Science & Technology Facilities Council

Preprint

RAL-P-2014-014

Worst-case evaluation complexity of regularization methods for smooth unconstrained optimization using Hölder continuous gradients

C Cartis, NIM Gould, PL Toint

December 2014

Submitted for publication in Optimization Methods and Software

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

Worst-case evaluation complexity of regularization methods for smooth unconstrained optimization using Hölder continuous gradients

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

ABSTRACT

The worst-case behaviour of a general class of regularization algorithms is considered in the case where only objective function values and associated gradient vectors are evaluated. Upper bounds are derived on the number of such evaluations that are needed for the algorithm to produce an approximate first-order critical point whose accuracy is within a user-defined threshold. The analysis covers the entire range of meaningful powers in the regularization term as well as for the Hölder exponent for the gradient. The resulting complexity bounds vary according to the regularization power and the assumed Hölder exponent, recovering known results when available.

---

[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/I013067/1.

[5] Namur Center for Complex Systems (naXys) and Department of Mathematics, 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

20 Nov 2014

1 Introduction

The complexity analysis of algorithms for smooth, possibly non-convex, unconstrained optimization has been the subject of a burgeoning literature over the past few years (see the contributions by Nesterov [8, 10], Gratton, Sartenaer and Toint [7], Cartis, Gould and Toint [1, 2, 3, 4], Ueda [11], Ueda and Yamashita[12, 13, 14], Grapiglia, Yuan and Yuan [5, 6], and Vicente [15], for instance). The present contribution belongs to this active trend and focuses on the analysis of the worst-case behaviour of regularization methods where only objective function values and associated gradient vectors are evaluated. It proposes upper bounds on the number of such evaluations that are needed for the algorithm to produce an approximate first-order critical point whose accuracy is within a user-defined threshold.

An analysis of this type is already available for the case where the objective function's gradient is assumed to be Lipschitz-continuous and where the regularization uses the second or third power of the norm of the computed step at a given iteration (see the paper by Nesterov [9] for the former and those of Cartis *et al.* [2, 3] for both cases). The novelty of the present approach is to extend the analysis to cover problems whose objective gradients are simply Hölder continuous and methods that allow weaker regularization than in the Lipschitz case. The resulting complexity bounds vary according to the regularization power and the assumed Hölder exponent, providing a unified view and recovering known results when available.

The paper is organized as follows. Section 2 presents the problem and the class of algorithms considered. The complexity analysis itself is given in Section 3 and briefly discussed in Section 4.

2 The problem and algorithm

We consider the problem of finding an approximate solution of the optimization problem

$$\min_x f(x) \tag{2.1}$$

where $x \in \mathbb{R}^n$ is the vector of optimization variables and f is a function from $\mathbb{R}^n$ into $\mathbb{R}$ that is assumed to be bounded below and continuously differentiable with Hölder continuous gradients. If we denote $g(x) \stackrel{\rm def}{=} \nabla_x f(x)$, the latter says that the inequality

$$\|g(x) - g(y)\| \leq L_\beta \|x - y\|^\beta \tag{2.2}$$

holds for all $x, y \in \mathbb{R}^n$, where $L_\beta \geq 0$ and $\beta > 0$ are constants independent of x and y and where $\|\cdot\|$ is the Euclidean norm on $\mathbb{R}^n$. As explained in Lemma 3.1 below, we will assume, without loss of generality, that $\beta \leq 1$. In our context, an approximate solution for problem (2.1) is a vector x_ϵ such that

$$\|g(x_\epsilon)\| \leq \epsilon \quad \text{or} \quad f(x) \leq f_{\rm target} \tag{2.3}$$

where $\epsilon > 0$ is a user-specified accuracy threshold and $f_{\rm target} \leq f(x_0)$ is a threshold value under which the reduction of the objective function is deemed sufficient by the user. The first case in (2.3) corresponds to finding an approximate first-order-critical point. If a suitable value for $f_{\rm target}$ is not known, minus infinity can be used instead, in effect reducing (2.3) to its first part.

The class of regularization methods that we consider for computing an x satisfying (2.3) consists of iterative algorithms where, at each iteration, a local (linear or quadratic) model of f around the current iterate x_k is constructed, regularized by a term using the p-th power of the norm of the step, and then approximately minimized (in the "Cauchy point" sense) to provide a trial step s_k. The quality of this step is then measured in order to accept the resulting trial point $x_k + s_k$ as the next iterate, or to reject it and adjust the strength of the regularization.

More specifically, a regularized model of $f(x_k + s)$ of the form

$$m_k(x_k + s) = f(x_k) + g_k^T s + \tfrac{1}{2} s^T B_k s + \frac{\sigma_k}{p}\|s\|^p \tag{2.4}$$

is considered around the k-th iterate x_k, where we have defined $g_k \stackrel{\rm def}{=} g(x_k)$, where B_k is a symmetric $n \times n$ matrix, where $\sigma_k > 0$ is the regularization parameter at iteration k and where $p > 1$ is the (iteration independent) regularization power. This model is the approximately minimized in the sense that the trial step s_k is computed such that

$$m_k(x_k + s_k) \leq m_k(x_k + s_k^C), \tag{2.5}$$

where the "Cauchy step" s_k^C is defined by

$$s_k^C = -\alpha_k^C g_k \;\; \text{with} \;\; \alpha_k^C = \arg\min_{\alpha \geq 0} m_k(x_k - \alpha g_k). \tag{2.6}$$

We will choose the regularization power p in (2.4) in order to guarantee that m_k is bounded below and grows at infinity, thereby ensuring that (2.6) is well-defined. In particular, this imposes the restriction $p > 1$ and furthermore

$$p > 2 \;\; \text{whenever } B_k \text{ is allowed to not be positive semi-definite.} \tag{2.7}$$

Notice that (2.5) and (2.6) together imply that

$$m_k(x_k + s_k) \leq m_k(x_k + s_k^C) < f(x_k) \tag{2.8}$$

provided $g(x_k) \neq 0$. We may now describe our class of algorithms more formally as Algorithm 2.1 on the facing page.

Iterations of Algorithm 2.1 where $\rho_k \geq \eta_1$ are called "successful" and their index set is denoted by $\mathcal{S}$. Note that the mechanism of the algorithm ensures that $\sigma_k > 0$ for all $k \geq 0$. Note also that each iteration of the algorithm involves a single evaluation of the objective function and (for successful iterations only) of its gradient. The evaluation complexity can therefore be carried out by measuring how many *iterations* are needed before an approximate first-order critical point is found or the objective value decreases below the required target.

Algorithm 2.1: A Class of First-Order Adaptive Regularization Methods

Step 0: Initialization. An initial point x_0 , a target objective function value $f_{\rm target} \leq f(x_0)$ and an initial regularization parameter $\sigma_0 > 0$ are given, as well as an accuracy level ϵ. The constants η_1, η_2, γ_1, γ_2 and γ_3 are also given and satisfy

$$0 < \eta_1 \leq \eta_2 < 1 \ \text{ and } \ 0 < \gamma_1 < 1 < \gamma_2 < \gamma_3. \tag{2.9}$$

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

Step 1: Test for termination. If $\|g_k\| \leq \epsilon$ or $f(x_k) \leq f_{\rm target}$, terminate with the approximate solution $x_\epsilon = x_k$.

Step 2: Step calculation. Compute the step s_k approximately by minimizing the model (2.4) in the sense that conditions (2.5) and (2.6) hold.

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

$$\rho_k = \frac{f(x_k) - f(x_k + s_k)}{m_k(x_k) - m_k(x_k + s_k)}. \tag{2.10}$$

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

Step 4: Regularization parameter update. Set

$$\sigma_{k+1} \in \begin{cases} [\gamma_1\sigma_k, \sigma_k] & \text{if } \ \rho_k \geq \eta_2, \\ [\sigma_k, \gamma_2\sigma_k] & \text{if } \ \rho_k \in [\eta_1, \eta_2), \\ [\gamma_2\sigma_k, \gamma_3\sigma_k] & \text{if } \ \rho_k < \eta_1. \end{cases} \tag{2.11}$$

Increment k by one and go to Step 1.

3 Worst-case evaluation complexity analysis

In order to analyze the worst-case complexity of Algorithm 2.1, we need to specify our assumptions.

AS.1 The objective function f is continuously differentiable on $\mathbb{R}^n$.

AS.2 $g = \nabla_x f$ is Hölder continuous in the sense that (2.2) holds for all $x, y \in \mathbb{R}^n$ and some constants $L_\beta \geq 0$ and $\beta > 0$.

AS.3 There exists a constant $f_{\rm low}$ (possibly equal to minus infinity) such that, for all $x \in \mathbb{R}^n$,

$$f(x) \geq f_{\rm low} \quad \text{and} \quad f_* \stackrel{\rm def}{=} \max[f_{\rm low}, f_{\rm target}] > -\infty$$

AS.4 There exists constants $\kappa_{gl} \geq 0$ and $\kappa_{gu} > 0$ such that

$$\kappa_{gl} \leq \|g(x)\| \leq \kappa_{gu} \text{ for all } x \in \mathbb{R}^n \text{ such that } \quad f_* \leq f(x) \leq f(x_0).$$

AS.5 There exists a constant $\kappa_B \geq 0$ such that, for all $k \geq 0$,

$$\|B_k\| \leq \kappa_B.$$

AS.1 and AS.2 formalize our framework, as described in the introduction while AS.5 is standard in similar contexts and avoids possibly infinite curvature of the model, which would make the regularization irrelevant. AS.3 states that, if no target value is specified by the user, then there must exists a global lower bound on the objective function's values to make the minimization problem meaningful. The role of AS.4 is discussed below, but we immediately note that, when $f_* = f_{\text{target}} > f_{\text{low}}$, it may well happen that no single $x \in \mathbb{R}^n$ satisfies both conditions in (2.3), and thus that that the first termination criterion in (2.3) cannot be satisfied by our minimization algorithm before the second. We take this possibility into account by allowing $\kappa_{gl} > 0$, and expresssing the complexity results in terms of

$$\epsilon_* \stackrel{\text{def}}{=} \max[\,\epsilon,\, \kappa_{gl}] \tag{3.1}$$

which is the "attainable" gradient accuracy for the problem.

We start by deriving consequences of our assumptions, which are independent of the algorithm. The first is intended to explore the consequence of a value of β exceeding 1.

Lemma 3.1. Suppose that AS.1 holds and that AS.2 holds for some $\beta > 1$. Then f is linear in $\mathbb{R}^n$, AS.2 holds for all $\beta > 0$ with $L_\beta = 0$ and AS.4 holds with $\kappa_{gl} = \kappa_{gu} = \|g(x_0)\|$.

Proof. If e_i is the i-th vector of the canonical basis, we have, using the Cauchy-Schwarz inequality and the Hölder condition (2.2), that, for all $i = 1, \ldots, n$ and all $x \in \mathbb{R}^n$,

$$\frac{|(g_i(x + te_i) - g_i(x)|}{|t|} \leq \frac{\|g(x + te_i) - g(x)\|}{\|x + te_i - x\|} \leq L_\beta |t|^{\beta - 1}$$

and $\beta - 1 > 0$. Taking the limit when $t \to 0$ gives that the directional derivative of each g_i exists and is zero for all i and all x. Thus the gradient is constant in $\mathbb{R}^n$, f is linear and AS.2 obviously holds with $L_\beta = 0$ for all $\beta > 0$ since $\|g(x) - g(y)\|$ is identically zero for all $x, y \in \mathbb{R}^n$. □

This justifies our choice to restrict our attention to the case where $\beta \in (0, 1]$ for the rest of our analysis. The second result indicates common circumstances in which AS.4 holds.

Lemma 3.2. Suppose that AS.1, AS.2 and that there exists a constant $f_{\text{low}} > -\infty$ such that

$$f(x) \geq f_{\text{low}} \tag{3.2}$$

for all $x \in \mathcal{L}_0 \stackrel{\text{def}}{=} \{x \in \mathbb{R}^n \mid f(x) \leq f(x_0)\}$ for all $x \in \mathbb{R}^n$. Then AS.4 holds.

Proof. Let $x \in \mathcal{L}_0$. AS.1, the mean-value theorem, and AS.2 then ensure that, for all s,

$$\begin{aligned} f_{\text{low}} &\leq f(x+s) \\ &\leq f(x) + g(x)^T s + \int_0^1 (g(x+\xi s) - g(x))^T s \, d\xi \\ &\leq f(x) + g(x)^T s + \frac{L_\beta}{1+\beta}\|s\|^{\beta+1} \stackrel{\text{def}}{=} h(s) \end{aligned} \tag{3.3}$$

Given that the minimizer of the convex function $h(s)$ is given by

$$s_* = -\frac{g(x)}{L_\beta^{1/\beta}\|g(x)\|^{\frac{1-\beta}{\beta}}},$$

we obtain that

$$\min_s h(s) = h(s_*) = f(x) - \frac{\beta L_\beta^{-\frac{1}{\beta}}}{1+\beta}\|g(x)\|^{1+\frac{1}{\beta}}.$$

As a consequence, we obtain, using the fact that $f(x) \leq f(x_0)$ since $x \in \mathcal{L}_0$ and (3.3), that

$$f_{\text{low}} \leq f(x_0) - \frac{\beta L_\beta^{-\frac{1}{\beta}}}{1+\beta}\|g(x)\|^{1+\frac{1}{\beta}},$$

which in turn implies that

$$\|g(x)\| \leq \left[L_\beta \left(1+\frac{1}{\beta}\right)^\beta (f(x_0) - f_{\text{low}})^\beta\right]^{\frac{1}{1+\beta}} \stackrel{\text{def}}{=} \kappa_{gu},$$

yielding the desired conclusion, irrespective of the value of f_{target}. □

Note that (3.2) is indeed very common. For instance, $f_{\text{low}} = 0$ for all nonlinear least-squares problems. Hence the form of AS.4 should not be viewed as overly restrictive and also allows for the case where (3.2) fails but the objective function's gradient remains reasonably well-behaved.

We now turn to the analysis of the algorithm's properties. But, before we start in earnest, it is useful to introduce some specific notation. In a number of occurrences, we need to include some of the terms in formulae only if certain conditions apply. We will indicate this by underbracing the conditional part of the formula, the text below the

underbrace then specifying the relevant condition. For instance we may have an expression of the type

$$\max[\underbrace{a^{-1}}_{a>0}, b, c],$$

meaning that the maximum should include the first term if and only if $a > 0$ (making the term well-defined in this case).

We first derive two bounds of the step length.

Lemma 3.3. We have that, for all $k \geq 0$,

$$\|s_k\| \leq \max\left[\underbrace{\left(\frac{p}{\sigma_k}\|B_k\|\right)^{\frac{1}{p-2}}}_{B_k \not\succeq 0}, \left(\frac{2p}{\sigma_k}\|g_k\|\right)^{\frac{1}{p-1}}\right]. \tag{3.4}$$

Moreover,

$$\|s_k\| \leq \left(\frac{2p}{\sigma_k}\|g_k\|\right)^{\frac{1}{p-1}} \tag{3.5}$$

provided

$$\sigma_k \geq \frac{(p\|B_k\|)^{p-1}}{(2p\|g_k\|)^{p-2}}. \tag{3.6}$$

Proof. Observe first that (2.8) and $g_k \neq 0$ ensure that

$$m_k(x_k + s_k) - f(x_k) = g_k^T s_k + \tfrac{1}{2} s_k^T B_k s_k + \frac{\sigma_k}{p}\|s_k\|^p < 0 \tag{3.7}$$

Assume first that $s_k^T B_k s_k > 0$. Then we must have that

$$g_k^T s_k + \frac{\sigma_k}{p}\|s_k\|^p < 0,$$

and therefore (remembering that $\sigma_k > 0$ and that $g_k^T s_k \geq -\|g_k\|\,\|s_k\|$)

$$\|s_k\| < \left(\frac{p}{\sigma_k}\|g_k\|\right)^{\frac{1}{p-1}} < \left(\frac{2p}{\sigma_k}\|g_k\|\right)^{\frac{1}{p-1}}. \tag{3.8}$$

If $s_k^T B_k s_k \leq 0$, we may rewrite (3.7) as

$$\left[g_k^T s_k + \frac{\sigma_k}{2p}\|s_k\|^p\right] + \left[\tfrac{1}{2} s_k^T B_k s_k + \frac{\sigma_k}{2p}\|s_k\|^p\right] < 0$$

and the left-hand side of this inequality can only be negative if at least one of the bracketed expressions is negative, giving that

$$\|s_k\| \leq \max\left[\left(\frac{p}{\sigma_k}\|B_k\|\right)^{\frac{1}{p-2}}, \left(\frac{2p}{\sigma_k}\|g_k\|\right)^{\frac{1}{p-1}}\right],$$

where we also used that $g_k^T s_k \geq -\|g_k\| \, \|s_k\|$ and $s_k^T B_k s_k \geq -\|B_k\| \, \|s_k\|^2$. Combining this with (3.8) then yields (3.4). Checking (3.5) subject to (3.6) is straightforward. □

We now turn to the task of finding a lower bound on the model decrease $f(x_k) - m_k(x_k + s_k)$ resulting from (2.5)-(2.6). The first step is to find a suitable positive lower bound on the step α_k^C as defined in (2.6).

Lemma 3.4. We have that

$$m_k(x_k + s_k^C) \leq m_k(x_k - \alpha_k^* g_k) < f(x_k) \tag{3.9}$$

where

$$\alpha_k^* \stackrel{\text{def}}{=} \min \left[\underbrace{\frac{\|g_k\|^2}{2 g_k^T B_k g_k}}_{g_k^T B_k g_k > 0}, \left(\frac{p}{2\sigma_k} \frac{1}{\|g_k\|^{p-2}} \right)^{\frac{1}{p-1}} \right]$$

Proof. Substituting the definition $s = -\alpha g_k$ into (2.4), we obtain from (2.5)-(2.6) that, for all $\alpha > 0$,

$$m_k(x_k - \alpha g_k) - f(x_k) = \alpha \left(-\|g_k\|^2 + \tfrac{1}{2} \alpha g_k^T B_k g_k + \frac{\sigma_k}{p} \alpha^{p-1} \|g_k\|^p \right). \tag{3.10}$$

Assume first that $g_k^T B_k g_k \leq 0$. Then

$$-\|g_k\|^2 + \frac{\sigma_k}{p} \alpha^{p-1} \|g_k\|^p < 0$$

for all $\alpha \in (0, \bar{\alpha}_k]$ where

$$\bar{\alpha}_k = \left(\frac{p}{\sigma_k} \frac{1}{\|g_k\|^{p-2}} \right)^{\frac{1}{p-1}}. \tag{3.11}$$

and, because $\alpha > 0$ and $g_k^T B_k g_k \leq 0$, we also obtain from (3.10) that $m_k(x_k - \alpha g_k) < f(x_k)$ for all $\alpha \in (0, \bar{\alpha}_k]$. In particular, this yields that $m_k(x_k - \alpha_k^* g_k) < f(x_k)$, where

$$\alpha_k^* = \left(\frac{p}{2\sigma_k} \frac{1}{\|g_k\|^{p-2}} \right)^{\frac{1}{p-1}}. \tag{3.12}$$

Condition (2.6) then ensures that (3.9) holds as desired.

Assume next that $g_k^T B_k g_k > 0$ and, in this case, define

$$\alpha_k^* \stackrel{\text{def}}{=} \min \left[\frac{\|g_k\|^2}{2 g_k^T B_k g_k}, \left(\frac{p}{2\sigma_k} \frac{1}{\|g_k\|^{p-2}} \right)^{\frac{1}{p-1}} \right] \tag{3.13}$$

Then it is easy to verify that both bracketed expressions in

$$\left[-\tfrac{1}{2}\|g_k\|^2+\tfrac{1}{2}\alpha_k^* g_k^T B_k g_k\right]+\left[-\tfrac{1}{2}\|g_k\|^2+\frac{\sigma_k}{p}(\alpha_k^*)^{p-1}\|g_k\|^p\right]=\frac{1}{\alpha_k^*}\Big(m_k(x_k-\alpha_k^* g_k)-f(x_k)\Big)$$

are negative and thus, because $\alpha_k^* > 0$, that $m_k(x_k - \alpha_k^* g_k) < f(x_k)$. The desired conclusion can now be obtained by invoking (2.6). □

We now translate the conclusions of the last lemma in terms of the model reduction at the Cauchy point and beyond.

Lemma 3.5. We have that

$$f(x_k)-m_k(x_k+s_k)\geq\frac{1}{4}\min\left[\underbrace{\frac{\|g_k\|^4}{2g_k^T B_k g_k}}_{g_k^T B_k g_k>0},\left(\frac{p}{2\sigma_k}\|g_k\|^p\right)^{\frac{1}{p-1}}\right] \tag{3.14}$$

Proof. If $g_k^T B_k g_k \leq 0$, substituting (3.12) into (3.10) immediately yields that

$$f(x_k)-m_k(x_k-\alpha_k^* g_k)\geq\left(\frac{p}{2\sigma_k}\frac{1}{\|g_k\|^{p-2}}\right)^{\frac{1}{p-1}}\left[\|g_k\|^2-\tfrac{1}{2}\|g_k\|^2\right]=\frac{1}{2}\left(\frac{p}{2\sigma_k}\|g_k\|^p\right)^{\frac{1}{p-1}}. \tag{3.15}$$

If $g_k^T B_k g_k > 0$, we have from (3.10) and (3.13) that

$$\begin{aligned}
&f(x_k)-m_k(x_k-\alpha_k^* g_k)\\
&\quad\geq\ \alpha_k^*\left[\|g_k\|^2-\frac{1}{2}\left(\frac{\|g_k\|^2}{2g_k^T B_k g_k}\right)g_k^T B_k g_k-\frac{\sigma_k}{p}\left(\frac{p}{2\sigma_k}\frac{1}{\|g_k\|^{p-2}}\right)\|g_k\|^p\right]\\
&\quad=\ \min\left[\frac{\|g_k\|^2}{2g_k^T B_k g_k},\left(\frac{p}{2\sigma_k}\frac{1}{\|g_k\|^{p-2}}\right)^{\frac{1}{p-1}}\right]\left[\|g_k\|^2-\tfrac{1}{4}\|g_k\|^2-\tfrac{1}{2}\|g_k\|^2\right]\\
&\quad=\ \frac{1}{4}\min\left[\frac{\|g_k\|^4}{2g_k^T B_k g_k},\left(\frac{p}{2\sigma_k}\|g_k\|^p\right)^{\frac{1}{p-1}}\right].
\end{aligned}$$

Combining this last inequality with (3.15) and using (2.5) then gives (3.14). □

The model decrease specified by (3.14) turns out to be useful if the value of σ_k (appearing at the denominator of the second term in the min) can be bounded above across all iterations. We obtain this result in two stages, the first being to determine conditions under which the regularized model (2.4) is an overestimation of the objective function at the trial point $x_k + s_k$.

Lemma 3.6. Suppose that AS.1 and AS.2 hold. Suppose also that

$$\sigma_k \geq \max\left[\underbrace{1, \frac{(p\|B_k\|)^{p-1}}{(2p\|g_k\|)^{p-2}}, \kappa_1\|B_k\|^{\frac{p-1}{\beta}}\|g_k\|^{\frac{2-p}{\beta}}}_{B_k \not\succeq 0}, \kappa_2\|g_k\|^{\frac{1+\beta-p}{\beta}}\right] \tag{3.16}$$

where

$$\kappa_1 \stackrel{\rm def}{=} (2^p p)^{\frac{1}{\beta}} \quad \text{and} \quad \kappa_2 \stackrel{\rm def}{=} 2p^{\frac{p-1}{\beta}}(2p)^{\frac{p+\beta-1}{\beta}}\left(\frac{L_\beta}{1+\beta}\right)^{\frac{p-1}{\beta}}. \tag{3.17}$$

Then we have that

$$f(x_k+s_k) \leq m_k(x_k+s_k). \tag{3.18}$$

Proof. First notice that AS.1 and the mean-value theorem implies that

$$f(x_k+s_k) - m_k(x_k+s_k) = \int_0^1 (g(x_k+\xi s_k)-g_k)^T s_k\, d\xi - \tfrac{1}{2}s_k^T B_k s_k - \frac{\sigma_k}{p}\|s_k\|^p.$$

Using now AS.2, we obtain that

$$f(x_k+s_k) - m_k(x_k+s_k) \leq \frac{L_\beta}{1+\beta}\|s_k\|^{1+\beta} - \tfrac{1}{2}s_k^T B_k s_k - \frac{\sigma_k}{p}\|s_k\|^p. \tag{3.19}$$

Assume first that $B \succeq 0$. Then (3.18) holds if

$$\sigma_k \geq \frac{pL_\beta}{1+\beta}\|s_k\|^{1+\beta-p},$$

which, in view of (3.4) and $B_k \succeq 0$, holds if

$$\sigma_k \geq \frac{pL_\beta}{1+\beta}\left(\frac{2p}{\sigma_k}\|g_k\|\right)^{\frac{1+\beta-p}{p-1}},$$

that is if

$$\sigma_k \geq 2p\,\frac{L_\beta}{1+\beta}^{\frac{p-1}{\beta}}\|g_k\|^{\frac{1+\beta-p}{\beta}}. \tag{3.20}$$

Assume now that $B \not\succeq 0$, in which case we cannot guarantee that $s_k^T B_k s_k \geq 0$ in (3.19). Then, using the Cauchy-Schwarz inequality, (3.18) holds if

$$\sigma_k \geq \frac{pL_\beta}{1+\beta}\|s_k\|^{1+\beta-p} + \frac{p}{2}\|B_k\|\,\|s_k\|^{2-p}.$$

If we now assume that σ_k is large enough to ensure (3.6), (3.5) ensures that (3.18) holds if

$$\sigma_k \geq \frac{2pL_\beta}{1+\beta}\left(2p\frac{\|g_k\|}{\sigma_k}\right)^{\frac{1+\beta-p}{p-1}} + \frac{p}{2}\|B_k\|\left(2p\frac{\|g_k\|}{\sigma_k}\right)^{\frac{2-p}{p-1}}$$

Observe now that $B_k \not\succeq 0$ and (2.7) imply that $\max[1, p-1] = p-1$, and hence that $1+\beta-p \leq 2-p \leq 0$, which, if we additionally assume that $\sigma_k \geq 1$, implies that $\sigma_k^{-\frac{1+\beta-p}{p-1}} \geq \sigma_k^{-\frac{2-p}{p-1}}$. Hence (3.18) holds if

$$\sigma_k \geq \frac{2pL_\beta}{1+\beta}(2p\|g_k\|)^{\frac{1+\beta-p}{p-1}}\left(\frac{1}{\sigma_k}\right)^{\frac{1+\beta-p}{p-1}} + \frac{p}{2}\|B_k\|(2p\|g_k\|)^{\frac{2-p}{p-1}}\left(\frac{1}{\sigma_k}\right)^{\frac{1+\beta-p}{p-1}}$$

which is equivalent to requiring that

$$\sigma_k^{\frac{\beta}{p-1}} \geq \frac{2pL_\beta}{1+\beta}(2p\|g_k\|)^{\frac{1+\beta-p}{p-1}} + \frac{p}{2}\|B_k\|(2p\|g_k\|)^{\frac{2-p}{p-1}}$$

Using the fact that

$$\frac{2pL_\beta}{1+\beta}(2p\|g_k\|)^{\frac{1+\beta-p}{p-1}}+\frac{p}{2}\|B_k\|(2p\|g_k\|)^{\frac{2-p}{p-1}} \leq 4p\max\left[\frac{L_\beta}{1+\beta}(2p\|g_k\|)^{\frac{1+\beta-p}{p-1}}, \|B_k\|(2p\|g_k\|)^{\frac{2-p}{p-1}}\right]$$

and taking the $\beta/(p-1)$-th root of this last inequality, we finally conclude that (3.18) holds (in the case where $B \not\succeq 0$) if the inequality

$$\sigma_k \geq (4p)^{\frac{p-1}{\beta}}\max\left[\left(\frac{L_\beta}{1+\beta}\right)^{\frac{p-1}{\beta}}(2p\|g_k\|)^{\frac{1+\beta-p}{\beta}}, \|B_k\|^{\frac{p-1}{\beta}}(2p\|g_k\|)^{\frac{2-p}{\beta}}\right] \tag{3.21}$$

holds in addition to (3.6) and $\sigma_k \geq 1$. The proof of the lemma is now completed by combining these last two additional conditions, (3.20) and (3.21). □

We are now in position to prove an iteration-independent upper bound on the value of σ_k.

Lemma 3.7. Suppose that AS.1–AS.5 hold. Then, as long as the algorithm does not terminate, and given the constants

$$\kappa_1^\sigma \stackrel{\rm def}{=} \gamma_3\max\left[1, \frac{(p\kappa_B)^{p-1}}{(2p)^{p-2}}, \kappa_1\kappa_B^{\frac{p-1}{\beta}}, \kappa_2\right] \quad \text{and} \quad \kappa_2^\sigma \stackrel{\rm def}{=} \max\left[\sigma_0, \gamma_3\kappa_2\kappa_{gu}^{\frac{1+\beta-p}{\beta}}, \underbrace{\kappa_1^\sigma}_{1+\beta<p}\right] \tag{3.22}$$

with κ_1 and κ_2 defined in (3.17), we have that

$$\sigma_k \leq \max\left[\kappa_2^\sigma, \underbrace{\kappa_1^\sigma\epsilon_*^{\frac{1+\beta-p}{\beta}}}_{1+\beta<p}\right]. \tag{3.23}$$

Proof. The mechanism of the algorithm ensures that σ_k is not increased at iteration k if $f(x_k+s_k) \leq m_k(x_k+s_k)$, which we know from Lemma 3.6 is ensured if (3.16) holds.

We now distinguish two cases. Assume first that

$$1 + \beta \geq p,$$

which in turn implies that $p \in (1, 2]$ and thus, in view of (2.7), that $B_k \succeq 0$ for all k. Then Lemma 3.2 and condition (3.16) imply that $f(x_k + s_k) \leq m_k(x_k + s_k)$ provided

$$\sigma_k \geq \kappa_2 \kappa_{gu}^{\frac{1+\beta-p}{\beta}}, \tag{3.24}$$

which is a constant independent of k and ϵ. The second, more complicated case is when

$$1 + \beta < p,$$

in which we again distinguish two subcases. The first of these subcases in when $p \in (1, 2]$, which, because of (2.7), implies that $B_k \succeq 0$ for all k. In order to derive a suitable iteration independent bound for the regularization parameter, we recall that, as long as the algorithm has not terminated, we have that $\|g_k\| > \epsilon_*$ and thus, from (3.16), that $f(x_k + s_k) \leq m_k(x_k + s_k)$ provided

$$\sigma_k \geq \kappa_2 \epsilon_*^{\frac{1+\beta-p}{\beta}}. \tag{3.25}$$

Consider now the second subcase, where $p > 2$, allowing B_k to be non positive semidefinite. Taking the inequality $\|g_k\| > \epsilon_*$ and AS.5 into account, we derive from (3.16) that $f(x_k + s_k) \leq m_k(x_k + s_k)$ provided

$$\sigma_k \geq \max\left[1, \frac{(p\kappa_B)^{p-1}}{(2p)^{p-2}} \epsilon_*^{2-p}, \kappa_1 \kappa_B^{\frac{p-1}{\beta}} \epsilon_*^{\frac{2-p}{\beta}}, \kappa_2 \epsilon_*^{\frac{1+\beta-p}{\beta}}\right] \tag{3.26}$$

Note now that, because $\beta \leq 1$,

$$0 > 2 - p \geq \frac{2-p}{\beta} \geq \frac{1+\beta-p}{\beta}. \tag{3.27}$$

We may then conclude from (3.26) and (3.27) that $f(x_k + s_k) \leq m_k(x_k + s_k)$ provided

$$\sigma_k \geq \max\left[1, \frac{(p\kappa_B)^{p-1}}{(2p)^{p-2}}, \kappa_1 \kappa_B^{\frac{p-1}{\beta}}, \kappa_2\right] \max\left[1, \epsilon_*^{\frac{1+\beta-p}{\beta}}\right]. \tag{3.28}$$

We may combine all cases together and obtain from (3.24) , (3.25) and (3.28) that $f(x_k + s_k) \leq m_k(x_k + s_k)$ if

$$\sigma_k \geq \max\left[\kappa_2 \kappa_{gu}^{\frac{1+\beta-p}{\beta}}, \underbrace{\kappa_1^\sigma/\gamma_3}_{1+\beta<p}, \underbrace{(\kappa_1^\sigma/\gamma_3)\epsilon_*^{\frac{1+\beta-p}{\beta}}}_{1+\beta<p}\right], \tag{3.29}$$

where κ_1^σ is defined in (3.22). The proof of (3.23) is then completed by taking into account that the initial parameter σ_0 may exceed the bound given by the right-hand side of (3.29), and also that (3.29) may just fail by a small margin at an unsuccessful iteration, resulting in an increase of σ_k by a factor γ_3 before (3.29) applies. □

Having now derived an iteration independent upper bound on σ_k, we may return to the model decrease given by Lemma 3.5.

Lemma 3.8. Suppose that AS.1– AS.5 hold. Then, as long as the algorithm does not terminate, and given the constant

$$\kappa_m \stackrel{\rm def}{=} \frac{1}{4}\min\Big[\frac{1}{2\kappa_B}, \left(\frac{p}{2\kappa_2^\sigma}\right)^{\frac{1}{p-1}}\Big], \tag{3.30}$$

- if $1+\beta \geq p$, then

$$f(x_k) - m(x_k+s_k) \geq \kappa_m \min\left(\epsilon_*^2,\, \epsilon_*^{\frac{p}{p-1}}\right). \tag{3.31}$$

- if $1+\beta < p$, then

$$f(x_k) - m(x_k+s_k) \geq \kappa_m \min\left(\epsilon_*^2,\, \epsilon^{\frac{p}{p-1}},\, \epsilon_*^{1+\frac{1}{\beta}}\right). \tag{3.32}$$

Proof. Assume first that $1+\beta \geq p$ This implies that $p \in [1,2]$ and hence, because of (2.7), that $g_k^T B_k g_k \geq 0$. Taking into account that, when $g_k^T B_k s_k > 0$,

$$g_k^T B_k g_k \leq \kappa_B \|g_k\|^2,$$

because of AS.5, substituting (3.23) into (3.14) and using (3.23) and the fact that $\|g_k\| \geq \epsilon_*$ as long as the algorithm has not terminated, yields that

$$\begin{aligned} f(x_k) - m_k(x_k+s_k) &\geq \tfrac{1}{4}\min\left[\frac{\epsilon_*^2}{2\kappa_B}, \left(\frac{p}{2\kappa_2^\sigma}\right)^{\frac{1}{p-1}} \epsilon_*^{\frac{p}{p-1}}\right] \\ &\geq \tfrac{1}{4}\min\left[\frac{1}{2\kappa_B}, \left(\frac{p}{2\kappa_2^\sigma}\right)^{\frac{1}{p-1}}\right] \min\left(\epsilon_*^2, \epsilon_*^{\frac{p}{p-1}}\right) \end{aligned}$$

and (3.31) follows.

Consider now the case where $1+\beta < p$. Again substituting (3.23) into (3.14), using (3.23), the fact that $\kappa_2^\sigma \geq \kappa_1^\sigma$ because of (3.22), AS.5 and the fact that $\|g_k\| \geq \epsilon_*$ as long as the algorithm has not terminated, we obtain that

$$\begin{aligned} f(x_k) - m_k(x_k+s_k) &\geq \frac{1}{4}\min\left[\underbrace{\frac{\epsilon_*^2}{2\kappa_B}}_{g_k^T B_k g_k > 0}, \left(\frac{p\,\epsilon_*^p}{2\max\left[\kappa_2^\sigma, \kappa_1^\sigma \epsilon_*^{\frac{1+\beta-p}{\beta}}\right]}\right)^{\frac{1}{p-1}}\right] \\ &\geq \frac{1}{4}\min\left[\frac{1}{2\kappa_B}, \left(\frac{p}{2\kappa_2^\sigma}\right)^{\frac{1}{p-1}}\right] \min\left(\epsilon_*^2,\, \epsilon_*^{\frac{p}{p-1}},\, \epsilon_*^{1+\frac{1}{\beta}}\right). \end{aligned}$$

which is (3.32) . □

We now recall an important technical lemma which, in effect, gives a bound on the total number of unsuccessful iterations before iteration k as a function of the number of successful ones.

Lemma 3.9. The mechanism of Algorithm 2.1 guarantees that, if

$$\sigma_k \leq \sigma_{\max}, \tag{3.33}$$

for some $\sigma_{\max} > 0$, then

$$k \leq |\mathcal{S}_k| \left(1 + \frac{|\log \gamma_1|}{\log \gamma_2}\right) + \frac{1}{\log \gamma_2} \log \left(\frac{\sigma_{\max}}{\sigma_0}\right). \tag{3.34}$$

Proof. See [2]. □

We are now ready to prove our main result on the worst-case complexity of Algorithm 2.1.

Theorem 3.10. Suppose that AS.1–AS.5 hold and define ϵ_* as in (3.1).

1. If $1 + \beta \geq p$, there exist constants κ_p^s, κ_p^a and κ_p^c such that, for any $\epsilon > 0$, Algorithm 2.1 requires at most

$$\left\lfloor \kappa_p^s \frac{f(x_0) - f_*}{\min\left(\epsilon_*^2, \epsilon_*^{\frac{p}{p-1}}\right)} \right\rfloor \tag{3.35}$$

successful iterations (and gradient evaluations),and a total of

$$\left\lfloor \kappa_p^a \frac{f(x_0) - f_*}{\min\left(\epsilon_*^2, \epsilon_*^{\frac{p}{p-1}}\right)} + \kappa_p^c \right\rfloor \tag{3.36}$$

iterations (and objective function evaluations) before producing an iterate x_ϵ such that $\|g(x_\epsilon)\| \leq \epsilon_*$ or $f(x_\epsilon) \leq f_{\text{target}}$.

2. If $1+\beta < p$, there exist constants κ^s_β, κ^a_β, κ^b_β and κ^c_β such that, for all $\epsilon > 0$, Algorithm 2.1 requires at most

$$\left\lfloor \kappa^s_\beta \frac{f(x_0) - f_*}{\min\left(\epsilon_*^2,\, \epsilon_*^{\frac{p}{p-1}},\, \epsilon_*^{1+\frac{1}{\beta}}\right)} \right\rfloor \tag{3.37}$$

successful iterations (and gradient evaluations) and a total of

$$\left\lfloor \kappa^a_\beta \frac{f(x_0) - f_*}{\min\left(\epsilon_*^2,\, \epsilon_*^{\frac{p}{p-1}},\, \epsilon_*^{1+\frac{1}{\beta}}\right)} + \kappa^b_\beta \,|\log \epsilon_*| + \kappa^c_\beta \right\rfloor \tag{3.38}$$

iterations (and objective function evaluations) before producing an iterate x_ϵ such that $\|g(x_\epsilon)\| \le \epsilon_*$ or $f(x_\epsilon) \le f_{\rm target}$.

In the above statements the constants are given by

$$\kappa^s_p = \kappa^s_\beta \stackrel{\rm def}{=} \frac{1}{\eta_1 \kappa_m}, \tag{3.39}$$

$$\kappa^a_p \stackrel{\rm def}{=} \frac{1}{\eta_1 \kappa_m}\left(1 + \frac{|\log \gamma_1|}{\log \gamma_2}\right), \quad \kappa^c_p \stackrel{\rm def}{=} \frac{1}{\log \gamma_2} \log\left(\frac{\kappa_2^\sigma}{\sigma_0}\right), \tag{3.40}$$

$$\kappa^a_\beta \stackrel{\rm def}{=} \frac{1}{\eta_1 \kappa_m}\left(1 + \frac{|\log \gamma_1|}{\log \gamma_2}\right), \quad \kappa^b_\beta \stackrel{\rm def}{=} \frac{p-\beta-1}{\beta \log \gamma_2} \tag{3.41}$$

and

$$\kappa^c_\beta \stackrel{\rm def}{=} \frac{1}{\log \gamma_2}\Big(\log\left(\max[1, \kappa_2^\sigma, \kappa_1^\sigma]\right) + |\log(\sigma_0)|\Big), \tag{3.42}$$

where

$$\kappa_1 \stackrel{\rm def}{=} (2^p p)^{\frac{1}{\beta}}, \quad \kappa_2 \stackrel{\rm def}{=} 2p^{\frac{p-1}{\beta}} (2p)^{\frac{p+\beta-1}{\beta}} \left(\frac{L_\beta}{1+\beta}\right)^{\frac{p-1}{\beta}},$$

$$\kappa_1^\sigma \stackrel{\rm def}{=} \gamma_3 \max\left[1, \frac{(p\kappa_B)^{p-1}}{(2p)^{p-2}}, \kappa_1 \kappa_B^{\frac{p-1}{\beta}}, \kappa_2\right], \quad \kappa_2^\sigma \stackrel{\rm def}{=} \max\left[\sigma_0, \gamma_3 \kappa_2 \kappa_{gu}^{\frac{1+\beta-p}{\beta}}, \underbrace{\kappa_1^\sigma}_{1+\beta<p}\right]$$

and

$$\kappa_m \stackrel{\rm def}{=} \frac{1}{4} \min\Big[\frac{1}{2\kappa_B}, \left(\frac{p}{2\kappa_2^\sigma}\right)^{\frac{1}{p-1}}\Big].$$

Proof. Consider first the case where $1+\beta \ge p$. We then deduce from (3.31) in Lemma 3.8, AS.4 and the definition of a successful iteration, that, as long as the

algorithm has not terminated,

$$\begin{aligned} f(x_0) - f_* &\geq f(x_0) - f(x_{k+1}) \\ &= \sum_{j \in \mathcal{S}_k} [f(x_j) - f(x_j + s_j)] \\ &\geq \eta_1 \sum_{j \in \mathcal{S}_k} [f(x_j) - m_j(x_j + s_j)] \\ &> \eta_1 \, \kappa_m \, \min\left(\epsilon_*^2, \epsilon_*^{\frac{p}{p-1}}\right) |\mathcal{S}_k|, \end{aligned} \tag{3.43}$$

where $|\mathcal{S}_k|$ is the cardinality of $\mathcal{S}_k \stackrel{\rm def}{=} \{j \in \mathcal{S} \mid j \leq k\}$, that is the number of successful iterations up to iteration k. This provides an upper bound on $|\mathcal{S}_k|$ which is independent of k and ϵ, from which we obtain the bound (3.35) with (3.39). Calling now upon Lemma 3.9 and (3.23), we deduce that the total number of iterations (and function evaluations) cannot exceed

$$\kappa_p^s \frac{f(x_0) - f_*}{\epsilon_*^{\frac{p}{p-1}}} \left(1 + \frac{|\log \gamma_1|}{\log \gamma_2}\right) + \frac{1}{\log \gamma_2} \log\left(\frac{\kappa_2^\sigma}{\sigma_0}\right),$$

which then gives the bound (3.36) with (3.40).

The proof for the case where $1 + \beta < p$ is derived in a manner entirely similar to that used for the case where $1 + \beta \geq p$, replacing $\epsilon^{\frac{p}{p-1}}$ by $\epsilon^{1+\frac{1}{\beta}}$ in (3.43), the use of (3.31) by that of (3.32) and taking into account that

$$\max\left[\kappa_2^\sigma, \kappa_1^\sigma \epsilon_*^{\frac{1+\beta-p}{\beta}}\right] \leq \max[\kappa_2^\sigma, \kappa_1^\sigma] \epsilon_*^{\frac{1+\beta-p}{\beta}}$$

(where κ_2^σ and κ_1^σ are defined by (3.17) and (3.22)) and thus that

$$\log\left(\frac{\max\left[\kappa_2^\sigma, \kappa_1^\sigma \epsilon_*^{\frac{1+\beta-p}{\beta}}\right]}{\sigma_0}\right) \leq \left|\frac{1+\beta-p}{\beta}\right| |\log \epsilon_*| + \log(\max[1, \kappa_2^\sigma, \kappa_1^\sigma]) + |\log(\sigma_0)|.$$

We may thus deduce that (3.37) holds with (3.39) and that (3.38) holds with (3.41)–(3.42). □

Which power of ϵ_* dominates in the complexity bounds of the theorem is summarized in Table 3.1 and illustrated in Figure 3.1.

Note that, when $\epsilon_* \leq 1$, $\epsilon^{-\frac{p}{p-1}} > \epsilon^{-(1+\frac{1}{\beta})}$ in the triangle for which $1 + \beta \geq p$ and $p \leq 2$. If $\epsilon_* > 1$ the powers of ϵ_* are bounded above by one and the complexity bounds are dominated by the difference $f(x_0) - f_*$ and the constants defined in (3.39)–(3.30).

4 Discussion

It is interesting to note that the worst-case evaluation complexity of our general class of regularized method does depend on the relative values of p and β. In the bounds derived

	$p \leq 2$	$p > 2$
$1+\beta \geq p$	$O\left(\epsilon_*^{-\frac{p}{p-1}}\right)$	-
$1+\beta < p$	$O\left(\epsilon_*^{-(1+\frac{1}{\beta})}\right)$	$O\left(\epsilon_*^{-(1+\frac{1}{\beta})}\right)$

Table 3.1: The complexity order as a function of $\epsilon_* < 1$ in the statement of Theorem 3.10

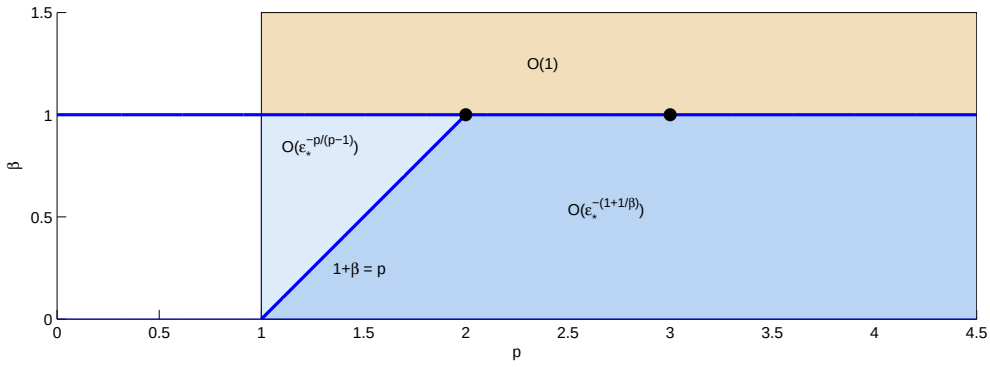

Figure 3.1: Worst-case evaluation complexity as a function of β and p in the cases where $\epsilon_* \leq 1$

in Theorem 3.10, the terms in ϵ_*^{-2} dominate only if $\epsilon_* > 1$ that is if either $\epsilon > 1$ or $\kappa_{gl} > 1$. They can be ignored in the more standard case where $\kappa_{gl} = 0$ and ϵ is a small number between 0 and 1, implying that $\epsilon_* = \epsilon$. Finally note that Lemma 3.1 allows us to equate $\beta > 1$ with $\beta = 1$ and $\kappa_{gl} = \|g(x_0)\|$. In this case, either $\epsilon_* = \epsilon > \|g(x_0)\|$ and Algorithm 2.1 stops at iteration 0, or $\epsilon_* = \|g(x_0)\|$ and the bounds of Theorem 3.10 become independent of ϵ, resulting in a bound on the number of iterations and evaluations directly proportional to $f(x_0) - f_{\text{target}}$, as expected.

We conclude by observing that the theory presented above recovers known results (see [2] for the case where $p = 3$ and $\beta = 1$ and [9, 3] for the case where $p = 2$ and $\beta = 1$); these cases correspond to the thick dots in the upper part of Figure 3.1.

References

[1] 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.

[2] 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.

[3] C. Cartis, N. I. M. Gould, and Ph. L. Toint. On the evaluation complexity of composite function minimization with applications to nonconvex nonlinear programming. *SIAM Journal on Optimization*, 21(4):1721–1739, 2011.

[4] 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.

[5] G. N. Grapiglia, J. Yuan, and Y. Yuan. Global convergence and worst-case complexity of a derivative-free trust-region algorithm for composite nonsmooth optimization. Technical report, University of Parana, Curitiba, Brasil, 2014.

[6] 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*, (to appear), 2014.

[7] S. Gratton, A. Sartenaer, and Ph. L. Toint. Recursive trust-region methods for multiscale nonlinear optimization. *SIAM Journal on Optimization*, 19(1):414–444, 2008.

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

[9] Yu. Nesterov. Gradient methods for minimising composite objective functions. *Mathematical Programming, Series A*, 140(1):125–161, 2013.

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

[11] K. Ueda. *A Regularized Newton Method without Line Search for Unconstrained Optimization*. PhD thesis, Department of Applied Mathematics and Physics, Graduate School of Informatics, Kyoto University, Kyoto, Japan, 2009.

[12] K. Ueda and N. Yamashita. Convergence properties of the regularized Newton method for the unconstrained nonconvex optimization. *Applied Mathematics & Optimization*, 62(1):27–46, 2009.

[13] K. Ueda and N. Yamashita. Convergence properties of the regularized newton method for the unconstrained nonconvex optimization. *Applied Mathematics and Optimization*, 62(1):27–46, 2010.

[14] K. Ueda and N. Yamashita. On a global complexity bound of the levenberg-marquardt method. *Journal of Optimization Theory and Applications*, 147:443–453, 2010.

[15] L. N. Vicente. Worst case complexity of direct search. *EURO Journal on Computational Optimization*, (to appear), 2013.