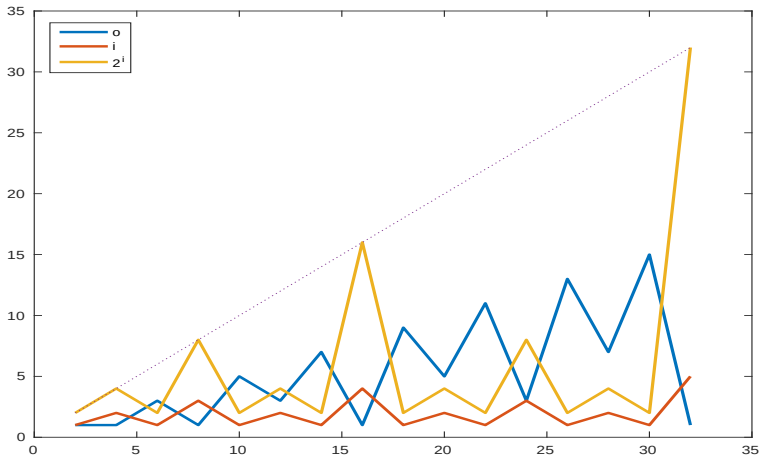
UNIVERSITÉ DE NAMUR

naxys
Namur Center for Complex Systems

Improved worst-case evaluation complexity for potentially rank-deficient nonlinear least-Euclidean-norm problems using higher-order regularized models

by C. Cartis, N. I. M. Gould and Ph. L. Toint

Report NAXYS-12-2015 17 November 2015



University of Namur, 61, rue de Bruxelles, B5000 Namur (Belgium)

http://www.unamur.be/sciences/naxys

Improved worst-case evaluation complexity for potentially rank-deficient nonlinear least-Euclidean-norm problems using higher-order regularized models

C. Cartis* N. I. M. Gould† and Ph. L. Toint‡

17 November 2015

Abstract

Given a sufficiently smooth vector-valued function $r(x)$, a local minimizer of $\|r(x)\|_2$ within a closed, non-empty, convex set $\mathcal{F}$ is sought by modelling $\|r(x)\|_2^q/q$ with a p-th order Taylor-series approximation plus a $(p+1)$-st order regularization term for given even p and some appropriate associated q. The resulting algorithm is guaranteed to find a value $\bar{x}$ for which $\|r(\bar{x})\|_2 \leq \epsilon_p$ or $\chi(\bar{x}) \leq \epsilon_d$, for some first-order criticality measure $\chi(x)$ of $\|r(x)\|_2$ within $\mathcal{F}$, using at most $O(\max\{\max(\epsilon_d, \chi_{\min})^{-(p+1)/p}, \max(\epsilon_p, r_{\min})^{-1/2^i}\})$ evaluations of $r(x)$ and its derivatives; here $r_{\min}$ and $\chi_{\min} \geq 0$ are any lower bounds on $\|r(x)\|_2$ and $\chi(x)$, respectively, and 2^i is the highest power of 2 that divides p. An improved bound is possible under a suitable full-rank assumption.

1 Introduction

Consider a given, sufficiently smooth, vector-valued function $r : \mathbb{R}^n \longrightarrow \mathbb{R}^m$. A ubiquitous problem is to find the value of x within a closed, convex, non-empty subset $\mathcal{F} \subseteq \mathbb{R}^n$ so that $\|r(x)\|$ is a small as possible, where here and elsewhere $\|\cdot\|$ is the Euclidean norm. A common approach is to consider instead the equivalent problem of minimizing

$$\Phi_q(x) := \tfrac{1}{q}\|r(x)\|^q, \tag{1.1}$$

for some integer q chosen so that Φ_q inherits the smoothness of r, with $q = 2$ being the usual least-squares choice. The problem of minimizing (1.1) is thereafter tackled using a generic method for unconstrained optimization, or one that exploits the special structure of Φ_q.

A question of interest in general smooth unconstrained optimization is how many evaluations of an objective function, $f(x)$, and its derivatives are necessary to reduce some measure of optimality below a specified (small) $\epsilon > 0$ from some arbitrary initial guess. If the measure is $\|g(x)\|$, where $g(x) := \nabla_x f(x)$, it is known that some well-known schemes (including steepest descent

*Mathematical Institute, Oxford University, Oxford OX2 6GG, Great Britain. Email: coralia.cartis@maths.ox.ac.uk.

†Numerical Analysis Group, Rutherford Appleton Laboratory, Chilton OX11 0QX, Great Britain. Email: nick.gould@stfc.ac.uk.

‡Namur Center for Complex Systems (naXys) and Department of Mathematics, University of Namur, 61, rue de Bruxelles, B-5000 Namur, Belgium. Email: philippe.toint@unamur.be.

and generic second-order trust-region methods) may require $\Theta(\epsilon^{-2})$ evaluations under standard assumptions [2], while this may be improved to $\Theta(\epsilon^{-3/2})$ evaluations for second-order methods with cubic regularization or using specialised trust-region tools [3, 8, 9]. Here and hereafter $O(\cdot)$ indicates a term that is of at worst a multiple of its argument, while $\Theta(\cdot)$ indicates additionally there are instances for which the bound holds.

For the problem we consider here, an obvious approach is to apply the aforementioned algorithms to minimize (1.1), and to terminate when

$$\nabla_x \Phi_q(x) = \|r(x)\|^{q-2} \nabla_x \Phi_2(x), \quad \text{where } \nabla_x \Phi_2(x) = J^T(x) r(x) \text{ and } J(x) := \nabla_x r(x), \tag{1.2}$$

is small. However, it has been argued that this ignores the possibility that it may suffice to stop instead when $r(x)$ is small, and that a more sensible criterion is to terminate when

$$\|r(x)\| \leq \epsilon_p \quad \text{or} \quad \|g_r(x)\| \leq \epsilon_d, \tag{1.3}$$

where ϵ_p and ϵ_d are (possibly different) required accuracy tolerances in $(0, 1)$ and where

$$g_r(x) := \begin{cases} \dfrac{\nabla_x \Phi_2(x)}{\|r(x)\|} & \text{whenever } r(x) \neq 0 \text{ and} \\ 0 & \text{otherwise.} \end{cases} \tag{1.4}$$

Note that the scaled gradient $g_r(x)$ in (1.4) is precisely the gradient of $\|r(x)\|$ whenever $r(x) \neq 0$, while if $r(x) = 0$, we are at the global minimum of r and so $g_r(x) = 0 \in \partial(\|r(x)\|)$. It has been shown that a second-order minimization method based on cubic regularization will satisfy (1.3) after at most $O\left(\max(\epsilon_d^{-3/2}, \epsilon_p^{-1/2})\right)$ evaluations [5, Thm.3.2].

Since it is easy to do so, we shall consider the more general problem in which the solution is additionally constrained to lie within a closed, convex non-empty set $\mathcal{F}$. As in [4], our generalization is based upon a suitable continuous first-order criticality measure for the constrained problem of minimizing a given function $f : \mathbb{R}^n \to \mathbb{R}$ over $\mathcal{F}$. For an arbitrary $x \in \mathcal{F}$, this criticality measure is given by

$$\chi_f(x) := \left| \min_{x+d \in \mathcal{F}, \|d\|_\chi \leq 1} \langle \nabla_x f(x), d \rangle \right|, \tag{1.5}$$

where $\langle \cdot, \cdot \rangle$ denotes the Euclidean inner product, and $\|\cdot\|_\chi$ is any fixed norm, possibly chosen to make the computation of $\chi_h(x)$ easier. Let $\kappa_n > 0$ be the norm equivalence constant such that

$$\|v\| \leq \kappa_n \|v\|_\chi \text{ for all } v \in \mathbb{R}^n. \tag{1.6}$$

Observe that $\chi_f(x)$ depends on the geometry of $\mathcal{F}$ rather than any specific parameterization using constraint functions, and that x is a first-order critical point of the problem

$$\underset{x \in \mathcal{F}}{\text{minimize}} \quad f(x) \tag{1.7}$$

if and only if $\chi_f(x) = 0$ [7, Thm.12.1.6]. Also note that $\chi_f(x) = \|\nabla_x f(x)\|$ whenever $\mathcal{F} = \mathbb{R}^n$ and $\|\cdot\|_\chi = \|\cdot\|$.

For the problem in hand, for which $f(x) = \Phi_q(x)$, and in view of (1.3)–(1.4), we prefer to judge approximate first-order critically by terminating our proposed algorithm as soon as

$$\|r(x)\| \leq \epsilon_p \quad \text{or} \quad \chi_r(x) \leq \epsilon_d, \tag{1.8}$$

where

$$\chi_r(x) := \begin{cases} \dfrac{\chi_{\Phi_2}(x)}{\|r(x)\|} & \text{whenever } r(x) \neq 0 \text{ and} \\ 0 & \text{otherwise.} \end{cases} \tag{1.9}$$

Notice in particular that these conditions are equivalent to (1.3)–(1.4) whenever $\mathcal{F} = \mathbb{R}^n$.

Very recently, it has been shown that the worst-case evaluation complexity of smooth unconstrained minimization [1] and subsequently of convexly-constrained constrained minimization [6] improves if one is prepared to model the objective at each iteration by a higher-order model and appropriate regularization. In particular, a p-th order Taylor model with a $(p+1)$ regularization term leads to at most $O(\epsilon^{-(p+1)/p})$ evaluations under standard assumptions. Thus the purpose of this short paper is to examine the consequences of this complexity breakthrough for the (convexly-constrained) norm-minimization problem, and particularly for the termination rule (1.8). Of particular note is the way that our analysis links the choice of the order p of the Taylor model to the power of the norm q in (1.1) in an unusual, number-theoretic way. Background material on our generic method for convexly-constrained minimization is presented in §2, while its application to the least-norm problem of interest here follows in §3.

2 General smooth minimization subject to convex constraints

In this section, we consider the general problem (1.7) for which we assume that $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ is p-times continuously differentiable, bounded from below, and has Lipschitz continuous p-th derivatives; for the q-th derivative of f to be Lipschitz continuous on the set $\mathcal{S} \subseteq \mathbb{R}^n$, we require that there exists a constant $L_{f,q} \geq 0$ such that, for all $x, y \in \mathcal{S}$,

$$\|\nabla_x^q f(x) - \nabla_x^q f(y)\|_T \leq (q-1)!\, L_{f,q} \|x - y\|$$

where $\|\cdot\|_T$ is the recursively induced Euclidean norm on the space of q-th order tensors. As we indicated in §1, we also assume that the feasible set $\mathcal{F}$ is closed, convex and non-empty. Note that this formulation covers standard inequality (and linear equality) constrained optimization in its different forms: the set $\mathcal{F}$ may be defined by simple bounds, and both polyhedral and more general convex constraints. We remark though that we are tacitly assuming here that the cost of evaluating constraint functions and their derivatives is negligible. Moreover, we presume that we are able to find an $x_0 \in \mathcal{F}$.

The algorithm considered in this paper is iterative. Let $T_p(x_k, s)$ be the p-th order Taylor-series approximation to $f(x_k + s)$ at some iterate $x_k \in \mathbb{R}^n$, and define the local regularized model at x_k by

$$m_k(x_k + s) := T_p(x_k, s) + \frac{\sigma_k}{p+1}\|s\|^{p+1}, \tag{2.1}$$

where $\sigma_k > 0$ is the regularization parameter. Note that $m_k(x_k) = T_p(x_k, 0) = f(x_k)$. The approach used in [4] (when $p = 2$) seeks to define a new iterate x_{k+1} from the preceding one by computing an approximate solution of the subproblem

$$\underset{x_k + s \in \mathcal{F}}{\text{minimize}} \quad m_k(x_k + s)$$

using a modified version of the Adaptive Regularization with Cubics (ARC) method for unconstrained minimization. By contrast, and in common with [6], the method we now examine is based on the ARp algorithm of [1], and our aim is to inherit its interesting features.

We now describe Algorithm 2.1, a variant of ARp for Convex Constraints, on this page.

Algorithm 2.1: Adaptive Regularization using p-th order models for minimization subject to convex constraints

A feasible starting point $x_0 \in \mathcal{F}$, an initial and a minimal regularization parameter $\sigma_0 \geq \sigma_{\min} > 0$, algorithmic parameters $\theta > 0$, $\gamma_3 \geq \gamma_2 > 1 > \gamma_1 > 0$ and $1 > \eta_2 \geq \eta_1 > 0$, are given, as well as an accuracy threshold $\epsilon \in (0, 1]$. Evaluate $f(x_0)$.
For $k = 0, 1, \ldots,$ do:

1. Evaluate $\nabla_x f(x_k)$. If
$$\chi_f(x_k) \leq \epsilon, \tag{2.2}$$
terminate with $x_\epsilon = x_k$. Otherwise compute derivatives of f of order 2 to p at x_k.

2. Compute a step s_k by approximately minimizing $m_k(x_k + s)$ over $s \in \mathcal{F}$ so that
$$x_k + s_k \in \mathcal{F}, \tag{2.3}$$
$$m_k(x_k + s_k) < m_k(x_k) \tag{2.4}$$
and
$$\chi_{m_k}(x_k + s_k) \leq \theta \, \|s_k\|^p. \tag{2.5}$$

3. Compute $f(x_k + s_k)$ and
$$\rho_k = \frac{f(x_k) - f(x_k + s_k)}{T_p(x_k, 0) - T_p(x_k, s_k)}. \tag{2.6}$$
If $\rho_k \geq \eta_1$, set $x_{k+1} = x_k + s_k$. Otherwise set $x_{k+1} = x_k$.

4. Set
$$\sigma_{k+1} \in \begin{cases} [\max(\sigma_{\min}, \gamma_1\sigma_k)\sigma_k] & \text{if } \rho_k > \eta_2 & \text{[very succesx sful iteration]} \\ [\sigma_k, \gamma_2\sigma_k] & \text{if } \eta_1 \leq \rho_k \leq \eta_2 & \text{[successful iteration]} \\ [\gamma_2\sigma_k, \gamma_3\sigma_k] & \text{otherwise.} & \text{[unsuccessful iteration],} \end{cases} \tag{2.7}$$
and go to step 2 if $\rho_k < \eta_1$.

We first state a useful property of Algorithm 2.1 that ensures that a fixed fraction of the iterations $1, 2, \ldots, k$ must be either successful or very successful.

Lemma 2.1. [1, Lem.2.4; 5, Thm.2.2]. Assume that, for some $\sigma_{\max} > 0$, $\sigma_j \leq \sigma_{\max}$ for all $0 \leq j \leq k$. Then Algorithm 2.1 ensures that

$$k \leq \kappa_u |\mathcal{S}_k|, \quad \text{where } \kappa_u := \left\lceil \left(1 + \frac{|\log \gamma_1|}{\log \gamma_2}\right) + \frac{1}{\log \gamma_2} \log\left(\frac{\sigma_{\max}}{\sigma_0}\right) \right\rceil, \tag{2.8}$$

where $\mathcal{S}_k$ is the number of successful and very successful iterations, in the sense of (2.7), up to iteration k.

We start our worst-case analysis by formalizing our assumptions.

AS.1 The objective function f is p times continuously differentiable on an open set containing $\mathcal{F}$.

AS.2 The p-th derivative of f is Lipschitz continuous on $\mathcal{F}$.

AS.3 The feasible set $\mathcal{F}$ is closed, convex and non-empty.

Algorithm 2.1 is required to start from a feasible $x_0 \in \mathcal{F}$, which, together with the fact that the subproblem solution in Step 2 involves minimization over $\mathcal{F}$, leads to AS.3.

We now recall some simple results whose proof can be found in [1] in the context of the original ARp algorithm.

Lemma 2.2. Suppose that AS.1–AS.3 hold, and that Algorithm 2.1 is applied to problem (1.7). Then, for each $k \geq 0$,

(i)
$$f(x_k + s_k) \leq T_p(x_k, s_k) + \frac{L_{f,p}}{p} \|s_k\|^{p+1} \tag{2.9}$$

and
$$\|\nabla_x f(x_k + s_k) - \nabla_s T(x_k, s_k)\| \leq L_{f,p} \|s_k\|^p; \tag{2.10}$$

(ii)
$$T_p(x_k, 0) - T_p(x_k, s_k) \geq \frac{\sigma_k}{p+1} \|s_k\|^{p+1}; \tag{2.11}$$

(iii)
$$\sigma_k \leq \sigma_{\max,f} := \max\left[\sigma_0, \frac{\gamma_3 L_{f,p}(p+1)}{p\,(1-\eta_2)}\right]. \tag{2.12}$$

Proof. See [1] for the proofs of (2.9) and (2.10), which crucially depend on AS.1 and AS.2 being valid on the segment $[x_k, x_k + s_k]$. Observe also that (2.1) and (2.4) ensure (2.11). Assume now that

$$\sigma_k \geq \frac{L_{f,p}(p+1)}{p\,(1-\eta_2)}. \tag{2.13}$$

Using (2.9) and (2.11), we may then deduce that

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

and thus that $\rho_k \geq \eta_2$. Then iteration k is very successful in that $\rho_k \geq \eta_2$ and $\sigma_{k+1} \leq \sigma_k$. As a consequence, the mechanism of the algorithm ensures that (2.12) holds. □

Next, we prove that, at successful iterations, the step at iteration k must be bounded below by a multiple of the p-th root of the criticality measure at iteration $k+1$.

Lemma 2.3. Suppose that AS.1–AS.3 hold, and that Algorithm 2.1 is applied to problem (1.7). Then

$$\|s_k\| \geq \left[\frac{\chi_f(x_{k+1})}{2\kappa_n(L_{f,p} + \theta + \sigma_{\max,f})}\right]^{\frac{1}{p}} \quad \text{for all} \quad k \in \mathcal{S}. \tag{2.14}$$

Proof. Since $k \in \mathcal{S}$ and by definition of the trial point, we have that $x_{k+1} = x_k + s_k$. Observe now that (2.10) and (2.12) imply that

$$\|\nabla f(x_{k+1}) - \nabla_x m_k(x_{k+1})\| \leq L_{f,p}\|s_k\|^p + \sigma_k\|s_k\|^p \leq (L_{f,p} + \sigma_{\max,f})\|s_k\|^p, \tag{2.15}$$

and also that

$$\begin{array}{rcl} \chi_f(x_{k+1}) & := & |\langle \nabla_x f(x_{k+1}), d_{k+1}\rangle| \\ & \leq & |\langle \nabla_x f(x_{k+1}) - \nabla_s m_k(x_{k+1}), d_{k+1}\rangle| + |\langle \nabla_s m_k(x_{k+1}), d_{k+1}\rangle|, \end{array} \tag{2.16}$$

where the first equality defines the vector d_{k+1} with

$$\|d_{k+1}\|_\chi \leq 1. \tag{2.17}$$

Assume now, for the purpose of deriving a contradiction, that (2.14) fails at iteration $k \in \mathcal{S}$. Using the Cauchy-Schwarz inequality, (1.6), (2.17), (2.15), the failure of (2.14) and the first part of (2.16) successively, we then obtain that

$$\begin{array}{rcl} \langle \nabla_s m_k(x_{k+1}), d_{k+1}\rangle - \langle \nabla_x f(x_{k+1}), d_{k+1}\rangle & & \\ & \leq & |\langle \nabla_x f(x_{k+1}), d_{k+1}\rangle - \langle \nabla_s m_k(x_{k+1}), d_{k+1}\rangle| \\ & \leq & \|\nabla_x f(x_{k+1}) - \nabla_s m_k(x_k + s_l)\|\,\|d_{k+1}\| \\ & \leq & \kappa_n(L_{p,f} + \sigma_{\max,f})\|s_k\|^p \\ & \leq & \kappa_n(L_{p,f} + \theta + \sigma_{\max,f})\|s_k\|^p \\ & \leq & \frac{1}{2}\chi_f(x_{k+1}) \\ & = & -\frac{1}{2}\langle \nabla_x f(x_{k+1}), d_{k+1}\rangle, \end{array}$$

which in turn ensures that

$$\langle \nabla_s m_k(x_{k+1}), d_{k+1}\rangle \leq \frac{1}{2}\langle \nabla_x f(x_{k+1}), d_{k+1}\rangle < 0.$$

Moreover, $x_{k+1} + d_{k+1} \in \mathcal{F}$ by definition of $\chi_f(x_{k+1})$, and hence, using (2.17),

$$|\langle \nabla_s m_k(x_{k+1}), d_{k+1} \rangle| \leq \chi_{m_k}(x_{k+1}). \tag{2.18}$$

We may then substitute this inequality in (2.16) and use the Cauchy-Schwarz inequality, (1.6) and (2.17) again to deduce that

$$\chi_f(x_{k+1}) \leq \|\nabla_x f(x_{k+1}) - \nabla_s m_k(x_{k+1})\| + \chi_{m_k}(x_{k+1}) \leq \kappa_n(L_p + \alpha + \sigma_{\max,f})\|s_k\|^p \tag{2.19}$$

where the last inequality results from (2.15), the identity $x_{k+1} = x_k + s_k$ and (2.5). But this contradicts our assumption that (2.14) fails. Hence (2.14) must hold. □

We now consolidate the previous results by deriving a lower bound on the objective function decrease at successful iterations; we note that this does not depend on the history of the algorithm, but simply on the smoothness of the objective function between x_k and x_{k+1}.

Lemma 2.4. Suppose that AS.1–AS.3 hold, and that Algorithm 2.1 is applied to problem (1.7). Then, if iteration k is successful,

$$f(x_k) - f(x_{k+1}) \geq \frac{1}{\kappa_{s,f}} \chi_f(x_{k+1})^{\frac{p+1}{p}}$$

where

$$\kappa_{s,f} := \frac{p+1}{\eta_1 \sigma_{\min}} \Big[2\kappa_n(L_{f,p} + \theta + \sigma_{\max,f}) \Big]^{\frac{p+1}{p}}. \tag{2.20}$$

Proof. If iteration k is successful, we have, using (2.6), (2.11), (2.7), (2.14) and (2.12) successively, that

$$\begin{aligned} f(x_k) - f(x_{k+1}) &\geq \eta_1 [\, T_p(x_k, 0) - T_p(x_k, s_k) \,] \\ &\geq \frac{\eta_1 \sigma_{\min}}{p+1} \|s_k\|^{p+1} \\ &\geq \frac{\eta_1 \sigma_{\min}}{(p+1)[2\kappa_n(L_{f,p} + \theta + \sigma_{\max,f})]^{\frac{p+1}{p}}} \chi_f(x_{k+1})^{\frac{p+1}{p}}. \end{aligned}$$

□

A worst-case evaluation complexity result can now be proved by combining this last result with the fact that $\chi_f(x_k)$ cannot be smaller than ϵ before termination.

Theorem 2.5. Suppose that AS.1–AS.3 hold and let $f_{\rm low}$ be a lower bound on f on $\mathcal{F}$. Then, given $\epsilon > 0$, Algorithm 2.1 applied to problem (1.7) needs at most

$$\left\lfloor \kappa_{s,f}\ [f(x_0) - f_{\rm low}]\epsilon^{-\frac{p+1}{p}} \right\rfloor$$

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

$$\kappa_{u,f} \left\lfloor \kappa_{s,f}[f(x_0) - f_{\rm low}]\epsilon^{-\frac{p+1}{p}} \right\rfloor$$

iterations in total to produce an iterate x_ϵ such that $\chi_f(x_\epsilon) \leq \epsilon$, where $\kappa_{u,f}$ is given by (2.8) with $\sigma_{\max} = \sigma_{\max,f}$ defined by (2.12).

Proof. At each successful iteration, we have, using Lemma 2.4, that

$$f(x_k) - f(x_{k+1}) \geq (\kappa_{s,f})^{-1}\chi_f(x_{k+1})^{\frac{p+1}{p}} \geq (\kappa_{s,f})^{-1}\epsilon^{\frac{p+1}{p}},$$

where we used the fact that $\chi_f(x_{k+1}) \geq \epsilon$ before termination to deduce the last inequality. Thus we deduce that, as long as termination does not occur,

$$f(x_0) - f(x_{k+1}) = \sum_{j\in\mathcal{S}_k} [f(x_j) - f(x_j + s_j)] \geq \frac{|\mathcal{S}_k|}{\kappa_{s,f}}\ \epsilon^{\frac{p+1}{p}},$$

from which the desired bound on the number of successful iterations follows. Lemma 2.1 is then invoked to compute the upper bound on the total number of iterations. □

We note that essentially the same complexity result has been established [6] for a variant of Algorithm 2.1 in which first-order criticality is measured instead by

$$\pi_f(x) := \|P_{\mathcal{F}}[x - \nabla_x f(x)] - x\|, \tag{2.21}$$

where $P_{\mathcal{F}}$ denotes the orthogonal projection onto $\mathcal{F}$, and (2.5) is replaced by $\pi_{m_k}(x_k + s_k) \leq \theta\,\|s_k\|^p$.

3 Euclidean-norm minimization subject to convex constraints

In what follows, we shall apply Algorithm 2.1 to the function $f(x) = \Phi_q(x)$ from (1.1), and replace the stopping rule (2.2) by (1.8). In this case, using Lemma 2.4, the identity (1.2), and the definition (1.9), we obtain the following key estimate.

Lemma 3.1. Suppose that $\Phi_q(x) \in C^p$ with Lipschitz continuous p-th derivatives, that AS.3 holds, and that we apply Algorithm 2.1 to $\Phi_q(x)$. Then if iteration k is successful,

$$\|r(x_k)\|^q - \|r(x_{k+1})\|^q \geq q\kappa^{-1}_{s,\Phi_q}[\chi_r(x_{k+1})]^{\frac{p+1}{p}} \cdot \|r(x_{k+1})\|^{\frac{(q-1)(p+1)}{p}}$$

where κ_{s,Φ_q} is as in Lemma 2.4 but using the Lipschitz constant $L_{\Phi_q,p}$ for Φ_q.

We now turn to the vital choice of q. We use the following elementary result.

Lemma 3.2. Let $p \in \mathbb{N}$. Then p may be expressed uniquely as the product of an odd natural number and a power of 2, i.e.,

$$p = o \cdot 2^i \tag{3.1}$$

for some odd positive integer o and $i \in \mathbb{Z}_+$.

Proof. If $p = 1$, $t = 1 \cdot 2^0$, so assume inductively that the result is true for all $p \in [1, ..., k-1]$. Then if k is odd, then $k = k \cdot 2^0$. By contrast, if k is even, then $k = 2r$ for some $1 \leq r < k$ for which $r = o \cdot 2^j$ for some odd o and $j \in \mathbb{Z}_+$. Hence $k = 2r = o \cdot 2^{j+1}$. Thus the identity holds for $p = k$ and hence for all p. Uniqueness follows from the uniqueness of the prime factorization of p. □

Note that to find o and i from p, one simply successively divides p by increasing powers of 2 until the result is no longer an integer.

Given o and i, we now choose

$$q = 1 + o \cdot (2^i - 1). \tag{3.2}$$

Observe that if p is odd, $i = 0$ and $q = 1$, while if p is even, $i > 0$ and q is even. Since we require Φ_q to be smooth, we cannot allow q to be odd, and thus we henceforth restrict our attention to even p. Armed with q, our next result improves on the estimate provided by Lemma 3.1.

Lemma 3.3. Given even p, i from (3.1), and q from (3.2), suppose that $\Phi_q(x) \in C^p$ with Lipschitz continuous p-th derivatives, that AS.3 holds, and that we apply Algorithm 2.1 to $\Phi_q(x)$. Then if $r(x_k) \neq 0$ and iteration k is successful, we have

$$\begin{aligned} &\|r(x_k)\|^{1/2^i} - \|r(x_{k+1})\|^{1/2^i} \\ &\qquad \geq \min\left\{ 2^{-i} \kappa_{s,\Phi_q}^{-1} \beta^{\frac{(q-1)(p+1)}{p}} [\chi_r(x_{k+1})]^{(p+1)/p}, (\beta^{-1/2^i} - 1)\|r(x_{k+1})\|^{1/2^i} \right\}, \end{aligned} \tag{3.3}$$

where κ_{s,Φ_q} is as in Lemma 3.1 and $\beta \in (0,1)$ is any fixed problem-independent constant.

Proof. Suppose that $r(x_k) \neq 0$, let $\beta \in (0,1)$ and denote

$$\mathcal{S}_\beta := \{k \in \mathcal{S} : \ \|r(x_{k+1})\| > \beta \|r(x_k)\|\}, \tag{3.4}$$

where

$$\mathcal{S} := \{k \geq 0 : \text{iteration } k \text{ is successful or very successful in the sense of (2.7)}\}. \tag{3.5}$$

We first analyze the function decrease for iterations $k \in \mathcal{S}_\beta$ and then, for the ones in $\mathcal{S} \setminus \mathcal{S}_\beta$.

Let $k \in \mathcal{S}_\beta$; then $r(x_{k+1}) \neq 0$ since $r(x_k) \neq 0$. From Lemma 3.1 and (3.4), we deduce

$$\begin{aligned} \|r(x_k)\|^q - \|r(x_{k+1})\|^q &\geq q\kappa_{s,\Phi_q}^{-1}[\chi_r(x_{k+1})]^{\frac{p+1}{p}} \cdot \|r(x_{k+1})\|^{\frac{(q-1)(p+1)}{p}} \\ &\geq q\kappa_{s,\Phi_q}^{-1}\beta^{\frac{(q-1)(p+1)}{p}}[\chi_r(x_{k+1})]^{\frac{p+1}{p}} \cdot \|r(x_k)\|^{\frac{(q-1)(p+1)}{p}}. \end{aligned} \tag{3.6}$$

But since $\|r(x_k)\| \geq \|r(x_{k+1})\|$, we have that

$$\begin{aligned} \|r(x_k)\|^q - \|r(x_{k+1})\|^q &= (\|r(x_k)\| - \|r(x_{k+1})\|) \cdot \textstyle\sum_{i=0}^{q-1} \|r(x_k)\|^i \|r(x_{k+1})\|^{q-i-1} \\ &\leq (\|r(x_k)\| - \|r(x_{k+1})\|) \cdot q\|r(x_k)\|^{q-1} \end{aligned}$$

and thus from (3.6) that

$$\begin{aligned} \|r(x_k)\| - \|r(x_{k+1})\| &\geq \frac{\|r(x_k)\|^q - \|r(x_{k+1})\|^q}{q\|r(x_k)\|^{q-1}} \\ &\geq \kappa_{s,\Phi_q}^{-1}\beta^{\frac{(q-1)(p+1)}{p}}[\chi_r(x_{k+1})]^{\frac{p+1}{p}} \cdot \|r(x_k)\|^{\frac{q-1}{p}} \quad \text{for all } k \in \mathcal{S}_\beta. \end{aligned} \tag{3.7}$$

Furthermore, conjugacy properties and the monotonic decrease of $\|r(x)\|$ give that

$$\begin{aligned} \|r(x_k)\|^{1/2^{j-1}} - \|r(x_{k+1})\|^{1/2^{j-1}} &= (\|r(x_k)\|^{1/2^j} + \|r(x_{k+1})\|^{1/2^j})(\|r(x_k)\|^{1/2^j} - \|r(x_{k+1})\|^{1/2^j}) \\ &\leq 2\|r(x_k)\|^{1/2^j}(\|r(x_k)\|^{1/2^j} - \|r(x_{k+1})\|^{1/2^j}) \end{aligned}$$

for all $j \geq 1$, and therefore in particular

$$\begin{aligned} \|r(x_k)\| - \|r(x_{k+1})\| &\leq 2^i\|r(x_k)\|^{1/2+\cdots+1/2^i}(\|r(x_k)\|^{1/2^i} - \|r(x_{k+1})\|^{1/2^i}) \\ &= 2^i\|r(x_k)\|^{(2^i-1)/2^i}(\|r(x_k)\|^{1/2^i} - \|r(x_{k+1})\|^{1/2^i}). \end{aligned} \tag{3.8}$$

Thus combining (3.7) and (3.8), we find that

$$\begin{aligned} \|r(x_k)\|^{1/2^i} - \|r(x_{k+1})\|^{1/2^i} &\geq \frac{\|r(x_k)\| - \|r(x_{k+1})\|}{2^i\|r(x_k)\|^{(2^i-1)/2^i}} \\ &\geq 2^{-i}\kappa_{s,\Phi_q}^{-1}\beta^{\frac{(q-1)(p+1)}{p}}[\chi_r(x_{k+1})]^{\frac{p+1}{p}} \cdot \|r(x_k)\|^{\frac{q-1}{p}-\frac{2^i-1}{2^i}}. \end{aligned} \tag{3.9}$$

But

$$\frac{q-1}{p} - \frac{2^i-1}{2^i} = \frac{q-1-\frac{p(2^i-1)}{2^i}}{p} = 0$$

from the definitions (3.1) and (3.2). Thus (3.9) gives that

$$\|r(x_k)\|^{1/2^i} - \|r(x_{k+1})\|^{1/2^i} \geq 2^{-i}\kappa_{s,\Phi_q}^{-1}\beta^{\frac{(q-1)(p+1)}{p}}[\chi_r(x_{k+1})]^{\frac{p+1}{p}} \quad \text{for all } k \in \mathcal{S}_\beta. \tag{3.10}$$

Conversely, let $k \in \mathcal{S} \setminus \mathcal{S}_\beta$, which gives

$$\|r(x_{k+1})\| \leq \beta\|r(x_k)\|, \text{ and thus } \|r(x_{k+1})\|^{1/2^i} \leq \beta^{1/2^i}\|r(x_k)\|^{1/2^i}, \tag{3.11}$$

and so the residuals decrease linearly on such iterations. It follows from (3.11) that on such iterations we have the following function decrease

$$\begin{aligned} \|r(x_k)\|^{1/2^i} - \|r(x_{k+1})\|^{1/2^i} &\geq (1-\beta^{1/2^i})\|r(x_k)\|^{1/2^i} \\ &\geq \frac{1-\beta^{1/2^i}}{\beta^{1/2^i}}\|r(x_{k+1})\|^{1/2^i} \quad \text{for all } k \in \mathcal{S} \setminus \mathcal{S}_\beta. \end{aligned} \tag{3.12}$$

(Note that (3.12) continues to hold if $r(x_{k+1}) = 0$.) The bound (3.3) now follows from (3.10) and (3.12). □

The following theorem is our main result, a general evaluation complexity bound for Algorithm 2.1 applied to (1.1) when the termination condition (1.8) is employed.

Theorem 3.4. Given even p, i from (3.1), and q from (3.2), suppose that $\Phi_q(x) \in C^p$ with Lipschitz continuous p-th derivatives, that AS.3 holds, and that we apply Algorithm 2.1 to $\Phi_q(x)$, with the termination condition (1.8) replacing (2.2). Then the algorithm terminates after at most

$$k_\epsilon + 1, \quad \text{where} \quad k_\epsilon := \left\lceil \max\{\kappa_1 \max(\epsilon_d, \chi_{\min})^{-(p+1)/p}, \kappa_2 \max(\epsilon_p, r_{\min})^{-1/2^i}\} \right\rceil, \tag{3.13}$$

successful iterations—or, equivalently, derivative evaluations—and at most

$$\kappa_u k_\epsilon + 1 \tag{3.14}$$

total (successful and unsuccessful) iterations—or, equivalently, residual-evaluations, where

$$\kappa_1 := (\|r(x_0)\|^{1/2^i} - r_{\min}^{1/2^i}) 2^i \kappa_{s,\Phi_q} \beta^{-\frac{(q-1)(p+1)}{p}} \quad \text{and} \quad \kappa_2 := (\|r(x_0)\|^{1/2^i} - r_{\min}^{1/2^i})(\beta^{-1/2^i} - 1)^{-1}, \tag{3.15}$$

κ_{s,Φ_q} is as in Lemma 3.1, $r_{\min} \geq 0$ and $\chi_{\min} \geq 0$ are any lower bounds on $\|r(x)\|$ and $\chi_{\Phi_2}(x)/\|r(x)\|$, respectively, that are independent of ϵ_p and ϵ_d, and $\beta \in (0,1)$ is a fixed problem-independent constant.

Proof. Clearly, if (1.3) is satisfied at the starting point, there is nothing left to prove. Assume now that (1.3) fails at $k = 0$. For any iteration $(k+1)$ at which the algorithm does not terminate, it follows from (1.3) that we have

$$\|r(x_{k+1})\| > \max(\epsilon_p, r_{\min}) \quad \text{and} \quad \chi_r(x_{k+1}) > \max(\epsilon_d, \chi_{\min}). \tag{3.16}$$

From (3.3) and (3.16), we deduce

$$\begin{aligned} &\|r(x_k)\|^{1/2^i} - \|r(x_{k+1})\|^{1/2^i} \geq \\ &\quad \min\left\{ 2^{-i} \kappa_{s,\Phi_q}^{-1} \beta^{\frac{(q-1)(p+1)}{p}} \max(\epsilon_d, \chi_{\min})^{(p+1)/p}, (\beta^{-1/2^i} - 1) \max(\epsilon_p, r_{\min})^{1/2^i} \right\} \end{aligned} \tag{3.17}$$

for all $k \in \mathcal{S}$ for which (3.16) holds. Summing up (3.17) over all iterations $k \in \mathcal{S}$ for which (3.16) holds, with say $j_\epsilon \leq \infty$ as the largest index, and using that the iterates are unchanged over unsuccessful iterations, we obtain

$$\begin{aligned} \|r(x_0)\|^{1/2^i} - \|r(x_{j_\epsilon})\|^{1/2^i} &= \sum_{k=0, k\in\mathcal{S}}^{j_\epsilon - 1} \left[\|r(x_k)\|^{1/2^i} - \|r(x_{k+1})\|^{1/2^i} \right] \\ &\geq |\mathcal{S}_\epsilon| \min\left\{ 2^{-i} \kappa_{s,\Phi_q}^{-1} \beta^{\frac{(q-1)(p+1)}{p}} \max(\epsilon_d, \chi_{\min})^{(p+1)/p}, (\beta^{-1/2^i} - 1) \max(\epsilon_p, r_{\min})^{1/2^i} \right\} \end{aligned} \tag{3.18}$$

where $|\mathcal{S}_\epsilon|$ denotes the number of successful iterations up to iteration j_ϵ. As $\|r(x_{j_\epsilon})\|^{1/2^i}$

$\geq r_{\min}^{1/2^i}$, (3.18) ensures that $j_\epsilon < \infty$ and that

$$|\mathcal{S}_\epsilon| \leq \frac{\|r(x_0)\|^{1/2^i} - r_{\min}^{1/2^i}}{\min\left\{2^{-i}\kappa_{s,\Phi_q}^{-1}\beta^{\frac{(q-1)(p+1)}{p}}\max(\epsilon_d,\chi_{\min})^{(p+1)/p}, (\beta^{-1/2^i}-1)\max(\epsilon_p, r_{\min})^{1/2^i}\right\}},$$

which gives (3.13) since $|\mathcal{S}_\epsilon|$ must be an integer and since the termination condition is checked at the next iteration; see [3, (5.21), (5.22)] for full details. To derive (3.14), we apply Lemma 2.1 and recall that ϵ_p, $\epsilon_d \in (0,1)$. □

The best bound in Theorem 3.4 occurs when p is a power of 2, since then $q = p = 2^i$ for some i. We state this as follows.

Corollary 3.5. Suppose that p is a power of two, that $\Phi_p(x) \in C^p$ with Lipschitz p-th derivatives, that AS.3 holds, and that we apply Algorithm 2.1 to $\Phi_p(x)$, with the termination condition (1.8) replacing (2.2). Then Algorithm 2.1 terminates after at most

$$k_\epsilon + 1, \quad \text{where} \quad k_\epsilon := \left\lceil \max\{\kappa_1 \max(\epsilon_d, \chi_{\min})^{-(p+1)/p}, \kappa_2 \max(\epsilon_p, r_{\min})^{-1/p}\}\right\rceil,$$

successful iterations—or, equivalently, derivative evaluations—and at most $\kappa_u k_\epsilon + 1$ total (successful and unsuccessful) iterations—or, equivalently, residual-evaluations, where

$$\kappa_1 := (\|r(x_0)\|^{1/p} - r_{\min}^{1/p})p\kappa_{s,\Phi_q}\beta^{-(p^2-1)/p} \quad \text{and} \quad \kappa_2 := (\|r(x_0)\|^{1/p} - r_{\min}^{1/p})(\beta^{-1/p}-1)^{-1},$$

κ_{s,Φ_q} is as in Lemma 3.1, $r_{\min} \geq 0$ and $\chi_{\min} \geq 0$ are any lower bounds on $\|r(x)\|$ and $\chi_{\Phi_2}(x)/\|r(x)\|$, respectively, that are independent of ϵ_p and ϵ_d, and $\beta \in (0,1)$ is a fixed problem-independent constant.

A much stronger result is possible if the lower bound $\chi_{\min}$ on $\chi_{\Phi_2}(x)/\|r(x)\|$ in the statement of Theorem 3.4 is strictly positive. To see this, we show that such a restriction implies that the sequence $\{\|r(x_k)\|\}_{k\geq k_0}$ decreases linearly on successful iterations once $\|r(x_{k_0})\|$ is sufficiently small.

Theorem 3.6. Given even p, i from (3.1), and q from (3.2), suppose that $\Phi_q(x) \in C^p$ with Lipschitz p-th derivatives, that AS.3 holds, that we apply Algorithm 2.1 to $\Phi_q(x)$ with the termination condition (1.8) replacing (2.2), and that

$$\chi_r(x_k) \geq \chi_{\min} > 0 \text{ for all } k \text{ until termination.} \tag{3.19}$$

Then Algorithm 2.1 terminates after at most

$$k_\epsilon + 1 \tag{3.20}$$

successful iterations—or, equivalently, derivative evaluations—and at most

$$\kappa_u k_\epsilon + 1 \tag{3.21}$$

total (successful and unsuccessful) iterations—or, equivalently, residual-evaluations, where

$$k_\epsilon := \begin{cases} \lceil \max\{\kappa_1 \max(\epsilon_d, \chi_{\min})^{-(p+1)/p}, \kappa_2 \rho^{-1}\} \rceil + \left\lceil |\log_\beta(\epsilon_p/\rho^{2^i})| \right\rceil & \text{if } \epsilon_p < \rho^{2^i} \\ \left\lceil \max\{\kappa_1 \max(\epsilon_d, \chi_{\min})^{-(p+1)/p}, \kappa_2 \epsilon_p^{-1/2^i}\} \right\rceil & \text{otherwise,} \end{cases} \tag{3.22}$$

$$\rho := 0.99 \cdot 2^{-i} \kappa_{s,\Phi_q}^{-1} \beta^{\frac{(q-1)(p+1)}{p}} \chi_{\min}^{\frac{(p+1)}{p}}, \tag{3.23}$$

$$\kappa_1 := (\|r(x_0)\|^{1/2^i} - \rho^{1/2^i}) 2^i \kappa_{s,\Phi_q} \beta^{-\frac{(q-1)(p+1)}{p}} \text{ and } \kappa_2 := (\|r(x_0)\|^{1/2^i} - \rho^{1/2^i})(\beta^{-1/2^i} - 1)^{-1}, \tag{3.24}$$

κ_{s,Φ_q} is as in Lemma 3.1 and $\beta \in (0,1)$ is a fixed problem-independent constant.

Proof. First, observe that since Theorem 3.4 shows that Algorithm 2.1 ensures (1.8) for any given $\epsilon_p, \epsilon_d > 0$, and as (3.19) forces $\chi_r(x_k) > \epsilon_d$ whenever $\epsilon_d < \chi_{\min}$, it must be that the algorithm terminates with $\|r(x)\| \leq \epsilon_p$. As this is true for arbitrary ϵ_p, we conclude that $\|r(x)\|$ may be made arbitrarily small within $\mathcal{F}$ by picking ϵ_p appropriately small, and thus certainly $r_{\min} = 0$.

Now consider the set $\mathcal{S}_\beta$ just as in (3.4) in the proof of Lemma 3.3, and suppose that k_0 is the smallest k for which

$$\|r(x_k)\| \leq \rho^{2^i}. \tag{3.25}$$

where ρ is defined in (3.23). Then if $k \in \mathcal{S}_\beta$, (3.10) holds, and thus

$$\|r(x_k)\|^{1/2^i} - \|r(x_{k+1})\|^{1/2^i} \geq 2^{-i} \kappa_{s,\Phi_q}^{-1} \beta^{\frac{(q-1)(p+1)}{p}} \chi_{\min}^{\frac{(p+1)}{p}} > \rho \geq \|r(x_k)\|^{1/2^i}. \tag{3.26}$$

because of (3.19) and (3.25). Since this then implies $\|r(x_{k+1})\|^{1/2^i} < 0$ which is impossible, we must have that $k \in \mathcal{S} \setminus \mathcal{S}_\beta$ for all successful $k \geq k_0$, and thus

$$\|r(x_{k+1})\| \leq \beta \|r(x_k)\| \text{ for all } k \geq k_0 \in \mathcal{S}. \tag{3.27}$$

We may then invoke Theorem 3.4 to deduce that Algorithm 2.1 will achieve

$$\|r(x_k)\| \leq \max(\epsilon_p, \rho^{2^i}) \quad \text{or} \quad \chi_r(x_k) \leq \epsilon_d$$

after at most

$$k_\rho + 1, \;\; \text{where} \;\; k_\rho := \left\lceil \max\{\kappa_1 \max(\epsilon_d, \chi_{\min})^{-(p+1)/p}, \kappa_2 \max(\epsilon_p, \rho^{2^i})^{-1/2^i}\} \right\rceil,$$

successful iterations, and thus from Lemma 2.1 that

$$k_0 \leq \kappa_u k_\rho + 1,$$

where κ_1 and κ_2 are given by (3.24). If $\epsilon_p < \rho^{2^i}$, the linear rate of convergence of subsequent successful iterations from (3.27) implies that most a further $\left\lceil |\log_\beta(\epsilon_p/\rho^{2^i})| \right\rceil$ successful iterations will be required. This gives (3.20), and (3.21) then follows from Lemma 2.1. □

The reader might be concerned that (3.22) indicates an evaluation bound dependence on ϵ_d and ϵ_p via $\max(\epsilon_d, \chi_{\min})^{-(p+1)/p}$ and $\epsilon_p^{-1/2^i}$ when $\epsilon_p \geq \rho^{2^i}$, but of course the (weaker) bounds

$$\max(\epsilon_d, \chi_{\min})^{-(p+1)/p} \leq \chi_{\min}^{-(p+1)/p} \;\; \text{and} \;\; \epsilon_p^{-1/2^i} \leq \rho^{-1}$$

in this case should allay any such fears. The only true dependence on either tolerance is via the very mild term $\lceil |\log_\beta(\epsilon_p/\rho^{2^i})| \rceil$ when $\epsilon_p < \rho^{2^i}$.

When $\mathcal{F} = \mathbb{R}^n$, the condition (3.19) is of course nothing other than the well-know assumption that the smallest singular value of $J(x)$ is bounded away from zero at points encountered. It is also easy to infer directly from Lemma 3.1 that

$$\|r(x_k)\|^q \geq q\kappa_{s,\Phi_q}^{-1} \chi_{\min}^{\frac{(p+1)}{p}} \|r(x_k + s_k)\|^{\frac{(q-1)(p+1)}{p}},$$

or equivalently that

$$\|r(x_k + s_k)\| \leq [q\kappa_{s,\Phi_q}^{-1}]^{\frac{-p}{(q-1)(p+1)}} \chi_{\min}^{\frac{-1}{(q-1)}} \|r(x_k)\|^{\frac{pq}{(q-1)(p+1)}},$$

when condition (3.19) holds, and thus from (3.2), that the successful iterates then ultimately converge superlinearly (with Q-factor $pq/(q-1)(p+1) = pq/(pq-o)$, or $p^2/(p^2-1)$ when p is a power of 2); such a result is given by [9, Thm.7] in the case $p = 2$. While this may lead us to improve the iteration count marginally over that given in Theorem 3.6, we feel such a general result is uninformative and the effort needed is thus unwarranted.

Once again, the best bound in Theorem 3.6 occurs when p is a power of 2, as follows.

Corollary 3.7. Suppose that p is a power of two, that $\Phi_p(x) \in C^p$ with Lipschitz p-th derivatives, that AS.3 holds, that we apply Algorithm 2.1 to $\Phi_p(x)$ with the termination condition (1.8) replacing (2.2), and that (3.19) holds. Then Algorithm 2.1 terminates after at most $k_\epsilon + 1$ successful iterations—or, equivalently, derivative evaluations—and at most $\kappa_u k_\epsilon + 1$ total (successful and unsuccessful) iterations—or, equivalently, residual-evaluations, where

$$k_\epsilon := \begin{cases} \lceil \max\{\kappa_1 \max(\epsilon_d, \chi_{\min})^{-(p+1)/p}, \kappa_2 \rho^{-1}\} \rceil + \lceil |\log_\beta(\epsilon_p/\rho^p)| \rceil & \text{if } \epsilon_p < \rho^p \\ \left\lceil \max\{\kappa_1 \max(\epsilon_d, \chi_{\min})^{-(p+1)/p}, \kappa_2 \epsilon_p^{-1/p}\} \right\rceil & \text{otherwise,} \end{cases}$$

$$\rho := 0.99 \cdot p^{-1} \kappa_{s,\Phi_q}^{-1} \beta^{(p^2-1)/p} \chi_{\min}^{(p+1)/p},$$

$$\kappa_1 := (\|r(x_0)\|^{1/p} - \rho^{1/p}) p \kappa_{s,\Phi_q} \beta^{-(p^2-1)/p} \text{ and } \kappa_2 := (\|r(x_0)\|^{1/p} - \rho^{1/p})(\beta^{-1/p} - 1)^{-1},$$

κ_{s,Φ_p} is as in Lemma 3.1 and $\beta \in (0,1)$ is a fixed problem-independent constant.

Finally, we comment on our blanket assumption that $\Phi_q(x) \in C^p$ with Lipschitz p-th derivatives. These allow us to specify the constants in our complexity bounds very succinctly, but it is also straightforward to show that the assumptions are implied by the more straightforward assumption that

AS.4 each component $r_i(x)$ of $r(x)$ is p-times continuously differentiable, and each derivative of order 0 to p is Lipschitz continuous.

To see why this might be so, since p is even, so is q and we then have $q = 2j$ for some integer j. Thus

$$\Phi_q(x) = \frac{2^{j-1}}{j} [\Phi_2(x)]^j. \tag{3.28}$$

We note that assumptions AS.1 and AS.2 were only actually required to hold for $f(x)$ in Lemma 2.2—and thus to establish the complexity bounds in Theorem 2.5—on segments $[x_k, x_k + s_k]$ generated by the algorithm, and thus the same is required of $\Phi_q(x)$ to establish Theorems 3.4 and 3.6. It has been shown [6, Lem.3.1] that AS.4 implies that $\Phi_2(x)$ and its p derivatives are Lipschitz on $[x_k, x_k + s_k]$. The identity (3.28) implies that to show the same for $\Phi_q(x)$ simply requires that we use the chain rule on $[\Phi_2(x)]^j$. The details are fiendishly complicated (and omitted); the p derivatives of $\Phi_q(x)$ involve weighted sums of products of $\Phi_2(x)$ and its derivatives up to order p, and just as in the proof of [6, Lem.3.1], Lipschitz continuity and boundedness of derivatives of $\Phi_2(x)$ that are implied by [6, Lem.3.1], together with boundedness of $\Phi_2(x_k)$ that follow since the algorithm generates monotonically-decreasing $\|r(x)\|$, give the result.

4 Comments and conclusions

We have demonstrated that it is possible to design an algorithm for least-Euclidean-norm minimization that is guaranteed to find a value $\bar{x}$ for which $\|r(\bar{x})\|_2 \leq \epsilon_p$ or $\|\partial(\|r(\bar{x})\|_2)\|_2 \leq \epsilon_d$ using at most $O(\max(\{\max(\epsilon_d, \chi_{\min})^{-(p+1)/p}, \max(\epsilon_p, r_{\min})^{-1/2^i}))$ evaluations of $r(x)$ and its derivatives; here $r_{\min} \geq 0$ is any lower bound on $\|r(x)\|_2$, $\chi_{\min} \geq 0$ is any lower bound on $\chi_{\Phi_2}(x)/\|r(x)\|$, and

2^i is the highest power of 2 that divides p. The algorithm relies on using a globally-convergent algorithm to approximately minimize a model $m(x,s,\sigma) := T_p(x,s) + \frac{\sigma}{p+1}\|s\|_2^{p+1}$ in which $T_p(x,s)$ be the p-th order Taylor-series approximation to $\frac{1}{q}\|r(x)\|_2^q$, $q = 1 + p(2^i-1)/2^i$, and $\sigma > 0$ is an iteration-dependent parameter. The evaluation-complexity bound may be improved significantly if a suitable full-rank assumption holds as then the algorithm will ultimately converge (super)linearly.

The reader may be interested in the relationship between a given p and the implied o, i (and 2^i) from (3.1) and q from (3.2). We illustrate these for the first 16 positive even integers:

p	2	4	6	8	10	12	14	16	18	20	22	24	26	28	30	32
o	1	1	3	1	5	3	7	1	9	5	11	3	13	7	15	1
i	1	2	1	3	1	2	1	4	1	2	1	3	1	2	1	5
2^i	2	4	2	8	2	4	2	16	2	4	2	8	2	4	2	32
q	2	4	4	8	6	10	8	16	10	16	12	22	14	22	16	32

It is easy to show that $1+p/2 \leq q \leq p$, and clearly $q = p$ if and only if $p = 2^i$; the best complexity bounds in Theorems 3.4 and 3.6 occur when p is a power of two.

Acknowledgements

The work of the second author was supported by EPSRC grants EP/I013067/1 and EP/M025179/1. The third author gratefully acknowledges the financial support of the Belgian Fund for Scientific Research, the Leverhume Trust and Balliol College (Oxford).

References

[1] E. G. Birgin, J.L. Gardenghi, J.M. Martinez, S.A. Santos, and Ph. L. Toint. Worst-case evaluation complexity for unconstrained nonlinear optimization using high-order regularized models. Report naXys-05-2015, University of Namur, Belgium, 2015.

[2] C. Cartis, N. I. M. Gould, and Ph. L. Toint. On the complexity of steepest descent, Newton's method and regularized Newton's methods for nonconvex unconstrained optimization problems. *SIAM Journal on Optimization*, 20(6):2833–2852, 2010.

[3] C. Cartis, N. I. M. Gould, and Ph. L. Toint. Adaptive cubic regularisation methods for unconstrained optimization. Part II: worst-case function and derivative-evaluation complexity. *Mathematical Programming, Series A*, 130(2):295–319, 2011.

[4] 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 Journal of Numerical Analysis*, 32(4):1662–1695, 2012.

[5] 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 Journal on Optimization*, 23(3):1553–1574, 2013.

[6] C. Cartis, N. I. M. Gould, and Ph. L. Toint. Evaluation complexity bounds for smooth constrained nonlinear optimization using scaled KKT conditions and high-order models. Report naXys-11-2015(R1), University of Namur, Belgium, 2015.

[7] A. R. Conn, N. I. M. Gould, and Ph. L. Toint. *Trust-Region Methods*. SIAM, Philadelphia, 2000.

[8] F. E. Curtis and D. P. Robinson. A trust region algorithm with a worst-case iteration complexity of $O(\epsilon^{-3/2})$ for nonconvex optimization. COR@L Technical Report 14T-009, Lehigh University,, Bethlehem, PA, USA, 2014.

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