Worst-case evaluation complexity for unconstrained nonlinear optimization using high-order regularized models*

E. G. Birgin†, J. L. Gardenghi†, J. M. Martínez‡, S. A. Santos‡ and Ph. L. Toint§

21 April 2016

Abstract

The worst-case evaluation complexity for smooth (possibly nonconvex) unconstrained optimization is considered. It is shown that, if one is willing to use derivatives of the objective function up to order p (for $p \geq 1$) and to assume Lipschitz continuity of the p-th derivative, then an ϵ-approximate first-order critical point can be computed in at most $O(\epsilon^{-(p+1)/p})$ evaluations of the problem's objective function and its derivatives. This generalizes and subsumes results known for $p = 1$ and $p = 2$.

1 Introduction

Recent years have seen a surge of interest in the analysis of worst-case evaluation complexity of optimization algorithms for nonconvex problems (see, for instance, Vavasis [17], Nesterov and Polyak [16], Nesterov [14, 15], Gratton, Sartenaer and Toint [13], Cartis, Gould and Toint [3, 4, 5, 8], Bian, Chen and Ye [2], Bellavia, Cartis, Gould, Morini and Toint [1], Grapiglia, Yuan and Yuan [12], Vicente [18]). In particular the paper [16] was the first to show that a method using second derivatives can find an ϵ-approximate first-order critical point for an unconstrained problem with Lipschitz continuous Hessians in at most $O(\epsilon^{-3/2})$ evaluations of the objective function (and its derivatives), in contrast with methods using first-derivatives only, whose evaluation complexity was known [14] to be $O(\epsilon^{-2})$ for problems with Lipschitz continuous gradients. The purpose of the present short paper is to show that, if one is willing to use derivatives up to order p (for $p \geq 1$) and to assume Lipschitz continuity of the p-th derivative, then an ϵ-approximate first-order critical point can be computed in at most $O(\epsilon^{-(p+1)/p})$ evaluations of the objective function and its derivatives. This is achieved by the use of a regularization method very much in the spirit of the first- and second-order ARC methods described in [4, 5].

*This work has been partially supported by the Brazilian agencies FAPESP (grants 2010/10133-0, 2013/03447-6, 2013/05475-7, 2013/07375-0, and 2013/23494-9) and CNPq (grants 304032/2010-7, 309517/2014-1, 303750/2014-6, and 490326/2013-7) and by the Belgian Fund for Scientific Research (FNRS).

†Department of Computer Science, Institute of Mathematics and Statistics, University of São Paulo, Rua do Matão, 1010, Cidade Universitária, 05508-090, São Paulo, SP, Brazil. e-mail: {egbirgin | john}@ime.usp.br

‡Department of Applied Mathematics, Institute of Mathematics, Statistics, and Scientific Computing, University of Campinas, Campinas, SP, Brazil. e-mail: {martinez | sandra}@ime.unicamp.br

§Namur Center for Complex Systems (naXys) and Department of Mathematics, University of Namur, 61, rue de Bruxelles, B-5000 Namur, Belgium. Email: philippe.toint@unamur.be

2 A regularized p-th order model and algorithm

For $p \geq 1$, p integer, consider the problem

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

where we assume that f from $\mathbb{R}^n$ to $\mathbb{R}$ is bounded below and p-times continuously differentiable. We also assume that its p-th derivative at x, the p-th order tensor

$$\nabla_x^p f(x) = \left[\frac{\partial^p f}{\partial x_{i_1} \ldots \partial x_{i_p}} \right]_{i_j \in \{1,\ldots,n\}, j=1,\ldots,p} (x),$$

is Lipschitz continuous, i.e. that there exists a constant $L \geq 0$ such that, for all $x, y \in \mathbb{R}^n$,

$$\|\nabla_x^p f(x) - \nabla_x^p f(y)\|_{[p]} \leq (p-1)!\, L \|x - y\|. \tag{2.2}$$

In (2.2), $\|\cdot\|_{[p]}$ is the tensor norm recursively induced by the Euclidean norm $\|\cdot\|$ on the space of p-th order tensors, which is given by

$$\|T\|_{[p]} \stackrel{\text{def}}{=} \max_{\|v_1\| = \cdots = \|v_p\| = 1} |T[v_1, \ldots, v_p]|, \tag{2.3}$$

where $T[v_1, \ldots, v_j]$ stands for the tensor of order $q - j \geq 0$ resulting from the application of the q-th order tensor T to the vectors $v_1, \ldots, v_j$. Let $T_p(x, s)$ be the Taylor series of the function $f(x + s)$ at x truncated at order p

$$T_p(x, s) \stackrel{\text{def}}{=} f(x) + \sum_{j=1}^{p} \frac{1}{j!} \nabla_x^j f(x)[s]^j, \tag{2.4}$$

where the notation $T[s]^j$ stands for the tensor T applied j times to the vector s. Then Taylor's theorem, the identity

$$\int_0^1 (1 - \xi)^{p-1}\, d\xi = \frac{1}{p}, \tag{2.5}$$

the induced nature of $\|\cdot\|_{[p]}$ and (2.2) imply that, for all $x, s \in \mathbb{R}^n$,

$$\begin{aligned}
f(x+s) &= T_{p-1}(x,s) + \frac{1}{(p-1)!} \int_0^1 (1-\xi)^{p-1} \nabla_x^p f(x + \xi s)[s]^p \, d\xi \\
&\leq T_p(x,s) + \frac{1}{(p-1)!} \left| \int_0^1 (1-\xi)^{p-1} (\nabla_x^p f(x+\xi s)[s]^p - \nabla_x^p f(x)[s]^p) \, d\xi \right| \\
&\leq T_p(x,s) + \frac{1}{(p-1)!} \int_0^1 (1-\xi)^{p-1} |\nabla_x^p f(x+\xi s)[s]^p - \nabla_x^p f(x)[s]^p \, d\xi \\
&\leq T_p(x,s) + \left[\int_0^1 \frac{(1-\xi)^{p-1}}{(p-1)!} \, d\xi \right] \max_{\xi \in [0,1]} |\nabla_x^p f(x+\xi s)[s]^p - \nabla_x^p f(x)[s]^p| \\
&\leq T_p(x,s) + \frac{1}{p!} \|s\|^p \max_{\xi \in [0,1]} \|\nabla_x^p f(x+\xi s) - \nabla_x^p f(x)\|_{[p]} \\
&\leq T_p(x,s) + \frac{L}{p} \|s\|^{p+1}.
\end{aligned} \tag{2.6}$$

Following the more general argument developed by Cartis, Gould and Toint [10], consider now, for an arbitrary unit vector v, $\phi(\alpha) = \nabla_x^1 f(x+\alpha s)[v]$ and $\tau_{p-1}(\alpha) = \sum_{i=0}^{p-1} \phi^{(i)}(0)\alpha^i/i!$. Taylor's identity then gives that

$$\phi(1) - \tau_{p-1}(1) = \frac{1}{(p-2)!}\int_0^1 (1-\xi)^{p-2}[\phi^{(p-1)}(\xi) - \phi^{(p-1)}(0)]\,d\xi.$$

Hence, since $\tau_{p-1}(1) = \nabla_s^1 T_p(x,s)[v]$,

$$(\nabla_x^1 f(x+s) - \nabla_s^1 T_p(x,s))[v] = \frac{1}{(p-2)!}\int_0^1 (1-\xi)^{p-2}[\nabla_x^p f(x+\xi s) - \nabla_x^p f(x)][s]^{p-1}[v]\,d\xi.$$

Thus, using the symmetry of the derivative tensors, picking v to maximize the absolute value of the left-hand side and using (2.5), (2.3) and (2.2) successively, we obtain that

$$\begin{aligned}
&\|\nabla_x^1 f(x+s) - \nabla_s^1 T_p(x,s)\| \\
&\quad = \frac{1}{(p-2)!}\left|\int_0^1 (1-\xi)^{p-2}(\nabla_x^p f(x+\xi s) - \nabla_x^p f(x))[v]\left[\frac{s}{\|s\|}\right]^{p-1}\|s\|^{p-1}d\xi\right| \\
&\quad \le \frac{1}{(p-2)!}\left[\int_0^1 (1-\xi)^{p-2}d\xi\right]\max_{\xi\in[0,1]}\left|(\nabla_x^p f(x+\xi s) - \nabla_x^p f(x))[v]\left[\frac{s}{\|s\|}\right]^{p-1}\right|\|s\|^{p-1} \qquad (2.7)\\
&\quad \le \frac{1}{(p-1)!}\max_{\xi\in[0,1]}\max_{\|w_1\|=\cdots=\|w_p\|=1}|(\nabla_x^p f(x+\xi s) - \nabla_x^p f(x))[w_1,\ldots,w_p]|\,\|s\|^{p-1} \\
&\quad = \frac{1}{(p-1)!}\max_{\xi\in[0,1]}\|\nabla_x^p f(x+\xi s) - \nabla_x^p f(x)\|_{[p]}\|s\|^{p-1} \\
&\quad \le L\|s\|^p.
\end{aligned}$$

In order to describe our algorithm, we also define the regularized Taylor series

$$m(x,s,\sigma) = T_p(x,s) + \frac{\sigma}{p+1}\|s\|^{p+1}, \qquad (2.8)$$

whose gradient is

$$\nabla_s^1 m(x,s,\sigma) = \nabla_s^1 T_p(x,s) + \sigma\|s\|^p\frac{s}{\|s\|}. \qquad (2.9)$$

Note that

$$m(x,0,\sigma) = T_p(x,0) = f(x). \qquad (2.10)$$

The minimization algorithm we consider is now detailed as Algorithm 1 on the following page.

Each iteration of this algorithm requires the approximate minimization of $m(x_k,s,\sigma_k)$, but we may note that conditions (2.12) and (2.13) are relatively weak, in that they only require a decrease of the regularized p-th order model and an approximate first-order stationary point: no global optimization of this possibly nonconvex model is needed. Fortunately, *this approximate minimization does not involve additional computations of f or of its derivatives at other points than at x_k, and therefore the exact method used and the resulting effort spent in Step 2 have no impact on the evaluation complexity.* Also note that the numerator and

Algorithm 1: ARp

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

$$\theta > 0, \quad \sigma_{\min} \in (0, \sigma_0], \quad 0 < \eta_1 \leq \eta_2 < 1 \text{ and } 0 < \gamma_1 < 1 < \gamma_2 < \gamma_3. \tag{2.11}$$

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

Step 1: Test for termination. Evaluate $\nabla_x^1 f(x_k)$. If $\|\nabla_x^1 f(x_k)\| \leq \epsilon$, terminate with the approximate solution $x_\epsilon = x_k$. Otherwise compute derivatives of f from order 2 to p at x_k.

Step 2: Step calculation. Compute the step s_k by approximately minimizing the model $m(x_k, s, \sigma_k)$ with respect to s in the sense that the conditions

$$m(x_k, s_k, \sigma_k) < m(x_k, 0, \sigma_k) \tag{2.12}$$

and

$$\|\nabla_s^1 m(x_k, s_k, \sigma_k)\| \leq \theta \|s_k\|^p \tag{2.13}$$

hold.

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

$$\rho_k = \frac{f(x_k) - f(x_k + s_k)}{T_p(x_k, 0) - T_p(x_k, s_k)}. \tag{2.14}$$

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

Step 4: Regularization parameter update. Set

$$\sigma_{k+1} \in \begin{cases} [\max(\sigma_{\min}, \gamma_1 \sigma_k), \sigma_k] & \text{if } \rho_k \geq \eta_2, \\ [\sigma_k, \gamma_2 \sigma_k] & \text{if } \rho_k \in [\eta_1, \eta_2), \\ [\gamma_2 \sigma_k, \gamma_3 \sigma_k] & \text{if } \rho_k < \eta_1. \end{cases} \tag{2.15}$$

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

denominator in (2.14) are strictly comparable, the latter being Taylor's approximation of the former, without the regularization parameter playing any role.

Iterations for which $\rho_k \geq \eta_1$ (and hence $x_{k+1} = x_k + s_k$) are called "successful" and we denote by $\mathcal{S}_k \stackrel{\text{def}}{=} \{0 \leq j \leq k \mid \rho_j \geq \eta_1\}$ the index set of all successful iterations between 0 and k. We also denote by $\mathcal{U}_k$ its complement in $\{0, \ldots, k\}$, which corresponds to the index set of "unsuccessful" iterations between 0 and k. Note that, before termination, each successful iteration requires the evaluation of f and its first p derivatives, while only the evaluation of f is needed at unsuccessful ones.

We first derive a very simple result on the model decrease obtained under condition (2.12).

Lemma 2.1 The mechanism of Algorithm 1 then guarantees that, for all $k \geq 0$,

$$T_p(x_k, 0) - T_p(x_k, s_k) \geq \frac{\sigma_k}{p+1} \|s_k\|^{p+1}. \tag{2.16}$$

Proof. Observe that, because of (2.12) and (2.8),

$$0 \leq m(x_k, 0, \sigma_k) - m(x_k, s_k, \sigma_k) = T_p(x_k, 0) - T_p(x_k, s_k) - \frac{\sigma_k}{p+1} \|s_k\|^{p+1}$$

which implies the desired bound. □

As a result, we obtain that (2.14) is well-defined for all $k \geq 0$. We next deduce a simple upper bound on the regularization parameter σ_k.

Lemma 2.2 Suppose that f is p times continuously differentiable with Lipschitz continuous p-th derivative (i.e., that (2.2) holds). Then, for all $k \geq 0$,

$$\sigma_k \leq \sigma_{\max} \stackrel{\text{def}}{=} \max \left[\sigma_0, \frac{\gamma_3 L(p+1)}{p\,(1-\eta_2)} \right]. \tag{2.17}$$

Proof. Assume that

$$\sigma_k \geq \frac{L(p+1)}{p\,(1-\eta_2)}. \tag{2.18}$$

Using (2.6) and (2.16), we may then deduce that

$$|\rho_k - 1| \leq \frac{|f(x_k + s_k) - T_p(x_k, s_k)|}{|T_p(x_k, 0) - T_p(x_k, s_k)|} \leq \frac{L(p+1)}{p\,\sigma_k} \leq 1 - \eta_2$$

and thus that $\rho_k \geq \eta_2$. Then iteration k is very successful in that $\rho_k \geq \eta_2$ and $\sigma_{k+1} \leq \sigma_k$. As a consequence, the mechanism of the algorithm ensures that (2.17) holds. □

Our next step, very much in the line of the theory proposed in [5], is to show that the steplength cannot be arbitrarily small compared with the gradient of the objective function at the trial point $x_k + s_k$.

Lemma 2.3 Suppose that f is p times continuously differentiable with Lipschitz continuous p-th derivative (i.e., that (2.2) holds). Then, for all $k \geq 0$,

$$\|s_k\| \geq \left(\frac{\|\nabla_x^1 f(x_k + s_k)\|}{L + \theta + \sigma_k} \right)^{\frac{1}{p}}. \tag{2.19}$$

Proof. Using the triangle inequality, (2.7), (2.9) and (2.13), we obtain that

$$\begin{aligned} \|\nabla_x^1 f(x_k + s_k)\| &\leq \|\nabla_x^1 f(x_k + s_k) - \nabla_s^1 T_p(x_k, s_k)\| + \left\| \nabla_s^1 T_p(x_k, s_k) + \sigma_k \|s_k\|^p \frac{s_k}{\|s_k\|} \right\| \\ &\quad + \sigma_k \|s_k\|^p \\ &\leq L\|s_k\|^p + \|\nabla_s^1 m(x_k, s_k, \sigma_k)\| + \sigma_k \|s_k\|^p \\ &\leq [L + \theta + \sigma_k] \, \|s_k\|^p \end{aligned}$$

and (2.19) follows. □

We now bound the number of unsuccessful iterations as a function of the number of successful ones.

Lemma 2.4 [5, Theorem 2.1] The mechanism of Algorithm 1 guarantees that, if

$$\sigma_k \leq \sigma_{\max}, \tag{2.20}$$

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

$$k + 1 \leq |\mathcal{S}_k| \left(1 + \frac{|\log \gamma_1|}{\log \gamma_2} \right) + \frac{1}{\log \gamma_2} \log \left(\frac{\sigma_{\max}}{\sigma_0} \right). \tag{2.21}$$

Proof. The regularization parameter update (2.15) gives that, for each k,

$$\gamma_1 \sigma_j \leq \max[\gamma_1 \sigma_j, \sigma_{\min}] \leq \sigma_{j+1}, \quad j \in \mathcal{S}_k, \quad \text{and} \quad \gamma_2 \sigma_j \leq \sigma_{j+1}, \quad j \in \mathcal{U}_k.$$

Thus we deduce inductively that

$$\sigma_0 \gamma_1^{|\mathcal{S}_k|} \gamma_2^{|\mathcal{U}_k|} \leq \sigma_k.$$

We therefore obtain, using (2.20), that

$$|\mathcal{S}_k| \log \gamma_1 + |\mathcal{U}_k| \log \gamma_2 \leq \log \left(\frac{\sigma_{\max}}{\sigma_0} \right),$$

which then implies that

$$|\mathcal{U}_k| \leq -|\mathcal{S}_k| \frac{\log \gamma_1}{\log \gamma_2} + \frac{1}{\log \gamma_2} \log \left(\frac{\sigma_{\max}}{\sigma_0} \right),$$

since $\gamma_2 > 1$. The desired result (2.21) then follows from the equality $k+1 = |\mathcal{S}_k| + |\mathcal{U}_k|$ and the inequality $\gamma_1 < 1$ given by (2.11). □

Using all the above results, we are now in position to state our main evaluation complexity result.

Theorem 2.5 Suppose that f is p times continuously differentiable with Lipschitz continuous p-th derivative (i.e., that (2.2) holds), and let $f_{\rm low}$ be a lower bound on f. Then, given $\epsilon > 0$, Algorithm 1 needs at most

$$\left\lfloor \kappa_s \, \frac{f(x_0) - f_{\rm low}}{\epsilon^{\frac{p+1}{p}}} \right\rfloor$$

successful iterations (each involving one evaluation of f and its p first derivatives) and at most

$$\left\lfloor \kappa_s \, \frac{f(x_0) - f_{\rm low}}{\epsilon^{\frac{p+1}{p}}} \right\rfloor \left(1 + \frac{|\log \gamma_1|}{\log \gamma_2} \right) + \frac{1}{\log \gamma_2} \log \left(\frac{\sigma_{\max}}{\sigma_0} \right)$$

iterations in total to produce an iterate x_ϵ such that $\|\nabla_x^1 f(x_\epsilon)\| \le \epsilon$, where $\sigma_{\max}$ is given by (2.17) and where

$$\kappa_s \stackrel{\rm def}{=} \frac{p+1}{\eta_1 \sigma_{\min}} \, (L + \theta + \sigma_{\max})^{\frac{p+1}{p}} \, .$$

Proof. At each successful iteration before termination, we have that

$$\begin{aligned} f(x_k) - f(x_k + s_k) &\ge \eta_1 (T_p(x_k, 0) - T_p(x_k, s_k)) \\ &\ge \frac{\eta_1 \sigma_{\min}}{p+1} \, \|s_k\|^{p+1} \\ &\ge \frac{\eta_1 \sigma_{\min}}{(p+1)(L + \theta + \sigma_k)^{\frac{p+1}{p}}} \, \|\nabla_x^1 f(x_k + s_k)\|^{\frac{p+1}{p}} \\ &\ge \frac{\eta_1 \sigma_{\min}}{(p+1)(L + \theta + \sigma_{\max})^{\frac{p+1}{p}}} \, \epsilon^{\frac{p+1}{p}} , \end{aligned}$$

where we used (2.14), (2.16), (2.15), (2.19), (2.17) and the fact that $\|\nabla_x^1 f(x_k + s_k)\| \ge \epsilon$ before termination. Thus we deduce that, for successful iterations and as long as termination does not occur,

$$f(x_0) - f(x_{k+1}) = \sum_{j \in \mathcal{S}_k} [f(x_j) - f(x_j + s_j)] \ge \frac{|\mathcal{S}_k|}{\kappa_s} \, \epsilon^{\frac{p+1}{p}} ,$$

from which the desired bound on the number of successful iterations follows. Lemma 2.4 is then invoked to compute the upper bound on the total number of iterations. □

The complexity bound of Theorem 2.5 can also be stated as the fact that, for a p times continuously differentiable objective function with Lipschitz continuous p-th derivative, the global rate of convergence for the gradient's norm is $O(k^{-p/(p+1)})$.

3 Final comments

We have shown that, under suitable smoothness assumptions, an ϵ-approximate stationary point must be found by Algorithm 1 in at most $O(\epsilon^{-(p+1)/p})$ iterations and function evaluations. This extension of results known for $p = 1$ and $p = 2$ to arbitrary $p \geq 1$ is made possible by the introduction of two main innovations: weaker termination conditions on the model minimization subproblem (no global optimization is required at all) and a reformulation of the ratio of achieved versus predicted decreases where the model is limited to the Taylor approximation. Of course, each iteration of the proposed algorithm requires the approximate minimization of a typically nonconvex regularized p-th order model but this minimization does not involve additional computation of the objective function of the original problem or of its derivatives, and therefore its cost does not affect the evaluation complexity of Algorithm 1. What numerical procedure is best for this task is beyond the scope of the present note (for instance, one might think of applying an efficient first-order method on the model).

Once this paper was submitted, the authors became aware of the interesting contribution by Dussault [11] where the decrease measure (2.14) is also used to analyse a framework unifying the complexity analysis of the cubic regularization algorithm and trust-region methods.

It is of course interesting to consider if the extensions of the theories developed for the first- and second-order cubic regularization methods for second-order optimality [8] or convexly constrained problems [7] can be extended to higher-order regularization approaches. We also note that Cartis *et al.* showed in [6] that a worst-case evaluation complexity of order $O(\epsilon^{-3/2})$ is optimal for a large class of second-order methods applied on twice continuously differentiable problems with Hölder continuous Hessians. The generalization of this optimality result for $p > 2$ is also an open question.

Whether the approach presented here has practical implications remains to be seen, since the approximate model minimization could be costly even if computation of f is avoided, and computing p derivatives for $p > 2$ may often be out of reach.

We conclude this paper by mentioning a simple extension which we anticipate could be useful in other contexts. We may, instead of minimizing $f(x)$, split the objective function into two parts and consider minimizing $\Phi(x) = h(x) + f(x)$ where h is bounded below and continuously differentiable. In this case, we then replace the model defined by (2.8) by $m(x, s, \sigma) = h(x + s) + T_p(x, s) + \sigma\|s\|^{p+1}/(p + 1)$ and, provided we are ready to (approximately) minimize this augmented model in Step 2 of the algorithm, the above analysis remains unchanged. There are many possible interesting choices for $h(x)$: in the context of optimization with non-negative variables, a possibility is, for example, to choose $h(x) = [\max(x, 0)]^2$ as a reformulation of the constraint $x \geq 0$. What part of the objective function is "easy" enough to be included in the model $m(x, s, \sigma)$ explicitly and which part is better included using a Taylor series approximation may depend on the problem at hand, but it is interesting to note that the evaluation-complexity bound presented in Theorem 2.5 is unaffected.

Acknowledgements

The authors are pleased to thank Coralia Cartis and Nick Gould for valuable comments, in particular on the definition of the tensor Lipschitz condition and associated material. Two anonymous referees also helped to improve the final manuscript.

References

[1] S. Bellavia, C. Cartis, N. I. M. Gould, B. Morini, and Ph. L. Toint. Convergence of a regularized Euclidean residual algorithm for nonlinear least-squares. *SIAM Journal on Numerical Analysis*, 48(1):1–29, 2010.

[2] W. Bian, X. Chen, and Y. Ye. Complexity analysis of interior point algorithms for non-Lipschitz and nonconvex minimization. *Mathematical Programming, Series A*, 149:301–327, 2015.

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

[4] C. Cartis, N. I. M. Gould, and Ph. L. Toint. Adaptive cubic overestimation methods for unconstrained optimization. Part I: motivation, convergence and numerical results. *Mathematical Programming, Series A*, 127(2):245–295, 2011.

[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 optimization. Technical Report naXys-17-2011, Namur Center for Complex Systems (naXys), University of Namur, Namur, Belgium, 2011.

[7] C. Cartis, N. I. M. Gould, and Ph. L. Toint. An adaptive cubic regularization algorithm for nonconvex optimization with convex constraints and its function-evaluation complexity. *IMA Journal of Numerical Analysis*, 32(4):1662–1695, 2012.

[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. Evaluation complexity of adaptive cubic regularization methods for convex unconstrained optimization. *Optimization Methods and Software*, 27(2):197–219, 2012.

[10] C. Cartis, N. I. M. Gould, and Ph. L. Toint. Second-order optimality and (sometimes) beyond: characterization and evaluation complexity in nonconvex optimization. Technical Report (in preparation), Namur Center for Complex Systems (naXys), University of Namur, Namur, Belgium, 2016.

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

[12] 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.

[13] 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.

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

[15] Yu. Nesterov. Modified Gauss-Newton scheme with worst-case guarantees for global performance. *Optimization Methods and Software*, 22(3):469–483, 2007.

[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] S. A. Vavasis. Black-box complexity of local minimization. *SIAM Journal on Optimization*, 3(1):60–80, 1993.

[18] L. N. Vicente. Worst case complexity of direct search. *EURO Journal on Computational Optimization*, 1:143–153, 2013.