THE EVALUATION COMPLEXITY OF FINDING HIGH-ORDER MINIMIZERS OF NONCONVEX OPTIMIZATION

CORALIA CARTIS, NICHOLAS I. M. GOULD, AND PHILIPPE L. TOINT

ABSTRACT

We introduce the concept of strong high-order approximate minimizers of nonconvex optimization problems. These apply in both standard smooth and composite nonsmooth settings, and additionally allow convex or inexpensive constraints. An adaptive regularization algorithm is then proposed to find such approximate minimizers. Under suitable Lipschitz continuity assumptions, the evaluation complexity of this algorithm is investigated. The bounds obtained not only provide, to the best of our knowledge, the first known result for (unconstrained or inexpensively-constrained) composite problems for optimality orders exceeding one, but also give the first sharp bounds for high-order strong approximate qth order minimizers of standard (unconstrained and inexpensively constrained) smooth problems, thereby complementing known results for weak minimizers.

MATHEMATICS SUBJECT CLASSIFICATION 2020

Primary 90C60; Secondary 90C26, 90C30, 49M15, 68Q25

KEYWORDS

Nonconvex optimization, composite optimization, regularization methods, complexity bounds, global rates of convergence

INTERNATIONAL CONGRESS
OF MATHEMATICIANS
2022 JULY 6–14

Proc. Int. Cong. Math. 2022, Vol. 7, pp. 5256–5289
DOI 10.4171/ICM2022/95

1. INTRODUCTION

We consider composite optimization problems of the form

$$\min_{x\in\mathcal{F}} w(x) \stackrel{\text{def}}{=} f(x) + h\big(c(x)\big), \tag{1.1}$$

where f, c are smooth and h possibly nonsmooth but Lipschitz continuous, and where $\mathcal{F}$ is a feasible set associated with inexpensive constraints (which are discussed in the next paragraph). Such problems have attracted considerable attention, due to the their occurrence in important applications such as LASSO methods in computational statistics [26], Tikhonov regularization of underdetermined estimation problems [21], compressed sensing [16], artificial intelligence [22], penalty or projection methods for constrained optimization [8], least Euclidean distance and continuous location problems [17], reduced-precision deep-learning [27], image processing [2], to cite but a few examples. We refer the reader to the thorough review in [23]. In these applications, the function h is typically globally Lipschitz continuous and cheap to compute—common examples include the Euclidean, ℓ_1, or ℓ_∞ norms.

Inexpensive constraints defining the feasible set $\mathcal{F}$ are constraints whose evaluation or enforcement has negligible cost compared to that of evaluating f, c and/or their derivatives. They are of interest here since the evaluation complexity of solving inexpensively constrained problems is dominated solely by the number of evaluations of f, c and their derivatives. Inexpensive constraints include, but are not limited to, convex constraints with cheap projections (such as bounds or the ordered simplex). Such constraints have already been considered elsewhere [3,12].

Of course, problem (1.1) may be viewed as a general nonsmooth optimization problem, to which a battery of existing methods may be applied (for example, subgradient, proximal gradient, and bundle methods). However, this avenue ignores the problem's special structure, which may be viewed as a drawback. More importantly for our purpose, this approach essentially limits the type of approximate minimizers one can reasonably hope for to first-order points (see [18, CHAPTER 14] for a discussion of second-order optimality conditions and [8,20] for examples of structure-exploiting first-order complexity analysis). However, our first objective in this paper is to cover *approximate minimizers of arbitrary order* (obviously including first- and second-order ones), in a sense that we describe below. This, as far we know, precludes a view of (1.1) that ignores the structure present in h.

It is also clear that any result we can obtain for problem (1.1) also applies to standard smooth problems (by letting h be the zero function), for which evaluation complexity results are available. Most of these results cover first- and second-order approximate minimizers (see [7,10,15,24,25] for a few references), but two recent papers [11,12] propose an analysis covering our stated objective to cover arbitrary-order minimizers for smooth nonconvex functions. However, these two proposals significantly differ, in that they use different definitions of high-order minimizers, by no means a trivial concept. The first paper, focusing on trust-region methods, uses a much stronger definition than the second one which covers adaptive regularization algorithms. Our second objective in the present paper is to strengthen these latter results to *use the stronger definition of optimality for adaptive regu-*

larization algorithms and therefore bridge the gap between the two previous approaches in the more general framework of composite problems.

Contributions and motivation. The main contributions of this paper may be summarized as follows:

(1) We formalize the notion of strong approximate minimizer of arbitrary order for standard (noncomposite) smooth problems and extend it to composite ones, including the case where the composition function is nonsmooth, and additionally allow inexpensive constraints. This notion is stronger than that of "weak" approximate minimizers used in [3,12,14].

(2) We provide a conceptual adaptive regularization algorithm whose purpose is to compute such strong approximate minimizers.

(3) We analyze the worst-case complexity of this conceptual algorithm both for composite and standard problems, allowing arbitrary optimality order and any degree of the model used within the algorithm. For composite problems, these bounds are the first ones available for approximate minimizers of order exceeding one. For smooth problems, the bounds are shown to improve on those derived in [11] for trust-region methods, while being less favorable (for orders beyond the second) than those in [12] for approximate minimizers of the weaker sort. These bounds are summarized in Table 1.1 in the case where all ε_j are identical. Each table entry also mentions existing references for the quoted result. Sharpness (in the order of ε) is also reported when known.

We acknowledge upfront that our approach is essentially theoretical, because it depends, in its present incarnation, on computing global minimizers of Taylor series within Euclidean balls, a problem which is known to be a very hard for high orders [1]. Although these calculations do not involve any evaluation of the problem's objective function or of its derivatives (and thus do not affect evaluation complexity bounds), this is a significant hurdle. While realistic algorithms may have to resort to inexact global minimization (we discuss the necessity and impact of such approximations in Section 7), the case of exact ones can be viewed as an idealized, aspirational setting and the complexity results derived therein as "best possible." Thus we ask here the mathematically important questions: what would be achievable in this idealized setting? Or if constrained global minimizers of polynomials were computable because of special problem structure? A second motivation is that high-order models have already proved their usefulness in practice, in particular in the solution of highly nonlinear low-dimensional least-squares problems [19], even if implementing algorithms using them is far from obvious [4]. The identification of approximate minimizers of orders matching the degree of the models is, in our view, an obvious, yet unexplored question. Moreover, the consideration of such approximate minimizers results in new insights in the definition of approximate minimizers and prompts a proposal for a new approximate optimality measure (see Section 2). At variance with standard ones, this proposal has the

	inexpensive constraints	Weak minimizers	Strong minimizers		
		smooth ($h = 0$)	smooth ($h = 0$)	composite	
				h convex	h nonconvex
$q = 1$	none	$\varepsilon^{-\frac{p+1}{p}}$ sharp [5,12]	$\varepsilon^{-\frac{p+1}{p}}$ sharp [5,12]	$\varepsilon^{-\frac{p+1}{p}}$ sharp †	ε^{-2} [8,20]
	convex	$\varepsilon^{-\frac{p+1}{p}}$ sharp [5,12]	$\varepsilon^{-\frac{p+1}{p}}$ sharp [5,12]	$\varepsilon^{-\frac{p+1}{p}}$ sharp	ε^{-2}
	nonconvex	$\varepsilon^{-\frac{p+1}{p}}$ sharp [5,12]	$\varepsilon^{-\frac{p+1}{p}}$ sharp [5,12]	ε^{-2}	ε^{-2}
$q = 2$	none	$\varepsilon^{-\frac{p+1}{p-1}}$ sharp [12]	$\varepsilon^{-\frac{p+1}{p-1}}$ sharp [12]	ε^{-3}	ε^{-3}
	convex	$\varepsilon^{-\frac{p+1}{p-1}}$ sharp [12]	$\varepsilon^{-\frac{2(p+1)}{p}}$ sharp	ε^{-3}	ε^{-3}
	nonconvex	$\varepsilon^{-\frac{p+1}{p-1}}$ sharp [12]	$\varepsilon^{-\frac{2(p+1)}{p}}$ sharp	ε^{-3}	ε^{-3}
$q > 2$	none, or general	$\varepsilon^{-\frac{p+1}{p-q+1}}$ sharp [12]	$\varepsilon^{-\frac{q(p+1)}{p}}$ sharp	$\varepsilon^{-(q+1)}$	$\varepsilon^{-(q+1)}$

TABLE 1.1
Order bounds (as multiples of powers of the accuracy ε) on the worst-case evaluation complexity of finding weak/strong (ε, δ)-approximate minimizers for composite and smooth problems, as a function of optimality order (q), model degree (p), convexity of the composition function h and presence/absence/convexity of inexpensive constraints. The dagger indicates that this bound for the special case when $h(\cdot) = \|\cdot\|_2$ and $f = 0$ is already known [9].

advantage of being well-defined and consistent across all orders and it is obviously also applicable (and computationally cheap) for orders one and two.

Outline. The paper is organized as follows. Section 2 outlines some useful background and motivation on high-order optimality measures. In Section 3, we describe our problem more formally and introduce the notions of weak and strong high-order approximate minimizers. We describe an adaptive regularization algorithm for problem (1.1) in Section 4, while Section 5 discusses the associated evaluation complexity analysis. Section 6 then shows that several of the obtained complexity bounds are sharp, while Section 7 discusses the necessity of global minimizations and the impact of allowing them to be inexact. Some conclusions and perspectives are finally outlined in Section 8.

2. A DISCUSSION OF qTH-ORDER NECESSARY OPTIMALITY CONDITIONS

Before going any further, it is best to put our second objective (establishing strong complexity bounds for arbitrary qth order using an adaptive regularization method) in perspective by briefly discussing high-order optimality measures. For this purpose, we now digress slightly and first focus on the standard unconstrained (smooth) optimization problem where one tries to minimize an objective function f over $\mathbb{R}^n$. The definition of a jth-order approximate minimizer of a general (sufficiently) smooth function f is a delicate question. It was argued in [11] that expressing the necessary optimality conditions at a given point x in terms of individual derivatives of f at x leads to extremely complicated expressions involving the potential decrease of the function along all possible feasible arcs emanating from x. To avoid this, an alternative based on Taylor expansions was proposed. Such an expansion is given by

$$T_{f,q}(x,d) = \sum_{\ell=0}^{q} \frac{1}{\ell!} \nabla_x^\ell f(x)[d]^\ell \tag{2.1}$$

where $\nabla_x^\ell f(x)[d]^\ell$ denotes the ℓth-order cubically[1] symmetric derivative tensor (of dimension ℓ) of f at x applied to ℓ copies of the vector d. The idea of the *approximate* necessary condition that we use is that, if x is a local minimizer and $q \le p$ is an integer, there should be a neighborhood of x of radius $\delta_j \in (0,1]$ in which the decrease in (2.1), which we measure by

$$\phi_{f,j}^{\delta_j}(x) \stackrel{\text{def}}{=} f(x) - \min_{d\in\mathbb{R}^n, \|d\|\le\delta_j} T_{f,j}(x,d), \tag{2.2}$$

must be small. In fact, it can be shown [11, LEMMA 3.4] that

$$\lim_{\delta_j\to 0} \frac{\phi_{f,j}^{\delta_j}(x)}{\delta_j^j} = 0, \tag{2.3}$$

whenever x is a local minimizer of f. Making the ratio in this limit small for small enough δ_j therefore seems reasonable. Let ε_j is a prescribed order-dependent accuracy parameter, and $\varepsilon \stackrel{\text{def}}{=} (\varepsilon_1, \dots, \varepsilon_q)$. Also let $\delta \stackrel{\text{def}}{=} (\delta_1, \dots, \delta_q)$ be a vector of associated "optimality radii." Then we will say that x is a *strong* (ε,δ)-approximate qth-order minimizer if, for all $j \in \{1,\dots,q\}$, there exists a $\delta_j > 0$ such that

$$\phi_{f,j}^{\delta_j}(x) \le \varepsilon_j \frac{\delta_j^j}{j!}. \tag{2.4}$$

(The factor $j!$ is introduced for notational convenience.) The δ_j are called optimality radii because they are the radii of the *neighborhood of x in which the Taylor series $T_{f,j}(x,d)$ cannot decrease more than ε_j (appropriately scaled)*. Thus δ_j and ε_j are tightly linked (see Lemma 4.4 below) and the limit (2.3) (which applies at true local minimizers) is conceptually achieved when ε_j itself tends to zero. Note that (2.4) reduces to the condition $\|\nabla_x^1 f(x)\| \le \varepsilon_1$ for $j = 1$, and that, for $j = 2$, $\phi_{f,2}^{\delta_j}(x)$ is obtained by solving a trust-region subproblem,

[1] Meaning all its dimensions are the same.

a process whose cost is comparable to that of computing the leftmost eigenvalue of the Hessian, as would be required for the standard second-order measure.

The definition (2.4) should be contrasted with notion of weak minimizers introduced in [12]. Formally, x is a *weak* (ε,δ)-approximate qth-order minimizer if there exists $\delta_q \in \mathbb{R}$ such that

$$\phi_{f,q}^{\delta_q}(x) \le \varepsilon_q \chi_q(\delta_q) \quad \text{where } \chi_q(\delta) \overset{\text{def}}{=} \sum_{\ell=1}^{q} \frac{\delta^\ell}{\ell!}. \tag{2.5}$$

Obviously, (2.5) is less restrictive than (2.4) since it is easy to show that $\chi_q(\delta) \in [\delta, 2\delta)$ and is thus significantly larger than $\delta_q^q/q!$ for small δ_q. Moreover, (2.5) is a single condition, while (2.4) has to hold for all $j \in \{1,\ldots,q\}$. The interest of considering weak approximate minimizers is that they can be computed faster than strong ones. It is shown in [12] that the evaluation complexity bound for finding them is $O(\varepsilon^{-\frac{p+1}{p-q+1}})$, thereby providing a smooth extension to high-order of the complexity bounds known for $q \in \{1,2\}$. However, the major drawback of using the weak notion is that, at variance with (2.4), it is not coherent with the scaling implied by (2.3).[2] Obtaining this coherence therefore comes at a cost for orders beyond two, as will be clear in our developments below.

If we now consider that inexpensive constraints are present in the problem, it is easy to adapt the notions of weak and strong optimality for this case by (re)defining

$$\phi_{f,j}^{\delta_j}(x) \overset{\text{def}}{=} f(x) - \min_{x+d\in\mathcal{F},\,\|d\|\le\delta_j} T_{f,j}(x,d), \tag{2.6}$$

where $\mathcal{F}$ is the feasible set. We then say that x is a strong inexpensively constrained (ε,δ)-approximate qth-order minimizer if, for all $j \in \{1,\ldots,q\}$, there exists a $\delta_j > 0$ such that (2.4) holds with this new definition.

3. THE COMPOSITE PROBLEM AND ITS PROPERTIES

We now return to the more general composite optimization (1.1), and make our assumptions more specific:

AS.1 The function f from $\mathbb{R}^n$ to $\mathbb{R}$ is p times continuously differentiable and each of its derivatives $\nabla_x^\ell f(x)$ of order $\ell \in \{1,\ldots,p\}$ are Lipschitz continuous in a convex open neighborhood of $\mathcal{F}$, that is, for every $j \in \{1,\ldots,p\}$, there exists a constant $L_{f,j} \ge 1$ such that, for all x, y in that neighborhood,

$$\|\nabla_x^j f(x) - \nabla_x^j f(y)\| \le L_{f,j}\|x-y\|, \tag{3.1}$$

where $\|\cdot\|$ denotes the Euclidean norm for vectors and the induced operator norm for matrices and tensors.

AS.2 The function c from $\mathbb{R}^n$ to $\mathbb{R}^m$ is p times continuously differentiable and each of its derivatives $\nabla_x^\ell c(x)$ of order $\ell \in \{1,\ldots,p\}$ are Lipschitz continuous in

[2] In the worst case, it may lead to the origin being accepted as a second-order approximate minimizer of $-x^2$.

a convex open neighborhood of $\mathcal{F}$, that is, for every $j \in \{1, \dots, p\}$ there exists a constant $L_{c,j} \geq 1$ such that, for all x, y in that neighborhood,

$$\left\| \nabla_x^j c(x) - \nabla_x^j c(y) \right\| \leq L_{c,j} \|x - y\|, \tag{3.2}$$

AS.3 The function h from $\mathbb{R}^m$ to $\mathbb{R}$ is Lipschitz continuous, subbadditive, and zero at zero, that is, there exists a constant $L_{h,0} \geq 0$ such that, for all $x, y \in \mathbb{R}^m$,

$$\left\| h(x) - h(y) \right\| \leq L_{h,0} \|x - y\|, \tag{3.3}$$

$$h(x + y) \leq h(x) + h(y) \quad \text{and} \quad h(0) = 0. \tag{3.4}$$

AS.4 There is a constant w_{low} such that $w(x) \geq w_{\text{low}}$ for all $x \in \mathcal{F}$.

Assumption AS.3 allows a fairly general class of composition functions. Examples include the popular $\|\cdot\|_1$, $\|\cdot\|$, and $\|\cdot\|_\infty$ norms, concave functions vanishing at zero and, in the unidimensional case, the ReLu function $\max[0, \cdot]$ and the periodic $|\sin(\cdot)|$. As these examples show, nonconvexity and nondifferentiability are allowed (but not necessary). Note that finite sums of functions satisfying AS.3 also satisfy AS.3. Note also that h being subadditive does not imply that h^α is also subadditive for $\alpha \geq 1$ ($h(c) = c$ is, but $h(c)^2$ is not), or that it is concave [6]. Observe finally that equality always holds in (3.4) when h is odd.[3]

When h is smooth, problem (1.1) can be viewed either as composite or smooth. Does the composite view present any advantage in this case? The answer is that the assumptions needed on h in the composite case are weaker in that Lipschitz continuity is only required for h itself, not for its derivatives of orders 1 to p. If any of these derivatives are costly, unbounded or nonexistent, this can be a significant advantage. However, as we will see below (in Theorems 5.5 and 5.6) this comes at the price of a worse evaluation complexity bound. For example, the case of linear h is simple to assess, since in that case $h(c)$ amounts to a linear combination of the c_i, and there is obviously no costly or unbounded derivative involved: a smooth approach is therefore preferable from a complexity perspective.

Observe also that AS.1 and AS.2 imply, in particular, that

$$\left\| \nabla_x^j f(x) \right\| \leq L_{f,j-1} \quad \text{and} \quad \left\| \nabla_x^j c(x) \right\| \leq L_{c,j-1} \quad \text{for } j \in \{2, \dots, p\} \tag{3.5}$$

Observe also that AS.3 ensures that, for all $x \in \mathbb{R}^m$,

$$\left| h(x) \right| = \left| h(x) - h(0) \right| \leq L_{h,0} \|x - 0\| = L_{h,0} \|x\|. \tag{3.6}$$

For future reference, we define

$$L_w \stackrel{\text{def}}{=} \max_{j \in \{1, \dots, p\}} (L_{f,j-1} + L_{h,0} L_{c,j-1}). \tag{3.7}$$

We note that AS.4 makes the problem well-defined in that its objective function is bounded below. We now state a useful lemma on the Taylor expansion's error for a general function r with Lipschitz continuous derivative.

[3] Indeed, $h(-x - y) \leq h(-x) + h(-y)$ and thus, since h is odd, $-h(x + y) \leq -h(x) - h(y)$, which, combined with (3.4), gives that $h(x + y) = h(x) + h(y)$.

Lemma 3.1. *Let $r : \mathbb{R}^n \to \mathbb{R}$ be p times continuously differentiable and suppose that $\nabla_x^p r(x)$ is Lipschitz continuous with Lipschitz constant $L_{r,p}$, Let $T_{r,p}(x,s)$ be the pth degree Taylor approximation of $r(x+s)$ about x given by (2.1). Then for all $x, s \in \mathbb{R}^n$,*

$$\left|r(x+s) - T_{r,p}(x,s)\right| \leq \frac{L_{r,p}}{(p+1)!}\|s\|^{p+1}, \tag{3.8}$$

$$\left\|\nabla_x^j r(x+s) - \nabla_s^j T_{r,p}(x,s)\right\| \leq \frac{L_{r,p}}{(p-j+1)!}\|s\|^{p-j+1} \quad (j = 1, \dots, p). \tag{3.9}$$

Proof. See [12, LEMMA 2.1] with $\beta = 1$. ■

We now extend the concepts and notation of Section 2 to the case of composite optimization. Abusing notation slightly, we denote, for $j \in \{1, \dots, p\}$,

$$T_{w,j}(x,s) \stackrel{\text{def}}{=} T_{f,j}(x,s) + h\big(T_{c,j}(x,s)\big) \tag{3.10}$$

($T_{w,j}(x,s)$ it is *not* a Taylor expansion). We also define, for $j \in \{1, \dots, q\}$,

$$\begin{aligned}\phi_{w,j}^{\delta}(x) &\stackrel{\text{def}}{=} w(x) - \min_{x+d \in \mathcal{F}, \|d\| \leq \delta}\big[T_{f,j}(x,s) + h\big(T_{c,j}(x,s)\big)\big] \\ &= w(x) - \min_{x+d \in \mathcal{F}, \|d\| \leq \delta} T_{w,j}(x,s)\end{aligned} \tag{3.11}$$

by analogy with (2.6). This definition allows us to *consider (approximate) high-order minimizers of w, despite h being potentially nonsmooth*, because we have left h unchanged in the optimality measure (3.11), rather than using a Taylor expansion of h.

We now state a simple first-order necessary optimality condition for a global minimizer of composite problems of the form (1.1) with convex h.

Lemma 3.2. *Suppose that f and c are continuously differentiable and that AS.3 holds. Suppose in addition that h is convex and that x_* is a global minimizer of w. Then the origin is a global minimizer of $T_{w,1}(x_*, s)$ and $\phi_{w,1}^{\delta}(x_*) = 0$ for all $\delta > 0$.*

Proof. Suppose now that the origin is not a global minimizer of $T_{w,1}(x_*, s)$, but that there exists an $s_1 \neq 0$ with $T_{w,1}(x_*, s_1) < T_{w,1}(x_*, 0) = w(x_*)$. By Taylor's theorem, we obtain that, for $\alpha \in [0,1]$,

$$f(x_* + \alpha s_1) = T_{f,1}(x_*, \alpha s_1) + o(\alpha), \quad c(x_* + \alpha s_1) = T_{c,1}(x_*, \alpha s_1) + o(\alpha) \tag{3.12}$$

and, using AS.3 and (3.6),

$$\begin{aligned}h\big(c(x_* + \alpha s_1)\big) &= h\big(T_{c,1}(x_*, \alpha s_1) + o\big(\alpha\|s_1\|\big)\big) \leq h\big(T_{c,1}(x_*, \alpha s_1)\big) + h\big(o(\alpha)\|s_1\|\big) \\ &\leq h\big(T_{c,1}(x_*, \alpha s_1)\big) + o(\alpha)L_{h,0}\|s_1\| = h\big(T_{c,1}(x_*, \alpha s_1)\big) + o(\alpha).\end{aligned} \tag{3.13}$$

Now note that the convexity of h and the linearity of $T_{f,1}(x_*, s)$ and $T_{c,1}(x_*, s)$ imply that $T_{w,1}(x_*, s)$ is convex, and thus that $T_{w,1}(x_*, \alpha s_1) - w(x_*) \leq \alpha[T_{w,1}(x_*, s_1) - w(x_*)]$. Hence, using (3.12) and (3.13), we deduce that

$$\begin{aligned}0 \leq w(x_* + \alpha s_1) - w(x_*) &\leq T_{w,1}(x_*, \alpha s_1) - w(x_*) + o(\alpha) \\ &\leq \alpha\big[T_{w,1}(x_*, s_1) - w(x_*)\big] + o(\alpha),\end{aligned}$$

which is impossible for α sufficiently small, since $T_{w,1}(x_*, s_1) - w(x_*) < 0$ by construction of s_1. As a consequence, the origin must be a global minimizer of the convex $T_{w,1}(x_*, s)$ and therefore $\phi^{\delta}_{w,1}(x_*) = 0$ for all $\delta > 0$. ■

Unfortunately, this result does not extend to $\phi^{\delta}_{w,q}(x)$ when $q = 2$, as is shown by the following example. Consider the univariate $w(x) = -\frac{2}{5}x + |x - x^2 + 2x^3|$, where h is the (convex) absolute value function satisfying AS.3. Then $x_* = 0$ is a global minimizer of w (plotted as unbroken line in Figure 3.1) and yet

$$T_{w,2}(x_*, s) = T_{f,2}(x_*, s) + \left|T_{c,2}(x_*, s)\right| = -\frac{2}{5}s + \left|s - s^2\right|$$

(plotted as dashed line in the figure) admits a global minimum for $s = 1$ whose value $(-\frac{2}{5})$ is smaller that $w(x_*) = 0$. Thus $\phi^{1}_{w,2}(x_*) > 0$ despite x_* being a global minimizer. But it is clear in the figure that $\phi^{\delta}_{w,2}(x_*) = 0$ for δ smaller than $\frac{1}{2}$.

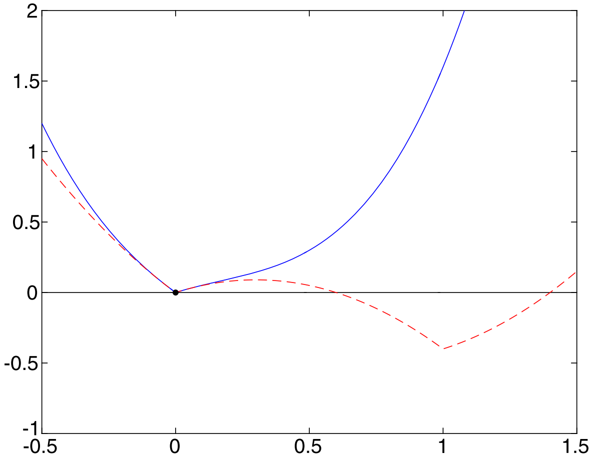

FIGURE 3.1
Functions $w(x)$ (unbroken) and $T_{w,2}(0, s) = T_{f,2}(0, s) + |T_{c,2}(0, s)|$ (dashed).

In the smooth ($h = 0$) case, Lemma 3.2 may be extended for unconstrained (i.e., $\mathcal{F} = \mathbb{R}^n$) twice-continuously differentiable f since then standard second-order optimality conditions at a global minimizer x_* of f imply that $T_{f,j}(x_*, d)$ is convex for $j = 1, 2$ and thus that $\phi^{\delta}_{f,1}(x_*) = \phi^{\delta}_{f,2}(x_*) = 0$. When constraints are present (i.e., $\mathcal{F} \subset \mathbb{R}^n$), unfortunately, this may require that we restrict δ. For example, the global minimizer of $f(x) = -(x - 1/3)^2 + 2/3x^3$ for $x \in [0, 1]$ lies at $x_* = 0$, but $T_{f,2}(x_*, d) = -(d - 1/3)^2$ which has its constrained global minimizer at $d = 1$ with $T_{f,2}(x_*, 1) < T_{f,2}(x_*, 0)$ and we would need $\delta \leq 2/3$ to ensure that $\phi^{\delta}_{f,2}(x_*) = 0$.

4. AN ADAPTIVE REGULARIZATION ALGORITHM FOR COMPOSITE OPTIMIZATION

We now consider an adaptive regularization algorithm to search for a (strong) (ε,δ)-approximate qth-order minimizer for problem (1.1), that is a point $x_k \in \mathcal{F}$ such that

$$\phi_{w,j}^{\delta}(x_k) \le \varepsilon_j \frac{\delta_j^j}{j!} \quad \text{for } j \in \{1,\dots,q\}, \tag{4.1}$$

where $\phi_{w,q}^{\delta}(x)$ is defined in (3.11). At each iteration, the algorithm seeks a feasible approximate minimizer of the (possibly nonsmooth) regularized model

$$\begin{aligned} m_k(s) &= T_{f,p}(x_k,s) + h\big(T_{c,p}(x_k,s)\big) + \frac{\sigma_k}{(p+1)!}\|s\|^{p+1} \\ &= T_{w,p}(x_k,s) + \frac{\sigma_k}{(p+1)!}\|s\|^{p+1} \end{aligned} \tag{4.2}$$

and this process is allowed to terminate whenever

$$m_k(s) \le m_k(0) \tag{4.3}$$

and, for each $j \in \{1,\dots,q\}$,

$$\phi_{m_k,j}^{\delta_{s,j}}(s) \le \theta\varepsilon_j \frac{\delta_{s,j}^j}{j!} \tag{4.4}$$

for some $\theta \in (0,1)$. Observe that $m_k(s)$ is bounded below since (3.6) ensures that the regularization term of degree $p+1$ dominates for large steps. Obviously, the inclusion of h in the definition of the model (4.2) implicitly assumes that, as is common, the cost of evaluating h is small compared with that of evaluating f or c. It also implies that computing $\phi_{w,j}^{\delta_j}(x)$ and $\phi_{m_k,j}^{\delta_{s,j}}(s)$ is potentially more complicated than in the smooth case, although it does not impact the evaluation complexity of the algorithm because the model's approximate minimization does not involve evaluating f, c or any of their derivatives.

The rest of the algorithm, that we shall refer to as AR$qp$C, follows the standard pattern of adaptive regularization algorithms, and is stated on this page. As everywhere in this paper, we assume that $q \in \{1,\dots,p\}$.

AR$qp$C algorithm for finding an (ε,δ)-approximate qth-order minimizer of the composite function w in (1.1)

Step 0: Initialization. An initial point x_0 and an initial regularization parameter $\sigma_0 > 0$ are given, as well as an accuracy level $\varepsilon \in (0,1)^q$. The constants δ_0, θ, η_1, η_2, γ_1, γ_2, γ_3, and $\sigma_{\min}$ are also given and satisfy

$$\begin{aligned} &\theta \in (0,1), \quad \delta_0 \in (0,1]^q, \quad \sigma_{\min} \in (0,\sigma_0], \quad 0 < \eta_1 \le \eta_2 < 1, \\ &\text{and} \quad 0 < \gamma_1 < 1 < \gamma_2 < \gamma_3. \end{aligned} \tag{4.5}$$

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

Step 1: Test for termination. Evaluate $\{\nabla_x^i f(x_k)\}_{i=1}^q$ and $\{\nabla_x^i c(x_k)\}_{i=1}^q$. If (4.1) holds with $\delta = \delta_k$, terminate with the approximate solution $x_\varepsilon = x_k$. Otherwise compute $\{\nabla_x^i f(x_k)\}_{i=q+1}^p$ and $\{\nabla_x^i c(x_k)\}_{i=q+1}^p$.

Step 2: Step calculation. Attempt to compute an approximate minimizer s_k of model $m_k(s)$ given in (4.2) such that $x_k + s_k \in \mathcal{F}$ and optimality radii $\delta_{s_k} \in (0, 1]^q$ exist such that (4.3) holds and (4.4) holds for $j \in \{1, \dots, q\}$ and $s = s_k$. If no such step exists, terminate with the approximate solution $x_\varepsilon = x_k$.

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

$$\rho_k = \frac{w(x_k) - w(x_k + s_k)}{w(x_k) - T_{w,p}(x_k, s)}. \tag{4.6}$$

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

Step 4: Regularization parameter update. Set

$$\sigma_{k+1} \in \begin{cases} [\max(\sigma_{\min}, \gamma_1 \sigma_k), \sigma_k] & \text{if } \rho_k \geq \eta_2, \\ [\sigma_k, \gamma_2 \sigma_k] & \text{if } \rho_k \in [\eta_1, \eta_2), \\ [\gamma_2 \sigma_k, \gamma_3 \sigma_k] & \text{if } \rho_k < \eta_1. \end{cases} \tag{4.7}$$

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

As expected, the AR$qp$C algorithm shows obvious similarities with that discussed in [12], but differs from it in significant ways. Beyond the fact that it now handles composite objective functions, the main one being that the termination criterion in Step 1 now tests for strong approximate minimizers, rather than weak ones.

As is standard for adaptive regularization algorithms, we say that an iteration is successful when $\rho_k \geq \eta_1$ (and $x_{k+1} = x_k + s_k$) and that it is unsuccessful otherwise. We denote by $\mathcal{S}_k$ the index set of all successful iterations from 0 to k, that is,

$$\mathcal{S}_k = \big\{ j \in \{0, \dots, k\} \mid \rho_j \geq \eta_1 \big\},$$

and then obtain a well-known result ensuring that successful iterations up to iteration k do not amount to a vanishingly small proportion of these iterations.

Lemma 4.1. *The mechanism of the ARqpC algorithm guarantees that, if*

$$\sigma_k \leq \sigma_{\max}, \tag{4.8}$$

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

$$k + 1 \leq |\mathcal{S}_k| \left(1 + \frac{|\log \gamma_1|}{\log \gamma_2} \right) + \frac{1}{\log \gamma_2} \log \left(\frac{\sigma_{\max}}{\sigma_0} \right). \tag{4.9}$$

Proof. See [5, THEOREM 2.4]. ■

We also have the following identity for the norm of the successive derivatives of the regularization term.

Lemma 4.2. *Let s be a vector of $\mathbb{R}^n$. Then*

$$\left\| \nabla_s^j \left(\|s\|^{p+1} \right) \right\| = \frac{(p+1)!}{(p-j+1)!} \|s\|^{p-j+1} \quad \textit{for } j \in \{0, \dots, p+1\}. \tag{4.10}$$

Moreover, $\nabla_s^j(\|s\|^{p+1})[d]^j$ is a convex function in d for any d orthogonal to s. It is also convex for any multiple of s whenever j is even.

Proof. See [12, LEMMA 2.4] with $\beta = 1$. ■

As the conditions for accepting a pair (s_k, δ_s) in Step 2 are stronger than previously considered (in particular, they are stronger than those discussed in [12]), we must ensure that such acceptable pairs exist. We start by recalling a result discussed in [12] for the smooth case.

Lemma 4.3. *Suppose that*

$$\begin{array}{llll} (q = 1, & \mathcal{F} \text{ is convex}, & \text{and} & h \text{ is convex}), \text{ or} \\ (q = 2, & \mathcal{F} = \mathbb{R}^n & \text{and} & h = 0). \end{array} \tag{4.11}$$

Suppose in addition that $s_k^* \neq 0$ is a global minimizer of $m_k(s)$ for $x_k + s \in \mathcal{F}$. Then there exist a feasible neighborhood of s_k^* such that (4.3) and (4.4) hold for any s_k in this neighborhood with $\delta_s = 1$.

Proof. Consider first the composite case where $q = 1$. We have that

$$\begin{aligned} T_{m_k,1}(s_k^*, d) = {} & T_{f,p}(x_k, s_k^*) + \nabla_s^1 T_{f,p}(x_k, s_k^*)[d] + h\big(T_{c,p}(x_k, s_k^*) + \nabla_s^1 T_{c,p}(x_k, s_k^*)[d]\big) \\ & + \frac{\sigma_k}{(p+1)!}\left(\big|\big|s_k^*\big|\big|^{p+1} + \nabla_s^1 \big|\big|s_k^*\big|\big|^{p+1}[d]\right) \end{aligned}$$

is a convex function of d (since h is convex, all terms in the above right-hand side are). Suppose now that it has a feasible global minimizer d_* such that $T_{m_k,1}(s_k^*, d_*) < T_{m_k,1}(s_k^*, 0) = m_k(s_k^*)$. Since $\mathcal{F}$ is convex, $T_{m_k,1}(s_k^*, d') < m_k(s_k^*)$ for all d' in the segment $(0, d_*]$. But (3.5) implies that

$$\begin{aligned} \big\|\nabla_s^\ell T_{f,p}(x_k, s_k^*)\big\| &\le \sum_{i=\ell}^{p} \frac{1}{(i-\ell)!}\big\|\nabla_x^i f(x_k)\big\|\,\big\|s_k^*\big\|^{i-\ell} \le \max_{j\in\{2,\dots,p\}} L_{f,j-1} \sum_{i=\ell}^{p} \frac{\|s_k^*\|^{i-\ell}}{(i-\ell)!}, \\ \big\|\nabla_s^\ell T_{c,p}(x_k, s_k^*)\big\| &\le \sum_{i=\ell}^{p} \frac{1}{(i-\ell)!}\big\|\nabla_x^i c(x_k)\big\|\,\big\|s_k^*\big\|^{i-\ell} \le \max_{j\in\{2,\dots,p\}} L_{c,j-1} \sum_{i=\ell}^{p} \frac{\|s_k^*\|^{i-\ell}}{(i-\ell)!} \end{aligned}$$

and both must be bounded for s_k^* given. Thus, $T_{m_k,1}(s_k^*, d')$ approximates $m_k(s_k^* + d')$ arbitrarily well for small enough $\|d'\|$, and therefore $m_k(s_k^* + d') < m_k(s_k^*)$ for small enough $\|d'\|$, which is impossible since s_k^* is a global minimizer of $m_k(s)$. As a consequence $d = 0$ must be a global minimizer of $T_{m_k,1}(s_k^*, d)$. Thus $\phi_{m_k,1}^\delta(s_k^*) = 0$ for all $\delta > 0$, and in particular for $\delta = 1$, which, by continuity, yields the desired conclusion.

Consider now the case where $q = 2$, $h = 0$ and $\mathcal{F} = \mathbb{R}^n$. Suppose that $j = 1$ ($j = 2$). Then the jth order Taylor expansion of the model at s_k^* is a linear (positive semidefinite quadratic) polynomial, which is a convex function. As a consequence, we obtain as above that $\phi_{m_k,j}^\delta(s_k^*) = 0$ for all $\delta_{s,j} > 0$ and the conclusion then again follows. ■

Alas, the example given at the end of Section 3 implies that δ_s may have to be chosen smaller than one for $q = 2$ and when h is nonzero, even if it is convex. Fortunately, the existence of a step is still guaranteed in general, even without assuming convexity of h. To state our result, we first define ξ to be an arbitrary constant in $(0, 1)$ independent of ε, which we specify later.

Lemma 4.4. *Let $\xi \in (0,1)$ and suppose that s_k^* is a global minimizer of $m_k(s)$ for $x_k + s \in \mathcal{F}$ such that $m_k(s_k^*) < m_k(0)$. Then there exists a pair $(\bar{s}, \delta_s)$ such that (4.3) and (4.4) hold. Moreover, one has that either $\|\bar{s}\| \geq \xi$ or (4.3) and (4.4) hold for $\bar{s}$ for all $\delta_{s,j}$ $(j \in \{1,\dots,q\})$, for which*

$$0 < \delta_{s,j} \leq \frac{\theta}{q!(6L_w + 2\sigma_k)}\,\varepsilon_j. \tag{4.12}$$

Proof. We first need to show that a pair $(\bar{s}, \delta_s)$ satisfying (4.3) and (4.4) exists. Since $m_k(s_k^*) < m_k(0)$, we have that $s_k^* \neq 0$. By Taylor's theorem, we have that, for all d,

$$\begin{aligned}
0 &\leq m_k(s_k^* + d) - m_k(s_k^*)\\
&= \sum_{\ell=1}^{p} \frac{1}{\ell!}\nabla_s^\ell T_{f,p}(x_k, s_k^*)[d]^\ell + h\left(\sum_{\ell=0}^{p} \frac{1}{\ell!}\nabla_s^\ell T_{c,p}(x_k, s_k^*)[d]^\ell\right) - h\big(T_{c,p}(x_k, s_k^*)\big)\\
&\quad + \frac{\sigma_k}{(p+1)!}\left[\sum_{\ell=1}^{p} \frac{1}{\ell!}\nabla_s^\ell\big(\|s_k^*\|^{p+1}\big)[d]^\ell + \frac{1}{(p+1)!}\nabla_s^{p+1}\big(\|s_k^* + \tau d\|^{p+1}\big)[d]^{p+1}\right]
\end{aligned} \tag{4.13}$$

for some $\tau \in (0,1)$. Using (4.10) in (4.13) and the subadditivity of h ensured by AS.3 then yields that, for any $j \in \{1,\dots,q\}$ and all d,

$$\begin{aligned}
&-\sum_{\ell=1}^{j} \frac{1}{\ell!}\nabla_s^\ell T_{f,p}(x_k, s_k^*)[d]^\ell + h\big(T_{c,p}(x_k, s_k^*)\big) - h\left(\sum_{\ell=0}^{j} \frac{1}{\ell!}\nabla_s^\ell T_{c,p}(x_k, s_k^*)[d]^\ell\right)\\
&\quad - \frac{\sigma_k}{(p+1)!}\sum_{\ell=1}^{j} \nabla_s^\ell \|s_k^*\|^{p+1}[d]^\ell\\
&\leq \sum_{\ell=j+1}^{p} \frac{1}{\ell!}\nabla_s^\ell T_{f,p}(x_k, s_k^*)[d]^\ell + h\left(\sum_{\ell=j+1}^{q} \frac{1}{\ell!}\nabla_s^\ell T_{c,p}(x_k, s_k^*)[d]^\ell\right)\\
&\quad + \frac{\sigma_k}{(p+1)!}\left[\sum_{\ell=j+1}^{p} \frac{1}{\ell!}\nabla_s^\ell \|s_k^*\|^{p+1}[d]^\ell + \|d\|^{p+1}\right].
\end{aligned} \tag{4.14}$$

Since $s_k^* \neq 0$, and using (3.6), we may then choose $\delta_{s,j} \in (0,1]$ such that, for every d with $\|d\| \leq \delta_{s,j}$, we have

$$\begin{aligned}
&\sum_{\ell=j+1}^{p} \frac{1}{\ell!}\nabla_s^\ell T_{f,p}(x_k, s_k^*)[d]^\ell + h\left(\sum_{\ell=j+1}^{p} \frac{1}{\ell!}\nabla_s^\ell T_{c,p}(x_k, s_k^*)[d]^\ell\right)\\
&\quad + \frac{\sigma_k}{(p+1)!}\left[\sum_{\ell=j+1}^{p} \frac{1}{\ell!}\nabla_s^\ell \|s_k^*\|^{p+1}[d]^\ell + \|d\|^{p+1}\right] \leq \frac{1}{2}\theta\varepsilon_j\,\frac{\delta_{s,j}^j}{j!}.
\end{aligned} \tag{4.15}$$

As a consequence, we obtain that if $\delta_{s,j}$ is small enough to ensure (4.15), then (4.14) implies

$$\begin{aligned}
&-\sum_{\ell=1}^{j} \frac{1}{\ell!}\nabla_s^\ell T_{f,p}(x_k, s_k^*)[d]^\ell + h\big(T_{c,p}(x_k, s_k^*)\big) - h\left(\sum_{\ell=0}^{j} \frac{1}{\ell!}\nabla_s^\ell T_{c,p}(x_k, s_k^*)[d]^\ell\right)\\
&\quad - \frac{\sigma_k}{(p+1)!}\sum_{\ell=1}^{j} \nabla_s^\ell \|s_k^*\|^{p+1}[d]^\ell \leq \frac{1}{2}\theta\varepsilon_j\,\frac{\delta_{s,j}^j}{j!}.
\end{aligned} \tag{4.16}$$

The fact that, by definition,

$$\phi_{m_k,j}^{\delta_{s,j}}(s) = \max\left[0, \max_{\|d\|\le\delta_{s,j}}\left\{-\sum_{\ell=1}^{j}\frac{1}{\ell!}\nabla_s^\ell T_{f,p}(x_k,s)[d]^\ell + h\big(T_{c,p}(x_k,s_k)\big)\right.\right.$$
$$\left.\left. - h\left(\sum_{\ell=0}^{j}\frac{1}{\ell!}\nabla_s^\ell T_{c,p}(x_k,s)[d]^\ell\right) - \frac{\sigma_k}{(p+1)!}\sum_{\ell=1}^{j}\frac{1}{\ell!}\nabla_s^\ell\|s\|^{p+1}[d]^\ell\right\}\right], \tag{4.17}$$

continuity of $T_{f,p}(x_k, s)$ and $T_{c,p}(x_k, s)$ and their derivatives and the inequality $m_k(s_k^*) < m_k(0)$ then ensure the existence of a feasible neighborhood of $s_k^* \neq 0$ in which $\overline{s}$ can be chosen such that (4.3) and (4.4) hold for $s = \overline{s}$, concluding the first part of the proof.

To prove the second part, assume first that $\|s_k^*\| \ge 1$. We may then restrict the neighborhood of s_k^* in which $\overline{s}$ can be chosen enough to ensure that $\|\overline{s}\| \ge \xi$. Assume therefore that $\|s_k^*\| \le 1$. Remembering that, by definition and the triangle inequality,

$$\big\|\nabla_s^\ell T_{f,p}(x_k,s_k^*)\big\| \le \sum_{j=\ell}^{p}\frac{1}{(j-\ell)!}\big\|\nabla_x^j f(x_k)\big\|\,\|s_k^*\|^{j-\ell},$$
$$\big\|\nabla_s^\ell T_{c,p}(x_k,s_k^*)\big\| \le \sum_{j=\ell}^{p}\frac{1}{(j-\ell)!}\big\|\nabla_x^j c(x_k)\big\|\,\|s_k^*\|^{j-\ell},$$

for $\ell \in \{q+1,\dots,p\}$, and thus, using (3.6), (3.5), and (4.10), we deduce that

$$\begin{aligned}
&\sum_{\ell=j+1}^{p}\frac{1}{\ell!}\nabla_s^\ell T_{f,p}(x_k,s_k^*)[d]^\ell + h\left(\sum_{\ell=j+1}^{p}\frac{1}{\ell!}\nabla_s^\ell T_{c,p}(x_k,s_k^*)[d]^\ell\right)\\
&\quad + \frac{\sigma_k}{(p+1)!}\left[\sum_{\ell=j+1}^{p}\nabla_s^\ell\|s_k^*\|^{p+1}[d]^\ell\right]\\
&\le \sum_{\ell=j+1}^{p}\frac{1}{\ell!}\nabla_s^\ell T_{f,p}(x_k,s_k^*)[d]^\ell + L_{h,0}\left\|\sum_{\ell=j+1}^{p}\frac{1}{\ell!}\nabla_s^\ell T_{c,p}(x_k,s_k^*)[d]^\ell\right\|\\
&\quad + \frac{\sigma_k}{(p+1)!}\left[\sum_{\ell=j+1}^{p}\nabla_s^\ell\|s_k^*\|^{p+1}[d]^\ell\right]\\
&\le \sum_{\ell=j+1}^{p}\frac{\|d\|^\ell}{\ell!}\left[\sum_{i=\ell}^{p}\frac{\|s_k^*\|^{i-\ell}}{(i-\ell)!}\big(\|\nabla_x^i f(x_k)\| + L_{h,0}\|\nabla_x^i c(x_k)\|\big) + \frac{\sigma_k\|s_k^*\|^{p-\ell+1}}{(p-\ell+1)!}\right]\\
&\le \sum_{\ell=j+1}^{p}\frac{\|d\|^\ell}{\ell!}\left[L_w\sum_{i=\ell}^{p}\frac{\|s_k^*\|^{i-\ell}}{(i-\ell)!} + \frac{\sigma_k\|s_k^*\|^{p-\ell+1}}{(p-\ell+1)!}\right],
\end{aligned}$$

where L_w is defined in (3.7). We therefore obtain from (4.15) that any pair $(s_k^*, \delta_{s,j})$ satisfies (4.16) for $\|d\| \le \delta_{s,j}$ if

$$\sum_{\ell=j+1}^{p}\frac{\delta_{s,j}^\ell}{\ell!}\left[L_w\sum_{i=\ell}^{p}\frac{1}{(i-\ell)!}\big\|s_k^*\big\|^{i-\ell} + \frac{\sigma_k\|s_k^*\|^{p-\ell+1}}{(p-\ell+1)!}\right] + \sigma_k\frac{\delta_{s,j}^{p+1}}{(p+1)!} \le \frac{1}{2}\theta\varepsilon_j\frac{\delta_{s,j}^j}{j!}. \tag{4.18}$$

which, because $\|s_k^*\| \leq 1$, is in turn ensured by the inequality

$$\sum_{\ell=j+1}^{p} \frac{\delta_{s,j}^{\ell}}{\ell!}\left[L_w \sum_{i=\ell}^{p} \frac{1}{(i-\ell)!} + \sigma_k\right] + \sigma_k \frac{\delta_{s,j}^{p+1}}{(p+1)!} \leq \frac{1}{2}\theta\varepsilon_j \frac{\delta_{s,j}^{j}}{j!}. \tag{4.19}$$

Observe now that, since $\delta_{s,j} \in [0,1]$, we have $\delta_{s,j}^{\ell} \leq \delta_{s,j}^{j+1}$ for $\ell \in \{j+1,\dots,p\}$. Moreover, we have that,

$$\sum_{i=\ell}^{p} \frac{1}{(i-\ell)!} \leq e < 3 \quad \big(\ell \in \{j+1,\dots,p+1\}\big), \qquad \sum_{\ell=j+1}^{p+1} \frac{1}{\ell!} \leq e-1 < 2,$$

and therefore (4.19) is (safely) guaranteed by the condition

$$j!(6L_w + 2\sigma_k)\,\delta_{s,j} \leq \frac{1}{2}\theta\varepsilon_j, \tag{4.20}$$

which means that the pair (s_k^*, δ_s) satisfies (4.16) for all $j \in \{1,\dots,q\}$ whenever

$$\delta_{s,j} \leq \frac{\frac{1}{2}\theta\varepsilon_j}{q!(6L_w + 2\sigma_k)} \stackrel{\text{def}}{=} \frac{1}{2}\delta_{\min,k}.$$

We may thus again invoke the continuity of the derivatives of m_k and (4.17) to deduce that there exists a neighborhood of s_k^* such that, for every $\overline{s}$ in this neighborhood, $m_k(\overline{s}) < m_k(0)$ and the pair $(\overline{s}, \delta_{\min,k})$ satisfies $\phi_{m_k,j}^{\delta_{\min,k}}(\overline{s}) \leq \theta\varepsilon_j \frac{\delta_{\min,k}^{j}}{j!}$, yielding the desired conclusion. ■

This lemma indicates that either the norm of the step is larger than ξ, or the range of acceptable $\delta_{s,j}$ is not too small in that any positive value at most equal to the right-hand side of (4.12) can be chosen. Thus any value larger than a fixed fraction (a half, say) of (4.12) is also acceptable. Such a value is, for instance, guaranteed if $\delta_{s,j}$ is chosen according to the technique described as Algorithm 4.1.

A detailed Step 2 for the AR$qp$C algorithm (Algorithm 4.1)

```
Step 2: Step calculation.
    Step 2.1: Compute a descent step $s_k$ such that
                  $m_k(s_k) < m_k(0)$
              and either $\|s_k\| \geq 1$ or $s_k$ is the global minimizer of $m_k(s)$ for
              $\|s\| \leq 1$. If no such step exists, terminate the AR$qp$C algorithm
              with the approximate solution $x_\varepsilon = x_k$.
    Step 2.2: For $j \in \{1,\dots,q\}$, set $\delta_{s,j} = 1$.
    Step 2.3: If $\|s_k\| > 1$, return the pair $(s_k, \delta_s)$.
    Step 2.4: For each $j \in \{3,\dots,q\}$,
        (i)  compute the global minimum of $T_{m_k,j}(s_k, d)$ over all $d$
             such that $\|d\| \leq \delta_{s,j}$;
```

(ii) if

$$\phi_{m_k,j}^{\delta_{s,j}}(s_k) \le \theta\varepsilon_j \, \frac{\delta_{s,j}^j}{j!}$$

consider the next value of j; else set $\delta_{s,j} = \frac{1}{2}\delta_{s,j}$ and return to Step 2.4(ii).

Step 2.5: Return the pair (s_k, δ_s). ■

Lemma 4.4 then ensures that this conceptual algorithm is well-defined (and, in particular, that the loop within Step 2.4 is finite for each j). We therefore assume, without loss of generality, that, if some constant $\sigma_{\max}$ is given such that $\sigma_k \le \sigma_{\max}$ for all k, then the AR$qp$C algorithm ensures that

$$\delta_{s,j} \ge \kappa_{\delta,\min}\, \varepsilon_j \quad \text{with} \quad \kappa_{\delta,\min} \overset{\text{def}}{=} \frac{\theta}{2q!(6L_w + 2\sigma_{\max})} \in \left(0, \frac{1}{2}\right) \tag{4.21}$$

for $j \in \{1, \dots, q\}$ whenever $\|s_k\| \le \xi$.

We also need to establish that the possibility of termination in Step 2 of the AR$qp$C algorithm is a satisfactory outcome.

Lemma 4.5. *Termination cannot occur in Step 2 of the ARqpC algorithm if $q = 1$ and h is convex. In other cases, if the ARqpC algorithm terminates in Step 2 of iteration k with $x_\varepsilon = x_k$, then there exists a $\delta \in (0, 1]^q$ such that (4.1) holds for $x = x_\varepsilon$ and x_ε is a strong (ε, δ)-approximate qth-order-necessary minimizer.*

Proof. Given Lemma 4.4, if the algorithm terminates within Step 2, it must be because every (feasible) global minimizer s_k^* of $m_k(s)$ is such that $m_k(s_k^*) \ge m_k(0)$. In that case, $s_k^* = 0$ is one such global minimizer. If $q = 1$ and h is convex, Lemma 3.2 ensures that termination must have happened in Step 1, and termination in Step 2 is thus impossible. Otherwise, we have that, for any $j \in \{1, \dots, q\}$ and all d with $x_k + d \in \mathcal{F}$,

$$\begin{aligned}
0 \le m_k(d) - m_k(0) &= \sum_{\ell=1}^{j} \frac{1}{\ell!}\nabla_x^\ell f(x_k)[d]^\ell + \sum_{\ell=j+1}^{p} \frac{1}{\ell!}\nabla_x^\ell f(x_k)[d]^\ell \\
&\quad + h\left(c(x_k) + \sum_{\ell=1}^{j} \frac{1}{\ell!}\nabla_x^\ell c(x_k)[d]^\ell + \sum_{\ell=j+1}^{p} \frac{1}{\ell!}\nabla_x^\ell c(x_k)[d]^\ell\right) \\
&\quad + \frac{\sigma_k}{(p+1)!}\|d\|^{p+1} - h\big(c(x_k)\big) \\
&\le \sum_{\ell=1}^{j} \frac{1}{\ell!}\nabla_x^\ell f(x_k)[d]^\ell + \sum_{\ell=j+1}^{p} \frac{1}{\ell!}\nabla_x^\ell f(x_k)[d]^\ell + h\left(\sum_{\ell=1}^{j} \frac{1}{\ell!}\nabla_x^\ell c(x_k)[d]^\ell\right) \\
&\quad + h\left(\sum_{\ell=j+1}^{p} \frac{1}{\ell!}\nabla_x^\ell c(x_k)[d]^\ell\right) + \frac{\sigma_k}{(p+1)!}\|d\|^{p+1},
\end{aligned}$$

where we used the subadditivity of h (ensured by AS.3) to derive the last inequality. Hence

$$\begin{aligned}
-\sum_{\ell=1}^{j}\frac{1}{\ell!}\nabla_x^\ell f(x_k)[d]^\ell - h\left(\sum_{\ell=1}^{j}\frac{1}{\ell!}\nabla_x^\ell c(x_k)[d]^\ell\right)&\\
\leq \sum_{\ell=j+1}^{p}\frac{1}{\ell!}\nabla_x^\ell f(x_k)[d]^\ell + h\left(\sum_{\ell=j+1}^{p}\frac{1}{\ell!}\nabla_x^\ell c(x_k)[d]^\ell\right) + \frac{\sigma_k}{(p+1)!}\|d\|^{p+1}.&
\end{aligned}$$

Using (3.6), we may now choose each $\delta_j \in (0,1]$ for $j \in \{1,\dots,q\}$ small enough to ensure that the absolute value of the last right-hand side is at most $\varepsilon_j \delta_{k,j}^j / j!$ for all d with $\|d\| \leq \delta_{k,j}$ and $x_k + d \in \mathcal{F}$, which, in view of (3.11), implies (4.1). ■

5. EVALUATION COMPLEXITY

To analyze the evaluation complexity of the AR$qp$C algorithm, we first derive the predicted decrease in the unregularized model from (4.2).

Lemma 5.1. *At every iteration k of the AR$qp$C algorithm, one has that*

$$w(x_k) - T_{w,p}(x_k, s_k) \geq \frac{\sigma_k}{(p+1)!}\|s_k\|^{p+1}. \tag{5.1}$$

Proof. Immediate from (4.2) and (3.10), the fact that $m_k(0) = w(x_k)$ and (4.3). ■

We next derive the existence of an upper bound on the regularization parameter for the structured composite problem. The proof of this result hinges on the fact that, once the regularization parameter σ_k exceeds the relevant Lipschitz constant ($L_{w,p}$ here), there is no need to increase it any further because the model then provides an overestimation of the objective function.

Lemma 5.2. *Suppose that AS.1–AS.3 hold. Then, for all $k \geq 0$,*

$$\sigma_k \leq \sigma_{\max} \overset{\text{def}}{=} \max\left[\sigma_0, \frac{\gamma_3 L_{w,p}}{1-\eta_2}\right], \tag{5.2}$$

where $L_{w,p} = L_{f,p} + L_{h,0}L_{c,p}$.

Proof. Successively using (4.6), Theorem 3.1 applied to f and c, and (5.1), we deduce that, at iteration k,

$$\begin{aligned}
|\rho_k - 1| &= \left|\frac{w(x_k) - w(x_k+s_k)}{w(x_k) - T_{w,p}(x_k,s)} - 1\right|\\
&= \frac{|f(x_k+s_k) + h(c(x_k+s_k)) - T_{f,p}(x_k,s) - h(T_{c,p}(x_k,s))|}{w(x_k) - T_{w,p}(x_k,s)}\\
&\leq \frac{\frac{L_{f,p}\|s_k\|^{p+1}}{(p+1)!} + L_{h,0}\|c(x_k+s_k) - T_{c,p}(x_k,s)\|}{w(x_k) - T_{w,p}(x_k,s)}\\
&\leq \frac{\frac{L_{f,p}+L_{h,0}L_{c,p}}{(p+1)!}\|s_k\|^{p+1}}{\frac{\sigma_k}{(p+1)!}\|s_k\|^{p+1}} = \frac{L_{f,p}+L_{h,0}L_{c,p}}{\sigma_k}.
\end{aligned}$$

Thus, if $\sigma_k \geq L_{w,p}/(1-\eta_2)$, then iteration k is successful, $x_{k+1} = x_k$, and (4.7) implies that $\sigma_{k+1} \leq \sigma_k$. The conclusion then follows from the mechanism of (4.7). ■

We now establish an important inequality derived from our smoothness assumptions.

Lemma 5.3. *Suppose that AS.1–AS.3 hold. Suppose also that iteration k is successful and that the ARqpC algorithm does not terminate at iteration $k+1$. Then there exists a $j \in \{1,\dots,q\}$ such that*

$$(1-\theta)\,\varepsilon\,\frac{\delta_{k+1,j}^j}{j!} \leq (L_{w,p}+\sigma_{\max})\sum_{\ell=1}^{j}\frac{\delta_{k+1,j}^\ell}{\ell!}\|s_k\|^{p-\ell+1} + 2\frac{L_{h,0}L_{c,p}}{(p+1)!}\|s_k\|^{p+1}. \quad (5.3)$$

Proof. If the algorithm does not terminate at iteration $k+1$, there must exist a $j \in \{1,\dots,q\}$ such that (4.1) fails at order j at iteration $k+1$. Consider such a j and let d be the argument of the minimization in the definition of $\phi_{w,j}^{\delta_{k+1,j}}(x_{k+1})$. Then $x_k + d \in \mathcal{F}$ and $\|d\| \leq \delta_{k+1,j} \leq 1$. The definition of $\phi_{w,j}^{\delta_{k+1,j}}(x_{k+1})$ in (3.11) then gives that

$$\begin{aligned}
\varepsilon\frac{\delta_{k+1,j}^j}{j!} &< \phi_{w,j}^{\delta_{k+1,j}}(x_{k+1})\\
&= -\sum_{\ell=1}^{j}\frac{1}{\ell!}\nabla_x^\ell f(x_{k+1})[d]^\ell + h\big(c(x_{k+1})\big) - h\left(\sum_{\ell=0}^{j}\frac{1}{\ell!}\nabla_x^\ell c(x_{k+1})[d]^\ell\right)\\
&= -\sum_{\ell=1}^{j}\frac{1}{\ell!}\nabla_x^\ell f(x_{k+1})[d]^\ell + \sum_{\ell=1}^{j}\frac{1}{\ell!}\nabla_s^\ell T_{f,p}(x_k,s_k)[d]^\ell + h\big(c(x_{k+1})\big)\\
&\quad - h\big(T_{c,p}(x_k,s_k)\big) - h\left(\sum_{\ell=0}^{j}\frac{1}{\ell!}\nabla_x^\ell c(x_{k+1})[d]^\ell\right)\\
&\quad + h\left(\sum_{\ell=0}^{j}\frac{1}{\ell!}\nabla_s^\ell T_{c,p}(x_k,s_k)[d]^\ell\right) - \sum_{\ell=1}^{j}\frac{1}{\ell!}\nabla_s^\ell T_{f,p}(x_k,s_k)[d]^\ell\\
&\quad + h\big(T_{c,p}(x_k,s_k)\big) - h\left(\sum_{\ell=0}^{j}\frac{1}{\ell!}\nabla_s^\ell T_{c,p}(x_k,s_k)[d]^\ell\right)\\
&\quad - \sum_{\ell=1}^{j}\frac{\sigma_k\|s_k\|^{p-\ell+1}[d]^\ell}{\ell!(p-\ell+1)!} + \sum_{\ell=1}^{j}\frac{\sigma_k\|s_k\|^{p-\ell+1}[d]^\ell}{\ell!(p-\ell+1)!}.
\end{aligned} \quad (5.4)$$

Now, using Theorem 3.1 for $r = f$ yields

$$\begin{aligned}
&-\sum_{\ell=1}^{j}\frac{1}{\ell!}\nabla_x^\ell f(x_{k+1})[d]^\ell + \sum_{\ell=1}^{j}\frac{1}{\ell!}\nabla_s^\ell T_{f,p}(x_k,s_k)[d]^\ell\\
&\quad\leq \sum_{\ell=1}^{j}\frac{\delta_{k+1,j}^\ell}{\ell!}\left\|\nabla_x^\ell f(x_{k+1}) - \nabla_s^\ell T_{f,p}(x_k,s_k)\right\|\\
&\quad\leq L_{f,p}\sum_{\ell=1}^{j}\frac{\delta_{k+1,j}^\ell}{\ell!(p-\ell+1)!}\|s_k\|^{p-\ell+1}.
\end{aligned} \quad (5.5)$$

In the same spirit, also using AS.3 and applying Theorem 3.1 to c, we obtain

$$\begin{aligned}
&-h\left(\sum_{\ell=0}^{j}\frac{1}{\ell!}\nabla_x^\ell c(x_{k+1})[d]^\ell\right)+h\left(\sum_{\ell=0}^{j}\frac{1}{\ell!}\nabla_s^\ell T_{c,p}(x_k,s_k)[d]^\ell\right)\\
&\quad\le L_{h,0}\left\|\sum_{\ell=0}^{j}\frac{1}{\ell!}\left[\nabla_x^\ell c(x_{k+1})-\nabla_s^\ell T_{c,p}(x_k,s_k)\right][d]^\ell\right\|\\
&\quad\le L_{h,0}\sum_{\ell=0}^{j}\frac{\delta_{k+1,j}^\ell}{\ell!}\left\|\nabla_x^\ell c(x_{k+1})-\nabla_s^\ell T_{c,p}(x_k,s_k)\right\|\\
&\quad\le L_{h,0}L_{c,p}\sum_{\ell=0}^{j}\frac{\delta_{k+1,j}^\ell}{\ell!(p-\ell+1)!}\|s_k\|^{p-\ell+1}
\end{aligned}\tag{5.6}$$

and

$$h\big(c(x_{k+1})\big)-h\big(T_{c,p}(x_k,s_k)\big)\le L_{h,0}\left\|c(x_{k+1})-T_{c,p}(x_k,s_k)\right\|\le\frac{L_{h,0}L_{c,p}}{(p+1)!}\|s_k\|^{p+1}.\tag{5.7}$$

Because of Lemma 5.2 we also have that

$$\sum_{\ell=1}^{j}\frac{\sigma_k\|s_k\|^{p-\ell+1}\delta_{k+1,j}^\ell}{\ell!(p-\ell+1)!}\le\sigma_{\max}\sum_{\ell=1}^{j}\frac{\|s_k\|^{p-\ell+1}\delta_{k+1,j}^\ell}{\ell!(p-\ell+1)!}.\tag{5.8}$$

Moreover, in view of (4.2) and (4.4),

$$\begin{aligned}
&-\sum_{\ell=1}^{j}\frac{1}{\ell!}\nabla_s^\ell T_{f,p}(x_k,s_k)[d]^\ell+h\big(T_{c,p}(x_k,s_k)\big)-h\left(\sum_{\ell=0}^{j}\frac{1}{\ell!}\nabla_s^\ell T_{c,p}(x_k,s_k)[d]^\ell\right)\\
&-\sum_{\ell=1}^{j}\frac{\sigma_k}{\ell!(p-\ell+1)!}\|s_k\|^{p-\ell+1}\delta_{k+1,j}^\ell\le\phi_{m_k,j}^{\delta_{s,j}}(s_k)=\theta\varepsilon\,\frac{\delta_{k+1,j}^j}{j!},
\end{aligned}\tag{5.9}$$

where the last equality is derived using $\delta_{s,j}=\delta_{k+1,j}$ if iteration k is successful. We may now substitute (5.5)–(5.9) into (5.4) and use the inequality $(p-\ell+1)!\ge 1$ to obtain (5.3). ■

Lemma 5.4. *Suppose that AS.1–AS.3 hold, that iteration k is successful, and that the ARqpC algorithm does not terminate at iteration $k+1$. Suppose also that the algorithm ensures, for each k, that either $\delta_{k+1,j}=1$ for $j\in\{1,\dots,q\}$ if (4.11) holds (as allowed by Lemma 4.3), or that (4.21) holds (as allowed by Lemma 4.4) otherwise. Then there exists a $j\in\{1,\dots,q\}$ such that*

$$\|s_k\|\ge\begin{cases}\left(\frac{1-\theta}{3j!(L_{w,p}+\sigma_{\max})}\right)^{\frac{1}{p-j+1}}\varepsilon_j^{\frac{1}{p-j+1}} & \textit{if (4.11) holds,}\\[1ex]
\left(\frac{(1-\theta)\kappa_{\delta,\min}^{j-1}}{3j!(L_{w,p}+\sigma_{\max})}\right)^{\frac{1}{p}}\varepsilon_j^{\frac{j}{p}} & \textit{if (4.11) fails but } h=0,\\[1ex]
\left(\frac{(1-\theta)\kappa_{\delta,\min}^{j}}{3j!(L_{w,p}+\sigma_{\max})}\right)^{\frac{1}{p+1}}\varepsilon_j^{\frac{j+1}{p+1}} & \textit{if (4.11) fails and } h\neq 0,\end{cases}\tag{5.10}$$

where $\kappa_{\delta,\min}$ is defined in (4.21).

Proof. We now use our freedom to choose $\xi\in(0,1)$. Let

$$\xi\overset{\text{def}}{=}\left(\frac{1-\theta}{3q!(L_{w,p}+\sigma_{\max})}\right)^{\frac{1}{p-q+1}}=\min_{j\in\{1,\dots,q\}}\left(\frac{1-\theta}{3j!(L_{w,p}+\sigma_{\max})}\right)^{\frac{1}{p-j+1}}\in(0,1).$$

If $\|s_k\| \geq \xi$, then (5.10) clearly holds since $\varepsilon \leq 1$ and $\kappa_{\delta,\min} < 1$. We therefore assume that $\|s_k\| < \xi$. Because the algorithm has not terminated, Lemma 5.3 ensures that (5.3) holds for some $j \in \{1, \dots, q\}$. It is easy to verify that this inequality is equivalent to

$$\alpha\,\varepsilon\,\delta_{k+1,j}^j \leq \|s_k\|^{p+1}\chi_j\left(\frac{\delta_{k+1,j}}{\|s_k\|}\right) + \beta\|s_k\|^{p+1} \tag{5.11}$$

where the function χ_j is defined in (2.5) and where we have set

$$\alpha = \frac{1-\theta}{j!(L_{w,p} + \sigma_{\max})} \quad \text{and} \quad \beta = \frac{2}{(p+1)!}\,\frac{L_{h,0}L_{c,p}}{L_{w,p} + \sigma_{\max}} \in [0, 1),$$

the last inclusion resulting from the definition of $L_{w,p}$ in Lemma 5.2. In particular, since $\chi_j(t) \leq 2t^j$ for $t \geq 1$ and $\beta < 1$, we have that, when $\|s_k\| \leq \delta_{k+1,j}$,

$$\alpha\,\varepsilon \leq 2\|s_k\|^{p+1}\left(\frac{1}{\|s_k\|}\right)^j + \left(\frac{\|s_k\|}{\delta_{k+1,j}}\right)^j \|s_k\|^{p-j+1} \leq 3\|s_k\|^{p-j+1}. \tag{5.12}$$

Suppose first that (4.11) hold. Then, from our assumptions, $\delta_{k+1,j} = 1$ and $\|s_k\| \leq \xi < 1 = \delta_{k+1,j}$. Thus (5.12) yields the first case of (5.10). Suppose now that (4.11) fails. Then our assumptions imply that (4.21) holds. If $\|s_k\| \leq \delta_{k+1,j}$, we may again deduce from (5.12) that the first case of (5.10) holds, which implies, because $\kappa_{\delta,\min} < 1$, that the second and third cases also hold. Consider therefore the case where $\|s_k\| > \delta_{k+1,j}$ and suppose first that $\beta = 0$. Then (5.11) and the fact that $\chi_j(t) < 2t$ for $t \in [0, 1]$ give that

$$\alpha\varepsilon\delta_{k+1,j}^j \leq 2\|s_k\|^{p+1}\left(\frac{\delta_{k+1,j}}{\|s_k\|}\right),$$

which, with (4.21), implies the second case of (5.10). Finally, if $\beta > 0$, (5.11), the bound $\beta \leq 1$, and $\chi_j(t) < 2$ for $t \in [0, 1]$ ensure that

$$\alpha\varepsilon\delta_{k+1,j}^j \leq 2\|s_k\|^{p+1} + \|s_k\|^{p+1},$$

the third case of (5.10) then follows from (4.21). ■

Note that the proof of this lemma ensures the better lower bound given by the first case of (5.10) whenever $\|s_k\| \leq \delta_{k+1,j}$. Unfortunately, there is no guarantee that this inequality holds when (4.11) fails.

We may then derive our final evaluation complexity results. To make them clearer, we provide separate statements for the standard smooth and for the general composite cases.

Theorem 5.5 (Smooth case). *Suppose that AS.1 and AS.4 hold and that $h = 0$. Suppose also that the algorithm ensures, for each k, that either $\delta_{k+1,j} = 1$ for $j \in \{1, \dots, q\}$ if (4.11) holds (as allowed by Lemma 4.3), or that (4.21) holds (as allowed by Lemma 4.4) otherwise.*

1. *Suppose that $\mathcal{F}$ is convex and $q = 1$ or that $\mathcal{F} = \mathbb{R}^n$ and $q = 2$. Then there exist positive constants $\kappa^{s,1}_{\text{ARqp}}$, $\kappa^{a,1}_{\text{ARqp}}$, and κ^c_{ARqp} such that, for any $\varepsilon \in (0, 1]^q$, the ARqpC algorithm requires at most*

$$\kappa^{a,1}_{\text{ARqp}}\frac{w(x_0) - w_{\text{low}}}{\min_{j\in\{1,\dots,q\}} \varepsilon_j^{\frac{p+1}{p-j+1}}} + \kappa^c_{\text{ARqp}} = \mathcal{O}\left(\max_{j\in\{1,\dots,q\}} \varepsilon_j^{-\frac{p+1}{p-j+1}}\right) \tag{5.13}$$

evaluations of f and c, and at most

$$\kappa^{s,1}_{\mathrm{ARqp}}\frac{w(x_0)-w_{\mathrm{low}}}{\min_{j\in\{1,\ldots,q\}}\varepsilon_j^{\frac{p+1}{p-j+1}}}+1=\mathcal{O}\Big(\max_{j\in\{1,\ldots,q\}}\varepsilon_j^{-\frac{p+1}{p-j+1}}\Big) \tag{5.14}$$

evaluations of the derivatives of f of orders 1 to p to produce an iterate x_ε such that $\phi^1_{f,j}(x_\varepsilon)\le\varepsilon_j/j!$ for all $j\in\{1,\ldots,q\}$.

2. *Suppose that either $\mathcal{F}\subset\mathbb{R}^n$ and $q=2$, or that $\mathcal{F}$ is nonconvex or that $q>2$. Then there exist positive constants $\kappa^{s,2}_{\mathrm{ARqp}}$, $\kappa^{a,2}_{\mathrm{ARqp}}$, and $\kappa^{c}_{\mathrm{ARqp}}$ such that, for any $\varepsilon\in(0,1]^q$, the ARqpC algorithm requires at most*

$$\kappa^{a,2}_{\mathrm{ARqp}}\frac{w(x_0)-w_{\mathrm{low}}}{\min_{j\in\{1,\ldots,q\}}\varepsilon_j^{\frac{j(p+1)}{p}}}+\kappa^{c}_{\mathrm{ARqp}}=\mathcal{O}\Big(\max_{j\in\{1,\ldots,q\}}\varepsilon_j^{-\frac{j(p+1)}{p}}\Big) \tag{5.15}$$

evaluations of f and c, and at most

$$\kappa^{s,2}_{\mathrm{ARqp}}\frac{w(x_0)-w_{\mathrm{low}}}{\min_{j\in\{1,\ldots,q\}}\varepsilon_j^{\frac{j(p+1)}{p}}}+1=\mathcal{O}\Big(\max_{j\in\{1,\ldots,q\}}\varepsilon_j^{-\frac{j(p+1)}{p}}\Big) \tag{5.16}$$

evaluations of the derivatives of f of orders 1 to p to produce an iterate x_ε such that $\phi^{\delta_\varepsilon}_{f,j}(x_\varepsilon)\le\varepsilon_j\,\delta^j_{\varepsilon,j}/j!$ for some $\delta_\varepsilon\in(0,1]^q$ and all $j\in\{1,\ldots,q\}$.

Theorem 5.6 (Composite case). *Suppose that AS.1–AS.4 hold. Suppose also that the algorithm ensures, for each k, that either $\delta_{k+1,j}=1$ for $j\in\{1,\ldots,q\}$ if (4.11) holds (as allowed by Lemma 4.3), or that (4.21) holds (as allowed by Lemma 4.4) otherwise.*

1. *Suppose that $\mathcal{F}$ is convex, $q=1$, and h is convex. Then there exist positive constants $\kappa^{s,1}_{\mathrm{ARqpC}}$, $\kappa^{a,1}_{\mathrm{ARqpC}}$, and $\kappa^{c}_{\mathrm{ARqpC}}$ such that, for any $\varepsilon_1\in(0,1]$, the ARqpC algorithm requires at most*

$$\kappa^{a,1}_{\mathrm{ARqpC}}\frac{w(x_0)-w_{\mathrm{low}}}{\varepsilon_1^{\frac{p+1}{p}}}+\kappa^{c,1}_{\mathrm{ARqpC}}=\mathcal{O}\big(\varepsilon_1^{-\frac{p+1}{p}}\big) \tag{5.17}$$

evaluations of f and c, and at most

$$\kappa^{s,1}_{\mathrm{ARqpC}}\frac{w(x_0)-w_{\mathrm{low}}}{\varepsilon_1^{\frac{p+1}{p}}}+1=\mathcal{O}\big(\varepsilon_1^{-\frac{p+1}{p}}\big) \tag{5.18}$$

evaluations of the derivatives of f and c of orders 1 to p to produce an iterate x_ε such that $\phi^1_{w,j}(x_\varepsilon)\le\varepsilon_1$ for all $j\in\{1,\ldots,q\}$.

2. *Suppose that $\mathcal{F}$ is nonconvex or that h is nonconvex, or that $q>1$. Then there exist positive constants $\kappa^{s,2}_{\mathrm{ARqp}}$, $\kappa^{a,2}_{\mathrm{ARqp}}$, and $\kappa^{c}_{\mathrm{ARqp}}$ such that, for any $\varepsilon\in(0,1]^q$, the ARqpC algorithm requires at most*

$$\kappa^{a,2}_{\mathrm{ARqpC}}\frac{w(x_0)-w_{\mathrm{low}}}{\min_{j\in\{1,\ldots,q\}}\varepsilon_j^{j+1}}+\kappa^{c}_{\mathrm{ARqpC}}=\mathcal{O}\Big(\max_{j\in\{1,\ldots,q\}}\varepsilon_j^{-(j+1)}\Big) \tag{5.19}$$

evaluations of f and c, and at most

$$\kappa^{s,2}_{\mathrm{ARqpC}}\frac{w(x_0)-w_{\mathrm{low}}}{\min_{j\in\{1,\ldots,q\}}\varepsilon_j^{j+1}}+1=\mathcal{O}\Big(\max_{j\in\{1,\ldots,q\}}\varepsilon_j^{-(j+1)}\Big) \tag{5.20}$$

evaluations of the derivatives of f and c of orders 1 to p to produce an iterate x_ε such that $\phi^{\delta_\varepsilon}_{w,j}(x_\varepsilon) \le \varepsilon_j\, \delta^j_{\varepsilon,j}/j!$ for some $\delta_\varepsilon \in (0,1]^q$ and all $j \in \{1,\dots,q\}$.

Proof. We prove Theorems 5.5 and 5.6 together. At each successful iteration k of the AR$qp$C algorithm before termination, we have the guaranteed decrease

$$w(x_k) - w(x_{k+1}) \ge \eta_1\big(T_{w,p}(x_k,0) - T_{w,p}(x_k,s_k)\big) \ge \frac{\eta_1\sigma_{\min}}{(p+1)!}\,\|s_k\|^{p+1} \tag{5.21}$$

where we used (5.1) and (4.7). We now wish to substitute the bounds given by Lemma 5.4 in (5.21), and deduce that, for some $j \in \{1,\dots,q\}$,

$$w(x_k) - w(x_{k+1}) \ge \kappa^{-1}\varepsilon_j^\omega \tag{5.22}$$

where the definition of κ and ω depends on q and h. Specifically,

$$\kappa \overset{\text{def}}{=} \begin{cases} \kappa^{s,1}_{\text{ARqp}} = \kappa^{s,1}_{\text{ARqpC}} \overset{\text{def}}{=} \frac{(p+1)!}{\eta_1\sigma_{\min}}\left(\frac{1-\theta}{3j!(L_{w,p}+\sigma_{\max})}\right)^{-\frac{p+1}{p-j+1}} \\ \quad \text{if } (q=1, h \text{ and } \mathcal{F} \text{ are convex}),\ \text{and} \\ \quad \text{if } (q\in\{1,2\}, \mathcal{F} \text{ is convex and } h=0), \\ \kappa^{s,2}_{\text{ARqp}} \overset{\text{def}}{=} \frac{(p+1)!}{\eta_1\sigma_{\min}}\left(\frac{(1-\theta)\kappa^{j-1}_{\delta,\min}}{3j!(L_{w,p}+\sigma_{\max})}\right)^{-\frac{p+1}{p}} \\ \quad \text{if } h=0 \text{ and} \\ \quad ((q=2 \text{ and } \mathcal{F}\subset\mathbb{R}^n) \text{ or } q>2 \text{ or } \mathcal{F} \text{ is nonconvex}) \\ \kappa^{s,2}_{\text{ARqpC}} \overset{\text{def}}{=} \frac{(p+1)!}{\eta_1\sigma_{\min}}\left(\frac{(1-\theta)\kappa^{j}_{\delta,\min}}{3j!(L_{w,p}+\sigma_{\max})}\right)^{-1} \\ \quad \text{if } h\ne 0 \text{ and } (q>1 \text{ or } \mathcal{F} \text{ is nonconvex}), \end{cases}$$

where $\kappa_{\delta,\min}$ is given by (4.21), and

$$\omega \overset{\text{def}}{=} \begin{cases} \frac{p+1}{p-q+1} & \text{if } (q=1, h \text{ and } \mathcal{F} \text{ are convex}), \text{and} \\ & \text{if } (q=2, \mathcal{F}=\mathbb{R}^n \text{ and } h=0), \\ \frac{q(p+1)}{p} & \text{if } h=0 \text{ and} \\ & ((q=2 \text{ and } \mathcal{F}\subset\mathbb{R}^n) \text{ or } q>2 \text{ or } \mathcal{F} \text{ is nonconvex}) \\ q+1 & \text{if } h\ne 0 \text{ and } (q>1 \text{ or } \mathcal{F} \text{ is nonconvex}). \end{cases} \tag{5.23}$$

Thus, since $\{w(x_k)\}$ decreases monotonically,

$$w(x_0) - w(x_{k+1}) \ge \kappa^{-1} \min_{j\in\{1,\dots,q\}} \varepsilon_j^\omega\, |\mathcal{S}_k|.$$

Using AS.4, we conclude that

$$|\mathcal{S}_k| \le \kappa\, \frac{w(x_0) - w_{\text{low}}}{\min_{j\in\{1,\dots,q\}} \varepsilon_j^\omega} \tag{5.24}$$

until termination, bounding the number of successful iterations. Lemma 4.1 is then invoked to compute the upper bound on the total number of iterations, yielding the constants

$$\kappa^{a,1}_{\text{ARqp}} \overset{\text{def}}{=} \kappa^{s,1}_{\text{ARqp}}\left(1+\frac{|\log\gamma_1|}{\log\gamma_2}\right), \quad \kappa^{a,2}_{\text{ARqp}} \overset{\text{def}}{=} \kappa^{s,2}_{\text{ARqp}}\left(1+\frac{|\log\gamma_1|}{\log\gamma_2}\right),$$
$$\kappa^{a,1}_{\text{ARqpC}} \overset{\text{def}}{=} \kappa^{s,1}_{\text{ARqpC}}\left(1+\frac{|\log\gamma_1|}{\log\gamma_2}\right), \quad \kappa^{a,2}_{\text{ARqpC}} \overset{\text{def}}{=} \kappa^{s,2}_{\text{ARqpC}}\left(1+\frac{|\log\gamma_1|}{\log\gamma_2}\right),$$

and

$$\kappa^c_{\mathsf{ARqp}} = \kappa^c_{\mathsf{ARqpC}} \overset{\text{def}}{=} \frac{1}{\log \gamma_2} \log\left(\frac{\sigma_{\max}}{\sigma_0}\right),$$

where $\sigma_{\max} = \max[\sigma_0, \frac{\gamma_3 L_{w,p}}{1-\eta_2}]$ (see (5.2)). The desired conclusions then follow from the fact that each iteration involves one evaluation of f and each successful iteration one evaluation of its derivatives. ■

For the standard smooth case, Theorem 5.5 provides the first results on the complexity of finding *strong* minimizers of arbitrary orders using adaptive regularization algorithms that we are aware of. By comparison, [12] provides similar results but for the convergence to *weak* minimizers (see (2.5)). Unsurprisingly, the worst-case complexity bounds for weak minimizers are better than those for strong ones: the $\mathcal{O}(\epsilon^{-(p+1)/(p-q+1)})$ bound which we have derived for $q \in \{1,2\}$ then extends to any order q. Moreover, the full power of AS.1 is not needed for these results since it is sufficient to assume that $\nabla_x^p f(x)$ is Lipschitz continuous. It is interesting to note that the results for weak and strong approximate minimizers coincide for first and second order. The results of Theorem 5.5 may also be compared with the bound in $\mathcal{O}(\epsilon^{-(q+1)})$ which was proved for trust-region methods in [11]. While these trust-region bounds do not depend on the degree of the model, those derived above for the ARqpC algorithm show that worst-case performance improves with p and is always better than that of trust-region methods. It is also interesting to note that the bound obtained in Theorem 5.5 for order q is identical to that which would be obtained for first-order but using ϵ^q instead of ϵ. This reflects the observation that, different from the weak approximate optimality, the very definition of strong approximate optimality in (2.4) requires very high accuracy on the (usually dominant) low order terms of the Taylor series while the requirement lessens as the order increases.

An interesting feature of the algorithm discussed in [12] is that computing and testing the value of $\phi^{\delta}_{m_k,j}(s_k)$ is unnecessary if the length of the step is large enough. The same feature can easily be introduced into the ARqpC algorithm. Specifically, we may redefine Step 2 to accept a step as soon as (4.3) holds and

$$\|s_k\| \geq \begin{cases} \varpi \min_{j\in\{1,\ldots,q\}} \epsilon_j^{\frac{1}{p-q+1}} & \text{if } (q = 1, h \text{ and } \mathcal{F} \text{ are convex}), \text{and} \\ & \text{if } (q = 2, \mathcal{F} = \mathbb{R}^n \text{ and } h = 0), \\ \varpi \min_{j\in\{1,\ldots,q\}} \epsilon_j^{\frac{q}{p}} & \text{if } h = 0 \text{ and} \\ & ((q = 2 \text{ and } \mathcal{F} \subset \mathbb{R}^n) \text{ or } q > 2 \text{ or } \mathcal{F} \text{ is nonconvex}), \\ \varpi \min_{j\in\{1,\ldots,q\}} \epsilon_j^{\frac{q+1}{p+1}} & \text{if } h \neq 0 \text{ and } (q > 1 \text{ or } \mathcal{F} \text{ is nonconvex}), \end{cases}$$

for some $\varpi \in (\theta, 1]$. If these conditions fail, then one still needs to verify the requirements (4.3) and (4.4), as we have done previously. Given Lemma 5.1 and the proof of Theorems 5.5 and 5.6, it is easy to verify that this modification does not affect the conclusions of these complexity theorems, while potentially avoiding significant computations.

Existing complexity results for (possibly nonsmooth) composite problems are few [8, 13, 14, 20]. Theorem 5.6 provides, to the best of our knowledge, the first upper complexity bounds for optimality orders exceeding one, with the exception of [13] (but this paper requires strong specific assumptions on $\mathcal{F}$). While equivalent to those of Theorem 5.5 for the standard case when $q = 1$, they are not as good and match those obtained for the trust-region methods when $q > 1$. They could be made identical in order of ε_j to those of Theorem 5.5 if one is ready to assume that $L_{h,0}L_{c,p}$ is sufficiently small (for instance, if c is a polynomial of degree less than p). In this case, the constant β in Lemma 5.11 will of the order of $\delta_{k+1,j}/\|s_k\|$, leading to the better bound.

6. SHARPNESS

We now show that the upper complexity bounds in Theorem 5.5 and the first part of Theorem 5.6 are sharp. Since it is sufficient for our purposes, we assume in this section that $\varepsilon_j = \varepsilon$ for all $j \in \{1, \dots, q\}$.

We first consider a first class of problems, where the choice of $\delta_{k,j} = 1$ is allowed. Since it is proved in [12] that the order in ε given by the Theorem 5.5 is sharp for finding weak approximate minimizers for the standard (smooth) case, it is not surprising that this order is also sharp for the stronger concept of optimality whenever the same bound applies, that is when $q \in \{1, 2\}$. However, the ARqpC algorithm slightly differs from the algorithm discussed in [12]. Not only are the termination tests for the algorithm itself and those for the step computation weaker in [12], but the algorithm there makes a provision to avoid computing $\phi^{\delta}_{m_k,j}$ whenever the step is large enough, as discussed at the end of the last section. It is thus impossible to use the example of slow convergence provided in [12, SECTION 5.2] directly, but we now propose a variant that fits our present framework.

Theorem 6.1. *Suppose that $h = 0$ and that the choice $\delta_{k,j} = 1$ is possible (and made) for all k and all $j \in \{1, \dots, q\}$. Then the ARqpC algorithm applied to minimize f may require*

$$\varepsilon^{-\frac{p+1}{p-q+1}}$$

iterations and evaluations of f and of its derivatives of order 1 up to p to produce a point x_ε such that $\phi^1_{w,q}(x_\varepsilon) \leq \varepsilon/j!$ for all $j \in \{1, \dots, q\}$.

Proof. Our aim is to show that, for each choice of $p \geq 1$, there exists an objective function satisfying AS.1 and AS.4 such that obtaining a strong (ε, δ)-approximate qth-order-necessary minimizer may require at least $\varepsilon^{-(p+1)/(p-q+1)}$ evaluations of the objective function and its derivatives using the AR$qp$C algorithm. Also note that, in this context, $\phi^{\delta_j}_{w,q}(x) = \phi^{\delta_j}_{f,q}(x)$ and (4.1) reduces to (2.4).

Given a model degree $p \geq 1$ and an optimality order q, we also define the sequences $\{f^{(j)}_k\}$ for $j \in \{0, \dots, p\}$ and $k \in \{0, \dots, k_\varepsilon\}$ by

$$k_\varepsilon = \left\lceil \varepsilon^{-\frac{p+1}{p-q+1}} \right\rceil \tag{6.1}$$

and

$$\omega_k = \varepsilon \frac{k_\varepsilon - k}{k_\varepsilon} \in [0, \varepsilon], \tag{6.2}$$

as well as

$$f_k^{(j)} = 0 \quad \text{for } j \in \{1,\dots,q-1\} \cup \{q+1,\dots,p\} \quad \text{and} \quad f_k^{(q)} = -(\varepsilon + \omega_k) < 0.$$

Thus

$$T_{f,p}(x_k, s) = \sum_{j=0}^{p} \frac{f_k^{(j)}}{j!} s^j = f_k^{(0)} - (\varepsilon + \omega_k)\frac{s^q}{q!}. \tag{6.3}$$

We also set $\sigma_k = p!/(q-1)!$ for all $k \in \{0,\dots,k_\varepsilon\}$ (we verify below that is acceptable). It is easy to verify using (6.3) that the model (4.2) is then globally minimized for

$$s_k = \left|f_k^{(q)}\right|^{\frac{1}{p-q+1}} = [\varepsilon + \omega_k]^{\frac{1}{p-q+1}} > \varepsilon^{\frac{1}{p-q+1}} \quad \big(k \in \{0,\dots,k_\varepsilon\}\big). \tag{6.4}$$

We then assume that Step 2 of the ARqpC algorithm returns, for all $k \in \{0,\dots,k_\varepsilon\}$, the step s_k given by (6.4) and the optimality radius $\delta_{k,j} = 1$ for $j \in \{1,\dots,q\}$ (as allowed by our assumption). Thus implies that

$$\phi_{f,q}^{\delta_{k,q}}(x_k) = (\varepsilon + \omega_k)\frac{\delta_{k,q}^q}{q!}, \tag{6.5}$$

and therefore that

$$\omega_k \in (0,\varepsilon], \quad \phi_{f,j}^{\delta_{k,j}}(x_k) = 0 \quad (j = 1,\dots,q-1) \quad \text{and} \quad \phi_{f,q}^{\delta_{k,q}}(x_k) > \varepsilon\frac{\delta_{k,q}^q}{q!} \tag{6.6}$$

(and (2.4) fails at x_k) for $k \in \{0,\dots,k_\varepsilon - 1\}$, while

$$\omega_{k_\varepsilon} = 0, \quad \phi_{f,j}^{\delta_{k,j}}(x_{k_\varepsilon}) = 0 \quad (j = 1,\dots,q-1) \quad \text{and} \quad \phi_{f,q}^{\delta_{k,q}}(x_{k_\varepsilon}) = \varepsilon\frac{\delta_{k,q}^q}{q!} \tag{6.7}$$

(and (2.4) holds at x_{k_ε}). The step (6.4) yields that

$$\begin{aligned} m_k(s_k) &= f_k^{(0)} - \frac{\varepsilon + \omega_k}{q!}[\varepsilon + \omega_k]^{\frac{q}{p-q+1}} + \frac{\sigma_k}{(p+1)!}[\varepsilon + \omega_k]^{\frac{p+1}{p-q+1}} \\ &= f_k^{(0)} - \frac{\varepsilon + \omega_k}{q!}[\varepsilon + \omega_k]^{\frac{q}{p-q+1}} + \frac{1}{(p+1)(q-1)!}[\varepsilon + \omega_k]^{\frac{p+1}{p-q+1}} \\ &= f_k^{(0)} - \zeta(q,p)[\varepsilon + \omega_k]^{\frac{p+1}{p-q+1}} \end{aligned} \tag{6.8}$$

where

$$\zeta(q,p) \stackrel{\text{def}}{=} \frac{p-q+1}{(p+1)q!} \in (0,1). \tag{6.9}$$

Thus $m_k(s_k) < m_k(0)$ and (4.3) holds. We then define

$$f_0^{(0)} = 2^{1+\frac{p+1}{p-q+1}} \quad \text{and} \quad f_{k+1}^{(0)} = f_k^{(0)} - \zeta(q,p)[\varepsilon + \omega_k]^{\frac{p+1}{p-q+1}}, \tag{6.10}$$

which provides the identity

$$m_k(s_k) = f_{k+1}^{(0)} \tag{6.11}$$

(ensuring that iteration k is successful because $\rho_k = 1$ in (4.6) and thus that our choice of a constant σ_k is acceptable). In addition, using (6.2), (6.10), (6.6), (6.9) and the inequality $k_\varepsilon \le 1 + \varepsilon^{-\frac{p+1}{p-q+1}}$ resulting from (6.1), gives that, for $k \in \{0,\dots,k_\varepsilon\}$,

$$\begin{aligned} f_0^{(0)} \ge f_k^{(0)} &\ge f_0^{(0)} - k\zeta(q,p)[2\varepsilon]^{\frac{p+1}{p-q+1}} > f_0^{(0)} - k_\varepsilon \varepsilon^{\frac{p+1}{p-q+1}} 2^{\frac{p+1}{p-q+1}} \\ &\ge f_0^{(0)} - \left(1 + \varepsilon^{\frac{p+1}{p-q+1}}\right) 2^{\frac{p+1}{p-q+1}} \ge f_0^{(0)} - 2^{1+\frac{p+1}{p-q+1}}, \end{aligned}$$

and hence that

$$f_k^{(0)} \in (0, 2^{1+\frac{p+1}{p-q+1}}] \quad \text{for } k \in \{0, \dots, k_\varepsilon\}. \tag{6.12}$$

We also set

$$x_0 = 0 \quad \text{and} \quad x_k = \sum_{i=0}^{k-1} s_i.$$

Then (6.11) and (4.2) give that

$$\left| f_{k+1}^{(0)} - T_{f,p}(x_k, s_k) \right| = \frac{1}{(p+1)(q-1)!} |s_k|^{p+1} \leq |s_k|^{p+1}. \tag{6.13}$$

Now note that, using (6.3) and the first equality in (6.4),

$$T_{f,p}^{(j)}(x_k, s_k) = \frac{f_k^{(q)}}{(q-j)!} s_k^{q-j} \delta_{[j \leq q]} = -\frac{1}{(q-j)!} s_k^{p-j+1} \delta_{[j \leq q]},$$

where $\delta_{[\cdot]}$ is the standard indicator function. We now see that, for $j \in \{1, \dots, q-1\}$,

$$\left| f_{k+1}^{(j)} - T_{f,p}^{(j)}(x_k, s_k) \right| = \left| 0 - T_{f,p}^{(j)}(x_k, s_k) \right| \leq \frac{1}{(q-j)!} |s_k|^{p-j+1} \leq |s_k|^{p-j+1}, \tag{6.14}$$

while, for $j = q$, we have that

$$\left| f_{k+1}^{(q)} - T_{f,p}^{(q)}(x_k, s_k) \right| = \left| -s_k^{p-q+1} + s_k^{p-q+1} \right| = 0 \tag{6.15}$$

and, for $j \in \{q+1, \dots, p\}$,

$$\left| f_{k+1}^{(j)} - T_{f,p}^{(j)}(x_k, s_k) \right| = |0 - 0| = 0. \tag{6.16}$$

Combining (6.13)–(6.16), we may then apply classical Hermite interpolation (see [12, THEOREM 5.2] with $\kappa_f = 1$), and deduce the existence of a p times continuously differentiable function f_{ARqpC} from $\mathbb{R}$ to $\mathbb{R}$ with Lipschitz continuous derivatives of order 0 to p (hence satisfying AS.1) which interpolates $\{f_k^{(j)}\}$ at $\{x_k\}$ for $k \in \{0, \dots, k_\varepsilon\}$ and $j \in \{0, \dots, p\}$. Moreover, (6.12), (6.3), (6.4), and the same Hermite interpolation theorem imply that $|f^{(j)}(x)|$ is bounded by a constant only depending on p and q, for all $x \in \mathbb{R}$ and $j \in \{0, \dots, p\}$ (and thus AS.1 holds) and that f_{ARqpC} is bounded below (ensuring AS.4.) and that its range only depends on p and q. This concludes our proof. ■

This immediately provides the following important corollary.

Corollary 6.2. *Suppose that $h = 0$ and that either $q = 1$ and $\mathcal{F}$ is convex, or $q = 2$ and $\mathcal{F} = \mathbb{R}^n$. Then the ARqpC algorithm applied to minimize f may require*

$$\varepsilon^{-\frac{p+1}{p-q+1}}$$

iterations and evaluations of f and of its derivatives of order 1 up to p to produce a point x_ε such that $\phi_{w,q}^1(x_\varepsilon) \leq \varepsilon/j!$ for all $j \in \{1, \dots, q\}$.

Proof. We start by noting that, in both cases covered by our assumptions, Lemma 4.3 allows the choice $\delta_{k,j} = 1$ for all k and all $j \in \{1, \dots, q\}$. We conclude by applying Theorem 6.1. ■

It is then possible to derive a lower complexity bound for the simple composite case where h is nonzero but convex and $q = 1$.

Corollary 6.3. *Suppose that $q = 1$ and that h is convex. Then the ARqpC algorithm applied to minimize w may require*

$$\varepsilon^{-\frac{p+1}{p}}$$

iterations and evaluations of f and c and of their derivatives of order 1 up to p to produce a point x_ε such that $\phi^1_{w,1}(x_\varepsilon) \leq \varepsilon$.

Proof. It is enough to consider the unconstrained problem where $w = h(c(x))$ with $h(x) = |x|$ and c is the positive function f constructed in the proof of Theorem 6.1. ■

We now turn to the high-order smooth case.

Theorem 6.4. *Suppose that $h = 0$ and that either $q > 2$, or $q = 2$ and $\mathcal{F} = \mathbb{R}^n$. If $\varepsilon \in (0, 1)$ is sufficiently small and if the ARqpC algorithm applied to minimize f allows the choice of an arbitrary $\delta_{k,j} > 0$ satisfying (4.21), it may then require*

$$\varepsilon^{-\frac{q(p+1)}{p}}$$

iterations and evaluations of f and of its derivatives of order 1 up to p to produce a point x_ε such that $\phi^{\delta_{\varepsilon,j}}_{f,j}(x_\varepsilon) \leq \varepsilon\delta^j_{\varepsilon,j}/j!$ for all $j \in \{1, \dots, q\}$ and some $\delta_\varepsilon \in (0, 1]^q$.

Proof. As this is sufficient, we focus on the case where $\mathcal{F} = \mathbb{R}^n$. Our aim is now to show that, for each choice of $p \geq 1$ and $q > 2$, there exists an objective function satisfying AS.1 and AS.4 such that obtaining a strong (ε, δ)-approximate qth-order-necessary minimizer may require at least $\varepsilon^{-q(p+1)/p}$ evaluations of the objective function and its derivatives using the AR$qp$C algorithm. As in Theorem 6.1, we have to construct f such that it satisfies AS.1 and is globally bounded below, which then ensures AS.4. Again, we note that, in this context, $\phi^{\delta_j}_{f,q}(x) = \phi^{\delta_j}_{f,q}(x)$ and (4.1) reduces to (2.4).

Without loss of generality, we assume that $\varepsilon \leq \frac{1}{2}$. Given a model degree $p \geq 1$ and an optimality order $q > 2$, we set

$$k_\varepsilon = \left\lceil \varepsilon^{-\frac{q(p+1)}{p}} \right\rceil \tag{6.17}$$

and

$$\omega_k = \varepsilon^q \, \frac{k_\varepsilon - k}{k_\varepsilon} \in \left[0, \varepsilon^q\right] \quad \big(k \in \{0, \dots, k_\varepsilon\}\big). \tag{6.18}$$

Moreover, for $j \in \{0, \dots, p\}$ and each $k \in \{0, \dots, k_\varepsilon\}$, we define the sequences $\{f^{(j)}_k\}$ by

$$f^{(1)}_k = -\frac{\varepsilon^q + \omega_k}{q!} < 0 \quad \text{and} \quad f^{(j)}_k = 0 \quad \text{for } j \in \{2, \dots, p\}, \tag{6.19}$$

and therefore

$$T_{f,p}(x_k, s) = \sum_{j=0}^{p} \frac{f^{(j)}_k}{j!} s^j = f^{(0)}_k - \frac{\varepsilon^q + \omega_k}{q!} s. \tag{6.20}$$

This definition and the choice $\sigma_k = p!$ ($k \in \{0, \dots, k_\varepsilon\}$) (we verify below that this is acceptable) then allow us to define the model (4.2) by

$$m_k(s) = f_k^{(0)} - \frac{\varepsilon^q + \omega_k}{q!} s + \frac{|s|^{p+1}}{p+1}. \tag{6.21}$$

We now assume that, for each k, Step 2 returns the model's global minimizer

$$s_k = \left[\frac{\varepsilon^q + \omega_k}{q!}\right]^{\frac{1}{p}} \quad \bigl(k \in \{0, \dots, k_\varepsilon\}\bigr) \tag{6.22}$$

and the optimality radius

$$\delta_{k,j} = \varepsilon \quad \bigl(j \in \{1, \dots, q\}\bigr). \tag{6.23}$$

Indeed, a simple calculation shows that we may choose $\delta_{k,j}$ at least as large as

$$\delta_{k,j} = \frac{3|s_k|}{p-1} = \frac{3}{p-1}\left[\frac{\varepsilon^q + \omega_k}{q!}\right]^{\frac{1}{p}}. \tag{6.24}$$

which is clearly the case for (6.23) under our assumption on ε. Let us show that the above choice (6.24) is correct. Consider the model (6.21) and let $\beta = (\varepsilon^q + \omega_k)/q!$. We may then compute $T_{m_k,j}(s_k, \alpha s_k)$ the jth degree Taylor expansion of this model at s_k for $j \in \{1, \dots, q\}$. Since $\nabla_s^1 m_k(s_k) = 0$, we obtain from Lemma 4.2 that

$$\begin{aligned}
T_{m_k,j}(s_k, \alpha s_k) = \sum_{\ell=0}^{j} \frac{\nabla_s^\ell m_k(s_k)[\alpha s_k]^\ell}{\ell!} &= m_k(s_k) + \frac{\sigma_k}{(p+1)!} \sum_{\ell=2}^{j} \frac{\nabla_s^\ell(|s_k^*|^{p+1})[\alpha s_k]^\ell}{\ell!} \\
&= m_k(s_k) + \frac{\sigma_k}{(p+1)!} \sum_{\ell=2}^{j} \frac{\alpha^\ell \nabla_s^\ell(|s_k|^{p+1})[s_k]^\ell}{\ell!} \\
&= m_k(s_k) + \sigma_k \sum_{\ell=2}^{j} \frac{\alpha^\ell |s_k|^{p+1}}{\ell!(p+1-\ell)!}.
\end{aligned}$$

Clearly, $T_{m_k,1}(s_k, \alpha s_k) = m_k(s_k)$ for all α because the standard first-order optimality condition at s_k gives that $\nabla_d^1 T_{m_k,1}(s_k, 0) = 0$. The second-order optimality condition implies that $T_{m_k,2}(s_k, \alpha s_k)$ is convex in α, but, given that α can be negative, approximations of degree larger than 2 are no longer convex for odd values of j. We are now interested in computing an upper bound on $\delta_{s_k,j}$ so that (4.4) holds and for odd j (and thus for all j). Consider the case where $j = 3$: choosing $\beta = 1$ (and thus $s_k^* = e_1$) as above, (4.4) then requires that, for all $|\alpha s_k| \le \delta_{s_k,3}$,

$$T_{m_k,3}(s_k, 0) - T_{m_k,3}(s_k^*, \alpha s_k) < \theta\varepsilon \frac{|\alpha s_k|^3}{6},$$

which is obviously satisfied for any $\delta_{s_k,3}$ smaller or equal to absolute value of the root $\alpha_{*,3}$ of the equation $T_{m_k,3}(s_k, 0) = T_{m_k,3}(s_k^*, \alpha s_k)$. Using the expression of $T_{m_k,3}(s_k^*, \alpha s_k)$ derived above, one verifies that

$$\alpha_{*,3} = -\frac{3p|s_k|^{p+1}}{p(p-1)|s_k|^{p+1}} = -\frac{3}{p-1}.$$

The cases $j = 5, 9, \dots, p$ are less restrictive because the corresponding roots $\alpha_{*,j}$ are all smaller than $\alpha_{*,3}$. As a consequence, (4.4) holds for $j \in \{1, \dots, q\}$ and $\delta_{k,j} = \frac{3|s_k|}{p-1}$.

Thus, from (6.24), (6.20) and (6.23),

$$\phi_{f,j}^{\delta_{k,j}}(x_k) = (\varepsilon^q + \omega_k)\frac{\varepsilon}{q!}$$

for $j \in \{1, \ldots, q\}$ and $k \in \{0, \ldots, k_\varepsilon\}$. Using (6.23), (6.17), and the fact that, for $j \in \{1, \ldots, q-1\}$,

$$\frac{\varepsilon^q + \omega_k}{q!} \le \frac{2\varepsilon^q}{q!} \le \frac{\varepsilon^j}{j!} = \frac{\delta_{k,j}^j}{j!} \tag{6.25}$$

when $q \ge 2$ and $\varepsilon \le \frac{1}{2}$, we then obtain that

$$\phi_{f,j}^{\delta_{k,j}}(x_k) \le \varepsilon\frac{\delta_{k,j}^j}{j!} \quad (j = 1 \ldots, q-1) \quad \text{and} \quad \phi_{f,q}^{\delta_{k,q}}(x_k) > \varepsilon\frac{\delta_{k,q}^q}{q!}$$

(and (2.4) fails at x_k) for $k \in \{0, \ldots, k_\varepsilon - 1\}$, while

$$\phi_{f,j}^{\delta_{k,j}}(x_{k_\varepsilon}) < \varepsilon\frac{\delta_{k,j}^j}{j!} \quad (j = 1 \ldots, q-1) \quad \text{and} \quad \phi_{f,q}^{\delta_{k,q}}(x_{k_\varepsilon}) = \varepsilon\frac{\delta_{k,q}^q}{q!}$$

(and (2.4) holds at x_{k_ε}). Now (6.21) and (6.22) give that

$$\begin{aligned} m_k(s_k) &= f_k^{(0)} - \frac{\varepsilon^q + \omega_k}{q!}\left[\frac{\varepsilon^q + \omega_k}{q!}\right]^{\frac{1}{p}} + \frac{1}{p+1}\left[\frac{\varepsilon^q + \omega_k}{q!}\right]^{\frac{p+1}{p}} \\ &= f_k^{(0)} - \frac{p}{p+1}\left[\frac{\varepsilon^q + \omega_k}{q!}\right]^{\frac{p+1}{p}}. \end{aligned}$$

Thus $m_k(s_k) < m_k(0)$ and (4.3) holds. We then define

$$f_0^{(0)} = 2^{1+\frac{q(p+1)}{p}} \quad \text{and} \quad f_{k+1}^{(0)} = f_k^{(0)} - \frac{p}{p+1}\left[\frac{\varepsilon^q + \omega_k}{q!}\right]^{\frac{p+1}{p}}, \tag{6.26}$$

which provides the identity

$$m_k(s_k) = f_{k+1}^{(0)} \tag{6.27}$$

(ensuring that iteration k is successful because $\rho_k = 1$ in (4.6) and thus that our choice of a constant σ_k is acceptable). In addition, using (6.18), (6.26), and the inequality $k_\varepsilon \le 1 + \varepsilon^{-q(p+1)/p}$ resulting from (6.17), (6.26) gives that, for $k \in \{0, \ldots, k_\varepsilon\}$,

$$\begin{aligned} f_0^{(0)} \ge f_k^{(0)} \ge f_0^{(0)} - k[2\varepsilon]^{\frac{q(p+1)}{p}} &\ge f_0^{(0)} - k_\varepsilon \varepsilon^{\frac{q(p+1)}{p}} 2^{\frac{q(p+1)}{p}} \\ &\ge f_0^{(0)} - \left(1 + \varepsilon^{\frac{q(p+1)}{p}}\right)2^{\frac{q(p+1)}{p}} \ge f_0^{(0)} - 2^{1+\frac{q(p+1)}{p}}, \end{aligned}$$

and hence that

$$f_k^{(0)} \in \left[0, 2^{1+\frac{q(p+1)}{p}}\right] \quad \text{for } k \in \{0, \ldots, k_\varepsilon\}. \tag{6.28}$$

As in Theorem 6.1, we set $x_0 = 0$ and $x_k = \sum_{i=0}^{k-1} s_i$. Then (6.11) and (4.2) give that

$$\left|f_{k+1}^{(0)} - T_{f,p}(x_k, s_k)\right| = \frac{1}{p}|s_k|^{p+1}. \tag{6.29}$$

Using (6.20), we also see that

$$\left|f_{k+1}^{(1)} - T_{f,p}^{(1)}(x_k, s_k)\right| = \left|-\frac{(\varepsilon^q + \omega_{k+1})}{q!} + \frac{(\varepsilon^q + \omega_k)}{q!}\right| \le |s_k|^p\left[1 - \frac{\varepsilon^q + \omega_{k+1}}{\varepsilon^q + \omega_k}\right] < |s_k|^p, \tag{6.30}$$

while, for $j \in \{2, \dots, p\}$,

$$\left| f_{k+1}^{(j)} - T_{f,p}^{(j)}(x_k, s_k) \right| = |0 - 0| < |s_k|^{p-j+1}. \tag{6.31}$$

The proof is concluded as in Theorem 6.1. Combining (6.29)–(6.31), we may then apply classical Hermite interpolation (see [12, THEOREM 5.2] with $\kappa_f = 1$) and deduce the existence of a p-times continuously differentiable function f_{ARqpC} from $\mathbb{R}$ to $\mathbb{R}$ with Lipschitz continuous derivatives of order 0 to p (hence satisfying AS.1) which interpolates $\{f_k^{(j)}\}$ at $\{x_k\}$ for $k \in \{0, \dots, k_\varepsilon\}$ and $j \in \{0, \dots, p\}$. Moreover, the Hermite theorem, (6.19), and (6.22) also guarantee that $|f^{(j)}(x)|$ is bounded by a constant only depending on p and q, for all $x \in \mathbb{R}$ and $j \in \{0, \dots, p\}$. As a consequence, AS.1, AS.2, and AS.4 hold. This concludes the proof. ■

Whether the bound (5.20) is sharp remains an open question at present.

7. INEXACT GLOBAL MINIMIZATION

We finally discuss the necessity of performing global minimization when calculating the (objective and model) optimality measures and, when relevant, the effect of performing such computations inexactly. We start by recalling that such minimization problems potentially occur in two parts of the algorithm: in Step 1 (for deciding termination) and in Step 2 (during the step computation).

Step computation. Consider the step computation first and remember that the ultimate purpose of Step 2 is to find a step s_k guaranteeing a sufficient decrease of the Taylor series at x_k, in that

$$T_{w,j}(x_k, 0) - T_{w,j}(x_k, s_k) \geq \kappa_{\text{decr}} \varepsilon_j^{\omega} \tag{7.1}$$

for some fixed $\kappa_{\text{decr}} > 0$ and $j \in \{1, \dots, q\}$, where ω is defined in (5.23) (this argument is used in the proof of Theorems 5.5 and 5.6). Of course, if a step s_k that satisfies (7.1) for some given κ_{decr} can be found simply,[4] without resorting to global optimization, so much the better (and we may then choose $\delta_{k,j} = 1$ for $j \in \{1, \dots, q\}$). In other cases, the decrease guarantee (7.1) is obtained in one of two possible ways: if $\|s_k\| \geq 1$ and given that $\varepsilon_j \in (0, 1]$, sufficient decrease follows from Lemma 5.1 with $\kappa_{\text{decr}} = \sigma_{\min}/(p+1)!$. Alternatively, if $\|s_k\| \leq 1$, we then have to enforce (4.4) for some $\delta_{s,j} \in (0, 1]$, and use the more complicated Lemma 5.4 to reach the desired conclusion. In our development, the constant 1 in the inequality $\|s_k\| \geq 1$ was chosen solely for simplicity of exposition, but can be replaced by any constant independent of k. In particular, it can be replaced by $\kappa_{\text{decr}}^{1/(p+1)} \varepsilon_{\min}^{\omega/(p+1)}$ where $\varepsilon_{\min} = \min_{j \in \{1,\dots,q\}} \varepsilon_j$, so that sufficient decrease still immediately follows from Lemma 5.1 if

$$\|s_k\| \geq \kappa_{\text{decr}}^{1/(p+1)} \varepsilon_{\min}^{\omega/(p+1)}. \tag{7.2}$$

As a consequence, we see that performing any global optimization in Step 2 is only necessary whenever a descent step cannot be found that satisfies either (7.1) or (7.2). From a practical point of view, the failure of these two conditions could be considered as a reasonable ter-

[4] Say, by applying some trusted local minimization method.

mination rule for small enough $\varepsilon_{\min}$, even if there is then no guarantee that the iterate x_k at which the algorithm appears to be stuck is an approximate minimizer.

If one now insists on true optimality, the details of Algorithm 4.1 become relevant. In this algorithm, the sole purpose of the global minimization in Step 2.1 is to ensure that Lemma 4.4 can be applied to guarantee finite termination of the loop within Step 2.2. Thus, if Step 2.1 cannot be performed exactly, it may happen that this loop does not terminate (even assuming feasibility of the additional global minimizations within the loop). A practical algorithm would terminate this loop if $\delta_{s,j}$ becomes too small or if a maximum number of inner iterations have been taken, returning a value of $\delta_{s,j}$ which is potentially too large for the computed step (compared to what would have resulted if global minimization had always been successful). This is also the outcome of Step 2.2 if the global minimizations involved within this step become too costly and the jth loop must be terminated prematurely. Thus, given that $\delta_{k+1} = \delta_s$ at successful iterations, we next have to consider what happens in Step 1 of iteration $k + 1$ when one or more of the $\delta_{k+1,j}$ is too large. In this case, the definition of $\phi_{w,j}^{\delta_{k+1,j}}(x_{k+1})$ (see (2.2)) implies that there might exist a move $d_{k+1,j}$ with $\|d_{k+1,j}\| \le \delta_{k+1,j}$ such that

$$T_{w,j}(x_{k+1}, 0) - T_{w,j}(x_{k+1}, d_{k+1,j}) > \varepsilon_j \frac{\delta_{k+1,j}^j}{j!},$$

preventing termination even if x_{k+1} is a suitable (ε, δ)-approximate minimizer. This is obviously a serious problem from the point of view of bounding evaluation complexity, since the algorithm will continue and evaluate further, unnecessary, values of f, c, and their derivatives. Two possibilities may then occur. Either iteration $k + 1$ is unsuccessful, σ_k increased causing a subsequent stepsize reduction and, if the behavior persists, forcing convergence to x_k, or it is successful,[5] yielding a further objective function reduction and allowing the algorithm to progress towards an alternative approximate minimizer with a lower objective function value. The complexity bound is maintained if (7.1) or (7.2) holds, or if an insufficient decrease only occurs at most a number of times independent of $\varepsilon_{\min}$. However, even if this is not the case and the complexity bound we have derived evaporates as a consequence, the fact that the algorithm moves on can be viewed as beneficial for the optimization process from a more global perspective.

Termination test. One also needs global minimization to compute the optimality measure $\phi_{w,j}^{\delta_{k,j}}(x_k)$ in Step 1. Clearly, the global optimization defining $\phi_{w,j}^{\delta_{k,j}}(x_k)$ in (2.2) may be terminated as soon as an approximate solution d is found such that

$$\phi_{w,j}^{\delta_{k,j}}(x_k) > \varepsilon_j \frac{\delta_{k,j}^j}{j!},$$

thereby avoiding a full-accuracy computation of the global minimizer. When far from the solution, we expect the optimality measure to be large, and hence such an approximate

[5] As suggested by the fact that minimization in Step 2 of iteration $k + 1$ may obviously be started from $x_{k+1} + d_{k+1,j}$, a point already providing descent on a good approximation of w.

solution d to be found quickly. Suppose now that the solution is approached, and that the minimization of $T_{w,j}(x_k, d)$ within the ball of radius $\delta_{k,j}$ can only be performed inexactly in that one can only find a move d such that

$$T_{w,j}(x_k, d) - T_{w,j}(x_k, d_*) \leq \varepsilon_{\phi,j} \frac{\delta_{k,j}^j}{j!}, \tag{7.3}$$

where d_* is the elusive constrained global minimizer and $\varepsilon_{\phi_j} \in (0, 1]$. Then the only effect of this computational constraint is to limit the achievable accuracy on the approximate minimizer by imposing that $\varepsilon_j \geq \varepsilon_{\phi,j}$. However, achieving (7.3) for small $\delta_{k,j}$ might also be too challenging: one is then left (as above) with the option of using a larger value of $\delta_{k,j}$, possibly missing the identification of x_k as an (ε, δ)-approximate minimizer, which potentially leads to an alternative better one but destroys the complexity guarantee.

To summarize this discussion, the need for global optimization in Steps 1 and 2.4 is driven by the desire to obtain a good evaluation complexity bound (by avoiding further evaluations if a suitable approximate minimizer has been found). The algorithm could still employ approximate calculations, but at the price of losing the complexity guarantee or limiting the achievable accuracy.

8. CONCLUSIONS AND PERSPECTIVES

We have presented an adaptive regularization algorithm for the minimization of nonconvex, nonsmooth composite functions, and proved bounds detailed in Table 1.1 on the evaluation complexity (as a function of accuracy) for composite and smooth problems and for arbitrary model degree and optimality orders.

These results complement the bound proved in [12] for weak approximate minimizers of inexpensively constrained smooth problems (third column of Table 1.1) by providing corresponding results for strong approximate minimizers. They also provide the first complexity results for the convergence to minimizers of order larger than one for (possibly nonsmooth and inexpensively constrained) composite ones.

The fact that high-order approximate minimizers for nonsmooth composite problems can be defined and computed in a quantifiable way opens up interesting possibilities. In particular, these results may be applied in the case of expensively-constrained optimization problems, where exact penalty functions result in composite subproblems of the type studied here.

REFERENCES

[1] A. A. Ahmadi and J. Zhang, Complexity aspects of local minima and related notions. *Adv. Math.* (2021), online.

[2] A. Beck and M. Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.* **2** (2009), 183–202.

[3] S. Bellavia, G. Gurioli, B. Morini, and Ph. L. Toint, Adaptive regularization algorithms with inexact evaluations for nonconvex optimization. *SIAM J. Optim.* **29** (2019), no. 4, 2881–2915.

[4] E. G. Birgin, J.-L. Gardenghi, J. M. Martínez, and S. A. Santos, On the use of third-order models with fourth-order regularization for unconstrained optimization. *Optim. Lett.* **14** (2020), 815–838.

[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. *Math. Program., Ser. A* **163** (2017), no. 1, 359–368.

[6] A. M. Bruckner and E. Ostrow, Some function classes related to the class of convex functions. *Pacific J. Math.* **12** (1962), no. 4, 1203–1215.

[7] C. Cartis, N. I. M. Gould, and Ph. L. Toint, Adaptive cubic overestimation methods for unconstrained optimization. Part II: worst-case function-evaluation complexity. *Math. Program., Ser. A* **130** (2011), no. 2, 295–319.

[8] C. Cartis, N. I. M. Gould, and Ph. L. Toint, On the evaluation complexity of composite function minimization with applications to nonconvex nonlinear programming. *SIAM J. Optim.* **21** (2011), no. 4, 1721–1739.

[9] C. Cartis, N. I. M. Gould, and Ph. L. Toint, Improved worst-case evaluation complexity for potentially rank-deficient nonlinear least-Euclidean-norm problems using higher-order regularized models. Technical Report naXys-12-2015, Namur Center for Complex Systems (naXys), University of Namur, Namur, Belgium, 2015.

[10] C. Cartis, N. I. M. Gould, and Ph. L. Toint, Worst-case evaluation complexity of regularization methods for smooth unconstrained optimization using Hölder continuous gradients. *Optim. Methods Softw.* **6** (2017), no. 6, 1273–1298.

[11] C. Cartis, N. I. M. Gould, and Ph. L. Toint, Second-order optimality and beyond: characterization and evaluation complexity in convexly-constrained nonlinear optimization. *Found. Comput. Math.* **18** (2018), no. 5, 1073–1107.

[12] 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 J. Optim.* **30** (2020), no. 1, 513–541.

[13] X. Chen and Ph. L. Toint, High-order evaluation complexity for convexly-constrained optimization with non-Lipschitzian group sparsity terms. *Math. Program., Ser. A* **187** (2021), 47–78.

[14] X. Chen, Ph. L. Toint, and H. Wang, Partially separable convexly-constrained optimization with non-Lipschitzian singularities and its complexity. *SIAM J. Optim.* **29** (2019), 874–903.

[15] F. E. Curtis, D. P. Robinson, and M. Samadi, An inexact regularized Newton framework with a worst-case iteration complexity of $O(\varepsilon^{-3/2})$ for nonconvex optimization. *IMA J. Numer. Anal.* **00** (2018), 1–32.

[16] D. L. Donoho, Compressed sensing. *IEEE Trans. Inf. Theory* **52** (2006), no. 4, 1289–1306.

[17] Z. Drezner and H. W. Hamacher, *Facility location: applications and theory*. Springer, Heidelberg, Berlin, New York, 2002.

[18] R. Fletcher, *Practical Methods of Optimization: Constrained Optimization*. J. Wiley and Sons, Chichester, England, 1981.

[19] N. I. M. Gould, T. Rees, and J. A. Scott, Convergence and evaluation-complexity analysis of a regularized tensor-Newton method for solving nonlinear least-squares problems. *Comput. Optim. Appl.* **73** (2019), no. 1, 1–35.

[20] S. Gratton, E. Simon, and Ph. L. Toint, An algorithm for the minimization of nonsmooth nonconvex functions using inexact evaluations and its worst-case complexity. *Math. Program., Ser. A* **187** (2021), 1–24.

[21] P. C. Hansen, *Rank-deficient and discrete ill-posed problems: numerical aspects of linear inversion*. SIAM, Philadelphia, USA, 1998.

[22] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, Gradient-based learning applied to document recognition. *Proc. IEEE* **86** (1998), no. 11, 2278–2324.

[23] A. S. Lewis and S. J. Wright, A proximal method for composite minimization. *Math. Program., Ser. A* **158** (2016), 501–546.

[24] Yu. Nesterov and B. T. Polyak, Cubic regularization of Newton method and its global performance. *Math. Program., Ser. A* **108** (2006), no. 1, 177–205.

[25] C. W. Royer and S. J. Wright, Complexity analysis of second-order line-search algorithms for smooth nonconvex optimization. *SIAM J. Optim.* **28** (2018), no. 2, 1448–1477.

[26] R. Tibshirani, Regression shrinkage and selection via the LASSO. *J. Roy. Statist. Soc. Ser. B* **58** (1996), no. 1, 267–288.

[27] N. Wang, J. Choi, D. Brand, C.-Y. Chen, and K. Gopalakrishnan, Training deep neural networks with 8-bit floating point numbers. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 7686–7695, 2018.

CORALIA CARTIS

Mathematical Institute, University of Oxford, Woodstock Road, Oxford OX2 6GG, UK, coralia.cartis@maths.ox.ac.uk

NICHOLAS I. M. GOULD

Computational Mathematics Group, STFC-Rutherford Appleton Laboratory, Chilton OX11 0QX, UK, nick.gould@stfc.ac.uk

PHILIPPE L. TOINT

Namur Centre for Complex Systems (naXys), University of Namur, 61, rue de Bruxelles, B-5000 Namur, Belgium, philippe.toint@unamur.be