Convergence properties of an Objective-Function-Free Optimization regularization algorithm, including an $\mathcal{O}(\epsilon^{-3/2})$ complexity bound

S. Gratton*, S. Jerad† and Ph. L. Toint‡

18 III 2022

Abstract

An adaptive regularization algorithm for unconstrained nonconvex optimization is presented in which the objective function is never evaluated, but only derivatives are used. This algorithm belongs to the class of adaptive regularization methods, for which optimal worst-case complexity results are known for the standard framework where the objective function is evaluated. It is shown in this paper that these excellent complexity bounds are also valid for the new algorithm, despite the fact that significantly less information is used. In particular, it is shown that, if derivatives of degree one to p are used, the algorithm will find a ϵ_1-approximate first-order minimizer in at most $\mathcal{O}(\epsilon_1^{-(p+1)/p})$ iterations, and an (ϵ_1, ϵ_2)-approximate second-order minimizer in at most $\mathcal{O}(\max[\epsilon^{-(p+1)/p}, \epsilon_2^{-(p+1)/(p-1)}])$ iterations. As a special case, the new algorithm using first and second derivatives, when applied to functions with Lipschitz continuous Hessian, will find an iterate x_k at which the gradient's norm is less than ϵ_1 in at most $\mathcal{O}(\epsilon_1^{-3/2})$ iterations.

Keywords: nonlinear optimization, adaptive regularization methods, evaluation complexity, objective-function-free optimization (OFFO).

1 Introduction

This paper is about the (complexity-wise) fastest known optimization method which does not evaluate the objective function. Such methods, coined OFFO for Objective-Function-Free Optimization, have recently been very popular in the context of noisy problems, in particular in deep learning applications (see [24, 17, 30, 29] among many others), where they have shown remarkable insensitivity to the noise level. This is a first motivation to consider them, and it is our point of view that their deterministic (noiseless) counterparts are good stepping stones to understand their behaviour. Another motivation is the observation that other more standard methods (using objective function evaluations) have been proposed in the noisy case, but

*Université de Toulouse, INP, IRIT, Toulouse, France. Email: serge.gratton@enseeiht.fr. Work partially supported by 3IA Artificial and Natural Intelligence Toulouse Institute (ANITI), French "Investing for the Future - PIA3" program under the Grant agreement ANR-19-PI3A-0004"

†ANITI, Université de Toulouse, INP, IRIT, Toulouse, France. Email: sadok.jerad@toulouse-inp.fr

‡NAXYS, University of Namur, Namur, Belgium. Email: philippe.toint@unamur.be. Partly supported by ANITI.

typically require the noise on the function values to be tightly controlled at a level lower than that allowed for derivatives [12, 13, 6, 15, 4, 3, 2, 1]

The convergence anaysis of OFFO algorithms is not a new subject, and has been considered for instance in [16, 29, 21, 20, 18, 30]. However, as far as the authors are aware, the existing theory focuses on the case where only gradients are used (with the exception of [23]) and establish a worst-case iteration complexity of, at best, $\mathcal{O}(\epsilon^{-2})$ for finding an ϵ-approximate first-order stationary point [26] . It is already remarkable that this bound is, in order and for the same goal, identical to that of standard methods using function values. But methods using second-derivatives have proved to be globally more efficient in this latter context, , and the (complexity-wise) fastest such method is known to have an $\mathcal{O}(\epsilon^{-3/2})$ complexity bound [27, 14, 28, 8, 5, 11]. Moreover, this better bound was shown to be sharp and optimal among a large class of optimization algorithms using second-derivatives for the noiseless case [9]. Is such an improvement in complexity also possible for (noiseless) OFFO algorithms? We answer this question positively in what follows.

The theory developed here combines elements of standard adaptive regularization methods such as ARp [5] and of the OFFO approaches of [30] and [18]. We exhibit an OFFO regularization method whose iteration complexity is identical to that obtained when objective function values are used. In particular, we consider convergence to approximate first-order and second-order critical points, and provide sharp complexity bounds depending on the degree of derivatives used.

The paper is organized as follows. After introducing the new algorithm in Section 2, we present a first-order worst-case complexity analysis in Section 3, while convergence to approximate second-order minimizers is considered in Section 4. The results are then discussed in Section 5 and some conclusions and perspectives outlined in Section 6.

2 An OFFO adaptive regularization algorithm

We now consider the problem of finding approximate minimizers of the unconstrained non-convex optimization problem

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

where f is a sufficiently smooth function from $\mathbb{R}^n$ into $\mathbb{R}$. As motivated in the introduction, our aim is to design an algorithm in which the objective function value is never computed. Our approach is based on regularization methods. In such methods, a model of the objective function is build by "regularizing" a truncated Taylor expansion of degree $p \geq 1$. We now detail the assumption on the problems that ensure this approach makes sense.

AS.1 f is p times continuously Fréchet differentiable.

AS.2 There exists a constant $f_{\rm low}$ such that $f(x) \geq f_{\rm low}$ for all $x \in \mathbb{R}^n$.

AS.3 The pth derivative of f is globally Lipshitz continuous, that is, there exist a non-negative constant L_p such that

$$\|\nabla_x^p f(x) - \nabla_x^p f(y)\| \leq L_p\|x-y\| \ \text{ for all } x,y \in \mathbb{R}^n, \tag{2.2}$$

where $\|.\|$ denotes the usual Euclidean norm in $\mathbb{R}^n$.

AS.4 If $p > 1$, there exists a constant $\kappa_{\rm high} \geq 0$ such that

$$\min_{\|d\|\leq 1} \nabla_x^i f(x)[d]^i \geq -\kappa_{\rm high} \ \text{ for all } \ x \in \mathbb{R}^n \ \text{ and } \ i \in \{2,\ldots,p\}, \tag{2.3}$$

where $\nabla_x^i f(x)$ is the ith derivative tensor of f computed at x, and where $T[d]^i$ denotes the i-dimensional tensor T applied on i copies of the vector d. (For notational convenience, we set $\kappa_{\rm high} = 0$ if $p = 1$.)

We note that AS.4 is weaker than assuming uniform boundedness of the derivative tensors of degree two and above (there is no upper bound on the value of $\nabla_x^i f(x)[d]^i$), or, equivalently, Lipschitz continuity of derivatives of degree one to $p-1$.

2.1 The OFFARp algorithm

Adaptive regularization methods are iterative schemes which compute a step from an iterate x_k to the next by approximately minimizing a pth degree regularized model $m_k(s)$ of $f(x_k+s)$ of the form

$$m_k(s) \stackrel{\rm def}{=} T_{f,p}(x_k, s) + \frac{\sigma_k}{(p+1)!}\|s\|^{p+1}, \tag{2.4}$$

where $T_{f,p}(x, s)$ is the pth order Taylor expansion of functional f at x truncated at order p, that is,

$$T_{f,p}(x, s) \stackrel{\rm def}{=} f(x) + \sum_{i=1}^{p} \frac{1}{i!}\nabla_x^i f(x)[s]^i. \tag{2.5}$$

In (2.4), the pth order Taylor series is "regularized" by adding the term $\frac{\sigma_k}{(p+1)!}\|s\|^{p+1}$, where σ_k is known as the "regularization parameter". This term guarantees that $m_k(s)$ is bounded below and thus makes the procedure of finding a step s_k by (approximately) minimizing $m_k(s)$ well-defined. Our proposed algorithm follows the outline line of existing ARp regularization methods [8, 5, 11], with the significant difference that the objective function $f(x_k)$ is never computed, and therefore that the ratio of achieved to predicted reduction (a standard feature for these methods) cannot be used to accept or reject a potential new iterate and to update the regularization parameter. Instead, such potential iterates are always accepted and the regularization parameter is updated in a manner independent of this ratio. We now state the resulting OFFARp algorithm in detail.

The test (2.9) follows [22] and extends the more usual condition where the step s_k is chosen to ensure that

$$\|\nabla_s^1 m_k(s_k)\| \leq \theta_1 \|s_k\|^p.$$

It is indeed easy to verify that (2.9) holds at a local minimizer of m_k with $\theta_1 \geq 1$ (see [22] for details).

3 Evaluation complexity for the OFFARp algorithm

Before discussing our analysis of evaluation complexity, we first restate some classical lemmas of ARp algorithms, starting with Lipschitz error bounds.

Algorithm 2.1: OFFO adaptive regularization of degree p (OFFARp)

Step 0: Initialization: An initial point $x_0 \in \mathbb{R}^n$, a regularization parameter $v_0 = \sigma_0 > 0$ and a requested final gradient accuracy $\epsilon_1 \in (0, 1]$ are given, as well as the parameters

$$\theta_1 > 1 \quad \text{and} \quad \vartheta \in (0, 1]. \tag{2.6}$$

Set $k = 0$.

Step 1: Check for termination: Evaluate $g_k = \nabla_x^1 f(x_k)$. Terminate with $x_\epsilon = x_k$ if

$$\|g_k\| \leq \epsilon_1. \tag{2.7}$$

Else, evaluate $\{\nabla_x^i f(x_k)\}_{i=2}^p$.

Step 2: Step calculation: Compute a step s_k which sufficiently reduces the model m_k defined in (2.4) in the sense that

$$m_k(s_k) - m_k(0) < 0 \tag{2.8}$$

and

$$\|\nabla_s^1 T_{f,p}(x_k, s_k)\| \leq \theta_1 \frac{\sigma_k}{p!} \|s_k\|^p. \tag{2.9}$$

Step 3: Updates. Set

$$x_{k+1} = x_k + s_k, \tag{2.10}$$

$$v_{k+1} = v_k + v_k \|s_k\|^{p+1} \tag{2.11}$$

and select

$$\sigma_{k+1} \in [\vartheta v_{k+1}, v_{k+1}]. \tag{2.12}$$

Increment k by one and go to Step 1.

Lemma 3.1 Suppose that AS.1 and AS.3 hold. Then

$$|f(x_{k+1}) - T_{f,p}(x_k, s_k)| \leq \frac{L_p}{(p+1)!}\|s_k\|^{p+1}, \tag{3.1}$$

and

$$\|g_{k+1} - \nabla_s^1 T_{f,p}(x_k, s_k)\| \leq \frac{L_p}{p!}\|s_k\|^p. \tag{3.2}$$

Proof. This is a standard result (see [10, Lemma 2.1] for instance). □

We start by stating a simple lower bound on the Taylor series' decrease.

Lemma 3.2

$$\Delta T_{f,p}(x_k, s_k) \stackrel{\rm def}{=} T_{f,p}(x_k, 0) - T_{f,p}(x_k, s_k) \geq \frac{\sigma_k}{(p+1)!}\|s_k\|^{p+1}. \tag{3.3}$$

Proof. The bound directly results from (2.8) and (2.4). □

This and AS.2 allow us to establish a lower bound on the decrease in the objective function (although it is never computed).

Lemma 3.3 Suppose that AS.1 and AS.3 hold and that $\sigma_k \geq 2L_p$. Then

$$f(x_k) - f(x_{k+1}) \geq \frac{\sigma_k}{2(p+1)!}\|s_k\|^{p+1}. \tag{3.4}$$

Proof. From (3.1) and (3.3), we obtain that

$$f(x_k) - f(x_{k+1}) \geq \frac{\sigma_k - L_p}{(p+1)!}\|s_k\|^{p+1}$$

and (3.4) immediately follows from our assumption on σ_k. □

The next lemma provides a useful lower bound on the step length, in the spirit of [5, Lemma 2.3] or [22].

Lemma 3.4 Suppose that AS.1 and AS.3 hold. Then

$$\|s_k\|^p > \frac{p!}{L_p + \theta_1\sigma_k}\|g(x_{k+1})\|. \tag{3.5}$$

Proof. Successively using the triangle inequality,condition (2.9) and (3.2), we deduce that

$$\|g(x_{k+1})\| \leq \|g(x_{k+1}) - \nabla_s^1 T_{f,p}(x_k, s_k)\| + \|\nabla_s^1 T_{f,p}(x_k, s_k)\| \leq \frac{1}{p!} L_p \|s_k\|^p + \theta_1 \frac{\sigma_k}{p!} \|s_k\|^p.$$

The inequality (3.5) follows by rearranging the terms. □

Inspired by [18, Lemma 7], we now establish an upper bound on the number of iterations needed to enter the algorithm's phase where Lemma 3.3 applies and thus all iterations produce a decrease in the objective function.

Lemma 3.5 Suppose that AS.1 and AS.3 hold, and that the OFFARp algorithm does not terminate before or at iteration of index

$$k \geq k_* \stackrel{\rm def}{=} \left\lceil \left(\frac{2L_p(L_p + \theta_1 \sigma_0)}{p! \vartheta \sigma_0 \epsilon_1} \right)^{\frac{p+1}{p}} \right\rceil. \tag{3.6}$$

Then,

$$v_k \geq \frac{2L_p}{\vartheta} \tag{3.7}$$

which implies that

$$\sigma_k \geq 2L_p. \tag{3.8}$$

Proof. Note that (3.8) is a direct consequence of (2.12) if (3.7) is true. Suppose the opposite and that for some $k \geq k_*$, $v_k < \frac{2L_p}{\vartheta}$. Since v_k is a non-decreasing sequence, we have that $v_j < \frac{2L_p}{\vartheta}$ for $j \in \{0, \ldots, k\}$. Successively using the form of the v_k update rule (2.11), (3.5), (2.12) and the fact that, if the algorithm has reached iteration k_*, it must be that (2.7) has failed for all iterations of index at most k_*, we derive that

$$\begin{aligned}
v_k &> \sum_{j=0}^{k-1} v_j \|s_j\|^{p+1} \geq \sum_{j=0}^{k-1} v_j \left(\frac{p! \|g(x_{j+1})\|}{L_p + \theta_1 \sigma_j} \right)^{\frac{p+1}{p}} \geq \sum_{j=0}^{k-1} v_j \left(\frac{p! \|g(x_{j+1})\|}{L_p + \theta_1 v_j} \right)^{\frac{p+1}{p}} \\
&= \sum_{j=0}^{k-1} v_j^{-\frac{1}{p}} \left(\frac{p! \|g(x_{j+1})\|}{\frac{L_p}{v_j} + \theta_1} \right)^{\frac{p+1}{p}} > \sum_{j=0}^{k-1} v_j^{-\frac{1}{p}} \left(\frac{p! \|g(x_{j+1})\|}{\frac{L_p}{\sigma_0} + \theta_1} \right)^{\frac{p+1}{p}} \\
&= \sum_{j=0}^{k-1} v_j^{-\frac{1}{p}} \left(\frac{p! \sigma_0 \|g(x_{j+1})\|}{L_p + \theta_1 \sigma_0} \right)^{\frac{p+1}{p}} > \frac{k_* \, \vartheta^{\frac{1}{p}} (p! \sigma_0 \epsilon_1)^{\frac{p+1}{p}}}{(2L_p)^{\frac{1}{p}} (L_p + \theta_1 \sigma_0)^{\frac{p+1}{p}}}.
\end{aligned}$$

Substituting the definition of k_* in the last inequality, we obtain that

$$\frac{2L_p}{\vartheta} < v_{k_*} < \frac{2L_p}{\vartheta},$$

which is impossible. Hence no index $k \geq k_*$ exists such that $v_k < \frac{2L_p}{\vartheta}$ and (3.7) and (3.8) hold. □

We now define

$$k_1 = \min\left\{ k \geq 1 \mid v_k \geq \frac{2L_p}{\vartheta} \right\}, \tag{3.9}$$

the first iterate such that significant objective function decrease is guaranteed. The next series of Lemmas provide bounds on $f(x_{k_1})$ and σ_{k_1}, which in turn will allow establishing an upper bound on the regularization paremeter. We start by proving an upper bound on s_k generalizing those proposed in [7, 22] to the case where p is arbitrary.

Lemma 3.6 Suppose that AS.1 and AS.4 holds. At each iteration k, we have that for

$$\|s_k\| \leq 2\eta + 2\left(\frac{(p+1)!\|g_k\|}{\sigma_k}\right)^{\frac{1}{p}}, \tag{3.10}$$

where

$$\eta = \sum_{i=2}^{p} \left[\frac{\max[0, -\kappa_{\rm high}](p+1)!}{i!\,\vartheta v_0}\right]^{\frac{1}{i}}. \tag{3.11}$$

Proof. If $p = 1$, we obtain from (2.8) and the Cauchy-Schwarz inequality that

$$\tfrac{1}{2}\sigma_k\|s_k\|^2 < -g_k^T s_k \leq \|g_k\|\,\|s_k\|$$

and (3.10) holds with $\eta = 0$. Suppose now that $p > 1$. Again (2.8) gives that

$$\frac{\sigma_k}{(p+1)!}\|s_k\|^{p+1} \leq -g_k^T s_k - \sum_{i=2}^{p} \nabla_x^i f(x_k)[s_k]^i \leq \|g_k\|\|s_k\| + \sum_{i=2}^{p} \frac{\max[0, -\kappa_{\rm high}]}{i!}\|s_k\|^i.$$

Applying now the Lagrange bound for polynomial roots [31, Lecture VI, Lemma 5] with $x = \|s_k\|$, $n = p+1$, $a_0 = 0$, $a_1 = \|g_k\|$, $a_i = \max[0, -\kappa_{\rm high}]/i!$ $i \in \{2, \ldots, p\}$ and $a_{p+1} = \sigma_k/(p+1)!$, we know from (2.8) that the equation $\sum_{i=0}^{n} a_i x^i = 0$ admits at least a strictly positive root, and we may thus derive that

$$\begin{aligned}
\|s_k\| &\leq 2\left(\frac{(p+1)!\|g_k\|}{\sigma_k}\right)^{\frac{1}{p}} + 2\sum_{i=2}^{p}\left[\frac{\max[0, -\kappa_{\rm high}](p+1)!}{i!\,\sigma_k}\right]^{\frac{1}{i}} \\
&\leq 2\left(\frac{(p+1)!\|g_k\|}{\sigma_k}\right)^{\frac{1}{p}} + 2\sum_{i=2}^{p}\left[\frac{\max[0, -\kappa_{\rm high}](p+1)!}{i!\,\vartheta v_k}\right]^{\frac{1}{i}} \\
&\leq 2\left(\frac{(p+1)!\|g_k\|}{\sigma_k}\right)^{\frac{1}{p}} + 2\sum_{i=2}^{p}\left[\frac{\max[0, -\kappa_{\rm high}](p+1)!}{i!\,\vartheta v_0}\right]^{\frac{1}{i}},
\end{aligned}$$

and (3.10) holds with (3.11). □

Our next step is to prove that v_{k_1} is bounded by constants only depending on the problem and the fixed algorithmic parameters.

Lemma 3.7 Suppose that AS.1, AS.3 and AS.4 hold. Let k_1 be defined by (3.9). We have that,

$$v_{k_1} \leq v_{\max} = \max\left[\sigma_0 + \sigma_0\left(2\eta + 2\left(\frac{(p+1)!\|g_0\|}{\sigma_0}\right)^{\frac{1}{p}}\right)^{p+1}, \kappa_1\frac{2L_p}{\vartheta}\right] \tag{3.12}$$

where

$$\kappa_1 \stackrel{\rm def}{=} 1 + 2^{2p+1}\eta^{p+1} + 2^{2p+1}\left[\frac{(p+1)\left(\theta_1 + \frac{L_p}{\vartheta\sigma_0}\right)}{\vartheta}\right]^{\frac{p+1}{p}}. \tag{3.13}$$

Proof. if $k_1 = 1$, we have that

$$v_1 = \sigma_0 + \sigma_0\|s_0\|^{p+1}.$$

Using Lemma 3.6 to bound $\|s_0\|^{p+1}$, we derive the first part of (3.12). Suppose now that $k_1 \geq 2$. Successively using (2.11), Lemma 3.6, the fact that $(x+y)^{p+1} \leq 2^p(x^{p+1}+y^{p+1})$, the updates rule of v_k (2.11) and σ_k (2.12) and Lemma 3.4, we derive that,

$$\begin{aligned}
v_{k_1} &= v_{k_1-1} + v_{k_1-1}\|s_{k_1-1}\|^{p+1} \\
&\leq v_{k_1-1} + v_{k_1-1}\left(2\left((p+1)!\frac{\|g_k\|}{\sigma_k}\right)^{\frac{1}{p}} + 2\eta\right)^{p+1} \\
&\leq v_{k_1-1} + 2^p v_{k_1-1}\left[2^{p+1}\eta^{p+1} + 2^{p+1}\left(\frac{(p+1)!\|g_{k_1-1}\|}{\sigma_{k_1-1}}\right)^{\frac{p+1}{p}}\right] \\
&\leq v_{k_1-1} + 2^{2p+1}v_{k_1-1}\left[\eta^{p+1} + \left(\frac{(p+1)!\|g_{k_1-1}\|}{\vartheta v_{k_1-1}}\right)^{\frac{p+1}{p}}\right] \\
&\leq v_{k_1-1} + 2^{2p+1}v_{k_1-1}\eta^{p+1} + 2^{2p+1}\left(\frac{(p+1)!}{\vartheta}\right)^{\frac{p+1}{p}}\frac{\|g_{k_1-1}\|^{\frac{p+1}{p}}}{v_{k_1-1}^{\frac{1}{p}}} \\
&\leq v_{k_1-1} + 2^{2p+1}v_{k_1-1}\eta^{p+1} \\
&\qquad + 2^{2p+1}\left[\frac{(p+1)!(L_p+\theta_1\sigma_{k_1-2})}{\vartheta p!}\right]^{\frac{p+1}{p}} v_{k_1-1}^{-\frac{1}{p}}\|s_{k_1-2}\|^{p+1}.
\end{aligned}$$

Now v_k is a non decreasing sequence, and therefore

$$\begin{aligned}
v_{k_1} &\leq v_{k_1-1} + 2^{2p+1} v_{k_1-1} \eta^{p+1} \\
&\quad + 2^{2p+1} \left[\frac{(p+1)!(L_p + \theta_1 \sigma_{k_1-2})}{\vartheta p!} \right]^{\frac{p+1}{p}} v_{k_1-2}^{-\frac{1}{p}} \|s_{k_1-2}\|^{p+1} \\
&\leq v_{k_1-1} + 2^{2p+1} v_{k_1-1} \eta^{p+1} \\
&\quad + 2^{2p+1} \left[\frac{(p+1)! \left(\theta_1 + \frac{L_p}{\sigma_{k_1-2}}\right)}{\vartheta p!} \right]^{\frac{p+1}{p}} v_{k_1-2}^{-\frac{1}{p}} \sigma_{k_1-2}^{\frac{p+1}{p}} \|s_{k_1-2}\|^{p+1} \\
&\leq v_{k_1-1} + 2^{2p+1} v_{k_1-1} \eta^{p+1} + 2^{2p+1} \left[\frac{(p+1)! \left(\theta_1 + \frac{L_p}{\vartheta \sigma_0}\right)}{\vartheta p!} \right]^{\frac{p+1}{p}} v_{k_1-2} \|s_{k_1-2}\|^{p+1} \\
&\leq v_{k_1-1} + 2^{2p+1} v_{k_1-1} \eta^{p+1} + 2^{2p+1} \left[\frac{(p+1)! \left(\theta_1 + \frac{L_p}{\vartheta \sigma_0}\right)}{\vartheta p!} \right]^{\frac{p+1}{p}} (v_{k_1-1} - v_{k_1-2}) \\
&\leq v_{k_1-1} + 2^{2p+1} v_{k_1-1} \eta^{p+1} + 2^{2p+1} \left[\frac{(p+1)! \left(\theta_1 + \frac{L_p}{\vartheta \sigma_0}\right)}{\vartheta p!} \right]^{\frac{p+1}{p}} v_{k_1-1}.
\end{aligned}$$

We then obtain the second part of (3.12) by observing that $v_{k_1-1} \leq \frac{2L_p}{\vartheta}$. □

This result allows us to establish an upperbound on $f(x_{k_1})$ as a function of $v_{\max}$.

Lemma 3.8 Suppose that AS.1, AS.3 and AS.4 hold. Then

$$f(x_{k_1}) \leq f(x_0) + \frac{L_p v_{\max} + \vartheta \sigma_0^2}{(p+1)! \sigma_0}. \tag{3.14}$$

Proof. From (3.1) and (3.3), we know that

$$f(x_{j+1}) - f(x_j) \leq (L_p - \sigma_j) \frac{\|s_j\|^{p+1}}{(p+1)!}. \tag{3.15}$$

Using now (2.11) and the fact that v_k is a non-decreasing function, we derive that

$$v_{k_1} \geq \sigma_0 + \sigma_0 \sum_{j=0}^{k_1-1} \|s_j\|^{p+1}. \tag{3.16}$$

Summing the inequality (3.15) for $j \in \{0, \ldots, k_1 - 1\}$ and using (3.16), (2.11) and (2.12),

we deduce that

$$\begin{aligned} f(x_{k_1}) &\leq f(x_0) + \frac{L_p}{(p+1)!}\sum_{j=0}^{k_1-1}\|s_j\|^{p+1} - \frac{1}{(p+1)!}\sum_{j=0}^{k_1-1}\sigma_j\|s_j\|^{p+1} \\ &\leq f(x_0) + \frac{L_p}{(p+1)!}\left(\frac{v_{k_1}-\sigma_0}{\sigma_0}\right) - \frac{1}{(p+1)!}\sum_{j=0}^{k_1-1}\vartheta v_j\|s_j\|^{p+1} \\ &\leq f(x_0) + \frac{L_p}{(p+1)!}\left(\frac{v_{k_1}-\sigma_0}{\sigma_0}\right) - \frac{\vartheta}{(p+1)!}(v_{k_1}-\sigma_0). \end{aligned}$$

We then obtain (3.14) by ignoring the negative terms in the right-hand side of this last inequality and using Lemma 3.7 to bound v_{k_1}. □

The two bounds in Lemma 3.8 and Lemma 3.7 are useful in that they now imply an upper bound on the regularization parameter, a crucial step in standard theory for regularization methods.

Lemma 3.9 Suppose that AS.1, AS.3 and AS.4 hold. Suppose also that $k \geq k_1$. Then

$$\sigma_k \leq \sigma_{\max} \stackrel{\rm def}{=} \frac{2(p+1)!}{\vartheta}\left(f(x_0) - f_{\rm low} + \frac{L_p v_{\max} + \vartheta\sigma_0^2}{(p+1)!\sigma_0}\right) + v_{\max}. \tag{3.17}$$

Proof. Let $j \in \{k_1, \ldots, k\}$. By the definition of k_1 in (3.9), $\sigma_j \geq 2L_p$. From Lemma 3.3, we then have that

$$f(x_j) - f(x_{j+1}) \geq \frac{\sigma_j}{2(p+1)!}\|s_j\|^{p+1} \geq \vartheta\frac{v_j}{2(p+1)!}\|s_j\|^{p+1}.$$

Summing the previous inequality from $j = k_1$ to $k-1$ and using the v_j update rule (2.11) and AS.2, we deduce that

$$f(x_{k_1}) - f_{\rm low} \geq f(x_{k_1}) - f(x_k) \geq \frac{\vartheta}{2(p+1)!}(v_k - v_{k_1}).$$

Rearranging the previous inequality and using Lemma 3.7,

$$v_k \leq \frac{2(p+1)!}{\vartheta}\left(f(x_{k_1}) - f_{\rm low}\right) + v_{\max}.$$

Combining now Lemma 3.8 (to bound $f(x_{k_1})$) and the fact that $\sigma_j \leq v_j$ gives (3.17). □

We may now resort to the standard "telescoping sum" argument to obtain the desired evaluation complexity bound.

Theorem 3.10 Suppose that AS.1–AS.4 hold. Then the OFFARp algorithm requires at most

$$\left[\kappa_{\text{OFFARp}}\left(f(x_0)-f_{\text{low}}+\frac{L_p v_{\max}+\vartheta\sigma_0^2}{(p+1)!\sigma_0}\right)+\left(\frac{2L_p(L_p+\theta_1\sigma_0)}{p!\vartheta\sigma_0}\right)^{\frac{p+1}{p}}\right]\epsilon_1^{-\frac{p+1}{p}}+2$$

iterations and evaluations of $\{\nabla_x^i f\}_{i=1}^p$ to produce a vector $x_\epsilon\in\mathbb{R}^n$ such that $\|g(x_\epsilon)\|\leq\epsilon_1$, where

$$\kappa_{\text{OFFARp}}\stackrel{\text{def}}{=}2(p+1)!\,\sigma_{\max}^{1/p}\left(\frac{L_p+\vartheta\theta_1\sigma_0}{\vartheta p!\sigma_0}\right)^{\frac{p+1}{p}}$$

where $\sigma_{\max}$ is defined in Lemma 3.9 and $v_{\max}$ is defined in Lemma 3.7.

Proof. Suppose that the algorithm terminates at an iteration $k<k_1$, where k_1 is given by (3.9). The desired conclusion then follows from the fact that, by this definition and Lemma 3.5,

$$k_1\leq k_*\leq\left(\frac{2L_p(L_p+\theta_1\sigma_0)}{p!\vartheta\sigma_0\epsilon_1}\right)^{\frac{p+1}{p}}+1. \tag{3.18}$$

Suppose now that the algorithm has not terminated at iteration k_1 and consider an iteration $j\geq k_1$. From k_1 definition (3.9) and Lemma 3.9, we have that $2L_p\leq\sigma_j\leq\sigma_{\max}$. Since $\sigma_j\geq 2L_p$, Lemma 3.3 is valid for iteration j. But $\sigma_j\in[\vartheta\sigma_0,\sigma_{\max}]$ because of Lemma 3.9 and $\|g(x_{j+1})\|\geq\epsilon_1$ before termination, and we therefore deduce that

$$f(x_j)-f(x_{j+1})\geq\frac{\sigma_j\|s_j\|^{p+1}}{2(p+1)!}\geq\frac{\sigma_j(p!)^{\frac{p+1}{p}}\|g(x_{j+1})\|^{\frac{p+1}{p}}}{2(p+1)!(L_p+\theta_1\sigma_j)^{\frac{p+1}{p}}}\geq\frac{(p!)^{\frac{p+1}{p}}\epsilon_1^{\frac{p+1}{p}}}{2(p+1)!\sigma_{\max}^{\frac{1}{p}}\left(\frac{L_p}{\vartheta\sigma_0}+\theta_1\right)^{\frac{p+1}{p}}}. \tag{3.19}$$

Summing this inequality from k_1 to $k\geq k_1$ and using AS.3, we obtain that

$$f(x_{k_1})-f_{\text{low}}\geq f(x_{k_1})-f(x_k)\geq\frac{(k-k_1)}{\kappa_{\text{OFFARp}}}\epsilon_1^{\frac{p+1}{p}}. \tag{3.20}$$

Rearranging the terms of the last inequality and using (3.18) and Lemma 3.8 then yields the desired result. □

While this theorem covers all model's degrees, it is worthwhile to isolate the most commonly used cases.

Corollary 1 Suppose that AS.1–AS.4 hold and that $p = 1$. Then the OFFAR1 algorithm requires at most

$$\left[4\sigma_{\max}\left(\frac{L_1+\vartheta\theta_1\sigma_0}{\vartheta\sigma_0}\right)^2\left(f(x_0)-f_{\low}+\frac{L_1 v_{\max}+\vartheta\sigma_0^2}{2\sigma_0}\right)+\left(\frac{2L_1(L_1+\theta_1\sigma_0)}{\vartheta\sigma_0}\right)^2\right]\epsilon_1^{-2}+2$$

iterations and evaluations of the gradient to produce a vector $x_\epsilon \in \mathbb{R}^n$ such that $\|g(x_\epsilon)\| \leq \epsilon_1$, where $\sigma_{\max}$ is defined in Lemma 3.9 and $v_{\max}$ is defined in Lemma 3.7. If $p = 2$, the OFFAR2 algorithm requires at most

$$\left[12\sigma_{\max}^{1/2}\left(\frac{L_2+\vartheta\theta_1\sigma_0}{2\vartheta\sigma_0}\right)^{\frac{3}{2}}\left(f(x_0)-f_{\low}+\frac{L_2 v_{\max}+\vartheta\sigma_0^2}{6\sigma_0}\right)+\left(\frac{2L_2(L_2+\theta_1\sigma_0)}{2\vartheta\sigma_0}\right)^{\frac{3}{2}}\right]\epsilon_1^{-\frac{3}{2}}+2$$

iterations and evaluations of the gradient and Hessian to achieve the same result.

We now prove that the complexity bound stated by Theorem 3.10 is sharp in order.

Theorem 3.11 Let $\epsilon_1 \in (0,1]$ and $p \geq 1$. Then there exists a p times continuously differentiable function f_p from $\mathbb{R}$ into $\mathbb{R}$ such that the OFFARp applied to f_p starting from the origin takes exactly $k_\epsilon = \lceil \epsilon_1^{-\frac{p+1}{p}} \rceil$ iterations and derivative's evaluations to produce an iterate x_{k_ϵ} such that $|\nabla_x^1 f_p(x_{k_\epsilon})| \leq \epsilon_1$.

Proof. To prove this result, we first define a sequence of function and derivatives' values such that the gradients converge sufficiently slowly and then show that these sequences can be generated by the OFFARp algorithm and also that there exists a function f_p satisfying AS.1–AS.4 which interpolate them.

First select $\vartheta = 1$ (implying that $\sigma_k = v_k$ for all k), some $\sigma_0 = v_0 > 0$ and define, for all $k \in \{0, \ldots, k_\epsilon\}$,

$$\omega_k = \epsilon_1 \frac{k_\epsilon - k}{k_\epsilon} \in [0, \epsilon_1] \tag{3.21}$$

and

$$g_k = -(\epsilon_1 + \omega_k) \quad \text{and} \quad D_{i,k} = 0, \quad (i = 2, \ldots, p), \tag{3.22}$$

so that

$$|g_k| \in [\epsilon_1, 2\epsilon_1] \subset [0, 2] \ \text{ for all } \ k \in \{0, \ldots, k_\epsilon\}. \tag{3.23}$$

We then set, for all $k \in \{0, \ldots, k_\epsilon\}$,

$$s_k = \left(\frac{p!|g_k|}{\sigma_k}\right)^{\frac{1}{p}}, \tag{3.24}$$

so that

$$\begin{aligned}\sigma_k &\stackrel{\rm def}{=} \sigma_0 + \sum_{j=0}^{k-1} \sigma_j |s_j|^{p+1} \qquad (3.25)\\ &= \sigma_0 + \sum_{j=0}^{k-1} \sigma_j \left(\frac{p!|g_j|}{\sigma_j}\right)^{\frac{p+1}{p}} = \sigma_0 + (p!)^{\frac{p+1}{p}} \sum_{j=0}^{k-1} \frac{\left(\epsilon_1 + \omega_j\right)^{\frac{p+1}{p}}}{\sigma_j^{\frac{1}{p}}}\\ &\le \sigma_0 + \left(\frac{(2p!)^{p+1}}{\sigma_0}\right)^{\frac{1}{p}} \sum_{j=0}^{k-1} \epsilon_1^{\frac{p+1}{p}} \le \sigma_0 + \left(\frac{(2p!)^{p+1}}{\sigma_0}\right)^{\frac{1}{p}} k_\epsilon \epsilon_1^{\frac{p+1}{p}} \le \sigma_0 + 2\left(\frac{(2p!)^{p+1}}{\sigma_0}\right)^{\frac{1}{p}}\\ &\stackrel{\rm def}{=} \sigma_{\max},\end{aligned}$$

where we successively used (3.24), (3.22), (3.21) and the definition of k_ϵ. We finally set

$$f_0 = 2^{\frac{2p+1}{p}} \left(\frac{p!}{\sigma_0}\right)^{\frac{1}{p}} \text{ and } f_{k+1} \stackrel{\rm def}{=} f_k + g_k s_k + \sum_{i=2}^{p} \frac{1}{i!} D_{i,k}[s_k]^i = f_k - \left(\frac{p!}{\sigma_k}\right)^{\frac{1}{p}} \left(\epsilon_1 + \omega_k\right)^{\frac{p+1}{p}},$$

yielding, using (3.25) and the definition of k_ϵ, that

$$f_0 - f_{k_\epsilon} = \sum_{k=0}^{k_\epsilon - 1} \left(\frac{p!}{\sigma_k}\right)^{\frac{1}{p}} \left(\epsilon_1 + \omega_k\right)^{\frac{p+1}{p}} \le 2^{\frac{p+1}{p}} \left(\frac{p!}{\sigma_0}\right)^{\frac{1}{p}} k_\epsilon \epsilon_1^{\frac{p+1}{p}} \le 2^{\frac{2p+1}{p}} \left(\frac{p!}{\sigma_0}\right)^{\frac{1}{p}}.$$

As a consequence

$$f_k \in [0, f_0] \text{ for all } k \in \{0, \ldots, k_\epsilon\} \qquad (3.26)$$

Observe that (3.24) satisfies (2.8) (for the model (2.4)) and (2.9) for $\theta_1 = 1$. Moreover (3.25) is the same as (2.11)-(2.12). Hence the sequence $\{x_k\}$ generated by

$$x_0 = 0 \text{ and } x_{k+1} = x_k + s_k$$

may be viewed as produced by the OFFARp algorithm given (3.22). Observe also that

$$|f_{k+1} - f_k| \le (p!)^{\frac{1}{p}} \sigma_{\max} \left(\frac{\epsilon_1 + \omega_k}{\sigma_k}\right)^{\frac{p+1}{p}} \le \frac{\sigma_{\max}}{p!} |s_k|^{p+1} \qquad (3.27)$$

and

$$|g_{k+1} - g_k| \le |\omega_k - \omega_{k+1}| = \frac{\epsilon_1}{k_\epsilon} \le \epsilon_1^{\frac{2p+1}{p}} \le \frac{\sigma_{\max}}{\sigma_k}(\epsilon_1 + \omega_k) = \frac{\sigma_{\max}}{p!}|s_k|^p \qquad (3.28)$$

(we used $k_\epsilon \le \epsilon_1^{-\frac{p+1}{p}} + 1$ and $\epsilon_1 \le 1$), while, if $p > 1$,

$$|D_{i,k+1} - D_{i,k}| = 0 \le \frac{\sigma_{\max}}{p!} |s_k|^{p+1-i} \qquad (3.29)$$

for $i = 2, \ldots, p$. In view of (3.23), (3.26) and (3.27)-(3.29), we may then apply classical Hermite interpolation to the data given by $\{(x_k, f_k, g_k, D_{2,k}, \ldots, D_{p,k})\}_{k=0}^{k_\epsilon}$ (see [11, Theorem A.9.2] with $\kappa_f = \max[2, f_0, \sigma_{\max}/p!]$, for instance) and deduce that there exists a

p times continuously differentiable piecewise polynomial function f_p satisfying AS.1–AS.4 and such that, for $k \in \{0, \ldots, k_\epsilon\}$,

$$f_k = f_p(x_k), \quad g_k = \nabla_x^1 f_p(x_k) \text{ and } D_{i,k} = \nabla_x^i f_p(x_k), \quad (i = 2, \ldots, p).$$

The sequence $\{x_k\}$ may thus be interpreted as being produced by the OFFARp algorithm applied to f_p starting from $x_0 = 0$. The desired conclusion then follows by observing that, from (3.21) and (3.22),

$$|g_k| > \epsilon_1 \text{ for } k \in \{0, \ldots, k_\epsilon - 1\} \text{ and } |g_{k_\epsilon}| = \epsilon_1.$$

□

4 Second-order optimality

If second-derivatives are available and $p \geq 2$, it is also possible to modify the OFFARp algorithm to obtain second-order optimality guarantees. We thus assume in this section that $p \geq 2$ and restate the algorithm as follows.

The modified algorithm only differs from that of page 15 by the addition of condition (4.5) on the step s_k. As was the case for (2.9)/(4.4), note that (4.5) holds with $\theta_2 = 1$ at a second-order minimizer of the model $m_k(s)$, and is thus achievable for $\theta_2 > 1$. Moreover, because the modified algorithm subsumes the original one, all properties derived in the previous section continue to hold. In addition, we may complete the bounds of Lemma 3.1 by noting that AS.3 for $p > 1$ also implies that

$$\|\nabla_x^2 f(x_{k+1}) - \nabla_s^2 T_{f,p}(x_k, s_k)\| \leq \frac{L_p}{(p-1)!}\|s_k\|^{p-1}. \tag{4.9}$$

We now derive a second-order analog of the step lower bound of Lemma 3.4.

Lemma 4.1 Suppose that AS.1 and AS.3 hold and that the modified algorithm is applied. Then

$$\|s_k\|^{p-1} > \frac{(p-1)!}{L_p + \theta_2 \sigma_k} \max\left[0, -\lambda_{\min}[\nabla_x^2 f(x_{k+1})]\right]. \tag{4.10}$$

Algorithm 4.1: Modified OFFO adaptive regularization of degree p

Step 0: Initialization: An initial point $x_0 \in \mathbb{R}^n$, a regularization parameter $v_0 = \sigma_0 > 0$, a requested final gradient accuracy $\epsilon_1 \in (0,1]$ and a requested final curvature accuracy $\epsilon_2 \in (0,1]$ are given, as well as the parameters

$$\theta_1, \theta_2 > 1 \quad \text{and} \quad \vartheta \in (0,1] \tag{4.1}$$

Set $k = 0$.

Step 1: Check for termination: Evaluate $g_k = \nabla_x^1 f(x_k)$ and $\nabla_x^2 f(x_k)$. Terminate with $x_\epsilon = x_k$ if

$$\|g_k\| \leq \epsilon_1 \quad \text{and} \quad \lambda_{\min}[\nabla_x^2 f(x_k)] \geq -\epsilon_2. \tag{4.2}$$

Else, evaluate $\{\nabla_x^i f(x_k)\}_{i=3}^p$.

Step 2: Step calculation: Compute a step s_k which sufficiently reduces the model m_k defined in (2.4) in the sense that

$$m_k(s_k) - m_k(0) < 0, \tag{4.3}$$

$$\|\nabla_s^1 T_{f,p}(x_k, s_k)\| \leq \theta_1 \frac{\sigma_k}{p!} \|s_k\|^p \tag{4.4}$$

and

$$\lambda_{\min}[\nabla_s^2 T_{f,p}(x_k, s_k)] \geq -\theta_2 \frac{\sigma_k}{(p-1)!} \|s_k\|^{p-1}. \tag{4.5}$$

Step 3: Updates. Set

$$x_{k+1} = x_k + s_k, \tag{4.6}$$

$$v_{k+1} = v_k + v_k \|s_k\|^{p+1} \tag{4.7}$$

and select

$$\sigma_{k+1} \in [\vartheta v_{k+1},\, v_{k+1}]\,. \tag{4.8}$$

Increment k by one and go to Step 1.

Proof. Successively using the triangle inequality, (4.9) and (4.5), we obtain that

$$\begin{aligned}
\lambda_{\min}[\nabla_x^2 f(x_{k+1})] &= \min_{\|d\|\leq 1} \nabla_x^2 f(x_{k+1})[d]^2 \\
&= \min_{\|d\|\leq 1} \left[\nabla_x^2 f(x_{k+1})[d]^2 - \nabla_s^2 T_{f,p}(x_k,s_k)[d]^2 + \nabla_s^2 T_{f,p}(x_k,s_k)[d]^2\right] \\
&\geq \min_{\|d\|\leq 1} \left[\nabla_x^2 f(x_{k+1})[d]^2 - \nabla_s^2 T_{f,p}(x_k,s_k)[d]^2\right] + \min_{\|d\|\leq 1} \nabla_s^2 T_{f,p}(x_k,s_k)[d]^2 \\
&= \min_{\|d\|\leq 1} \left[(\nabla_x^2 f(x_{k+1}) - \nabla_s^2 T_{f,p}(x_k,s_k))[d]^2\right] + \lambda_{\min}[\nabla_s^2 T_{f,p}(x_k,s_k)] \\
&\geq -\|\nabla_x^2 f(x_{k+1}) - \nabla_s^2 T_{f,p}(x_k,s_k)\| - \theta_2 \frac{\sigma_k}{(p-1)!}\|s_k\|^{p-1} \\
&= -\|\nabla_x^2 f(x_{k+1}) - \nabla_s^2 T_{f,p}(x_k,s_k)\| - \theta_2 \frac{\sigma_k}{(p-1)!}\|s_k\|^{p-1} \\
&= -\frac{L_p}{(p-1)!}\|s_k\|^{p-1} - \theta_2 \frac{\sigma_k}{(p-1)!}\|s_k\|^{p-1},
\end{aligned}$$

which proves (4.10). □

We now have to adapt our argument since the termination test (4.2) may fail if either its first or its second part fails. Lemma 3.4 then gives a lower bound on the step if the first part fails, while we have to use Lemma 4.1 if the second part fails. This is formalized in the following lemma.

Lemma 4.2 Suppose that AS.1 and AS.3 hold, and that the OFFARp algorithm has reached iteration of index

$$k \geq k_{**} \stackrel{\rm def}{=} \left\lceil \frac{2L_p}{\kappa_{\rm both}^{p+1}\vartheta} \max\left[\left(\frac{2L_p}{\vartheta}\right)^{\frac{1}{p}}, \left(\frac{2L_p}{\vartheta}\right)^{\frac{2}{p-1}}\right] \max\left[\epsilon_1^{-\frac{p+1}{p}}, \epsilon_2^{-\frac{p+1}{p-1}}\right]\right\rceil, \tag{4.11}$$

where

$$\kappa_{\rm both} \stackrel{\rm def}{=} \min\left[\left(\frac{p!}{\frac{L_p}{\vartheta\sigma_0}+\theta_1}\right)^{\frac{1}{p}}, \left(\frac{(p-1)!}{\frac{L_p}{\vartheta\sigma_0}+\theta_2}\right)^{\frac{1}{p-1}}\right]. \tag{4.12}$$

Then

$$v_k \geq \frac{2L_p}{\vartheta}, \tag{4.13}$$

which implies that

$$\sigma_k \geq 2L_p. \tag{4.14}$$

Proof. As in Lemma 3.5, (4.14) is a direct consequence of (4.8) if (4.13) is true. In order to adapt the proof of Lemma 3.5, we observe that, at iteration k, (3.5) and (4.10) hold and

$$\|s_k\| > \min\left[\left(\frac{p!}{L_p+\theta_1\sigma_k}\|g(x_{k+1})\|\right)^{\frac{1}{p}}, \left(\frac{(p-1)!}{L_p+\theta_2\sigma_k}\max\left[0, -\lambda_{\min}[\nabla_x^2 f(x_{k+1})]\right]\right)^{\frac{1}{p-1}}\right]$$

which, given termination has not yet occured and $v_k \geq \sigma_k \geq \vartheta\sigma_0$ implies that

$$\begin{aligned}\|s_k\| &> \min\left[\sigma_k^{-\frac{1}{p}}\left(\frac{p!}{\frac{L_p}{\vartheta\sigma_0}+\theta_1}\right)^{\frac{1}{p}}, \sigma_k^{-\frac{1}{p-1}}\left(\frac{(p-1)!}{\frac{L_p}{\vartheta\sigma_0}+\theta_2}\right)^{\frac{1}{p-1}}\right]\min\left[\epsilon_1^{\frac{1}{p}}, \epsilon_2^{\frac{1}{p-1}}\right] \\ &\geq \kappa_{\rm both}\min\left[v_k^{-\frac{1}{p}}, v_k^{-\frac{1}{p-1}}\right]\min\left[\epsilon_1^{\frac{1}{p}}, \epsilon_2^{\frac{1}{p-1}}\right]. \qquad (4.15)\end{aligned}$$

Suppose now that (4.13) fails, i.e. that for some $k \geq k_{**}$, $v_k < \frac{2L_p}{\vartheta}$. Since v_k is a non-decreasing sequence, we have that $v_j < \frac{2L_p}{\vartheta}$ for $j \in \{0,\ldots,k\}$. Successively using (4.7) and (4.15), we obtain that

$$\begin{aligned}v_k &> \sum_{j=0}^{k-1} v_j\|s_j\|^{p+1} \geq \sum_{j=0}^{k-1}\kappa_{\rm both}^{p+1}\min\left[v_j^{-\frac{1}{p}}, v_j^{-\frac{2}{p-1}}\right]\min\left[\epsilon_1^{\frac{1}{p}}, \epsilon_2^{\frac{1}{p-1}}\right]^{p+1} \\ &\geq \sum_{j=0}^{k-1}\kappa_{\rm both}^{p+1}\min\left[\left(\frac{2L_p}{\vartheta}\right)^{-\frac{1}{p}}, \left(\frac{2L_p}{\vartheta}\right)^{-\frac{2}{p-1}}\right]\min\left[\epsilon_1^{\frac{1}{p}}, \epsilon_2^{\frac{1}{p-1}}\right]^{p+1} \\ &= k_{**}\kappa_{\rm both}^{p+1}\min\left[\left(\frac{2L_p}{\vartheta}\right)^{-\frac{1}{p}}, \left(\frac{2L_p}{\vartheta}\right)^{-\frac{2}{p-1}}\right]\min\left[\epsilon_1^{\frac{1}{p}}, \epsilon_2^{\frac{1}{p-1}}\right]^{p+1}.\end{aligned}$$

Using the definition of k_{**} in the last inequality, we see that

$$\frac{2L_p}{\vartheta} < v_{k_{**}} < \frac{2L_p}{\vartheta},$$

which is impossible. Hence no index $k \geq k_{**}$ exists such that $v_k < \frac{2L_p}{\vartheta}$ and (4.13) and (4.14) hold. □

We then continue to use the theory of the previous section with a value of k_1 now satisfying the improved bound

$$k_1 \leq k_{**}, \qquad (4.16)$$

instead of $k_1 \leq k_*$. This directly leads us to the following strengthened complexity result.

Theorem 4.3 Suppose that AS.1–AS.4 hold and that $p > 1$ Then the modified OFFARp algorithm requires at most

$$\left[\kappa_{\rm MOFFARp}\left(f(x_0) - f_{\rm low} + \frac{L_p v_{\max} + \vartheta\sigma_0^2}{(p+1)!\sigma_0}\right)\right] + \frac{2L_p}{\kappa_{\rm both}^{p+1}\vartheta}\max\left[\left(\frac{2L_p}{\vartheta}\right)^{\frac{1}{p}}, \left(\frac{2L_p}{\vartheta}\right)^{\frac{2}{p-1}}\right] \times \max\left[\epsilon_1^{-\frac{p+1}{p}}, \epsilon_2^{-\frac{p+1}{p-1}}\right] + 2$$

iterations and evaluations of $\{\nabla_x^i f\}_{i=1}^p$ to produce a vector $x_\epsilon \in \mathbb{R}^n$ such that $\|g(x_\epsilon)\| \leq \epsilon_1$ and $\lambda_{\min}[\nabla_x^2 f(x_\epsilon)] \geq -\epsilon_2$, where

$$\kappa_{\rm MOFFARp} \stackrel{\rm def}{=} 2(p+1)!\max\left[\sigma_{\max}^{1/p}\left(\frac{L_p + \vartheta\theta_1\sigma_0}{\vartheta p!\sigma_0}\right)^{\frac{p+1}{p}}, \sigma_{\max}^{2/p-1}\left(\frac{L_p + \vartheta\theta_2\sigma_0}{\vartheta(p-1)!\sigma_0}\right)^{\frac{p+1}{p-1}}\right]$$

and where $\sigma_{\max}$ is defined in Lemma 3.9, $v_{\max}$ is defined in Lemma 3.7 and $\kappa_{\rm both}$ in (4.12).

Proof. The bound of Theorem 3.10 remains valid for obtaining a vector $x_\epsilon \in \mathbb{R}^n$ such that $\|g(x_\epsilon)\| \leq \epsilon_1$, but we are now interested to satisfy the second part of (4.2) as well. Using (4.10) instead of (3.5), we deduce (in parallel to (3.19)) that before termination,

$$\begin{aligned} f(x_j) - f(x_{j+1}) &\geq \frac{\sigma_j\|s_j\|^{p+1}}{2(p+1)!} \\ &\geq \frac{\sigma_j((p-1)!)^{\frac{p+1}{p-1}}\max[0, -\lambda_{\min}[\nabla_x^2 f(x_{k+1})]]^{\frac{p+1}{p-1}}}{2(p+1)!(L_p + \theta_2\sigma_j)^{\frac{p+1}{p-1}}} \\ &\geq \frac{((p-1)!)^{\frac{p+1}{p-1}}\epsilon_2^{\frac{p+1}{p-1}}}{2(p+1)!\sigma_{\max}^{\frac{2}{p-1}}\left(\frac{L_p}{\vartheta\sigma_0} + \theta_2\right)^{\frac{p+1}{p}}}, \end{aligned}$$

so that, summing this inequality from k_1 to $k \geq k_1$ and using AS.3 now gives (in parallel to (3.20)) that, before the second part of (4.2) is satisfied,

$$f(x_{k_1}) - f_{\rm low} \geq f(x_{k_1}) - f(x_k) \geq \frac{(k-k_1)}{\kappa_{\rm 2nd}}\epsilon_2^{\frac{p+1}{p-1}}$$

where

$$\kappa_{\rm 2nd} \stackrel{\rm def}{=} 2(p+1)!\,\sigma_{\max}^{2/p-1}\left(\frac{L_p + \vartheta\theta_2\sigma_0}{\vartheta(p-1)!\sigma_0}\right)^{\frac{p+1}{p-1}}.$$

As a consequence, we deduce, using (4.16), that the second part of (4.2) must hold at the latest after

$$\left[\kappa_{\rm 2nd}\left(f(x_0) - f_{\rm low} + \frac{L_p v_{\max} + \vartheta\sigma_0^2}{(p+1)!\sigma_0}\right)\right]\epsilon_2^{-\frac{p+1}{p-1}} + k_{**} + 2$$

iterations and evaluations of the derivatives, where k_{**} is defined in (4.11). Combining this result with that of Theorem 3.10 then yields the desired conclusion. □

Focusing again on the case where $p = 2$ and upperbounding complicated constants, we may state the following corollary.

Corollary 2 Suppose that AS.1–AS.4 hold and that $p = 2$. Then there exists constants κ_* such that the modified OFFAR1 algorithm requires at most

$$\kappa_* \max\left[\epsilon_1^{-3/2}, \epsilon_2^{-3}\right]$$

iterations and evaluations of the gradient and Hessian to produce a vector $x_\epsilon \in \mathbb{R}^n$ such that $\|g(x_\epsilon)\| \leq \epsilon_1$ and $\lambda_{\min}[\nabla_x^2 f(x_{k_\epsilon})] \geq -\epsilon_2$.

We finally prove that the complexity for reaching approximate second order points, as stated by Theorem 4.3, is also sharp.

Theorem 4.4 Let $\epsilon_1, \epsilon_2 \in (0, 1]$ and $p > 1$. Then there exists a p times continuously differentiable function f_p from $\mathbb{R}$ into $\mathbb{R}$ such that the modified OFFARp applied to f_p starting from the origin takes exactly $k_\epsilon = \lceil \epsilon_2^{-\frac{p+1}{p-1}} \rceil$ iterations and derivative's evaluations to produce an iterate x_{k_ϵ} such that $|\nabla_x^1 f_p(x_{k_\epsilon})| \leq \epsilon_1$ and $\lambda_{\min}[\nabla_x^2 f(x_{k_\epsilon})] \geq -\epsilon_2$.

Proof. The proof is very similar to that of Theorem 3.11, this time taking a uniformly zero gradient but a minimal eigenvalue of the Hessian slowly converging to $-\epsilon_2$ from below. It is detailed in appendix. □

5 Discussion

It is remarkable that the complexity bound stated by Theorems 3.10 and 4.3 are identical (in order) to that known for the standard setting where the objective function is evaluated at each iteration. Moreover, the $\mathcal{O}(\epsilon^{-3/2})$ bound for $p = 2$ was shown in [9] to be optimal within a very large class of second-order methods. One then concludes that, from the sole viewpoint of evaluation complexity, the computation of the objective function's values is an unnecessary effort for achieving convergence at optimal speed.

The above results may be extended in different ways, which we have not included in our development to avoid too much generality and reduce the notational burden. The first is to allow errors in derivatives of orders 2 to p. If we denote by $\widehat{\nabla_x^i f}$ the approximation of $\nabla_x^i f$, it is easily seen in the proof of Lemma 3.4 that the argument remains valid as long as, for some $\kappa_D \geq 0$,

$$\|\widehat{\nabla_x^i f}(x_k) - \nabla_x^i f(x_k)\| \leq \kappa_D \|s_k\|^{p+1-i}. \tag{5.1}$$

Since the accuracy of derivatives of degree larger than one only occurs in this lemma, we conclude that our results still hold if (5.1) holds.

The second extension is to replace the gradient Lipschitz continuity in AS.3 by a weaker Hölder continuity, namely that there exist non-negative constant L_p and $\beta \in (0,1]$ such that

$$\|\nabla_x^p f(x) - \nabla_x^p f(y)\| \leq L_p \|x-y\|^\beta \text{ for all } x, y \in \mathbb{R}^n. \tag{5.2}$$

It then possible to verify that all our result remain valid with $p+1$ replaced by $p+\beta$.

A third possibility is to consider optimization in infinite-dimensional smooth Banach spaces, a development presented for the standard framework in [19]. This requires specific techniques for computing the step and a careful handling of the norms involved.

We may also consider non-smooth norms, as in [22], or imposing convex constraints on the variables [11, Chapter 6].

Finally, an extension to guarantee third-order optimality conditions (in the case where third derivatives are available) may be possible along the lines discussed in [11, Chapter 4].

6 Conclusions

We have presented an adaptive regularization algorithm for nonconvex unconstrained minimization where the objective function is never calculated and which has, for a given degree of used derivatives, the best-known worst-case complexity order, not only among OFFO methods, but also among all known optimization algorithms. In particular, the algorithm using gradients and Hessians requires at most $\mathcal{O}(\epsilon_1^{-3/2})$ iterations to produce an iterate such that $\|\nabla_x^1 f(x_k)\| \leq \epsilon_1$, and at most $\mathcal{O}(\epsilon_2^{-3})$ iterations to additionally ensure that $\lambda_{\min}[\nabla_x^2 f(x_k)] \geq -\epsilon_2$. Moreover, all stated complexity bounds are sharp.

Given the prowess of OFFO methods on noisy problems, the transition from the present deterministic theory to the noisy context is clearly of interest and is the object of ongoing research.

References

[1] S. Bellavia, G. Gurioli, B. Morini, and Ph. L. Toint. A stochastic ARC method with inexact function and random derivatives evaluations. In *Proceedings of the International Conference on Machine Learning (ICML2020)*, 2020.

[2] S. Bellavia, G. Gurioli, B. Morini, and Ph. L. Toint. Quadratic and cubic regularization methods with inexact function and random derivatives for finite-sum minimization. In *Proceedings of the ICCSA 2021*, 2021.

[3] S. Bellavia, G. Gurioli, B. Morini, and Ph. L. Toint. Adaptive regularization algorithm for nonconvex optimization using inexact function evaluations and randomly perturbed derivatives. *Journal of Complexity*, 68, 2022.

[4] A. Berahas, L. Cao, and K. Scheinberg. Global convergence rate analysis of a generic line search algorithm with noise. *SIAM Journal on Optimization*, 31:1489–1518, 2021.

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

[6] J. Blanchet, C. Cartis, M. Menickelly, and K. Scheinberg. Convergence rate analysis of a stochastic trust region method via supermartingales. *INFORMS Journal on Optimization*, 1(2):92–119, 2019.

[7] 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.

[8] 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.

[9] C. Cartis, N. I. M. Gould, and Ph. L. Toint. Worst-case evaluation complexity and optimality of second-order methods for nonconvex smooth optimization. In B. Sirakov, P. de Souza, and M. Viana, editors, *Invited Lectures, Proceedings of the 2018 International Conference of Mathematicians (ICM 2018), vol. 4, Rio de Janeiro*, pages 3729–3768. World Scientific Publishing Co Pte Ltd, 2018.

[10] C. Cartis, N. I. M. Gould, and Ph. L. Toint. Sharp worst-case evaluation complexity bounds for arbitrary-order nonconvex optimization with inexpensive constraints. *SIAM Journal on Optimization*, 30(1):513–541, 2020.

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

[12] C. Cartis and K. Scheinberg. Global convergence rate analysis of unconstrained optimization methods based on probabilistic models. *Mathematical Programming, Series A*, 159(2):337–375, 2018.

[13] R. Chen, M. Menickelly, and K. Scheinberg. Stochastic optimization using a trust-region method and random models. *Mathematical Programming, Series A*, 169(2):447–487, 2018.

[14] F. E. Curtis, D. P. Robinson, and M. Samadi. A trust region algorithm with a worst-case iteration complexity of $\mathrm{O}(\epsilon^{-3/2})$ for nonconvex optimization. *Mathematical Programming, Series A*, 162(1):1–32, 2017.

[15] F. E. Curtis, K. Scheinberg, and R. Shi. A stochastic trust region algorithm based on careful step normalization. *INFORMS Journal on Optimization*, 1(3):200–220, 2019.

[16] A. Défossez, L. Bottou, F. Bach, and N. Usunier. A simple convergence proof for Adam and Adagrad. arXiv:2003.02395v2, 2020.

[17] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12, July 2011.

[18] G. N. Grapiglia and G. F. D. Stella. An adaptive trust-region method without function evaluation. Optimization Online, February 2022.

[19] S. Gratton, S. Jerad, and Ph. L. Toint. Hölder gradient descent and adaptive regularization methods in banach spaces for first-order points. arXiv:2104.02564, 2021.

[20] S. Gratton, S. Jerad, and Ph. L. Toint. First-order objective-function-free optimization algorithms and their complexity. arXiv:2203.01757, 2022.

[21] S. Gratton, S. Jerad, and Ph. L. Toint. Parametric complexity analysis for a class of first-order Adagrad-like algorithms. arXiv:2203.01647, 2022.

[22] S. Gratton and Ph. L. Toint. Adaptive regularization minimization algorithms with non-smooth norms. *IMA Journal of Numerical Analysis*, (to appear), 2022.

[23] S. Gratton and Ph. L. Toint. OFFO minimization algorithms for second-order optimality and their complexity. arXiv:2203.03351, 2022.

[24] D. Kingma and J. Ba. Adam: A method for stochastic optimization. In *Proceedings in the International Conference on Learning Representations (ICLR)*, 2015.

[25] J. L. Lagrange. *Traité de la résolution des équations numériques de tous les degrés: avec des notes sur plusieurs points de la théorie des équations algébriques*. Courcier, Paris, 1806.

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

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

[28] C. W. Royer and S. J. Wright. Complexity analysis of second-order line-search algorithms for smooth nonconvex optimization. *SIAM Journal on Optimization*, 28(2):1448–1477, 2018.

[29] R. Ward, X. Wu, and L. Bottou. Adagrad stepsizes: sharp convergence over nonconvex landscapes. In *Proceedings in the International Conference on Machine Learning (ICML2019)*, 2019.

[30] X. Wu, R. Ward, and L. Bottou. WNGRAD: Learn the learning rate in gradient descent. arXiv:1803.02865, 2018.

[31] C. K. Yap. *Fundamental Problems of Algorithmic Algebra*. Oxford University Press, Oxford, United Kingdom, 1999.

Appendix

We give the detailed proof of Theorem 4.4.

Theorem A.1 Let $\epsilon_2 \in (0,1]$ and $p > 1$. Then there exists a p times continuously differentiable function f_p from $\mathbb{R}$ into $\mathbb{R}$ such that the modified OFFARp applied to f_p starting from the origin takes exactly $k_\epsilon = \lceil \epsilon_2^{-\frac{p+1}{p-1}} \rceil$ iterations and derivative's evaluations to produce an iterate x_{k_ϵ} such that $|\nabla_x^1 f_p(x_{k_\epsilon})| \leq \epsilon_1$ and $\lambda_{\min}[\nabla_x^2 f(x_{k_\epsilon})] \geq -\epsilon_2$.

Proof. The proof of this result closely follows that of Theorem 3.11. First select $\vartheta = 1$ (implying that $\sigma_k = v_k$ for all k), some $\sigma_0 = v_0 > 0$ and define, for all $k \in \{0, \ldots, k_\epsilon\}$,

$$\omega_k = \epsilon_2 \frac{k_\epsilon - k}{k_\epsilon} \in [0, \epsilon_2] \tag{A.1}$$

and

$$g_k = 0, \quad H_k = -(\epsilon_2 + \omega_k) \quad \text{and} \quad D_{i,k} = 0, \quad (i = 3, \ldots, p), \tag{A.2}$$

so that

$$|H_k| \in [\epsilon_2, 2\epsilon_2] \subset [0, 2] \ \text{ for all } \ k \in \{0, \ldots, k_\epsilon\}. \tag{A.3}$$

We then set, for all $k \in \{0, \ldots, k_\epsilon\}$,

$$s_k = \left(\frac{p!|H_k|}{\sigma_k} \right)^{\frac{1}{p-1}}, \tag{A.4}$$

so that

$$\begin{aligned}
\sigma_k &\stackrel{\rm def}{=} \sigma_0 + \sum_{j=0}^{k-1} \sigma_j |s_j|^{p+1} \qquad (A.5)\\
&= \sigma_0 + \sum_{j=0}^{k-1} \sigma_j \left(\frac{p!|H_j|}{\sigma_j} \right)^{\frac{p+1}{p-1}} = \sigma_0 + (p!)^{\frac{p+1}{p-1}} \sum_{j=0}^{k-1} \frac{(\epsilon_2 + \omega_j)^{\frac{p+1}{p-1}}}{\sigma_j^{\frac{2}{p-1}}} \\
&\leq \sigma_0 + \left(\frac{(2p!)^{p+1}}{\sigma_0^2} \right)^{\frac{1}{p-1}} \sum_{j=0}^{k-1} \epsilon_2^{\frac{p+1}{p-1}} \leq \sigma_0 + \left(\frac{(2p!)^{p+1}}{\sigma_0^2} \right)^{\frac{1}{p-1}} k_\epsilon \epsilon_2^{\frac{p+1}{p-1}} \leq \sigma_0 + 2 \left(\frac{(2p!)^{p+1}}{\sigma_0^2} \right)^{\frac{1}{p-1}} \\
&\stackrel{\rm def}{=} \sigma_{\max},
\end{aligned}$$

where we successively used (A.4), (A.2), (A.1) and the definition of k_ϵ. We finally set

$$f_0 = 2^{\frac{p+1}{p-1}} \left(\frac{p!}{\sigma_0} \right)^{\frac{2}{p-1}} \ \text{ and } \ f_{k+1} \stackrel{\rm def}{=} f_k + \tfrac{1}{2} H_k s_k^2 + \sum_{i=2}^{p} \frac{1}{i!} D_{i,k} [s_k]^i = f_k - \tfrac{1}{2} \left(\frac{p!}{\sigma_k} \right)^{\frac{2}{p-1}} (\epsilon_2 + \omega_k)^{\frac{p+1}{p-1}},$$

yielding, using (3.25) and the definition of k_ϵ, that

$$f_0 - f_{k_\epsilon} = \frac{1}{2} \sum_{k=0}^{k_\epsilon - 1} \left(\frac{p!}{\sigma_k} \right)^{\frac{2}{p-1}} (\epsilon_2 + \omega_k)^{\frac{p+1}{p-1}} \leq 2^{\frac{2}{p-1}} \left(\frac{p!}{\sigma_0} \right)^{\frac{2}{p-1}} k_\epsilon \epsilon_2^{\frac{p+1}{p}} \leq 2^{\frac{p+1}{p-1}} \left(\frac{p!}{\sigma_0} \right)^{\frac{2}{p-1}}.$$

As a consequence

$$f_k \in [0, f_0] \text{ for all } k \in \{0, \ldots, k_\epsilon\}. \tag{A.6}$$

Observe that (A.4) satisfies (4.3) (for the model (2.4)), (4.4) for $\theta_1 = 1$ and (4.5) for $\theta_2 = 1$. Moreover (A.5) is the same as (4.7)-(4.8). Hence the sequence $\{x_k\}$ generated by

$$x_0 = 0 \text{ and } x_{k+1} = x_k + s_k$$

may be viewed as produced by the modified OFFARp algorithm given (A.2). Observe also that

$$|f_{k+1} - f_k| \leq (p!)^{\frac{2}{p-1}} \sigma_{\max} \left(\frac{\epsilon_2 + \omega_k}{\sigma_k} \right)^{\frac{p+1}{p-1}} \leq \frac{\sigma_{\max}}{p!} |s_k|^{p+1}, \tag{A.7}$$

$$|g_{k+1} - g_k| = 0 \leq \frac{\sigma_{\max}}{p!} |s_k|^p, \tag{A.8}$$

and

$$|H_{k+1} - H_k| \leq |\omega_k - \omega_{k+1}| = \frac{\epsilon_2}{k_\epsilon} \leq \epsilon_2^{\frac{2p}{p-1}} \leq \frac{\sigma_{\max}}{\sigma_k}(\epsilon_2 + \omega_k) = \frac{\sigma_{\max}}{p!}|s_k|^{p-1} \tag{A.9}$$

(we used $k_\epsilon \leq \epsilon_2^{-\frac{p+1}{p-1}} + 1$ and $\epsilon_2 \leq 1$), while, if $p > 2$,

$$|D_{i,k+1} - D_{i,k}| = 0 \leq \frac{\sigma_{\max}}{p!} |s_k|^{p+1-i} \tag{A.10}$$

for $i = 3, \ldots, p$. In view of (A.3), (A.6) and (A.7)-(A.10), we may then apply classical Hermite interpolation to the data given by $\{(x_k, f_k, g_k, H_k, D_{3,k}, \ldots, D_{p,k})\}_{k=0}^{k_\epsilon}$ (see [11, Theorem A.9.2] with $\kappa_f = \max[2, f_0, \sigma_{\max}/p!]$, for instance) and deduce that there exists a p times continuously differentiable piecewise polynomial function f_p satisfying AS.1–AS.4 and such that, for $k \in \{0, \ldots, k_\epsilon\}$,

$$f_k = f_p(x_k), \quad g_k = \nabla_x^1 f_p(x_k), \quad H_k = \nabla_x^2 f_p(x_k) \text{ and } D_{i,k} = \nabla_x^i f_p(x_k), \quad (i = 3, \ldots, p).$$

The sequence $\{x_k\}$ may thus be interpreted as being produced by the OFFARp algorithm applied to f_p starting from $x_0 = 0$. The desired conclusion then follows by observing that, from (A.1) and (A.2), $g_k = 0 < \epsilon_1$ for all k while

$$\lambda_{\min}[H_k] = H_k < -\epsilon_2 \text{ for } k \in \{0, \ldots, k_\epsilon - 1\} \quad \text{and} \quad \lambda_{\min}[H_{k_\epsilon}] = H_{k_\epsilon} = -\epsilon_2.$$

□