Examples of slow convergence for adaptive regularization optimization methods are not isolated

Philippe L. Toint*

24 IX 2024

Abstract

The adaptive regularization algorithm for unconstrained nonconvex optimization was shown in [20, 7] to require, under standard assumptions, at most $\mathcal{O}(\epsilon^{3/(3-q)})$ evaluations of the objective function and its derivatives of degrees one and two to produce an ϵ-approximate critical point of order $q \in \{1, 2\}$. This bound was shown to be sharp in [6, 5] for $q = 1$ and in [11] for arbitrary $q \in \{1, 2\}$. This note revisits these results and shows that the example for which slow convergence is exhibited is not isolated, but that this behaviour occurs for a subset of univariate functions of nonzero measure.

Keywords: complexity theory, adaptive regularization methods, approximate criticality, nonconvex optimization.

1 Introduction

Adaptive regularization algorithms for unconstrained nonconvex minimization have been extensively studied in recent years, focusing first on the case where a cubically regularized quadratic model is used (see [18, 20, 21, 14, 7, 13, 8, 9, 3, 15, 19, 2, 1, 10], for instance) and then extending the concept to methods using models of arbitrary degree [4, 12]. We refer the reader to [11] for an extensive coverage of this class of algorithms. A remarkable feature of methods in this class is their optimal worst-case evaluation complexity: it can indeed be shown that an ϵ-approximate critical point of order q must be obtained by an adaptive regularization algorithm using models of degree $p \geq q$ in at most $\mathcal{O}(\epsilon^{(p+1)/(p-q+1)})$ evaluations of the objective function and its derivatives of degrees 1 to p (see [4] for $q = 1$ and [11] for arbitrary $q \leq p$). This complexity bound was shown to be sharp in [6] for the case where $p = 2$ and $q = 1$, and this result was extended in [11] to arbitrary p and $q \leq p$ and, independently, in [5] for the special case where $q = 1$. These sharpness results all hinge on exhibiting of a typically quite contrived function on which the algorithm under consideration uses exactly as many evaluations as allowed by the complexity bound to achieve approximate criticality. Because of this contrived nature, one is naturally led to the question(1) of how "exceptional" these examples are. The purpose of the present note is to show that, albeit not common, such examples are not isolated, but rather form a set of nonzero measure. To do so, we focus on the case where $p = 2$.

*naXys, University of Namur, Namur, Belgium. Email: philippe.toint@unamur.be

(1)Adressed to the author in several occasions.

To reach this conclusion, we first need to set the stage in Section 2 where we briefly recall the approximate criticality measures, restate the AR2 regularization algorithm and recall the relevant worst-case complexity result. We next revisit, in Section 3, the example of slow convergence proposed in [6] and modify it to clarify the amount of freedom available in its construction. A brief discussion and conclusion is finally proposed in Section 4.

2 The context

In what follows, we consider the unconstrained minimization problem

$$\min_{x \in \mathbb{R}^n} f(x) \tag{2.1}$$

where f is twice times continously differentiable function from $\mathbb{R}^n$ into $\mathbb{R}$, with globally Lipschitz second derivative, meaning that there exist a constants $L_j \geq 0$ such that

$$\|\nabla_x^2 f(x) - \nabla_x^2 f(y)\| \leq L\|x-y\|,$$

for all $x, y \in \mathbb{R}^n$. To make the problem well-defined, we also assume that f is bounded below. We denote by $T_{f,2}(x,s)$ the second order Taylor's series of such a function computed at x for a step s, that is

$$T_{f,1}(x,s) = f(x) + \nabla_x^1 f(x)^T s \quad \text{and} \quad T_{f,2}(x,s) = f(x) + \nabla_x^1 f(x)^T s + \tfrac{1}{2} s^T \nabla_x^2 f(x) s.$$

For $j \in \{1,2\}$, we next define the *j-th order criticality measure*

$$\phi_{f,j}(x) = f(x) - \min_{\|d\| \leq 1} T_{f,j}(x,d), \tag{2.2}$$

that is the *largest decrease of the j-th order Taylor's series of f at x that is achievable in an Euclidean neighbouhood of radius one centered at x.* The measure $\phi_{f,j}(x)$ is a continuous function of x for $j = 1,2$ and we note that it is also independent of $f(x)$. Following [11, Section 12.1.5], we use this measure to express the notion of approximate first- and second-order criticality and say that, given $\epsilon = (\epsilon_1, \ldots, \epsilon_q)$, the point $x \in \mathbb{R}^n$ is ϵ,-approximate critical of order q (with $q \in \{1,2\}$), whenever

$$\phi_{f,j}(x) \leq \frac{\epsilon_j}{j} \quad \text{for} \quad j \in \{1, \ldots, q\}.$$

Theorem 12.1.3 in [11] describes the (natural) way in which $\phi_{f,j}(x)$ relates to the familiar criticality measures for the values $j = 1$ and $q = 2$ of interest here. In particular, we have that, for all $x \in \mathbb{R}^n$, $\phi_{f,1}(x) = \|\nabla_x^1 f(x)\|$ and $\phi_{f,2}(x) = |\min[0, \lambda_{\min}[\nabla_x^2 f(x)]|$ whenever $\phi_{f,1}(x) = 0$, where $\lambda_{\min}[H]$ denotes the leftmost eigenvalue of the symmetric matrix H. Theorem 12.1.8 in the same reference also gives a lower bound on the values of (the derivatives of) f in a neighbourhood of x.

Having clarified our objective (finding an approximate q-order critical point of f), we recall the definition of the AR2 algorithm, the relevant generalized adaptive regularization minimization method. As alluded to in the introduction, this algorithm proceeds, at each iterate x_k, by minimizing, the regularized quadratic model given by

$$m_k(s) = T_{f,2}(x_k, s) + \frac{\sigma_k}{6}|s|^3. \tag{2.3}$$

Note that $m_k(s)$ is bounded below, which makes it minimization in $\mathbb{R}^n$ well-defined. The minimization of m_k need not be exact (although we will not use this freedom in what follows) but can be terminated as soon as the criticality measures $\phi_{m_k,j}$ *associated with the model* are small enough or the step is large enough (see [11] or [17] for details).

The AR2 algorithm (with exact step) is then specified as follows.

Algorithm 2.1: AR2 for ϵ-approximate q-th-order minimizers

Step 0: Initialization. A criticality order $q \in \{1,2\}$, an initial point x_0 and an initial regularization parameter $\sigma_0 > 0$ are given, as well as accuracy levels $\epsilon_1 \in (0,1]$ and $\epsilon_2 \in (0,1]$ The constants θ, η_1, η_2, γ_1, γ_2, γ_3 and $\sigma_{\min}$ are also given, and satisfy

$$\theta \in (0,1), \quad \sigma_{\min} \in (0,\sigma_0], \quad 0 < \eta_1 \le \eta_2 < 1 \;\text{ and }\; 0 < \gamma_1 < 1 < \gamma_2 < \gamma_3.$$

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

Step 1: Test for termination. For $j = 1, \ldots, q$,

1. Evaluate $\nabla_x^j f(x_k)$ and compute $\phi_{f,j}(x_k)$.
2. If
$$\phi_{f,j}(x_k) \ge \frac{\epsilon_j}{j},$$
go to Step 2.

Terminate with the approximate solution $x_\epsilon = x_k$.

Step 2: Step calculation. If not available, evaluate $\nabla_x^2 f(x_k)$ and compute a global minimizer s_k of the model $m_k(s)$ given by (2.3).

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

$$\rho_k = \frac{f(x_k) - f(x_k+s_k)}{T_{f,2}(x_k,0) - T_{f,2}(x_k,s_k)}. \tag{2.4}$$

If $\rho_k \ge \eta_1$, then define $x_{k+1} = x_k + s_k$. Otherwise define $x_{k+1} = x_k$.

Step 4: Regularization parameter update. Set

$$\sigma_{k+1} \in \left\{ \begin{array}{ll} [\max(\sigma_{\min}, \gamma_1\sigma_k), \sigma_k] & \text{if } \rho_k \ge \eta_2, \\ [\sigma_k, \gamma_2\sigma_k] & \text{if } \rho_k \in [\eta_1,\eta_2), \\ [\gamma_2\sigma_k, \gamma_3\sigma_k] & \text{if } \rho_k < \eta_1. \end{array} \right. \tag{2.5}$$

Increment k by one and go to Step 1 if $\rho_k \ge \eta_1$, or to Step 2 otherwise.

The evaluation complexity of this algorithm is stated in [11, Theorem 12.2.14], which we summarize as follows.

Theorem 2.1 Under the assumptions on f stated at the beginning of this section and given a criticality order $q \in \{1,2\}$, the AR2 algorithm requires at most

$$\mathcal{O}\left(\max_{j\in\{1,\ldots,q\}} \epsilon_j^{-\frac{3}{3-j}}\right) \tag{2.6}$$

evaluations of f, and its derivatives of orders one and two to produce an iterate x_ϵ such that $\phi_{f,j}(x_\epsilon) \le \epsilon_j/j$ for $j \in \{1,2\}$.

3 The example

We now describe a one-dimensional example for which the convergence of the AR2 algorithm is "slow". Given $q \in \{1,2\}$ and $\epsilon_1 = \epsilon_2 = \epsilon \in (0, \frac{1}{4}]$, we wish to construct a function from $\mathbb{R}$ to $\mathbb{R}$ such that the AR2 algorithm, when applied to minimize this function, will take exactly

$$k_\epsilon = \left\lceil \epsilon^{-\frac{3}{3-q}} \right\rceil \tag{3.7}$$

iterations and evaluations of the objective function and its first and second derivatives to achieve ϵ-approximate criticality of order q. This function is constructed by first defining sequences of iterates $\{x_k\}_{k=0}^{k_\epsilon}$, function and derivatives values $\{\{f_k^{(j)}\}_{j=0}^2\}_{k=0}^{k_\epsilon}$ and associated "slowly converging" criticality measures $\{\{\phi_{f,j}(x_k)\}_{j=1}^2\}_{k=0}^{k_\epsilon}$. We then show that these sequences can be viewed as being generated by the AR2 algorithm applied to a piecewise polynomial function from $\mathbb{R}$ to $\mathbb{R}$ interpolating the values $f_k^{(j)}$. The general idea is to consider an example in the spirit of [6], but allowing additional perturbations to the function's and derivative's values in a way controlled to maintain the example's mechanism. Note that ensuring termination of the algorithm at iteration k_ϵ requires

$$\phi_{f,1}(x_k) > \epsilon \;\text{ for }\; k \in \{0,\ldots,k_\epsilon - 1\} \;\text{ and }\; \phi_{f,1}(x_{k_\epsilon}) \le \epsilon. \tag{3.8}$$

if $q = 1$. If $q = 2$, one instead needs that

$$\text{either } \phi_{f,1}(x_k) > \epsilon \quad\text{or}\quad \phi_{f,2}(x_k) \ge \frac{\epsilon}{2} \;\text{ for }\; k \in \{0,\ldots,k_\epsilon - 1\} \tag{3.9}$$

and

$$\phi_{f,1}(x_{k_\epsilon}) \le \epsilon \quad\text{and}\quad \phi_{f,2}(x_{k_\epsilon}) \le \frac{\epsilon}{2}. \tag{3.10}$$

We first set $p \stackrel{\rm def}{=} 3 - q$ (so that $\{p,q\} = \{1,2\}$) and define the sequences of derivatives values for $j \in \{1,2\}$ and $k \in \{0,\ldots,k_\epsilon\}$ by

$$f_k^{(j)} = \left\{\begin{array}{ll} -\beta_{q,k}\alpha_k\epsilon - \alpha_k\epsilon & \text{for } j = q, \\ -\beta_{p,k}[\alpha_k\epsilon]^{\frac{q}{p}} & \text{for } j = p, \end{array}\right. . \tag{3.11}$$

The sets $\mathcal{A} = \{(\alpha_k - 1)\}_{k=0}^{k_\epsilon}$ and $\mathcal{B} = \{\{\beta_{j,k}\}_{j=1}^2\}_{k=0}^{k_\epsilon}$ define the perturbations of the derivative's values. That they are acceptable for the example depends crucially on the conditions

$$\alpha_k \in [1,2] \;\text{ for }\; k \in \{0,\ldots,k_\epsilon - 1\} \;\text{ and }\; \alpha_{k_\epsilon} = 0 \tag{3.12}$$

and

$$\beta_{q,k} \in [0, \tfrac{1}{2}], \quad \beta_{p,k} = -\beta_{q,k} \;\text{ for }\; k \in \{0, \ldots, k_\epsilon - 1\} \;\text{ and }\; \beta_{q,k_\epsilon} = \beta_{p,k_\epsilon} = 0. \tag{3.13}$$

Given (3.11), one verifies that, for sequences $\{x_k\}$ and $\{f_k^{(0)}\}$ yet to be defined,

$$T_{f,2}(x_k, s) = f_k^{(0)} - \alpha_k \epsilon \frac{s^q}{q} - \beta_{q,k}[\alpha_k \epsilon]\, \frac{s^q}{q} - \beta_{p,k}[\alpha_k \epsilon]^{\frac{q}{p}}\, \frac{s^p}{p} \tag{3.14}$$

so that the model (2.3) is given by

$$m_k(s) = f_k^{(0)} - \alpha_k \epsilon \frac{s^q}{q} - \beta_{q,k}[\alpha_k \epsilon] \frac{s^q}{q} - \beta_{p,k}[\alpha_k \epsilon]^{\frac{q}{p}} \frac{s^p}{p} + \frac{\sigma_k}{6}|s|^3. \tag{3.15}$$

We furthermore choose (as we show below is possible) a constant regularization parameter

$$\sigma_k = 2 \quad \text{for all } \; k. \tag{3.16}$$

Moreover, we consider using the step

$$s_k = [\alpha_k \epsilon]^{\frac{1}{p}} \;\text{ for }\; k \in \{0, \ldots, k_\epsilon - 1\} \tag{3.17}$$

and now verify that this step defines a global minimizer for the model. From (3.15), (3.16), (3.17) and (3.13), we deduce that

$$\begin{aligned} m_k^{(1)}(s_k) &= -\alpha_k \epsilon s_k^{q-1} - \beta_{q,k}[\alpha_k \epsilon] s^{q-1} - \beta_{p,k}[\alpha_k \epsilon]^{\frac{q}{p}} s^{p-1} + \frac{\sigma_k\, s^2}{2} \\ &= -[\alpha_k \epsilon]^{\frac{2}{p}} - \beta_{q,k}[\alpha_k \epsilon]^{\frac{2}{p}} - \beta_{p,k}[\alpha_k \epsilon]^{\frac{2}{p}} + \frac{\sigma_k\, [\alpha_k \epsilon]^{\frac{2}{p}}}{2} \\ &= -[\alpha_k \epsilon]^{\frac{2}{p}} \left(1 - \tfrac{1}{2}\sigma_k + \beta_{q,k} + \beta_{p,k}\right) \\ &= 0. \end{aligned} \tag{3.18}$$

and m_k thus admits a global minimizer for $s = s_k$ because the model is the sum of a quadratic with non-positive slope at $s = 0$ and positive cubic regularization term. We also observe that, because of (3.13),

$$\frac{1}{q} + \frac{\beta_{q,k}}{q} + \frac{\beta_{p,k}}{p} = \begin{cases} 1 + \frac{1}{2}\beta_{q,k} \geq 1 & \text{if } q = 1, \\ \frac{1}{2} - \frac{1}{2}\beta_{q,k} \geq \frac{1}{4} & \text{if } q = 2. \end{cases} \tag{3.19}$$

As a consequence

$$T_{f,2}(x_k, s_k) = f_k^{(0)} - [\alpha_k \epsilon]^{\frac{3}{p}} \left(\frac{1}{q} + \frac{\beta_{q,k}}{q} + \frac{\beta_{p,k}}{p}\right) \leq f_k^{(0)} - \frac{s_k^3}{4}, \tag{3.20}$$

irrespective of the value of q. This gives us the necessary information to define the sequence of function values by

$$f_0^{(0)} = 3 \times 2^{\frac{3}{p}} \tag{3.21}$$

and

$$f_{k+1}^{(0)} = T_{f,2}(x_k, s_k) + \beta_{0,k+1} s_k^3 \tag{3.22}$$

for some $\beta_{0,k+1}$ (the perturbation on function values) such that

$$\beta_{0,0} = 0 \quad \text{and} \quad |\beta_{0,k}| \leq \frac{1 - \eta_1}{4} \;\text{ for }\; k \in \{1, \ldots, k_\epsilon\}. \tag{3.23}$$

As a consequence, we have from (2.4) and (3.20) that

$$|\rho_k - 1| = \left| \frac{f_{k+1}^{(0)} - T_{f,2}(x_k, s_k)}{f_k^{(0)} - T_{f,2}(x_k, s_k)} \right| =\leq 4\,|\beta_{0,k+1}| \leq 1 - \eta_1,$$

and thus $\rho_k \geq \eta_1$ for $k \in \{0, \ldots, k_\epsilon - 1\}$, so that every iteration is successful. Our choice of a constant σ_k is therefore acceptable in view of (2.5) and $x_{k+1} = x_k + s_k$ for $k \in \{0, \ldots, k_\epsilon - 1\}$. We may then set

$$x_0 = 0 \quad \text{and} \quad x_k = \sum_{i=0}^{k-1} s_i \;\text{ for }\; k \in \{1, \ldots, k_\epsilon\}. \tag{3.24}$$

A simple calculation shows that

$$\phi_{f,1}(x_k) = \left\{ \begin{array}{ll} \alpha_k\epsilon(\beta_{q,k} + 1) & \text{if } q = 1 \\ [\alpha_k\epsilon]^2 \beta_{p,k} & \text{if } q = 2 \end{array} \right. \tag{3.25}$$

If $q = 2$, (3.13) implies that $\beta_{2,k} + 1 \geq 1$ and thus

$$T_{f,2}(x_k, d) = f_k^{(0)} - \beta_{1,k}(\alpha_k\epsilon)^2 d - \tfrac{1}{2}\alpha_k\epsilon(\beta_{2,k} + 1)d^2$$

is a concave quadratic in d with nonpositive slope at the origin. Hence its global minimizer in the interval $[-1, 1]$ (see (2.2)) is achieved for $d = 1$, yielding

$$\phi_{f,2}(x_k) = \beta_{1,k}(\alpha_k\epsilon)^2 + \tfrac{1}{2}\alpha_k\epsilon(\beta_{2,k} + 1) = (\alpha_k\epsilon)\big(\beta_{p,k}(\alpha_k\epsilon) + \tfrac{1}{2}(\beta_{q,k} + 1)\big) \tag{3.26}$$

(the value of $\phi_{f,2}(x_k)$ is irrelevant if $q = 1$).
Let us now consider the values of $\phi_{f,1}(x_k)$ (if $q = 1$) or $\phi_{f,1}(x_k)$ and $\phi_{f,2}(x_k)$ (if $q = 2$) as a function of k. Consider the case where $q = 1$ first. Then (3.25), (3.12) and (3.13) give that

$$\phi_{f,1}(x_k) > \epsilon \;\text{ for }\; k \in \{0, \ldots, k_\epsilon - 1\} \;\text{ and }\; \phi_{f,1}(x_{k_\epsilon}) = 0 \tag{3.27}$$

and (3.8) holds. For the case $q = 2$, (3.13), the inequality $\epsilon \leq \frac{1}{4}$ and (3.12) give that

$$\beta_{p,k}(\alpha_k\epsilon) + \tfrac{1}{2}(\beta_{q,k} + 1) = -\beta_{q,k}(\alpha_k\epsilon) + \tfrac{1}{2}(\beta_{q,k} + 1) \geq \tfrac{1}{2}$$

Substituting this inequality in (3.26) and using (3.12), we obtain that

$$\phi_{f,2}(x_k) > \tfrac{1}{2}\epsilon \;\text{ for }\; k \in \{0, \ldots, k_\epsilon - 1\} \;\text{ and }\; \phi_{f,2}(x_{k_\epsilon}) = 0, \tag{3.28}$$

where we again used (3.12) to derive the last equality. Moreover, from (3.25) and (3.12),

$$\phi_{f,1}(x_{k_\epsilon}) = 0 \tag{3.29}$$

(when $q = 2$, the value of $\phi_{f,1}(x_k)$ for $k \in \{0, \ldots, k_\epsilon - 1\}$ is irrelevant since $\phi_{f,2}(x_k) > \frac{1}{2}\epsilon$ for these k's). Thus termination conditions (3.9) and (3.10) hold.
Turning now to the function values, we observe that (3.19) implies that

$$\frac{1}{q} + \frac{\beta_{q,k}}{q} + \frac{\beta_{p,k}}{p} \leq 1 + \left|\frac{1}{q} - \frac{1}{p}\right| \beta_{q,k} \leq \frac{5}{4},$$

and therefore, successively using (3.22), (3.20), (3.17), (3.23) and (3.12), that

$$f_k^{(0)} - f_{k+1}^{(0)} = [\alpha_k\epsilon]^{\frac{3}{p}}\left(\frac{1}{q}+\frac{\beta_{q,k}}{q}+\frac{\beta_{p,k}}{p}\right) - \beta_{0,k+1}[\alpha_k\epsilon]^{\frac{3}{p}} \leq [\alpha_k\epsilon]^{\frac{3}{p}}\left(\frac{5}{4}+\frac{1-\eta_1}{4}\right) \leq \frac{3}{2}[2\epsilon]^{\frac{3}{p}}.$$

Thus, using the inequalilty $\epsilon < 1$, we obtain that

$$\begin{array}{rcl} f_0^{(0)} - f_{k_\epsilon}^{(0)} & \leq & f_0^{(0)} - \frac{3}{2}\times 2^{\frac{3}{p}} \times k_\epsilon \epsilon^{\frac{3}{p}} \\ & = & f_0^{(0)} - \frac{3}{2}\times 2^{\frac{3}{p}} \times \lceil \epsilon^{-\frac{3}{p}} \rceil \epsilon^{\frac{3}{p}} \\ & \leq & f_0^{(0)} - \frac{3}{2}\times 2^{\frac{3}{p}} \times (1+\epsilon^{-\frac{3}{p}})\epsilon^{\frac{3}{p}} \\ & = & f_0^{(0)} - \frac{3}{2}\times 2^{\frac{3}{p}} \times (1+\epsilon^{\frac{3}{p}}) \\ & < & f_0^{(0)} - 3\times 2^{\frac{3}{p}}. \end{array} \tag{3.30}$$

Moreover

$$f_k^{(0)} - f_{k+1}^{(0)} = [\alpha_k\epsilon]^{\frac{3}{p}}\left(\frac{1}{q}+\frac{\beta_{q,k}}{q}+\frac{\beta_{p,k}}{p}\right) - \beta_{0,k+1}[\alpha_k\epsilon]^{\frac{3}{p}} \geq [\alpha_k\epsilon]^{\frac{3}{p}}\left(\frac{5}{4}-\frac{1-\eta_1}{4}\right) \geq [\alpha_k\epsilon]^{\frac{3}{p}} > 0,$$

implying that the sequence $\{f_k^{(0)}\}$ is decreasing. Thus, in view of (3.30) and (3.21), we derive that

$$f_k^{(0)} \in \left[0, 3\times 2^{\frac{3}{p}}\right] \quad \text{for} \;\; k \in \{0,\ldots,k_\epsilon\}. \tag{3.31}$$

We now turn to verifying the conditions of [11, Theorem A.9.2] allowing interpolation by a piecewise polynomial with Lipschitz Hessian. We first verify, using (3.22) and (3.23), that

$$|f_{k+1}^{(0)} - T_{f,2}(s_k)| \leq \left(\frac{1-\eta_1}{4}\right) s_k^3. \tag{3.32}$$

We also have from (3.14) that

$$\begin{array}{rcl} T_{f,2}^{(1)}(x_k,s_k) & = & -\alpha_k\epsilon s^{q-1} - \beta_{q,k}[\alpha_k\epsilon]s^{q-1} - \beta_{p,k}[\alpha_k\epsilon]^{\frac{q}{p}}s^{p-1} \\ & = & -[\alpha_k\epsilon]^{\frac{2}{p}}(1+\beta_{q,k}+\beta_{p,k}) = -s_k^2 \end{array} \tag{3.33}$$

and

$$\begin{array}{rcl} T_{f,2}^{(2)}(x_k,s_k) & = & -(q-1)\alpha_k\epsilon s^{q-2} - (q-1)\beta_{q,k}[\alpha_k\epsilon]s^{q-2} - (p-1)\beta_{p,k}[\alpha_k\epsilon]^{\frac{q}{p}}s^{p-2} \\ & = & -(q-1)[\alpha_k\epsilon]^{\frac{1}{p}}(1+\beta_{q,k}) - (p-1)\beta_{p,k}[\alpha_k\epsilon]^{\frac{1}{p}} \\ & = & s_k[-(q-1)(1+\beta_{q,k}) - (p-1)\beta_{p,k}]. \end{array} \tag{3.34}$$

Moreover,

$$\frac{s_{k+1}^p}{s_k^p} = \frac{\alpha_{k+1}}{\alpha_k} \leq 2 \tag{3.35}$$

because of (3.12). Using now (3.33),(3.35) and (3.13) for $q = 1$, we obtain that, for $k \in \{0,\ldots,k_\epsilon - 1\}$,

$$|f_{k+1}^{(1)} - T_{f,2}^{(1)}(x_k,s_k)| = |-s_{k+1}^2(1+\beta_{q,k}) - s_k^2| \leq |s_k^2 + 2s_k^2(1+\tfrac{1}{2})| = 4\,s_k^2 \tag{3.36}$$

while using (3.34),(3.35) and (3.13) for $q = 1$ gives that, for $k \in \{0,\ldots,k_\epsilon - 1\}$,

$$|f_{k+1}^{(2)} - T_{f,2}^{(2)}(x_k,s_k)| = |-\beta_{p,k+1}s_{k+1} + \beta_{p,k}s_k| \leq \beta_{q,k+1}2^{\frac{1}{2}}s_k + \beta_{q,k}s_k \leq 2\,s_k. \tag{3.37}$$

Similarly, using (3.33),(3.35) and (3.13) for $q = 2$ yields that, for $k \in \{0, \dots, k_\epsilon - 1\}$,

$$|f_{k+1}^{(1)} - T_{f,2}^{(1)}(x_k, s_k)| = |-\beta_{p,k+1} s_{k+1}^2 - s_k^2| \le 4\beta_{q,k+1} s_k^2 + s_k^2 \le 3\, s_k^2 \tag{3.38}$$

and using (3.34),(3.35) and (3.13) for $q = 2$ implies that, for $k \in \{0, \dots, k_\epsilon - 1\}$,

$$\begin{aligned} |f_{k+1}^{(2)} - T_{f,2}^{(2)}(x_k, s_k)| &= |-s_{k+1}(1+\beta_{q,k+1}) + s_k(1+\beta_{q,k})| \\ &\le s_k[2(1+\beta_{q,k+1}) + 1 + \beta_{q,k}] \\ &< \tfrac{9}{2}\, s_k \end{aligned} \tag{3.39}$$

Combining (3.36)–(3.39), we see that, in all cases and for all $k \in \{0, \dots, k_\epsilon - 1\}$,

$$|f_{k+1}^{(1)} - T_{f,2}^{(1)}(x_k, s_k)| \le \tfrac{9}{2}\, s_k^2 \;\text{ and }\; |f_{k+1}^{(2)} - T_{f,2}^{(2)}(x_k, s_k)| \le \tfrac{9}{2}\, s_k. \tag{3.40}$$

In addition, (3.11), (3.12), (3.13), the fact that $\epsilon \le \frac{1}{2}$ and (3.31) yield that, for all $j \in \{0, 1, 2\}$ and all $k \in \{0, \dots, k_\epsilon\}$,

$$|f_k^{(j)}| \le \max\left[\frac{5}{2},\, 3 \times 2^{\frac{3}{p}}\right] \quad \text{and} \quad |s_k| \le 1, \tag{3.41}$$

where we used (3.17) and (3.12) to derive the second inequality.
Combining (3.32), (3.40) and (3.41) , we may now apply [11, Theorem A.9.2] for Hermite interpolation with

$$\kappa_f = \max\left[\frac{9}{2},\, \frac{5}{2},\, 3 \times 2^{\frac{3}{p}},\, 1\right] = 9 \times 2^{\frac{3}{p}-1} \approx \left\{ \begin{array}{ll} 8.485 & \text{if } q = 1 \\ 24 & \text{if } q = 2 \end{array} \right.$$

and deduce the existence of a twice times continuously differentiable piecewise polynomial function $f_{\mathcal{A},\mathcal{B}}$ from $\mathbb{R}$ to $\mathbb{R}$ with bounded continuous derivatives of degrees zero to two and Lipschitz continuous Hessian, which interpolates the data given by $\{\{f_k^{(j)}\}_{j=0}^2\}_{k=0}^{k_\epsilon}$ at the iterates $\{x_k\}_{k=0}^{k_\epsilon}$. Moreover, its Hessian's Lipschitz constant L only depends on κ_f. The same theorem, together with (3.31), also ensures that $|f_{\mathcal{A},\mathcal{B}}^{(j)}(x)|$ is uniformly bounded for all $j \in \{0, 1, 2\}$ and all $x \in \mathbb{R}$. Finally, (3.17), (3.12), the inequality $\epsilon \le \frac{1}{4}$ and (3.24) imply that

$$x_k \in [0, \tfrac{1}{2} k_\epsilon] \quad \text{for all} \quad k \in \{0, \dots, k_\epsilon\}, \tag{3.42}$$

irrespective of the choice of $\mathcal{A}$ and $\mathcal{B}$. This conclude the construction of our example.

4 Discussion

We note that the example of [11, Section 12.2.2.4] corresponds to selecting

$$\beta_{k,j} = 0 \quad \text{and} \quad \alpha_k = 1 + \frac{k_\epsilon - k}{k_\epsilon}$$

for all k and j in the above development[(2)].

[(2)] And also corrects a minor error in equation (12.2.78) of [11].

However, the very fact that we may choose

$$\beta_{0,k+1} \in \left[-\frac{1-\eta_1}{4}, \frac{1-\eta_1}{4}\right], \quad \beta_{p,k} = -\beta_{q,k} \in [0, \tfrac{1}{2}] \quad \text{and} \quad \alpha_k \in [1,2] \tag{4.43}$$

for each k and each j (see (3.12) (3.13) and (3.23)) tells us, in conjunction with (3.11) and (3.22), that the interpolation data at x_k may be chosen, for $k \in \{0,\ldots,k_\epsilon - 1\}$, arbitrarily in an interval of radius $2(1-\eta_1)s_k^{p+1}/4$ for the objective function see (3.32)), and in intervals of radius at least ϵ for the derivative of degree q and $\frac{1}{2}\epsilon^{q/p}$ for the derivative of degree p (see (3.11)–(3.13)). Moreover, the proof of [11, Theorem A.9.2] reveals that, in each interval $[x_k, x_{k+1}]$, $f_{\mathcal{A},\mathcal{B}}$ is a linear combination of the monomials 1, s, s^2, s^3, s^4 and s^5 whose coefficients depend continuously and bijectively on the values of $f_k^{(j)}$ and $f_{k+1}^{(j)}$ for $j \in \{0,1,2\}$. Now observe that, in our example,

$$\begin{pmatrix} f_k^{(0)} \\ f_k^{(q)} \\ f_k^{(p)} \end{pmatrix} = \begin{pmatrix} T_{2,f}(x_{k-1}, s_{k-1}) + \beta_{0,k}[\alpha_{k-1}\epsilon]^{\frac{3}{p}} \\ -[\alpha_k\epsilon](1+\beta_{q,k}) \\ [\alpha_k\epsilon]^{\frac{q}{p}}\beta_{q,k} \end{pmatrix} \stackrel{\rm def}{=} \Theta_k(\beta_{0,k}, \alpha_k, \beta_{q,k}).$$

Since the determinant of the Jacobian of Θ_k is given by

$$\left|\begin{pmatrix} [\alpha_{k-1}\epsilon]^{\frac{3}{p}} & 0 & 0 \\ 0 & -\epsilon(1+\beta_{q,k}) & -\alpha_k\epsilon \\ 0 & \frac{q}{p}\beta_{q,k}\epsilon[\alpha_k\epsilon]^{\frac{q}{p}-1} & [\alpha_k\epsilon]^{\frac{q}{p}} \end{pmatrix}\right| = -[\alpha_{k-1}\epsilon]^{\frac{3}{p}}\epsilon^{\frac{3}{p}}\alpha_k^{\frac{q}{p}}\left[1 + \left(1-\frac{q}{p}\right)\beta_{q,k}\right]$$

and is nonzero for all values of $(\beta_{0,k}, \alpha_k, \beta_{q,k})$ in the ranges (4.43), J_k is nonsingular in these ranges and we deduce that Θ_k is continuous and bijective for all k. As a consequence, different values of $\beta_{0,k}$, α_k and $\beta_{q,k}$ for a at least one k result in piecewise polynomials $f_{\mathcal{A},\mathcal{B}}$ which (continuoulsy) differ in the neighbourhood (in x) of at least one iterate, the radius of this neighbourhood depending on the parameter-independent Lipschitz constant L. We may therefore conclude that

the set of functions on which the AR2 *algorithm can take* $\mathcal{O}(\epsilon^{-3/p})$ *evaluations is of nonzero measure (in the standard topology for continuous functions from* $\mathbb{R}$ *to* $\mathbb{R}$*)*

although the measure of this set may shrink when ϵ tends to zero.

Figure 1 illustrates the diversity of possible examples by showing the graphs of a few $f_{\mathcal{A},\mathcal{B}}$ for $q = 1$ corresponding to different choices of $\mathcal{A}$ and $\mathcal{B}$ with $\beta_{0,k} = 0$ for all k. The unpertubed example (i.e. with $\alpha_k = 1$ and $\beta_{j,k} = 0$ for all k is the top fatter curve while the bottom dashed one corresponds to setting $\alpha_k = 2$ and $\beta_{q,k} = 0$ for all k. The thin continuous curves show what happens if $\beta_{q,k} = \frac{1}{2}$ is used for all k with $\alpha_k = 1$ or $\alpha_k = 2$. Curves have different maximum abscissa for the same number of iterations because the steplength s_k given by (3.17) varies with α_k.

The reader may also wonder if our example contradicts the result of [16], which states that the AR2 algorithm terminates in fact in $o(\epsilon^{3/p})$ rather that in $O(\epsilon^{3/p})$ iterations and evaluations. Fortunately, this is not the case, because the contexts in which these results are derived differ. More specifically, the example presented above constructs a different function for every choice of ϵ (and possible perturbations). It states that, if ϵ is fixed, then there exists a set of sufficiently smooth functions depending on ϵ causing slow convergence of the AR2

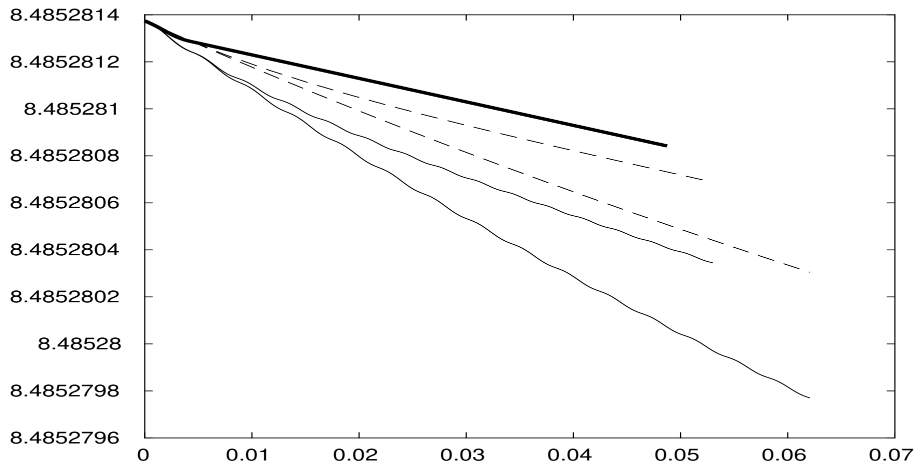

Figure 1: A few members of the set of functions causing slow convergence of the adaptive regularization algorithm ($\epsilon = 10^{-5}$, $q = 1$, showing the first 15 iterations)

algorithm. In contrast, the result of [16] states that, for any sufficiently smooth (fixed) function, the number of iterations (and evaluations) required to achieve q-th order ϵ-approximate criticality *for this function* increases slower than $\epsilon^{-3/p}$ when ϵ tends to zero.

Acknowledgement

Thanks to Geovani Grapiglia for persuading the author to write this note during an interesting discussion at the Brazopt XIV conference in Rio de Janeiro, and to the organizers of this meeting for their support. Thanks also to Serge Gratton for his encouragements, and to Benedetta Morini, Stefania Bellavia and Margherita Porcelli for helping to check the details.

References

[1] S. Bellavia, G. Gurioli, and B. Morini. Adaptive cubic regularization methods with dynamic inexact Hessian information and applications to finite-sum minimization. *IMA Journal of Numerical Analysis*, 41(1):764–799, 2021.

[2] E. Bergou, Y. Diouane, and S. Gratton. On the use of the energy norm in trust-region and adaptive cubic regularization subproblems. *Computational Optimization and Applications*, 68:533–554, 2017.

[3] T. Bianconcini, G. Liuzzi, B. Morini, and M. Sciandrone. On the use of iterative methods in cubic regularization for unconstrained optimization. *Computational Optimization and Applications*, 60(1):35–57, 2015.

[4] E. G. Birgin, J. L. Gardenghi, J. M. Martínez, S. A. Santos, and Ph. L. Toint. Worst-case evaluation complexity for unconstrained nonlinear optimization using high-order regularized models. *Mathematical Programming, Series A*, 163(1):359–368, 2017.

[5] Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford. Lower bounds for finding stationary points I. *Mathematical Programming, Series A*, 184:71–120, 2020.

[6] 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.

[7] C. Cartis, N. I. M. Gould, and Ph. L. Toint. Adaptive cubic overestimation methods for unconstrained optimization. Part II: worst-case function-evaluation complexity. *Mathematical Programming, Series A*, 130(2):295–319, 2011.

[8] C. Cartis, N. I. M. Gould, and Ph. L. Toint. Complexity bounds for second-order optimality in unconstrained optimization. *Journal of Complexity*, 28:93–108, 2012.

[9] C. Cartis, N. I. M. Gould, and Ph. L. Toint. On the evaluation complexity of cubic regularization methods for potentially rank-deficient nonlinear least-squares problems and its relevance to constrained nonlinear optimization. *SIAM Journal on Optimization*, 23(3):1553–1574, 2013.

[10] C. Cartis, N. I. M. Gould, and Ph. L. Toint. A concise second-order evaluation complexity for unconstrained nonlinear optimization using high-order regularized models. *Optimization Methods and Software*, 35(2):243–256, 2020.

[11] C. Cartis, N. I. M. Gould, and Ph. L. Toint. *Evaluation complexity of algorithms for nonconvex optimization.* Number 30 in MOS-SIAM Series on Optimization. SIAM, Philadelphia, USA, June 2022.

[12] J. P. Dussault. Simple unified convergence proofs for the trust-region and a new ARC variant. Technical report, University of Sherbrooke, Sherbrooke, Canada, 2015.

[13] N. I. M. Gould, M. Porcelli, and Ph. L. Toint. Updating the regularization parameter in the adaptive cubic regularization algorithm. *Computational Optimization and Applications*, 53(1):1–22, 2012.

[14] N. I. M. Gould, D. P. Robinson, and H. S. Thorne. On solving trust-region and other regularised subproblems in optimization. *Mathematical Programming, Series C*, 2(1):21–57, 2010.

[15] G. N. Grapiglia, J. Yuan, and Y. Yuan. On the convergence and worst-case complexity of trust-region and regularization methods for unconstrained optimization. *Mathematical Programming, Series A*, 152:491–520, 2015.

[16] S. Gratton, C.-K. Sim, and Ph. L. Toint. Refining asymptotic complexity bounds for nonconvex optimization methods, including why steepest descent is $o(\epsilon^{-2})$ rather than $\mathcal{O}(\epsilon^{-2})$. arXiv:2408.09124, 2024.

[17] S. Gratton and Ph. L. Toint. Adaptive regularization minimization algorithms with non-smooth norms. *IMA Journal of Numerical Analysis*, 43(2):920–949, 2023.

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

[19] J. M. Martínez and M. Raydan. Cubic-regularization counterpart of a variable-norm trust-region method for unconstrained minimization. *Journal of Global Optimization*, 68:367–385, 2017.

[20] Yu. Nesterov and B. T. Polyak. Cubic regularization of Newton method and its global performance. *Mathematical Programming, Series A*, 108(1):177–205, 2006.

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