Science & Technology
Facilities Council

Preprint
RAL-P-2016-010

Universal regularization methods- varying the power, the smoothness and the accuracy

C Cartis, NIM Gould, PL Toint

December 2016

Submitted for publication in SIAM Journal on Optimization

RAL Library
STFC Rutherford Appleton Laboratory
R61
Harwell Oxford
Didcot
OX11 0QX

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

ISSN 1361- 4762

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.

Universal regularization methods—varying the power, the smoothness and the accuracy

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

ABSTRACT

Adaptive cubic regularization methods have emerged as a credible alternative to linesearch and trust-region for smooth nonconvex optimization, with optimal complexity amongst second-order methods. Here we consider a general/new class of adaptive regularization methods, that use first- or higher-order local Taylor models of the objective regularized by a(ny) power of the step size and applied to convexly-constrained optimization problems. We investigate the worst-case evaluation complexity/global rate of convergence of these algorithms, when the level of sufficient smoothness of the objective may be unknown or may even be absent. We find that the methods accurately reflect in their complexity the degree of smoothness of the objective and satisfy increasingly better bounds with improving accuracy of the models. The bounds vary continuously and robustly with respect to the regularization power and accuracy of the model and the degree of smoothness of the objective.

[1] Mathematical Institute, Andrew Wiles Building, University of Oxford, Oxford,
OX2 6GG, England, EU. Email: coralia.cartis@maths.ox.ac.uk .
Current reports available from "http://eprints.maths.ox.ac.uk/view/groups/nag/".

[2] This work was supported by the EPSRC grant EP/I028854/1.

[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".

[4] This work was supported by the EPSRC grant EP/M025179/1.

[5] Namur Center for Complex Systems (naXys), University of Namur, 61, rue de Bruxelles,
B-5000 Namur, Belgium, EU. Email : philippe.toint@unamur.be .
Current reports available from "http://www.fundp.ac.be/∼phtoint/pht/publications.html".

Scientific Computing Department
Rutherford Appleton Laboratory
Oxfordshire OX11 0QX

1 Introduction

We consider the (possibly) convexly-constrained optimization problem

$$\min_{x\in\mathcal{F}} f(x) \tag{1.1}$$

where $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ is a smooth, possibly nonconvex, objective and where the feasible set $\mathcal{F} \subset \mathbb{R}^n$ is closed, convex and non-empty (for example, the set $\mathcal{F}$ could be described by simple bounds and both polyhedral and more general convex constraints)[1]. Clearly, the case of unconstrained optimization is covered here by letting $\mathcal{F} = \mathbb{R}^n$. We are interested in the case when $f \in \mathcal{C}^{p,\beta_p}(\mathcal{F})$, namely, f is $p-$times continuously differentiable in $\mathcal{F}$ with the pth derivative being Hölder continuous of (unknown) degree $\beta_p \in [0,1]$[2]. We consider adaptive regularization methods applied to problem (1.1) that generate feasible iterates x_k that are (possibly very) approximate minimizers over $\mathcal{F}$ of local models of the form

$$m_k(x_k+s) = T_p(x_k,s) + \frac{\sigma_k}{r}\|s\|_2^r,$$

where $T_p(x_k,s)$ is the pth order Taylor polynomial of f at x_k and $r > p \geq 1$. The parameter $\sigma_k > 0$ is adjusted to ensure sufficient decrease in f happens when the model value is decreased. In this paper, we derive evaluation complexity bounds for finding first-order critical points of (1.1) using higher-order adaptive regularization methods. Despite the higher order of the models, the model minimization is performed only approximately, generalizing the approach in [3]. The proposed methods also ensure that the steps are 'sufficiently long', in a new way, generalizing ideas in [11]. The ensuing complexity analysis shows the robust interplay of the regularization power r, the model accuracy p and the degree of smoothness β_p of the objective, with some surprising results. In particular, we find that the degree of smoothness of the objective—which is often unknown and is even allowed to be absent here—is accurately reflected in the complexity of the methods, independently of the regularization power, provided the latter is sufficiently large. Furthermore, for all possible powers r, the methods satisfy increasingly better bounds as the accuracy p of the models and smoothness level β_p are increased. All bounds vary continuously as a function of the regularization power and smoothness level. Table 4.1 in Section 4 summarizes our complexity bounds.

We now review existing literature in detail and further clarify our approach, motivation and contributions. Cubic regularization for the (unconstrained) minimization of $f(x)$ for $x \in \mathbb{R}^n$ was proposed independently by [12,16,17], with [16] showing it has better global worst-case function evaluation complexity than the method of steepest descent. Extending [16], we proposed some practical variants – Adaptive Regularization with Cubics (ARC) [5] – that satisfy the same complexity bound as the regularization methods in [16], namely at most $\mathcal{O}(\epsilon^{-\frac{3}{2}})$ evaluations are needed to find a point x for which

$$\|\nabla_x f(x)\| \leq \epsilon, \tag{1.2}$$

under milder requirements on the algorithm (specifically, inexact model minimization). We further showed in [4,6] that this complexity bound for ARC is sharp and optimal for a large class

[1] We are tacitly assuming that the cost of evaluating constraint functions and their derivatives is negligible.

[2] Note that if $\beta_p > 1$, then the resulting class of objectives is restricted to multivariate polynomials of degree p. If $p = 1$, we only allow $\beta_1 \in (0,1]$, for reasons to be explained later in the paper.

of second-order methods when applied to functions with globally Lipschitz-continuous second derivatives. Quadratic regularization, namely, a first order accurate model of the objective regularized by a quadratic term, has also been extensively studied, and shown to satisfy the complexity bound of steepest descent, namely, $\mathcal{O}(\epsilon^{-2})$ evaluations to obtain (1.2) [14]. It was also shown in [5] that one can loosen the requirement that global Lipschitz continuity of the second derivative holds, to just global Hölder continuity of the same derivative with exponent $\beta_2 \in (0,1]$. Then, if one also regularizes the quadratic objective model by the power $2+\beta_2$ of the step, involving the (often unknown) Hölder exponent, the resulting method requires $\mathcal{O}(\epsilon^{-\frac{2+\beta_2}{1+\beta_2}})$ evaluations, which just as a function of ϵ, belongs to the interval $\left[\epsilon^{-\frac{3}{2}}, \epsilon^{-2}\right]$; these bounds are sharp and optimal for objectives with corresponding level of smoothness of the Hessian [6]. Note that this bound also holds if $\beta_2 = 0$.

An important related question and extension was answered in [3]: if higher-order derivatives are available, can one improve the complexity of regularization methods? It was shown in [3] that if one considers approximately minimizing a $(r-1)$th order Taylor model of the objective regularized by the (weighted) rth power of the (Euclidean) norm of the step in each iteration (so $r = p+1$), the complexity of the resulting adaptive regularization method is $\mathcal{O}(\epsilon^{-\frac{r}{r-1}})$ evaluations to obtain (1.2), under the assumption that the $(r-1)$th derivative tensor is globally Lipschitz continuous. The method proposed in [3] measures progress of each iteration by comparing the Taylor model decrease (without the regularization term) to that of the true function decrease and only requiring mild approximate (local) minimization of the regularized model. Here, we generalize these higher-order regularization methods from [3] to allow for an arbitrary local Taylor model, an arbitrary regularization power of the step and varying levels of smoothness of the highest-order derivative in the Taylor model.

The interest in considering relaxations of Lipschitz continuity to Hölder continuity of derivatives comes not only from the needs of some engineering applications (such as flows in gas pipelines [10, Section 17] and properties of nonlinear PDE problems [1]), but also in its own right in optimization theory, as a bridging case between the smooth and non-smooth classes of problems [13, 15]. In particular, a zero Hölder exponent for a Hölder continuous derivative corresponds to a bounded derivative, an exponent in $(0,1)$ corresponds to a continuous but not necessarily differentiable derivative, while an exponent of 1 corresponds to a Lipschitz continuous derivative that can be differentiated again. For the case of function with Hölder-continuous gradients, methods have already been devised, and their complexity analysed, both as a weaker set of assumptions and as an attempt to have a 'smooth' transition between the smooth and nonsmooth (convex) problem classes, without knowing a priori the level of smoothness of the gradient (i.e., the Hölder exponent) [9, 15]; even lower complexity bounds are known [13]. In [7] we considered regularization methods applied to nonconvex objectives with Hölder continuous gradients (with unknown exponent $\beta_1 \in (0,1]$), that employ a first-order quadratic model of the objective regularized by the rth power of the step. We showed that the worst-case complexity of the resulting regularization methods varies depending on $\min\{r, 1+\beta_1\}$. In particular, when $1 < r \leq 1+\beta_1$, the methods take at most $\mathcal{O}\left(\epsilon^{-\frac{r}{r-1}}\right)$ evaluations/iterations until termination, and otherwise, at most $\mathcal{O}\left(\epsilon^{-\frac{1+\beta_1}{\beta_1}}\right)$ evaluations/iterations to achieve the same condition. The latter complexity bound reflects the smoothness of the objective's landscape, without prior knowledge or use of it in the algorithm, and is independent of the regularization power. Here we generalize the approach in [7] to pth order Taylor models and find that similar bounds can be obtained. Also, we are able to allow $\beta_p = 0$ provided $p \geq 2$.

Recently, [11] proposed a new cubic regularization scheme that yields a *universal* algorithm in the sense that its complexity reflects the (possibly unknown or even absent) degree of sufficient smoothness of the objective (for this, $p = 2$, $r = 3$ and $\beta_2 \in [0, 1]$). Our ARp algorithm includes a modification in a similar (but not identical) vein to that in [11]. In particular, our approach checks/ensures that the length of the step is sufficiently large on all iterations on which the objective is sufficiently decreased, while the technique in [11] uses a specific/new sufficient decrease condition of the objective on each iteration. We generalize the approach in [11] and achieve complexity bounds with similar universal properties for varying r, p and unknown $\beta_p \in [0, 1]$, provided $r \geq p + \beta_p$. We are also able to analyze ARp's complexity in the regime $p < r \leq p + \beta_p$ providing continuously varying results with r and β_p.

Our algorithm can be applied to convexly-constrained optimization problems with nonconvex objectives, where the constraint/feasibility evaluations are inexpensive, offering another generalization of proposals in [3] and [11] which are presented for the unconstrained case only; we also extend [11] by allowing inexact subproblem solution.

The structure of the paper is as follows. Section 2 describes our main algorithmic framework, ARp. Section 3 presents our complexity analysis while Section 4 concludes with a summary of our complexity bounds (see Table 4.1) and a discussion of the results.

2 A universal adaptive regularization framework - ARp

Let $f \in \mathcal{C}^p(\mathcal{F})$, with p integer, $p \geq 1$; let $r \in \mathbb{R}$, $r > p \geq 1$. We measure optimality using a suitable continuous first-order criticality measure for (1.1). We define this measure for a general function $h : \mathbb{R}^n \longrightarrow \mathbb{R}$ on $\mathcal{F}$: for an arbitrary $x \in \mathcal{F}$, the criticality measure is given by

$$\pi_h(x) \stackrel{\text{def}}{=} \|P_{\mathcal{F}}[x - \nabla_x h(x)] - x\|, \tag{2.1}$$

where $P_{\mathcal{F}}$ denotes the orthogonal projection onto $\mathcal{F}$ and $\|\cdot\|$ the Euclidean norm. Letting $h(x) := f(x)$ in (2.1), it is known that x is a first-order critical point of problem (1.1) if and only if $\pi_f(x) = 0$. Also note that

$$\pi_f(x) = \|\nabla_x f(x)\| \text{ whenever } \mathcal{F} = \mathbb{R}^n.$$

For more properties of this measure see [2, 8].

Our ARp algorithm generates feasible iterates x_k that (possibly very) approximately minimize the local model

$$m_k(x_k + s) = T_p(x_k, s) + \frac{\sigma_k}{r}\|s\|^r \text{ subject to } x_k + s \in \mathcal{F}, \tag{2.2}$$

which is a regularization of the pth order Taylor model of f around x_k,

$$T_p(x_k, s) = f(x_k) + \sum_{j=1}^{p} \frac{1}{j!}\nabla_x^j f(x^k)[s]^j, \tag{2.3}$$

where $\nabla_x^j f(x^k)[s]^j$ is the jth order tensor $\nabla_x^j f(x^k)$ of f at x_k applied to the vector s repeated j times. Note that $T_p(x_k, 0) = f(x_k)$. We will also use the measure (2.1) with $h(s) := m_k(x_k + s)$

for terminating the approximate minimization of $m_k(x_k+s)$, and for which we have again

$$\pi_{m_k}(x_k+s) = \|\nabla_s m_k(x_k+s)\| \;\text{ whenever }\; \mathcal{F} = \mathbb{R}^n.$$

A summary of the main algorithmic framework is as follows.

Algorithm 2.1: A universal ARp variant.

Step 0: Initialization. An initial point $x_0 \in \mathcal{F}$ and an initial regularization parameter $\sigma_0 > 0$ are given, as well as an accuracy level $\epsilon > 0$. The constants η_1, η_2, γ_1, γ_2 and γ_3, θ, $\sigma_{\min}$ and α, are also given and satisfy

$$\theta > 0, \quad \sigma_{\min} \in (0, \sigma_0], \quad 0 < \eta_1 \le \eta_2 < 1 \;\text{ and }\; 0 < \gamma_3 < 1 < \gamma_1 < \gamma_2 \;\text{ and }\; \alpha \in \left(0, \frac{1}{3}\right]. \tag{2.4}$$

Compute $f(x_0)$, $\nabla_x f(x_0)$ and set $k = 0$. If $\pi_f(x_0) < \epsilon$, terminate. Else, for $k \ge 0$, do:

Step 1: Model set-up. Compute derivatives of f of order 2 to p at x_k.

Step 2: Step calculation. Compute the step s_k by approximately minimizing the model $m_k(x_k+s)$ in (2.2) over $x_k + s \in \mathcal{F}$ such that the following conditions hold,

$$x_k + s_k \in \mathcal{F}, \tag{2.5}$$

$$m_k(x_k+s_k) < f(x_k) \tag{2.6}$$

and

$$\pi_{m_k}(x_k+s_k) \le \theta \|s_k\|^{r-1}. \tag{2.7}$$

Step 3: Test for termination. Compute $\nabla_x f(x_k+s_k)$. If $\pi_f(x_k+s_k) < \epsilon$, terminate with the approximate solution $x_\epsilon = x_k + s_k$.

Step 4: Acceptance of the trial point. Compute $f(x_k+s_k)$ and define

$$\rho_k = \frac{f(x_k) - f(x_k+s_k)}{f(x_k) - T_p(x_k, s_k)}. \tag{2.8}$$

If $\rho_k \ge \eta_1$, check whether

$$\sigma_k \|s_k\|^{r-1} \ge \alpha\epsilon. \tag{2.9}$$

If both $\rho_k \ge \eta_1$ and (2.9) hold, then define $x_{k+1} = x_k + s_k$; otherwise define $x_{k+1} = x_k$.

Step 5: Regularization parameter update. Set

$$\sigma_{k+1} \in \left\{ \begin{array}{ll} [\max(\sigma_{\min}, \gamma_3\sigma_k), \sigma_k] & \text{if } \rho_k \ge \eta_2 \text{ and (2.9) holds,} \\ [\sigma_k, \gamma_1\sigma_k] & \text{if } \rho_k \in [\eta_1, \eta_2) \text{ and (2.9) holds,} \\ [\gamma_1\sigma_k, \gamma_2\sigma_k] & \text{if } \rho_k < \eta_1 \text{ or (2.9) fails.} \end{array} \right. \tag{2.10}$$

Increment k by one, and go to Step 1 if $\rho_k \ge \eta_1$ and (2.9) hold, and to Step 2 otherwise.

Iterations for which $\rho_k \geq \eta_1$ and (2.9) hold (and so $x_{k+1} = x_k + s_k$) are called *successful*, those for which $\rho_k \geq \eta_2$ and (2.9) hold are referred to as *very successful*, while the remaining ones are *unsuccessful.* For a(ny) $j \geq 0$, we denote the set of successful iterations up to j by $\mathcal{S}_j = \{0 \leq k \leq j : \rho_k \geq \eta_1 \text{ and } (2.9) \text{ holds}\}$ and the set of unsuccessful ones by $\mathcal{U}_j = \{0, \ldots, j\} \setminus \mathcal{S}_j$. We have the following simple lemma that relates the number of successful and unsuccessful iterations and that is ensured by the mechanism of the Algorithm 2.1.

Lemma 2.1. [5, Theorem 2.1] For any fixed $j \geq 0$ until termination, let $\sigma_{\text{up}} > 0$ be such that $\sigma_k \leq \sigma_{\text{up}}$ for all $k \leq j$ in Algorithm 2.1. Then

$$|\mathcal{U}_j| \leq \frac{|\log \gamma_3|}{\log \gamma_1} |\mathcal{S}_j| + \frac{1}{\log \gamma_1} \log\left(\frac{\sigma_{\text{up}}}{\sigma_0}\right), \tag{2.11}$$

where $|\cdot|$ denotes the cardinality of the respective index set.

Proof. The proof of (2.11) follows identically to the given reference; note that the sets $\mathcal{S}_j$ and $\mathcal{U}_j$ are not identical to the usual ARC ones in [5] but the mechanism for modifying σ_k in ARp coincides with the one in ARC on these iterations and that is why the proof of this lemma follows identically to [5, Theorem 2.1]. □

Now we comment on the construction of the ARp algorithm. Note that the model minimization conditions (Step 2) and the definition of ρ in Step 4 are straightforward generalizations of the approach in [3] to pth order Taylor models regularized by different powers r of the norm of the step. However, there are two main differences to the by-now standard approaches to (cubic or higher order) regularization methods. Firstly, we check whether the gradient goes below ϵ at each trial points, and if so, terminate on possibly unsuccessful iterations (Step 3). Secondly, we check whether the step s_k is sufficiently long (in the sense of condition (2.9)) on every successful steps, and only allow such sufficiently long steps to be taken by the algorithm; if the step is not sufficiently long (or $\rho_k \leq \eta_1$), σ_k is increased. Note that though the length of the step s_k decreases as σ_k is increased, this is not the case for the expression $\sigma_k \|s_k\|^{r-1}$ in (2.9), which increases with σ_k, as Lemma 3.4 implies. These two additional ingredients—the gradient calculation at each trial point and the step length condition (2.9)—are directly related to trying to achieve universality of ARp, extending ideas from [11].[3]

Remarks. Instead of requiring (2.9) on each successful step, we could ask that each model minimization step calculated in Step 2 satisfies (2.9); if (2.9) failed, σ_k would be increased at the end of Step 2 and the model minimization step would be repeated. This approach may result in an unnecessarily small step in practice. Similarly, following [11], one could replace (2.9) with a different definition for ρ_k, namely, the denominator in ρ_k would be replaced by a rational function in ϵ and σ_k to achieve the desired order of model/function decrease for universal complexity and behaviour. We found the implicit way of controlling the length of the step to be less intuitive but accept this is simply a matter of opinion. According to our calculations, qualitatively similar

[3] We note that without the condition (2.9) on the length of the step, or a similar measure of progress, the complexity of ARp would dramatically worsen (according to our calculations) in the regime when $r > p + \beta_p$.

complexity bounds would be obtained for these two ARp variants.

3 Worst-case complexity analysis of ARp

3.1 Some preliminary properties

We have the following simple consequence of (2.6).

Lemma 3.1. On each iteration of Algorithm 2.1, we have the decrease

$$f(x_k) - T_q(x_k, s_k) \geq \frac{\sigma_k}{r}\|s_k\|^r. \tag{3.1}$$

Proof. Note that condition (2.6) and the definition of $m_k(s)$ in (2.2) immediately give (3.1). □

We have the following upper bound on s_k.

Lemma 3.2. On each iteration of Algorithm 2.1, we have

$$\|s_k\| \leq \max_{1\leq j\leq p}\left\{\left(\frac{pr}{j!\sigma_k}\|\nabla_x^j f(x_k)\|\right)^{\frac{1}{r-j}}\right\}. \tag{3.2}$$

Proof. It follows from (2.6), (2.2) and (2.3) that

$$s_k^T\nabla_x f(x_k) + \frac{1}{2}\nabla_x^2 f(x_k)[s_k, s_k] + \ldots + \frac{1}{p!}\nabla_x^p f(x^k)[s_k, s_k, \ldots, s_k] + \frac{\sigma_k}{r}\|s_k\|^r < 0,$$

which from Cauchy-Schwarz and norm properties, further implies

$$-\|s_k\|\cdot\|\nabla_x f(x_k)\| - \frac{1}{2}\|s_k\|^2\cdot\|\nabla_x^2 f(x_k)\| - \ldots - \frac{1}{p!}\|s_k\|^p\cdot\|\nabla_x^p f(x^k)\| + \frac{\sigma_k}{r}\|s_k\|^r < 0,$$

or equivalently,

$$\sum_{j=1}^{p}\left(\frac{\sigma_k}{pr}\|s_k\|^r - \frac{1}{j!}\|s_k\|^j\cdot\|\nabla_x^j f(x^k)\|\right) < 0.$$

The last displayed equation cannot hold unless at least one of the terms on the left-hand side is negative, which is equivalent to (3.2), using also that $r > p \geq 1$. □

Let us assume that $f \in \mathcal{C}^{p,\beta_p}$, namely,

A.1 $f \in C^p(\mathcal{F})$ and $\nabla_x^p f$ is Hölder continuous on the path of the iterates and trial points, namely,

$$\|\nabla_x^p f(y) - \nabla_x^p f(x_k)\|_T \leq (p-1)!L_p\|y - x_k\|^{\beta_p}$$

holds for all $y \in [x_k, x_k + s_k]$, $k \geq 0$ and some constants $L_p \geq 0$ and $\beta_p \in [0,1]$, where $\|\cdot\|$ is the Euclidean norm on $\mathbb{R}^n$ and $\|\cdot\|_T$ is recursively induced by this norm on the space of the pth order tensors.

A simple consequence of A.1 is that

$$|f(x_k + s_k) - T_p(x_k, s_k)| \leq \frac{L_p}{p(\beta_p + 1)} \|s_k\|^{p+\beta_p}, \quad k \geq 0, \tag{3.3}$$

and

$$\|\nabla_x f(x_k + s_k) - \nabla_s T_p(x_k, s_k)\| \leq \frac{L_p}{\beta_p + 1} \|s_k\|^{p+\beta_p - 1}, \quad k \geq 0; \tag{3.4}$$

see [2] for a proof of (3.3) and (3.4), with A.1 replacing Lipschitz continuity of the pth derivative.

Remark Note that throughout the paper we assume $r > p \geq 1$, $r \in \mathbb{R}$ and $p \in \mathbb{N}$; and that either $p \geq 1$ and $\beta_p \in (0,1]$ or $p \geq 2$ and $\beta_p \in [0,1]$. Thus in both cases $p + \beta_p - 1 > 0$. □

Two useful preliminary lemmas follow.

Lemma 3.3. Assume that A.1 holds. Then on each iteration of Algorithm 2.1, we have

$$\pi_f(x_k + s_k) \leq \frac{L_p}{\beta_p + 1} \|s_k\|^{p+\beta_p - 1} + (\sigma_k + \theta)\|s_k\|^{r-1}. \tag{3.5}$$

Proof. We have that

$$\nabla_s m_k(x_k + s) = \nabla_s T_p(x_k, s) + \sigma_k \|s\|^{r-1} \frac{s}{\|s\|}$$

and so

$$\begin{aligned} \|\nabla_x f(x_k + s_k) - \nabla_s m_k(x_k + s_k)\| &\leq \|\nabla_x f(x_k + s_k) - \nabla_s T_p(x_k, s_k)\| + \sigma_k \|s_k\|^{r-1} \\ &\leq \tfrac{L_p}{\beta_p + 1} \|s_k\|^{p+\beta_p - 1} + \sigma_k \|s_k\|^{r-1}, \end{aligned} \tag{3.6}$$

where we also used (3.4). Now using the contractive property of the projection operator $P_{\mathcal{F}}$ and triangle inequality, we have

$$\begin{aligned} \pi_f(x_k + s_k) &= \|P_{\mathcal{F}}[x_k + s_k - \nabla_x f(x_k + s_k)] - P_{\mathcal{F}}[x_k + s_k - \nabla_s m_k(x_k + s_k)] \\ &\quad + P_{\mathcal{F}}[x_k + s_k - \nabla_s m_k(x_k + s_k)] - (x_k + s_k)\| \\ &\leq \|P_{\mathcal{F}}[x_k + s_k - \nabla_x f(x_k + s_k)] - P_{\mathcal{F}}[x_k + s_k - \nabla_s m_k(x_k + s_k)]\| + \pi_{m_k}(x_k + s_k) \\ &\leq \|\nabla_x f(x_k + s_k)] - \nabla_s m_k(x_k + s_k)\| + \theta \|s_k\|^{r-1}, \end{aligned}$$

where in the last inequality, we also employed the termination condition (2.7). Now (3.5) follows from (3.6). □

Lemma 3.4. Assume that A.1 holds and that Algorithm 2.1 has not terminated. Then, if

$$\sigma_k \geq \max\left\{\theta, \kappa_2 \|s_k\|^{p+\beta_p-r}\right\}, \tag{3.7}$$

where

$$\kappa_2 \stackrel{\rm def}{=} \frac{rL_p}{p(1+\beta_p)(1-\eta_2)}, \tag{3.8}$$

both $\rho_k \geq \eta_2$ and (2.9) hold, and so iteration k is very successful.

Proof. Evaluating ρ_k in (2.8), we deduce

$$|\rho_k - 1| \leq \frac{|f(x_k+s_k) - T_p(x_k,s_k)|}{f(x_k) - T_p(x_k,s_k)} \leq \frac{\frac{L_p}{p(\beta_p+1)}\|s_k\|^{p+\beta_p}}{\frac{\sigma_k}{r}\|s_k\|^r} = \frac{rL_p}{p(\beta_p+1)\sigma_k}\|s_k\|^{p+\beta_p-r}$$

where we also used (3.1) and (3.3). It follows from (2.8) that if $|1-\rho_k| \leq 1-\eta_2$, then $\rho_k \geq \eta_2$. The former condition is satisfied if (3.7) holds.

It remains to show that while Algorithm 2.1 does not terminate, (3.7) also implies (2.9). It follows from (3.5) and $\sigma_k \geq \theta$ that

$$\pi_f(x_k+s_k) \leq \|s_k\|^{p+\beta_p-1}\left(\frac{L_p}{\beta_p+1} + 2\sigma_k\|s_k\|^{r-p-\beta_p}\right) \leq \|s_k\|^{p+\beta_p-1}\left(\kappa_2 + 2\sigma_k\|s_k\|^{r-p-\beta_p}\right),$$

where in the last inequality, we used the definition of κ_2, $r > p$ and $\eta_2 \in (0,1)$. Using (3.7), the last displayed inequality further becomes

$$\pi_f(x_k+s_k) \leq \|s_k\|^{p+\beta_p-1}\left(3\sigma_k\|s_k\|^{r-p-\beta_p}\right) = 3\sigma_k\|s_k\|^{r-1}.$$

Thus $\sigma_k\|s_k\|^{r-1} \geq \frac{1}{3}\pi_f(x_k+s_k)$, which in turn implies (2.9) since $\alpha \leq \frac{1}{3}$ and $\pi_f(x_k+s_k) \geq \epsilon$ as Algorithm 2.1 has not terminated. □

3.2 The case when $r > p + \beta_p$

Using Lemmas 3.3 and 3.4, we have the following result, which was inspired by and generalizes the result in [11, Lemma 4].

Lemma 3.5. Let $r > p + \beta_p$ and assume A.1. While Algorithm 2.1 has not terminated, if

$$\sigma_k \geq \max\left\{\theta, \kappa_1 \epsilon^{\frac{p+\beta_p-r}{p+\beta_p-1}}\right\}, \tag{3.9}$$

where

$$\kappa_1 \stackrel{\rm def}{=} \left(3^{r-p-\beta_p}\kappa_2^{r-1}\right)^{\frac{1}{p+\beta_p-1}} \quad \text{and } \kappa_2 \text{ is defined in (3.8)}, \tag{3.10}$$

then (3.7) holds, and so iteration k is very successful.

Proof. While Algorithm 2.1 does not terminate, we have $\pi_f(x_k+s_k) \geq \epsilon$. Assume that (3.7) does not hold on iteration k, and so

$$\sigma_k \|s_k\|^{r-p-\beta_p} < \kappa_2. \tag{3.11}$$

Then (3.5), $\sigma_k \geq \theta$ and $\pi_f(x_k+s_k) \geq \epsilon$ imply

$$\epsilon \leq \frac{L_p}{\beta_p+1}\|s_k\|^{p+\beta_p-1} + 2\sigma_k\|s_k\|^{r-1} \leq \|s_k\|^{p+\beta_p-1}\left(\frac{L_p}{\beta_p+1} + 2\sigma_k\|s_k\|^{r-p-\beta_p}\right),$$

and further using (3.11), $r > p$ and $\eta_2 \in (0,1)$, and that $r > p+\beta_p > 1$,

$$\epsilon < \|s_k\|^{p+\beta_p-1}\left(\frac{L_p}{\beta_p+1} + 2\kappa_2\right) < \left(\frac{\kappa_2}{\sigma_k}\right)^{\frac{p+\beta_p-1}{r-p-\beta_p}}\left(\frac{L_p}{\beta_p+1} + 2\kappa_2\right) < \left(\frac{\kappa_2}{\sigma_k}\right)^{\frac{p+\beta_p-1}{r-p-\beta_p}} \cdot (3\kappa_2).$$

The latter inequality implies $\sigma_k < \kappa_1 \epsilon^{\frac{p+\beta_p-r}{p+\beta_p-1}}$, which contradicts (3.9). Thus (3.7) must hold and Lemma 3.4 implies that $\rho_k \geq \eta_2$ and (2.9) hold, and so k is very successful. □

Lemma 3.6. Let $r > p+\beta_p$ and assume A.1. Then, while Algorithm 2.1 has not terminated, we have

$$\sigma_k \leq \max\left\{\sigma_0, \gamma_2\theta, \gamma_2\kappa_1\epsilon^{\frac{p+\beta_p-r}{p+\beta_p-1}}\right\}, \tag{3.12}$$

where κ_1 is defined in (3.10).

Proof. Let the right-hand side of (3.9) be denoted by $\overline{\sigma}$. It follows from Lemma 3.5 and the mechanism of the algorithm that

$$\sigma_k \geq \overline{\sigma} \quad \Longrightarrow \quad \sigma_{k+1} \leq \sigma_k. \tag{3.13}$$

Thus, when $\sigma_0 \leq \gamma_2\overline{\sigma}$, it follows that $\sigma_k \leq \gamma_2\overline{\sigma}$, where the factor γ_2 is introduced for the case when σ_k is less than $\overline{\sigma}$ and the iteration k is not very successful. Letting $k=0$ in (3.13) gives (3.12) when $\sigma_0 \geq \gamma_2\overline{\sigma}$ since $\gamma_2 > 1$. □

We are ready to establish an upper bound on the number of successful iterations until termination.

Theorem 3.7. Let $r > p + \beta_p$, assume A.1 and that $\{f(x_k)\}$ is bounded below by $f_{\rm low}$ and $\epsilon \in (0,1]$. Then for all successful iterations k until the termination of Algorithm 2.1, we have

$$f(x_k) - f(x_{k+1}) \geq \kappa_{s,p}\epsilon^{\frac{p+\beta_p}{p+\beta_p-1}}, \tag{3.14}$$

where

$$\kappa_{s,p} \stackrel{\rm def}{=} \frac{\eta_1}{r}\left(\frac{\alpha^r}{\sigma_{\max}}\right)^{\frac{1}{r-1}}, \quad \sigma_{\max} \stackrel{\rm def}{=} \max\{\sigma_0, \gamma_2\theta, \gamma_2\kappa_1\}, \tag{3.15}$$

and κ_1 is defined in (3.10). Thus Algorithm 2.1 takes at most

$$\left\lfloor \frac{f(x_k) - f_{\rm low}}{\kappa_{s,p}}\epsilon^{-\frac{p+\beta_p}{p+\beta_p-1}} \right\rfloor \tag{3.16}$$

successful iterations/evaluations of derivatives of degree 2 and above of f until termination.

Proof. On every successful iteration k, we have $\rho_k \geq \eta_1$; this and Lemma 3.1 imply

$$f(x_k) - f(x_{k+1}) \geq \eta_1(f(x_k) - T_p(x_k, s_k)) \geq \eta_1\frac{\sigma_k}{r}\|s_k\|^r. \tag{3.17}$$

On every successful iteration k we also have that (2.9) holds. Thus, applying the latter inequality twice, we deduce

$$f(x_k) - f(x_{k+1}) \geq \frac{\eta_1}{r}(\sigma_k\|s_k\|^{r-1})\|s_k\| \geq \frac{\eta_1}{r}\alpha\epsilon\|s_k\| \geq \frac{\eta_1}{r}\alpha\epsilon\left(\frac{\alpha\epsilon}{\sigma_k}\right)^{\frac{1}{r-1}} = \frac{\eta_1}{r}\frac{(\alpha\epsilon)^{\frac{r}{r-1}}}{\sigma_k^{\frac{1}{r-1}}}. \tag{3.18}$$

We use that $\epsilon \in (0,1]$ in (3.12) to deduce that

$$\sigma_k \leq \sigma_{\max}\epsilon^{\frac{p+\beta_p-r}{p+\beta_p-1}}, \tag{3.19}$$

where $\sigma_{\max}$ is defined in (3.15). We combine this upper bound with (3.18) to see that

$$f(x_k) - f(x_{k+1}) \geq \frac{\eta_1}{r}(\alpha\epsilon)^{\frac{r}{r-1}}\sigma_{\max}^{-\frac{1}{r-1}}\epsilon^{\frac{r-p-\beta_p}{(p+\beta_p-1)(r-1)}} = \frac{\eta_1}{r}\left(\frac{\alpha^r}{\sigma_{\max}}\right)^{\frac{1}{r-1}} \cdot \epsilon^{\frac{p+\beta_p}{p+\beta_p-1}},$$

which gives (3.14). Using that $f(x_k) = f(x_{k+1})$ on unsuccessful iterations, and that $f(x_k) \geq f_{\rm low}$ for all k, we can sum up over all successful iterations to deduce (3.16). □

We are left with counting the number of unsuccessul iterations until termination, and the total iteration and evaluation upper bound.

Lemma 3.8. Let $r > p + \beta_p$ and $\epsilon \in (0,1]$. Then, for any fixed $j \geq 0$ until termination, Algorithm 2.1 satisfies

$$|\mathcal{U}_j| \leq \frac{|\log\gamma_3|}{\log\gamma_1}|\mathcal{S}_j| + \frac{1}{\log\gamma_1}\log\frac{\sigma_{\max}}{\sigma_0} + \frac{r-p-\beta_p}{(p+\beta_p-1)\log\gamma_1}|\log\epsilon|, \tag{3.20}$$

where $\sigma_{\max}$ is defined in (3.15).

Proof. We apply Lemma 2.1. To prove (3.20), we use $\epsilon \in (0,1]$ and the upper bound (3.19) in place of σ_{up} in (2.11). □

Corollary 3.9. Let $r > p+\beta_p$ and assume A.1, that $\{f(x_k)\}$ is bounded below by f_{low} and $\epsilon \in (0,1]$. Then Algorithm 2.1 takes at most

$$\left\lfloor \frac{f(x_k)-f_{\text{low}}}{\kappa_{s,p}}\left(1+\frac{|\log\gamma_3|}{\log\gamma_1}\right)\epsilon^{-\frac{p+\beta_p}{p+\beta_p-1}} + \frac{r-p-\beta_p}{(p+\beta_p-1)\log\gamma_1}|\log\epsilon| + \frac{1}{\log\gamma_1}\log\frac{\sigma_{\max}}{\sigma_0}\right\rfloor \tag{3.21}$$

iterations/evaluations of f and its derivatives until termination, where $\kappa_{s,p}$ and $\sigma_{\max}$ are defined in (3.15).

Proof. The proof follows from Theorem 3.7 and (3.20), where we let j denote the first iteration with $\pi_f(x_j+s_j) < \epsilon$ (so the iteration where ARp terminates) and we use $j = |\mathcal{S}_j| + |\mathcal{U}_j|$. □

We note that the lower bound on σ_k, $\sigma_k \geq \sigma_{\min} > 0$ for all k, imposed in (2.10), has not been employed in the above proofs. It seems that in the case $r \geq p+\beta_p$, such a lower bound on σ_k may follow implicitly from (2.9). However, the requirement involving $\sigma_{\min}$ is needed for the case $r < p+\beta_p$.

3.3 The case when $p < r \leq p+\beta_p$

Note that $p < r \leq p+\beta_p$ imposes that $\beta_p > 0$ in this case. Also, note that the proof of Lemma 3.5 fails to hold for $r \leq p+\beta_p$. Thus we need a different approach here to upper bounding σ_k. In particular, we need the following additional assumption (for the case when $r < p+\beta_p$).

A.2 For $j \in \{1,\dots,p\}$, the derivative $\{\nabla^j f(x_k)\}$ is uniformly bounded above with respect to k, namely,

$$\|\nabla^j f(x_k)\| \leq M_j \text{ for all } k \geq 0, \quad j \in \{1,\dots,p\}.$$

We let $M \overset{\text{def}}{=} \max\limits_{1\leq j\leq p}\left\{\left(\frac{rp}{j!\sigma_{\min}}M_j\right)^{\frac{1}{r-j}}\right\}$ where $\sigma_{\min}$ is defined in (2.10).

Lemma 3.10. Let $r \leq p+\beta_p$ and assume A.1. If $r < p+\beta_p$ assume also A.2. While Algorithm 2.1 has not terminated, if

$$\sigma_k \geq \max\left\{\theta, \kappa_2 M^{p+\beta_p-r}\right\}, \tag{3.22}$$

where κ_2 and M are defined in (3.8) and A.2, respectively, then (3.7) holds, and so iteration k is very successful.

Proof. If $r = p + \beta_p$, then (3.22) clearly implies (3.7) and so Lemma 3.4 applies.

If $r < p+\beta_p$, then we upper bound $\|s_k\|$ by using A.2 in (3.2), as well as $\sigma_k \geq \sigma_{\min}$, to deduce that $\|s_k\| \leq M$ where M is defined in A.2. Now (3.22) implies (3.7) and so Lemma 3.4 again applies, yielding that iteration k is very successful. □

We are ready to bound σ_k from above for all iterations.

Lemma 3.11. Let $r \leq p + \beta_p$ and assume A.1. If $r < p + \beta_p$ assume also A.2. While Algorithm 2.1 has not terminated, we have

$$\sigma_k \leq \max\left\{\sigma_0, \gamma_2\theta, \gamma_2\kappa_2 M^{p+\beta_p-r}\right\} \stackrel{\rm def}{=} \sigma_{\rm up}, \tag{3.23}$$

where κ_2 and M are defined in (3.8) and A.2, respectively.

Proof. The proof follows a similar argument to that of Lemma 3.6, with (3.9) replaced by (3.22). □

We are now ready to upper bound the number of successful iterations of Algorithm 2.1 until termination.

Theorem 3.12. Let $r \leq p + \beta_p$, assume A.1 and that $\{f(x_k)\}$ is bounded below by $f_{\rm low}$. If $r < p + \beta_p$ assume also A.2. Then for all successful iterations k until the termination of Algorithm 2.1, we have

$$f(x_k) - f(x_{k+1}) \geq \kappa_{s,r}\epsilon^{\frac{r}{r-1}}, \tag{3.24}$$

where

$$\kappa_{s,r} \stackrel{\rm def}{=} \frac{\eta_1}{r}\left(\frac{\alpha^r}{\sigma_{\rm up}}\right)^{\frac{1}{r-1}}, \tag{3.25}$$

and $\sigma_{\rm up}$ is defined in (3.23). Thus Algorithm 2.1 takes at most

$$\left\lfloor \frac{f(x_k) - f_{\rm low}}{\kappa_{s,r}}\epsilon^{-\frac{r}{r-1}} \right\rfloor \tag{3.26}$$

successful iterations/evaluations of derivatives of degree 2 and higher of f until termination.

Proof. Note that (3.17) and (3.18) continue to hold in this case (they only use general ARp properties and the mechanism of the algorithm). Applying (3.23) in (3.18), we deduce

$$f(x_k) - f(x_{k+1}) \geq \frac{\eta_1}{r}(\alpha\epsilon)^{\frac{r}{r-1}}\sigma_{\rm up}^{-\frac{1}{r-1}} = \frac{\eta_1}{r}\left(\frac{\alpha^r}{\sigma_{\rm up}}\right)^{\frac{1}{r-1}} \cdot \epsilon^{\frac{r}{r-1}}, \tag{3.27}$$

which gives (3.24).

Algorithm	$p < r \leq p + \beta_p$	$p + \beta_p < r$
ARp with $p = 1$	$\mathcal{O}\left(\epsilon^{-\frac{r}{r-1}}\right) = \left[\mathcal{O}\left(\epsilon^{-\frac{1+\beta_1}{\beta_1}}\right), \infty\right)$	$\mathcal{O}\left(\epsilon^{-\frac{1+\beta_1}{\beta_1}}\right)$
ARp with $p = 2$	$\mathcal{O}\left(\epsilon^{-\frac{r}{r-1}}\right) = \left[\mathcal{O}\left(\epsilon^{-\frac{2+\beta_2}{1+\beta_2}}\right), \mathcal{O}\left(\epsilon^{-2}\right)\right)$	$\mathcal{O}\left(\epsilon^{-\frac{2+\beta_2}{1+\beta_2}}\right)$
ARp with $p = 3$	$\mathcal{O}\left(\epsilon^{-\frac{r}{r-1}}\right) = \left[\mathcal{O}\left(\epsilon^{-\frac{3+\beta_3}{2+\beta_3}}\right), \mathcal{O}\left(\epsilon^{-\frac{3}{2}}\right)\right)$	$\mathcal{O}\left(\epsilon^{-\frac{3+\beta_3}{2+\beta_3}}\right)$
...	...	...
ARp with $p \geq 2$	$\mathcal{O}\left(\epsilon^{-\frac{r}{r-1}}\right) = \left[\mathcal{O}\left(\epsilon^{-\frac{p+\beta_p}{p+\beta_p-1}}\right), \mathcal{O}\left(\epsilon^{-\frac{p}{p-1}}\right)\right)$	$\mathcal{O}\left(\epsilon^{-\frac{p+\beta_p}{p+\beta_p-1}}\right)$

Table 4.1: Summary of complexity bounds for regularization methods for ranges of r. Recall we assumed that $\epsilon \in (0, 1]$, $r > p \geq 1$, $r \in \mathbb{R}$ and $p \in \mathbb{N}$; and that either $p \geq 1$ and $\beta_p \in (0, 1]$, or $p \geq 2$ and $\beta_p \in [0, 1]$. Also, the ranges in the second column are as a function of the dominating terms in ϵ and varying r in the appropriate interval and they are plotting the changing bound $\mathcal{O}(\epsilon^{\frac{r}{r-1}})$.

Using that $f(x_k) = f(x_{k+1})$ on unsuccessful iterations, and that $f(x_k) \geq f_{\text{low}}$ for all k, we can sum up over all successful iterations to deduce (3.26). □

We are left with counting the number of total iterations and evaluations.

Corollary 3.13. Let $r \leq p + \beta_p$, assume A.1 and that $\{f(x_k)\}$ is bounded below by f_{low}. If $r < p + \beta_p$ assume also A.2. Then Algorithm 2.1 takes at most

$$\left\lfloor \frac{f(x_k) - f_{\text{low}}}{\kappa_{s,r}} \left(1 + \frac{|\log \gamma_3|}{\log \gamma_1}\right) \epsilon^{-\frac{r}{r-1}} + \frac{1}{\log \gamma_1} \log \frac{\sigma_{\text{up}}}{\sigma_0} \right\rfloor \tag{3.28}$$

iterations/evaluations of f and its derivatives until termination, where $\kappa_{s,r}$ and σ_{up} are defined in (3.26) and (3.23), respectively.

Proof. We first upper bound the total number of unsuccessful iterations; for this, we apply Lemma 2.1 to upper bound $|\mathcal{U}_j|$ with σ_{up} defined in (3.23). To prove (3.28), use (3.26) and (2.11), where we let j denote the first iteration with $\pi_f(x_j + s_j) < \epsilon$ (so the iteration where ARp terminates), and we use $j = |\mathcal{S}_j| + |\mathcal{U}_j|$. □

4 Discussion of complexity bounds

Table 4 gives a summary of our complexity bounds as a function of r and q.

Several remarks and comparisons are in order concerning these bounds.

- **The first-order case.** Note that the case $p = 1$ is also covered, with a more general quadratic model and using a Cauchy analysis, in [7]; the same complexity bounds ensue (as a function of the accuracy) as in Table 4 for $p = 1$; the case $\beta_1 = 0$ is also not covered in [7].

- **Sharpness.** The bound for the case $p = 1$ and $r \geq 1 + \beta_1$ was shown to be sharp in [7]. Also, the bounds for ARp with $p = 2$ and $2 < r \leq 2 + \beta_2$ are sharp and optimal for the corresponding smoothness classes [6].

- **Continuity.** All bounds vary continuously with r and $\beta_p \in [0, 1]$. In particular, when $r = p + \beta_p$, the complexity bounds in the second and third column match (for a given p and β_p).

- **Universality [11, 13, 15].** For fixed p and β_p, the best complexity bounds are obtained when $r \geq p + \beta_p$. These bounds do not depend on the regularization power r, and even though the smoothness parameter β_p is (usually) unknown, its value is captured accurately in the complexity, even for the case when $\beta_p = 0$ and $p \geq 2$. Note that the values of the complexity bounds as a function of the accuracy indicate that one should choose $r \geq p + 1$ to achieve the best complexity when β_p is unknown; and there seems to be little reason, from an evaluation complexity point of view, to pick anything other than $r = p + 1$.

- **Complexity values in the order of the accuracy.** Table 4 shows the increasingly good complexity obtained as p grows and $\beta_p \in (0, 1]$, namely, the more derivatives are available and the smoother these derivatives are. In particular, purely as a function of ϵ and as r varies, we obtain the following ranges of complexity powers : $[\epsilon^{-2}, \infty)$ $(p = 1)$; $[\epsilon^{-\frac{3}{2}}, \epsilon^{-2}]$ $(p = 2)$; $[\epsilon^{-\frac{4}{3}}, \epsilon^{-\frac{3}{2}}]$ $(p = 3)$; $[\epsilon^{-\frac{5}{4}}, \epsilon^{-\frac{4}{3}}]$ $(p = 4)$; and so on.

- **Loss of smoothness** Note that for fixed $p \geq 2$, $\beta_p = 0$ corresponds to the case when the objective has the highest level of non-smoothness compared to $\beta_p \in (0, 1]$. Then ARp can still be applied, and the good complexity bounds for the case $r \geq p + \beta_p \geq 2$ hold.

5 Conclusions

We have generalized and modified the regularization methods in [3] to allow for varying regularization power, accuracy of Taylor polynomials and different (Hölder) smoothness levels of derivatives. Our results show the robustness of the evaluation complexity bounds with respect to such perturbations. We found that complexity bounds of regularization methods improve with growing accuracy of the Taylor models and increasing smoothness levels of the objective. Furthermore, when the regularization power r is sufficiently large (say $r \geq p + 1$) our modification to ARp in the spirit of [11] allows ARp's worst-case behaviour to be independent of the regularization power and to accurately reflect the (often unknown) smoothness level of the objective. We have also generalized [3] and [11] to problems with convex constraints and inexact subproblem solutions. The question as to whether the complexity bounds we obtained are sharp remains open when $r \neq p + \beta_p$ and $p \geq 3$. This question is particularly poignant in the case when $p < r < p + \beta_p$: could a suitable modification of ARp achieve an (improved) evaluation complexity bound that is independent of the regularization power in this case as well?

References

[1] Alain Bensoussan and Jens Frehse. *Regularity results for nonlinear elliptic systems and applications.* Springer Verlag, Heidelberg, Berlin, New York, 2002.

[2] D.P.Bertsekas. *Nonlinear Programming.* Athena Scientific, Belmont, Massachusetts, USA, 2nd edition, 1999.

[3] 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. Technical Report naXys-05-2015, Namur Center for Complex Systems (naXys), University of Namur, Namur, Belgium, 2015.

[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. Adaptive cubic overestimation methods for unconstrained optimization. Part II: worst-case function-evaluation complexity. *Mathematical Programming, Series A*, 130(2):295–319, 2011.

[6] C. Cartis, N. I. M. Gould and Ph. L. Toint. Optimal Newton-type methods for nonconvex smooth optimization problems. ERGO Technical Report 11-009, School of Mathematics, University of Edinburgh, 2011.

[7] C. Cartis, N. I. M. Gould and Ph. L. Toint. Worst-case evaluation complexity of regularization methods for smooth unconstrained optimization using Holder continuous gradients. Technical Report 1874, Maths EPrints Archive, Mathematical Institute, University of Oxford, 2014.

[8] A.R. Conn, N.I.M. Gould and Ph. L. Toint. *Trust region methods.* MOS-SIAM series on Optimization, 2000.

[9] O. Devolder. Exactness, Inexactness and Stochasticity in First-Order Methods for Large-Scale Convex Optimization. PhD Thesis, ICTEAM and CORE, Université Catholique de Louvain, 2013.

[10] Gas Processors and Suppliers Association. *Engineering Data Book. Vol. 2.* GPSA, Tulsa, USA, 1994.

[11] G. N. Grapiglia and Yu. Nesterov. Globally-convergent second-order schemes for minimizing twice-differentiable functions. CORE Discussion Paper 2016/28, Université Catholique de Louvain, Louvain, Belgium, 2016.

[12] 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.

[13] A. S. Nemirovski and D. B. Yudin. *Problem Complexity and Method Efficiency in Optimization* Wiley Interscience Series in Discrete Mathematics, 1983.

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

[15] Yu. Nesterov. Universal gradient methods for convex optimization problems. Technical Report DP 2013/26140, CORE, Catholic University of Louvain, Louvain-la-Neuve, Belgium, 2013.

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

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