Institutional Repository - Research Portal
Dépôt Institutionnel - Portail de la Recherche
researchportal.unamur.be
University of Namur

RESEARCH OUTPUTS / RÉSULTATS DE RECHERCHE

Sharp worst-case evaluation complexity bounds for arbitrary-order nonconvex optimization with inexpensive constraints

Toint, Philippe; Cartis, Coralia; Gould, N. I. M.

Published in:
SIAM Journal on Optimization

DOI:
10.1137/17M1144854

Publication date:
2020

Document Version
Publisher's PDF, also known as Version of record

Link to publication

Citation for pulished version (HARVARD):
Toint, P, Cartis, C & Gould, NIM 2020, 'Sharp worst-case evaluation complexity bounds for arbitrary-order nonconvex optimization with inexpensive constraints', *SIAM Journal on Optimization*, vol. 30, no. 1, pp. 513-541. https://doi.org/10.1137/17M1144854

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal ?

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 04. Jun. 2024

SIAM J. OPTIM.
Vol. 30, No. 1, pp. 513–541
© 2020 Society for Industrial and Applied Mathematics

SHARP WORST-CASE EVALUATION COMPLEXITY BOUNDS FOR ARBITRARY-ORDER NONCONVEX OPTIMIZATION WITH INEXPENSIVE CONSTRAINTS*

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

Abstract. We provide sharp worst-case evaluation complexity bounds for nonconvex minimization problems with general inexpensive constraints, i.e., problems where the cost of evaluating/enforcing of the (possibly nonconvex or even disconnected) constraints, if any, is negligible compared to that of evaluating the objective function. These bounds unify, extend, or improve all known upper and lower complexity bounds for nonconvex unconstrained and convexly constrained problems. It is shown that, given an accuracy level ϵ, a degree of highest available Lipschitz continuous derivatives p, and a desired optimality order q between one and p, a conceptual regularization algorithm requires no more than $O(\epsilon^{-\frac{p+1}{p-q+1}})$ evaluations of the objective function and its derivatives to compute a suitably approximate qth order minimizer. With an appropriate choice of the regularization, a similar result also holds if the pth derivative is merely Hölder rather than Lipschitz continuous. We provide an example that shows that the above complexity bound is sharp for unconstrained and a wide class of constrained problems; we also give reasons for the optimality of regularization methods from a worst-case complexity point of view, within a large class of algorithms that use the same derivative information.

Key words. nonlinear optimization, complexity analysis, regularization methods

AMS subject classifications. 90C30, 68Q25, 65K05, 49M37, 90C60

DOI. 10.1137/17M1144854

1. Introduction.

Ever since the seminal paper by Vavasis [23] on the complexity of finding first-order critical points in unconstrained nonlinear optimization was published 25 years ago, the question of the optimal worst-case complexity of optimization methods has been of interest to mathematicians and also, because of its strong connection with deep learning, to computer scientists. Of late, there has been a growing interest in this research field, both for convex and nonconvex problems. This paper focuses on the latter class and follows a now substantial[1] research trend that derives bounds on the worst-case evaluation complexity (or oracle complexity) of first- and (more rarely) second-order-necessary minimizers[2] of nonlinear nonconvex unconstrained optimization problems [23, 20, 16, 21, 5]. These papers all provide *upper* evaluation complexity bounds: they show that, to obtain an ϵ-approximate first-order-necessary minimizer (for unconstrained problem, this is a point at which the gradient of the objective function is less than ϵ in norm), *at most* $O(\epsilon^{-2})$ evaluations of the

*Received by the editors August 25, 2017; accepted for publication (in revised form) October 21, 2019; published electronically February 20, 2020.
https://doi.org/10.1137/17M1144854
Funding: The work of the first author was supported by the Alan Turing Institute of Data Science, London, UK and of the second author by EPSRC grant EP/M025179/1.
†Mathematical Institute, University of Oxford, Oxford, OX2 6GG, UK (Coralia.Cartis@maths.ox.ac.uk).
‡Computational Science and Engineering Department, STFC-Rutherford Appleton Laboratory, Chilton, England, OX11 0QX, UK (nick.gould@stfc.ac.uk).
§Namur Research Center on Complex Systems (NAXYS), University of Namur, Namur, B5000, Belgium (philippe.toint@unamur.be).
[1]See [13] for a more complete list of references.
[2]That is, points satisfying the first- or second-order necessary optimality conditions for minimization.

objective function[3] are needed if a model involving first derivatives is used, and *at most* $O(\epsilon^{-3/2})$ evaluations are needed if using second derivatives is permitted. This result was extended to convexly constrained problems in [7]. A broader framework allowing the use of Taylor series of degree p was more recently proposed in [2], in which case the worst-case evaluation complexity bound for ϵ-first-order-necessary unconstrained minimizer is shown to be $O(\epsilon^{-\frac{p+1}{p}})$, thereby generalizing the previous results for this case. Complexity for obtaining ϵ-approximate second-order-necessary unconstrained minimizers was considered in [21, 5], where a bound of $O(\epsilon^{-3})$ evaluations was proved to obtain an ϵ-second-order-necessary minimizer using a Taylor's model of degree two, and a bound of $O(\epsilon^{-\frac{p+1}{p-1}})$ evaluations was shown in [9] for the case where a Taylor model of degree p is used. Defining qth-order-necessary minimizers for $q > 2$ was considered in [12], where the difficulty of stating and verifying necessary optimality was discussed. In particular, it was concluded in this latter reference that defining and computing ϵ-approximate qth-order-necessary minimizers for $q > 2$ is likely to remain elusive, essentially because of the nonlinearity and lack of continuity of the kernels of the derivatives involved. A more general Taylor-based definition of optimality was introduced instead, which allowed to show an upper bound of $O(\epsilon^{-(q+1)})$ on evaluation complexity for convexly constrained problems, in particular improving on the bound of $O(\epsilon^{-9/2})$ stated in [1] for the case $p = q = 3$.

The unconstrained and convexly constrained cases where the assumption of Lipschitz continuity is replaced by the weaker β-Hölder continuity ($\beta \in (0, 1]$) have also been studied for $q = 1$ in [15, 8, 10]. These references show that *at most* $O(\epsilon^{-\frac{p+\beta}{p-1+\beta}})$ evaluations are needed for obtaining an ϵ-first-order-necessary minimizer.

While upper complexity bounds are important as they provide a handle on the intrinsic difficulty of the considered problem, they do so at the condition of not being overly pessimistic. To address this last point, *lower* bounds on the evaluation complexity of unconstrained nonconvex optimization problems and methods were derived in [4, 20] and [13], where it was shown that the known upper complexity bounds are sharp (irrespective of the problem's dimension) for most known methods using Taylor's models of degree one or two. That is to say that there are examples for which the complexity order predicted by the upper bound is actually achieved. More recently, Carmon et al. [3] provided an elaborate construction showing that *at least* a multiple of $\epsilon^{-\frac{p+1}{p}}$ function evaluations may be needed to obtain an ϵ-first-order-necessary unconstrained minimizer where derivatives of order at most p are used. This result, which matches in order the upper bound of [2], covers a very wide class of potential optimization methods[4] but has the drawback of being only valid for problems whose dimension essentially exceeds the number of iterations needed, which can be very large and quickly grows when ϵ tends to zero.

Contributions. The present paper aims at unifying and generalizing all the above results in a single framework, providing, for problems with inexpensive or no constraints, provably optimal evaluation complexity bounds for arbitrary optimality order, all relevant model degrees and levels of smoothness of the objective function. By "inexpensive constraints," we mean general set constraints whose enforcement and evaluation[5] cost is negligible compared to the cost of evaluating the objective function. As a consequence, the evaluation complexity for such problems is meaningfully captured by focusing of the number of evaluations of this latter function. This class of minimization problems contains important cases such as bound-constrained

[3] And its available derivatives.

[4] In particular, it covers randomized methods, which we do not consider in this paper.

[5] Constraints' values and that of their derivatives, if relevant.

problems and convexly constrained problems (when the projection onto the feasible set is inexpensive) but also allows possibly nonconvex or even disconnected feasible sets.

In order to achieve these objectives, we first revisit the Taylor-based optimality measure of [12] and define (ϵ,δ)-qth-order-necessary minimizers, a notion extending the standard ϵ-first- and ϵ-second-order cases to arbitrary orders. We then present a conceptual regularization algorithm using degree p models and show that this algorithm requires at most $O(\epsilon^{-\frac{p+\beta}{p-q+\beta}})$ evaluations of f and its derivatives to find such an (ϵ,δ)-qth-order-necessary minimizer when the pth derivative of f is assumed to be β-Hölder continuous. (If the pth derivative is assumed to be Lipschitz continuous, the bound becomes $O(\epsilon^{-\frac{p+1}{p-q+1}})$.) This bound matches the best known lower bounds for first and second order and improves on the bound in $O(\epsilon^{-(q+1)})$ given by [12]. We then show that this bound is sharp in order for unconstrained problems with Lipschitz continuous pth derivative by completing and extending the result of [3] in two ways. The first is to show that the lower worst-case bound of order $\epsilon^{-\frac{p+1}{p}}$ evaluations for obtaining a first-order-necessary minimizer using at most p derivatives is also valid for problems of every dimension, and the second is to show that this bound can be generalized to a multiple of $\epsilon^{-\frac{p+1}{p-q+1}}$ for obtaining a qth-order-necessary minimizer of any order $q \leq p$. In particular, this result matches in order the upper bound obtained in the first part of the paper and subsumes or improves known lower bounds for first- and second-order-necessary minimizers. While our lower bounds are derived for regularization algorithms applied to unconstrained problems, we also indicate that they may be extended to a much wider class of minimization methods and to a significant class of constrained problems.

The paper is organized as follows. Section 2 introduces the (possibly constrained) minimization problem of interest and the concept of (ϵ,δ)-approximate qth-order-necessary minimizers. It also presents a variant of the Adaptive Regularization algorithm using degree p Taylor's models (ARp) whose purpose is to find such minimizers. Section 3 then provides an upper bound on the evaluation complexity for the ARp algorithm to achieve this task. Section 4 then discusses specialization of this result to the case where ϵ-approximate second-order-necessary minimizers are sought. The complexity upper bound of section 3 is then proved to be sharp in section 5 for the Lipschitz-continuous cases where the feasible set contains a ray. Some conclusions are finally presented in section 6.

Notation. Throughout the paper, $\|v\|$ denotes the standard Euclidean norm of a vector $v \in \mathbb{R}^n$. For a symmetric tensor S of order p, $S[v_1, \ldots, v_p]$ is the result of applying S to the vectors $v_1, \ldots, v_p$, $S[v]^p$ is the result of applying S to p copies of the vector v, and

$$\|S\|_{[p]} \stackrel{\text{def}}{=} \max_{\|v\|=1} |S[v]^p| = \max_{\|v_1\|=\cdots=\|v_p\|=1} |S[v_1, \ldots, v_p]| \tag{1.1}$$

(where the second equality results from Theorem 2.1 in [25]) is the associated induced norm for such tensors. If S_1 and S_2 are tensors, $S_1 \otimes S_2$ is their tensor product and $S_1^{k\otimes}$ is the product of S_1 k times with itself. For a real, sufficiently differentiable univariate function f, $f^{(i)}$ denotes its ith derivative and $f^{(0)}$ is a synonym for f. For an integer k and a real $\beta \in (0, 1]$, we define $(k+\beta)! \stackrel{\text{def}}{=} \prod_{\ell=1}^{k}(\beta+\ell)$ (this coincides with the standard factorial if $\beta = 1$) and $\beta! = 1$. As is usual, we also define $0! = 1$. If M is a symmetric matrix, $\lambda_{\min}(M)$ is its leftmost eigenvalue. If α is a real, $\lceil\alpha\rceil$ and $\lfloor\alpha\rfloor$ denote the smallest integer not smaller than α and the largest integer not

exceeding α, respectively. Finally $\operatorname{globmin}_{x\in\mathcal{S}} f(x)$ denotes the smallest value of $f(x)$ over $x \in \mathcal{S}$.

2. High-order necessary conditions for optimality and the ARp algorithm.

2.1. A high-order optimality measure. Given $p \geq 1$, this paper considers the set-constrained optimization problem

$$\min_{x\in\mathcal{F}} f(x), \tag{2.1}$$

where we assume that $\mathcal{F} \subseteq \mathbb{R}^n$ is closed and nonempty, and where $f \in \mathcal{C}^{p,\beta}(\mathbb{R}^n)$, namely, that

- f is p-times continuously differentiable,
- the pth derivative tensor of f at x is globally Hölder continuous, that is, there exist constants $L \geq 0$ and $\beta \in (0, 1]$ such that, for all $x, y \in \mathbb{R}^n$,

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

Observe that convexity or even connectedness of $\mathcal{F}$ is not requested. Observe also that the more usual case of *Lipschitz continuous pth derivative corresponds to* $\beta = 1$. We note that our assumption covers the continuous range of the objective function's smoothness from Hölder continuous gradients to Lipschitz continuous pth derivatives. In what follows, we assume that β is known.

If $T_p(x, s)$ is the standard pth degree Taylor's expansion of f about x computed for the increment s, that is,

$$T_p(x, s) \stackrel{\text{def}}{=} f(x) + \sum_{\ell=1}^{p} \frac{1}{\ell!} \nabla_x^\ell f(x)[s]^\ell, \tag{2.3}$$

(2.2) provides crucial approximation bounds, whose proof can be found in the appendix.

LEMMA 2.1. *Let $f \in \mathcal{C}^{p,\beta}(\mathbb{R}^n)$, and let $T_p(x, s)$ be the Taylor approximation of $f(x + s)$ about x given by* (2.3). *Then for all $x, s \in \mathbb{R}^n$,*

$$f(x + s) \leq T_p(x, s) + \frac{L}{(p + \beta)!} \|s\|^{p+\beta}, \tag{2.4}$$

$$\|\nabla_x^j f(x + s) - \nabla_s^j T_p(x, s)\|_{[j]} \leq \frac{L}{(p - j + \beta)!} \|s\|^{p-j+\beta}. \quad (j = 1, \ldots, p). \tag{2.5}$$

In order to characterize minimizers of (2.1), we follow [12] and introduce, for given $\delta \in (0, 1]$ and $j \leq p$,

$$\phi_{f,j}^\delta(x) \stackrel{\text{def}}{=} f(x) - \operatorname*{globmin}_{\substack{x+d\in\mathcal{F} \\ \|d\|\leq\delta}} T_j(x, d), \tag{2.6}$$

which can be interpreted as the *magnitude of the largest decrease achievable on the Taylor's expansion of degree j within the intersection of a ball of radius δ with the feasible set.* It was shown in [12] that $\phi_{f,j}^\delta(x)$ is a proper generalization of well-known unconstrained optimality measures for low orders. In particular, for any $\delta > 0$, we have

$$\phi_{f,1}^\delta(x) = \|\nabla_x^1 f(x)\|\, \delta, \tag{2.7}$$

$$\phi^{\delta}_{f,2}(x) = \frac{1}{2}\left|\min[0, \lambda_{\min}\left(\nabla^2_x f(x)\right)\right| \, \delta^2 \tag{2.8}$$

provided $\nabla^1_x f(x) = 0$, and also, if additionally $\nabla^2_x f(x)$ is positive semidefinite, that

$$\phi^{\delta}_{f,3} = \| \text{ projection of } \nabla^3_x f(x) \text{ onto the nullspace of } \nabla^2_x f(x) \, \| \, \delta^3. \tag{2.9}$$

At variance with other optimality measures, $\phi^{\delta}_{j,f}(x)$ is well-defined for any order $j \geq 1$ and varies continuously when x varies continuously in $\mathcal{F}$. The role of the "optimality radius" δ in (2.6) merits some discussion. It follows from (2.7)–(2.9) that the choice of $\delta = 1$ is adequate for retrieving known optimality conditions in the unconstrained case for $j = 1$, and $j = 2$ provided $\nabla^1_x f(x) = 0$, and $j = 3$ provided additionally $\nabla^2_x f(x)$ is positive semidefinite. However, δ becomes important in other cases. Corollary 3.6 in [12] indicates that, when $\mathcal{F}$ is convex, qth-order necessary "path-based" optimality conditions hold if

$$\lim_{\delta \to 0} \frac{\phi^{\delta}_{f,j}(x)}{\delta^j} = 0 \quad \text{for} \quad j = 1, \ldots, q. \tag{2.10}$$

The limit for $\delta \to 0$ is necessary to capture the notion of local minimizer for (2.1). This implies that δ should be seen as a truly *local* quantity associated with x. However, considering $\phi^{\delta}_{f,j}(x)$ for nonvanishing δ has substantial advantages from the point of view of optimization: while it may fail to indicate that x is a local minimizer, it does so only by providing a direction leading to values of f below $f(x)$, thereby helping to avoid local but nonglobal approximate solutions. We refer the reader to [12] for a further discussion but conclude that considering large δ has strong advantages when solving (2.1).

A special case is when x is an isolated feasible point, that is, a point which is the sole intersection between $\mathcal{F}$ and any sufficiently small neighborhood of x. Such a point is clearly a local minimizer, and this is reflected by the fact that $\phi^{\delta}_{f,q}(x) = 0$ for any f, any q, and any sufficiently small δ.

The main drawback of using $\phi^{\delta}_{f,j}(x)$ is, of course, that its computation requires the global minimization of $T_p(x, d)$ in the intersection of the ball of radius δ with $\mathcal{F}$. We are not aware of an easy way to do this in general[6] when $n > 1$, which is why our analysis remains of an essentially theoretical nature, as was the case for [12]. Note, however, that, albeit potentially very difficult, solving this global minimization problem does not involve calculating the value of f or of any of its derivatives. In that sense, this drawback is thus irrelevant for the worst-case evaluation complexity which solely focuses on these evaluations.

Observe now that, if we were to relax the first-order condition $\nabla^1_x f(x) = 0$ for unconstrained problems to $\|\nabla^1_x f(x)\| \leq \epsilon$ and, at the same time, relax the second-order condition to $|\min[0, \lambda_{\min}(\nabla^2_x f(x))]| \leq \epsilon$, we then deduce that

$$\phi^{\delta}_{f,2}(x) \leq \epsilon\delta + \frac{1}{2}\epsilon\delta^2 = \epsilon \sum_{\ell=1}^{2} \frac{\delta^\ell}{\ell!}. \tag{2.11}$$

A natural generalization of this observation is to define an (ϵ, δ)-*approximate* q*th-order-necessary minimizer* of f as a point x such that

$$\phi^{\delta}_{f,q}(x) \leq \epsilon\chi_q(\delta), \tag{2.12}$$

[6] A small value of δ might help, but this computation remains NP-hard in most cases.

where

$$\chi_q(\delta) \stackrel{\text{def}}{=} \sum_{\ell=1}^{q} \frac{\delta^\ell}{\ell!}. \tag{2.13}$$

Because (2.12) is a new way to look at approximate optimality and is crucial for the rest of this paper, it is worthwhile to motivate and discuss it further.

1. When $\epsilon = 0$, (2.12) implies that the complicated path-based necessary optimality conditions derived in [12] do hold. This results from the fact that these latter conditions merely express that the Taylor's model of order q cannot decrease close enough to x along any feasible polynomial path emanating from x, which is clearly the case if x is a global minimizer of the same models in the intersection of the feasible set and a ball of radius δ centered at x. By continuity, these path-based conditions must therefore hold in the limit under (2.12) when ϵ tends to zero. The role of (2.12) as a condition for approximate minimization is thus coherent and consistent with known necessary conditions.
2. Inspired by (2.10), the stronger approximate optimality condition

$$\phi^\delta_{f,j}(x) \leq \epsilon\, \delta^j \quad \text{for} \quad j \in \{1, \ldots, q\} \tag{2.14}$$

was used in [12] instead of (2.12). Our main reason to prefer (2.12) is the following. Observe that (2.14) implies in particular that $\phi^\delta_{f,q}(x) \leq \epsilon\delta^q$, which in turn implies, for δ small enough for the first-order term to dominate, that $\phi^\delta_{f,1}(x) \leq \epsilon\delta^q$. In the unconstrained case (for example), this requires $\|\nabla^1_x f(x_k)\| \leq \epsilon\delta^{q-1}$, imposing an inordinate level of first-order optimality, much stronger than the standard condition $\|\nabla^1_x f(x_k)\| \leq \epsilon$. No such difficulty arises with (2.12) because the right-hand side of the condition involves all powers of δ, which is not the case of the right-hand side of (2.14). Note, however, that the vital continuity properties of $\phi^\delta_{f,q}$ are not affected by the choice of the right-hand side and are thus inherited by (2.12).
3. For given $\delta \in (0, 1]$, (2.12) does not imply that $\phi^\delta_{f,j}(x) \leq \epsilon\chi_j(\delta)$ for $j \in \{1, \ldots, q-1\}$, although the violation of this condition tends to zero with δ.[7] This slight blemish can be cured by requiring that $\phi^\delta_{f,j}(x) \leq \epsilon\chi_j(\delta)$ for $j \in \{1, \ldots, q\}$ instead of (2.12).
4. Since $\delta \in (0, 1]$, we note that $\chi_q(\delta) = \Theta(\delta)$, and so an equivalent alternative to the termination condition (2.12) is to require that $\phi^\delta_{f,q}(x) \leq \epsilon\delta$. We use (2.12) as it naturally occurs in subsequent proofs.

In order to further justify (2.12), we now make more explicit the "minimizing guarantees" provided by this approximate optimality condition, by formulating a result analogous to Theorem 3.7 in [12]. This result gives a lower bound on the value of $f(x)$ in the feasible neighborhood of an (ϵ, δ)-approximate qth-order-necessary minimizer.

THEOREM 2.2. *Suppose that f is p times continuously differentiable and that $\nabla^q_x f$ is β-Hölder continuous with constant L (in the sense of* (2.2) *with $p = q$) in an open neighborhood of radius $\delta \in (0, 1]$ of some $x \in \mathcal{F}$. Suppose also that x is an (ϵ, δ)-approximate qth-order-necessary minimizer of f in the sense of* (2.12)*. Then*

[7] When δ tends to zero, the terms of orders $j + 1$ and higher in the Taylor's expansion defining $\phi^\delta_{f,q}(x)$ and $\chi_q(\delta)$ become negligible compared to the first j.

$$f(x+d) \geq f(x) - 2\epsilon\chi_q(\delta) \quad \textit{for all } d \textit{ with } \quad x+d \in \mathcal{F} \quad \textit{and} \quad \|d\| \leq \min\left[\delta, \left(\frac{(q+\beta)!\,\epsilon}{L}\right)^{\frac{1}{q+\beta-1}}\right]. \tag{2.15}$$

Proof. Using the triangle inequality, (2.2), (2.4), and (2.12), we obtain that, for $\|d\| \leq \delta$,

$$\begin{aligned} f(x+d) &\geq f(x+d) - T_q(x,d) + T_q(x,d) \\ &\geq -|f(x+d) - T_q(x,d)| + T_q(x,0) - \phi^{\delta}_{f,q}(x) \\ &\geq -\frac{L}{(q+\beta)!}\|d\|^{q+\beta} + f(x) - \epsilon\chi_q(\delta). \end{aligned}$$

Thus,

$$f(x+d) \geq f(x) - \frac{L}{(q+\beta)!}\|d\|^{q+\beta-1}\,\delta - \epsilon\chi_q(\delta),$$

and the desired bound follows from the fact that $\delta \leq \chi_q(\delta)$. ∎

2.2. The ARp algorithm for high-order criticality. In order to find (ϵ, δ)-approximate qth-order-necessary minimizers, we consider applying a variant of the ARp algorithm to (2.1). This algorithm, described as Algorithm 2.1 on the next page, is of the regularization type in that, at each iterate x_k, a step s_k is computed which approximately minimizes (in a sense defined below) the model

$$m_k(s) = T_p(x_k, s) + \frac{\sigma_k}{(p+\beta)!}\|s\|^{p+\beta} \tag{2.16}$$

subject to $x_k + s \in \mathcal{F}$, where p in an integer such that $p \geq q$ and $\sigma_k \geq \sigma_{\min}$ is a "regularization parameter."

A few comments are useful at this stage.

1. Since $\sigma_k \geq \sigma_{\min}$ by (2.22), we have that $m_k(s)$ is bounded below as a function of s and the existence of a constrained global minimizer s^*_k is guaranteed because $\beta > 0$. To the best of our knowledge, the only methods that can use $\beta = 0$ are universal methods [8, 15, 19], where a higher power of regularization is required and additional precautions are taken when computing a step, and which we do not cover here.
2. Conditions (2.19) and (2.20) essentially ensure that the step is long enough, which will be important for proving the important lower bound on the step-length in Lemma 3.3 below. If (2.19) holds, the possibly expensive computation of $\phi^{\delta_s}_{m_k,q}(s_k)$ in (2.20) is unnecessary and δ_s may be chosen arbitrarily in $(0, 1]$.
3. Our choice to update δ_{k+1} in parallel with x_{k+1} reflects our earlier comment on the fact that δ is a local quantity: hence δ_{k+1} should be consistent with its corresponding value at $x_{k+1} = x_k + s_k$, which is δ_s.
4. We assume the availability of a feasible starting point, which is without loss of generality for inexpensive constraints.
5. Before termination, each successful iteration requires the evaluation of f and its first p derivative tensors, while only the evaluation of f is needed at unsuccessful ones.
6. The mechanism of the algorithm ensures the nonincreasing nature of the sequence $\{f(x_k)\}_{k\geq 0}$.

Algorithm 2.1. ARp for (ϵ, δ)-approximate qth-order-necessary minimizers.

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

$$\varpi \in (0,1], \ \ \theta > 0, \ \ \delta_0 \in (0,1], \ \ \sigma_{\min} \in (0, \sigma_0], \ \ 0 < \eta_1 \leq \eta_2 < 1, \text{ and } \ 0 < \gamma_1 < 1 < \gamma_2 < \gamma_3. \tag{2.17}$$

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

Step 1: Test for termination. Evaluate $\{\nabla_x^i f(x_k)\}_{i=1}^q$. If (2.12) holds with $\delta = \delta_k$, terminate with the approximate solution $x_\epsilon = x_k$. Otherwise compute $\{\nabla_x^i f(x_k)\}_{i=q+1}^p$.

Step 2: Step calculation. Attempt to compute an approximate minimizer s_k of model $m_k(s)$ and an optimality radius $\delta_s \in (0,1]$ such that $x_k + s_k \in \mathcal{F}$,

$$m_k(s_k) < m_k(0), \tag{2.18}$$

and either

$$\|s_k\| \geq \varpi \epsilon^{\frac{1}{p-q+\beta}} \tag{2.19}$$

or

$$\phi_{m_k,q}^{\delta_s}(s_k) \leq \frac{\theta \|s_k\|^{p-q+\beta}}{(p-q+\beta)!} \chi_q(\delta_s). \tag{2.20}$$

If such a step does not exist, terminate with the approximate solution $x_\epsilon = x_k$.

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

$$\rho_k = \frac{f(x_k) - f(x_k + s_k)}{T_p(x_k, 0) - T_p(x_k, s_k)}. \tag{2.21}$$

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

Step 4: Regularization parameter update. Set

$$\sigma_{k+1} \in \begin{cases} [\max(\sigma_{\min}, \gamma_1 \sigma_k), \sigma_k] & \text{if } \ \rho_k \geq \eta_2, \\ [\sigma_k, \gamma_2 \sigma_k] & \text{if } \ \rho_k \in [\eta_1, \eta_2), \\ [\gamma_2 \sigma_k, \gamma_3 \sigma_k] & \text{if } \ \rho_k < \eta_1. \end{cases} \tag{2.22}$$

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

Iterations for which $\rho_k \geq \eta_1$ (and hence $x_{k+1} = x_k + s_k$) are called "successful," and we denote by $\mathcal{S}_k \stackrel{\text{def}}{=} \{0 \leq j \leq k \mid \rho_j \geq \eta_1\}$ the index set of all successful iterations between 0 and k. We immediately observe that the total number of iterations (successful or not) can be bounded as a function of the number of successful ones (and include a proof in the appendix).

LEMMA 2.3 (see [2, Theorem 2.4]). *The mechanism of Algorithm* 2.1 *guarantees that, if*

$$\sigma_k \leq \sigma_{\max} \tag{2.23}$$

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

$$k+1 \leq |\mathcal{S}_k| \left(1 + \frac{|\log \gamma_1|}{\log \gamma_2}\right) + \frac{1}{\log \gamma_2} \log\left(\frac{\sigma_{\max}}{\sigma_0}\right). \tag{2.24}$$

2.3. Specific properties of the ARp algorithm. We now need to verify that the algorithm is well-defined in the sense that either a step s_k and associated δ_s satisfying (2.18)–(2.20) can always be found, or termination is justified. For unconstrained problems with $q \in \{1, 2\}$, the first possibility directly results from the observation that $\phi^{\delta_s}_{m_k,j}(s_k)$ (as given by (2.7)–(2.9) for $f = m_k$ and $j \in \{1, 2, 3\}$) can be made suitably small at a global minimizer of the model. In those cases $\delta_s = 1$ is always acceptable. (More details for the case $q = 2$ are provided in section 4). The situation is more complicated for $q \geq 3$ because a global minimizer of the model (2.16) may not be a global minimizer of its qth order Taylor's expansion in the intersection of $\mathcal{F}$ and a ball of arbitrary radius: we may have to restrict this radius for this important property to hold. In order to clarify this issue, we first state a useful technical lemma, whose proof is in the appendix.

LEMMA 2.4. *Let* s *be a vector of* $\mathbb{R}^n$*. Then*

$$\| \nabla_s^j \big(\|s\|^{p+\beta}\big) \|_{[j]} = \frac{(p+\beta)!}{(p-j+\beta)!} \|s\|^{p-j+\beta} \quad \text{for } j \in \{0, \ldots, p\} \tag{2.25}$$

and

$$\| \nabla_s^{p+1} \big(\|s\|^{p+\beta}\big) \|_{[p+1]} = \beta\,(p+\beta)!\,\|s\|^{\beta-1}. \tag{2.26}$$

We next provide reasonable sufficient conditions for a nonzero step s_k and an optimality radius δ_s to satisfy (2.18)–(2.20).

LEMMA 2.5. *Suppose that* s_k^* *is a global minimizer of* $m_k(s)$ *under the constraint that* $x_k + s \in \mathcal{F}$ *such that* $m_k(s_k^*) < m_k(0)$*. Then there exist a neighborhood of* s_k^* *and a range of sufficiently small* δ *such that* (2.18) *and* (2.20) *hold for any* s_k *in the intersection of this neighborhood with* $\mathcal{F}$ *and any* δ_s *in this range.*

Proof. Let s_k^* be the global minimizer of the model $m_k(s)$ over all s such that $x_k + s \in \mathcal{F}$. 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,

$$0 \leq m_k(s_k^* + d) - m_k(s_k^*) = \sum_{\ell=1}^{p} \frac{1}{\ell!} \nabla_s^\ell m_k(s_k^*)[d]^\ell + \frac{1}{(p+1)!} \nabla_s^{p+1} m_k(s_k^* + \xi d)[d]^{p+1}$$

for some $\xi \in (0, 1)$. Thus, using the triangle inequality, (2.16), and (2.26),

$$\begin{aligned} -\sum_{\ell=1}^{q} \frac{1}{\ell!} \nabla_s^\ell m_k(s_k^*)[d]^\ell &\leq \sum_{\ell=q+1}^{p} \frac{\|d\|^\ell}{\ell!} \big\|\nabla_s^\ell m_k(s_k^*)\big\|_{[\ell]} + \frac{\|d\|^{p+1}}{(p+1)!} \big\|\nabla_s^{p+1} m_k(s_k^* + \xi d)\big\|_{[p+1]} \\ &= \sum_{\ell=q+1}^{p} \frac{\|d\|^\ell}{\ell!} \big\|\nabla_s^\ell m_k(s_k^*)\big\|_{[\ell]} + \beta\sigma_k \frac{\|d\|^{p+1}}{(p+1)!} \|s_k^* + \xi d\|^{\beta-1}. \end{aligned} \tag{2.27}$$

Since $s_k^* \neq 0$, we may then choose $\delta_s < \|s_k^*\|$ such that, for every d with $\|d\| \leq \delta_s$, $\|s_k^* + \xi d\| \geq \frac{1}{2}\|s_k^*\| > 0$ and

$$\sum_{\ell=q+1}^{p} \frac{\|d\|^\ell}{\ell!} \|\nabla_s^\ell m_k(s_k^*)\|_{[\ell]} + 2^{1-\beta}\beta\sigma_k \frac{\|d\|^{p+1}}{(p+1)!} \|s_k^*\|^{\beta-1} \leq \frac{\theta \|s_k^*\|^{p-q+\beta}}{2(p-q+\beta)!} \|d\|. \tag{2.28}$$

Hence we deduce from (2.27) and (2.28) that, for $\|d\| \leq \delta_s$,

$$-\sum_{\ell=1}^{q} \frac{1}{\ell!} \nabla_s^\ell m_k(s_k^*)[d]^\ell \leq \frac{\theta \|s_k^*\|^{p-q+\beta}}{2(p-q+\beta)!}\, \delta_s \leq \frac{\theta \|s_k^*\|^{p-q+\beta}}{2(p-q+\beta)!}\, \chi_q(\delta_s),$$

where the last inequality follows from (2.13). Continuity of m_k and its derivatives and the inequality $m_k(s_k^*) < m_k(0)$ then imply that there exists a neighborhood of $s_k^* \neq 0$ such that (2.18) holds and

$$-\sum_{\ell=1}^{q} \frac{1}{\ell!} \nabla_s^\ell m_k(s)[d]^\ell \leq \frac{\theta \|s\|^{p-q+\beta}}{(p-q+\beta)!}\, \chi_q(\delta_s)$$

for all s in this neighborhood and all d with $\|d\| \leq \delta_s$. This yields that, for all such s with $x_k + s \in \mathcal{F}$,

$$\phi_{m_k,q}^{\delta_s}(s) = \max\left[0, \operatorname*{globmax}_{\substack{\|d\| \leq \delta_s \\ x_k + d \in \mathcal{F}}} \left(-\sum_{\ell=1}^{q} \frac{1}{\ell!} \nabla_s^\ell m_k(s)[d]^\ell \right)\right] \leq \frac{\theta \|s\|^{p-q+\beta}}{(p-q+\beta)!}\, \chi_q(\delta_s),$$

as requested. ∎

As can be seen in the proof of this lemma, δ_s may need to be small if any of the tensors

$$\nabla_s^\ell m_k(s_k^*) = \sum_{j=\ell}^{p} \frac{1}{j!} \nabla_s^j m_k(0)[s_k^*]^{j-\ell}$$

for $\ell \in \{1, \ldots, p+1\}$ has a large norm. This may occur in particular if β and $\|s_k^*\|$ are both close to zero, as is shown by the last term in the left-hand side of (2.28). We also note that (2.20) obviously holds for $s_k = s_k^*$ if $x_k + s_k^*$ is an isolated feasible point.

That one needs to consider the second case in Step 2 (where no step exists satisfying (2.18)–(2.20)) can be seen by examining the following one-dimensional example. Let $p = q = 3$ and $\beta = 1$, and suppose that $\delta_k = 1$, $T_q(x_k, s) = s^2 - 2s^3$, and $\sigma_k = 4! = 24$. Then $m_k(s) = s^2 - 2s^3 + s^4 = s^2(s-1)^2$ and the origin is a global minimizer of the model (and a local minimizer of $T_q(x_k, s)$) but yet $T_q(x_k, \delta_k) = -1$, yielding that $\phi_{f,q}^{\delta_k}(x_k) = 1 > \epsilon\chi_q(1)$ for $\epsilon \leq 1/\chi_q(1) = \frac{4}{7}$. Thus, Step 1 with $\delta_k = 1$ has failed to identify that termination was possible. It now remains to verify that it is justified to terminate in Step 2 when no suitable nonzero step can be found.

LEMMA 2.6. *Suppose that the algorithm terminates in Step* 2 *of iteration* k *with* $x_\epsilon = x_k$. *Then there exists a* $\delta \in (0, 1]$ *such that* (2.12) *holds for* $x = x_\epsilon$ *and* x_ϵ *is an* (ϵ, δ)*-approximate qth-order-necessary minimizer.*

Proof. Given Lemma 2.5, if the algorithm terminates within Step 2, it must be because every global minimizer s_k^* of $m_k(s)$ under the constraints $x_k + s \in \mathcal{F}$ is such that $m_k(s_k^*) \geq m_k(0)$. In that case, $s_k^* = 0$ is one such global minimizer and we have that, for all d,

$$0 \leq m_k(d) - m_k(0) = \sum_{\ell=1}^{q} \frac{1}{\ell!} \nabla_x^j f(x_k)[d]^j + \sum_{\ell=q+1}^{p} \frac{1}{\ell!} \nabla_x^j f(x_k)[d]^j + \frac{\sigma_k}{(p+\beta)!} \|d\|^{p+\beta}.$$

We may now choose $\delta \in (0,1]$ small enough to ensure that, for all d with $\|d\| \leq \delta$,

$$\left| \sum_{\ell=q+1}^{p} \frac{1}{\ell!} \nabla_x^j f(x_k)[d]^j + \frac{\sigma_k}{(p+\beta)!} \|d\|^{p+\beta} \right| \leq \epsilon \|d\| \leq \epsilon \, \chi_q(\delta), \tag{2.29}$$

which in turn implies that, for all d with $\|d\| \leq \delta$,

$$\phi_{f,q}^{\delta}(x_k) = \max \left[0, \operatorname*{globmax}_{\substack{\|d\| \leq \delta \\ x_k + d \in \mathcal{F}}} \left(- \sum_{\ell=1}^{q} \frac{1}{\ell!} \nabla_x^{\ell} f(x_k)[d]^{\ell} \right) \right] \leq \epsilon \, \chi_q(\delta),$$

concluding the proof. ∎

Observe that, in this proof, we could have chosen δ small enough to ensure

$$\frac{\sigma_k}{(p+\beta)!} \|d\|^{p+\beta} \leq \epsilon \chi_p(\delta)$$

instead of (2.29), yielding $\phi_{f,p}^{\delta}(x_k) \leq \epsilon \chi_p(\delta)$, which is a stronger necessary optimality condition than (2.12). Together, Lemmas 2.5 and 2.6 ensure that Algorithm 2.1 is well-defined.

However, none of the inner step and criticality computations involve the (re-) evaluation of f or its derivatives, and therefore the evaluation complexity bound presented in the next section is unaffected.

3. An upper bound on the evaluation complexity. The proofs of the following two lemmas are very similar to corresponding results in [2], and hence we again defer them to the appendix (but still include them for completeness, as the algorithm has changed).

LEMMA 3.1. *The mechanism of Algorithm* 2.1 *guarantees that, for all* $k \geq 0$,

$$T_p(x_k, 0) - T_p(x_k, s_k) \geq \frac{\sigma_k}{(p+\beta)!} \|s_k\|^{p+\beta}, \tag{3.1}$$

and so (2.21) *is well-defined.*

LEMMA 3.2. *Let* $f \in \mathcal{C}^{p,\beta}(\mathbb{R}^n)$. *Then, for all* $k \geq 0$,

$$\sigma_k \leq \sigma_{\max} \stackrel{\text{def}}{=} \max \left[\sigma_0, \frac{\gamma_3 L}{1 - \eta_2} \right]. \tag{3.2}$$

We are now in position to prove the crucial lower bound on the step length.

LEMMA 3.3. *Let* $f \in \mathcal{C}^{p,\beta}(\mathbb{R}^n)$. *Then, for all* $k \geq 0$ *such that iteration* k *is successful and Algorithm* 2.1 *does not terminate at iteration* $k+1$,

$$\|s_k\| \geq \kappa_s \epsilon^{\frac{1}{p-q+\beta}}, \tag{3.3}$$

where

$$\kappa_s \stackrel{\text{def}}{=} \min \left[\varpi, \left(\frac{(p-q+\beta)!}{(L + \sigma_{\max} + \theta)} \right)^{\frac{1}{p-q+\beta}} \right]. \tag{3.4}$$

Proof. If $\|s_k\| > \varpi\epsilon^{\frac{1}{p-q+\beta}}$ (i.e., (2.19) holds), the result is obvious. Suppose now that $\|s_k\| \leq \varpi\epsilon^{\frac{1}{p-q+\beta}}$, which, in view of Step 2 of the algorithm, implies that (2.20) holds. Since the algorithm does not terminate at iteration $k+1$, we have that

$$\phi_{f,q}^{\delta_{k+1}}(x_{k+1}) > \epsilon\chi_q(\delta_{k+1}). \tag{3.5}$$

Let the global minimum in the definition of $\phi_{f,q}^{\delta_{k+1}}(x_{k+1})$ be achieved at d with $\|d\| \leq \delta_{k+1}$. Since $\phi_{f,q}^{\delta_{k+1}}(x_{k+1}) > 0$, we have from (2.6) that

$$\sum_{\ell=1}^{q} \frac{1}{\ell!}\nabla_x^\ell f(x_{k+1})[d]^\ell < 0.$$

Then, successively using (2.6) for f at x_{k+1}, the triangle inequality, (2.16), (1.1), and (2.25), we deduce that

(3.6)

$$\begin{aligned}
&\phi_{f,q}^{\delta_{k+1}}(x_{k+1}) \\
&\quad = -\sum_{\ell=1}^{q} \frac{1}{\ell!}\nabla_x^\ell f(x_{k+1})[d]^\ell \\
&\quad = -\sum_{\ell=1}^{q} \frac{1}{\ell!}\nabla_x^\ell f(x_{k+1})[d]^\ell + \sum_{\ell=1}^{q} \frac{1}{\ell!}\nabla_s^\ell T_p(x_k,s_k)[d]^\ell - \sum_{\ell=1}^{q} \frac{1}{\ell!}\nabla_s^\ell T_p(x_k,s_k)[d]^\ell \\
&\qquad - \frac{\sigma_k}{(p+\beta)!}\sum_{\ell=1}^{q} \frac{1}{\ell!}\left(\nabla_s^\ell\left[\|s\|^{p+\beta}\right](s_k)\right)[d]^\ell + \frac{\sigma_k}{(p+\beta)!}\sum_{\ell=1}^{q} \frac{1}{\ell!}\left(\nabla_s^\ell\left[\|s\|^{p+\beta}\right](s_k)\right)[d]^\ell \\
&\leq \left|\sum_{\ell=1}^{q} \frac{1}{\ell!}\left[\nabla_x^\ell f(x_{k+1}) - \nabla_s^\ell T_p(x_k,s_k)\right][d]^\ell\right| \\
&\qquad - \sum_{\ell=1}^{q} \frac{1}{\ell!}\left(\nabla_s^\ell\left[T_p(x_k,s) + \frac{\sigma_k}{(p+\beta)!}\|s\|^{p+\beta}\right]_{s=s_k}\right)[d]^\ell \\
&\qquad + \frac{\sigma_k}{(p+\beta)!}\left|\sum_{\ell=1}^{q} \frac{1}{\ell!}\left(\nabla_s^\ell\left[\|s\|^{p+\beta}\right]_{s=s_k}\right)[d]^\ell\right| \\
&\quad \leq \sum_{\ell=1}^{q} \frac{L}{\ell!(p-\ell+\beta)!}\|s_k\|^{p-\ell+\beta}\delta_{k+1}^\ell \\
&\qquad - \sum_{\ell=1}^{q} \frac{1}{\ell!}\nabla_s^\ell m_k(s_k)[d]^\ell + \sum_{\ell=1}^{q} \frac{\sigma_k}{\ell!(p-\ell+\beta)!}\|s_k\|^{p-\ell+\beta}\delta_{k+1}^\ell.
\end{aligned}$$

Now, since $\|d\| \leq \delta_{k+1}$, and using (2.6) for m_k at s_k,

$$-\sum_{\ell=1}^{q} \frac{1}{\ell!}\nabla_s^\ell m_k(s_k)[d]^\ell \leq \max\left[0, -\sum_{\ell=1}^{q} \frac{1}{\ell!}\nabla_s^\ell m_k(s_k)[d]^\ell\right] \leq \phi_{m_k,q}^{\delta_{k+1}}(s_k).$$

Using the fact that iteration k is successful, and thus $\delta_{k+1} = \delta_s$, we obtain, from (2.20) and (3.6), that

$$\begin{aligned}(3.7)\qquad \phi_{f,q}^{\delta_{k+1}}(x_{k+1}) &\leq \sum_{\ell=1}^{q} \frac{L}{\ell!(p-\ell+\beta)!}\|s_k\|^{p-\ell+\beta}\delta_{k+1}^{\ell} + \frac{\theta\,\chi_q(\delta_{k+1})}{(p-q+\beta)!}\|s_k\|^{p-q+\beta} \\ &\quad + \sum_{\ell=1}^{q} \frac{\sigma_k}{\ell!(p-\ell+\beta)!}\|s_k\|^{p-\ell+\beta}\delta_{k+1}^{\ell} \\ &\leq \frac{\Big[L+\sigma_k+\theta\Big]\chi_q(\delta_{k+1})}{(p-q+\beta)!}\|s_k\|^{p-q+\beta},\end{aligned}$$

where we have used the fact that $\|s_k\| \leq \varpi\epsilon^{\frac{1}{p-q+\beta}} \leq 1$ to deduce the last inequality. As a consequence, (3.5) implies that

$$\|s_k\| \geq \left[\frac{\epsilon(p-q+\beta)!}{(L+\sigma_k+\theta)}\right]^{\frac{1}{p-q+\beta}},$$

and (3.3) then immediately follows from (3.2). ∎

The bound given in the above lemma is another indication that choosing θ of the order of L (when this is known a priori) makes sense. Observe also that the statement of the above lemma is completely independent of δ_{k+1}.

We now combine all the above results to deduce an upper bound on the maximum number of successful iterations, from which a final complexity bound immediately follows.

THEOREM 3.4. *Let $f \in \mathcal{C}^{p,\beta}(\mathbb{R}^n)$, and suppose that $f(x) \geq f_{\rm low}$ for all $x \in \mathbb{R}^n$. Then, given $\epsilon \in (0,1)$, Algorithm* 2.1 *needs at most*

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

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

$$(3.8)\qquad \left\lfloor \left\lfloor \kappa_p(f(x_0)-f_{\rm low})\left(\epsilon^{-\frac{p+\beta}{p-q+\beta}}\right)+1\right\rfloor \left(1+\frac{|\log\gamma_1|}{\log\gamma_2}\right)+\frac{1}{\log\gamma_2}\log\left(\frac{\sigma_{\max}}{\sigma_0}\right)\right\rfloor$$

iterations in total to produce an iterate x_ϵ such that (2.12) *holds, where $\sigma_{\max}$ is given by* (3.2) *and where*

$$\kappa_p \stackrel{\rm def}{=} \frac{(p+\beta)!}{\eta_1\sigma_{\min}}\max\left\{\varpi^{-(p+\beta)}, \left[\frac{(L+\sigma_{\max}+\theta)}{(p-q+\beta)!}\right]^{\frac{p+\beta}{p-q+\beta}}\right\}.$$

Proof. At each successful iteration k before termination, we have the guaranteed decrease

$$(3.9)\qquad f(x_k)-f(x_{k+1}) \geq \eta_1(T_p(x_k,0)-T_p(x_k,s_k)) \geq \frac{\eta_1\sigma_{\min}}{(p+\beta)!}\,\|s_k\|^{p+\beta},$$

where we used (2.21), (3.1), and (2.22). Moreover we deduce from (3.9), (3.3), and (3.2) that

$$(3.10)\qquad f(x_k)-f(x_{k+1}) \geq \kappa_p^{-1}\epsilon^{\frac{p+\beta}{p-q+\beta}}, \quad \text{where } \kappa_p^{-1} \stackrel{\rm def}{=} \frac{\eta_1\sigma_{\min}\kappa_s^{p+\beta}}{(p+\beta)!}.$$

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

$$f(x_0) - f(x_{k+1}) \geq \kappa_p^{-1} \epsilon^{\frac{p+\beta}{p-q+\beta}} |\mathcal{S}_k|.$$

Using that f is bounded below by $f_{\rm low}$, we conclude

$$|\mathcal{S}_k| \leq \frac{f(x_0) - f_{\rm low}}{\kappa_p^{-1}} \epsilon^{-\frac{p+\beta}{p-q+\beta}} \tag{3.11}$$

until termination. The desired bound on the number of successful iterations follows from combining (3.11). Lemma 2.3 is then invoked to compute the upper bound on the total number of iterations. ∎

In particular, if the pth derivative of f is assumed to be globally Lipschitz rather than merely Hölder continuous (i.e., if $\beta = 1$), the bound (3.8) on the maximum number of evaluations becomes

$$\left\lfloor \left\lfloor \kappa_p (f(x_0) - f_{\rm low}) \left(\epsilon^{-\frac{p+1}{p-q+1}} \right) + 1 \right\rfloor \left(1 + \frac{|\log \gamma_1|}{\log \gamma_2} \right) + \frac{1}{\log \gamma_2} \log \left(\frac{\sigma_{\max}}{\sigma_0} \right) \right\rfloor, \tag{3.12}$$

where

$$\kappa_p \stackrel{\rm def}{=} \frac{(p+1)!}{\eta_1 \sigma_{\min}} \max \left\{ \varpi^{p+\beta}, \left[\frac{q!(L + \sigma_{\max} + \theta)(e-1)}{(p-q+1)!} \right]^{\frac{p+1}{p-q+1}} \right\}.$$

This worst-case evaluation bound generalizes known bounds for $q = 1$ (see [2]) or $q = 2$ (see [9]) and significantly improves upon the bounds in $O(\epsilon^{-(q+1)})$ given by [12] for a more stringent termination rule. It also extends the results obtained in [7] for convexly constrained problems with $q = 1$ by allowing the significantly broader class of inexpensive constraints.

We also note that it is possible to weaken the assumption that $\nabla_x^p f$ must satisfy the Hölder inequality (2.2) for every $x, y \in \mathbb{R}^n$ (as required in the beginning of section 2). The weakest possible smoothness assumption is to require that (2.2) holds only for points belonging to the same segment of the "tree of iterates" $\cup_{k \geq 0} [x_k, x_k + s_k]$ (this is necessary for the proof of Lemma 2.1). As this path joining feasible iterates may be hard to predict a priori, one may instead require (2.2) to hold in the whole of $\mathcal{F}$, which must then be convex to ensure the desired Hölder property on every segment $[x_k, x_k + s_k]$.

4. Seeking ϵ-approximate second-order-necessary minimizers. We now discuss the particular and much-studied case where second-order minimizers are sought for unconstrained problems with Lipschitz continuous Hessians (that is, $p \geq q = 2$, $\mathcal{F} = \mathbb{R}^n$, and $\beta = 1$). As we now show, a specialization of Algorithm 2.1 to this case is very close (but not identical) to well-known methods. Let us consider Step 1 first. The computation of $\phi_{f,2}^{\delta_k}(x_k)$ then reduces to

$$\phi_{f,2}^{\delta_k}(x_k) = \max \left[0, - \operatorname*{globmin}_{\|d\| \leq \delta_k} \left(\nabla_x^1 f(x_k)^T d + \frac{1}{2} d^T \nabla_x^2 f(x_k) d \right) \right], \tag{4.1}$$

which amounts to solving a standard trust-region subproblem with radius δ_k (see [14]). Hence verifying (4.1) or testing the more usual approximate second-order criteria

$$\|\nabla_x^1 f(x_k)\| \leq \epsilon \ \text{ and } \ \lambda_{\min} \left(\nabla_x^2 f(x_k) \right) \geq -\epsilon \tag{4.2}$$

have very similar numerical costs (remember that finding the leftmost eigenvalue of the Hessian is the same as finding the global minimizer of the associated Rayleigh quotient). If we now turn to the computation of s_k in Step 2, Algorithm 2.1 then computes such a step by attempting to minimize the model

$$T_p(x_k, s) + \frac{\sigma_k}{(p+1)!}\|s\|^{p+1}, \tag{4.3}$$

as has already been proposed before for general p [2, 9]. Moreover, the failure of (2.12) in Step 1 is enough, when $q \leq 2$, to guarantee the existence of nonzero global minimizers of $T_p(x_k, s)$ and $m_k(s)$ and thus to ensure that a nonzero s_k is possible. The approximate model minimization is stopped as soon as (2.19) or (2.20) holds, the latter then reducing to checking that

$$\phi^{\delta_s}_{m_k,2}(s_k) = \max\left[0, -\operatorname*{globmin}_{\|d\|\leq\delta_s}\left(\nabla^1_s m_k(s_k)^T d + \frac{1}{2}d^T\nabla^2_s m_k(s_k)d\right)\right] \leq \frac{\theta\|s_k\|^{p-1}}{(p-1)!}\,\chi_2(\delta_s) \tag{4.4}$$

for some $\delta_s \in (0, 1]$. For each potential s_k, finding $\delta_s \in (0, 1]$ requires solving (possibly approximately)

$$-\operatorname*{globmin}_{\|d\|\leq\delta_s}\left(\nabla^1_s m_k(s_k)^T d + \frac{1}{2}d^T\nabla^2_s m_k(s_k)d\right) \leq \frac{\theta\|s_k\|^{p-1}}{(p-1)!}\,\chi_2(\delta_s).$$

While this could be acceptable without affecting the overall evaluation complexity of the algorithm, a simpler alternative is available for $q = 2$. We may consider terminating the model minimization when either (2.19) holds or

$$0 > \operatorname*{globmin}_{\|d\|\leq 1}\left(\nabla^1_s m_k(s_k)^T d + \frac{1}{2}d^T\nabla^2_s m_k(s_k)d\right) \geq -\frac{\theta\|s_k\|^{p-1}}{(p-1)!}\,\chi_2(1) = -\frac{3\theta\|s_k\|^{p-1}}{2(p-1)!}. \tag{4.5}$$

The inequality is guaranteed to hold when s_k is close enough to s^*_k, a global minimizer of the model $m_k(s)$, since then $\nabla^1_s m_k(s^*_k) = 0$ and $\nabla^2_s m_k(s^*_k)$ is positive semidefinite, and then $d = 0$ provides the global minimizer of the second-order Taylor model of $m_k(s)$ around s_k. Verifying (4.5) only requires at most one trust-region calculation for each potential step and ensures (4.4) with $\delta_s = 1$, making the choice $\delta_{k+1} = 1$ acceptable. The cost of this technique is comparable to that proposed in [9] where an eigenvalue computation is required for each potential step. Combining these observations, Algorithm 2.1 then becomes Algorithm 4.1.

If $p = q = 2$, computing s_k in Step 2 amounts to approximately minimizing the now well-known cubic model of [17, 21, 24, 5]. In addition, if s_k is the exact global minimizer of this model, the above argument shows that (4.5) automatically holds at s_k, and checking this inequality by solving a trust-region subproblem is thus unnecessary. The only difference between our proposed algorithm and the more usual cubic regularization (ARC) method with exact global minimization is that the latter would check (4.2) for termination, while the algorithm presented here would instead check (4.1) with $\delta_k = 1$ by solving a trust-region subproblem. As observed above, both techniques have comparable numerical cost.

The bound (3.12) then ensures that Algorithm 4.1 terminates in at most $O(\epsilon^{-\frac{p+1}{p-1}})$ evaluations of f, its gradient and Hessian. This algorithm thus shares[8] the upper

[8]For a marginally weaker (see footnote 7 and Theorem 2.2) but still necessary and, in our view, more sensible approximate optimality condition.

Algorithm 4.1. ARp for ϵ-approximate second-order-necessary minimizers.

Step 0: Initialization. An initial point $x_0 \in \mathcal{F}$ and an initial regularization parameter $\sigma_0 > 0$ are given, as well as an accuracy level $\epsilon \in (0,1)$. The constants ϖ, θ, η_1, η_2, γ_1, γ_2, γ_3, and $\sigma_{\min}$ are also given and satisfy (2.17). Compute $f(x_0)$ and set $k = 0$.

Step 1: Test for termination. Evaluate $\{\nabla_x^i f(x_k)\}_{i=1}^2$. If $\phi_{f,2}^1(x_k) \leq \epsilon\chi_2(1)$, with $\phi_{f,2}^1(x_k)$ given by (4.1) and $\chi_2(1)$ by (2.13), terminate with the approximate solution $x_\epsilon = x_k$. Otherwise compute $\{\nabla_x^i f(x_k)\}_{i=3}^p$.

Step 2: Step calculation. Compute a step $s_k \neq 0$ by approximately minimizing the model (4.3) in the sense that (2.18) holds and

$$\|s_k\| \geq \varpi\epsilon^{\frac{1}{p-2+\beta}} \quad \text{or (4.5) holds.}$$

Step 3: Acceptance of the trial point. Compute $f(x_k + s_k)$ and define ρ_k as in (2.21). If $\rho_k \geq \eta_1$, then define $x_{k+1} = x_k + s_k$; otherwise define $x_{k+1} = x_k$.

Step 4: Regularization parameter update. Compute σ_{k+1} as in (2.22). Increment k by one and go to Step 1 if $\rho_k \geq \eta_1$, or to Step 2 otherwise.

complexity bounds stated in [9] for general p with different values of ϵ for first and second order, and in [21, 5] for $p = 2$.

5. A matching lower bound on the evaluation complexity for the Lipschitz continuous case. We now intend to show that the upper bound on evaluation complexity of Theorem 3.4 is tight in terms of the order given for unconstrained and a broad class of constrained problems with Lipschitz continuous pth derivative (i.e., $\beta = 1$[9]). This objective is attained by defining a variant of the high-degree Hermite interpolation technique developed in [12] and then using this technique to build, for any number p of available derivatives of the objective function and any optimality order q, an unconstrained univariate example of suitably slow convergence (i.e., for which the order in ϵ given by (3.12) is achieved). This example is then embedded in higher dimensions to provide general lower bounds.

5.1. High-degree univariate Hermite interpolation. We start by investigating some useful properties of Hermite interpolation. Let us assume that we wish to construct a univariate Hermite interpolant π of degree $2(p+1)$ of the form

$$\pi(\tau) = \sum_{i=0}^{2p+1} c_i \, \tau^i \tag{5.1}$$

on the interval $[0, s]$ satisfying the $2(p+1)$ conditions

$$\pi^{(i)}(0) = f_0^{(i)}, \quad \pi^{(i)}(s) = f_1^{(i)} \ \text{ for } \ i \in \{0, \ldots, p\}, \tag{5.2}$$

where $f_0^{(i)}$ and $f_1^{(i)}$ are given. The values of the coefficients $c_0, \ldots, c_p$ may then be obtained by

$$c_i = \frac{f_0^{(i)}}{i!} \ \text{ for } \ i \in \{0, \ldots, p\}$$

[9] A example of slow convergence for general β and $p > 1 + \beta$ is provided in [10].

while the remaining ones satisfy the linear system

(5.3)
$$\begin{pmatrix} a_{0,0}s^{p+1} & a_{0,1}s^{p+2} & \cdots & a_{0,p-1}s^{2p} & a_{1,p}s^{2p+1} \\ a_{1,0}s^{p} & a_{2,2}s^{p+1} & \cdots & a_{2,p-1}s^{2p-1} & a_{2,p}s^{2p} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{p,0}s & a_{p,1}s^{2} & \cdots & a_{p,p-1}s^{p} & a_{p,p}s^{p+1} \end{pmatrix} \begin{pmatrix} c_{p+1} \\ c_{p+2} \\ \vdots \\ c_{2p+1} \end{pmatrix} = \begin{pmatrix} f_1^{(0)} - T_p^{(0)}(0,s) \\ f_1^{(1)} - T_p^{(1)}(0,s) \\ \vdots \\ f_1^{(p)} - T_p^{(p)}(0,s)) \end{pmatrix},$$

where

$$T_p(0,s) = \sum_{i=0}^{p} \frac{f_0^{(i)}}{i!}\, s^i \quad \text{and} \quad a_{i,j} = \frac{(p+j+1)!}{(p+j+1-i)!} \quad (i,j = 0,\ldots,p).$$

Observe that (5.3) can be rewritten as

$$\begin{pmatrix} s^p & & & \\ & s^{p-1} & & \\ & & \ddots & \\ & & & 1 \end{pmatrix} A_p \begin{pmatrix} s & & & \\ & s^2 & & \\ & & \ddots & \\ & & & s^{p+1} \end{pmatrix} \begin{pmatrix} c_{p+1} \\ c_{p+2} \\ \vdots \\ c_{2p+1} \end{pmatrix} = \begin{pmatrix} f_1^{(0)} - T_p^{(0)}(0,s) \\ f_1^{(1)} - T_p^{(1)}(0,s) \\ \vdots \\ f_1^{(p)} - T_p^{(p)}(0,s)) \end{pmatrix}$$

with A_p is the matrix whose (i,j)th entry is $a_{i,j}$, which only depends on p. It was show in [12, Appendix] that A_p is nonsingular. Therefore

$$\begin{pmatrix} c_{p+1}\, s \\ c_{p+2}\, s^2 \\ \vdots \\ c_{2p+1}\, s^{p+1} \end{pmatrix} = A_p^{-1} \begin{pmatrix} \frac{1}{s^p}\left[f_1^{(0)} - T_p^{(0)}(0,s)\right] \\ \frac{1}{s^{p-1}}\left[f_1^{(1)} - T_p^{(1)}(0,s)\right] \\ \vdots \\ f_1^{(p)} - T_p^{(p)}(0,s) \end{pmatrix}.$$

We therefore deduce that, for any $\tau \in [0,s]$,

$$\begin{aligned} |\pi^{(p+1)}(\tau)| &= \left| \sum_{i=0}^{p} \frac{(p+1+i)!}{i!} c_{p+1+i}\, \tau^i \right| \\ &\leq \sum_{i=0}^{p} \frac{(p+1+i)!}{i!} \left(|c_{p+1+i}|\, s^{i+1} \right) s^{-1} \\ &\leq \frac{(p+1)(2p+1)!}{p!} \|A_p^{-1}\|_\infty \max_{j=0,\ldots,p} \left| \frac{f_1^{(j)} - T_p^{(j)}(0,s)}{s^{p-j+1}} \right|. \end{aligned}$$

The mean-value theorem then implies that, for any $0 \leq \tau_2 \leq \tau_1 \leq s$ and some $\xi \in [\tau_2, \tau_1] \subseteq [0,s]$,

$$\begin{aligned} \frac{|\pi^{(p)}(\tau_1) - \pi^{(p)}(\tau_2)|}{|\tau_1 - \tau_2|} &= |\pi^{(p+1)}(\xi)| \\ &\leq \max_{\tau \in [0,s]} |\pi^{(p+1)}(\tau)| \\ &\leq \frac{(p+1)(2p+1)!}{p!} \|A_p^{-1}\|_\infty \max_{j=0,\ldots,p} \left| \frac{f_1^{(j)} - T_p^{(j)}(0,s)}{s^{p-j+1}} \right|. \end{aligned} \tag{5.4}$$

This development thus leads us to the following conclusion.

THEOREM 5.1. *Suppose that* $\{f_\ell^{(j)}\}$ *are given for* $\ell \in \{0,1\}$ *and* $j \in \{0,\ldots,p\}$. *Suppose also that there exists a constant* $\kappa_f \geq 0$ *such that, for all* $j \in \{0,,\ldots,p\}$,

$$\left|f_1^{(j)} - T_p^{(j)}(0,s)\right| \leq \kappa_f\, s^{p-j+1}. \tag{5.5}$$

Then the Hermite interpolation polynomial $\pi(\tau)$ *on* $[0,s]$ *given by* (5.1) *and satisfying* (5.2) *admits a Lipschitz continuous pth derivative on* $[0,s]$, *with Lipschitz constant given by*

$$L_p \stackrel{\rm def}{=} \frac{(p+1)(2p+1)!}{p!}\|A_p^{-1}\|_\infty \kappa_f,$$

which only depends on p *and* κ_f.

Proof. The proof directly results from (5.4) and (5.5). ∎

Observe that (5.5) is identical to (2.5) when $\beta = 1$ and $n = 1$. This means that the conditions of Theorem 5.1 automatically hold if the interpolation data $\{f_i^{(j)}\}$ is itself extracted from a function having a Lipschitz continuous pth derivative.

Applying the above results to several interpolation intervals then yields the existence of a smooth Hermite interpolant.

THEOREM 5.2. *Suppose that, for some integer* $k_e > 0$ *and* $p > 0$, *the data* $\{f_k^{(j)}\}$ *and* $\{x_k\}$ *is given for* $k \in \{0,\ldots,k_e\}$ *and* $j \in \{0,\ldots,p\}$. *Suppose also that* $s_k = x_{k+1} - x_k \in (0,\kappa_s]$ *for* $k \in \{0,\ldots,k_e\}$ *and some* $\kappa_s > 0$, *and that, for some constant* $\kappa_f \geq 0$ *and* $k \in \{0,\ldots,k_e-1\}$,

$$\left|f_{k+1}^{(j)} - T_{k,p}^{(j)}(x_k,s_k)\right| \leq \kappa_f\, s_k^{p-j+1}, \tag{5.6}$$

where $T_{k,p}(x_k,s) = \sum_{i=0}^p f_k^{(i)}\, s^i/i!$. *Then there exists a* p *times continuously differentiable function* f *from* $\mathbb{R}$ *to* $\mathbb{R}$ *with Lipschitz continuous pth derivative such that, for* $k \in \{0,\ldots,k_e\}$,

$$f^{(j)}(x_k) = f_k^{(j)} \;\; \textit{for} \;\; j \in \{0,\ldots,p\}.$$

Moreover, the range of f *only depends on* p, κ_f, $\max_k f_k^{(0)}$, *and* $\min_k f_k^{(0)}$.

Proof. We first use Theorem 5.1 to define a Hermite interpolant $\pi_k(s)$ of the form (5.1) on each interval $[x_k,x_{k+1}] = [x_k,x_k+s_k]$ ($k \in \{0,\ldots,k_e\}$) using $f_0^{(j)} = f_k^{(j)}$ and $f_1^{(j)} = f_{k+1}^{(j)}$ for $j \in \{0,\ldots,p\}$, and then set

$$f(x_k+s) = \pi_k(s)$$

for any $s \in [0,s_k]$. We may then smoothly prolongate f for $x \in \mathbb{R}$ by defining two additional interpolation intervals $[x_{-1},x_0] = [-s_{-1},0]$ and $[x_{k_e},x_{k_e}+s_{k_e}]$ with end conditions

$$f_{-1} = f_0^{(0)}, \quad f_{k_e+1} = f_{k_e}^{(0)} \;\text{ and }\; f_{-1}^{(j)} = f_{k_e+1}^{(j)} = 0 \;\text{ for }\; j \in \{1,\ldots,p\},$$

and where s_{-1} and s_{k_e} are chosen sufficiently large to ensure that (5.6) also holds on intervals -1 and k_e. We next set

$$f(x) = \begin{cases} f_0^{(0)} & \text{for} \quad x \leq x_{-1}, \\ \pi_k(x-x_k) & \text{for} \quad x \in [x_k,x_{k+1}] \;\text{ and }\; k \in \{-1,\ldots,n\}, \\ f_{k_e}^{(0)} & \text{for} \quad x \geq x_{k_e}+s_{k_e}. \end{cases}$$

∎

5.2. Slow convergence to (ϵ,δ)-approximate qth-order-necessary minimizers. We now consider an unconstrained univariate instance of problem (2.1). Our aim is first to show that, for each choice of $p \geq 1$ and $q \in \{1, \ldots, p\}$, there exists an objective function $f \in C^{p,1}(\mathbb{R})$ (i.e., $\beta = 1$) for problem (2.1) which is bounded below and such that obtaining an (ϵ, δ)-approximate qth-order-necessary minimizer may require at least

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

evaluations of the objective function and its derivatives using Algorithm 2.1, matching, in order of $\epsilon \in (0, 1]$, the upper bound (3.12). Our development follows the broad outline of [13] but extends it to approximate minimizers of arbitrary order. Given a model degree $p \geq 1$ and an optimality order $q \in \{1, \ldots, p\}$, we first define the sequences $\{f_k^{(j)}\}$ for $j \in \{0, \ldots, p\}$ and $k \in \{0, \ldots, k_\epsilon\}$ with

$$k_\epsilon = \left\lceil \epsilon^{-\frac{p+1}{p-q+1}} \right\rceil \tag{5.7}$$

by

$$\omega_k = \epsilon \, \frac{k_\epsilon - k}{k_\epsilon}, \tag{5.8}$$

as well as

$$f_k^{(j)} = 0 \;\; \text{for} \;\; j \in \{1, \ldots, q-1\} \cup \{q+1, \ldots, p\} \tag{5.9}$$

and

$$f_k^{(q)} = -(\epsilon + \omega_k) \, q! \, \chi_q(1) < 0. \tag{5.10}$$

Thus

$$T_p(x_k, s) = \sum_{j=0}^{p} \frac{f_k^{(j)}}{j!} s^j = f_k^{(0)} - (\epsilon + \omega_k) \chi_q(1) s^q \tag{5.11}$$

and, assuming $\delta_k = 1$ for all k (we verify below that this is acceptable),

$$\phi_{f,q}^{\delta_k}(x_k) = (\epsilon + \omega_k) \chi_q(\delta_k). \tag{5.12}$$

We also set $\sigma_k = p!$ for all $k \in \{0, \ldots, k_\epsilon\}$ (we again verify below that is acceptable). Note that

$$\omega_k \in (0, \epsilon] \quad \text{and} \quad \phi_{f,q}^{\delta_k}(x_k) > \epsilon \chi_q(\delta_k) \;\; \text{for} \;\; k \in \{0, \ldots, k_\epsilon - 1\} \tag{5.13}$$

(and (2.12) fails at x_k), while

$$\omega_{k_\epsilon} = 0 \quad \text{and} \quad \phi_{f,q}^{\delta_k}(x_{k_\epsilon}) = \epsilon \chi_q(\delta_k) \tag{5.14}$$

(and (2.12) holds at x_{k_ϵ}). It is easy to verify using (5.11) that the model (2.16) is then globally minimized for

$$s_k = \left[\frac{|f_k^{(q)}|}{(q-1)!} \right]^{\frac{1}{p-q+1}} = [q(\epsilon + \omega_k)\chi_q(1)]^{\frac{1}{p-q+1}} > \epsilon^{\frac{1}{p-q+1}} \quad (k \in \{0, \ldots, k_\epsilon\}). \tag{5.15}$$

Hence this step satisfies (2.19) if we choose $\varpi = 1$. Because of this fact, we are free to choose δ_s arbitrarily in $(0, 1]$, and we choose $\delta_s = 1$. The step (5.15) yields that

$$m_k(s_k) = f_k^{(0)} - (\epsilon + \omega_k)\chi_q(1)[q(\epsilon + \omega_k)\chi_q(1)]^{\frac{q}{p-q+1}} + \frac{1}{p+1}[q(\epsilon + \omega_k)\chi_q(1)]^{\frac{p+1}{p-q+1}}$$

$$= f_k^{(0)} - \zeta(q,p)[q(\epsilon + \omega_k)\chi_q(1)]^{\frac{p+1}{p-q+1}}, \tag{5.16}$$

where

$$\zeta(q,p) \stackrel{\text{def}}{=} \frac{p-q+1}{q(p+1)} \in (0,1). \tag{5.17}$$

Thus $m_k(s_k) < m_k(0)$ and (2.18) holds. We then define

$$f_0^{(0)} = 2[2q\chi_q(1)]^{\frac{p+1}{p-q+1}} \quad \text{and} \quad f_{k+1}^{(0)} = f_k^{(0)} - \zeta(q,p)[q(\epsilon + \omega_k)\chi_q(1)]^{\frac{p+1}{p-q+1}}, \tag{5.18}$$

which provides the identity

$$m_k(s_k) = f_{k+1}^{(0)} \tag{5.19}$$

(ensuring that iteration k is successful because $\rho_k = 1$ in (2.21) and thus that our choice of a constant σ_k is acceptable and also that, provided we choose $\delta_0 = 1$ to ensure (5.12) for $k = 0$, the value $\delta_k = 1$ is admissible for all k). In addition, using (5.18), (5.13), (5.17), the inequality $k_\epsilon \leq 1 + \epsilon^{-\frac{p+1}{p-q+1}}$ from (5.7) gives that, for $k \in \{0, \ldots, k_\epsilon\}$,

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

and hence that

$$f_k^{(0)} \in \left[0, 2[2q\chi_q(1)]^{\frac{p+1}{p-q+1}}\right] \quad \text{for} \quad k \in \{0, \ldots, k_\epsilon\}. \tag{5.20}$$

We also set

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

Then (5.19) and (2.16) give that

$$\left| f_{k+1}^{(0)} - T_p(x_k, s_k) \right| = \frac{1}{p+1}|s_k|^{p+1}. \tag{5.21}$$

Now note that, using (5.11) and the first equality in (5.15),

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

where $I_{[\cdot]}$ is the standard indicator function. We may now verify that, for $j \in \{1, \ldots, q-1\}$,

(5.22)
$$\left|f_{k+1}^{(j)} - T_p^{(j)}(x_k, s_k)\right| = \left|0 - T_p^{(j)}(x_k, s_k)\right| \leq \left|\frac{(q-1)!}{(q-j)!}\right| |s_k|^{p-j+1} \leq (q-1)!\,|s_k|^{p-j+1},$$

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

$$\left|f_{k+1}^{(q)} - T_p^{(q)}(x_k, s_k)\right| = \left|-(q-1)!\, s_k^{p-q+1} + (q-1)!\, s_k^{p-q+1}\right| = 0 \tag{5.23}$$

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

$$\left|f_{k+1}^{(j)} - T_p^{(j)}(x_k, s_k)\right| = |0 - 0| = 0. \tag{5.24}$$

Combining (5.21), (5.22), (5.23), and (5.24), we deduce that (5.6) holds with $\kappa_f = (q-1)!$. We may thus apply Theorem 5.2 with $\beta = 1$, $\kappa_f = (q-1)!$, and $\kappa_s = 1$ and deduce the existence of a p times continuously differentiable function f from $\mathbb{R}$ to $\mathbb{R}$ with Lipschitz continuous derivatives of order 0 to p which interpolates the $\{f_k^{(j)}\}$ at $\{x_k\}$ for $k \in \{0, \ldots, k_\epsilon\}$ and $j \in \{0, \ldots, p\}$. Moreover, (5.20) and Theorem 5.2 imply that f is bounded below and that its range only depends on p and q. In addition, (5.19) ensures that every iteration is successful and thus, because of (2.22), that the value $\sigma_k = p!$ may be used at all iterations.

This argument allows us to state the following lower bound on the complexity of the regularization algorithm using a pth degree model.

LEMMA 5.3. *Given any $p \in \mathbb{N}_0$ and $q \in \{1, \ldots, p\}$, there exists a p times continuously differentiable function f from $\mathbb{R}$ to $\mathbb{R}$ with range only depending on p and q and Lipschitz continuous pth derivative such that, when the regularization algorithm with pth degree model (Algorithm 2.1) is applied to minimize f without constraints, it takes exactly*

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

iterations (and evaluations of the objective function and its derivatives) to find an (ϵ,δ)-approximate qth-order-necessary minimizer.

This implies the following important consequence for higher dimensional problems.

THEOREM 5.4. *Given any $n \in \mathbb{N}_0$, $p \in \mathbb{N}_0$, and $q \in \{1, \ldots, p\}$, there exists a p times continuously differentiable function f from $\mathbb{R}^n$ to $\mathbb{R}$ with range only depending on p and q and Lipschitz continuous pth derivative tensor such that, when the regularization algorithm with pth degree model (Algorithm 2.1) is applied to minimize f without constraints, it takes exactly*

$$k_\epsilon = \left\lceil \epsilon^{-\frac{p+1}{p-q+1}} \right\rceil \tag{5.25}$$

iterations (and evaluations of the objective function and its derivatives) to find an (ϵ, δ)-approximate qth-order-necessary minimizer. Furthermore, the same conclusion holds if the optimization problem under consideration involves constraints provided the feasible set $\mathcal{F}$ contains a ray.

Proof. The first conclusion directly follows from Lemma 5.3 since it is always possible to include the unimodal example as an independent component of a multivariate one.

The second conclusion follows from the observation that our univariate example of slow convergence is only defined on $\mathbb{R}^+$ (even if Theorem 5.2 provides an extension to the complete real line). As a consequence, it may be used on any feasible ray. ∎

We now make a few observations.

1. In the above example, we have restricted our attention to the Lipschitz continuous case where $\beta = 1$. It is possible that it could be extended to cover a more general choice of $\beta \in (0, 1]$: for example, [6] develops precisely such examples for second-order methods and Hölder continuous functions.
2. Theorem 5.4 generalizes to arbitrary values of q, the bound obtained in [3] for the case $q = 1$ and also shows that, at variance with the result derived in this reference, the generalized bound applies for arbitrary problem dimension but depends on ϵ, p, and q.
3. For simplicity, we have chosen, in the above example, to minimize the model $m_k(s)$ globally at every iteration, but we might consider other pairs (s, δ_s). A similar example of slow convergence may in fact be constructed along the lines used above[10] for any sequence of acceptable[11] model reducing steps and associated optimality radii (in the sense of Lemma 2.5), provided the optimality radii remain bounded away from zero. This means that *our example of slow convergence applies* not only to Algorithm 2.1 but also to a much broader class of minimization methods containing *all known methods using Taylor series that attempt to achieve approximate qth order criticality* as defined here; note that much in the definition of the example's function is independent of the algorithm and one could, for instance, replace regularization with trust region or linesearch (of course, the complexity would be worse for the latter two if one uses standard frameworks; see next paragraph). Moreover, it is also possible to weaken the constraints on the step further by relaxing (5.19) and only insisting on acceptable decrease of the objective function value in Step 3 of the algorithm.

   In [3], the authors derive their upper bound for $q = 1$ for the general class of "zero-preserving" algorithms, which are algorithms that "never explore (from x_k) coordinates which appear not to affect the function," that is, directions d along which $T_p(x_k, \cdot)$ is constant. This property is obviously shared by Algorithm 2.1 because it attempts to reduce the Taylor's expansion of f around the current iterate (the presence of the isotropic regularization term is irrelevant for this).
4. Our example does apply, for instance, to a linesearch method using univariate minimization along a descent search direction computed from the Taylor expansion of f, which is another zero-preserving method. Note, however, that such a method, just as every other linesearch method along descent directions (including possibly randomized coordinate searches), is bound to fail when attempting to compute approximate minimizers of order beyond three for problems whose dimension exceeds one,[12] because then the Taylor expansion at a nonoptimal point need no longer decrease along straight lines. This is demonstrated by the following old example [18, 22]. Let

[10] At the price of possibly larger constants.

[11] Remember that $\delta = 1$ is always possible for $q = 1$. It thus unsurprising that no such condition appears in [3].

[12] Linesearch methods remain relevant for unidimensional problems, obviously, which is why we have mentioned them in relation with our slow convergence example.

$$f(x_1, x_2) = \left(\frac{1}{2}x_1^2 - x_2\right)\left(x_1^2 - x_2\right).$$

Then $f(0,0) = 0$, and the origin is not a minimizer since f decreases along the arc $x_2 = \frac{3}{4}x_1^2$ (the origin is a saddle point in this case). Yet the origin is a strong local minimizer along every straight line passing through the origin, preventing any linesearch method based on descent directions from progressing from $(0,0)$.

Let us now consider an alternative unconstrained minimization method which would attempt to reduce the *unregularized* model (that is, (2.16) with $\sigma_k = 0$) in order to find an unconstrained first-order minimizer. It is easy to see that if one chooses

$$f_k^{(1)} = -(\epsilon + \omega_k), \quad f_k^{(i)} = 0 \text{ for } i \in \{2, \ldots, p-1\} \text{ and } f_k^{(p)} = p!,$$

the same reasoning as above yields that the largest obtainable decrease with this model occurs at

$$s_k = \left(\frac{\epsilon + \omega_k}{p}\right)^{\frac{1}{p-1}}$$

and is given by

$$f_k^{(0)} - m_k(s_k) = (p-1)\left(\frac{\epsilon + \omega_k}{p}\right)^{\frac{p}{p-1}}.$$

This then implies that at least a multiple of $\epsilon^{-\frac{p}{p-1}}$ evaluations may be needed to find approximate first-order-necessary minimizers, which is worse than the bound in $\epsilon^{-\frac{p+1}{p}}$ holding for the regularized algorithm. This is consistent with the known lower $O(\epsilon^{-2})$ bound for first-order points that holds for the (unregularized) Newton method (and hence the trust-region method), both of which use $p = 2$. Adding the regularization term thus not only provides a mechanism to limit the stepsize and make the step well-defined when $T_p(x_k, s)$ is unbounded below but also amounts to increasing the "useful degree" of the model by one, improving the worst-case complexity bound.

Summing up the above discussion, we conclude that an example of slow convergence requiring at least (5.25) evaluations can be built for any method whose steps decrease the regularized ($\sigma_k \geq \sigma_{\min}$) or unregularized ($\sigma_k = 0$) model (2.16) and whose approximate local optimality can be measured by (2.20) for some constant θ and $\delta_k = 1$ (which we can always enforce by adapting ϖ and (5.9)). For orders up to two, this includes most variants of steepest-descent and Newton's methods including those globalized with regularization, trust-region, a linesearch, or a mixture of these (see [13] for a discussion). General linesearch methods are excluded for high-order optimization as they may fail to converge to approximate minimizers of order four and beyond.

Finally, one may wonder at what would happen if, for the interpolation data (5.9)–(5.10), the model

$$m_k(s) = T_p(x_k, s) + \frac{\sigma_k}{m!}|s|^m$$

were used for some $m > p+1$, resulting in a shorter step. The global model minimizer would then occur at $s = [q(\epsilon + \omega_k)\chi_q(1)]^{1/(m-1)}$ and give an optimal model decrease equal to $[q(\epsilon + \omega_k)\chi_q(1)]^{m/(m-1)}(m-q)/m$. However, (5.6) would then fail for $j = 0$, and the argument leading to an example of slow convergence would break down.

6. Summary, further comments, and open questions. For any optimality order $q \geq 1$, we have provided the concept of an (ϵ, δ)-approximate qth-order-necessary minimizer for the very general set-constrained problem (2.1). We have then proposed a conceptual regularization algorithm to find such approximate minimizers and have shown that, if $\nabla_x^p f$ is β-Hölder continuous, this algorithm requires at most $O(\epsilon^{-\frac{p+\beta}{p-q+\beta}})$ evaluations of the objective function and its first p derivatives to terminate. When $\nabla_x^p f$ is Lipschitz continuous, we have used an unconstrained univariate version of the problem to show that this bound is sharp in terms of the order in ϵ for any feasible set containing a ray and any problem dimension.

In view of the results in [8, 15], one may wonder at what would happen if the regularization power (i.e., the power of $\|s\|$ used in the last term of the model (2.16)) were allowed to differ from $p+\beta$. The theory presented above must then be reexamined, and the crucial point is whether a global upper bound $\sigma_{\max}$ on the regularization parameter can still be ensured as in Lemma 3.2. One easily verifies that this is the case for regularization powers $r \in (p, p+\beta]$. Arguments parallel to those presented above then yield an upper bound of $O(\epsilon^{-\frac{r}{r-q}})$ evaluations,[13] recovering the bound given in section 3.3 of [8] for $q = 1$. The situation is, however, more complicated (and beyond the scope of the present paper) for $r > p+\beta$, and the determination of a suitable general complexity upper bound for this latter case has not been formalized at this stage, but the analysis for $q = 1$ discussed in section 3.2 of [8] suggests that an improvement of the bound for larger r is unlikely.

Although the results presented essentially solve the question of determining the optimal evaluation complexity for unconstrained problems and problems with general inexpensive constraints, some interesting issues remain open at this stage. A first such issue is whether an example of slow convergence for all $\epsilon \in (0,1)$ can be found for feasible domains not containing a ray. A second is to extend the general complexity theory for problems whose constraints are not inexpensive: the discussion in [11] indicates that this is a challenging research area.

Appendix A.

A.1. Proof of lemmas in section 2.

Proof of Lemma 2.1. We first establish the identity

(A.1)
$$I_{k-1,\beta} \stackrel{\rm def}{=} \int_0^1 \xi^\beta (1-\xi)^{k-1}\, d\xi = \frac{(k-1)!}{(k+\beta)!}, \quad \text{where } (k+\beta)! \stackrel{\rm def}{=} \prod_{i=1}^{k}(i+\beta) \text{ and } \beta! = 1.$$

If $k = 1$, the result directly follows from a simple integration. To see (A.1) for $k > 1$, integrating by parts, we have that

$$I_{k-1,\beta} = \left[\frac{\xi^{1+\beta}}{1+\beta}(1-\xi)^{k-1}\right]_0^1 + \frac{(k-1)}{(1+\beta)}\int_0^1 \xi^{1+\beta}(1-\xi)^{k-2}\, d\xi = \frac{(k-1)}{(1+\beta)} I_{k-2,1+\beta}$$

and thus, recursively, that

$$I_{k-1,\beta} = \frac{(k-1)!}{(k-1+\beta)!} I_{0,k-1+\beta} = \frac{(k-1)!}{(k-1+\beta)!}\int_0^1 \xi^{k-1+\beta}\, d\xi = \frac{(k-1)!}{(k+\beta)!}.$$

[13] We may even relax (2.20) slightly by replacing $\|s_k\|^{p-q+\beta}$ by $\|s_k\|^{r-q}$.

As in [12], consider the Taylor identity

$$\psi(1) - \tau_k(1) = \frac{1}{(k-1)!} \int_0^1 (1-\xi)^{k-1} [\psi^{(k)}(\xi) - \psi^{(k)}(0)] \, d\xi \tag{A.2}$$

involving a given univariate C^k function $\psi(t)$ and its kth order Taylor approximation

$$\tau_k(t) = \sum_{i=0}^{k} \psi^{(i)}(0) \frac{t^i}{i!}$$

expressed in terms of the value $\psi^{(0)} = \psi$ and ith derivatives $\psi^{(i)}$, $i = 1, \ldots, k$. Then, picking $\psi(t) = f(x + ts)$, for given $x, s \in \mathbb{R}^n$, and $k = p$, the identity (A.2) and the relationships $\psi^{(p)}(t) = \nabla_x^p f(x + ts)[s]^p$ and $\tau_p(1) = T_p(x, s)$ give that

$$f(x+s) - T_p(x,s) = \frac{1}{(p-1)!} \int_0^1 (1-\xi)^{k-1} \left(\nabla_x^p f(x + \xi s) - \nabla_x^p f(x) \right) [s]^p \, d\xi,$$

and thus from the definition of the tensor norm (1.1), the Hölder bound (2.2) and the identity (A.1) when $k = p$, we obtain that

$$\begin{aligned}
&f(x+s) - T_p(x,s) \\
&\quad \le \frac{1}{(p-1)!} \int_0^1 (1-\xi)^{k-1} \left| \left(\nabla_x^p f(x + \xi s) - \nabla_x^p f(x) \right) \left[\frac{s}{\|s\|} \right]^p \right| \|s\|^p \, d\xi \\
&\quad \le \frac{1}{(p-1)!} \int_0^1 (1-\xi)^{k-1} \max_{\|v\|=1} \left| \left(\nabla_x^p f(x + \xi s) - \nabla_x^p f(x) \right) [v]^p \right| \|s\|^p \, d\xi \\
&\quad = \frac{1}{(p-1)!} \int_0^1 (1-\xi)^{k-1} \|\nabla_x^p f(x + \xi s) - \nabla_x^p f(x)\|_{[p]} d\xi \cdot \|s\|^p \\
&\quad \le \frac{1}{(p-1)!} \int_0^1 \xi^\beta (1-\xi)^{p-1} \, d\xi \cdot L \|s\|^{p+\beta} = \frac{L}{(p+\beta)!} \|s\|^{p+\beta}
\end{aligned}$$

for all $x, s \in \mathbb{R}^n$, which is the required (2.4).

Likewise, for arbitrary unit vectors $v_1, \ldots, v_j$, choosing $\psi(t) = \nabla_x^j f(x+ts)[v_1, \ldots, v_j]$ and $k = p - j$, it follows from (A.2), the relationships $\psi^{(p-j)}(t) = \nabla_x^p f(x + ts)[v_1, \ldots, v_j][s]^{p-j}$, and $\tau_{p-j}(1) = \nabla_s^j T_p(x, s)$ that

$$\begin{aligned}
&(\nabla_x^j f(x+s) - \nabla_s^j T_p(x,s))[v_1, \ldots, v_j] \\
&\qquad = \frac{1}{(p-j-1)!} \int_0^1 (1-\xi)^{p-j-1} \left(\nabla_x^p f(x + \xi s) - \nabla_x^p f(x) \right) [v_1, \ldots, v_j][s]^{p-j} \, d\xi.
\end{aligned} \tag{A.3}$$

Then picking $v_1, \ldots, v_j$ to maximize the absolute value of left-hand side of (A.3) and using the tensor norm (1.1), the Hölder bound (2.2), and the identity (A.1) when $k = p - j$, we find that

$$\begin{aligned}
&\|\nabla_x^j f(x+s) - \nabla_s^j T_p(x,s)\|_{[j]} \\
&\quad \le \frac{1}{(p-j-1)!} \int_0^1 (1-\xi)^{p-j-1} \Bigg| (\nabla_x^p f(x + \xi s) \\
&\qquad\qquad - \nabla_x^p f(x))[v_1, \ldots, v_j] \left[\frac{s}{\|s\|} \right]^{p-j} \Bigg| \|s\|^{p-j} \, d\xi \\
&\quad \le \frac{1}{(p-j-1)!} \int_0^1 (1-\xi)^{p-j-1} \max_{\|v_1\| = \cdots = \|v_p\| = 1} \left| (\nabla_x^p f(x + \xi s) \right. \\
&\qquad\qquad \left. - \nabla_x^p f(x)) \, [v_1, \ldots, v_p] \right| \|s\|^{p-j} \, d\xi
\end{aligned}$$

$$= \frac{1}{(p-j-1)!}\int_0^1 (1-\xi)^{p-j-1}\|\nabla_x^p f(x+\xi s) - \nabla_x^p f(x)\|_{[p]}\, d\xi \cdot \|s\|^{p-j}$$
$$\leq \frac{1}{(p-j-1)!}\int_0^1 \xi^\beta (1-\xi)^{p-j-1}\, d\xi \cdot L\|s\|^{p-j+\beta} = \frac{L}{(p-j+\beta)!}\|s\|^{p-j+\beta}$$

for all $x, s \in \mathbb{R}^n$, which gives (2.5). ∎

Proof of Lemma 2.3. The regularization parameter update (2.22) gives that, for each k,

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

where $\mathcal{U}_k \stackrel{\text{def}}{=} \{0,\ldots,k\} \setminus \mathcal{S}_k$. Thus we deduce inductively that $\sigma_0\gamma_1^{|\mathcal{S}_k|}\gamma_2^{|\mathcal{U}_k|} \leq \sigma_k$. We therefore obtain, using (2.23), that

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

which then implies that

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

since $\gamma_2 > 1$. The desired result (2.24) then follows from the equality $k+1 = |\mathcal{S}_k|+|\mathcal{U}_k|$ and the inequality $\gamma_1 < 1$ given by (2.17). ∎

Proof of Lemma 2.4. We first observe that $\nabla_s^j\big(\|s\|^{p+\beta}\big)$ is a jth order tensor, whose norm is defined using (1.1). Moreover, using the relationships

$$\nabla_s\big(\|s\|^\tau\big) = \tau\,\|s\|^{\tau-2}s \quad \text{and} \quad \nabla_s\big(s^{\tau\otimes}\big) = \tau\, s^{(\tau-1)\otimes}\otimes I, \quad (\tau \in \mathbb{R}), \tag{A.4}$$

defining

$$\nu_0 \stackrel{\text{def}}{=} 1 \quad \text{and} \quad \nu_i \stackrel{\text{def}}{=} \prod_{\ell=1}^{i}(p+2-2\ell), \tag{A.5}$$

and proceeding by induction, we obtain that, for some $\mu_{j,i} \geq 0$ with $\mu_{1,1} = 1$,

$$\begin{aligned}
&\nabla_s\left[\nabla_s^{j-1}\big(\|s\|^{p+\beta}\big)\right] \\
&\quad = \nabla_s\left[\sum_{i=2}^{j}\mu_{j-1,i-1}\nu_{i-1}\|s\|^{p+\beta-2(i-1)}\, s^{(2(i-1)-(j-1))\otimes}\otimes I^{((j-1)-(i-1))\otimes}\right] \\
&\quad = \sum_{i=2}^{j}\mu_{j-1,i-1}\nu_{i-1}\Big[(p+\beta-2(i-1))\|s\|^{p+\beta-2(i-1)-2}\, s^{(2(i-1)-(j-1)+1)\otimes}\otimes I^{(j-i)\otimes} \\
&\qquad + ((2(i-1)-(j-1))\|s\|^{p+\beta-2(i-1)}\, s^{(2(i-1)-(j-1)-1)\otimes}\otimes I^{((j-1)-(i-1)+1)\otimes}\Big] \\
&\quad = \sum_{i=2}^{j}\mu_{j-1,i-1}\nu_{i-1}\Big[(p+\beta+2-2i)\|s\|^{p+\beta-2i}\, s^{(2i-j)\otimes}\otimes I^{(j-i)\otimes} \\
&\qquad + (2(i-1)-j+1)\|s\|^{p+\beta-2(i-1)}\, s^{(2(i-1)-j)\otimes}\otimes I^{(j-(i-1))\otimes}\Big]
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=2}^{j} \mu_{j-1,i-1}\nu_{i-1}(p+\beta+2-2i)\|s\|^{p+\beta-2i}\, s^{(2i-j)\otimes} \otimes I^{(j-i)\otimes} \\
&\quad + \sum_{i=1}^{j-1} (2i-j+1)\mu_{j-1,i}\nu_i \|s\|^{p+\beta-2i}\, s^{(2i-j)\otimes} \otimes I^{(j-i)\otimes} \\
&= \sum_{i=1}^{j} \big((p+\beta+2-2i)\mu_{j-1,i-1}\nu_{i-1} + (2i-j+1)\mu_{j-1,i}\nu_i\big) \|s\|^{p+\beta-2i}\, s^{(2i-j)\otimes} \otimes I^{(j-i)\otimes},
\end{aligned}
$$

where the last equation uses the convention that $\mu_{j,0} = 0$ for all j. Thus we may write

$$
\nabla_s^j\big(\|s\|^{p+\beta}\big) = \nabla_s\left[\nabla_s^{j-1}\big(\|s\|^{p+\beta}\big)\right] = \sum_{i=1}^{j} \mu_{j,i}\nu_i\, \|s\|^{p+\beta-2i}\, s^{(2i-j)\otimes} \otimes I^{(j-i)\otimes} \tag{A.6}
$$

with

$$
\begin{aligned}
\mu_{j,i}\nu_i &= (p+\beta+2-2i)\mu_{j-1,i-1}\nu_{i-1} + (2i-j+1)\mu_{j-1,i}\nu_i \\
&= \big[\mu_{j-1,i-1} + (2i-j+1)\mu_{j-1,i}\big]\nu_i,
\end{aligned} \tag{A.7}
$$

where we used the identity

$$
\nu_i = (p+\beta+2-2i)\nu_{i-1} \quad \text{for} \quad i = 1, \ldots, j \tag{A.8}
$$

to deduce the second equality. Now (A.6) gives that

$$
\nabla_s^j\big(\|s\|^{p+\beta}\big)[v]^j = \sum_{i=1}^{j} \mu_{j,i}\nu_i \|s\|^{p+\beta-j} \left(\frac{s^T v}{\|s\|}\right)^{2i-j} (v^T v)^{j-i}.
$$

It is then easy to see that the maximum in (1.1) is achieved for $v = s/\|s\|$, so that

$$
\|\, \nabla_s^j\big(\|s\|^{p+\beta}\big)\, \|_{[j]} = \left(\sum_{i=1}^{j} \mu_{j,i}\nu_i\right) \|s\|^{p+\beta-j} = \pi_j \|s\|^{p+\beta-j} \tag{A.9}
$$

with

$$
\pi_j \stackrel{\text{def}}{=} \sum_{i=1}^{j} \mu_{j,i}\, \nu_i. \tag{A.10}
$$

Successively using this definition, (A.7), (A.8) (twice), the identity $\mu_{j-1,j} = 0$, and (A.10) again, we then deduce that

$$
\begin{aligned}
\pi_j &= \sum_{i=1}^{j} \mu_{j-1,i-1}\nu_i + \sum_{i=1}^{j} (2i-j+1)\mu_{j-1,i}\nu_i \\
&= \sum_{i=1}^{j-1} \mu_{j-1,i}\nu_{i+1} + \sum_{i=1}^{j} (2i-j+1)\mu_{j-1,i}\nu_i \\
&= \sum_{i=1}^{j-1} \mu_{j-1,i}\big[\nu_{i+1} + (2i-j+1)\nu_i\big] \\
&= \sum_{i=1}^{j-1} \mu_{j-1,i}\big[(p+\beta+2-2(i+1))\nu_i + (2i-j+1)\nu_i\big]
\end{aligned} \tag{A.11}
$$

$$= (p+\beta+1-j)\sum_{i=1}^{j-1}\mu_{j-1,i}\,\nu_i$$
$$= (p+\beta+1-j)\pi_{j-1}.$$

Since $\pi_1 = p+\beta$ from the first part of (A.4), we obtain that $\pi_j = (p+\beta)!/(p-j+\beta)!$, which, combined with (A.9) and (A.10), gives (2.25). We obtain (2.26) from (A.9) and (A.10), the observation that $\pi_p = (p+\beta)!$, and (A.11) for $j = p+1$. ∎

A.2. Proof of lemmas in section 3.

Proof of Lemma 3.1 (see [2, Lemma 2.1]). Observe that, because of (2.18) and (2.16),

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

which implies the desired bound. Note that $s_k \neq 0$ as long as we can satisfy condition (2.18), and so (3.1) implies (2.21) is well defined. ∎

Proof of Lemma 3.2 (see [2, Lemma 2.2]). Assume that

$$\sigma_k \ge \frac{L}{1-\eta_2}. \tag{A.12}$$

Using (2.4) and (3.1), we may then deduce that

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

and thus that $\rho_k \ge \eta_2$. Then iteration k is very successful in that $\rho_k \ge \eta_2$ and $\sigma_{k+1} \le \sigma_k$. As a consequence, the mechanism of the algorithm ensures that (3.2) holds. ∎

REFERENCES

[1] A. Anandkumar and R. Ge, *Efficient approaches for escaping high-order saddle points in nonconvex optimization*, Proc. Mach. Learn. Soc., 49 (2016), pp. 81–102.

[2] 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), pp. 359–368.

[3] Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford, *Lower bounds for finding stationary points* I, Math. Program., in press.

[4] C. Cartis, N. I. M. Gould, and Ph. L. Toint, *On the complexity of steepest descent, Newton's and regularized Newton's methods for nonconvex unconstrained optimization*, SIAM J. Optim., 20 (2010), pp. 2833–2852.

[5] C. Cartis, N. I. M. Gould, and Ph. L. Toint, *Adaptive cubic overestimation methods for unconstrained optimization. Part* II*: worst-case function-evaluation complexity*, Math. Program. Ser. A, 130 (2011), pp. 295–319.

[6] C. Cartis, N. I. M. Gould, and Ph. L. Toint, *Optimal Newton-Type Methods for Nonconvex Smooth Optimization Problems*, ERGO Technical Report 11-009, School of Mathematics, University of Edinburgh, Edinburgh, UK, 2011.

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

[8] C. Cartis, N. I. M. Gould, and Ph. L. Toint, *Universal regularization methods – varying the power, the smoothness and the accuracy*, SIAM J. Optim., 29 (2019), pp. 595–615.

[9] C. Cartis, N. I. M. Gould, and Ph. L. Toint, *A concise second-order complexity analysis for unconstrained optimization using high-order regularized models*, Optim. Methods Softw., 35 (2020), pp. 1–14.

[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), pp. 1273–1298.

[11] C. Cartis, N. I. M. Gould, and Ph. L. Toint, *Optimality of orders one to three and beyond: Characterization and evaluation complexity in constrained nonconvex optimization*, J. Complexity, 53 (2019), pp. 68–94.

[12] 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), pp. 1073–1107.

[13] C. Cartis, N. I. M. Gould, and Ph. L. Toint, *Worst-case evaluation complexity and optimality of second-order methods for nonconvex smooth optimization*, in the Proceedings of the 2018 International Congress of Mathematicians (ICM 2018), Rio de Janeiro, 2018, pp. 3711–3750.

[14] A. R. Conn, N. I. M. Gould, and Ph. L. Toint, *Trust-Region Methods*, MPS-SIAM Ser. Optim., SIAM, Philadelphia, PA, 2000.

[15] G. N. Grapiglia and Yu. Nesterov, *Regularized Newton methods for minimizing functions with Hölder continuous Hessians*, SIAM J. Optim., 27 (2017), pp. 478–506.

[16] S. Gratton, A. Sartenaer, and Ph. L. Toint, *Recursive trust-region methods for multiscale nonlinear optimization*, SIAM J. Optim., 19 (2008), pp. 414–444.

[17] A. Griewank, *The Modification of Newton's Method for Unconstrained Optimization by Bounding Cubic Terms*, Technical Report NA/12, Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, UK, 1981.

[18] H. Hancock, *The Theory of Maxima and Minima*, The Athenaeum Press, Ginn & Co, New York, NY, 1917, https://archive.org/details/theoryofmaximami00hancroft.

[19] J. M. Martínez, *On higher-order model regularization for constrained optimization*, SIAM J. Optim., 27 (2017), pp. 2447–2458.

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

[21] Yu. Nesterov and B. T. Polyak, *Cubic regularization of Newton method and its global performance*, Math. Program. Ser. A, 108 (2006), pp. 177–205.

[22] G. Peano, *Calcolo differenziale e principii di calcolo integrale*, Fratelli Bocca, Roma, Italy, 1884.

[23] S. A. Vavasis, *Black-box complexity of local minimization*, SIAM J. Optim., 3 (1993), pp. 60–80.

[24] M. Weiser, P. Deuflhard, and B. Erdmann, *Affine conjugate adaptive Newton methods for nonlinear elastomechanics*, Optim. Methods Softw., 22 (2007), pp. 413–431.

[25] X. Zhang, C. Ling, and L. Qi, *The best rank-1 approximation of a symmetric tensor and related spherical optimization problems*, SIAM J. Matrix Anal. Appl., 33 (2012), pp. 806–821.