On High-order Model Regularization for Constrained Optimization

José Mario Martínez*

February 7, 2017

Abstract

In two recent papers regularization methods based on Taylor polynomial models for minimization were proposed that only rely on Hölder conditions on the higher order employed derivatives. Grapiglia and Nesterov considered cubic regularization with a sufficient descent condition that uses the current gradient and resembles the classical Armijo's criterion. Cartis, Gould, and Toint used Taylor models with arbitrary-order regularization and defined methods that tackle convex constraints employing the descent criterion that compares actual reduction with predicted reduction. The methods presented in this paper consider general (not necessarily Taylor) models of arbitrary order, employ a very mild descent criterion and handles general, non necessarily convex, constraints. Complexity results are compatible with the ones presented in the papers mentioned above.

1 Introduction

In this paper we will address the problem of minimizing a smooth and generally nonconvex function onto a region of the Euclidean finite dimensional space.

Algorithms that use regularized quadratic subproblems for unconstrained minimization or adapted trust-regions supbproblems with proven complexity $O(\varepsilon^{-3/2})$ were given in [7, 8, 9, 12, 14, 23, 24]. Cubic regularization was introduced in [20] and used for solving challenging engineering problems in [29]. The optimality of the complexity $O(\varepsilon^{-3/2})$ was proved in [9].

In [4], a generalization of cubic regularization to arbitrary p-th regularization for unconstrained optimization was given. At each iteration, the method introduced in [4] approximately minimizes, with a mild tolerance, a p-th Taylor polynomial around the current point plus a regularization term of order $p+1$. Using Lipschitz conditions on the derivatives of order p, it was proved that the method achieves a gradient norm smaller than ε in at most $O(\varepsilon^{-\frac{p+1}{p}})$ iterations and functional evaluations.

On the other hand, Grapiglia and Nesterov [18] considered unconstrained problems in which, instead of Lipschitz continuity of second derivatives, Hölder continuity was assumed with a

*Institute of Mathematics, Statistics and Scientific Computing, State University of Campinas, CP 6065, 13081-970 Campinas SP, Brazil. This work was supported by Cepid-Cemeai-Fapesp, PRONEX-CNPq/FAPERJ E-26/111.449/2010-APQ1, FAPESP (grants 2010/10133-0, 2013/03447-6, 2013/05475-7, and 2013/07375-0), and CNPq.

parameter $\beta \in [0,1]$. They proved that, by means of cubic regularization and a proper sufficient descent condition, ε-gradient norm stationarity was achieved using $O(\varepsilon^{-\frac{2+\beta}{1+\beta}})$ iterations and evaluations.

Cartis, Gould, and Toint [10] generalized the results of [4] considering convex constraints and relaxing Lipschitz-continuity of the derivatives of order p to Hölder-continuity. When regularization of order $p+1$ was used in the subproblems the complexity in terms of iterations and evaluations was $O(\varepsilon^{-\frac{p+\beta}{p+\beta-1}})$. In the discussion of their complexity results they conclude that, from the point of view of complexity, there is no reason for using a regularization order different from $p+1$.

In the present paper we will consider arbitrary (non-necessarily Taylor-based) models of the objective function and non-necessarily convex constraints. At each iteration we compute a mild approximate KKT point for the subproblem without requiring the fulfillment of any constraint qualification, according to the theory of sequential optimality conditions [1, 2, 6]. The functional decrease required at each iteration will be $\frac{\alpha\varepsilon^{\frac{p+1}{p}}}{(2p+4)^{\frac{p+1}{p}}\sigma^{\frac{1}{p}}}$ where σ is the regularization parameter and $\alpha \in (0,1)$. The amount of iterations and objective function evaluations necessary to achieve Approximate KKT conditions with precision ε will be $O(\varepsilon^{-\frac{p+\beta}{p+\beta-1}})$, as in [10].

As a by-product of the present research, the algorithmic description and the proofs of the main results are quite simple and may be followed without previous knowledge on Optimization Methods and Theory.

Notation

$\|\cdot\|$ denotes the Euclidean norm.

If $v, w \in \mathbb{R}^n$, $\min\{v, w\}$ denotes the vector with components $\min\{v_1, w_1\}, \ldots, \min\{v_n, w_n\}$.

2 Main results

Assume that $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R}^n \to \mathbb{R}^n$, $h_E : \mathbb{R}^n \to \mathbb{R}^m$, $h_I : \mathbb{R}^n \to \mathbb{R}^q$, $h'_E : \mathbb{R}^n \to \mathbb{R}^{m\times n}$, and $h'_I : \mathbb{R}^n \to \mathbb{R}^{q\times n}$. Our problem will be

$$\text{Minimize } f(x) \text{ subject to } h_E(x) = 0 \text{ and } h_I(x) \leq 0. \tag{1}$$

Throughout this paper we will assume that $p \in \{1, 2, 3, \ldots\}$, $L > 0$, $\epsilon > 0$, and $\beta \in [0, 1]$ with $p + \beta > 1$. We will also assume that g denotes the gradient of f, while $h'_E(x)$ and $h'_I(x)$ are the Jacobians of $h_E(x)$ and $h_I(x)$, respectively. However, we warn that the formal results that will be proved in this section are independent of those meanings. Analogously, ∇ will denote the gradient operator but it could be interpreted also as an arbitrary linear operator on the space of functions.

For all $\bar{x} \in \mathbb{R}^n$ we define $M_{\bar{x}} : \mathbb{R}^n \to \mathbb{R}$ (intended to be a "model" of $f(x)$ around $\bar{x}$) and $\nabla M_{\bar{x}} : \mathbb{R}^n \to \mathbb{R}^n$. Along this work we will invoke frequently the fulfillment of the following assumptions by $\bar{x}$ and x.

$$\|g(x) - \nabla M_{\bar{x}}(x)\| \leq L\|x - \bar{x}\|^{p+\beta-1} \tag{2}$$

and

$$M_{\bar{x}}(\bar{x}) = f(\bar{x}) \text{ and } f(x) \leq M_{\bar{x}}(x) + L\|x - \bar{x}\|^{p+\beta}. \tag{3}$$

The most usual interpretation of assumptions (2) and (3) corresponds to the case in which g is the gradient of f, and $\nabla M_{\bar{x}}$ is the gradient of $M_{\bar{x}}$. In this case condition (2) says that the gradient of f is approximated by the gradient of $M_{\bar{x}}$ with an error proportional to the $p+\beta-1$ power of the distance between x and $\bar{x}$. Analogously, (3) says that $M_{\bar{x}}(x)$ plus $L\|x - \bar{x}\|^{p+\beta}$ overestimates $f(x)$. The condition $p + \beta > 1$ ensures that the difference between $g(x)$ tends to zero as x tends to $\bar{x}$.

Note that, under the usual interpretation, (3) follows from (2). In fact, by integration, (2) implies the stronger statement $|f(x)-M_{\bar{x}}(x)| \leq \frac{L}{2}\|x-\bar{x}\|^{p+\beta}$, but we prefer to state (3) explicitly in order to preserve independence with respect to the interpretations for g and $\nabla M_{\bar{x}}$.

If $M_{\bar{x}}(x)$ is the Taylor polynomial of order p around $\bar{x}$ and the $p-$th derivatives satisfy a Hölder condition defined by β and L the assumptions (2) and (3) are satisfied for all $\bar{x}, x \in \mathbb{R}^n$ (see, for example, [3]).

However, more general situations concerning the model $M_{\bar{x}}(x)$ are addressed by the approach of the present paper. For example, suppose that the objective function of a minimization problem has the form $f(x) = \varphi(x) + \sum_{i=1}^{n} \max\{0, x_i\}^2$. In this case it may be convenient to employ the model given by $T_{p,\bar{x}}(x) + \sum_{i=1}^{n} \max\{0, x_i\}^2$, where $T_{p,\bar{x}}(x)$ is the Taylor p-th approximation of φ around $\bar{x}$. In other cases the model is polynomial but the terms of highest order are not defined by true derivatives. For example, classical quasi-Newton methods [13], in which only the first-order terms of the model are derivatives (see complexity analysis in [19]), tensor methods [27], in which high order derivatives are approximated by means of specific schemes, and variable norm methods that employ quadratic Taylor models and crude approximations of third order derivatives [22, 23]. An extreme example is the case in which $M_{\bar{x}}(x) = f(x)$. In this case (2) and (3) hold for all p and solving the subproblem (8) described below consists on minimizing f plus a regularization term. This is a form of the proximal point algorithm [25, 26], which is known to be useful for solving Inverse Problems and for defining Augmented Lagrangians. See [28] for a discussion of inexact solutions of subproblems and [21] for general coercive regularizations in the proximal point context .

The following conditions regarding the behavior of the model $M_{\bar{x}}(x)$ will be employed along this paper:

$$M_{\bar{x}}(x) + \sigma\|x - \bar{x}\|^{p+1} \leq f(\bar{x}) \tag{4}$$

and

$$\|\nabla[M_{\bar{x}}(x) + \sigma\|x - \bar{x}\|^{p+1}] + h'_E(x)^T\lambda + h'_I(x)^T\mu\| \leq \theta\|x - \bar{x}\|^p, \tag{5}$$

where

$$\lambda \in \mathbb{R}^m, \mu \in \mathbb{R}^q_+, \|\min\{\mu, -h_I(x)\}\| \leq \epsilon \tag{6}$$

$$\|h_E(x)\| \leq \epsilon, \text{ and } \|h_I(x)_+\| \leq \epsilon. \tag{7}$$

The standard interpretation conditions (4)–(7) comes from considering the minimization problem

$$\text{Minimize } M_{\bar{x}}(x) + \sigma\|x - \bar{x}\|^{p+1} \text{ subject to } h_E(x) = 0 \text{ and } h_I(x) \leq 0. \tag{8}$$

In fact, under smoothness assumptions, (4)–(7) may be seen as a stopping criterion related to the Approximate KKT conditions (AKKT) associated with problem (8). Such conditions are satisfied by local minimizers of constrained optimization problems independently of constraint qualifications [1, 2]. In the case that $M_{\bar{x}}(x)$ is the p-th Taylor polynomial and the feasible region defined by $h_E(x) = 0$ and $h_I(x) \leq 0$ is non-empty, the subproblem (8) has a solution that necessarily satisfies (5)–(7).

The definitions above make it possible to define the main algorithm of this paper.

Algorithm 2.1

Assume that $x^0 \in \mathbb{R}^n$, $\alpha \in (0, 1)$, $\varepsilon \in (0, 1)$, $\epsilon > 0$, $f_{target} \in \mathbb{R}$, $\theta > 0$, and $\sigma_{min} > 0$,

Initialize $k \leftarrow 0$ and $\sigma_0 = \sigma_{min}$.

Step 1. Set $\sigma \leftarrow \sigma_k$.

Step 2. Find $x \in \mathbb{R}^n$, $\lambda \in \mathbb{R}^m$, and $\mu \in \mathbb{R}^q_+$ such that (4) and (5) hold with $\bar{x} = x^k$.

Step 3. If $\|g(x) + h'_E(x)^T\lambda + h'_I(x)^T\mu\| \leq \varepsilon$ or $f(x) \leq f_{target}$, stop.

Step 4. Test the sufficient descent condition

$$f(x) \leq f(x^k) - \frac{\alpha}{(2p+4)^{\frac{p+1}{p}}} \frac{\varepsilon^{\frac{p+1}{p}}}{\sigma^{\frac{1}{p}}}. \tag{9}$$

If (9) does not hold, set $\sigma \leftarrow 2\sigma$ and go to Step 2. Else, continue at Step 5.

Step 5. Define $x^{k+1} = x$, $k \leftarrow k + 1$, $\sigma_k = \sigma$, and go to Step 1.

The following lemma states that, whenever a current point $\bar{x}$ does not satisfy a suitable stationarity criterion, significative decrease must be expected in the objective function value ((14) and (15). This decrease will support the complexity results that state the maximal number of iterations and evaluations being necessary to achieve the stationary criterion.

It is interesting to observe that the following lemma and its proof is completely algebraic. We will prove that the statements (2), (3), (4), (5), (6), (7), and (12) imply (10), (13), (14), and (15) independently of the analytical meaning of terms of these equations and without the employment of differential or continuity arguments.

Lemma 2.1 *Assume that $\alpha \in (0, 1)$ and (2), (3), (4), (5), (6), and (7) are satisfied by $\bar{x}$, x, λ, and μ. Then,*

$$\|g(x) + h'_E(x)^T\lambda + h'_I(x)^T\mu\| \leq (\theta + (p+1)\sigma)\|x - \bar{x}\|^p + L\|x - \bar{x}\|^{p+\beta-1}. \tag{10}$$

Moreover, if

$$\|g(x) + h'_E(x)^T\lambda + h'_I(x)^T\mu\| \geq \varepsilon_g > 0 \tag{11}$$

and

$$\sigma \geq \max\left\{\theta, \varepsilon_g^{\frac{\beta-1}{p+\beta-1}} \max\left\{\frac{2^{\frac{-\beta+1}{p+\beta-1}}}{p+2} L^{\frac{p}{p+\beta-1}}, \frac{L}{(1-\alpha)((2p+4)^{(\beta-1)/(p+\beta-1)}}\right\}\right\} \tag{12}$$

we have that

$$\sigma\|x - \bar{x}\|^p \geq \frac{\varepsilon_g}{2p+4}, \tag{13}$$

$$f(x) \leq f(\bar{x}) - \alpha\sigma\|x - \bar{x}\|^{p+1}, \tag{14}$$

and

$$f(x) \leq f(\bar{x}) - \frac{\alpha}{(2p+4)^{\frac{p+1}{p}}} \frac{\varepsilon_g^{\frac{p+1}{p}}}{\sigma^{\frac{1}{p}}}. \tag{15}$$

Proof By (2),

$$\|g(x) + h'_E(x)^T\lambda + h'_I(x)^T\mu\| \leq \|\nabla M_{\bar{x}}(x) + h'_E(x)^T\lambda + h'_I(x)^T\mu\| + L\|x - \bar{x}\|^{p+\beta-1}.$$

Then, by (5),

$$\begin{aligned}
&\|g(x) + h'_E(x)^T\lambda + h'_I(x)^T\mu\| \\
&\leq \|\nabla[M_{\bar{x}}(x) + \sigma\|x - \bar{x}\|^{p+1}] + h'_E(x)^T\lambda + h'_I(x)^T\mu\| + \|\nabla[\sigma\|x - \bar{x}\|^{p+1}]\| + L\|x - \bar{x}\|^{p+\beta-1} \\
&\leq \|\nabla[M_{\bar{x}}(x) + \sigma\|x - \bar{x}\|^{p+1}] + h'_E(x)^T\lambda + h'_I(x)^T\mu\| + (p+1)\sigma\|x - \bar{x}\|^p + L\|x - \bar{x}\|^{p+\beta-1} \\
&\leq \theta\|x - \bar{x}\|^p + (p+1)\sigma\|x - \bar{x}\|^p + L\|x - \bar{x}\|^{p+\beta-1}.
\end{aligned}$$

Therefore, (10) is proved.
By (10) and (12, since $\sigma \geq \theta$,

$$(p+2)\sigma\|x - \bar{x}\|^p + L\|x - \bar{x}\|^{p+\beta-1} \geq \varepsilon_g. \tag{16}$$

Defining $t = \sigma\|x - \bar{x}\|^p$, $\nu = \frac{p+\beta-1}{p}$, and $c_1 = L/((p+2)\sigma^\nu)$, (16) implies that

$$t + c_1 t^\nu \geq \varepsilon_g/(p+2).$$

If $t \geq c_1 t^\nu$ we clearly have that (13) holds. On the other hand, if $c_1 t^\nu > t$ we have that

$$c_1 t^\nu \geq \varepsilon_g/(2p+4).$$

Therefore,

$$t \geq \left[\frac{\varepsilon_g}{(2p+4)c_1}\right]^{1/\nu}.$$

Thus, in order to prove (13) it is enough to prove that

$$\frac{\varepsilon_g}{(2p+4)c_1} \geq \frac{\varepsilon_g^\nu}{(2p+4)^\nu}. \tag{17}$$

This is equivalent to

$$c_1 \leq \varepsilon_g^{1-\nu}(2p+4)^{\nu-1}.$$

Therefore, we wish to prove that

$$L/((p+2)\sigma^\nu) \leq \varepsilon_g^{1-\nu}(2p+4)^{\nu-1}. \tag{18}$$

Isolating σ in this inequation and using the definition of ν, we see that (18) is equivalent to:

$$\sigma \geq \varepsilon_g^{\frac{\beta-1}{p+\beta-1}} 2^{\frac{1-\beta}{p+\beta-1}} (p+2)^{-1} L^{\frac{p}{p+\beta-1}}.$$

By (12) this inequality holds. Therefore, (18) holds, implying (13).

Let us now prove (14). By (3) and (4),

$$f(x) \leq M_{\bar{x}}(x) + L\|x-\bar{x}\|^{p+\beta}$$

$$\leq M_{\bar{x}}(x) + \sigma\|x-\bar{x}\|^{p+1} - \sigma\|x-\bar{x}\|^{p+1} + L\|x-\bar{x}\|^{p+\beta} \leq f(\bar{x}) - \sigma\|x-\bar{x}\|^{p+1} + L\|x-\bar{x}\|^{p+\beta}.$$

Thus, for proving (14) it is enough to prove that

$$\sigma\|x-\bar{x}\|^{p+1} - L\|x-\bar{x}\|^{p+\beta} \geq \alpha\sigma\|x-\bar{x}\|^{p+1}. \tag{19}$$

This is equivalent to

$$\sigma(1-\alpha) \geq L\|x-\bar{x}\|^{\beta-1} \tag{20}$$

But, by (13) we have that $\|x-\bar{x}\|^{\beta-1} \leq \frac{\varepsilon_g^{(\beta-1)/p}}{((2p+4)\sigma)^{(\beta-1)/p}}$. Thus, a sufficient condition for the fulfillment of (20) is:

$$\sigma(1-\alpha) \geq L\frac{\varepsilon_g^{(\beta-1)/p}}{((2p+4)\sigma)^{(\beta-1)/p}}. \tag{21}$$

This condition is equivalent to

$$\sigma \geq \frac{1}{1-\alpha}[L\frac{\varepsilon_g^{(\beta-1)/(p+\beta-1)}}{((2p+4)^{(\beta-1)/(p+\beta-1)}}]. \tag{22}$$

Therefore, by (12), the proof of (14) is complete.

Now let us prove (15). By (13) we have that

$$\|x-\bar{x}\| \geq \frac{\varepsilon_g^{1/p}}{(2p+4)^{1/p}\sigma^{1/p}}.$$

Therefore,

$$\alpha\sigma\|x-\bar{x}\|^{p+1} \geq \frac{\alpha}{(2p+4)^{\frac{p+1}{p}}}\frac{\varepsilon_g^{(p+1)/p}}{\sigma^{1/p}}.$$

As a consequence, (15) follows from (14). This completes the proof. □

Thanks to Lemma 2.1 we are able to prove that, if the algorithm does not stop at Step 3, a functional descent $O(\varepsilon^{\frac{p+\beta}{p+\beta-1}}$ is obtained.

Theorem 2.1 *Assume that x^{k+1} is computed by Algorithm 2.1 and the assumptions (2)–(7) hold for $\bar{x} = x^k$ at all the trial points x computed at Step 2 of iteration k.*

Define

$$c_p = \min\left\{\frac{1}{(2p+4)^{\frac{p+1}{p}}}\frac{1}{(2\theta)^{\frac{1}{p}}}, \frac{1}{(2p+4)^{\frac{p+1}{p}}\left\{2\max\left\{\frac{2^{\frac{-\beta+1}{p+\beta-1}}}{p+2}L^{\frac{p}{p+\beta-1}}, \frac{L}{(1-\alpha)((2p+4)^{(\beta-1)/(p+\beta-1)}}\right\}\right\}^{\frac{1}{p}}}\right\}. \tag{23}$$

Then,

$$f(x^{k+1}) \leq f(x^k) - \alpha c_p \varepsilon^{\frac{p+\beta}{p+\beta-1}}, \tag{24}$$

Proof Define

$$\sigma_{max} = 2\max\left\{\theta, \varepsilon^{\frac{\beta-1}{p+\beta-1}}\max\left\{\frac{2^{\frac{-\beta+1}{p+\beta-1}}}{p+2}L^{\frac{p}{p+\beta-1}}, \frac{L}{(1-\alpha)((2p+4)^{(\beta-1)/(p+\beta-1)}}\right\}\right\} \tag{25}$$

By Lemma 2.1 we have that $\sigma_k \leq \sigma_{max}$ for all k. Therefore, at each iteration we have:

$$f(x^{k+1}) \leq f(x^k) - \frac{\alpha}{(2p+4)^{\frac{p+1}{p}}}\frac{\varepsilon^{\frac{p+1}{p}}}{\sigma_k^{\frac{1}{p}}} \leq f(x^k) - \frac{\alpha}{(2p+4)^{\frac{p+1}{p}}}\frac{\varepsilon^{\frac{p+1}{p}}}{\sigma_{max}^{\frac{1}{p}}}.$$

Therefore, at each iteration we have that either

$$f(x^{k+1}) \leq f(x^k) - \frac{\alpha}{(2p+4)^{\frac{p+1}{p}}}\frac{\varepsilon^{\frac{p+1}{p}}}{(2\theta)^{\frac{1}{p}}} \tag{26}$$

or

$$f(x^{k+1}) \leq f(x^k) - \alpha\frac{\varepsilon^{\frac{p+1}{p}}}{(2p+4)^{\frac{p+1}{p}}\left\{2\varepsilon^{\frac{\beta-1}{p+\beta-1}}\max\left\{\frac{2^{\frac{-\beta+1}{p+\beta-1}}}{p+2}L^{\frac{p}{p+\beta-1}}, \frac{L}{(1-\alpha)((2p+4)^{(\beta-1)/(p+\beta-1)}}\right\}\right\}^{\frac{1}{p}}}. \tag{27}$$

In the case of (26) the thesis holds by the definition of c_p and $\varepsilon \leq 1$. If (27) takes place, we have that:

$$f(x^{k+1}) \leq f(x^k) - \alpha\frac{\varepsilon^{\frac{p+1}{p}+\frac{1-\beta}{p(p+\beta-1)}}}{(2p+4)^{\frac{p+1}{p}}\left\{2\max\left\{\frac{2^{\frac{-\beta+1}{p+\beta-1}}}{p+2}L^{\frac{p}{p+\beta-1}}, \frac{L}{(1-\alpha)((2p+4)^{(\beta-1)/(p+\beta-1)}}\right\}\right\}^{\frac{1}{p}}}$$

$$\leq f(x^k) - \alpha c_p \varepsilon^{\frac{p+\beta}{p+\beta-1}}.$$

This completes the proof. □

In Theorem 2.2 we will prove that the number of iterations performed by Algorithm 2.1 is, at most, a multiple of $\varepsilon^{-\frac{p+\beta}{p+\beta-1}}$.

Theorem 2.2 *Assume that (2)–(7) hold for $\bar{x} = x^k$ and all the trial points x computed at every iteration k performed by Algorithm 2.1. Let c_p be defined by (23). Then, after, at most,*

$$(f(x^0) - f_{target}) \frac{\varepsilon^{-\frac{p+\beta}{p+\beta-1}}}{\alpha c_p}. \tag{28}$$

iterations, Algorithm 2.1 computes $x \in \mathbb{R}^n$, $\lambda \in \mathbb{R}^m$, and $\mu \in \mathbb{R}^q_+$ verifying $f(x) \leq f_{target}$ or

$$\|g(x) + h'_E(x)^T \lambda + h'_I(x)^T \mu\| \leq \varepsilon, \tag{29}$$

and

$$\|h_E(x)\| \leq \epsilon, \|h_I(x)_+\| \leq \epsilon, \text{ and } \|\min\{\mu, -h_I(x)\}\| \leq \epsilon. \tag{30}$$

Proof The desired result follows from (6), (7), Theorem 2.1 and the stopping criterion at Step 3 of the algorithm. □

Remark

According to Theorem 2.1 the case defined by $p = 1$ and $\beta \approx 0$ is the worst possible one in terms of expected decrease of the objective function at each iteration. In this case the worst expected decrease is $\alpha c_p \varepsilon^{1+\frac{1}{\beta}}$ which, for $\varepsilon < 1$, tends to zero as $\beta \to 0$, predicting a large number of iterations to achieve the stopping criterion. Now, the case $[p = 1, \beta \approx 0]$ corresponds to the fulfillment of a very mild Hölder condition on the gradient of f. Essentially, this implies that one expects only continuity of the gradient and, by (2), very mild approximations of the gradient of the function with respect to the gradient of the model. This situation resemble the case of derivative-free optimization methods in which a model of the objective function is built with very low expectations in terms of accuracy with respect to the objective function [11]. The practical consequence seems to be that, when one has very little knowledge about the correspondence between function and model, Algorithm 2.1, or some variation of it, could be used with $p = 1$.

In order to complete the complexity analysis in terms of functional evaluations, we must add the number of times in which σ is increased at Step 4. By Lemma 2.1, increasing σ will not necessary if $\sigma \geq \sigma_{max}$. Therefore, the maximal number of increases of σ will be bounded by a multiple of $-\log_2(\varepsilon)$. Recall that the commplexity analysis in this paper involves only evaluations of f and its derivatives and does not take into account the evaluations of the functions that define the constraints.

Theorem 2.3 *Assume that the hypotheses of Theorem 2.2 hold. Then, the number of evaluations of f employed by Algorithm 2.1 is bounded above by*

$$(f(x^0)-f_{target})\frac{\varepsilon^{-\frac{p+\beta}{p+\beta-1}}}{\alpha c_p}+\left[\max\left\{\log_2(\theta),\left(\frac{1-\beta}{p+\beta-1}\log_2(\varepsilon^{-1})+c_\ell\right)\right\}\right]-\log_2(\sigma_{min})+1. \quad (31)$$

where

$$c_\ell=\log_2(\max\left\{\frac{2^{\frac{-\beta+1}{p+\beta-1}}}{p+2}L^{\frac{p}{p+\beta-1}},\frac{L}{(1-\alpha)((2p+4)^{(\beta-1)/(p+\beta-1)}}\right\}).$$

Proof By Theorem 2.2, the evaluation of f at each iterate x^k is responsible for the first term of (31). The remaining functional evaluations are at trial points at which the descent criterion (9) does not hold. Each time (9) fails the regularization parameter σ is doubled. However, by Lemma 2.1, if $\sigma \geq \sigma_{max}$, (9) necessarily takes place. This means that the number of times at which σ is doubled is bounded above by $\log_2(\sigma_{max}/\sigma_{min})$. Therefore, by (25), the second term of (31) is

$$\log_2\left[2\max\left\{\theta,\varepsilon^{\frac{\beta-1}{p+\beta-1}}\max\left\{\frac{2^{\frac{-\beta+1}{p+\beta-1}}}{p+2}L^{\frac{p}{p+\beta-1}},\frac{L}{(1-\alpha)((2p+4)^{(\beta-1)/(p+\beta-1)}}\right\}\right\}\right]-\log_2(\sigma_{min})$$

$$=\left[\max\left\{\log_2(\theta),\left(\frac{\beta-1}{p+\beta-1}\log_2(\varepsilon)+c_\ell\right)\right\}\right]-\log_2(\sigma_{min})+1.$$

This completes the proof. □

3 Conclusions

When the paper [4] appeared, several optimization colleagues expressed their concerns with respect to the potential computational usefulness of high order complexity results. However, the two papers that inspired the present research [18, 10] and, hopefully, this contribution too, indicate that the complexity analysis [4] may contribute to motivate the introduction of useful computer algorithms. As far as we try to solve complex problems, we face functional structures in which nonsmooth and semismooth expressions are combined with algorithmically determined terms and the results of simulations. In such conditions good surrogate models are necessary in order to replace the objective function, with few limitations about their algebraic complications. The iterative minimization of relatively complex local models, aiming practical efficiency, should be stimulated by the existence of robust complexity results. Of course, the necessity of maintaining (almost) feasibility at every iteration is a limitation. However, on the one hand, important problems exist in which minimizing models with non-trivial constraints is affordable. For example, in the field of electronic structure calculations, one minimizes on the set symmetric idempotent matrices with regularization or trust-region schemes and acceptable results are frequently obtained [15, 16]. Moreover, problems of minimization over the Stiefel manifold were solved in [17] using regularization or trust regions. Fortunately, our theory indicates that subproblems may be solved with very mild accuracy preserving overall complexity. This establishes a connection between our almost-feasible methods with gradient-like constrained procedures as

[17], since very few iterations of gradient-like or gradient-projection methods may be used to solve the subproblems. On the other hand, the problem arises of extending the present results to situations in which feasibility is not easy to maintain, perhaps along the lines of [5].

In principle, few considerations need to be made with respect to practical implementations of Algorithm 2.1. In order to obtain the best possible complexity results, in the presented algorithm the sequence $\{\sigma_k\}$ is non-decreasing. So, the regularization parameter is multiplied by 2 when the descent criterion is not fulfilled, and the total number of function evaluations at points that do not satisfy (9) is bounded by a multiple of the logarithm of $1/\varepsilon$. (Of course, the constant 2 is arbitrary and may be replaced by any number bigger tnan 1.) The consequence is that the regularization parameter could be driven to be very large at early iterations and maintained on that level when this is not necessary anymore, causing unnecessarily small steps. So, on average, it is better to start every iteration with a (very) small regularization parameter in order to exploit possible accuracy of the model approximation, especially when the iterate is close to the solution. Such modification causes a marginal worsening of the worst-case complexity, i.e. the maximal number of evaluations would be the product of the number of iterations and a multiple of the logarithm of $1/\varepsilon$, but should be better in practice. Clearly, the potentiality of this approach relies on its applicability to many different particular situations, which would dictate many problem-motivated implementation features.

References

[1] R. Andreani, G. Haeser, and J. M. Martínez, On sequential optimality conditions for smooth constrained optimization, *Optimization* 60, pp. 627-641, 2011.

[2] R. Andreani, J. M. Martínez, and B. F. Svaiter, A new sequential optimality condition for constrained optimization and algorithmic consequences, *SIAM Journal on Optimization* 20, pp. 3533-3554, 2010.

[3] D. P. Bertsekas, Nonlinear Programming, Athena Scientific, Belmont, Massachussets, USA, 2nd Edition, 1999.

[4] E. G. Birgin, J. L. Gardenghi, J. M. Martínez, S. A. Santos, and Ph. L. Toint, Worst-case evaluation complexity for unconstrained nonlinear optimization using high-order regularized models, *Mathematical Programming*, to appear (DOI: 10.1007/s10107-016-1065-8).

[5] E. G. Birgin, J. L. Gardenghi, J. M. Martínez, S. A. Santos, and Ph. L. Toint, Evaluation complexity for nonlinear constrained optimization using unscaled KKT conditions and high-order models, *SIAM Journal on Optimization 26*, pp. 951-967, 2016.

[6] E. G. Birgin and J. M. Martínez,*Practical Augmented Lagrangian Methods for Constrained Otimization*, SIAM Publications, Series: Fundamentals of Algorithms, Philadelphia, 2014.

[7] 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, pp. 2833–2852, 2010.

[8] C. Cartis, N. I. M. Gould, and Ph. L. Toint, Adaptive cubic regularization methods for unconstrained optimization. Part I: motivation motivation, convergence and numerical results, *Mathematical Programming* 127, pp. 245–295, 2011.

[9] C. Cartis, N. I. M. Gould, and Ph. L. Toint, Adaptive cubic regularization methods for unconstrained optimization. Part II: worst-case function and derivative complexity, *Mathematical Programming* 130, pp. 295–319, 2011.

[10] C. Cartis, N. I. M. Gould, and Ph. L. Toint, Universal regularization methods - varying the power, the smoothness and the accuracy, Preprint RAL-P-2016-010, Rutherford Appleton Laboratory, Chilton, England, 2016.

[11] A. R. Conn, K. Scheinberg, and L. N. Vicente, *Introduction to Derivative-Free Optimization*, MPS-SIAM Series on Optimization, SIAM, Philadelphia, 2009.

[12] F. E. Curtis, D. P. Robinson, and M. Samadi, A trust-region algorithm with a worst-case iteration complexity of $O(\varepsilon^{-3/2})$, *Mathematical Programming*, to appear (DOI: 10.1007/s10107-016-1026-2).

[13] J. E. Dennis Jr. and R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Society for Industral and Applied Mathematics, Philadelphia, 1996.

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

[15] J. B. Francisco, J. M. Martínez and L. Martínez, Globally convergent Trust-Region methods for Self-Consistent Field electronic structure calculations. *Journal of Chemical Physics* 121, pp. 10863-10878, 2004.

[16] J. B. Francisco, J. M. Martínez and L. Martínez. Density-Based Globally Convergent Trust-Region Method for Self-Consistent Field Electronic Structure Calculations. *Journal of Mathematical Chemistry* 40, pp. 349-377, 2006.

[17] J. B. Francisco and F. S. Viloche-Bazán, Nonmonotone algorithm for minimization on closed sets with application to minimization on Stiefel manifolds, *Journal of Computational and Applied Mathematics* 236, pp. 2717-2727, 2012.

[18] 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.

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

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

[21] C. Humes and P. J. S. Silva, Convex Regularizations, Proximal Point and Augmented Lagrangians, *RAIRO Operations Research* 34 pp. 283-303, 2000

[22] J. M. Martínez and M. Raydan, Separable cubic modeling and a trust-region strategy for unconstrained minimization with impact in global optimization, *Journal of Global Optimization* 63, pp. 315-342, 2015.

[23] J. M. Martínez and M. Raydan, Cubic-regularization counterpart of a variable-norm trust-region method for unconstrained minimization, *Journal of Global Optimization*, to appear (DOI: 10.1007/s10898-016-0475-8).

[24] Y. Nesterov and B. T. Polyak, Cubic regularization of Newton's method and its global performance, *Mathematical Programming* 108, pp. 177–205, 2006.

[25] R. T. Rockafellar, Augmented Lagrangians and applications of the proximal point algorithm in convex programming, *Mathematics of Operations Research* 1, pp. 97-116, 1976.

[26] R. T. Rockafellar, Monotone operators and the proximal point algorithm, *SIAM Journal on Control and Optimization* 14, pp. 877–898, 1976.

[27] R. B. Schnabel and T.-T. Chow, Tensor methods for unconstrained optimization using second derivatives, *SIAM Journal on Optimization* 1, pp. 293-315, 1991.

[28] M. V. Solodov and B. F. Svaiter, A unified framework for some inexact proximal point algorithms, *Numerical Functional Analysis and Optimization*, pp. 1013-1035, 2001.

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