Science & Technology Facilities Council

Technical Report
RAL-TR-2010-030

Evaluation complexity of adaptive cubic regularization methods for convex unconstrained optimization

C Cartis, NIM Gould, PL Toint

November 2010

©2010 Science and Technology Facilities Council

Enquiries about copyright, reproduction and requests for additional copies of this report should be addressed to:

RAL Library
Science and Technology Facilities Council
Rutherford Appleton Laboratory
Harwell Science and Innovation Campus
Didcot
OX11 0QX

Tel: +44(0)1235 445384
Fax: +44(0)1235 446403
email: library@rl.ac.uk

Science and Technology Facilities Council reports are available online at: http://epubs.cclrc.ac.uk/

ISSN 1358-6254

Neither the Council nor the Laboratory accept any responsibility for loss or damage arising from the use of information contained in any of their reports or in any communication about their tests or investigations.

Evaluation complexity of adaptive cubic regularization methods for convex unconstrained optimization

Coralia Cartis[1,2], Nicholas I. M. Gould[2,3,4] and Philippe L. Toint[2,5]

ABSTRACT

algorithms described in Cartis, Gould & Toint (2009, 2010) 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 (2004, 2008) and Nesterov & Polyak (2006) for these same problem classes, without employing the Hessian's Lipschitz constant explicitly in the algorithms or requiring 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.

---

[1] School of Mathematics, The King's Buildings, University of Edinburgh, EH9 3JZ, Scotland, EU. Email: coralia.cartis@ed.ac.uk .

[2] The authors are grateful to the Royal Society for its support through the International Joint Project 14265.

[3] Computational Science and Engineering Department, Rutherford Appleton Laboratory, Chilton, Oxfordshire, OX11 0QX, England, EU. Email: nick.gould@stfc.ac.uk .
Current reports available from "http://www.numerical.rl.ac.uk/reports/reports.shtml".

[3] This work was supported by the EPSRC grant EP/E053351/1.

[5] Department of Mathematics, Facultés Universitaires ND de la Paix, 61, rue de Bruxelles, B-5000 Namur, Belgium, EU. Email : philippe.toint@fundp.ac.be .
Current reports available from "http://www.fundp.ac.be/∼phtoint/pht/publications.html".

Computational Science and Engineering Department
Atlas Centre
Rutherford Appleton Laboratory
Oxfordshire OX11 0QX

November 8, 2010

1 Introduction

State-of-the-art methods for unconstrained smooth optimization typically depend on trust-region [6] or line-search [7] techniques to globalise Newton-like iterations. Of late, a third alternative, in which a local cubic over-estimator of the objective is used as the basis for a regularization strategy for the step computation, has been proposed [9, 12, 13]; see [2, §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, 3]. Our objective in this paper is to show that the complexity 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 ARC (Adaptive Regularization with Cubics) framework, a possibly nonconvex model

$$m_k(s) \stackrel{\text{def}}{=} f(x_k) + s^T g_k + \tfrac{1}{2} s^T B_k s + \tfrac{1}{3}\sigma_k \|s\|^3, \tag{1.1}$$

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 $s \in \mathbb{R}^n$ [2, §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, 3, S2.1]. The generic ARC framework [2, 3, §2.1] may be summarised as follows:

Algorithm 1.1: Adaptive Regularization using Cubics (ARC) [2, 3].

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

$$m_k(s_k) \le m_k(s_k^{\mathrm{C}}), \tag{1.2}$$

where the Cauchy point

$$s_k^{\mathrm{C}} = -\alpha_k^{\mathrm{C}} g_k \ \text{ and } \ \alpha_k^{\mathrm{C}} = \arg\min_{\alpha \in \mathbb{R}_+} m_k(-\alpha g_k). \tag{1.3}$$

2. Compute $f(x_k + s_k)$ and

$$\rho_k = \frac{f(x_k) - f(x_k + s_k)}{f(x_k) - m_k(s_k)}. \tag{1.4}$$

3. Set

$$x_{k+1} = \begin{cases} x_k + s_k & \text{if } \rho_k \geq \eta_1 \\ x_k & \text{otherwise.} \end{cases} \tag{1.5}$$

4. Set

$$\sigma_{k+1} \in \begin{cases} (0, \sigma_k] & \text{if } \rho_k > \eta_2 & \text{[very successful iteration]} \\ [\sigma_k, \gamma_1 \sigma_k] & \text{if } \eta_1 \leq \rho_k \leq \eta_2 & \text{[successful iteration]} \\ [\gamma_1 \sigma_k, \gamma_2 \sigma_k] & \text{otherwise.} & \text{[unsuccessful iteration]} \end{cases} \tag{1.6}$$

For a detailed description of the algorithm construction, including a justification that (1.2)–(1.4) are well-defined until termination, see [2]. The above ARC framework only requires the Cauchy condition (1.2) for the calculation of the step s_k, which is sufficient to ensure global convergence of ARC both to first-order critical points [2, §2.1] and with steepest-descent-like function-evaluation complexity bounds of order ϵ^{-2} [3, §3] to guarantee

$$\|g_k\| \leq \epsilon. \tag{1.7}$$

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 (1.2) in a computationally inexpensive way (see [2, §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 §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 [2, Thm.3.1], [9, 12]. 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 (1.2) remains satisfied—until a suitable termination condition is achieved. (For instance, in our ARC implementation [2], 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 (1.10) and (1.11) below.

Algorithm 1.2: ARC$_{(\mathrm{h})}$ [2, 3].

In each iteration k of Algorithm 1.1, compute s_k in Step 1 as the global minimizer of

$$\min_{s\in\mathbb{R}^n} m_k(s) \ \text{ subject to } s \in \mathcal{L}_k, \tag{1.8}$$

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

$$\boxed{\textbf{TC.h}} \qquad \begin{array}{c} \|\nabla_s m_k(s_k)\| \le \theta_k \|g_k\|, \ \text{ where } \ \theta_k \stackrel{\text{def}}{=} \kappa_\theta \min(1, h_k) \\ \text{and } \ h_k \stackrel{\text{def}}{=} h_k(\|s_k\|, \|g_k\|) > 0, \end{array} \tag{1.9}$$

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 (1.8) may only need to hold along the one-dimensional subspace determined by s_k [2, (3.11), (3.12)], provided (1.2) holds.

For ARC$_{(\mathrm{h})}$ 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

$$\boxed{\textbf{TC.s}} \qquad \|\nabla_s m_k(s_k)\| \le \theta_k \|g_k\|, \ \text{ where } \ \theta_k = \kappa_\theta \min(1, \|s_k\|). \tag{1.10}$$

and

$$\boxed{\textbf{TC.g2}} \qquad \|\nabla_s m_k(s_k)\| \le \theta_k \|g_k\|, \ \text{ where } \ \theta_k = \kappa_\theta \min\left(1, \|g_k\|^2\right). \tag{1.11}$$

Forthwith, we refer to ARC$_{(\mathrm{h})}$ with TC.s and with TC.g2 as ARC$_{(\mathrm{S})}$ and ARC$_{(\mathrm{g2})}$, respectively. The benefit of requiring the more stringent conditions (1.8), and (1.10) or (1.11), in the above ARC variants is that ARC$_{(\mathrm{S})}$ and ARC$_{(\mathrm{g2})}$ are also guaranteed to converge locally Q-quadratically and globally to second-order critical points [2, §4.2,§5], and to have improved function-evaluation complexity of order $\epsilon^{-3/2}$ to ensure (1.7) [3, §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$_{(\mathrm{S})}$ and/or ARC$_{(\mathrm{g2})}$ 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 to the nonconvex case. Specifically, generic ARC (Algorithm 1.1) takes at most order ϵ^{-1} and $\log \epsilon^{-1}$ function-evaluations to reach the neighbourhood

$$f(x_k) - f_* \le \epsilon \tag{1.12}$$

of the (global) minimum f_* of convex and strongly convex objectives, respectively, with Lipschitz continuous gradients; unsurprinsingly, due to the simple Cauchy decrease condition (1.2) required on the step, these bounds match in order those for standard steepest-descent methods on the same classes of objectives [10].

When applied to convex objectives with bounded level sets and globally Lipschitz continuous Hessian, $\text{ARC}_{(\text{g2})}$ with $B_k = H(x_k)$ will reach approximate optimality in the (1.12) sense in at most $\mathcal{O}(\epsilon^{-1/2})$ function-evaluations; this matches in order the bound obtained in [11, 12] for cubic regularization on the same problem class when the exact subproblem solution is computed in each iteration. Note that asymptotically, in $\text{ARC}_{(\text{g2})}$, the subproblem is solved to higher accuracy than in $\text{ARC}_{(\text{S})}$, which seems to be crucial when deriving the improved bound compared to 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 $\text{ARC}_{(\text{S})}$ and $\text{ARC}_{(\text{g2})}$ (with approximate Hessians as B_k) require at most $\mathcal{O}(|\log\kappa| + |\log\log\epsilon|)$ function-evaluations to satisfy (1.12), 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 [11, 12] 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 labeling used throughout the paper was chosen to maintain consistency with notation introduced in [2, 3]. The structure of the paper is as follows. Section 2 analyzes the complexity of basic ARC, while Section 3 that of the second-order variants $\text{ARC}_{(\text{S})}$ and $\text{ARC}_{(\text{g2})}$, 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

AF.1 $$f \in C^1(\mathbb{R}^n), \tag{2.1}$$

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

AF.4 $$\|g(x) - g(y)\| \leq \kappa_{\text{H}}\|x - y\|, \text{ for all } x,\, y \in X, \text{ and some } \kappa_{\text{H}} \geq 1. \tag{2.2}$$

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

AM.1 $$\|B_k\| \leq \kappa_{\rm B}, \quad \text{for all } k \geq 0, \text{ and some } \kappa_{\rm B} \geq 1. \tag{2.3}$$

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

$$\Delta_k \stackrel{\rm def}{=} f(x_k) - f_*, \quad \text{for all } k \geq 0. \tag{2.4}$$

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 (1.4)), 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

$$\mathcal{S} \stackrel{\rm def}{=} \{k \geq 0 : \ k \text{ successful or very successful in the sense of (1.6)}\}, \tag{2.5}$$

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

$$\mathcal{S}_j \stackrel{\rm def}{=} \{k \leq j : k \in \mathcal{S}\}, \tag{2.6}$$

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

$$\sigma_{k+1} \geq \gamma_3 \sigma_k, \quad \text{for some } \gamma_3 \in (0, 1]. \tag{2.7}$$

Note that (2.7) 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

$$\sigma_k \geq \sigma_{\min}, \quad k \geq 0, \tag{2.8}$$

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 (2.6). Assume that (2.7) holds and let $\overline{\sigma} > 0$ be such that

$$\sigma_k \leq \overline{\sigma}, \quad \text{for all } k \leq j. \tag{2.9}$$

Then

$$j \leq \left\lceil 1 - \frac{\log \gamma_3}{\log \gamma_1} \right\rceil \cdot |\mathcal{S}_j| + \left\lceil \frac{1}{\log \gamma_1} \log \left(\frac{\overline{\sigma}}{\sigma_0} \right) \right\rceil. \tag{2.10}$$

In particular, if σ_k satisfies (2.8), then it also achieves (2.7) with $\gamma_3 = \sigma_{\min}/\overline{\sigma}$, and we have that

$$j \leq \left\lceil 1 + \frac{2}{\log \gamma_1} \log \left(\frac{\overline{\sigma}}{\sigma_{\min}} \right) \right\rceil \cdot |\mathcal{S}_j|. \tag{2.11}$$

Proof. Apply [3, 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 (2.9) are provided in (2.16) below, and under stronger assumptions, in (3.6). (Note that due to Lemmas 2.4 and 2.6, the condition required for (2.16) 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 that 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 (1.2).

i) [2, Lemma 2.1] Let AM.1 hold. Then for $k \geq 0$, we have that

$$f(x_k) - m_k(s_k) \geq \frac{\|g_k\|}{6\sqrt{2}} \min \left[\frac{\|g_k\|}{\kappa_{\text{B}}}, \frac{1}{2} \sqrt{\frac{\|g_k\|}{\sigma_k}} \right], \tag{2.12}$$

and so Δ_k in (2.4) is monotonically decreasing,

$$\Delta_{k+1} \leq \Delta_k, \quad k \geq 0. \tag{2.13}$$

ii) [3, Lemma 3.2] Let AF.1, AF.4 and AM.1 hold. Also, assume that

$$\sqrt{\sigma_k \|g_k\|} > \frac{108\sqrt{2}}{1-\eta_2}(\kappa_{\rm H} + \kappa_{\rm B}) \stackrel{\rm def}{=} \kappa_{\rm HB}. \tag{2.14}$$

Then iteration k is very successful and

$$\sigma_{k+1} \le \sigma_k. \tag{2.15}$$

iii) [3, Lemma 3.3] Let AF.1, AF.4 and AM.1 hold. For any $\epsilon > 0$ and $j \ge 0$ such that $\|g_k\| > \epsilon$ for all $k \in \{0, \ldots, j\}$, we have

$$\sigma_k \le \max\left(\sigma_0, \frac{\gamma_2 \kappa_{\rm HB}^2}{\epsilon}\right), \ 0 \le k \le j. \tag{2.16}$$

A generic property follows.

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

$$\Delta_k \le \kappa_{\rm c} \|g_k\|^p, \ \text{ for all } k \ge 0, \tag{2.17}$$

for some $\kappa_{\rm c} > 0$ and $p > 0$, with Δ_k defined in (2.4). Then

$$f(x_k) - m_k(s_k) \ge \kappa_{\rm m} \Delta_k^{2/p}, \ \text{ for all } k \ge 0, \tag{2.18}$$

where $\kappa_{\rm HB}$ is defined in (2.14) and

$$\kappa_{\rm m} \stackrel{\rm def}{=} \frac{1}{12\sqrt{2}\kappa_{\rm c}^{2/p}} \min\left(\sqrt{\frac{\kappa_{\rm c}^{1/p}}{\sigma_0 \Delta_0^{1/p}}}, \frac{1}{\sqrt{\gamma_2}\kappa_{\rm HB}}\right). \tag{2.19}$$

Proof. We first show that

$$\sigma_k \Delta_k^{1/p} \le \max\left(\sigma_0 \Delta_0^{1/p}, \gamma_2 \kappa_{\rm c}^{1/p} \kappa_{\rm HB}^2\right), \ \text{ for all } k \ge 0. \tag{2.20}$$

For this, we use the implication

$$\sigma_k \Delta_k^{1/p} > \kappa_{\rm c}^{1/p} \kappa_{\rm HB}^2 \quad \Longrightarrow \quad \sigma_{k+1} \Delta_{k+1}^{1/p} \le \sigma_k \Delta_k^{1/p}, \tag{2.21}$$

which follows from (2.15) in Lemma 2.2 ii), (2.17) and (2.13). Thus, when $\sigma_0 \Delta_0^{1/p} \le \gamma_2 \kappa_{\rm c}^{1/p} \kappa_{\rm HB}^2$, (2.21) implies $\sigma_k \Delta_k^{1/p} \le \gamma_2 \kappa_{\rm c}^{1/p} \kappa_{\rm HB}^2$, where the factor γ_2 is introduced for the case when $\sigma_k \Delta_k^{1/p}$ is less than $\kappa_{\rm c}^{1/p} \kappa_{\rm HB}^2$ and the iteration k is not very successful.

Letting $k = 0$ in (2.21) gives the first inequality in (2.20) when $\sigma_0 \Delta_0^{1/p} \geq \gamma_2 \kappa_{\rm c}^{1/p} \kappa_{\rm HB}^2$, since $\gamma_2 > 1$. Next we deduce from (2.12) and (2.17) that

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

which together with (2.20) and the definition of $\kappa_{\rm HB}$, gives (2.18) and (2.19). □

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; see [10, Theorems 2.1.14, 2.1.15].

2.3 Basic ARC complexity on convex objectives

Let us now assume that

AF.7 $$f \text{ is convex,} \tag{2.22}$$

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

AF.8 $$\|x - x_*\| \leq D, \quad \text{for all } x \text{ such that } f(x) \leq f(x_0), \tag{2.23}$$

where x_* is any global minimizer of f and $D \geq 1$. The following property specifies the values of p and $\kappa_{\rm c}$ for which (2.17) 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 (2.4),

$$\Delta_k \leq D\|g_k\|, \quad \text{for all } k \geq 0. \tag{2.24}$$

Proof. AF.7 implies $f(x) - f(y) \geq g(y)^T(x - y)$, 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 (2.24). □

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

$$\Delta_j = f(x_j) - f_* \leq \frac{1}{|\mathcal{S}_j|\eta_1\kappa_{\rm m}^c}, \quad j \geq 0, \tag{2.25}$$

where $\mathcal{S}_j$ is defined in (2.6), and $\kappa_{\rm m}^c$ has the expression

$$\kappa_{\rm m}^c \stackrel{\rm def}{=} \frac{1}{12\sqrt{2}D^2} \min\left(\sqrt{\frac{D}{\sigma_0\Delta_0}}, \frac{1}{\sqrt{\gamma_2}\kappa_{\rm HB}}\right). \tag{2.26}$$

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

$$\left\lceil \frac{\kappa_{\rm s}^c}{\epsilon} \right\rceil \tag{2.27}$$

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

Proof. From (1.4) and (1.5), we have

$$f(x_k) - f(x_{k+1}) \geq \eta_1(f(x_k) - m_k(s_k)), \quad k \in \mathcal{S}. \tag{2.28}$$

Lemma 2.4 implies that the conditions of Lemma 2.3 are satisfied with $p = 1$ and $\kappa_{\rm c} = D$, and so (2.18) and (2.28) imply

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

where $\kappa_{\rm m}^c$ is defined in (2.26). Thus, recalling (2.4), we have

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

or equivalently,

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

where in the last inequality, we used (2.13). Since $\Delta_k = \Delta_{k+1}$ for any $k \notin \mathcal{S}$, summing up the above inequalities up to j gives

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

which gives (2.25), and hence, also (2.27). □

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

AF.9 $$f(y) \geq f(x) + g(x)^T(y-x) + \frac{\mu}{2}\|y-x\|^2, \quad \forall\, x,\, y \in \mathbb{R}^n. \tag{2.29}$$

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

The next property specifies the values of p and $\kappa_{\rm c}$ for which (2.17) 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

$$\Delta_k \leq \frac{1}{2\mu}\|g_k\|^2, \quad \text{for all } k \geq 0. \tag{2.30}$$

Proof. AF.9 implies $f(y) \leq f(x) + g(x)^T(y-x) + \frac{1}{2\mu}\|g(x)-g(y)\|^2$, for all $x, y \in \mathbb{R}^n$; see [10, Theorem 2.1.10] and its proof. Letting $x = x_*$ and $y = x_k$ in the latter gives (2.30). □

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

$$\Delta_j = f(x_j) - f_* \leq (1 - \eta_1 \kappa_{\rm m}^{sc})^{|\mathcal{S}_j|}\, \Delta_0, \quad j \geq 0, \tag{2.31}$$

where $\mathcal{S}_j$ is defined in (2.6), and $\kappa_{\rm m}^{sc}$ has the expression

$$\kappa_{\rm m}^{sc} \stackrel{\rm def}{=} \frac{\mu}{6\sqrt{2}} \min\left(\frac{1}{\sqrt{\sigma_0\sqrt{2\mu\Delta_0}}}, \frac{1}{\sqrt{\gamma_2}\kappa_{\rm HB}}\right) \in (0,1). \tag{2.32}$$

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

$$\left\lceil \kappa_{\mathrm{s}}^{sc} \log \frac{\Delta_0}{\epsilon} \right\rceil \tag{2.33}$$

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

Proof. Lemma 2.6 implies that (2.17) 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 (2.18), (2.19), (2.28) 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 (2.32), which immediately gives (2.31) 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 (2.30) and from (2.34) with $x = x_k$. The bound (2.31) and the inequality $(1 - \eta_1 \kappa_{\mathrm{m}}^{sc})^{|\mathcal{S}_j|} \leq e^{-\eta_1 \kappa_{\mathrm{m}}^{sc} |\mathcal{S}_j|}$ imply that $\Delta_j \leq \epsilon$ provided $e^{-\eta_1 \kappa_{\mathrm{m}}^{sc} |\mathcal{S}_j|} \Delta_0 \leq \epsilon$, which then gives (2.33) by applying the logarithm. □

Some remarks on basic ARC's complexity for (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 first-order 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 (strongly) convex objectives with Lipschitz continuous gradients given in [10, Corollary 2.1.2, Theorem 2.1.15]. Furthermore, assuming the exact Hessian is used in place of B_k and that all iterations are successful, the bound (2.25) is essentially a problem-independent-constant multiple of the one in [10, Corollary 2.1.2]. Under the same assumptions, there is a slightly weaker condition-number matching between our result for strongly convex functions and the one in [10]. Namely, (2.31) becomes a constant multiple of $1/c(H)$, where $c(H) = \kappa_{\mathrm{H}}/\mu$ which is a uniform bound on the Hessian's condition number, while [10, Theorem 2.1.15] depends on the smaller quantity $(c(H)-1)^2/(c(H)+1)^2$. Note that optimal bounds for steepest descent methods applied to (strongly) convex objectives with Lipschitz continous gradient are better than ours, namely, of order $\mathcal{O}(1/\sqrt{\epsilon})$ for the convex case [10, Theorems 2.1.7, §2.2.1], and for strongly convex, of order $\mathcal{O}(\log \epsilon^{-1})$ with convergence factor $(\sqrt{c(H)} - 1)^2/(\sqrt{c(H)} + 1)^2$ [10, Theorem 2.1.13; §2.2.1].

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 (1.7). This issue is important as the latter can be used as a termination condition for ARC, while Δ_k in (2.4), 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

$$f(x) - f_* \geq f(x) - f(x - \alpha g(x)) \geq \frac{1}{2\kappa_{\rm H}}\|g(x)\|^2, \quad \text{for all } \alpha \geq 0 \text{ and } x \in \mathbb{R}^n. \tag{2.34}$$

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

$$\Delta_k \geq \frac{1}{2\kappa_{\rm H}}\|g_k\|^2, \quad k \geq 0, \tag{2.35}$$

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

$$f(x_j) - f_* \leq \frac{\epsilon^2}{2\kappa_{\rm H}}. \tag{2.36}$$

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

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

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

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

The minimum of the right-hand side of the above inequality is attained at $\alpha_* = 1/\kappa_{\rm H}$, giving (2.34). □

Under the conditions of Theorem 2.5, ARC will take at most $\mathcal{O}(\epsilon^{-2})$ successful iterations to ensure (2.36) when applied to convex objectives. For strongly convex functions, Theorem 2.7 implies the same order of complexity of $|\log \epsilon|$ for $\|g_j\| \leq \epsilon$. (Note that the term $f(x_0) - f_*$ in (2.25) and (2.31) can be replaced by $D\|g_0\|$ and $\|g_0\|^2/(2\mu)$, respectively.)

Now recall [3, 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 $\|g_j\| \leq \epsilon$. 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 (2.35) that once $\|g_k\| \leq \epsilon$, 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 (1.10) and (1.11), namely of the $\text{ARC}_{(\text{S})}$ and $\text{ARC}_{(\text{g2})}$ variants. For the remainder of the paper, we assume that

AF.3 $$f \in C^2(\mathbb{R}^n). \tag{3.1}$$

While no assumption on the Hessian of f being globally or locally Lipschitz continuous has been imposed in the complexity results of §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 $$\|H(x) - H(x_k)\| \leq L\|x - x_k\|, \quad \text{for all } x \in [x_k, x_k + s_k] \text{ and all } k \geq 0, \tag{3.2}$$

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

AM.4 $$\|(H(x_k) - B_k)s_k\| \leq C\|s_k\|^2, \quad \text{for all } k \geq 0, \text{ and some constant } C > 0. \tag{3.3}$$

By using finite differences on the gradient for computing B_k, we showed in [5] 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 $\text{ARC}_{(\text{h})}$, in particular, necessary conditions for the global subproblem solution (1.8) and expressions for the model decrease (see Lemma 3 i)); also, some general properties that hold for a large class of (nonconvex) functions (see Lemma 3 ii) and iii)).

Lemma 3.1. i) [2, Lemmas 3.2, 3.3] Let s_k be the global minimizer of (1.8) for any $k \geq 0$. Then

$$g_k^\top s_k + s_k^\top B_k s_k + \sigma_k \|s_k\|^3 = 0, \tag{3.4}$$

and

$$f(x_k) - m_k(s_k) = \frac{1}{2} s_k^T B_k s_k + \frac{2}{3}\sigma_k \|s_k\|^3. \tag{3.5}$$

ii) [2, Lemma 5.2] Let AF.3, AF.6 and AM.4 hold. Then

$$\sigma_k \leq \max\left(\sigma_0, \tfrac{3}{2}\gamma_2(C+L)\right) \stackrel{\text{def}}{=} L_0, \quad \text{for all } k \geq 0. \tag{3.6}$$

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

$$\|s_k\| \geq \kappa_g \sqrt{\|g_{k+1}\|} \quad \text{for all successful iterations } k, \tag{3.7}$$

where κ_g is the positive constant

$$\kappa_g \stackrel{\text{def}}{=} \sqrt{(1-\kappa_\theta)/(L+C+L_0+\kappa_\theta\kappa_{\text{H}})}. \tag{3.8}$$

Note that in our second-order ARC variants in [2, 3], we employ the more general condition (3.4) and an approximate nonnegative curvature requirement [2, (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 (1.8). When the model is convex, as it is often the case here, such situations do not arise.

The bound (3.7) ensures that the step s_k does not become too small compared to the size of the gradient, and it is a crucial ingredient for obtaining, as shown in [3, Corollary 5.3], an $\mathcal{O}(\epsilon^{-3/2})$ upper bound on the iteration count of ARC$_{(\text{S})}$ 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$_{(\text{g2})}$ complexity on convex objectives

Here, we prove an $\mathcal{O}(1/\sqrt{\epsilon})$ iteration upper bound for ARC$_{(\text{g2})}$ to achieve (1.12), 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

$$\boxed{\textbf{AF.6}'} \qquad \|H(x) - H(y)\| \leq L\|x-y\|, \quad \text{for all } x,\, y \in \mathbb{R}^n. \tag{3.9}$$

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 $\mathrm{ARC}_{(\mathrm{g2})}$,

$$B_k = H(x_k), \quad \text{for all } k \geq 0. \tag{3.10}$$

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 (1.8) 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

$$\min_{s \in \mathcal{L}_k} m_k(s) \leq f(x_k) - 2\kappa^c_{\mathrm{m(g2)}}[f(x_k) - f(x_k + s_k^*)]^{\frac{3}{2}}, \tag{3.11}$$

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

$$\kappa^c_{\mathrm{m(g2)}} \stackrel{\mathrm{def}}{=} \left(6D\sqrt{6DL_1}\right)^{-1} \quad \text{and} \quad L_1 \stackrel{\mathrm{def}}{=} \max(\sigma_0, \gamma_2 L, \kappa_{\mathrm{H}}). \tag{3.12}$$

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

$$\left| f(x_k + s) - f(x_k) - s^T g_k - \tfrac{1}{2} s^T H(x_k) s \right| \leq \frac{L}{6}\|s\|^3, \quad s \in \mathbb{R}^n, \tag{3.13}$$

and so, from (1.1) 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 (3.6) and $\gamma_2 \geq 1$, we further obtain

$$m_k(s) \leq f(x_k + s) + L_1\|s\|^3, \quad s \in \mathbb{R}^n, \tag{3.14}$$

where L_1 is defined in (3.12). (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 (3.14) gives the first inequality below

$$\min_{s \in \mathcal{L}_k} m_k(s) \leq \min_{s \in \mathcal{L}_k} \left\{ f(x_k + s) + L_1\|s\|^3 \right\} \leq \min_{\alpha \in [0,1]} \left\{ f(x_k + \alpha s_k^*) + L_1 \alpha^3 \|s_k^*\|^3 \right\}, \tag{3.15}$$

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 (3.15),

$$\min_{s \in \mathcal{L}_k} m_k(s) \leq f(x_k) + \min_{\alpha \in [0,1]} \left\{ \alpha[f(x_k + s_k^*) - f(x_k)] + L_1 \alpha^3 \|s_k^*\|^3 \right\}. \tag{3.16}$$

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 (3.16) implies

$$\min_{s \in \mathcal{L}_k} m_k(s) \leq f(x_k) + \min_{\alpha \in [0,1]} \left\{ \alpha [f(x_k + s_k^*) - f(x_k)] + 8\alpha^3 L_1 D^3 \right\}. \tag{3.17}$$

The minimum in the right-hand side of (3.17) is attained at

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

Let us show that $\hat{\alpha}_k \leq 1$, namely, $f(x_k) - f(x_k + s_k^*) \leq 24 L_1 D^3$. AF.7 gives the first inequality

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

where we also used the Cauchy-Schwarz inequality, the bound on s_k^* just before (3.17), AF.4 and AF.8. Since we assumed in AF.8 that $D \geq 1$, and the definition of L_1 implies $L_1 \leq \kappa_{\rm H}$, we conclude that $f(x_k + s_k^*) - f(x_k) \geq -2\kappa_{\rm H} D^3 \geq -2L_1 D^3 \geq -24 L_1 D^3$. Thus, $\alpha_k^* = \hat{\alpha}_k$ and substituting the above value of $\hat{\alpha}_k$ in (3.17), we deduce (3.11) with the notation (3.12). □

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$_{\rm (g2)}$ with the choices (2.8) and (3.10) to minimizing f. Then

$$\Delta_j = f(x_j) - f_* \leq \frac{1}{(|\mathcal{S}_j| \eta_1 \beta \kappa_{\rm m(g2)}^c)^2}, \quad j \geq 0, \tag{3.18}$$

where $\mathcal{S}_j$ is defined in (2.6), $\kappa_{\rm m(g2)}^c$ in (3.12) and

$$\beta \stackrel{\rm def}{=} \frac{1}{2} \min\left(1, \frac{\kappa_{\rm g}^{3/2}}{4(\kappa_{\rm H} D)^{3/2}}\right) \quad \text{with} \quad \kappa_{\rm g} \stackrel{\rm def}{=} \frac{\sigma_{\min}(\kappa_{\rm m(g2)}^c)^2}{4\kappa_\theta^2 \kappa_{\rm H}^3}. \tag{3.19}$$

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

$$\left\lceil \frac{\kappa_{\rm s(g2)}^c}{\sqrt{\epsilon}} \right\rceil \tag{3.20}$$

successful iterations and gradient evaluations to generate $f(x_j) - f_* \leq \epsilon$, where $\kappa_{\rm s(g2)}^c \stackrel{\rm def}{=} (\eta_1 \beta \kappa_{\rm m(g2)}^c)^{-1}$.

Proof. Let $k \in \mathcal{S}$. From (1.4), (1.5) and (2.5), we have

$$f(x_{k+1}) \leq (1-\eta_1)f(x_k) + \eta_1 m_k(s_k) = (1-\eta_1)f(x_k) + \eta_1[m_k(s_k) - m_k(s_k^m)] + \eta_1 m_k(s_k^m), \tag{3.21}$$

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

$$m_k(s_k) - m_k(s_k^m) \leq \nabla_s m_k(s_k)^T(s_k - s_k^m) \leq \|\nabla_s m_k(s_k)\| \cdot \|s_k - s_k^m\| \leq \kappa_\theta \|g_k\|^3 \cdot \|s_k - s_k^m\|. \tag{3.22}$$

where the second inequality follows from TC.g2 (1.11). To bound $\|s_k - s_k^m\|$, recall that both s_k and s_k^m satisfy (3.4), which implies due to (2.8) and $B_k = H(x_k)$ being positive semidefinite,

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

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 (3.22) now provide the first inequality below,

$$m_k(s_k) - m_k(s_k^m) \leq \frac{2\kappa_\theta}{\sqrt{\sigma_{\min}}}\|g_k\|^{\frac{7}{2}} \leq \frac{2\kappa_\theta\kappa_{\rm H}\sqrt{2\kappa_{\rm H}}}{\sqrt{\sigma_{\min}}}\sqrt{\|g_k\|} \cdot \Delta_k^{\frac{3}{2}}, \tag{3.23}$$

while the second inequality follows from (2.35). Recalling (3.21), 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 (3.11) implies

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

Substituting this bound and (3.23) into (3.21), we deduce

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

or equivalently, recalling (2.4) and (3.19),

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

Thus we have the implication

$$\|g_k\| \leq \frac{\kappa_{\rm g}}{2} \quad \Longrightarrow \quad \Delta_k - \Delta_{k+1} \geq \eta_1\kappa^c_{\rm m(g2)}\Delta_k^{\frac{3}{2}}. \tag{3.24}$$

It remains to prove a bound of the same form as the right-hand side of (3.24) when $\|g_k\| > \kappa_{\rm g}/2$. 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 (3.11) is equal to $m_k(s_k)$ in this case, we employ (3.11) to upper bound the first inequality in (3.21), and obtain

$$f(x_{k+1}) \leq f(x_k) - 2\eta_1 \kappa^c_{\mathrm{m(g2)}} [f(x_k) - f(x_k + s^*_k)]^{\frac{3}{2}}. \tag{3.25}$$

Since s^*_k is a global minimizer of $f(x_k + s)$ over $s \in \mathcal{L}_k$, and $g \in \mathcal{L}_k$, 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 (2.34) and from (2.24), respectively. It follows from (3.25) that

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

or equivalently,

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

Thus we have the implication

$$\|g_k\| > \frac{\kappa_{\mathrm{g}}}{2} \quad \Longrightarrow \quad \Delta_k - \Delta_{k+1} \geq \eta_1 \kappa^c_{\mathrm{m(g2)}} \frac{\kappa_{\mathrm{g}} \sqrt{\kappa_{\mathrm{g}}}}{4\kappa_{\mathrm{H}} D \sqrt{\kappa_{\mathrm{H}} D}} \Delta_k^{\frac{3}{2}}. \tag{3.26}$$

Finally, we conclude from (3.24) and (3.26) that

$$\Delta_k - \Delta_{k+1} \geq 2\eta_1 \beta \kappa^c_{\mathrm{m(g2)}} \Delta_k^{\frac{3}{2}}, \quad k \in \mathcal{S}, \tag{3.27}$$

where β is defined in (3.19). 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 (3.27) and (2.13), respectively. Thus, recalling that Δ_k remains unchanged on unsuccessful iterations and summing the above up to j, we deduce

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

which gives (3.18) and also (3.20). □

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-size or the Hessian at the iterates is sparse.

Note the two regimes of analysis in the above proof, namely in the model decreases (3.24) and (3.26). 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 $\text{ARC}_{(\text{g2})}$ 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 (3.26) also holds for $\text{ARC}_{(\text{S})}$, but the termination condition TC.s does not seem strong enough to ensure a similar property to (3.24) for the asymptotic regime of $\text{ARC}_{(\text{S})}$.

The iteration complexity of Nesterov & Polyak's cubic regularization algorithm applied to convex problems is analysed in [12, Theorem 4] and [11, Theorem 1], and an $\mathcal{O}(1/\sqrt{\epsilon})$ bound is obtained. Note however, that for both these results, the Hessian's Lipschitz constant is assumed to be available for the running of the algorithm; see [12, (3.3)] and [11, (3.11)].

Complexity of generating approximately-optimal gradient values The complexity of $\text{ARC}_{(\text{g2})}$ generating a gradient value $\|g_j\| \leq \epsilon$ can be obtained as described in Section 2.5, by using (2.36) 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 $\text{ARC}_{(\text{S})}$ complexity on strongly convex objectives

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

Let us now assume AF.9. Due to AF.3, (2.29) is equivalent to

$$u^T H(x) u \geq \mu \|u\|^2, \qquad \text{for all } u,\, x \in \mathbb{R}^n. \tag{3.28}$$

Employing (2.29) with $y = x$ and $x = x_*$, we deduce that AF.8 is implied by AF.9 with

$$D \leq \sqrt{2\Delta_0/\mu}. \tag{3.29}$$

The strong convexity of f implies that asymptotically, $\text{ARC}_{(\text{S})}$ 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$_{(S)}$ to minimizing f, and assume that the Rayleigh quotient of B_k along s_k is uniformly bounded away from zero, namely

$$R_k(s_k) \stackrel{\text{def}}{=} \frac{s_k^T B_k s_k}{\|s_k\|^2} \geq R_{\min} > 0, \quad \forall\, k \in \mathcal{S}. \tag{3.30}$$

Then, recalling κ_g defined in (3.8) and letting $\delta \stackrel{\text{def}}{=} \frac{1}{2}(\eta_1 R_{\min}\kappa_g^2\sqrt{\mu})^2$,

$$\mathcal{N}_f \stackrel{\text{def}}{=} \{x : \, f(x) - f(x_*) \leq \delta\} \tag{3.31}$$

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

$$\Delta_{k+1} \leq \frac{1}{\delta}\Delta_k^2, \quad \text{for all } k \in \mathcal{S} \text{ and } k \geq j_q. \tag{3.32}$$

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

$$\left\lceil \log_2 \log_2 \left(\frac{\delta}{\epsilon}\right) \right\rceil \tag{3.33}$$

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

Proof. Let $k \in \mathcal{S}$. Then (1.5), (3.5), (3.30) and (3.7) 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

$$\Delta_{k+1} \leq \frac{1}{\delta}\Delta_k^2, \quad \text{for all } k \in \mathcal{S}, \tag{3.34}$$

where δ is defined in (3.31). Thus the expression of $\mathcal{N}_f$ in (3.31) follows, as well as (3.32). Assuming that $x_{j_q} \in \mathcal{N}_f$ with $\Delta_{j_q} \leq \delta/2$, we deduce from (3.32) that

$$\Delta_j \leq \delta^{1-2^l}\Delta_{j_q}^{2^l}, \quad \text{for any } j \geq j_q, \tag{3.35}$$

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 (3.35) shows that $\Delta_j \leq \epsilon$ provided $2^{-2^l}\delta \leq \epsilon$, which gives the bound (3.33). □

Remark on satisfying (3.30). If exact Hessians are used so that $B_k = H(x_k)$ for all k, then AF.9 implies (3.30) due to (3.28). Alternatively, (3.30) can be ensured if AM.4 holds with a sufficiently small C. Namely, note that AF.9, AM.4 and (3.29) imply

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

Thus (3.30) holds provided $C < \mu/(2D)$. Recall our comments on satisfying AM.4 by finite differencing following (3.3). □

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

$$\Delta_{j_q} = f(x_{j_q}) - f(x_*) \leq \tfrac{1}{2}\delta, \tag{3.36}$$

with δ defined in (3.31). The choice of s_k in (1.8) with $g_k \in \mathcal{L}_k$ implies that ARC$_{(S)}$ always satisfies the Cauchy condition (1.2) 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 (3.30) holds. Then, given any $\epsilon > 0$, ARC$_{(S)}$ takes, in total, at most

$$\left\lceil \kappa_s^{sc} \log \frac{2\Delta_0}{\delta} + \log_2 \log_2 \left(\frac{\delta}{\epsilon}\right) \right\rceil \tag{3.37}$$

successful iterations and gradient evaluations to generate $f(x_j) - f(x_*) \leq \epsilon$, where κ_s^{sc} is defined in (2.33) and δ in (3.31).

Proof. The conditions of Theorem 2.7 are satisfied, and so letting $\epsilon = \delta/2$ in (2.33), we deduce that (3.36) 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. □

Note that for the non-asymptotic phase of $\mathrm{ARC}_{(\mathrm{S})}$, an $\mathcal{O}(1/\sqrt{\delta})$ bound can be deduced similarly to the proof of Theorem 3.3. Namely, using Lemma 3.2, which clearly holds for $\mathrm{ARC}_{(\mathrm{S})}$, we deduce (3.25); then employ (2.34) just as in the first displayed equation after (3.25) and use (2.30). Then the total $\mathrm{ARC}_{(\mathrm{S})}$ 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 [12, pages 203–204] and [11, pages 176–177]. Note that such bounds are weaker than the ones we obtained in Corollary 3.5.

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 $\mathrm{ARC}_{(\mathrm{S})}$ to minimizing f, assuming that (3.30) holds. Then $\mathcal{N}_g \stackrel{\mathrm{def}}{=} \{x : \|g(x)\| \leq (\frac{1}{2}\eta_1 R_{\min}\kappa_g)^2 \stackrel{\mathrm{def}}{=} \zeta\}$ 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

$$\|g_{k+1}\| \leq \frac{1}{\zeta}\|g_k\|^2, \quad \text{for all } k \in \mathcal{S} \text{ and } k \geq j_q. \tag{3.38}$$

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

$$\left\lceil \log_2 \log_2 \left(\frac{\zeta}{\epsilon}\right) \right\rceil \tag{3.39}$$

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

$$\left\lceil 2\kappa_s^{sc} \log \frac{\|g_0\|\sqrt{\kappa_{\mathrm{H}}}}{\zeta\sqrt{\mu}} \right\rceil \tag{3.40}$$

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 (3.39) and (3.40).

Proof. AF.9 implies AF.7 which gives

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

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

$$\|g_k\| \geq \tfrac{1}{2}\eta_1 R_{\min}\|s_k\| \geq \tfrac{1}{2}\eta_1 R_{\min}\kappa_{\mathrm{g}}\sqrt{\|g_{k+1}\|}, \quad k \in \mathcal{S}, \tag{3.41}$$

where the latter inequality follows from (3.7). The expression and properties of $\mathcal{N}_g$ follow. The bound (3.39) is obtained similarly to the proof of (3.33) in Lemma 3.4.

To deduce (3.40), let $\epsilon = \zeta$ in (2.36) and in (2.33), 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 [11] for Nesterov & 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

$$f(x) = e^{-x}, \quad \text{for } x \geq 0. \tag{3.42}$$

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

Lemma 3.7. The function (3.42) 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 $\kappa_{\text{H}} = L = 1$, thus satisfying AF.4, AF.6′ and AF.7.
Apply ARC to minimizing (3.42), 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 (1.1) with $B_k = f''(x_k)$ and with the (reasonable) choice

$$\sigma_k := \sigma \geq \frac{L}{2} = \frac{1}{2}, \quad \forall k \geq 0, \tag{3.43}$$

which ensures that every iteration is very successful and that (2.8) holds. Then AM.1 and AM.4 hold (with $\kappa_{\text{B}} = 1$ 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 & Polyak's cubic regularization algorithm [12, 11]; recall our choice of s_k and σ_k in the above. In particular, it satisfies all the

conditions in [11, 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 & 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 3.1).

- If Newton's method is applied to this example, the complexity would be better; see Figure 3.1. 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 [4] 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.

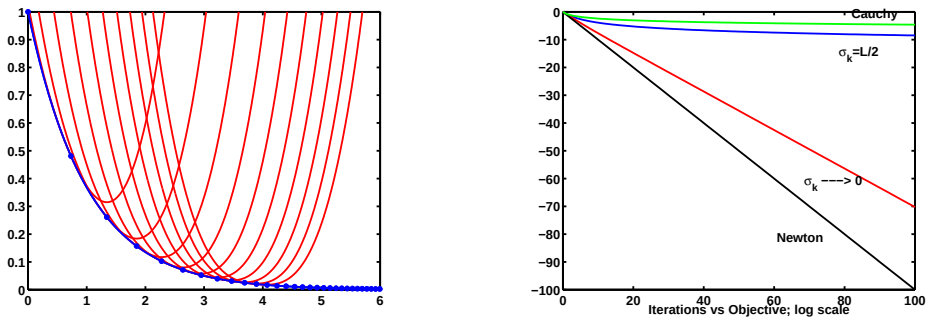

Figure 3.1: Graph of (3.42) and the local cubic regularizations at the ARC iterates (left-hand side). Plot of objective values at the iterates on a log scale for different ARC variants and for Newton's method (right-hand side).

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 $1/\sqrt{\epsilon}$ and $\log \kappa + \log |\log \epsilon|$ 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.

References

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

[2] C. Cartis, N. I. M. Gould and Ph. L. Toint. Adaptive cubic regularisation methods for unconstrained optimization. Part I: motivation, convergence and numerical results. *Mathematical Programming*, DOI: 10.1007/s10107-009-0286-5, 2009 (online).

[3] 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. *Mathematical Programming*, DOI: 10.1007/s10107-009-0337-y, 2010 (online).

[4] 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 Journal on Optimization*, 20(6): 2833–2852, 2010.

[5] 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 Technical Report 10-005, School of Mathematics, University of Edinburgh, 2010.

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

[7] J. E. Dennis and R. B. Schnabel. *Numerical methods for unconstrained optimization and nonlinear equations.* Prentice-Hall, Englewood Cliffs, New Jersey, USA, 1983. Reprinted as *Classics in Applied Mathematics 16*, SIAM, Philadelphia, USA, 1996.

[8] G. H. Golub and C. F. Van Loan. *Matrix Computations.* The John Hopkins University Press, Baltimore, USA, 1996.

[9] A. Griewank. The modification of Newton's method for unconstrained optimization by bounding cubic terms. Technical Report NA/12 (1981), Department of Applied Mathematics and Theoretical Physics, University of Cambridge, United Kingdom, 1981.

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

[11] Yu. Nesterov. Accelerating the cubic regularization of Newton's method on convex problems. *Mathematical Programming*, 112(1):159–181, 2008.

[12] Yu. Nesterov and B. T. Polyak. Cubic regularization of Newton's method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

[13] M. Weiser, P. Deuflhard and B. Erdmann. Affine conjugate adaptive Newton methods for nonlinear elastomechanics. *Optimization Methods and Software*, 22(3):413–431, 2007.

Appendix A

Proof of Lemma 3.7. Apply ARC to minimizing (3.42), 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

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

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

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

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

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

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

$$x_k \geq 0, \quad \forall\, k \geq 0; \tag{A.3}$$

so we only need to consider $f(x)$ for $x \geq 0$, which clearly satisfies AF.4 with $\kappa_{\text{H}} = 1$, AM.1 with $\kappa_{\text{B}} = 1$, AF.6′ with $L = 1$, AM.4 with $C = 0$, and AF.7. Furthermore, AF.6′, (1.6) and (3.13) provide the implication

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

This and (3.43) imply that all iterations k are very successful and that the iterates satisfy $x_{k+1} = x_k + s_k$, with s_k in (A.2), for all $k \geq 0$. Furthermore, (3.42) and $e^{-x} \in (0, 1]$ for $x \geq 0$, provide the following same-order bounds on s_k in (A.2)

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

which further become, by letting

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

$$c_1 e^{-\frac{1}{2}x_k} \geq s_k \geq c_2 e^{-\frac{1}{2}x_k}, \quad \forall\, k \geq 0. \tag{A.4}$$

From (3.42), we have

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

which further gives, by employing (A.4),

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

Employing again (3.42), we obtain

$$f(x_k) e^{-c_1\sqrt{f(x_k)}} \leq f(x_{k+1}) \leq f(x_k) e^{-c_2\sqrt{f(x_k)}}, \quad k \geq 0. \tag{A.5}$$

Since the following bounds hold for the exponential function

$$1 - y \leq e^{-y} \leq 1 - y + \frac{y^2}{2}, \quad y \in [0,1], \tag{A.6}$$

it follows from (A.5), (A.6) and $f_k = f(x_k) \in (0,1]$ that

$$f_k(1 - c_1\sqrt{f_k}) \leq f_{k+1} \leq f_k\left(1 - c_2\sqrt{f_k} + \frac{c_2^2}{2} f_k\right), \quad k \geq 0, \tag{A.7}$$

and so

$$c_2 f_k \sqrt{f_k}\left(1 - \frac{c_2}{2}\sqrt{f_k}\right) \leq f_k - f_{k+1} \leq c_1 f_k \sqrt{f_k}, \quad k \geq 0. \tag{A.8}$$

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

$$c_3 f_k \sqrt{f_k} \leq f_k - f_{k+1} \leq c_1 f_k \sqrt{f_k}, \quad k \geq 0, \tag{A.9}$$

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 (A.9) 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

$$\frac{c_3}{2} \leq \frac{1}{\sqrt{f_{k+1}}} - \frac{1}{\sqrt{f_k}} \leq \frac{c_1}{2} \cdot \frac{f_k}{f_{k+1}}, \quad k \geq 0. \tag{A.10}$$

Now let us give an upper bound on f_k/f_{k+1}. Using (A.5) and $f_k \in (0,1]$, we deduce

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

Thus (A.10) gives

$$\frac{c_3}{2} \leq \frac{1}{\sqrt{f_{k+1}}} - \frac{1}{\sqrt{f_k}} \leq \frac{c_1}{2}e^{c_1}, \quad k \geq 0. \tag{A.11}$$

Summing up (A.11) over $i \in \{0, \ldots, k\}$, we obtain

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

and thus,

$$k\frac{c_3}{2} \leq \frac{1}{\sqrt{f_k}} \leq k \max\left(c_1 e^{c_1}, 2e^{\frac{1}{2}x_0}\right), \quad k \geq 0. \tag{A.12}$$

Finally, (A.12) is equivalent to

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

which gives the desired complexity result of the Lemma. □