappa_{\mathrm{H}}). \tag{63}$$

Proof From AF.3 and AF.6′, we have the overestimation property

$$\left| f(x_k + s) - f(x_k) - s^{\mathrm{T}} g_k - \frac{1}{2} s^{\mathrm{T}} H(x_k) s \right| \leq \frac{L}{6}\|s\|^3, \quad s \in \mathbb{R}^n, \tag{64}$$

and so, from (1) and $B_k = H(x_k)$, we have

$$m_k(s) \leq f(x_k + s) + \frac{2\sigma_k + L}{6}\|s\|^3, \quad s \in \mathbb{R}^n.$$

Employing (57) and $\gamma_2 \geq 1$, we further obtain

$$m_k(s) \leq f(x_k + s) + L_1\|s\|^3, \quad s \in \mathbb{R}^n, \tag{65}$$

where L_1 is defined in (63). (Note that κ_{H} is not needed as yet in the definition of L_1; it will be useful later as we shall see.) Minimizing on both sides of (65) gives the first inequality below

$$\min_{s \in \mathcal{L}_k} m_k(s) \leq \min_{s \in \mathcal{L}_k}\{f(x_k + s) + L_1\|s\|^3\} \leq \min_{\alpha \in [0,1]}\{f(x_k + \alpha s_k^*) + L_1\alpha^3\|s_k^*\|^3\}, \tag{66}$$

where the second inequality follows from the definition of s_k^* which gives $\alpha s_k^* \in \mathcal{L}_k$, for all $\alpha \in [0, 1]$. From AF.7, we have $f(x_k + \alpha s_k^*) \leq (1 - \alpha)f(x_k) + \alpha f(x_k + s_k^*)$, for all $\alpha \in [0, 1]$, and so, from (66),

$$\min_{s \in \mathcal{L}_k} m_k(s) \leq f(x_k) + \min_{\alpha \in [0,1]}\{\alpha[f(x_k + s_k^*) - f(x_k)] + L_1\alpha^3\|s_k^*\|^3\}. \tag{67}$$

The construction of the algorithm implies $f(x_k) \leq f(x_0)$, so that $\|x_k - x_*\| \leq D$ due to AF.8. Furthermore, $f(x_k + s_k^*) \leq f(x_k)$, and so $\|x_k + s_k^* - x_*\| \leq D$. Thus, $\|s_k^*\| \leq \|x_k - x_*\| + \|x_k + s_k^* - x_*\| \leq 2D$, and (67) implies

$$\min_{s \in \mathcal{L}_k} m_k(s) \leq f(x_k) + \min_{\alpha \in [0,1]}\{\alpha[f(x_k + s_k^*) - f(x_k)] + 8\alpha^3 L_1 D^3\}. \tag{68}$$

The minimum in the right-hand side of (68) is attained at

$$\alpha_k^* = \min\{1, \hat{\alpha}_k\}, \quad \text{where } \hat{\alpha}_k := \frac{\sqrt{f(x_k) - f(x_k + s_k^*)}}{2D\sqrt{6L_1 D}}.$$

Let us show that $\hat{\alpha}_k \leq 1$, namely $f(x_k) - f(x_k + s_k^*) \leq 24L_1D^3$. AF.7 gives the first inequality

$$f(x_k + s_k^*) - f(x_k) \geq g_k^{\mathrm{T}} s_k^* \geq -\|g_k\| \cdot \|s_k^*\| \geq -2D\|g_k\| = -2D\|g_k - g(x^*)\| \geq -2\kappa_{\mathrm{H}} D^2,$$

where we also used the Cauchy–Schwarz inequality, the bound on s_k^* just before (68), AF.4 and AF.8. Since we assumed in AF.8 that $D \geq 1$, and the definition of L_1 implies $L_1 \leq \kappa_{\mathrm{H}}$, we conclude that $f(x_k + s_k^*) - f(x_k) \geq -2\kappa_{\mathrm{H}}D^3 \geq -2L_1D^3 \geq -24L_1D^3$. Thus, $\alpha_k^* = \hat{\alpha}_k$ and substituting the above value of $\hat{\alpha}_k$ in (68), we deduce (62) with the notation (63). ■

The main result of this section follows.

Theorem 3.3 *Let* AF.3, AF.6′ *and* AF.7–AF.8 *hold. Let* $f_* = f(x_*)$ *be the* (*global*) *minimum of* f. *Apply* ARC$_{(\mathrm{g2})}$ *with the choices* (20) *and* (61) *to minimizing* f. *Then,*

$$\Delta_j = f(x_j) - f_* \leq \frac{1}{(|\mathcal{S}_j|\eta_1\beta\kappa^c_{\mathrm{m(g2)}})^2}, \quad j \geq 0, \tag{69}$$

where $\mathcal{S}_j$ *is defined in* (18), $\kappa^c_{\mathrm{m(g2)}}$ *in* (63) *and*

$$\beta \overset{\text{def}}{=} \frac{1}{2}\min\left(1, \frac{\kappa_{\mathrm{G}}^{3/2}}{4(\kappa_{\mathrm{H}}D)^{3/2}}\right) \quad \textit{with} \quad \kappa_{\mathrm{G}} \overset{\text{def}}{=} \frac{\sigma_{\min}(\kappa^c_{\mathrm{m(g2)}})^2}{4\kappa_\theta^2\kappa_{\mathrm{H}}^3}. \tag{70}$$

Thus, given any $\epsilon > 0$, ARC$_{(\mathrm{g2})}$ *takes at most*

$$\left\lceil \frac{\kappa^c_{\mathrm{s(g2)}}}{\sqrt{\epsilon}} \right\rceil \tag{71}$$

successful iterations and gradient evaluations to generate $f(x_j) - f_* \leq \epsilon$, *where* $\kappa^c_{\mathrm{s(g2)}} \overset{\text{def}}{=} (\eta_1\beta\kappa^c_{\mathrm{m(g2)}})^{-1}$.

Proof Let $k \in \mathcal{S}$. From (4), (5) and (17), we have

$$f(x_{k+1}) \leq (1-\eta_1)f(x_k) + \eta_1 m_k(s_k) = (1-\eta_1)f(x_k) + \eta_1[m_k(s_k) - m_k(s_k^m)] + \eta_1 m_k(s_k^m), \tag{72}$$

where s_k^m denotes the global minimizer of $m_k(s)$ over $\mathbb{R}^n$. AF.7 implies $H(x_k)$ is positive semi-definite and so $m_k(s)$ is convex, which gives the first inequality below,

$$m_k(s_k) - m_k(s_k^m) \leq \nabla_s m_k(s_k)^{\mathrm{T}}(s_k - s_k^m) \leq \|\nabla_s m_k(s_k)\| \cdot \|s_k - s_k^m\| \leq \kappa_\theta\|g_k\|^3 \cdot \|s_k - s_k^m\|, \tag{73}$$

where the second inequality follows from TC.g2 (11). To bound $\|s_k - s_k^m\|$, recall that both s_k and s_k^m satisfy (55), which implies due to (20) and $B_k = H(x_k)$ being positive semi-definite,

$$\sigma_{\min}\|s\|^3 \leq \sigma_k\|s\|^3 \leq -g_k^{\mathrm{T}}s \leq \|g_k\| \cdot \|s\|, \quad \text{where } s = s_k \text{ or } s = s_k^m.$$

Thus, $\max\{\|s_k\|, \|s_k^m\|\} \leq \sqrt{\|g_k\|/\sigma_{\min}}$, and so

$$\|s_k - s_k^m\| \leq 2\sqrt{\frac{\|g_k\|}{\sigma_{\min}}}.$$

This and (73) now provide the first inequality below,

$$m_k(s_k) - m_k(s_k^m) \leq \frac{2\kappa_\theta}{\sqrt{\sigma_{\min}}}\|g_k\|^{7/2} \leq \frac{2\kappa_\theta\kappa_{\mathrm{H}}\sqrt{2\kappa_{\mathrm{H}}}}{\sqrt{\sigma_{\min}}}\sqrt{\|g_k\|} \cdot \Delta_k^{3/2}, \tag{74}$$

while the second inequality follows from (50). Recalling (72), we are left with bounding $m_k(s_k^m)$ above, for which we use Lemma 3.2 with $\mathcal{L}_k = \mathbb{R}^n$. Then, $s_k^* = x_* - x_k$ and so $f(x_k) - f(x_k +$

$s_k^*) = \Delta_k$, and (62) implies

$$m_k(s_k^m) \leq f(x_k) - 2\kappa^c_{\mathrm{m(g2)}}\Delta_k^{3/2}.$$

Substituting this bound and (74) into (72), we deduce

$$f(x_{k+1}) \leq f(x_k) + 2\eta_1\left(\frac{\kappa_\theta\kappa_{\mathrm{H}}\sqrt{2\kappa_{\mathrm{H}}}}{\sqrt{\sigma_{\min}}}\sqrt{\|g_k\|} - \kappa^c_{\mathrm{m(g2)}}\right)\Delta_k^{3/2},$$

or equivalently, recalling (16) and (70),

$$\Delta_k - \Delta_{k+1} \geq 2\eta_1\kappa^c_{\mathrm{m(g2)}}\left(1 - \sqrt{\frac{\|g_k\|}{2\kappa_{\mathrm{G}}}}\right)\Delta_k^{3/2}.$$

Thus, we have the implication

$$\|g_k\| \leq \frac{\kappa_{\mathrm{G}}}{2} \Longrightarrow \Delta_k - \Delta_{k+1} \geq \eta_1\kappa^c_{\mathrm{m(g2)}}\Delta_k^{3/2}. \tag{75}$$

It remains to prove a bound of the same form as the right-hand side of (75) when $\|g_k\| > \kappa_{\mathrm{G}}/2$. For this, we employ again Lemma 3.2, this time for s_k and the subspace $\mathcal{L}_k$ in the kth iteration of ARC$_{\mathrm{(g2)}}$ with $g \in \mathcal{L}_k$. Thus, noting that the left-hand side of (62) is equal to $m_k(s_k)$ in this case, we employ (62) to upper bound the first inequality in (72) and obtain

$$f(x_{k+1}) \leq f(x_k) - 2\eta_1\kappa^c_{\mathrm{m(g2)}}[f(x_k) - f(x_k + s_k^*)]^{3/2}. \tag{76}$$

Since s_k^* is a global minimizer of $f(x_k + s)$ over $s \in \mathcal{L}_k$ and since $g \in \mathcal{L}_k$, we have the first inequality below, for any $\alpha \geq 0$,

$$f(x_k) - f(x_k + s_k^*) \geq f(x_k) - f(x_k - \alpha g_k) \geq \frac{1}{2\kappa_{\mathrm{H}}}\|g_k\|^2 \geq \frac{\|g_k\|}{2\kappa_{\mathrm{H}}D}\Delta_k,$$

where the second and third inequalities follow from the second inequality in (49) and from (36), respectively. It follows from (76) that

$$f(x_{k+1}) \leq f(x_k) - \eta_1\kappa^c_{\mathrm{m(g2)}}\frac{\|g_k\|^{3/2}}{\kappa_{\mathrm{H}}D\sqrt{2\kappa_{\mathrm{H}}D}}\Delta_k^{3/2},$$

or equivalently,

$$\Delta_k - \Delta_{k+1} \geq \eta_1\kappa^c_{\mathrm{m(g2)}}\frac{\|g_k\|^{3/2}}{\kappa_{\mathrm{H}}D\sqrt{2\kappa_{\mathrm{H}}D}}\Delta_k^{3/2}.$$

Thus, we have the implication

$$\|g_k\| > \frac{\kappa_{\mathrm{G}}}{2} \Longrightarrow \Delta_k - \Delta_{k+1} \geq \eta_1\kappa^c_{\mathrm{m(g2)}}\frac{\kappa_{\mathrm{G}}\sqrt{\kappa_{\mathrm{G}}}}{4\kappa_{\mathrm{H}}D\sqrt{\kappa_{\mathrm{H}}D}}\Delta_k^{3/2}. \tag{77}$$

Finally, we conclude from (75) and (77) that

$$\Delta_k - \Delta_{k+1} \geq 2\eta_1\beta\kappa^c_{\mathrm{m(g2)}}\Delta_k^{3/2}, \quad k \in \mathcal{S}, \tag{78}$$

where β is defined in (70). For any $k \in \mathcal{S}$, we have the identity below

$$\begin{aligned}\frac{1}{\sqrt{\Delta_{k+1}}} - \frac{1}{\sqrt{\Delta_k}} &= \frac{\Delta_k - \Delta_{k+1}}{\sqrt{\Delta_k\Delta_{k+1}}(\sqrt{\Delta_k} + \sqrt{\Delta_{k+1}})}\\ &\geq 2\eta_1\beta\kappa^c_{\mathrm{m(g2)}}\frac{\Delta_k}{\sqrt{\Delta_{k+1}}(\sqrt{\Delta_k} + \sqrt{\Delta_{k+1}})} \geq \eta_1\beta\kappa^c_{\mathrm{m(g2)}},\end{aligned}$$

where we also used (78) and (25), respectively. Thus, recalling that Δ_k remains unchanged on unsuccessful iterations and summing the above up to j, we deduce

$$\frac{1}{\sqrt{\Delta_j}} \geq \frac{1}{\sqrt{\Delta_0}} + |\mathcal{S}_j|\eta_1\beta\kappa^c_{\mathrm{m(g2)}} \geq |\mathcal{S}_j|\eta_1\beta\kappa^c_{\mathrm{m(g2)}}, \quad j \geq 0,$$

which gives (69) and also (71). ■

As TC.g2 is satisfied at the global minimizer of the cubic model $m_k(s)$, the latter can be chosen as the step in our algorithm, which is an efficient choice as far as the cost of the subproblem solution is concerned, provided the problem is medium-sized or the Hessian at the iterates is sparse.

Note the two regimes of analysis in the above proof, namely in the model decreases (75) and (77). To obtain the former 'asymptotic' case, the termination criteria TC.g2 was used, while for the latter 'early stages' case, the first-order condition that the gradient be included in the subspace of minimization and the ensuing decrease along the steepest descent direction were essential. Thus, the construction of ARC$_{\mathrm{(g2)}}$ to behave like steepest descent early on and then naturally switch to higher accuracy as it approaches the solution is reflected in our complexity analysis, with the slight caveat that the (converging) gradient is nonmonotonic and so the distinction between the asymptotic and nonasymptotic regimes is not strict. Furthermore, the nonasymptotic result (77) also holds for ARC$_{\mathrm{(S)}}$, but the termination condition TC.s does not seem strong enough to ensure a similar property to (75) for the asymptotic regime of ARC$_{\mathrm{(S)}}$.

Assuming that σ_0 is chosen small enough, then the condition number $\kappa^c_{\mathrm{m(g2)}}$ in (63) and (69) that characterizes the asymptotic function decrease is a problem-independent constant multiple of $1/\sqrt{\max(\kappa_{\mathrm{H}}, L)D^3}$ while $\beta \in (0, 1)$ in (69) represents the fraction of this function decrease that can be ensured in the nonasymptotic regime when only a Cauchy decrease is achieved.

The iteration complexity of Nesterov and Polyak's cubic regularization algorithm applied to convex problems is analysed in [10, Theorem 1; 11, Theorem 4], and an $\mathcal{O}(1/\sqrt{\epsilon})$ bound is obtained. Here, we relax the requirement that the subproblem be solved globally and exactly, allowing approximate solutions to obtain a same-order bound.

3.1.1 *Complexity of generating approximately-optimal gradient values*

The complexity of ARC$_{\mathrm{(g2)}}$ generating a gradient value $\|g_j\| \leq \epsilon$ can be obtained as described in Section 2.5, by using (51) in Lemma 2.8, and an $\mathcal{O}(1/\epsilon)$ upper bound on the total number of iterations and gradient evaluations with $\|g_k\| > \epsilon$ ensues.

3.2 *ARC$_{\mathrm{(S)}}$ complexity on strongly convex objectives*

For generality purposes (since TC.s is a milder condition than TC.g2), we focus on ARC$_{\mathrm{(S)}}$ in this section, but similar results can be shown for ARC$_{\mathrm{(g2)}}$.

Let us now assume AF.9. Due to AF.3, (41) is equivalent to

$$u^{\mathrm{T}}H(x)u \geq \mu\|u\|^2, \quad \text{for all } u, x \in \mathbb{R}^n. \tag{79}$$

Employing (41) with $y = x$ and $x = x_*$, we deduce that AF.8 is implied by AF.9 with

$$D \leq \sqrt{\frac{2\Delta_0}{\mu}}. \tag{80}$$

The strong convexity of f implies that asymptotically ARC$_{(\mathrm{S})}$ converges Q-quadratically to the (global) minimizer and hence it possesses an associated evaluation complexity of order $\log_2|\log_2\epsilon|$ from some iteration $j_q \geq 0$ onwards [1, §9.5.3].

Lemma 3.4 *Assume* AF.3–AF.4, AF.6, AF.9 *and* AM.4 *hold, and let x_* be the global minimizer of f. Apply ARC$_{(\mathrm{S})}$ to minimizing f and assume that the Rayleigh quotient of B_k along s_k is uniformly bounded away from zero, namely*

$$R_k(s_k) \overset{\text{def}}{=} \frac{s_k^T B_k s_k}{\|s_k\|^2} \geq R_{\min} > 0, \quad \forall\, k \in \mathcal{S}. \tag{81}$$

Then, recalling κ_g defined in (59) *and letting $\delta \overset{\text{def}}{=} \frac{1}{2}(\eta_1 R_{\min}\kappa_g^2\sqrt{\mu})^2$,*

$$\mathcal{N}_f \overset{\text{def}}{=} \{x : f(x) - f(x_*) \leq \delta\} \tag{82}$$

is a neighbourhood of quadratic convergence for f, so that if there exists $j_q \geq 0$ such that $x_{j_q} \in \mathcal{N}_f$ with $\Delta_{j_q} \leq \delta/2$, then $x_k \in \mathcal{N}_f$ for all $k \geq j_q$, and

$$\Delta_{k+1} \leq \frac{1}{\delta}\Delta_k^2, \quad \textit{for all } k \in \mathcal{S} \textit{ and } k \geq j_q. \tag{83}$$

Furthermore, given $\epsilon > 0$, ARC$_{(\mathrm{S})}$ takes at most

$$\left\lceil \log_2 \log_2\left(\frac{\delta}{\epsilon}\right)\right\rceil \tag{84}$$

successful iterations and gradient evaluations from j_q onwards to generate $f(x_j) - f_* \leq \epsilon$.

Proof Let $k \in \mathcal{S}$. Then, (5), (56), (81) and (58) imply

$$\begin{aligned} f(x_k) - f(x_{k+1}) &\geq \eta_1(f(x_k) - m_k(s_k)) \geq \tfrac{1}{2}\eta_1 R_k(s_k)\|s_k\|^2 \geq \tfrac{1}{2}\eta_1 R_{\min}\|s_k\|^2 \\ &\geq \tfrac{1}{2}\eta_1 R_{\min}\kappa_{\mathrm{g}}^2\|g_{k+1}\|, \quad k \in \mathcal{S}. \end{aligned}$$

Lemma 2.6 applies at $k+1$ and so

$$\Delta_{k+1} \leq \frac{1}{2\mu}\|g_{k+1}\|^2.$$

The last two displayed equations further give

$$\Delta_k \geq f(x_k) - f(x_{k+1}) \geq \tfrac{1}{2}\eta_1 R_{\min}\kappa_{\mathrm{g}}^2\sqrt{2\mu\Delta_{k+1}},$$

and so

$$\Delta_{k+1} \leq \frac{1}{\delta}\Delta_k^2, \quad \text{for all } k \in \mathcal{S}, \tag{85}$$

where δ is defined in (82). Thus, the expression of $\mathcal{N}_f$ in (82) follows, as well as (83). Assuming that $x_{j_q} \in \mathcal{N}_f$ with $\Delta_{j_q} \leq \delta/2$, we deduce from (83) that

$$\Delta_j \leq \delta^{1-2^l}\Delta_{j_q}^{2^l}, \quad \text{for any } j \geq j_q, \tag{86}$$

where $l = |\{j_q, j_{q+1}, \ldots, j\} \cap \mathcal{S}|$ denotes the number of successful iterations from j_q up to j. Now employing $\Delta_{j_q} \leq \delta/2$ in (86) shows that $\Delta_j \leq \epsilon$ provided $2^{-2^l}\delta \leq \epsilon$, which gives the bound (84). ■

Remark on satisfying (81) If exact Hessians are used so that $B_k = H(x_k)$ for all k, then AF.9 implies (81) due to (79). Alternatively, (81) can be ensured if AM.4 holds with a sufficiently small C. Namely, note that AF.9, AM.4 and (80) imply

$$\mu \le \frac{s_k^T H_k s_k}{\|s_k\|^2} \le R_k(s_k) + \frac{s_k^T(H_k - B_k)s_k}{\|s_k\|^2} \le R_k(s_k) + C\|s_k\| \le R_k(s_k) + 2CD, \quad k \ge 0.$$

Thus, (81) holds provided $C < \mu/(2D)$. Recall our comments on satisfying AM.4 by finite differencing following (54). ■

We are left with bounding the successful iterations up to j_q, namely, the iterations ARC$_{(S)}$ takes until entering the region of quadratic convergence $\mathcal{N}_f$ (which must happen under the conditions of Corollary 3.5 as x_k converges to the unique global minimizer x_*). From the definition of j_q and $\mathcal{N}_f$ in Lemma 3.4, this is equivalent to counting the successful iterations until

$$\Delta_{j_q} = f(x_{j_q}) - f(x_*) \le \tfrac{1}{2}\delta, \tag{87}$$

with δ defined in (82). The choice of s_k in (8) with $g_k \in \mathcal{L}_k$ implies that ARC$_{(S)}$ always satisfies the Cauchy condition (2) and so the bound in Theorem 2.7 holds. This yields an upper bound on (the successful iterations up to) j_q of order $\log(\Delta_0/\delta)$ and emphasizes again that early on in the running of the algorithm, steepest-descent-like decrease is sufficient even from a worst-case complexity viewpoint. The bound on the total number of successful iterations is then obtained by adding up the bounds on the two distinct phases, up to and then inside the neighbourhood of quadratic convergence.

Corollary 3.5 *Assume* AF.3–AF.4, AF.6, AF.9, AM.1 *and* AM.4 *hold, and let x_* be the global minimizer of f. Apply ARC$_{(S)}$ to minimizing f, assuming that* (81) *holds. Then, given any $\epsilon > 0$, ARC$_{(S)}$ takes, in total, at most*

$$\left\lceil \kappa_s^{sc} \log \frac{2\Delta_0}{\delta} + \log_2 \log_2 \left(\frac{\delta}{\epsilon} \right) \right\rceil \tag{88}$$

successful iterations and gradient evaluations to generate $f(x_j) - f(x_*) \le \epsilon$, where κ_s^{sc} is defined in (45) *and δ in* (82).

Proof The conditions of Theorem 2.7 are satisfied, and so letting $\epsilon = \delta/2$ in (45), we deduce that (87) holds in at most $\lceil \kappa_s^{sc} \log(2\Delta_0/\delta) \rceil$ successful iterations. To bound the number of iterations from j_q to j, we employ Lemma 3.4. Thus, the total number of successful iterations up to j is the sum of these two bounds. ■

Recalling our comments following (46), let us interpret the condition numbers in (88). In particular, provided σ_0 is chosen sufficiently small, we obtain from (48) that κ_s^{sc} is a problem-independent multiple of the bound $c(H)$ in (47) on the condition number of the Hessian matrix $H(x)$. Additionally, if $B_k = H(x_k)$ so that $C = 0$ and $R_{\min} = \mu$, δ in (82) and (88) simplifies to a multiple of $\sqrt{\mu}/c(H)$.

Note that for the nonasymptotic phase of ARC$_{(S)}$, an $\mathcal{O}(1/\sqrt{\delta})$ bound can be deduced similar to the proof of Theorem 3.3. Namely, using Lemma 3.2, which clearly holds for ARC$_{(S)}$, we deduce (76); then employ (49) just as in the first displayed equation after (76) and use (42). Then, the total ARC$_{(S)}$ complexity would be of order $\delta^{-1/2} + \log_2 \log_2(\delta/\epsilon)$, which matches the bounds for cubic regularization with exact subproblem solution in [10, pp. 176–177; 11, pp. 203–204]. Note that such bounds are weaker than the ones we obtained in Corollary 3.5.

3.2.1 *Complexity of generating approximately optimal gradient values*

We have the following result, where the constants have already been defined in Corollary 3.5.

LEMMA 3.6 *Assume* AF.3–AF.4, AF.6, AF.9, AM.1 *and* AM.4 *hold. Apply ARC*$_{(\mathrm{S})}$ *to minimizing* f, *assuming that* (81) *holds. Then,* $\mathcal{N}_g \overset{\mathrm{def}}{=} \{x : \|g(x)\| \leq (\frac{1}{2}\eta_1 R_{\min}\kappa_g)^2 \overset{\mathrm{def}}{=} \zeta\}$ *is a neighbourhood of quadratic convergence for the gradient* g, *namely, there exists* j_q *such that* $x_{j_q} \in \mathcal{N}_g$ *with* $\|g_{j_q}\| \leq \zeta/2$, *then* $x_k \in \mathcal{N}_g$ *for all* $k \geq j_q$, *and*

$$\|g_{k+1}\| \leq \frac{1}{\zeta}\|g_k\|^2, \quad \textit{for all } k \in \mathcal{S} \textit{ and } k \geq j_q. \tag{89}$$

Thus, given $\epsilon > 0$, *ARC*$_{(\mathrm{S})}$ *takes at most*

$$\left\lceil \log_2 \log_2 \left(\frac{\zeta}{\epsilon}\right) \right\rceil \tag{90}$$

successful iterations from j_q *onwards to generate* $\|g_j\| \leq \epsilon$. *Furthermore, to generate* $\|g_{j_q}\| \leq \zeta$, *ARC*$_{(\mathrm{S})}$ *takes at most*

$$\left\lceil 2\kappa_s^{sc} \log \frac{\|g_0\|\sqrt{\kappa_{\mathrm{H}}}}{\zeta\sqrt{\mu}} \right\rceil \tag{91}$$

successful iterations, so that the total number of successful iterations and gradient evaluations required to generate $\|g_j\| \leq \epsilon$ *is at most equal to the sum of the bounds* (90) *and* (91).

Proof AF.9 implies AF.7 which gives

$$f(x_{k+1}) - f(x_k) \geq g_k^{\mathrm{T}} s_k \geq -\|g_k\| \cdot \|s_k\|, \quad k \geq 0.$$

This and the first set of displayed equations in the proof of Lemma 3.4 give the first inequality below

$$\|g_k\| \geq \tfrac{1}{2}\eta_1 R_{\min}\|s_k\| \geq \tfrac{1}{2}\eta_1 R_{\min}\kappa_{\mathrm{g}}\sqrt{\|g_{k+1}\|}, \quad k \in \mathcal{S}, \tag{92}$$

where the latter inequality follows from (58). The expression and properties of $\mathcal{N}_g$ follow. The bound (90) is obtained similar to the proof of (84) in Lemma 3.4. To deduce (91), let $\epsilon = \zeta$ in (51) and (45) and replace Δ_0 in the latter by its upper bound $\|g_0\|^2/(2\mu)$. ■

A similar estimate of a neighbourhood of quadratic convergence for the gradient can be found in [10] for Nesterov and Polyak's cubic regularization algorithm.

3.3 *On the tightness of ARC's complexity bounds*

The question arises as to whether the complexity bounds on ARC's performance on special problem classes presented in this section are too pessimistic, even for the worst case, and could potentially be improved. This is particularly relevant when it comes to the convex case and the corresponding bound of order $1/\sqrt{\epsilon}$ (Theorem 3.3), implying a sublinear rate of convergence of second-order ARC variants on convex functions. (For the strongly convex case, the $\log|\log\epsilon|$ bound can commonly be observed numerically when Q-quadratic convergence takes place.)

Here, we find a convex function that satisfies all the conditions of Theorem 3.3 apart from having bounded level sets and on which ARC takes precisely order $1/\sqrt{\epsilon}$ iterations (and function and gradient evaluations) to generate $f(x_j) - f_* \leq \epsilon$.

Consider a convex function $f \in \mathcal{C}^2(\mathbb{R})$, with

$$f(x) = \mathrm{e}^{-x}, \quad \text{for } x \geq 0. \tag{93}$$

We have the following complexity result, whose proof is given in the appendix.

LEMMA 3.7 *The function* (93) *is convex, bounded below by* $f_* = 0$ *and has bounded above and Lipschitz continuous second derivatives* $f''(x)$ *for* $x \in [0, \infty)$ *with constants* $\kappa_{\mathrm{H}} = L = 1$, *thus satisfying* AF.4, AF.6′ *and* AF.7.

Apply ARC to minimizing (93), *starting with* $x_0 \geq 0$. *On each iteration* k, *compute the step* s_k *as the global minimizer of the model* $m_k(s)$ *in* (1) *with* $B_k = f''(x_k)$ *and with the* (*reasonable*) *choice*

$$\sigma_k := \sigma \geq \frac{L}{2} = \frac{1}{2}, \quad \forall\, k \geq 0, \tag{94}$$

which ensures that every iteration is very successful and that (20) *holds. Then*, AM.1 *and* AM.4 *hold* (*with* $\kappa_{\mathrm{B}} = 1$ *and* $C = 0$), *and ARC takes* $\Theta(\epsilon^{-1/2})$ *total iterations to achieve* $f(x_k) \leq \epsilon$, *where* $\Theta(\cdot)$ *denotes upper and lower bounds of that order.*

Several remarks are in order concerning the above example.

- This example also applies to Nesterov and Polyak's cubic regularization algorithm [10,11]; recall our choice of s_k and σ_k in the above. In particular, it satisfies all the conditions in [10, Theorem 1] including $\sigma_k = L/2$ but except f having bounded level sets. The latter theorem establishes the $\mathcal{O}(\epsilon^{-1/2})$ iteration upper bound for Nesterov and Polyak's cubic regularization.
- Approximate termination criteria like TC.g2 and TC.s do not give better performance than the exact subproblem solution in this case (see the right-hand side plot of basic ARC with the Cauchy condition in Figure 1).
- If Newton's method is applied to this example, the complexity would be better (Figure 1). Similarly, if we allowed σ_k to decrease to zero so that the step approaches the Newton step, the complexity would again improve. Thus, the inefficient behaviour in this example is due to keeping the regularization always switched 'on', and always 'strongly' regularizing. However, we have shown in [3] that for nonconvex problems, Newton's method can behave worse than second-order ARC in the worst case, in fact, it can be as poor as steepest descent. It remains to see whether this is also possible for convex problems, or for problems with bounded level sets.

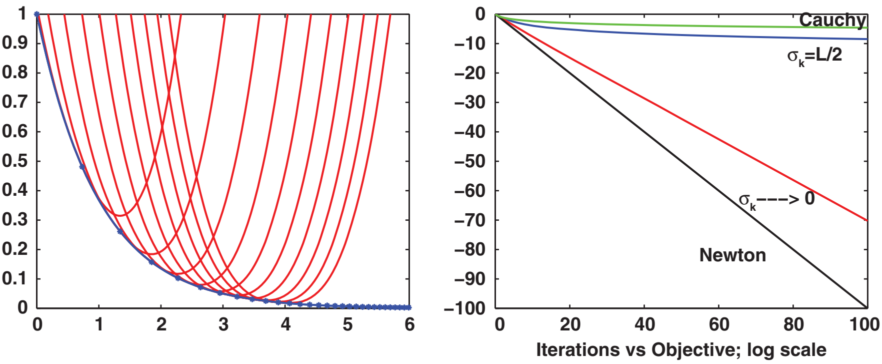

Figure 1. Graph of (93) and the local cubic regularizations at the ARC iterates (left-hand side). Plot of objective values at the iterates on a log scale for different ARC variants and for Newton's method (right-hand side).

4. Conclusions

The behaviour of ARC on some special problem classes was investigated and, as expected, improved complexity bounds were shown when additional structure was assumed to be present in the problem. In particular, upper bounds of order $\mathcal{O}(1/\sqrt{\epsilon})$ and $\mathcal{O}(|\log\kappa| + \log|\log\epsilon|)$ were proved for second-order ARC variants when applied to convex and strongly convex objectives, respectively. For the latter case, the fact that the constant number of steps before entering the region of quadratic convergence is a logarithmic function of condition numbers is an improvement over existing complexity bounds for second-order methods applied to such problems.

We have also given an example of (relatively) inefficient behaviour of second-order ARC on a convex problem with unbounded level sets which takes order $1/\sqrt{\epsilon}$ iterations to reach within ϵ of the optimum. Several open questions remain, such as whether a convex objective with bounded level sets can be found on which the latter iteration bound is attained, or whether Newton's method always has better worst-case complexity than ARC in the convex case.

Acknowledgements

All three authors are grateful to the Royal Society for its support through the International Joint Project 14265. The work of the second author was supported by the EPSRC grant EP/E053351/1.

References

[1] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, 2004.

[2] C. Cartis, N.I.M. Gould, and Ph.L. Toint, *On the oracle complexity of first-order and derivative-free algorithms for smooth nonconvex minimization*, ERGO Tech. Rep. 10-005, School of Mathematics, University of Edinburgh, Edinburgh, 2010.

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

[4] C. Cartis, N.I.M. Gould, and Ph.L. Toint, *Adaptive cubic regularisation methods for unconstrained optimization. Part II: Worst-case function- and derivative-evaluation complexity*, Math. Program. (2010), doi:10.1007/s10107-009-0337-y (online).

[5] C. Cartis, N.I.M. Gould, and Ph.L. Toint, *Adaptive cubic regularisation methods for unconstrained optimization. Part I: Motivation, convergence and numerical results*, Math. Program. 127(2) (2011), pp. 245–295.

[6] A.R. Conn, N.I.M. Gould, and Ph.L. Toint, *Trust-Region Methods*, SIAM, Philadelphia, PA, 2000.

[7] J.E. Dennis and R.B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1983. Reprinted as Classics in Applied Mathematics 16, SIAM, Philadelphia, PA, 1996.

[8] A. Griewank, *The modification of Newton's method for unconstrained optimization by bounding cubic terms*, Tech. Rep. NA/12 (1981), Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, 1981.

[9] Yu. Nesterov, *Introductory Lectures on Convex Optimization*, Kluwer Academic Publishers, Dordrecht, 2004.

[10] Yu. Nesterov, *Accelerating the cubic regularization of Newton's method on convex problems*, Math. Program. 112(1) (2008), pp. 159–181.

[11] Yu. Nesterov and B.T. Polyak, *Cubic regularization of Newton's method and its global performance*, Math. Program. 108(1) (2006), pp. 177–205.

[12] M. Weiser, P. Deuflhard, and B. Erdmann, *Affine conjugate adaptive Newton methods for nonlinear elastomechanics*, Optim. Methods Softw. 22(3) (2007), pp. 413–431.

Appendix

Proof of Lemma 3.7 Apply ARC to minimizing (93), starting at $x_0 \geq 0$, where each s_k is computed as the global minimizer of the cubic model $m_k(s)$, $s \in \mathbb{R}$, with $B_k = f''(x_k)$, which thus has the expression

$$m_k(s) = \mathrm{e}^{-x_k} - s\mathrm{e}^{-x_k} + \frac{1}{2}s^2\mathrm{e}^{-x_k} + \frac{\sigma_k}{3}|s|^3, \quad s \in \mathbb{R}. \tag{A1}$$

Let us compute an explicit expression for s_k from $x_k \geq 0$. We have

$$\nabla m_k(s) = -\mathrm{e}^{-x_k} + s\mathrm{e}^{-x_k} + \sigma_k s|s|, \quad s \in \mathbb{R}.$$

Distinguishing between the case $s \geq 0$ and $s < 0$, we deduce that there is no stationary point – and hence minimizer – in the latter case, and that the former case yields the unique solution

$$s_k = \frac{2}{1 + \sqrt{1 + 4\sigma_k \mathrm{e}^{x_k}}} \tag{A2}$$

to $\nabla m_k(s) = 0$. Thus, $s_k > 0$, and since $x_0 \geq 0$, all iterates satisfy

$$x_k \geq 0, \quad \forall\, k \geq 0; \tag{A3}$$

so we only need to consider$f(x)$ for $x \geq 0$, which clearly satisfies AF.4 with $\kappa_{\mathrm{H}} = 1$, AM.1 with $\kappa_{\mathrm{B}} = 1$, AF.6$'$ with $L = 1$, AM.4 with $C = 0$, and AF.7. Furthermore, AF.6$'$, (6) and (64) provide the implication

$$\sigma_k \geq \frac{L}{2} \Longrightarrow k \text{ is very successful.}$$

This and (94) imply that all iterations k are very successful and that the iterates satisfy $x_{k+1} = x_k + s_k$, with s_k in (A2), for all $k \geq 0$. Furthermore, (93) and $\mathrm{e}^{-x} \in (0, 1]$ for $x \geq 0$ provide the following same-order bounds on s_k in (A2)

$$\frac{1}{\sqrt{\sigma}}\mathrm{e}^{-(1/2)x_k} > s_k \geq \frac{2}{1 + \sqrt{1 + 4\sigma}}\mathrm{e}^{-(1/2)x_k}, \quad \forall\, k \geq 0,$$

which further become, by letting

$$c_1 := \frac{1}{\sqrt{\sigma}} \quad \text{and} \quad c_2 := \frac{2}{1 + \sqrt{1 + 4\sigma}},$$

$$c_1\mathrm{e}^{-(1/2)x_k} \geq s_k \geq c_2\mathrm{e}^{-(1/2)x_k}, \quad \forall\, k \geq 0. \tag{A4}$$

From (93), we have

$$f(x_{k+1}) = \mathrm{e}^{-x_k - s_k} = \mathrm{e}^{-x_k}\mathrm{e}^{-s_k} = f(x_k)\mathrm{e}^{-s_k},$$

which further gives, by employing (A4),

$$f(x_k)\mathrm{e}^{-c_1\mathrm{e}^{-(1/2)x_k}} \leq f(x_{k+1}) \leq f(x_k)\mathrm{e}^{-c_2\mathrm{e}^{-(1/2)x_k}}, \quad k \geq 0.$$

Employing again (93), we obtain

$$f(x_k)\mathrm{e}^{-c_1\sqrt{f(x_k)}} \leq f(x_{k+1}) \leq f(x_k)\mathrm{e}^{-c_2\sqrt{f(x_k)}}, \quad k \geq 0. \tag{A5}$$

Since the following bounds hold for the exponential function

$$1 - y \leq \mathrm{e}^{-y} \leq 1 - y + \frac{y^2}{2}, \quad y \in [0, 1], \tag{A6}$$

it follows from (A5), (A6) and$f_k = f(x_k) \in (0, 1]$ that

$$f_k(1 - c_1\sqrt{f_k}) \leq f_{k+1} \leq f_k\left(1 - c_2\sqrt{f_k} + \frac{c_2^2}{2}f_k\right), \quad k \geq 0, \tag{A7}$$

and so

$$c_2 f_k\sqrt{f_k}\left(1 - \frac{c_2}{2}\sqrt{f_k}\right) \leq f_k - f_{k+1} \leq c_1 f_k\sqrt{f_k}, \quad k \geq 0. \tag{A8}$$

Furthermore, using $c_2 \in (0,1)$ and $f_k \in (0,1]$, we obtain

$$c_3 f_k\sqrt{f_k} \leq f_k - f_{k+1} \leq c_1 f_k\sqrt{f_k}, \quad k \geq 0, \tag{A9}$$

where $c_3 := c_2(1 - c_2/2)$. Next, we deduce an explicit expression of f_k in terms of k.

It is easy to check that, for any $k \geq 0$, we have

$$\frac{1}{\sqrt{f_{k+1}}} - \frac{1}{\sqrt{f_k}} = \frac{f_k - f_{k+1}}{\sqrt{f_k f_{k+1}}(\sqrt{f_k} + \sqrt{f_{k+1}})},$$

and so, employing (A9) on the right-hand side of the above, we obtain

$$\frac{c_3 f_k}{\sqrt{f_{k+1}}(\sqrt{f_k} + \sqrt{f_{k+1}})} \leq \frac{1}{\sqrt{f_{k+1}}} - \frac{1}{\sqrt{f_k}} \leq \frac{c_1 f_k}{\sqrt{f_{k+1}}(\sqrt{f_k} + \sqrt{f_{k+1}})},$$

and furthermore, using $f_k \geq f_{k+1}$, we deduce

$$\frac{c_3}{2} \leq \frac{1}{\sqrt{f_{k+1}}} - \frac{1}{\sqrt{f_k}} \leq \frac{c_1}{2} \cdot \frac{f_k}{f_{k+1}}, \quad k \geq 0. \tag{A10}$$

Now let us give an upper bound on f_k/f_{k+1}. Using (A5) and $f_k \in (0,1]$, we deduce

$$\frac{f_k}{f_{k+1}} \leq \mathrm{e}^{c_1\sqrt{f_k}} \leq \mathrm{e}^{c_1}, \quad k \geq 0.$$

Thus, (A10) gives

$$\frac{c_3}{2} \leq \frac{1}{\sqrt{f_{k+1}}} - \frac{1}{\sqrt{f_k}} \leq \frac{c_1}{2}\mathrm{e}^{c_1}, \quad k \geq 0. \tag{A11}$$

Summing up (A11) over $i \in \{0, \ldots, k\}$, we obtain

$$k\frac{c_3}{2} + \frac{1}{\sqrt{f_0}} \leq \frac{1}{\sqrt{f_k}} \leq k\frac{c_1}{2}\mathrm{e}^{c_1} + \frac{1}{\sqrt{f_0}}, \quad k \geq 0,$$

and thus,

$$k\frac{c_3}{2} \leq \frac{1}{\sqrt{f_k}} \leq k\max(c_1\mathrm{e}^{c_1}, 2\mathrm{e}^{(1/2)x_0}), \quad k \geq 0. \tag{A12}$$

Finally, (A12) is equivalent to

$$\frac{1}{k^2}\min\left(\frac{1}{c_1^2}\mathrm{e}^{-2c_1}, \frac{1}{4}f_0\right) \leq f_k \leq \frac{1}{k^2} \cdot \frac{4}{c_3^2}, \quad k \geq 0,$$

which gives the desired complexity result of the lemma. ■