Optimization Methods and Software

ISSN: 1055-6788 (Print) 1029-4937 (Online) Journal homepage: http://www.tandfonline.com/loi/goms20

Worst-case evaluation complexity of regularization methods for smooth unconstrained optimization using Hölder continuous gradients

C. Cartis, N.I.M. Gould & Ph.L. Toint

To cite this article: C. Cartis, N.I.M. Gould & Ph.L. Toint (2017): Worst-case evaluation complexity of regularization methods for smooth unconstrained optimization using Hölder continuous gradients, Optimization Methods and Software, DOI: 10.1080/10556788.2016.1268136

To link to this article: http://dx.doi.org/10.1080/10556788.2016.1268136

Published online: 17 Apr 2017.

Submit your article to this journal

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
http://www.tandfonline.com/action/journalInformation?journalCode=goms20

Download by: [the Bodleian Libraries of the University of Oxford] **Date:** 18 April 2017, At: 01:40

Optimization Methods & Software, 2017
http://dx.doi.org/10.1080/10556788.2016.1268136

Worst-case evaluation complexity of regularization methods for smooth unconstrained optimization using Hölder continuous gradients

C. Cartis[a]*, N.I.M. Gould[b] and Ph.L. Toint[c]

[a]*Mathematical Institute, Oxford University, Oxford, UK;* [b]*Numerical Analysis Group, Rutherford Appleton Laboratory, Chilton, UK;* [c]*Namur Center for Complex Systems (naXys) and Department of Mathematics, University of Namur, Namur, Belgium*

(*Received 1 July 2015; accepted 30 November 2016*)

The worst-case behaviour of a general class of regularization algorithms is considered in the case where only objective function values and associated gradient vectors are evaluated. Upper bounds are derived on the number of such evaluations that are needed for the algorithm to produce an approximate first-order critical point whose accuracy is within a user-defined threshold. The analysis covers the entire range of meaningful powers in the regularization term as well as in the Hölder exponent for the gradient. The resulting complexity bounds vary according to the regularization power and the assumed Hölder exponent, recovering known results when available.

Keywords: nonlinear optimisation; regularization methods; complexity analysis

1. Introduction

The complexity analysis of algorithms for smooth, possibly nonconvex, unconstrained optimization has been the subject of a burgeoning literature over the past few years (see the contributions by Nesterov [15], Nesterov and Polyak [18], Gratton, Sartenaer and Toint [12], Cartis, Gould and Toint [3,5–7], Ueda [20], Ueda and Yamashita [21,22], Grapiglia, Yuan and Yuan [10,11], and Vicente [23], for instance). The present contribution belongs to this active trend and focuses on the analysis of the worst-case behaviour of regularization methods where only objective function values and associated gradient vectors are evaluated. It proposes upper bounds on the number of such evaluations that are needed for the algorithm to produce an approximate first-order critical point whose accuracy is within a user-defined threshold.

An analysis of this type is already available for the case where the objective function's gradient is assumed to be Lipschitz-continuous and where the regularization uses the second or third power of the norm of the computed step at a given iteration (see the paper by Nesterov [16] for the former and those of Cartis *et al.* [5,6] for both cases). The novelty of the present approach is to extend the analysis to cover problems whose objective gradients are simply Hölder continuous and methods that allow weaker regularization than in the Lipschitz case. The resulting

*Corresponding author. Email: coralia.cartis@maths.ox.ac.uk

© 2017 Informa UK Limited, trading as Taylor & Francis Group

complexity bounds vary according to the regularization power and the assumed Hölder exponent, providing a unified view and recovering known results when available.

We consider the problem of finding an approximate solution of the optimization problem

$$\min_x f(x), \tag{1}$$

where $x \in \Re^n$ is the vector of optimization variables and f is a function from $\Re^n$ into $\Re$ that is assumed to be bounded below and continuously differentiable with Hölder continuous gradients. If we denote $g(x) \stackrel{\text{def}}{=} \nabla_x f(x)$, the latter says that the inequality

$$\|g(x) - g(y)\| \leq L_\beta \|x - y\|^\beta \tag{2}$$

holds for all $x, y \in \Re^n$, where $L_\beta \geq 0$ and $\beta \geq 0$ are constants independent of x and y and where $\|\cdot\|$ is the Euclidean norm on $\Re^n$. Problems involving functions with Hölder continuous gradients are interesting in their own right, but can also be found in engineering practice, such as in the design of gas pipelines (the Panhandle law which governs such flows states that the gas flow rate in a pipeline is a power between 1 and 2 of the difference in squared pressures, see [19, Section 17], for instance). Such functions also appear in the solution of certain nonlinear PDE problems (see Bensoussan and Frehse [1]). In optimization, functions with Hölder continuous gradients are regarded as a bridging case between the smooth and nonsmooth problem classes [14,17]. In particular, the case $\beta = 0$ corresponds to possibly nonsmooth functions with bounded subgradients; when $\beta \in (0, 1)$, f is continuously differentiable but the Hessian may not exist; finally, $\beta = 1$ corresponds to f having Lipschitz continuous gradient and so its Hessian is guaranteed to exist, while if $\beta > 1$, the Hessian is zero and f is linear (see Lemma 3.1 below). For convex optimization, methods have already been devised, and their complexity analysed, for functions satisfying (2) both as a weaker set of assumptions and as an attempt to have a 'smooth' transition between the smooth and nonsmooth problem classes [8,17]; even lower complexity bounds are known [14]. For nonconvex optimization, a gradient method with linesearch is proposed and analysed in [24] when f satisfies (2) with $\beta > 0$, with restricted stepsize that requires a priori knowledge of problem parameters such as β. More recently, [13] analysed block-coordinate descent first-order methods for this class of functions.

In this paper, we consider a family of regularization methods that iteratively build, and approximately minimize, a local linear or quadratic model of f around the current iterate x_k, regularized by the rth power of the norm of the change to x_k, where $r > 1$. We apply these methods to (2) with $\beta > 0$, but the methods do not require and do not explicitly estimate β. We terminate when an approximate solution for problem (1) is found, which in our context, denotes a vector x_ϵ such that

$$\|g(x_\epsilon)\| \leq \epsilon \quad \text{or} \quad f(x) \leq f_{\text{target}}, \tag{3}$$

where $\epsilon > 0$ is a user-specified accuracy threshold and f_{target} is a threshold value – independent of ϵ – under which the reduction of the objective function is deemed sufficient by the user. The first case in (3) corresponds to finding an approximate first-order-critical point. If a suitable value for f_{target} is not known, minus infinity can be used instead, in effect making the second part of (3) impossible to satisfy and reducing this condition to its first part. Allowing a target value to be specified by the user on the value of f is an additional feature of our results (not a requirement, as explained above) and it is, to the best of our knowledge, novel in the context of complexity analysis; it attempts to give theoretical underpinnings for practical termination conditions. We show that the worst-case complexity of the resulting regularization methods when applied to (2) with $\beta \in (0, 1]$, varies depending on $\min\{r, 1 + \beta\}$. In particular, when $1 < r \leq 1 + \beta$, the methods take at most $O(\epsilon^{-r/(r-1)})$ evaluations/iterations to satisfy (3); and otherwise, at most $O(\epsilon^{-(1+\beta)/\beta})$

evaluations/iterations to achieve the same condition. The latter bound illustrates the 'ability' of the proposed methods to adapt to the smoothness of the landscape they are applied to, without prior knowledge of it.

The paper is organized as follows. Section 2 presents the class of algorithms considered. The complexity analysis itself is given in Section 3 and the sharpness of some of the obtained result is discussed in Section 4, with further details delegated to the Appendix. Section 5 finally provides some comments on the results.

Notations: In what follows, $\|\cdot\|$ denotes the Euclidean norm and the ^{T} superscript denotes transposition. If v is a vector in $\Re^n$, $[v]_i$ denotes its ith component.

2. The algorithm

The class of regularization methods that we consider for computing an x satisfying (3) consists of iterative algorithms where, at each iteration, a local (linear or quadratic) model of f around the current iterate x_k is constructed, regularized by a term using the rth power of the norm of the step, and then approximately minimized (in the 'Cauchy point' sense) to provide a trial step s_k. The quality of this step is then measured in order to accept the resulting trial point $x_k + s_k$ as the next iterate, or to reject it and adjust the strength of the regularization.

More specifically, a regularized model of $f(x_k + s)$ of the form

$$m_k(x_k + s) = f(x_k) + g_k^T s + \frac{1}{2} s^T B_k s + \frac{\sigma_k}{r}\|s\|^r \tag{4}$$

is considered around the kth iterate x_k, where we have defined $g_k \stackrel{\text{def}}{=} g(x_k)$, where B_k is a symmetric $n \times n$ matrix, where $\sigma_k > 0$ is the regularization parameter at iteration k and where $r > 1$ is the (iteration independent) user-defined regularization power. In practice, the matrix B_k may be chosen to provide suitable scaling of the variables (if known), for instance using quasi-Newton formulae. The model (4) is then approximately minimized in the sense that the trial step s_k is computed such that

$$m_k(x_k + s_k) \le m_k(x_k + {s_k}^{\mathrm{C}}), \tag{5}$$

where the 'Cauchy step' s_k^{C} is defined by

$${s_k}^{\mathrm{C}} = -{\alpha_k}^{\mathrm{C}} g_k \quad \text{with} \quad {\alpha_k}^{\mathrm{C}} = \arg\min_{\alpha \ge 0} m_k(x_k - \alpha g_k). \tag{6}$$

We will choose the regularization power r in (4) in order to guarantee that m_k is bounded below and grows at infinity, thereby ensuring that (6) is well-defined. In particular, this imposes the restriction $r > 1$ and furthermore

$$r > 2 \quad \text{whenever } B_k \text{ is allowed to not be positive semi-definite.} \tag{7}$$

Notice that (5) and (6) together imply that

$$m_k(x_k + s_k) \le m_k(x_k + {s_k}^{\mathrm{C}}) < f(x_k) \tag{8}$$

provided $g(x_k) \neq 0$. We may now describe our class of algorithms more formally as Algorithm 2.1.

ALGORITHM 2.1 (A class of first-order adaptive regularization methods)

Step 0: Initialization. *An initial point x_0, a target objective function value $f_{\text{target}} \leq f(x_0)$ and an initial regularization parameter $\sigma_0 > 0$ are given, as well as an accuracy level ϵ. The constants η_1, η_2, γ_1, γ_2 and γ_3 are also given and satisfy*

$$0 < \eta_1 \leq \eta_2 < 1 \quad \textit{and} \quad 0 < \gamma_1 < 1 < \gamma_2 < \gamma_3. \tag{9}$$

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

Step 1: Test for termination. *If $\|g_k\| \leq \epsilon$ or $f(x_k) \leq f_{\text{target}}$, terminate with the approximate solution $x_\epsilon = x_k$.*

Step 2: Step calculation. *Compute the step s_k approximately by minimizing the model* (4) *in the sense that conditions* (5) *and* (6) *hold.*

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

$$\rho_k = \frac{f(x_k) - f(x_k + s_k)}{m_k(x_k) - m_k(x_k + s_k)}. \tag{10}$$

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

Step 4: Regularization parameter update. *Set*

$$\sigma_{k+1} \in \begin{cases} [\gamma_1\sigma_k, \sigma_k] & \text{if } \rho_k \geq \eta_2, \\ [\sigma_k, \gamma_2\sigma_k] & \text{if } \rho_k \in [\eta_1, \eta_2), \\ [\gamma_2\sigma_k, \gamma_3\sigma_k] & \text{if } \rho_k < \eta_1. \end{cases} \tag{11}$$

Increment k by one and go to Step 1.

Iterations of Algorithm 2.1 where $\rho_k \geq \eta_1$ are called 'successful' and their index set is denoted by $\mathcal{S}$. Note that the mechanism of the algorithm ensures that $\sigma_k > 0$ for all $k \geq 0$. Note also that each iteration of the algorithm involves a single evaluation of the objective function and (for successful iterations only) of its gradient. The evaluation complexity can therefore be carried out by measuring how many *iterations* are needed before an approximate first-order critical point is found or the objective value decreases below the required target.

If $r = 2$ or $r = 3$, the model minimization occuring in Step 2 of the algorithm is typically easy to compute if one is happy with the minimum requirement that (5) and (6) hold: an efficient unidimensional linesearch technique using quadratic or cubic interpolation is all that is needed. Larger model decrease may be obtained by pursuing the minimization beyond the Cauchy point, and again efficient algorithms are known for quadratic and cubic regularizations (see Cartis *et al.* [4] for the latter case, the former being the well known problem of minimizing a quadratic function). Good methods are also available for more general values of r (in effect requiring the one-dimensional minimization of a rth order polynomial) : see Cartis *et al.* [2] for the case of regularized least-norm problems with general $r \geq 2$ or Gould, Robinson and Thorne [9] for even more general cases.

3. Worst-case evaluation complexity analysis

In order to analyse the worst-case complexity of Algorithm 2.1, we need to describe our assumptions and define some constants.

AS.1 The objective function f is continuously differentiable on $\Re^n$.

AS.2 $g = \nabla_x f$ is Hölder continuous in the sense that (2) holds for all $x, y \in \Re^n$ and some constants $L_\beta \geq 0$ and $\beta > 0$.

AS.3 Let f_{low} be any known value, possibly equal to minus infinity, such that $f(x) \geq f_{\text{low}}$, for all $x \in \Re^n$. We assume that

$$f_* \stackrel{\text{def}}{=} \max[f_{\text{low}}, f_{\text{target}}] > -\infty.$$

AS.4 Let $\kappa_{gl} \geq 0$ be any known value such that $\|g(x)\| \geq \kappa_{gl}$, for all x such that $f_* \leq f(x) \leq f(x_0)$. We assume that there exists $\kappa_{gu} \geq 1$ such that

$$\|g(x)\| \leq \kappa_{gu} \quad \text{for all } x \in \Re^n \text{ such that} f_* \leq f(x) \leq f(x_0).$$

AS.5 There exists a constant $\kappa_B \geq 0$ such that, for all $k \geq 0$,

$$\|B_k\| \leq \kappa_B.$$

AS.1 and AS.2 formalize our framework, as described in the introduction while AS.5 is standard in similar contexts and avoids possibly infinite curvature of the model, which would make the regularization irrelevant. Note that the values of $L_\beta \geq 0$ and $\beta > 0$ are often unknown to the user. AS.3 states that, if no target value is specified by the user, then there must exist a global lower bound on the objective function's values to make the minimization problem meaningful. The role of AS.4 is to take into account that, when $f_* = f_{\text{target}} > f_{\text{low}}$, it may well happen that no single $x \in \Re^n$ satisfies both conditions in (3), and thus that the first termination criterion in (3) cannot be satisfied by our minimization algorithm before the second. We take this possibility into account by allowing $\kappa_{gl} > 0$, and expresssing the complexity results in terms of

$$\epsilon_* \stackrel{\text{def}}{=} \max[\epsilon, \kappa_{gl}], \tag{12}$$

which is the 'attainable' gradient accuracy for the problem given f_{target}. For simplicity of exposition, we assume for now that $\epsilon_* < 1$, but comment on the case $\epsilon_* \geq 1$ at the end of the paper. We note that AS.4 automatically holds if the set $\{x \in \Re^n \mid f_* \leq f(x) \leq f(x_0)\}$ is bounded, but also, as we discuss in Lemma 3.2 below, in the frequent situation where $f(x)$ is bounded below on the level set $\{x \in \Re^n \mid f(x) \leq f(x_0)\}$.

We start by deriving consequences of our assumptions, which are independent of the algorithm. The first is intended to explore the consequence of a value of β exceeding 1.

LEMMA 3.1 *Suppose that* AS.1 *holds and that* AS.2 *holds for some* $\beta > 1$. *Then* f *is linear in* $\Re^n$, AS.2 *holds for all* $\beta > 0$ *with* $L_\beta = 0$ *and* AS.4 *holds with* $\kappa_{gl} = \kappa_{gu} = \|g(x_0)\|$.

Proof If e_i is the ith vector of the canonical basis and $[g(x)]_i$ the ith component of the gradient at x, we have, using the Cauchy-Schwarz inequality and the Hölder condition (2), that, for all $i = 1, \ldots, n$ and all $x \in \Re^n$,

$$\frac{|[g(x + te_i)]_i - [g(x)]_i|}{|t|} \leq \frac{\|g(x + te_i) - g(x)\|}{\|x + te_i - x\|} \leq L_\beta |t|^{\beta - 1}$$

and $\beta - 1 > 0$. Taking the limit when $t \to 0$ gives that the directional derivative of each $[g(\cdot)]_i$ exists and is zero for all i and at all x. Thus the gradient is constant in $\Re^n$, f is linear and AS.2 obviously holds with $L_\beta = 0$ for all $\beta > 0$ since $\|g(x) - g(y)\|$ is identically zero for all $x, y \in \Re^n$. ■

This justifies our choice to restrict our attention to the case where $\beta \in (0, 1]$ for the rest of our analysis. The second result indicates common circumstances in which AS.4 holds.

LEMMA 3.2 *Suppose that* AS.1 *and* AS.2 *hold, and that there exists a constant* $f_{\text{low}} > -\infty$ *such that*

$$f(x) \geq f_{\text{low}} \tag{13}$$

for all $x \in \mathcal{L}_0 \stackrel{\text{def}}{=} \{y \in \Re^n \mid f(y) \leq f(x_0)\}$. *Then* AS.4 *holds.*

Proof Let $x \in \mathcal{L}_0$. AS.1, the mean-value theorem, and AS.2 then ensure that, for all s,

$$\begin{aligned} f_{\text{low}} &\leq f(x+s) \\ &\leq f(x) + g(x)^T s + \int_0^1 (g(x+\xi s) - g(x))^T s \, \mathrm{d}\xi \\ &\leq f(x) + g(x)^T s + \frac{L_\beta}{1+\beta} \|s\|^{\beta+1} \stackrel{\text{def}}{=} h(s). \end{aligned} \tag{14}$$

Given that the minimizer of the convex function $h(s)$ is given by

$$s_* = -\frac{g(x)}{L_\beta^{1/\beta}} \|g(x)\|^{(1-\beta)/\beta},$$

we obtain that

$$\min_s h(s) = h(s_*) = f(x) - \frac{\beta L_\beta^{-1/\beta}}{1+\beta} \|g(x)\|^{1+1/\beta}.$$

As a consequence, we obtain, using the fact that $f(x) \leq f(x_0)$ since $x \in \mathcal{L}_0$ and (14), that

$$f_{\text{low}} \leq f(x_0) - \frac{\beta L_\beta^{-1/\beta}}{1+\beta} \|g(x)\|^{1+1/\beta},$$

which in turn implies that

$$\|g(x)\| \leq \left[L_\beta \left(1 + \frac{1}{\beta}\right)^\beta (f(x_0) - f_{\text{low}})^\beta \right]^{1/(1+\beta)} \stackrel{\text{def}}{=} \kappa_{gu},$$

irrespective of the value of f_{target}. This and the choice $\kappa_{gl} = 0$ yield the desired conclusion. ■

Note that (13) is indeed very common. For instance, $f_{\text{low}} = 0$ for all nonlinear least-squares problems. Hence the form of AS.4 should not be viewed as overly restrictive and also allows for the case where (13) fails but the objective function's gradient remains reasonably well-behaved. For instance, problems whose objective function is an indefinite quadratic are allowed provided $f_{\text{target}} > -\infty$.

We now turn to the analysis of the algorithm's properties. But, before we start in earnest, it is useful to introduce some specific notation. In a number of occurrences, we need to include some of the terms in formulae only if certain conditions apply. We will indicate this by underbracing the conditional part of the formula, the text below the underbrace then specifying the relevant condition. For instance we may have an expression of the type

$$\max \left[\underbrace{a^{-1}}_{a>0}, b, c \right],$$

meaning that the maximum should include the first term if and only if $a > 0$ (making the term well-defined in this case).

We first derive two bounds of the step length, generalizing Lemma 2.2 in [4].

LEMMA 3.3 *We have that, for all* $k \geq 0$,

$$\|s_k\| \leq \max\left[\underbrace{\left(\frac{r}{\sigma_k}\|B_k\|\right)^{1/(r-2)}}_{B_k \not\geq 0}, \left(\frac{2r}{\sigma_k}\|g_k\|\right)^{1/(r-1)}\right]. \tag{15}$$

Moreover,

$$\|s_k\| \leq \left(\frac{2r}{\sigma_k}\|g_k\|\right)^{1/(r-1)} \tag{16}$$

provided

$$\sigma_k \geq \frac{(r\|B_k\|)^{r-1}}{(2r\|g_k\|)^{r-2}}. \tag{17}$$

Proof Observe first that (4), (8) and $g_k \neq 0$ ensure that

$$m_k(x_k + s_k) - f(x_k) = {g_k}^{\mathrm{T}} s_k + \frac{1}{2}{s_k}^{\mathrm{T}} B_k s_k + \frac{\sigma_k}{r}\|s_k\|^r < 0. \tag{18}$$

Assume first that $s_k^{\mathrm{T}} B_k s_k > 0$. Then we must have that

$${g_k}^{\mathrm{T}} s_k + \frac{\sigma_k}{r}\|s_k\|^r < 0,$$

and therefore (remembering that $\sigma_k > 0$ and that $g_k^{\mathrm{T}} s_k \geq -\|g_k\|\,\|s_k\|$)

$$\|s_k\| < \left(\frac{r}{\sigma_k}\|g_k\|\right)^{1/(r-1)} < \left(\frac{2r}{\sigma_k}\|g_k\|\right)^{1/(r-1)}. \tag{19}$$

If $s_k^{\mathrm{T}} B_k s_k \leq 0$, we may rewrite (18) as

$$\left[{g_k}^{\mathrm{T}} s_k + \frac{\sigma_k}{2r}\|s_k\|^r\right] + \left[\frac{1}{2}{s_k}^{\mathrm{T}} B_k s_k + \frac{\sigma_k}{2r}\|s_k\|^r\right] < 0$$

and the left-hand side of this inequality can only be negative if at least one of the bracketed expressions is negative, giving that

$$\|s_k\| \leq \max\left[\left(\frac{r}{\sigma_k}\|B_k\|\right)^{1/(r-2)}, \left(\frac{2r}{\sigma_k}\|g_k\|\right)^{1/(r-1)}\right],$$

where we also used that $g_k^{\mathrm{T}} s_k \geq -\|g_k\|\,\|s_k\|$ and $s_k^{\mathrm{T}} B_k s_k \geq -\|B_k\|\,\|s_k\|^2$. Combining this with (19) then yields (15). Checking (16) subject to (17) is straightforward. ■

We now turn to the task of finding a lower bound on the model decrease $f(x_k) - m_k(x_k + s_k)$ resulting from (5)-(6). The first step is to find a suitable positive lower bound on the step α_k^{C} as defined in (6).

LEMMA 3.4 *We have that*

$$m_k(x_k + s_k{}^{\mathrm{C}}) \le m_k(x_k - \alpha_k^* g_k) < f(x_k), \tag{20}$$

where

$$\alpha_k^* \stackrel{\text{def}}{=} \min\left[\underbrace{\frac{\|g_k\|^2}{2g_k^{\mathrm{T}} B_k g_k}}_{g_k{}^{\mathrm{T}} B_k g_k > 0}, \left(\frac{r}{2\sigma_k}\frac{1}{\|g_k\|^{r-2}}\right)^{1/(r-1)}\right]. \tag{21}$$

Proof Substituting the definition $s = -\alpha g_k$ into (4), we obtain from (5) and (6) that, for all $\alpha > 0$,

$$m_k(x_k - \alpha g_k) - f(x_k) = \alpha\left(-\|g_k\|^2 + \frac{1}{2}\alpha g_k{}^{\mathrm{T}} B_k g_k + \frac{\sigma_k}{r}\alpha^{r-1}\|g_k\|^r\right). \tag{22}$$

Assume first that $g_k^{\mathrm{T}} B_k g_k \le 0$. Then

$$-\|g_k\|^2 + \frac{\sigma_k}{r}\alpha^{r-1}\|g_k\|^r < 0$$

for all $\alpha \in (0, \bar{\alpha}_k]$ where

$$\bar{\alpha}_k = \left(\frac{r}{\sigma_k}\frac{1}{\|g_k\|^{r-2}}\right)^{1/(r-1)} \tag{23}$$

and, because $\alpha > 0$ and $g_k^{\mathrm{T}} B_k g_k \le 0$, we also obtain from (24) that $m_k(x_k - \alpha g_k) < f(x_k)$ for all $\alpha \in (0, \bar{\alpha}_k]$. In particular, this yields that $m_k(x_k - \alpha_k^* g_k) < f(x_k)$, where

$$\alpha_k^* = \left(\frac{r}{2\sigma_k}\frac{1}{\|g_k\|^{r-2}}\right)^{1/(r-1)}. \tag{24}$$

Condition (6) then ensures that (20) holds as desired.

Assume next that $g_k^{\mathrm{T}} B_k g_k > 0$ and, in this case, define

$$\alpha_k^* \stackrel{\text{def}}{=} \min\left[\frac{\|g_k\|^2}{2g_k^{\mathrm{T}} B_k g_k}, \left(\frac{r}{2\sigma_k}\frac{1}{\|g_k\|^{r-2}}\right)^{1/(r-1)}\right]$$

Then it is easy to verify that both bracketed expressions in

$$\left[-\frac{1}{2}\|g_k\|^2 + \frac{1}{2}\alpha_k^* g_k{}^{\mathrm{T}} B_k g_k\right] + \left[-\frac{1}{2}\|g_k\|^2 + \frac{\sigma_k}{r}(\alpha_k^*)^{r-1}\|g_k\|^r\right] = \frac{1}{\alpha_k^*}(m_k(x_k - \alpha_k^* g_k) - f(x_k))$$

are negative and thus, because $\alpha_k^* > 0$, that $m_k(x_k - \alpha_k^* g_k) < f(x_k)$. The desired conclusion can now be obtained by invoking (6).

We now translate the conclusions of the last lemma in terms of the model reduction at the Cauchy point and beyond, generalizing Lemma 2.1 in [4]. ■

LEMMA 3.5 *We have that*

$$f(x_k) - m_k(x_k + s_k) \ge \frac{1}{4}\min\left[\underbrace{\frac{\|g_k\|^4}{2g_k^{\mathrm{T}} B_k g_k}}_{g_k{}^{\mathrm{T}} B_k g_k > 0}, \left(\frac{r}{2\sigma_k}\|g_k\|^r\right)^{1/(r-1)}\right]. \tag{25}$$

Proof If $g_k^{\mathrm{T}} B_k g_k \leq 0$, substituting (24) into (22) immediately yields that

$$f(x_k) - m_k(x_k - \alpha_k^* g_k) \geq \left(\frac{r}{2\sigma_k}\frac{1}{\|g_k\|^{r-2}}\right)^{1/(r-1)} [\|g_k\|^2 - \frac{1}{2}\|g_k\|^2] = \frac{1}{2}\left(\frac{r}{2\sigma_k}\|g_k\|^r\right)^{1/(r-1)}. \tag{26}$$

If $g_k^{\mathrm{T}} B_k g_k > 0$, we have from (22) and (21) that

$$\begin{aligned} & f(x_k) - m_k(x_k - \alpha_k^* g_k) \\ & \geq \alpha_k^* \left[\|g_k\|^2 - \frac{1}{2}\left(\frac{\|g_k\|^2}{2g_k^{\mathrm{T}} B_k g_k}\right) {g_k}^{\mathrm{T}} B_k g_k - \frac{\sigma_k}{r}\left(\frac{r}{2\sigma_k}\frac{1}{\|g_k\|^{r-2}}\right)\|g_k\|^r\right] \\ & = \min\left[\frac{\|g_k\|^2}{2g_k^{\mathrm{T}} B_k g_k}, \left(\frac{r}{2\sigma_k}\frac{1}{\|g_k\|^{r-2}}\right)^{1/(r-1)}\right][\|g_k\|^2 - \frac{1}{4}\|g_k\|^2 - \frac{1}{2}\|g_k\|^2] \\ & = \frac{1}{4}\min\left[\frac{\|g_k\|^4}{2g_k^{\mathrm{T}} B_k g_k}, \left(\frac{r}{2\sigma_k}\|g_k\|^r\right)^{1/(r-1)}\right]. \end{aligned}$$

Combining this last inequality with (26) and using (5) then gives (25). ■

The model decrease specified by (25) turns out to be useful if the value of σ_k (appearing at the denominator of the second term in the min) can be bounded above across all iterations. We obtain this result in two stages, the first being to determine conditions under which an iteration must be very successful.

LEMMA 3.6 *Suppose that* AS.1, AS.2 *and* AS.5 *hold. Then* $\rho_k \geq \eta_2$, *iteration* k *is very successful and* $\sigma_{k+1} \leq \sigma_k$

(i) *if* $1 + \beta \geq r$ *and*

$$\sigma_k \geq \kappa_1 \|g_k\|^{(1+\beta-r)/\beta} \tag{27}$$

where

$$\kappa_1 \stackrel{\text{def}}{=} 2r\left(\frac{L_\beta}{1+\beta}\right)^{(r-1)/\beta},$$

(ii) *if* $1 + \beta < r$ *and*

$$\sigma_k \geq \kappa_2 \max\left[\|g_k\|^{2-r}, \|g_k\|^{(1+\beta-r)/\beta}\right], \tag{28}$$

where

$$\kappa_2 \stackrel{\text{def}}{=} \max\left[2r(2\kappa_B)^{r-1}, 2^{(2+\beta)/\beta} r\kappa_3^{1/\beta}, 8r\kappa_3\right] \tag{29}$$

with

$$\kappa_3 \stackrel{\text{def}}{=} \left(\left[\frac{L_\beta}{1+\beta} + \frac{1}{2}\kappa_B\right]\left[\frac{4}{1-\eta_2}\right]\right)^{r-1}. \tag{30}$$

Proof First notice that AS.1, the mean-value theorem and (4) imply that

$$f(x_k+s_k)-m_k(x_k+s_k)=\int_0^1 (g(x_k+\xi s_k)-g_k)^{\mathrm{T}}s_k\,\mathrm{d}\xi-\frac{1}{2}{s_k}^{\mathrm{T}}B_k s_k-\frac{\sigma_k}{r}\|s_k\|^r.$$

Using now AS.2, we obtain that

$$f(x_k+s_k)-m_k(x_k+s_k)\le\frac{L_\beta}{1+\beta}\|s_k\|^{1+\beta}-\frac{1}{2}{s_k}^{\mathrm{T}}B_k s_k-\frac{\sigma_k}{r}\|s_k\|^r. \tag{31}$$

Assume first that $r\le 1+\beta$ (which implies that $B\succeq 0$ because of (7)). Then $f(x_k+s_k)\le m_k(x_k+s_k)$ (and thus $\rho_k\ge 1>\eta_2$) if

$$\sigma_k\ge\frac{rL_\beta}{1+\beta}\|s_k\|^{1+\beta-r},$$

which, in view of (15) and $B_k\succeq 0$, holds if

$$\sigma_k\ge\frac{rL_\beta}{1+\beta}\left(\frac{2r}{\sigma_k}\|g_k\|\right)^{(1+\beta-r)/(r-1)},$$

that is if

$$\sigma_k\ge 2r\left(\frac{L_\beta}{1+\beta}\right)^{(r-1)/\beta}\|g_k\|^{(1+\beta-r)/\beta}, \tag{32}$$

proving the first item in the lemma's statement.

Assume now that $r>1+\beta$, in which case B_k is allowed to be indefinite if $r>2$ and we cannot guarantee that $s_k^{\mathrm{T}}B_k s_k\ge 0$ in (31). Then $\rho_k\ge\eta_2$ if

$$r_k\stackrel{\mathrm{def}}{=}f(x_k+s_k)-m_k(x_k+s_k)-(1-\eta_2)(f(x_k)-m_k(x_k+s_k))<0.$$

Note that a lower bound on $f(x_k)-m_k(x_k+s_k)$ is given by Lemma 3.5. If we now assume that, whenever $g_k^{\mathrm{T}}B_k g_k>0$,

$$\sigma_k\ge\frac{r}{2}(2\kappa_B)^{r-1}\|g_k\|^{2-r}, \tag{33}$$

then we obtain that the minimum occurring in the right-hand side of (25) is achieved by the second term, yielding that

$$f(x_k)-m_k(x_k+s_k)\ge\frac{1}{4}\left(\frac{r}{2\sigma_k}\|g_k\|^r\right)^{1/(r-1)}.$$

As a consequence, we obtain from (31), the Cauchy–Schwarz inequality and AS.5 that

$$r_k\le\frac{L_\beta}{1+\beta}\|s_k\|^{1+\beta}+\frac{1}{2}\kappa_B\|s_k\|^2-\frac{1-\eta_2}{4}\left(\frac{r}{2\sigma_k}\|g_k\|^r\right)^{1/(r-1)}.$$

If we also assume that, whenever $B_k\not\succeq 0$, (17) also holds, then we may substitute the upper bound (16) in this equation and obtain that $r_k<0$ if

$$\frac{L_\beta}{1+\beta}\left(\frac{2r}{\sigma_k}\|g_k\|\right)^{(1+\beta)/(r-1)}+\frac{1}{2}\kappa_B\left(\frac{2r}{\sigma_k}\|g_k\|\right)^{2/(r-1)}<\frac{1-\eta_2}{4}\left(\frac{r}{2\sigma_k}\|g_k\|^r\right)^{1/(r-1)}.$$

Now, if, on one hand,

$$\left(\frac{2r}{\sigma_k}\|g_k\|\right)^{(1+\beta)/(r-1)} \geq \left(\frac{2r}{\sigma_k}\|g_k\|\right)^{2/(r-1)}, \tag{34}$$

then we obtain that $r_k < 0$ if

$$\left(\frac{L_\beta}{1+\beta} + \frac{1}{2}\kappa_B\right)\left(\frac{2r}{\sigma_k}\|g_k\|\right)^{(1+\beta)/(r-1)} < \frac{1-\eta_2}{4}\left(\frac{r}{2\sigma_k}\|g_k\|^r\right)^{1/(r-1)}.$$

Taking the $(r-1)$th power and rearranging, we obtain that $r_k < 0$ if

$$\sigma_k \geq 2^{(2+\beta)/\beta} r \left(\frac{L_\beta}{1+\beta} + \frac{1}{2}\kappa_B\right)^{(r-1)/\beta} \left(\frac{4}{1-\eta_2}\right)^{(r-1)/\beta} \|g_k\|^{(1+\beta-r)/\beta}. \tag{35}$$

If, on the other hand, (34) fails, then $r_k < 0$ if

$$\left(\frac{L_\beta}{1+\beta} + \frac{1}{2}\kappa_B\right)\left(\frac{2r}{\sigma_k}\|g_k\|\right)^{2/(r-1)} < \frac{1-\eta_2}{4}\left(\frac{r}{2\sigma_k}\|g_k\|^r\right)^{1/(r-1)}.$$

Once more taking the $(r-1)$th power and rearranging, we obtain that $r_k < 0$ if

$$\sigma_k \geq 8r\left(\frac{L_\beta}{1+\beta} + \frac{1}{2}\kappa_B\right)^{r-1}\left(\frac{4}{1-\eta_2}\right)^{r-1}\|g_k\|^{2-r}. \tag{36}$$

Thus $r_k < 0$ (and therefore $\rho_k \geq \eta_2$) when $r > 1+\beta$ provided (35) and (36) hold together with (17) (when $B_k \not\succeq 0$) and (33) (when $g_k^{\mathrm{T}} B_k g_k > 0$). This proves the second item in the lemma's statement if we note that

$$\frac{(r\kappa_B)^{r-1}}{(2r)^{r-2}} = 2r\left(\frac{1}{2}\kappa_B\right)^{r-1} < 2r(2\kappa_B)^{r-1} \quad \text{and} \quad \frac{r}{2}(2\kappa_B)^{r-1} < 2r(2\kappa_B)^{r-1}.$$

■

Note that the second part of the lemma extends the result of Lemma 3.1 in [5] to general r and β. We are now in position to prove an iteration-independent upper bound on the value of σ_k.

LEMMA 3.7 *Suppose that* AS.1–AS.5 *hold and that* $\epsilon_* < 1$. *Then, as long as the algorithm does not terminate, we have that, for all* $k \geq 0$,

(i) *if* $1+\beta \geq r$,

$$\sigma_k \leq \kappa_1^\sigma, \tag{37}$$

where

$$\kappa_1^\sigma \stackrel{\text{def}}{=} \max[\gamma_3\kappa_1\kappa_{gu}^{1+\beta-r}, \sigma_0]; \tag{38}$$

(ii) *if* $1+\beta < r$,

$$\sigma_k \le \max[\kappa_2^\sigma, \kappa_3^\sigma \epsilon_*^{(1+\beta-r)/\beta}], \tag{39}$$

where

$$\kappa_2^\sigma \stackrel{\text{def}}{=} \max\left[0, \underbrace{\gamma_3\kappa_2\kappa_{gu}^{2-r}}_{r\le 2}, \sigma_0\right] \quad \text{and} \quad \kappa_3^\sigma \stackrel{\text{def}}{=} \gamma_3\kappa_2 \tag{40}$$

with κ_1 *and* κ_2 *defined in* (29).

Proof We again distinguish two cases. Assume first that $1+\beta \ge r$, which in turn implies that $r \in (1,2]$ and thus, in view of (7), that $B_k \succeq 0$ for all k. Then AS.4 and Lemma 3.6(i) imply that $\sigma_{k+1} \le \sigma_k$ provided

$$\sigma_k \ge \kappa_1\kappa_{gu}^{(1+\beta-r)/\beta}, \tag{41}$$

which is a constant independent of k and ϵ.

The second case is when $1+\beta < r$. We first consider the subclass where $r \le 2$ where, using AS.4,

$$\|g_k\|^{2-r} \le \kappa_{gu}^{2-r}. \tag{42}$$

This bound, part (ii) of Lemma 3.6 and the fact that $\|g_k\| > \epsilon_*$ as long as the algorithm has not terminated then imply that $\sigma_{k+1} \le \sigma_k$ provided

$$\sigma_k \ge \kappa_2 \max\left[\underbrace{\kappa_{gu}^{2-r}}_{r\le 2}, \epsilon_*^{(1+\beta-r)/\beta}\right], \tag{43}$$

where we have used that $1+\beta-r < 0$. Alternatively, if $r > 2$, part (ii) of Lemma 3.6 and the fact that $\|g_k\| > \epsilon_*$ as long as the algorithm has not terminated then give that $\sigma_{k+1} \le \sigma_k$ provided

$$\sigma_k \ge \kappa_2 \max[\epsilon_*^{2-r}, \epsilon_*^{(1+\beta-r)/\beta}] = \kappa_2\epsilon_*^{(1+\beta-r)/\beta}, \tag{44}$$

where the last equality now results from the fact that, because $\beta \le 1$,

$$0 > 2-r \ge \frac{2-r}{\beta} \ge \frac{1+\beta-r}{\beta}.$$

The proof of (37) and (39) is then completed by taking into account that the initial parameter σ_0 may exceed the bound given by the right-hand side (41) (if $1+\beta \ge r$) or (43) (if $1+\beta < r$), and also that these bounds may just fail by a small margin at an unsuccessful iteration, resulting in an increase of σ_k by a factor γ_3 before the relevant bound applies. ■

Having now derived an iteration independent upper bound on σ_k, we may return to the model decrease given by Lemma 3.5.

LEMMA 3.8 *Suppose that* AS.1–AS.5 *hold and that* $\epsilon_* < 1$. *Then, as long as the algorithm does not terminate,*

- *if* $1+\beta \ge r$, *then*

$$f(x_k) - m(x_k+s_k) \ge \kappa_1^m \epsilon_*^{r/(r-1)}, \tag{45}$$

where

$$\kappa_1^m \stackrel{\text{def}}{=} \frac{1}{4}\min\left[\frac{1}{2\kappa_B}, \left(\frac{r}{2\kappa_1^\sigma}\right)^{1/(r-1)}\right], \tag{46}$$

- *if* $1+\beta < r$, *then*

$$f(x_k) - m(x_k + s_k) \geq \kappa_2^m \epsilon_*^{1+1/\beta}, \tag{47}$$

where

$$\kappa_2^m \stackrel{\text{def}}{=} \frac{1}{4} \min \left[\frac{1}{2\kappa_B}, \left(\frac{r}{2\kappa_2^\sigma} \right)^{1/(r-1)} \right]. \tag{48}$$

Proof Assume first that $1+\beta \geq r$. As above, this implies that $r \in [1,2]$ and hence, because of (7), that $g_k^{\mathrm{T}} B_k g_k \geq 0$. Taking into account that, in this case,

$$g_k{}^{\mathrm{T}} B_k g_k \leq \kappa_B \|g_k\|^2$$

because of AS.5, substituting (37) into (25) and using (37) and the fact that $\|g_k\| \geq \epsilon_*$ as long as the algorithm has not terminated, yields that

$$\begin{aligned} f(x_k) - m_k(x_k + s_k) &\geq \frac{1}{4} \min \left[\frac{\epsilon_*^2}{2\kappa_B}, \left(\frac{r}{2\kappa_1^\sigma} \right)^{1/(r-1)} \epsilon_*^{r/(r-1)} \right] \\ &\geq \frac{1}{4} \min \left[\frac{1}{2\kappa_B}, \left(\frac{r}{2\kappa_1^\sigma} \right)^{1/(r-1)} \right] \min \left(\epsilon_*^2, \epsilon_*^{r/(r-1)} \right) \end{aligned}$$

and (45) follows since $\epsilon_* < 1$ and

$$\frac{r}{r-1} \geq 2 \quad \text{for} \quad r \in [1,2].$$

Consider now the case where $1+\beta < r$. Substituting now (39) into (25), using (39), AS.5 and the fact that $\|g_k\| \geq \epsilon_*$ as long as the algorithm has not terminated, we obtain that

$$\begin{aligned} f(x_k) - m_k(x_k + s_k) &\geq \frac{1}{4} \min \left[\underbrace{\frac{\epsilon_*^2}{2\kappa_B}}_{g_k{}^{\mathrm{T}} B_k g_k > 0}, \left(\frac{r\epsilon_*^r}{2\max[\kappa_2^\sigma, \kappa_3^\sigma \epsilon_*^{(1+\beta-r)/\beta}]} \right)^{1/(r-1)} \right] \\ &\geq \frac{1}{4} \min \left[\frac{1}{2\kappa_B}, \left(\frac{r}{2\max[\kappa_2^\sigma, \kappa_3^\sigma]} \right)^{1/(r-1)} \right] \min \left(\epsilon_*^2, \epsilon_*^{r/(r-1)}, \epsilon_*^{1+1/\beta} \right), \end{aligned}$$

which yields (47) since $\epsilon_* < 1$ and, for $1+\beta < r$ and $\beta \in (0,1]$,

$$1 + \frac{1}{\beta} \geq \frac{r}{r-1} \quad \text{and} \quad 1 + \frac{1}{\beta} \geq 2.$$

■

We now recall an important technical lemma which, in effect, gives a bound on the total number of unsuccessful iterations before iteration k as a function of the number of successful ones.

LEMMA 3.9 *The mechanism of Algorithm* 2.1 *guarantees that, if*

$$\sigma_k \leq \sigma_{\max}, \tag{49}$$

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

$$k \leq |\mathcal{S}_k| \left(1 + \frac{|\log \gamma_1|}{\log \gamma_2}\right) + \frac{1}{\log \gamma_2} \log\left(\frac{\sigma_{\max}}{\sigma_0}\right), \tag{50}$$

where $|\mathcal{S}_k|$ *is the cardinality of* $\mathcal{S}_k \stackrel{\text{def}}{=} \{j \in \mathcal{S} \mid j \leq k\}$, *that is the number of successful iterations up to iteration* k.

Proof We first note that the construction of the algorithm implies that $k = |\mathcal{S}_k| + |\mathcal{U}_k|$, where $\mathcal{U}_k$ denotes the number of unsuccessful iterations up to k. The bound (50) now follows by upper bounding $|\mathcal{U}_k|$ using [5, (2.13)]; where we note that the same update for σ_{k+1} is used in [5, (2.13)] as here, provided we account for a change in notation (namely, γ_3 in [5] is γ_1 here, γ_2 in [5] corresponds to γ_3 and γ_1 to γ_2 here). ■

We are now ready to prove our main result on the worst-case complexity of Algorithm 2.1.

THEOREM 3.10 *Suppose that* AS.1–AS.5 *hold and that* ϵ_* *defined in* (12) *satisfies* $\epsilon_* < 1$.

1. *If* $1 + \beta \geq r$, *there exist constants* κ_r^s, κ_r^a *and* κ_r^c *such that, for any* $\epsilon > 0$, *Algorithm* 2.1 *requires at most*

$$\left\lfloor \kappa_r^s \frac{f(x_0) - f_*}{\epsilon_*^{r/(r-1)}} \right\rfloor \tag{51}$$

successful iterations (*and gradient evaluations*), *and a total of*

$$\left\lfloor \kappa_r^a \frac{f(x_0) - f_*}{\epsilon_*^{r/(r-1)}} + \kappa_r^c \right\rfloor \tag{52}$$

iterations (*and objective function evaluations*) *before producing an iterate* x_ϵ *such that* $\|g(x_\epsilon)\| \leq \epsilon_*$ *or* $f(x_\epsilon) \leq f_{\text{target}}$.

2. *If* $1 + \beta < r$, *there exist constants* κ_β^s, κ_β^a, κ_β^b *and* κ_β^c *such that, for all* $\epsilon > 0$, *Algorithm* 2.1 *requires at most*

$$\left\lfloor \kappa_\beta^s \frac{f(x_0) - f_*}{\epsilon_*^{1+1/\beta}} \right\rfloor \tag{53}$$

successful iterations (and gradient evaluations) and a total of

$$\left\lfloor \kappa_\beta^a \frac{f(x_0) - f_*}{\epsilon_*^{1+1/\beta}} + \kappa_\beta^b |\log \epsilon_*| + \kappa_\beta^c \right\rfloor \tag{54}$$

iterations (*and objective function evaluations*) *before producing an iterate* x_ϵ *such that* $\|g(x_\epsilon)\| \leq \epsilon_*$ *or* $f(x_\epsilon) \leq f_{\text{target}}$. *In the above statements the constants are given by*

$$\kappa_r^s = \kappa_\beta^s \stackrel{\text{def}}{=} \frac{1}{\eta_1 \kappa_m}, \tag{55}$$

$$\kappa_r^a \stackrel{\text{def}}{=} \frac{1}{\eta_1 \kappa_m}\left(1 + \frac{|\log \gamma_1|}{\log \gamma_2}\right), \quad \kappa_r^c \stackrel{\text{def}}{=} \frac{1}{\log \gamma_2} \log\left(\frac{\kappa_1^\sigma}{\sigma_0}\right), \tag{56}$$

$$\kappa_\beta^a \stackrel{\text{def}}{=} \frac{1}{\eta_1 \kappa_m}\left(1 + \frac{|\log \gamma_1|}{\log \gamma_2}\right), \quad \kappa_\beta^b \stackrel{\text{def}}{=} \frac{r - \beta - 1}{\beta \log \gamma_2} \tag{57}$$

and

$$\kappa_\beta^c \stackrel{\text{def}}{=} \frac{1}{\log \gamma_2}(\log(\max[1, \kappa_2^\sigma, \kappa_3^\sigma]) + |\log(\sigma_0)|), \tag{58}$$

where

$$\kappa_1 \stackrel{\text{def}}{=} 2r\left(\frac{L_\beta}{1+\beta}\right)^{(r-1)/\beta}, \quad \kappa_2 \stackrel{\text{def}}{=} \max[2r(2\kappa_B)^{r-1}, 2^{(2+\beta)/\beta} r\kappa_3^{1/\beta}, 8r\kappa_3] \tag{59}$$

with

$$\kappa_3 \stackrel{\text{def}}{=} \left(\left[\frac{L_\beta}{1+\beta} + \frac{1}{2}\kappa_B\right]\left[\frac{4}{1-\eta_2}\right]\right)^{r-1}, \tag{60}$$

$$\kappa_1^\sigma \stackrel{\text{def}}{=} \gamma_3\kappa_1\kappa_{gu}^{1+\beta-r}, \quad \kappa_2^\sigma \stackrel{\text{def}}{=} \gamma_3 \max\left[0, \underbrace{\kappa_2\kappa_{gu}^{2-r}}_{r\le 2}\right], \quad \kappa_2^\sigma \stackrel{\text{def}}{=} \max \gamma_3\kappa_2 \tag{61}$$

and

$$\kappa_1^m \stackrel{\text{def}}{=} \frac{1}{4}\min\left[\frac{1}{2\kappa_B}, \left(\frac{r}{2\kappa_1^\sigma}\right)^{1/(r-1)}\right] \quad \text{and} \quad \kappa_2^m \stackrel{\text{def}}{=} \frac{1}{4}\min\left[\frac{1}{2\kappa_B}, \left(\frac{r}{2\kappa_2^\sigma}\right)^{1/(r-1)}\right]. \tag{62}$$

Proof Consider first the case where $1+\beta \ge r$. We then deduce from AS.3, the definition of a successful iteration and (45) in Lemma 3.8, that, as long as the algorithm has not terminated,

$$\begin{aligned} f(x_0) - f_* \ge f(x_0) - f(x_{k+1}) &= \sum_{j\in\mathcal{S}_k}[f(x_j) - f(x_j+s_j)] \\ &\ge \eta_1 \sum_{j\in\mathcal{S}_k}[f(x_j) - m_j(x_j+s_j)] > \eta_1\kappa_m\epsilon_*^{r/(r-1)}|\mathcal{S}_k|. \end{aligned} \tag{63}$$

This provides an upper bound on $|\mathcal{S}_k|$ which is independent of k and ϵ_*, from which we obtain the bound (51) with (55). Calling now upon Lemma 3.9 and (37), we deduce that the total number of iterations (and function evaluations) cannot exceed

$$\kappa_r^s \frac{f(x_0) - f_*}{\epsilon_*^{r/(r-1)}}\left(1 + \frac{|\log\gamma_1|}{\log\gamma_2}\right) + \frac{1}{\log\gamma_2}\log\left(\frac{\kappa_1^\sigma}{\sigma_0}\right),$$

which then gives the bound (52) with (56).

The proof for the case where $1+\beta < r$ is derived in a manner entirely similar to that used for the case where $1+\beta \ge r$, replacing $\epsilon^{r/(r-1)}$ by $\epsilon^{1+1/\beta}$ in (63) since (45) is used instead of (47), and also noting that, when using (39) instead of (37) in Lemma 3.9,

$$\log\left(\frac{\max[\kappa_2^\sigma, \kappa_3^\sigma\epsilon_*^{(1+\beta-r)/\beta}]}{\sigma_0}\right) \le \left|\frac{1+\beta-r}{\beta}\right| |\log\epsilon_*| + \log(\max[1, \kappa_2^\sigma, \kappa_3^\sigma]) + |\log(\sigma_0)|.$$

We may thus deduce that (53) and (54) hold with (57)–(62). ■

A close look at the expressions of the constants in (55)– (62) reveals that the global upper bound on the gradient norm, κ_{gu}, only occurs in the case where $r < 2$. Therefore, AS.4 is only needed in this case since the existence of $\kappa_{gl} \ge 0$ is always ensured by the non-negativity of $\|g(x)\|$.

4. An example of sharpness

We now show that the bound specified by part (ii) of Theorem 3.10 is essentially sharp in the sense that we exhibit a class of one-dimensional examples where the number of iterations necessary to produce an approximate first-order critical point is arbitrarily close to the theorem's bound[1]. To achieve this goal, we first establish sequences of iterates $\{x_k\}$, function values $\{f(x_k)\}$, gradient values $\{g_k\}$ and regularization parameter values $\{\sigma_k\}$ which can be generated by Algorithm 2.1 and such that the gradient values converge to zero sufficiently slowly to attain the desired lower bound on the number of iterations (and evaluations). Once these are defined, we construct a function $f(x)$ which interpolates these function and gradient values and finally prove that all our assumptions are satisfied. Because the derivation of the complexity bound involves an increasing sequence of regularization parameters $\{\sigma_k\}$, our example is unfortunately somewhat complicated because it has to include both successful and unsuccessful iterations. We choose to construct it such that *all even iterations are unsuccessful and all odd ones are successful*.

Construction of $\{g_k\}$*, the sequence of gradient values at the iterates.* Let $r > 1+\beta$, $\tau \in (0,1)$ be arbitrarily small, and q a positive integer. For all $k \geq 0$, consider the gradient sequence defined by

$$g_{2k} = -\left(\frac{1}{k+q}\right)^{\beta/(1+\beta)+\tau}, \quad g_{2k+1} = g_{2k} \tag{64}$$

and observe that the sequence of gradient norms $\{\|g_k\|\}$ is non-increasing for any choice of q. Assume first that $q=1$. This definition implies that

$$\frac{|g_{2k+3}|}{|g_{2k+1}|} \to 1, \tag{65}$$

when k tends to infinity, and thus that

$$\omega_{2k-1} \stackrel{\text{def}}{=} \left(\frac{|g_{2k+1}|^{1/\beta}}{|g_{2k+1}|^{1/\beta} + \frac{1}{2}|g_{2k+3}|^{1/\beta}}\right)^{r-1} \left(\frac{|g_{2k}|^{1/\beta}}{|g_{2k-1}|^{1/\beta}}\right)^{1+\beta-r} \to \left(\frac{2}{3}\right)^{r-1}. \tag{66}$$

Hence, there exists an integer $\ell \geq 2$ such that

$$\omega_{2k-1} \in \left[\frac{1}{2}\left(\frac{2}{3}\right)^{r-1}, \left(\frac{5}{6}\right)^{r-1}\right] \subset (0,1) \qquad \text{for} \quad k \geq \ell. \tag{67}$$

We now (re)define q in (64) by setting $q = \ell$, in effect shifting the $\{k\}$ sequence by ℓ such that (65)-(67) holds with (64) for the complete shifted sequence. Note that q only depends on β and τ and is independent of ϵ. Observe also that the rate of (monotonic) convergence of the sequence $\{g_k\}$ to zero ensures that, for any $\epsilon \in (0,1)$, $|g_k| \leq \epsilon$ only for k larger than $2(\lfloor \epsilon^{-(1+\beta)/(\beta+\tau(1+\beta))} \rfloor - q)$.

Construction of the iterates $\{x_k\}$ *and the steps* $\{s_k\}$. For $k \geq 0$, the step s_k is computed as the global minimizer of the model $m_k(x_k+s)$ in (4) with $B_k = 0$, that is

$$m_k(x_{2k+1}+s) = f(x_k) + g_k s + \frac{\sigma_k}{r}|s|^r,$$

where the function value $f(x_k)$ and σ_k are still to be defined. A simple calculation shows that

$$s_k = \left(\frac{|g_k|}{\sigma_k}\right)^{1/(r-1)}, \tag{68}$$

and furthermore, that

$$\Delta m_k \stackrel{\text{def}}{=} m_k(x_k) - m_k(x_k + s_k) = \left(1 - \frac{1}{r}\right)\left(\frac{|g_k|^r}{\sigma_k}\right)^{1/(r-1)} = \left(1 - \frac{1}{r}\right)|g_k s_k|. \tag{69}$$

Recalling that we attempt to ensure that odd iterations are successful and even ones are not, we define the sequence of iterates $\{x_k\}$ by

$$x_0 = x_1 = 0, \quad x_{2k+2} = x_{2k+1} + s_{2k+1} = x_{2k+3} \quad (k \geq 0).$$

Construction of $\{\sigma_k\}$ *and the function values* $\{f(x_k)\}$ *at the iterates.* In order to ensure the proper rate of increase of σ_k, we choose to set

$$\sigma_{2k+1} = |g_{2k+1}|^{(1+\beta-r)/\beta} \tag{70}$$

for all $k \geq 0$ (remembering that odd iterations are successful), while the value of σ_{2k} is still to be determined within the constraints of (11). It follows from (68) that

$$s_{2k+1} = |g_{2k+1}|^{1/\beta} \leq |g_0|^{1/\beta} < 1 \tag{71}$$

and so (69) becomes $\Delta m_{2k+1} = m_{2k+1}(x_{2k+1}) - m_{2k+1}(x_{2k+1} + s_{2k+1}) = (1 - 1/r)|g_{2k+1}|^{(1+\beta)/\beta}$. The sequence of function values is then defined by

$$f(x_0) = f(x_1) = 0, \quad f(x_{2k+2}) = m_{2k+1}(x_{2k+1} + s_{2k+1}) = f(x_{2k+3}) \quad (k \geq 0), \tag{72}$$

where the second part guarantees the very successful nature of iteration $2k+1$. We observe that, for $k \geq 0$,

$$f(x_{2k}) - f(x_{2k+1}) = 0$$

since iteration $2k$ is unsuccessful, and

$$f(x_{2k+1}) - f(x_{2k+2}) = \Delta m_{2k+1} = \left(1 - \frac{1}{r}\right)|g_{2k+1}|^{(1+\beta)/\beta}, \tag{73}$$

yielding that, for every $k \geq 0$,

$$\begin{aligned} f(x_0) - f(x_{2k+2}) &= \sum_{j=0}^{k}[f(x_{2j+1}) - f(x_{2j+2})] \\ &= \left(1 - \frac{1}{r}\right)\sum_{j=0}^{k}|g_{2j+1}|^{(1+\beta)/\beta} \\ &= \left(1 - \frac{1}{r}\right)\sum_{j=0}^{k}\left(\frac{1}{j+q}\right)^{1+((1+\beta)/\beta)\tau}. \end{aligned}$$

Hence the sequence $\{f(x_k)\}$ is bounded below by

$$f_\infty \stackrel{\text{def}}{=} \left(1 - \frac{1}{r}\right)\left[-\zeta\left(1 + \frac{1+\beta}{\beta}\tau\right) + \sum_{j=1}^{q-1}\left(\frac{1}{j}\right)^{1+((1+\beta)/\beta)\tau}\right] > -\infty, \tag{74}$$

where $\zeta(\cdot)$ is the Riemann zeta function. We conclude the definition of the sequences involved in our example by selecting σ_{2k} in order to impose that, for all $k \geq 0$,

$$s_{2k} = s_{2k+1} + \frac{1}{2} s_{2k+3}, \tag{75}$$

where $\frac{1}{2} \in [\frac{1}{2}, 1)$ is chosen as when defining q above. Using (68), this is equivalent to asking that

$$\frac{|g_{2k}|}{\sigma_{2k}} = \left(s_{2k+1} + \frac{1}{2} s_{2k+3} \right)^{r-1},$$

which, in view of (70), is equivalent to requiring that

$$\frac{\sigma_{2k}}{\sigma_{2k-1}} = \frac{|g_{2k}|}{|g_{2k-1}|^{(1+\beta-r)/\beta}} \left(\frac{1}{s_{2k+1} + \frac{1}{2} s_{2k+3}} \right)^{r-1}.$$

If we now take (71), (66) and (67) into account, this amounts to imposing that

$$\frac{\sigma_{2k}}{\sigma_{2k-1}} = \omega_{2k-1} \in \left[\frac{1}{2} \left(\frac{2}{3} \right)^{r-1}, \left(\frac{5}{6} \right)^{r-1} \right],$$

therefore satisfying (11) at successful iterations for a choice of $\gamma_1 \leq \frac{1}{2}(\frac{2}{3})^{r-1}$. (In order to start the recursion, we (arbitrarily) define σ_{-1} by (70) with $k = -1$ and $g_{-1} = -[1/(q-1)]^{\beta/(1+\beta)+\tau}$.) We also observe that, for large enough k,

$$\frac{\sigma_{2k+1}}{\sigma_{2k}} = \frac{|g_{2k+1}|^{(1+\beta-r)/\beta}}{\omega_{2k-1}\sigma_{2k-1}} = \left(\frac{s_{2k+1} + \frac{1}{2} s_{2k+3}}{s_{2k+1}} \right)^{r-1} \in \left[\left(\frac{s_1 + \frac{1}{2} s_3}{s_1} \right)^{r-1}, \left(\frac{3}{2} \right)^{r-1} \right] \tag{76}$$

and (11) therefore also holds at unsuccessful iterations. As a consequence of this somewhat lengthy description, we may therefore deduce that the sequences $\{x_k\}$, $\{g_k\}$ $\{\sigma_k\}$ and $\{f(x_k)\}$ may be generated by Algorithm 2.1 provided only that iteration $2k$ is indeed unsuccessful, that is if

$$f(x_{2k}) - f(x_{2k} + s_{2k}) < \eta_1 \Delta m_{2k},$$

where $f(x_{2k} + s_{2k})$ is the still undefined value of our putative objective function at $x_{2k} + s_{2k} = x_{2k+3} + \frac{1}{2} s_{2k+3}$. This condition is obviously satisfied if we also impose that $f(x_{2k+3} + \frac{1}{2} s_{2k+3}) = f_{2k+3}^{2k}$, where

$$f_{2k+3}^{2k} \stackrel{\text{def}}{=} \max[f(x_{2k+3}), f(x_{2k}) - 0.99\eta_1 \Delta m_{2k}, f(x_{2k+4}) - \tfrac{1}{2} g_{2k+4} s_{2k+3}]. \tag{77}$$

Note that this last condition ensures that

$$f(x_{2k+2}) = f(x_{2k+3}) \leq f_{2k+3}^{2k} \tag{78}$$

and also, since $f(x_{2k}) = f(x_{2k+1}) > f(x_{2k+3})$, that

$$f_{2k+3}^{2k} \geq f(x_{2k+4}) - \tfrac{1}{2} g_{2k+4} s_{2k+3} \;\Rightarrow\; f_{2k+3}^{2k} \in [f(x_{2k+3}), f(x_{2k+1})]. \tag{79}$$

Construction of the objective function $f(x)$, $x \geq 0$. We now turn to the definition of the objective function $f(x)$ which must interpolate the (already-defined) function and gradient values at

the iterates. We start by noting that, for arbitrary $a > 0$ and $s > 0$, function values f_a and f_b and gradient values g_a and g_b, it is possible to construct a function

$$f_{as}(t) = f_a + g_a t + c_{as}[\sin(\phi_{as} t)]^{1+\beta} \tag{80}$$

on the interval $[a, a+s]$ where the parameters c_{as} and $\phi_{as} \in (0, \pi]$ can be determined to ensure that

$$f_{as}(0) = f_a, \quad g_{as}(0) = g_a, \quad f_a(s) = f_b \quad \text{and} \quad g_{as}(s) = g_b.$$

Indeed, since

$$g_{as}(t) = g_a + c_{as}(1+\beta)\phi_{as}[\sin(\phi_{as} t)]^{\beta} \cos(\phi_{as} t), \tag{81}$$

we deduce that

$$g_b - g_a = c_{as}(1+\beta)\phi_{as}[\sin(\phi_{as} s)]^{\beta} \cos(\phi_{as} s), \tag{82}$$

which may substitute in (80) to obtain that

$$f_b - f_a = g_a s + \frac{(g_b - g_a)\sin(\phi_{as} s)}{(1+\beta)\phi_{as}\cos(\phi_{as} s)}$$

and hence conclude that $\phi_{as} s$ is the smallest positive root θ_{as} of the nonlinear equation

$$\frac{\sin(\theta)}{\theta} = \nu_{as}\cos(\theta), \quad \text{where} \quad \nu_{as} = (1+\beta)\frac{f_b - f_a - g_a s}{(g_b - g_a)s}. \tag{83}$$

It is easy to check that such a root always exist in $(0, \pi/2]$ if $\nu_{as} > 1$. Given ϕ_{as}, or, equivalently, $\theta_{as} = \phi_{as} s$, we also obtain that $c_{as} = (f_b - f_a - g_a s)[\sin(\theta_{as})]^{-1-\beta}$.

We now use this interpolation technique on each of the sequence of intervals specified in Table 1. Observe that the function is interpolated for every successful step in two pieces with an intermediate point corresponding (for all iterations beyond the first) to the penultimate unsuccessful trial point, where condition (77) is imposed as well as a zero gradient. We also choose (arbitrarily)

$$f_1^{-2} = |g_{-1}|^{(1+\beta)/\beta} - 0.99\frac{3\eta_1}{2}\left(\frac{|g_{-1}|^r}{\sigma_{-1}}\right)^{1/(r-1)}$$

(corresponding to a fictitious unsuccessful iteration of index $k = -2$ with $g_{-2} = g_{-1}$ and $\sigma_{-2} = \sigma_{-1}/(1+\frac{1}{2})^{r-1}$).

Table 1. Interpolation conditions for successful iterations.

Iteration k	Interpolation interval $[a, a+s]$	Interpolated values f_a	g_a	f_b	g_b
1	$[x_1, x_1 + \frac{1}{2}s_1]$	$f(x_0) = f(x_1)$	g_1	f_1^{-2}	0
1	$[x_1 + \frac{1}{2}s_1, x_2]$	0	0	$f(x_2)$	g_2
3	$[x_3, x_3 + \frac{1}{2}s_3]$	$f(x_2) = f(x_3)$	g_3	f_3^0	0
3	$[x_3 + \frac{1}{2}s_3, x_4]$	f_3^0	0	$f(x_4)$	g_4
5	$[x_5, x_5 + \frac{1}{2}s_5]$	$f(x_4) = f(x_5)$	g_5	f_5^2	0
5	$[x_5 + \frac{1}{2}s_5, x_6]$	f_5^2	0	$f(x_6)$	g_6
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$2k+1$	$[x_{2k+1}, x_{2k+1} + \frac{1}{2}s_{2k+1}]$	$f(x_{2k}) = f(x_{2k+1})$	g_{2k+1}	f_{2k+1}^{2k-2}	0
$2k+1$	$[x_{2k+1} + \frac{1}{2}s_{2k+1}, x_{2k+2}]$	f_{2k+1}^{2k-2}	0	$f(x_{2k+2})$	g_{2k+2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

For the function (80) and its gradient (81) to be well-defined, we still need that $\nu_{as} > 1$ for each interpolation interval, which we show in the appendix, Lemma A.1.

Figure 1 shows the shape of the resulting function and its gradient, whose construction implies that AS.1 holds. Figure 1 also shows the shape of the models $m_{2k}(x_{2k} + s)$ on the intervals $[x_{2k}, x_{2k} + s_{2k}] = [x_{2k}, x_{2k+3} + \frac{1}{2}s_{2k+3}]$ (dashed lines), illustrating that the model is a bad predictor of the objective function value at the point $x_{2k} + s_{2k}$, causing the unsuccessful nature of iteration $2k$. Note that $f(x)$ may be extended smoothly into a decreasing function for $x < 0$.

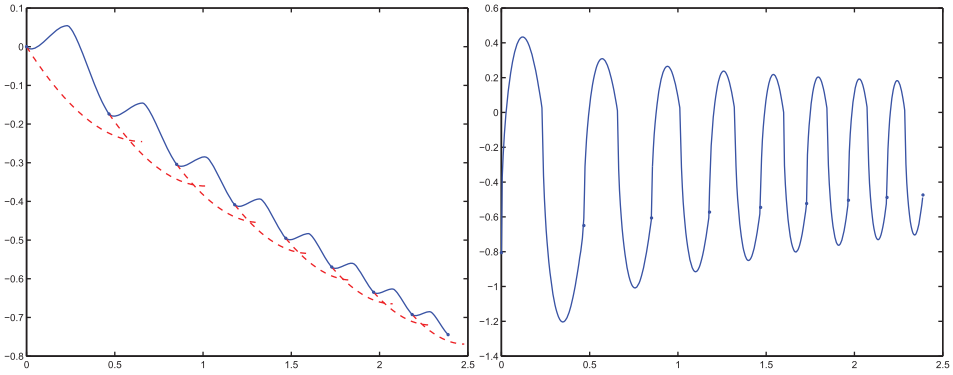

Figure 1. (Left) The shape of $f(x)$ for the first 8 successful iterations and the shape of the model at each unsuccessful iteration (dashed). (Right) The shape of $g(x)$ for the first 8 successful iterations. In both plots, we set $\beta = 0.45$, $r = 2.1$, $\tau = 0.001$, $\eta_1 = 0.6$ and $q = 3$.

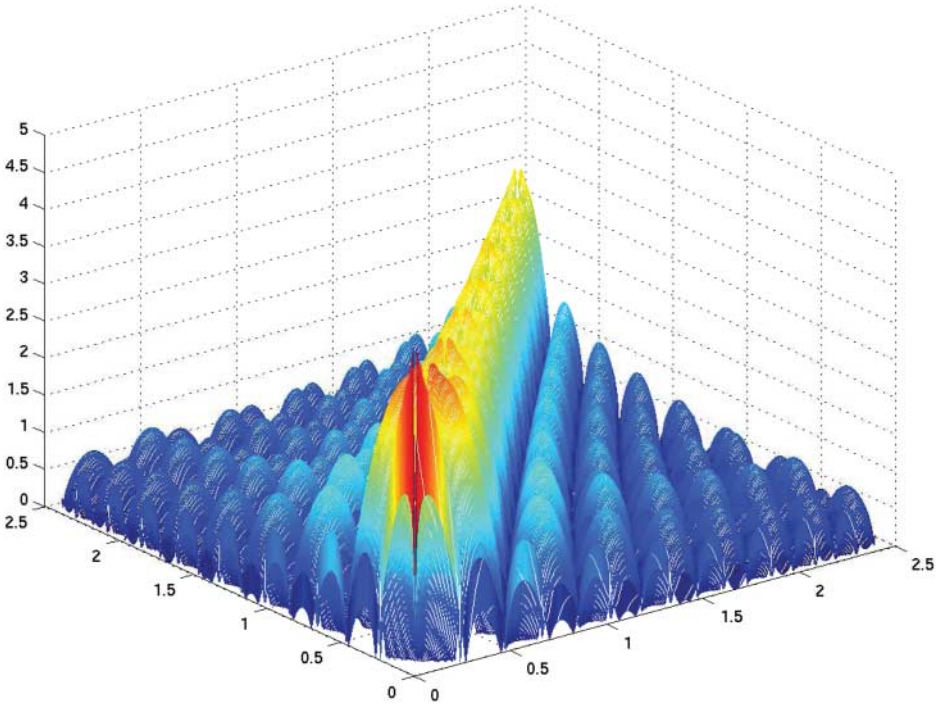

Figure 2. The shape of the function $|g(x) - g(y)|/|x - y|^{\beta}$ for the interval spanned by the first 8 successful iterations for $\beta = 0.45$, $r = 2.1$, $\tau = 0.001$, $\eta_1 = 0.6$ and $q = 3$.

As can be checked in these figures, $f(x)$ is nonconvex and continuously differentiable. The form (81) implies that $g(x)$ varies very quickly at the beginning of each interpolation interval, which is visible in Figure 1 (Right).

We show in the Appendix that assumptions AS.2–AS.5 are satisfied by the function we constructed. In particular, we prove that $g(x)$ is Hölder continuous with exponent β which we also illustrate in Figure 2. We draw the following conclusion for our example.

COROLLARY 4.1 *Let $r > 1 + \beta$ and $\tau \in (0, 1)$ arbitrarily small. Then there is a function $f(x)$ and a starting guess $x_0 = 0$ that satisfy assumptions* AS.1–AS.5 *and for which Algorithm* 2.1 *applied to the minimization of $f(x)$ starting from x_0 requires*

$$2(\lfloor \epsilon^{-(1+\beta)/(\beta+\tau(1+\beta))} \rfloor - q)$$

iterations (*and function evaluations*) *to obtain an iterate x_ϵ such that $\|g(x_\epsilon)\| \leq \epsilon$.*

Since q is independent of ϵ, this corollary shows that the complexity bound stated by part (ii) of Theorem 3.10 is essentially sharp.

5. Discussion

Figure 3 illustrates, as a function of r and β, which power of $\epsilon_* < 1$ dominates in the complexity bounds of Theorem 3.10. It is interesting to note that the worst-case evaluation complexity of our general class of regularized method does depend on the relative values of r and β. Observe also that, when $\epsilon_* < 1$, $\epsilon^{-r/(r-1)} > \epsilon^{-(1+1/\beta)}$ in the triangle for which $1 + \beta \geq r$ and $r \leq 2$. Thus, from the worst-case complexity point of view, there is little incentive to choose a regularization power $r < 2$. It is also interesting to observe that, if $r \geq 2$, the complexity no longer depends on the precise value of r, but only depends on the smoothness of the objective function as measured by the Hölder exponent β (whose knowledge is not required a priori). In that sense, the algorithm adapts itself to the problem at hand, without the need for further tuning (see also the "universal" gradient methods by Nesterov for the convex case [17]).

If $\epsilon_* \geq 1$ (that is if either $\epsilon \geq 1$ or $\kappa_{gl} \geq 1$), the results above simplify because negative powers of ϵ_* are bounded above by one. As a consequence, all terms involving such powers (which we kept explicit in the analysis for $\epsilon_* < 1$) are absorbed in the constants, and the complexity bounds of Theorem 3.10 essentially reduce to multiples of the difference $f(x_0) - f_*$.

Note also that Lemma 3.1 allows us to equate $\beta > 1$ with $\beta = 1$ and $\kappa_{gl} = \|g(x_0)\|$. In this case, either $\epsilon_* = \epsilon > \|g(x_0)\|$ and Algorithm 2.1 stops at iteration 0, or $\epsilon_* = \|g(x_0)\|$ and the

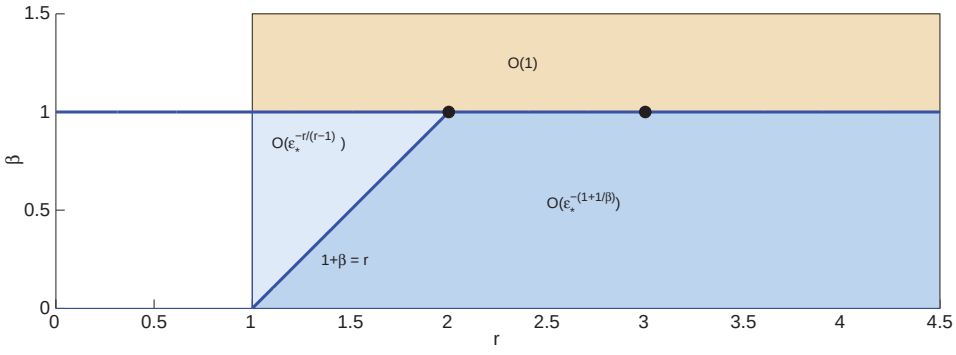

Figure 3. Worst-case evaluation complexity as a function of β and r in the cases where $\epsilon_* < 1$.

bounds of Theorem 3.10 become independent of ϵ, resulting in a bound on the number of iterations and evaluations directly proportional to $f(x_0) - f_{\text{target}}$, as expected.

We conclude by observing that the theory presented above recovers known results (see [5] for the case where $r = 3$ and $\beta = 1$ and [6,16] for the case where $r = 2$ and $\beta = 1$); these cases correspond to the thick dots in Figure 3.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by Engineering and Physical Sciences Research Council [Platform Grant EP/I01893X/1].

Note

1. Whether this can also be achieved for part (i) of the theorem is still unknown at this point.

References

[1] A. Bensoussan and J. Frehse, *Regularity Results for Nonlinear Elliptic Systems and Applications*, Springer Verlag, Heidelberg, 2002.
[2] C. Cartis, N.I.M. Gould, and Ph.L. Toint, *Trust-region and other regularization of linear least-squares problems*, BIT 49(1) (2009), pp. 21–53.
[3] C. Cartis, N.I.M. Gould, and Ph.L. Toint, *On the complexity of steepest descent, Newton's and regularized Newton's methods for nonconvex unconstrained optimization*, SIAM J. Optim. 20(6) (2010), pp. 2833–2852.
[4] C. Cartis, N.I.M. Gould, and Ph.L. Toint, *Adaptive cubic overestimation methods for unconstrained optimization. Part I: motivation, convergence and numerical results*, Math. Program. Ser. A 127(2) (2011), pp. 245–295.
[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(2) (2011), pp. 295–319.
[6] 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(4) (2011), pp. 1721–1739.
[7] C. Cartis, N.I.M. Gould, and Ph.L. Toint, *On the evaluation complexity of cubic regularization methods for potentially rank-deficient nonlinear least-squares problems and its relevance to constrained nonlinear optimization*, SIAM J. Optim. 23(3) (2013), pp. 1553–1574.
[8] O. Devolder, *Exactness, inexactness and stochasticity in first-order methods for large-scale convex optimization*, PhD Thesis, ICTEAM and CORE, Université Catholique de Louvain, 2013.
[9] N.I.M. Gould, D.P. Robinson, and H.S. Thorne, *On solving trust-region and other regularised subproblems in optimization*, Math. Program. Comput. 2(1) (2010), pp. 21–57.
[10] G.N. Grapiglia, J. Yuan, and Y. Yuan, *Global convergence and worst-case complexity of a derivative-free trust-region algorithm for composite nonsmooth optimization*, Technical report, University of Parana, Curitiba, Brasil, 2014.
[11] G.N. Grapiglia, J. Yuan, and Y. Yuan, *On the convergence and worst-case complexity of trust-region and regularization methods for unconstrained optimization*, Math. Program. Ser. A 152(1) (2015), pp. 491–520.
[12] S. Gratton, A. Sartenaer, and Ph.L. Toint, *Recursive trust-region methods for multiscale nonlinear optimization*, SIAM J. Optim. 19(1) (2008), pp. 414–444.
[13] B. Jiang, T. Lin, S. Ma, and S. Zhang, *Structured nonconvex and nonsmooth optimization: algorithms and iteration complexity analysis*, Technical Report, Optimization Online Repository, 2016.
[14] A.S. Nemirovski and D.B. Yudin, *Problem Complexity and Method Efficiency in Optimization*, Wiley Interscience Series in Discrete Mathematics, John Wiley and Sons, Chichester, 1983.
[15] Yu. Nesterov, *Introductory Lectures on Convex Optimization*, Applied Optimization, Kluwer Academic Publishers, Dordrecht, 2004.
[16] Yu. Nesterov, *Gradient methods for minimising composite objective functions*, Math. Program. Ser. A 140(1) (2013), pp. 125–161.
[17] Yu. Nesterov, *Universal gradient methods for convex optimization problems*, Technical Report DP 2013/26140, CORE, Catholic University of Louvain, Louvain-la-Neuve, Belgium, 2013.
[18] Yu. Nesterov and B.T. Polyak, *Cubic regularization of Newton method and its global performance*, Math. Program. Ser. A 108(1) (2006), pp. 177–205.
[19] Gas Processors and Suppliers Association. *Engineering Data Book. Vol. 2*. GPSA, Tulsa, OK, 1994.

[20] K. Ueda, *A regularized newton method without line search for unconstrained optimization*, PhD thesis, Department of Applied Mathematics and Physics, Graduate School of Informatics, Kyoto University, Kyoto, Japan, 2009.
[21] K. Ueda and N. Yamashita, *Convergence properties of the regularized Newton method for the unconstrained nonconvex optimization*, Appl. Math. Optim. 62(1) (2009), pp. 27–46.
[22] K. Ueda and N. Yamashita, *On a global complexity bound of the Levenberg-Marquardt method*, J. Optim. Theory Appl. 147 (2010), pp. 443–453.
[23] L.N. Vicente, *Worst case complexity of direct search*, EURO J. Comput. Optim. 1 (2013), pp. 143–153.
[24] M. Yashtini, *On the global convergence rate of the gradient descent method for functions with Hölder continuous gradients*, Optimization Letters, 2015, doi:10.1007/s11590-015-0936-x.

Appendix

In this Appendix, we prove various left-over results about the function we built by interpolation in Section 4. We first finalize this construction in the next lemma.

LEMMA A.1 *The function $f(x)$ constructed by interpolation on page 19 satisfies $\nu_{as} > 1$ for each interpolation interval in Table 1.*

Proof Consider the first such interval at iteration $2k+1$ ($k \geq 0$) and ν^1_{2k+1}, the value of ν_{as} corresponding to that interval. Using (78), we obtain that

$$\nu^1_{2k+1} = (1+\beta)\frac{f^{2k-1}_{2k+1} - f(x_{2k+1}) + \frac{1}{2}|g_{2k+1}s_{2k+1}|}{\frac{1}{2}|g_{2k+1}s_{2k+1}|} \geq 1+\beta > 1 \tag{A1}$$

as desired. For the second interpolation interval at iteration $2k+1$, we have that

$$\nu^2_{2k+1} = (1+\beta)\frac{f^{2k-2}_{2k+1} - f(x_{2k+2})}{(1-\frac{1}{2})|g_{2k+2}s_{2k+1}|} \geq (1+\beta)\frac{\frac{1}{2}|g_{2k+2}s_{2k+1}|}{\frac{1}{2}|g_{2k+2}s_{2k+1}|} = 1+\beta > 1, \tag{A2}$$

where we have used (77) to derive the inequality. We therefore obtain from (A1) and (A2) that, for all $k \geq 0$, the desired roots θ^1_{2k+1} and θ^2_{2k+1} exist and satisfy

$$\theta^1_{2k+1} \leq \frac{\pi}{2} \quad \text{and} \quad \theta^2_{2k+1} \leq \frac{\pi}{2}. \tag{A3}$$

As a consequence $\sin(\phi^i_{2k+1}t)$ is positive on each interpolation interval ($i=1,2$), and our interpolating function and its gradient are also well-defined for each interval. Moreover, since both ν^1_{2k+1} and ν^2_{2k+1} are bounded below by $1+\beta$, we obtain that there is a constant $\kappa_\theta > 0$ such that

$$\theta^1_{2k+1} \in \left[\kappa_\theta, \frac{\pi}{2}\right] \quad \text{and} \quad \theta^2_{2k+1} \in \left[\kappa_\theta, \frac{\pi}{2}\right], \tag{A4}$$

and thus that there exists a constant $\kappa_{\sin} > 0$ independent of k such that

$$\sin(\theta^1_{2k+1}) \geq \kappa_{\sin} \quad \text{and} \quad \sin(\theta^2_{2k+1}) \geq \kappa_{\sin}. \tag{A5}$$

■

We now investigate the properties of our interpolant further, and show that $f(x)$, $x \geq 0$, satisfies the properties for which the worst-case complexity upper bound applies.

THEOREM A.2 *The function $f(x)$ constructed by interpolation on page 19 satisfies assumptions* AS.2–AS.5.

Proof Note that, because of (73), (79), (69), the fact that $f(x_{2k}) = f(x_{2k+1})$ and the inequality $|g_{2k+2}s_{2k+1}| \leq |g_{2k-1}s_{2k-1}|$,

$$\begin{aligned} f^{2k-2}_{2k+1} - f(x_{2k+1}) &\leq \max[f(x_{2k-1}) - f(x_{2k+1}), f(x_{2k+2}) - \tfrac{1}{2}g_{2k+2}s_{2k+1} - f(x_{2k+1})] \\ &< \max[\Delta m_{2k-1}, \tfrac{1}{2}|g_{2k+2}s_{2k+1}|] \end{aligned}$$

$$< \max[\,|g_{2k-1}s_{2k-1}|, |g_{2k+2}s_{2k+1}|]$$
$$< |g_{2k-1}s_{2k-1}|$$

and hence that

$$\begin{aligned}
\nu^1_{2k+1} &= (1+\beta)\frac{f^{2k-2}_{2k+1} - f(x_{2k+1}) + \frac{1}{2}|g_{2k+1}s_{2k+1}|}{\frac{1}{2}|g_{2k+1}s_{2k+1}|} \\
&\le \frac{2(1+\beta)}{|g_{2k+1}s_{2k+1}|}(|g_{2k-1}s_{2k-1}| + \tfrac{1}{2}|g_{2k+1}s_{2k+1}|) \\
&< \frac{2(1+\beta)}{|g_{2k+1}s_{2k+1}|}(2|g_{2k-1}s_{2k-1}|) \\
&= 4(1+\beta)\left|\frac{g_{2k-1}}{g_{2k+1}}\right|^{(1+\beta)/\beta} \\
&\to 4(1+\beta),
\end{aligned} \tag{A6}$$

where we used (65). Similarly, using (79), (72), (69), (73) and (64) in succession, we obtain that

$$\begin{aligned}
\nu^2_{2k+1} &= (1+\beta)\frac{f^{2k-2}_{2k+1} - f(x_{2k+2})}{\frac{1}{2}|g_{2k+2}s_{2k+1}|} \\
&\le \frac{2(1+\beta)\max[\,f(x_{2k-1}) - f(x_{2k+2}), \frac{1}{2}|g_{2k+2}s_{2k+1}|]}{|g_{2k+2}s_{2k+1}|} \\
&\le \frac{2(1+\beta)\max[\,\Delta m_{2k-1} + \Delta m_{2k+1}, \frac{1}{2}|g_{2k+2}s_{2k+1}|\,]}{|g_{2k+2}s_{2k+1}|} \\
&\le 2(1+\beta)\max\left[\frac{2\Delta m_{2k-1}}{|g_{2k+2}s_{2k+1}|}, \frac{1}{2}\right]
\end{aligned}$$

and so, furthermore,

$$\begin{aligned}
\nu^2_{2k+1} &= 2(1+\beta)\max\left[2\left(1-\frac{1}{r}\right)\frac{\Delta m_{2k-1}}{\Delta m_{2k+1}}\frac{|g_{2k+1}s_{2k+1}|}{|g_{2k+2}s_{2k+1}|}, \frac{1}{2}\right] \\
&= 2(1+\beta)\max\left[2\left(1-\frac{1}{r}\right)\left|\frac{g_{2k-1}}{g_{2k+1}}\right|^{(1+\beta)/\beta}\frac{|g_{2k+1}s_{2k+1}|}{|g_{2k+2}s_{2k+1}|}, \frac{1}{2}\right] \\
&= 2(1+\beta)\max\left[2\left(1-\frac{1}{r}\right)\left|\frac{g_{2k-1}}{g_{2k+1}}\right|^{(1+\beta)/\beta}\frac{|g_{2k+1}|}{|g_{2k+2}|}, \frac{1}{2}\right] \\
&\to 2(1+\beta)\max\left[2\left(1-\frac{1}{r}\right), \frac{1}{2}\right].
\end{aligned} \tag{A7}$$

We may therefore deduce from (A6) and (A7) that there exists a constant $\kappa_\nu > 0$ independent of k such that, for all $k \ge 0$,

$$\nu^1_{2k+1} \le \kappa_\nu \quad \text{and} \quad \nu^2_{2k+1} \le \kappa_\nu.$$

As a consequence, and since the nonlinear equation in (83) can be written in the form

$$\tan(\theta) = \nu_{as}\theta,$$

we obtain that θ_{as} is uniformly bounded away from $\pi/2$ and hence that there exists a constant $\kappa_{\cos} > 0$ such that

$$\cos(\theta_{as}) = \cos(\phi_{as}s) \ge \kappa_{\cos} \tag{A8}$$

for every interpolation interval.

Consider now $0 \le t_1 < t_2 \le s$ for a given interpolation interval $[a, a+s]$. Because of (A3), we then have that

$$\begin{aligned}
|g(t_2) - g(t_1)| &= |c_{as}|(1+\beta)\phi_{as}\left\{|[\sin(\phi_{as}t_2)]^\beta\cos(\phi_{as}t_2) - [\sin(\phi_{as}t_1)]^\beta\cos(\phi_{as}t_1)|\right\} \\
&\le |c_{as}|(1+\beta)\phi_{as}\{|[\sin(\phi_{as}t_2)]^\beta\cos(\phi_{as}t_2) - [\sin(\phi_{as}t_2)]^\beta\cos(\phi_{as}t_1)| \\
&\quad + |[\sin(\phi_{as}t_2)]^\beta\cos(\phi_{as}t_1) - [\sin(\phi_{as}t_1)]^\beta\cos(\phi_{as}t_1)|\} \\
&= |c_{as}|(1+\beta)\phi_{as}\{|[\sin(\phi_{as}t_2)]^\beta|\,|\cos(\phi_{as}t_2) - \cos(\phi_{as}t_1)| \\
&\quad + |\cos(\phi_{as}t_1)||[\sin(\phi_{as}t_2)]^\beta - [\sin(\phi_{as}t_1)]^\beta|\} \\
&\le |c_{as}|(1+\beta)\phi_{as}\{|\cos(\phi_{as}t_2) - \cos(\phi_{as}t_1)| + |[\sin(\phi_{as}t_2)]^\beta - [\sin(\phi_{as}t_1)]^\beta|\}.
\end{aligned}$$

Now, using the mean-value theorem,

$$|\cos(\phi_{as}t_2) - \cos(\phi_{as}t_1)| = |\sin(\xi)|\phi_{as}|t_2 - t_1| \le \left(\frac{\pi}{2}\right)^{1-\beta} \phi_{as}^{\beta}|t_2 - t_1|^{\beta}, \tag{A9}$$

where $\xi \in (\phi_{as}t_1, \phi_{as}t_2)$ and where we have used the fact

$$\phi_{as}|t_2 - t_1| = \frac{\pi}{2}\left(\frac{2\phi_{as}|t_2 - t_1|}{\pi}\right) \le \frac{\pi}{2}\left(\frac{2\phi_{as}|t_2 - t_1|}{\pi}\right)^{\beta}$$

because $\phi_{as}|t_2 - t_1| \le \phi_{as}s \le \pi/2$. Moreover, using the inequality

$$|u^{\beta} - v^{\beta}| \le |u - v|^{\beta} \quad \text{forall } u, v \in [0, 1], \tag{A10}$$

and the fact that

$$\sin\left(\frac{\phi_{as}}{2}(t_2 - t_1)\right) < \frac{\phi_{as}}{2}(t_2 - t_1)$$

since $\phi_{as}(t_2 - t_1) \le \phi_{as}s \le \frac{\pi}{2}$, we deduce that

$$\begin{aligned}
|[\sin(\phi_{as}t_2)]^{\beta} - [\sin(\phi_{as}t_1)]^{\beta}| &\le |\sin(\phi_{as}t_2) - \sin(\phi_{as}t_1)|^{\beta} \\
&= 2^{\beta}\left|\cos\left(\frac{\phi_{as}}{2}(t_2 + t_1)\right)\right|^{\beta}\left|\sin\left(\frac{\phi_{as}}{2}(t_2 - t_1)\right)\right|^{\beta} \\
&\le 2^{\beta}\left|\sin\left(\frac{\phi_{as}}{2}(t_2 - t_1)\right)\right|^{\beta} \\
&< \phi_{as}^{\beta}|t_2 - t_1|^{\beta}.
\end{aligned}$$

Thus, combining this inequality with (A15), we obtain that

$$|g(t_2) - g(t_1)| \le \left[\left(\frac{\pi}{2}\right)^{1-\beta} + 1\right](1 + \beta)|c_{as}|\phi_{as}^{1+\beta}\,|t_2 - t_1|^{\beta}. \tag{A11}$$

But we know from (71) that, for all $k \ge 0$,

$$|g_{2k+1}| = s_{2k+1}^{\beta} \quad \text{and} \quad |g_{2k+2}| = |g_{2k+3}| = s_{2k+1}^{\beta}\frac{|g_{2k+3}|}{|g_{2k+1}|} \le s_{2k+1}^{\beta}.$$

As a consequence, we deduce using Table 1 that, for every interpolation interval,

$$|g_b - g_a| \le 2^{\beta}s^{\beta}$$

because the length s of each interval is equal to half that of the corresponding successful step. Using this inequality and (82), we obtain that

$$\begin{aligned}
|c_{as}|\phi_{as}^{1+\beta} &\le \frac{\phi_{as}^{\beta}|g_b - g_a|}{(1 + \beta)[\sin(\theta_s)]^{\beta}\cos(\theta_s)} \\
&\le \frac{2^{\beta}\phi_{as}^{\beta}s^{\beta}}{(1 + \beta)[\sin(\theta_s)]^{\beta}\cos(\theta_s)} \\
&\le \left(\frac{\pi}{2}\right)^{\beta}\frac{2^{\beta}}{(1 + \beta)[\kappa_{\sin}]^{\beta}\kappa_{\cos}},
\end{aligned} \tag{A12}$$

where we used the equality $\phi_{as}s = \theta_{as}$, (A3), (A5), and (A8) to derive the last inequality. Hence, we deduce from (A11) that, for x and y belonging to the same interpolation interval,

$$|g(x) - g(y)| \le \left[\frac{\pi}{2} + \left(\frac{\pi}{2}\right)^{\beta}\right]\frac{2^{\beta}}{[\kappa_{\sin}]^{\beta}\kappa_{\cos}}|x - y|^{\beta} \stackrel{\text{def}}{=} \frac{1}{2}L_{\beta}|x - y|^{\beta}. \tag{A13}$$

Consider now $0 \le x < y$ where x and y belong to different interpolation intervals and assume first that y belongs to the interpolation interval following that containing x. Then, if $z \in (x, y)$ is the junction point between the two successive intervals,

$$\begin{aligned}
|g(x) - g(y)| &\le |g(x) - g(z)| + |g(z) - g(y)| \\
&\le \tfrac{1}{2}L_{\beta}|x - z|^{\beta} + \tfrac{1}{2}L_{\beta}|z - y|^{\beta} \\
&\le L_{\beta}|x - y|^{\beta},
\end{aligned} \tag{A14}$$

where we use the triangle inequality, (A13) on each interval, and the fact that $u^{\beta} + v^{\beta} \le 2(u + v)^{\beta}$ for all $u, v \in [0, 1]$.

Consider finally $0 \leq x < y$ where x and y belong to different interpolation intervals, where y does not belong to the interval following that containing x. Let us denote by r_x the smallest root of g larger than x and by r_y the largest root smaller than y. Note that the existence of these roots is guaranteed by the construction of the interpolating function f which ensures that stationary point occurs at the junction between to two interpolation intervals covering a single successful step. It is easy to verify that x and r_x must belong either to the same interpolation interval or to two successive intervals. The same is true of r_y and y, yielding that

$$|x - r_x| \leq 1 \quad \text{and} \quad |r_y - y| \leq 1. \tag{A15}$$

Moreover, using either (A13) or (A14), we have that

$$|g(x) - g(r_x)| \leq L_\beta |x - r_a|^\beta \quad \text{and} \quad |g(r_y) - g(y)| \leq L_\beta |r_b - y|^\beta$$

and we may deduce, using (A15) and (A10), that

$$\begin{aligned} |g(x) - g(y)| &\leq |g(x) - g(r_x)| + |g(r_y) - g(y)| \\ &\leq L_\beta \left((r_x - x)^\beta + (y - r_y)^\beta\right) \\ &\leq L_\beta (r_x - x + y - r_y)^\beta \\ &\leq L_\beta |y - x|^\beta. \end{aligned} \tag{A16}$$

It then results from (A13), (A14) and (A16) that $g(x)$ is Hölder continuous and AS.2 is satisfied in our example.

We also note that, because of (A4), the definition of θ_{as}, the fact that $\frac{1}{2} < 1$, (71) and the decreasing nature of $\{\|g_k\|\}$, we have that, for every interpolation interval,

$$\phi_{as}^\beta > \left(\frac{\kappa_\theta}{s}\right)^\beta \geq \frac{\kappa_\theta^\beta}{|g_a|} \geq \frac{\kappa_\theta^\beta}{|g_0|}.$$

Hence (81) and (A12) ensure that $g(x)$ is bounded above for $x \geq 0$, which, together with the inequalities $f(x_k) \geq f_\infty > -\infty$, $s_k \leq 1$ and the mean-value theorem applied in each interval, guarantees that there exists a constant $f_{\text{low}} > -\infty$ such that $f(x) \geq f_{\text{low}}$ for all $x \geq 0$. Thus AS.3 holds with $f_{\text{target}} = -\infty$ and $f_* = f_{\text{low}}$. Moreover, AS.4 trivially follows with $\kappa_{gl} = 0$, $\kappa_{gu} = 1$ and $\epsilon_* = \epsilon$. AS.5 is satisfied by construction with $\kappa_B = 0$ since we set $B_k = 0$ for all $k \geq 0$. ■