Partially separable convexly-constrained optimization with non-Lipschitzian singularities and its complexity

X. Chen* Ph. L. Toint† and H. Wang‡

19 April 2017

Abstract

An adaptive regularization algorithm using high-order models is proposed for partially separable convexly constrained nonlinear optimization problems whose objective function contains non-Lipschitzian ℓ_q-norm regularization terms for $q \in (0,1)$. It is shown that the algorithm using an p-th order Taylor model for p odd needs in general at most $O(\epsilon^{-(p+1)/p})$ evaluations of the objective function and its derivatives (at points where they are defined) to produce an ϵ-approximate first-order critical point. This result is obtained either with Taylor models at the price of requiring the feasible set to be 'kernel-centered' (which includes bound constraints and many other cases of interest), or for non-Lipschitz models, at the price of passing the difficulty to the computation of the step. Since this complexity bound is identical in order to that already known for purely Lipschitzian minimization subject to convex constraints [9], the new result shows that introducing non-Lipschitzian singularities in the objective function may not affect the worst-case evaluation complexity order. The result also shows that using the problem's partially separable structure (if present) does not affect complexity order either. A final (worse) complexity bound is derived for the case where Taylor models are used with a general convex feasible set.

Keywords: complexity theory, nonlinear optimization, non-Lipschitz function, ℓ_q-norm regularization, partially separable problems.

1 Introduction

We consider the partially separable convexly constrained nonlinear optimization problem:

$$\min_{x\in\mathcal{F}} f(x) = \sum_{i\in\mathcal{N}} f_i(U_i x) + \sum_{i\in\mathcal{H}} |U_i x|^q = \sum_{i\in\mathcal{N}} f_i(x_i) + \sum_{i\in\mathcal{H}} f_i(x_i) \tag{1.1}$$

where $\mathcal{F}$ is a non-empty closed convex set, $\mathcal{N}\cup\mathcal{H} \stackrel{\rm def}{=} \mathcal{M}$, $\mathcal{N}\cap\mathcal{H}=\emptyset$, $f_i : \mathbb{R}^n \to \mathbb{R}$, $q\in(0,1)$, $f_i(x_i) = |x_i|^q = |U_i x|^q$ for $i \in \mathcal{H}$ and where, for $i \in \mathcal{M}$, $x_i \stackrel{\rm def}{=} U_i x$ with U_i a (fixed) $n_i \times n$

*Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong. Email: xiaojun.chen@polyu.edu.hk

†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

‡Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong. Email: hong.wang@connect.polyu.hk

matrix with $n_i \leq n$. Without loss of generality, we assume that $\|U_i\| = 1$ for all $i \in \mathcal{M}$ and that the ranges of the U_i^T for $i \in \mathcal{N}$ span $\mathbb{R}^n$ in that the intersection of the nullspaces of the U_i is reduced to the origin(1). In what follows, the "element functions" f_i ($i \in \mathcal{N}$) will be "well-behaved" smooth functions with Lipschitz continuous derivatives(2). If $\mathcal{H} \neq \emptyset$, we also require that

$$n_i = 1 \tag{1.2}$$

and (initially at least(3)) that the feasible set is 'kernel centered', in the sense that, if $P_{\mathcal{X}}[\cdot]$ is the orthogonal projection ont the convex set $\mathcal{X}$, then, for $i \in \mathcal{H}$,

$$P_{\ker(U_i)}[\mathcal{F}] \subseteq \mathcal{F} \quad \text{whenever} \quad \ker(U_i) \cap \mathcal{F} \neq \emptyset \tag{1.3}$$

in addition of $\mathcal{F}$ being convex, closed and non-empty. As will be discussed below (after Lemma 4.2), we may assume without loss of generality that, $\ker(U_i) \cap \mathcal{F} \neq \emptyset$ (and thus $P_{\ker(U_i)}[\mathcal{F}] \subseteq \mathcal{F}$) for all $i \in \mathcal{H}$. 'Kernel centered' feasible sets include boxes (corresponding to bound constrained problems), spheres/cylinders centered at the origin and other sets such as

$$\left\{(x_1, x_2) \in \mathbb{R}^{n_1+1} \mid x_1 \in \mathcal{F}_1 \text{ and } g_1(x_1) \leq x_2 \leq -g_2(x_1)\right\}, \quad \mathcal{H} = \{2\}, \tag{1.4}$$

where $\mathcal{F}_1$ is a non-empty closed convex set in $\mathbb{R}^{n_1}$ and $g_i(\cdot)$ are convex functions from $\mathbb{R}^{n_1}$ to $\mathbb{R}$ ($i = 1, 2,\ n_1 + 1 \leq n$) such that $g_i(x_1) \leq 0$ ($i = 1, 2$) for $x_1 \in \mathcal{F}_1$. Compositions using (1.4) recursively, rotations, cartesian products or intersections of such sets are also kernel-centered.

Problem (1.1) has many applications in engineering and science. Using the non-Lipschitz regularization function in the second term of the objective function f has remarkable advantages for the restoration of piecewise constant images and sparse signals [1, 4, 28], and sparse variable selection, for instance in bioinformatics [14, 27]. Theory and algorithms for solving q-norm regularized optimization problems have been developed in [12, 15, 29].

The partially separable structure defined in problem (1.1) is ubiquitous in applications of optimization. It is most useful in the frequent case where $n_i \ll n$ and subsumes that of sparse optimization (in the special case where the rows of each U_i are selected rows of the identity matrix). Moreover the decomposition in (1.1) has the advantage of being invariant for linear changes of variables (only the U_i matrices vary).

Partially separable optimization was proposed by Griewank and Toint in [26], studied for more than thirty years (see [10, 11, 20, 21, 30] for instance) and extensively used in the popular CUTE(st) testing environment [23] as well as in the AMPL [19], LANCELOT [17] and FILTRANE [24] packages, amongst others. In particular, the design of trust-region algorithms exploiting the partially separable decomposition (1.1) was investigated by Conn, Gould, Sartenaer and Toint in [16, 18] and Shahabuddin [33].

Focussing now on the nice multivariate element functions ($i \in \mathcal{N}$), we note that using the partially separable nature of a function f can be very useful if one wishes to use derivatives of

$$f_{\mathcal{N}}(x) \stackrel{\text{def}}{=} \sum_{i \in \mathcal{N}} f_i(U_i x) = \sum_{i \in \mathcal{N}} f_i(x_i) \tag{1.5}$$

(1)If the $\{U_i^T\}_{i \in \mathcal{N}}$ do not span $\mathbb{R}^n$, problem (1.1) can be modified without altering its optimal value by introducing an additional identically zero element term $f_0(U_0 x)$ (say) in $\mathcal{N}$ with associated U_0 such that $\cap_{i \in \mathcal{N}} \ker(U_i) \subseteq \operatorname{range}(U_0^T)$. It is clear that, since $f_0(x_0) = 0$, it is differentiable with Lipschitz continuous derivative for any order $p \geq 1$. Obviously, this covers the case where $\mathcal{N} = \emptyset \neq \mathcal{H}$.

(2)Hence the symbol $\mathcal{N}$ for "nice".

(3)We will drop this assumption in Section 5.

of order larger than one in the context of the p-th order Taylor series

$$T_{f_{\mathcal{N}},p}(x,s) = f_{\mathcal{N}}(x) + \sum_{j=1}^{p} \frac{1}{j!} \nabla_x^j f_{\mathcal{N}}(x)[s]^j. \tag{1.6}$$

Indeed, it may be verified that

$$\nabla_x^j f_{\mathcal{N}}(x)[s]^j = \sum_{i\in\mathcal{N}} \nabla_{x_i}^j f_i(x_i)[U_i s]^j. \tag{1.7}$$

This last expression indicates that only the $|\mathcal{N}|$ tensors $\{\nabla_{x_i}^j f_i(x_i)\}_{i\in\mathcal{N}}$ of dimension n_i^j needs to be computed and stored, a very substantial gain compared to the n^j-dimensional $\nabla_x^j f_{\mathcal{N}}(x)$ when (as is common) $n_i \ll n$ for all i. It may therefore be argued that exploiting derivative tensors of order larger than 2 — and thus using the high-order Taylor series (1.6) as a local model of $f(x+s)$ in the neighbourhood of x — may be practically feasible if f is partially separable. Of course the same comment applies to

$$f_{\mathcal{H}}(x) \stackrel{\rm def}{=} \sum_{i\in\mathcal{H}} f_i(U_i x) = \sum_{i\in\mathcal{H}} f_i(x_i) \tag{1.8}$$

whenever the required derivatives of $f_i(x_i) = |x_i|^p$ $(i \in \mathcal{H})$ exist.

Interestingly, the use of high-order Taylor models for optimization was recently investigated by Birgin *et al.* [3] in the context of adaptive regularization algorithms for unconstrained problems. Their proposal belongs to this emerging class of methods pioneered by Griewank [25], Nesterov and Polyak [32] and Cartis, Gould and Toint [6, 7] for the unconstrained case and by these last authors in [8] for the convexly constrained case of interest here. Such methods are distinguished by their excellent evaluation complexity, in that they need at most $O(\epsilon^{-(p+1)/p})$ evaluations of the objective function and their derivatives to produce an ϵ-approximate first-order critical point, compared to the $O(\epsilon^{-2})$ evaluations which might be necessary for the steepest descent and Newton's methods (see [31] and [5] for details). However, most adaptive regularization methods rely on a non-separable regularization term in the model of the objective function, making exploitation of structure difficult[(4)].

The purpose of the present paper is twofold. Its first aim is to show that worst-case evaluation complexity for nonconvex minimization subject to convex constraints is not affected by the introduction of non-Lipschitzian singularities in the objective function. The second and concurrent one is to show that this complexity is not affected either by the use of partially separable structure, if present in the problem.

The remaining of the paper is organized as follows. Section 2 establishes a necessary first-order optimality condition for the non-Lipschitzian case. Section 3 then introduces the partially separable adaptive regularization algorithm for this problem while Section 4 is devoted to its worst-case evaluation complexity analysis for the case where Taylor models are used with a kernel-centered feasible set. Section 5 drops the kernel-centered assumption for non-Lipschitz models and Taylor models. The results are discussed in Section 6 and some final conclusions and perspectives are presented in Section 7.

Notations. In what follow, $\|x\|$ denotes the Euclidean norm of the vector x and $\|T\|_p$ the recursively induced Euclidean norm on the p-th order tensor T (see [3, 9] for details). The notation $T[s]^i$ means that the tensor T is applied to i copies of the vector s. For any set $\mathcal{X}$, $|\mathcal{X}|$ denotes its cardinality. For any $\mathcal{I} \subseteq \mathcal{M}$, we also denote $f_{\mathcal{I}}(x) = \sum_{i\in\mathcal{I}} f_i(x)$.

[(4)] The only exception we are aware of is the unpublished note [22] in which a p-th order Taylor model is coupled with a regularization term involving the (totally separable) q-th power of the q norm $(q \geq 1)$.

2 First-order necessary conditions

In this section, we first present exact and approximate first-order necessary conditions for a local minimizer of problem (1.1). Such conditions for optimization problems with non-Lipschitzian singularities have been independently defined in the scaled form [15] or in subspaces [2, 14]. In a recent paper [13], KKT necessary optimality conditions for constrained optimization problems with non-Lipschitzan singularities are studied under the relaxed constant positive linear dependence and basic qualification. The above optimality conditions take the singularity into account by no longer requiring that the gradient (for unconstrained problems, say) nearly vanishes at an approximate solution x_ϵ (which would be impossible if the singularity is active) but by requiring that a scaled version of this requirement holds in that $\|X_\epsilon \nabla_x^1 f(x_\epsilon)\|$ is suitably small, where X_ϵ is a diagonal matrix whose diagonal entries are the components of x_ϵ. Unfortunately, if the i-th component of x_ϵ is small but not quite small enough to consider that the singularity is active for variable i (say it is equal to 2ϵ), the i-th component of $\nabla_x^1 f(x)$ can be as large as a multiple of ϵ^{-1}. As a result, comparing worst-case evaluation complexity bounds with those known for purely Lipschitz continuous problems (such as those proposed in [3] or [9]) may be misleading, since these latter conditions would never accept an approximate first-order critical point with such a large gradient. In order to avoid these pitfalls, we now propose a stronger definition of approximate first-order critical point for non-Lipschitzian problems where such "border-line" situations do not occur. The new definition is also makes use of subspaces but exactly reduces to the standard condition for Lipschitzian problems if the singularity is not active at x_ϵ, even if it is close to it.

Given a vector $x \in \mathbb{R}^n$ and $\epsilon \geq 0$, denote

$$\mathcal{C}(x,\epsilon) \stackrel{\rm def}{=} \{i \in \mathcal{H} \mid |U_i x| \leq \epsilon\}, \quad \mathcal{R}(x,\epsilon) \stackrel{\rm def}{=} \bigcap_{i \in \mathcal{C}(x,\epsilon)} \ker(U_i) = \left[\operatorname*{span}_{i \in \mathcal{C}(x,\epsilon)} \{U_i^T\}\right]^\perp$$

and

$$\mathcal{W}(x,\epsilon) \stackrel{\rm def}{=} \mathcal{N} \cup (\mathcal{H} \setminus \mathcal{C}(x,\epsilon)).$$

For convenience, if $\epsilon = 0$, we denote $\mathcal{C}(x) \stackrel{\rm def}{=} \mathcal{C}(x,0)$, $\mathcal{R}(x) \stackrel{\rm def}{=} \mathcal{R}(x,\epsilon)$ and $\mathcal{W}(x) \stackrel{\rm def}{=} \mathcal{W}(x,0)$.

Observe that the definition of $\mathcal{R}(x,\epsilon)$ above gives that

$$\mathcal{R}(x,\epsilon)^\perp \subseteq \operatorname*{span}_{i \in \mathcal{H}} \{U_i^T\}. \tag{2.1}$$

Also note that any $x \in \mathbb{R}^n$ can be decomposed uniquely as $x = y + z$ where $y \in \mathcal{R}(x)^\perp$ and $z \in \mathcal{R}(x)$. By the definition of $\mathcal{R}(x)$, it is not difficult to verify that

$$U_i z = 0, \ \forall i \in \mathcal{C}(x) \quad \text{and} \quad x \in \mathcal{R}(x).$$

Finaly note that, although $f(x)$ is nonsmooth if $\mathcal{H} \neq \emptyset$, $f_{\mathcal{W}(x,\epsilon)}(x)$ is as differentiable as the $f_i(x)$ for $i \in \mathcal{N}$ and any $\epsilon \geq 0$. This allows us to formulate our first-order necessary condition.

Theorem 2.1 If $x_* \in \mathcal{F}$ is a local minimizer of problem (1.1), then

$$\chi_f(x_*) = 0, \tag{2.2}$$

where, for any $x \in \mathcal{F}$,

$$\chi_f(x_*) = \chi_f(x_*, 0) \stackrel{\rm def}{=} \left| \min_{\substack{x+d\in\mathcal{F} \\ d\in\mathcal{R}(x), \|d\|\leq 1}} \nabla_x^1 f_{\mathcal{W}(x)}(x)^T d \right|. \tag{2.3}$$

Proof. Suppose first that $\mathcal{R}(x_*) = \{0\}$ (which happens if $x_* = 0 \in \mathcal{F}$ and $\text{span}_{i\in\mathcal{H}}\{U_i^T\} = \mathbb{R}^n$). Then (2.2)-(2.3) holds vacuously. Now suppose that $\mathcal{R}(x_*)$ contains at least one nonzero element. By assumption, there exists $\delta_{x_*} > 0$ such that

$$\begin{aligned}
f(x_*) &= \min\{f_{\mathcal{N}}(x) + f_{\mathcal{H}}(x) \mid x \in \mathcal{F},\ \|x - x_*\| \leq \delta_{x_*}\} \\
&= \min\{f_{\mathcal{N}}(y+z) + f_{\mathcal{H}}(y+z) \mid y+z \in \mathcal{F}, y \in \mathcal{R}(x_*)^{\perp}, z \in \mathcal{R}(x_*), \|y+z-x_*\| \leq \delta_{x_*}\} \\
&\leq \min\{f_{\mathcal{N}}(y+z) + \sum_{i\in\mathcal{H}} |U_i(y+z)|^q \mid y+z \in \mathcal{F}, y = 0, z \in \mathcal{R}(x_*), \|z - x_*\| \leq \delta_{x_*}\} \\
&= \min\{f_{\mathcal{N}}(z) + \sum_{i\in\mathcal{H}} |U_i z|^q \mid z \in \mathcal{F}\cap\mathcal{R}(x_*), \|z - x_*\| \leq \delta_{x_*}\} \\
&= \min\{f_{\mathcal{N}}(z) + \sum_{i\in\mathcal{H}\setminus\mathcal{C}(x_*)} |U_i z|^q \mid z \in \mathcal{F}\cap\mathcal{R}(x_*), \|z - x_*\| \leq \delta_{x_*}\}.
\end{aligned}$$

We now introduce a new problem, which is problem (1.1) reduced to $\mathcal{R}(x_*)$, namely,

$$\begin{aligned}
\min \quad & f_{\mathcal{W}(x_*)}(z) = f_{\mathcal{N}}(z) + \sum_{i\in\mathcal{H}\setminus\mathcal{C}(x_*)} |U_i z|^q, \\
\text{s.t.} \quad & z \in \mathcal{F}\cap\mathcal{R}(x_*)
\end{aligned} \tag{2.4}$$

where the gradient $\nabla_z^1 f_{\mathcal{W}(x_*)}(z)$ is locally Lipschitz continuous in some (bounded) neighborhood of x_*. It then follows from $x_* \in \mathcal{R}(x_*)$ that

$$f_{\mathcal{W}(x_*)}(x_*) = f_{\mathcal{N}}(x_*) + \sum_{i\in\mathcal{H}\setminus\mathcal{C}(x_*)} |U_i x_*|^q = f(x_*).$$

Therefore, we have that

$$f_{\mathcal{W}(x_*)}(x_*) \leq \min\{f_{\mathcal{W}(x_*)}(z) \mid z \in \mathcal{F}\cap\mathcal{C}(x_*), \|z - x_*\| \leq \delta_{x_*}\}$$

which implies that x_* is a local minimizer of problem (2.4). Hence, we have

$$\nabla_z^1 f_{\mathcal{W}(x_*)}(x_*)^T(z - x_*) \geq 0, \quad z \in \mathcal{F}\cap\mathcal{R}(x_*). \tag{2.5}$$

In addition,

$$\{d = 0\} \subseteq \{d \ \mid\ x_* + d \in \mathcal{F}, d \in \mathcal{R}(x_*), \|d\| \leq 1\} \subseteq \{d \ \mid\ x_* + d \in \mathcal{F}, d \in \mathcal{R}(x_*)\}$$

which gives the desired result (2.2)-(2.3). $\square$

We call x_* a *first-order stationary point* of (1.1), if x_* satisfies the relation (2.2) in Theorem 2.1. For $\epsilon > 0$, we call x_ϵ an *ϵ-approximate first-order stationary point* of (1.1), if x_ϵ satisfies

$$\chi_f(x_\epsilon, \epsilon) \stackrel{\rm def}{=} \left| \min_{\substack{x+d\in\mathcal{F} \\ d\in\mathcal{R}(x_\epsilon,\epsilon), \|d\|\leq 1}} \nabla_x^1 f_{\mathcal{W}(x_\epsilon,\epsilon)}(x_\epsilon)^T d \right| \leq \epsilon. \tag{2.6}$$

Theorem 2.2 Let x_ϵ be an ϵ-approximate first-order stationary point of (1.1). Then any cluster point of $\{x_\epsilon\}_{\epsilon>0}$ is a first-order stationary point of problem (1.1) as $\epsilon \to 0$.

Proof. Suppose that x_* is any cluster point of $\{x_\epsilon\}_{\epsilon>0}$. Hence there must exist an infinite sequence $\{\epsilon_k\}$ converging to zero and an infinite sequence $\{x_{\epsilon_k}\}_{k\geq 0} \subseteq \{x_\epsilon\}_{\epsilon>0}$ such that $x_* = \lim_{k\to\infty} x_{\epsilon_k}$ and x_{ϵ_k} is an ϵ_k-approximate first-order stationary point of (1.1) for eack $k \geq 0$. If $\mathcal{R}(x_*) = \{0\}$, (2.2) holds vacuously and hence x_* is a first-order stationary point. Suppose therefore that $\mathcal{R}(x_*)$ contains at least one nonzero element, implying that the dimension of $\mathcal{R}(x_*)$ is strictly positive.

First of all, we claim that there must exist $k_* \geq 0$ such that $\mathcal{R}(x_{\epsilon_k}, \epsilon_k)^\perp \subseteq \mathcal{R}(x_*)^\perp$ for any $k \geq k_*$. Indeed, if that is not the case, there exists a subsequence of $\{x_{\epsilon_k}\}$, say $\{x_{\epsilon_{k_j}}\}$, such that $\lim_{j\to\infty} \epsilon_{k_j} = 0$ and $\mathcal{R}(x_{\epsilon_{k_j}}, \epsilon_{k_j})^\perp \not\subseteq \mathcal{R}(x_*)^\perp$ for all j. Using now (2.1) and the fact that $\mathcal{H}$ is a finite set, we obtain that there must exist an $i_0 \in \mathcal{H}$ such that $i_0 \in \mathcal{C}(x_{\epsilon_{k_{j_t}}}, \epsilon_{k_{j_t}})$ but $i_0 \notin \mathcal{C}(x_*)$ where $\{k_{j_t}\} \subseteq \{k_j\}$ with $t = 1, 2, \cdots$. For convenience, we continue to use $\{k_j\}$ to denote its subsequence $\{k_{j_t}\}$. Hence, we have that

$$|U_{i_0} x_{\epsilon_{k_j}}| \leq \epsilon_{k_j}.$$

Let j go to infinity. It then follows from the above inequality that $|U_{i_0} x_*| = 0$, which contradicts the fact that $i_0 \notin \mathcal{C}(x_*)$. Thus, we conclude that, for some $k_* \geq 0$ and all $k \geq k_*$, $\mathcal{R}(x_{\epsilon_k}, \epsilon_k)^\perp \subseteq \mathcal{R}(x_*)^\perp$. Therefore we have that $\mathcal{R}(x_*) \subseteq \mathcal{R}(x_{\epsilon_k}, \epsilon_k)$ for $k \geq k_*$.

For any fixed ϵ_k approximate first-order stationary point x_{ϵ_k}, consider the following two minimization problems.

$$\begin{array}{ll} \min & \nabla_x^1 f_{\mathcal{W}(x_{\epsilon_k},\epsilon_k)}(x_{\epsilon_k})^T d, \\ \text{s.t.} & x_{\epsilon_k} + d \in \mathcal{F}, d \in \mathcal{R}(x_{\epsilon_k}, \epsilon_k), \|d\| \leq 1, \end{array} \tag{2.7}$$

and

$$\begin{array}{ll} \min & \nabla_x^1 f_{\mathcal{W}(x_{\epsilon_k},\epsilon_k)}(x_{\epsilon_k})^T d, \\ \text{s.t.} & x_{\epsilon_k} + d \in \mathcal{F}, d \in \mathcal{R}(x_*), \|d\| \leq 1. \end{array} \tag{2.8}$$

Since $d = 0$ is a feasible point of both problems (2.7) and (2.8), the minimum values of (2.7) and (2.8) are both nonpositive. Moreover, it follows from $\mathcal{R}(x_*) \subseteq \mathcal{R}(x_{\epsilon_k}, \epsilon_k)$ that the minimum value of (2.8) is not smaller than that of (2.7).

Hence, from (2.6), we have that for any x_{ϵ_k},

$$\left| \min_{\substack{x_{\epsilon_k}+d\in\mathcal{F} \\ d\in\mathcal{R}(x_*),\|d\|\le 1}} \nabla_x^1 f_{\mathcal{W}(x_{\epsilon_k},\epsilon_k)}(x_{\epsilon_k})^T d \right| \le \left| \min_{\substack{x_{\epsilon_k}+d\in\mathcal{F} \\ d\in\mathcal{R}(x_{\epsilon_k},\epsilon_k),\|d\|\le 1}} \nabla_x^1 f_{\mathcal{W}(x_{\epsilon_k},\epsilon_k)}(x_{\epsilon_k})^T d \right| \le \epsilon_k. \quad (2.9)$$

Suppose that d_{ϵ_k} is a minimizer of problem (2.8), then (2.9) implies that

$$-\epsilon_k \le \nabla_x^1 f_{\mathcal{W}(x_{\epsilon_k},\epsilon_k)}(x_{\epsilon_k})^T d_{\epsilon_k} \le 0, \quad (2.10)$$

where d_{ϵ_k} should satisfy that $x_{\epsilon_k} + d_{\epsilon_k} \in \mathcal{F}$, $d_{\epsilon_k} \in \mathcal{R}(x_*)$ and $\|d_{\epsilon_k}\| \le 1$. Note that, since $d_{\epsilon_k} \in \mathcal{R}(x_*)$,

$$\begin{aligned}
\nabla_x^1 f_{\mathcal{W}(x_{\epsilon_k},\epsilon_k)}(x_{\epsilon_k})^T d_{\epsilon_k} &= \left(\nabla_x f_{\mathcal{N}}(x_{\epsilon_k}) + \sum_{i\in\mathcal{H}\setminus\mathcal{C}(x_{\epsilon_k})} q|U_i x_{\epsilon_k}|^{q-1}\mathrm{sign}(U_i x_{\epsilon_k}) U_i^T \right)^T d_{\epsilon_k} \\
&= \nabla_x f_{\mathcal{N}}(x_{\epsilon_k})^T d_{\epsilon_k} + \sum_{i\in\mathcal{H}\setminus\mathcal{C}(x_{\epsilon_k})} q|U_i x_{\epsilon_k}|^{q-1}\mathrm{sign}(U_i x_{\epsilon_k}) U_i d_{\epsilon_k} \\
&= \nabla_x f_{\mathcal{N}}(x_{\epsilon_k})^T d_{\epsilon_k} + \sum_{i\in\mathcal{H}} q|U_i x_{\epsilon_k}|^{q-1}\mathrm{sign}(U_i x_{\epsilon_k}) U_i d_{\epsilon_k}.
\end{aligned} \quad (2.11)$$

From the compactness of $\{d \mid \|d\| \le 1\}$, we know that there must exist a subsequence of $\{d_{\epsilon_k}\}$ such that $d_{\epsilon_{k_j}} \to d_* \in \mathcal{R}(x_*)$ with $\|d_*\| \le 1$ as j goes to infinity. Since for $i \in \mathcal{H}\setminus\mathcal{C}(x_*)$, we have $\lim_{k\to\infty} |U_i x_{\epsilon_k}|^{q-1} = |U_i x_*|^{q-1}$. Let k go to infinity in (2.10) and (2.11), and we obtain that

$$0 = \nabla_x^1 f_{\mathcal{W}(x_{\epsilon_k},\epsilon_k)}(x_*)^T d_* = \nabla f_{\mathcal{N}}(x_*)^T d_* + \sum_{i\in\mathcal{H}\setminus\mathcal{C}(x_*)} q|U_i x_*|^{q-1}\mathrm{sign}(U_i x_*) U_i d_*,$$

which implies that

$$\min_{\substack{x_*+d\in\mathcal{F} \\ d\in\mathcal{R}(x_*),\|d\|\le 1}} \nabla_x^1 f_{\mathcal{W}(x_{\epsilon_k},\epsilon_k)}(x_*)^T d = \nabla_x^1 f_{\mathcal{W}(x_*)}(x_*)^T d_* = 0$$

and completes the proof. □

3 A partially separable regularization algorithm

We now examine the desired properties of the element functions f_i more closely. Assume first that, for $i \in \mathcal{N}$, each element function f_i is p times continuously differentiable and its p-th order derivative tensor $\nabla_x^p f_i$ is globally Lipschitz continuous with constant $L_i \ge 0$ in the sense that, for all $x_i, y_i \in \mathrm{range}(U_i)$,

$$\|\nabla_{x_i}^p f_i(x_i) - \nabla_{x_i}^p f_i(y_i)\|_p \le L_i \|x_i - y_i\|. \quad (3.1)$$

It can be shown (see (4.6) below) that this assumption implies that, for $i \in \mathcal{N}$,

$$f_i(x_i+s_i) = T_{f_i,p}(x_i,s_i) + \frac{1}{(p+1)!}\tau_i L_i \|s_i\|^{p+1} \quad \text{with} \quad |\tau_i| \le 1, \tag{3.2}$$

where $s_i = U_i s$.

Because the quantity $\tau_i L_i$ in (3.2) is usually unknown in practice, it is impossible to use (3.2) directly to model the objective function in a neighbourhood of x. However, we may replace this term with an adaptive parameter σ_i, which yields the following $(p+1)$-th order model for the i-th element ($i \in \mathcal{N}$):

$$m_i(x_i,s_i) = T_{f_i,p}(x_i,s_i) + \frac{1}{(p+1)!}\ \sigma_i \|s_i\|^{p+1}. \tag{3.3}$$

There is more than one possible choice for defining the element models for $i \in \mathcal{H}$. The first[(5)] is to pursue the line of polynomial Taylor-based models, for which we need the following technical result.

Lemma 3.1 We have that, for $i \in \mathcal{H}$ and all $x, s \in \mathbb{R}^n$ with $U_i x \neq 0 \neq U_i(x+s)$,

$$|x_i+s_i|^q = |x_i|^q + q\sum_{j=1}^{\infty}\frac{1}{j!}\left(\prod_{\ell=1}^{j-1}(q-\ell)\right)|x_i|^{q-j}\mu(x_i,s_i)^j, \tag{3.4}$$

where

$$\mu(x_i,s_i) \stackrel{\rm def}{=} \begin{cases} s_i & \text{if } x_i>0 \text{ and } x_i+s_i>0,\\ -s_i & \text{if } x_i<0 \text{ and } x_i+s_i<0,\\ -(2x_i+s_i) & \text{if } x_i>0 \text{ and } x_i+s_i<0,\\ 2x_i+s_i & \text{if } x_i<0 \text{ and } x_i+s_i>0.\end{cases} \tag{3.5}$$

Proof. If $y \in \mathbb{R}_+$, it can be verified that the Taylor expansion $|y+z|^q$ at $y \neq 0$ and $y+z \in \mathbb{R}_+$ is given by

$$[y+z]^q = y^q + q\sum_{j=1}^{\infty}\frac{1}{j!}\left[\prod_{\ell=1}^{j-1}(q-\ell)\right]y^{q-j}z^j. \tag{3.6}$$

Let us now consider $i \in \mathcal{H}$. Relation (3.6) yields that, if $x_i > 0$ and $x_i+s_i>0$,

$$|x_i+s_i|^q = |x_i|^q + q\sum_{j=1}^{\infty}\frac{1}{j!}\left[\prod_{\ell=1}^{j-1}(q-\ell)\right]|x_i|^{q-j}s_i^j. \tag{3.7}$$

By symmetry, if we have that if $x_i<0$ and $x_i+s_i<0$, then

$$|x_i+s_i|^q = |x_i|^q + q\sum_{j=1}^{\infty}\frac{(-1)^j}{j!}\left[\prod_{\ell=1}^{j-1}(q-\ell)\right]|x_i|^{q-j}s_i^j. \tag{3.8}$$

(5) Another choice is discussed in Section 5.

Moreover, if $x_i > 0$ and $x_i + s_i < 0$, then

$$|x_i + s_i|^q = |-x_i|^q + q\sum_{j=1}^{\infty}\frac{(-1)^j}{j!}\left[\prod_{\ell=1}^{j-1}(q-\ell)\right]|-x_i|^{q-j}(2x_i+s_i)^j. \tag{3.9}$$

Symmetrically, if $x_i < 0$ and $x_i + s_i > 0$, then again,

$$|x_i + s_i|^q = |-x_i|^q + q\sum_{j=1}^{\infty}\frac{1}{j!}\left[\prod_{\ell=1}^{j-1}(q-\ell)\right]|-x_i|^{q-j}(2x_i+s_i)^j \tag{3.10}$$

(3.4)-(3.5) then trivially results from (3.7)-(3.10) and the identity $|-x_i| = |x_i|$. $\square$

We now slightly abuse notation by defining

$$T_{|\cdot|^q,p}(x_i,s_i) \stackrel{\rm def}{=} \left\{\begin{array}{ll} T_{x^q,p}(x_i,s_i) & \text{if} \quad x_i > 0 \quad \text{and} \quad x_i+s_i > 0,\\ T_{(-x)^q,p}(x_i,-s_i) & \text{if} \quad x_i < 0 \quad \text{and} \quad x_i+s_i < 0,\\ T_{(-x)^q,p}(-x_i,2x_i+s_i) & \text{if} \quad x_i > 0 \quad \text{and} \quad x_i+s_i < 0,\\ T_{x^q,p}(-x_i,2x_i+s_i) & \text{if} \quad x_i < 0 \quad \text{and} \quad x_i+s_i > 0. \end{array}\right. \tag{3.11}$$

We are now in position to define the regularized "two-sided" model for the element function f_i ($i \in \mathcal{H}$) as

$$m_i(x_i,s_i) \stackrel{\rm def}{=} T_{|\cdot|^q,p}(x_i,s_i). \tag{3.12}$$

Figure 3.1 illustrates the two-sided model (3.11)-(3.12) for $x_i = -\frac{1}{2}$, $p = 3$, $q = \frac{1}{2}$.

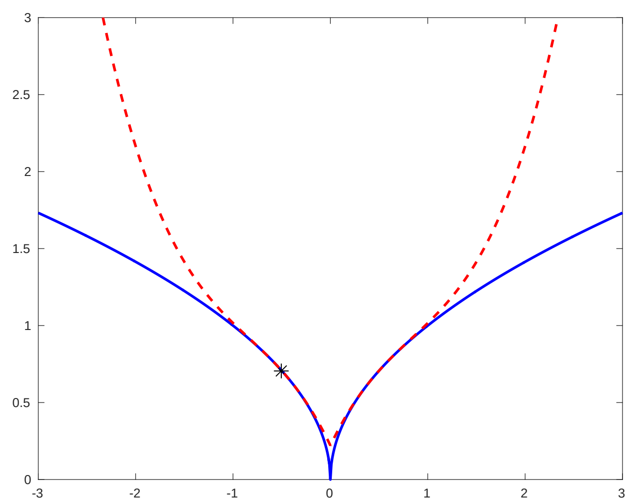

Figure 3.1: The square root function (continuous) and its two-sided model with $p = 3$ evaluated at $x_i = -\frac{1}{2}$ (dashed)

We may now build the complete model for f at x as

$$m(x,s) = \sum_{i\in\mathcal{M}} m_i(x_i,s_i). \tag{3.13}$$

The algorithm considered in this paper exploits the model (3.13) as follows. At each iteration k, the model (3.13) taken at the iterate $x = x_k$ is (approximately) minimized in order to define a step s_k. If the decrease in the objective function value along s_k is comparable to that predicted by the Taylor model, the trial point $x_k + s_k$ is accepted as the new iterate and the regularization parameters $\sigma_{i,k}$ (i.e. σ_i at iteration k) possibly updated. The process is terminated when an approximate local minimizer is found, that is when, for some $k \geq 0$,

$$\chi_f(x_k, \epsilon) \leq \epsilon. \tag{3.14}$$

In order to simplify notation in what follows, we make the following definitions:

$$\mathcal{C}_k \stackrel{\rm def}{=} \mathcal{C}(x_k, \epsilon), \quad \mathcal{R}_k \stackrel{\rm def}{=} \mathcal{R}(x_k, \epsilon), \quad \mathcal{W}_k \stackrel{\rm def}{=} \mathcal{W}(x_k, \epsilon),$$

and

$$\mathcal{C}_k^+ \stackrel{\rm def}{=} \mathcal{C}(x_k + s_k, \epsilon), \quad \mathcal{R}_k^+ \stackrel{\rm def}{=} \mathcal{R}(x_k + s_k, \epsilon), \quad \mathcal{W}_k^+ \stackrel{\rm def}{=} \mathcal{W}(x_k + s_k, \epsilon).$$

Having defined the criticality measure (2.3), it is natural to use this measure also for terminating the approximate model minimization: to find s_k, we therefore minimize $m(x_k, s)$ over s until, for some constant $\theta \geq 0$ and some exponent $r > 1$,

$$\chi_m(x_k, s_k, \epsilon) = \chi_{m_{\mathcal{W}_k^+}}(x_k, s_k, \epsilon) \leq \min\left[\tfrac{1}{4}q^2 \min_{i \in \mathcal{H} \cap \mathcal{W}_k^+} |U_i(x_k + s_k)|^r, \, \theta \|s_k\|^p\right] \tag{3.15}$$

where

$$\chi_{m_{\mathcal{W}_k^+}}(x_k, s_k, \epsilon) \stackrel{\rm def}{=} \left| \min_{\substack{x_k + s_k + d \in \mathcal{F} \\ d \in \mathcal{R}_k^+, \|d\| \leq 1}} \nabla_s^1 m_{\mathcal{W}_k^+}(x_k, s_k)^T d \right|. \tag{3.16}$$

We also require that, once $|U_i(x_k + s)| < \epsilon$ occurs for some $i \in \mathcal{H}$ in the course of the model minimization, it is fixed at this value, meaning that the remaining minimization is carried out in $\mathcal{R}(x_k + s, \epsilon)$. Thus the dimension of $\mathcal{R}(x_k + s, \epsilon)$ (and therefore of $\mathcal{R}(x_k, \epsilon)$) is monotonically non-increasing during the step computation and across iterations. Note that computing a step s_k satisfying (3.15) is always possible since the subspace $\mathcal{R}(x_k + s, \epsilon)$ can only become smaller during the model minimization and since we have seen in Section 2 that $\chi_m(x_k, s_k) = 0$ at any local minimizer of $m_{\mathcal{W}(x_k+s,\epsilon)}(x_k, s)$.

3.1 The algorithm

We now introduce some notation useful for describing our algorithm. Define

$$x_{i,k} \stackrel{\rm def}{=} U_i x_k, \quad s_{i,k} \stackrel{\rm def}{=} U_i s_k.$$

Also let

$$\delta f_{i,k} \stackrel{\rm def}{=} f_i(x_{i,k}) - f_i(x_{i,k} + s_{i,k})$$

$$\delta f_k \stackrel{\rm def}{=} f_{\mathcal{W}_k^+}(x_k) - f_{\mathcal{W}_k^+}(x_k + s_k) = \sum_{i \in \mathcal{W}_k^+} \delta f_{i,k},$$

$$\delta m_{i,k} \stackrel{\rm def}{=} m_i(x_{i,k}, 0) - m_i(x_{i,k}, s_{i,k}),$$

$$\delta m_k \stackrel{\rm def}{=} m_{\mathcal{W}_k^+}(x_k,0) - m_{\mathcal{W}_k^+}(x_k,s_k) = \sum_{i\in\mathcal{W}_k^+} \delta m_{i,k},$$

and

$$\begin{aligned}\delta T_k &\stackrel{\rm def}{=} T_{f_{\mathcal{W}_k^+},p}(x_k,0) - T_{f_{\mathcal{W}_k^+},p}(x_k,s_k)\\ &= [T_{f_{\mathcal{N}},p}(x_k,0) - T_{f_{\mathcal{N}},p}(x_k,s_k)] + [T_{|\cdot|_{\mathcal{H}\setminus\mathcal{C}_k^+},p}(x_k,0) - T_{|\cdot|_{\mathcal{H}\setminus\mathcal{C}_k^+},p}(x_k,s_k)]\\ &= \delta m_k + \frac{1}{(p+1)!}\sum_{i\in\mathcal{N}}\sigma_{i,k}\|s_{i,k}\|^{p+1}.\end{aligned} \tag{3.17}$$

The partially separable adaptive regularization algorithm is now formally stated as Algorithm 3.1 on the following page.

Note that an $x_0 \in \mathcal{F}$ can always be computed by projecting an infeasible starting point onto $\mathcal{F}$. The idea of the second and third parts of (3.21) and (3.22) is to identify cases where the model m_i overestimates the element function f_i to an excessive extent, leaving some space for reducing the regularization and hence allowing longer steps. The requirement that $\rho_k \geq \eta$ in both (3.21) and (3.22) is intended to prevent a situation where a particular regularization parameter is increased and another decreased at a given unsuccessful iteration, followed by the opposite situation at the next iteration, potentially leading to cycling. Other more elaborate mechanisms can be designed to achieve the same goal, such as attempting to reduce a given regularization parameter at most a fixed number of times before the occurence of a successful iteration, but we do not investigate those alternatives in detail here. The idea of the second and third parts of (3.21) and (3.22) is simply to identify cases where the model m_i overestimates the element function f_i to an excessive extent, leaving some space for reducing the regularization and hence allowing longer steps.

We note at this stage that the condition $s_k \in \mathcal{R}_k$ implies that

$$\mathcal{C}_k \subseteq \mathcal{C}_k^+ \quad\text{and}\quad \mathcal{W}_k^+ \subseteq \mathcal{W}_k.$$

Note that the above algorithm considerably simplifies in the Lipschitzian case where $\mathcal{H} = \emptyset$, since

$$f_{\mathcal{W}_k}(x) = f_{\mathcal{M}}(x) = f(x)$$

for all $k \geq 0$ and all $x \in \mathcal{F} = \mathcal{F}_{\mathcal{Q}}$.

4 Evaluation complexity for 'kernel-centered' fesible sets

We start our worst-case analysis by formalizing our assumptions for problem (1.1).

AS.1 The feasible set $\mathcal{F}$ is closed, convex and non-empty.

AS.2 Each element function f_i ($i \in \mathcal{N}$) is p times continuously differentiable in an open set containing $\mathcal{F}$, where p is odd whenever $\mathcal{H} \neq \emptyset$.

Algorithm 3.1: Partially Separable Adaptive Regularization

Step 0: Initialization: $x_0 \in \mathcal{F}$ and $\{\sigma_{0,i}\}_{i\in\mathcal{N}} > 0$ are given as well as the accuracy $\epsilon \in (0,1]$ and constants $0 < \gamma_0 < 1 < \gamma_1 \leq \gamma_2$, $\eta \in (0,1)$, $\theta \geq 0$, $\sigma_{\min} \in (0, \min_{i\in\mathcal{N}} \sigma_{0,i}]$ and $\kappa_{\rm big} > 1$. Set $k = 0$.

Step 1: Termination: Evaluate $f(x_k)$ and $\{\nabla_x^1 f_{\mathcal{W}_k}(x_k)\}$. If $\chi_f(x_k,\epsilon) \leq \epsilon$, return $x_\epsilon = x_k$ and terminate. Otherwise evaluate $\{\nabla_x^i f_{\mathcal{W}_k}(x_k)\}_{i=2}^p$.

Step 2: Step computation: Compute a step $s_k \in \mathcal{R}_k$ such that $x_k + s_k \in \mathcal{F}$, $m(x_k, s_k) < m(x_k, 0)$ and (3.15) holds.

Step 3: Step acceptance: Compute

$$\rho_k = \frac{\delta f_k}{\delta T_k} \tag{3.18}$$

and set $x_{k+1} = x_k$ if $\rho_k < \eta$, or $x_{k+1} = x_k + s_k$ if $\rho_k \geq \eta$.

Step 4: Update the "nice" regularization parameters: For $i \in \mathcal{N}$, if

$$f_i(x_{i,k} + s_{i,k}) > m_i(x_{i,k}, s_{i,k}) \tag{3.19}$$

set

$$\sigma_{i,k+1} \in [\gamma_1\sigma_{i,k}, \gamma_2\sigma_{i,k}]. \tag{3.20}$$

Otherwise, if either

$$\rho_k \geq \eta \quad \text{and} \quad \delta f_{i,k} \leq 0 \quad \text{and} \quad \delta f_{i,k} < \delta m_{i,k} - \kappa_{\rm big}|\delta f_k| \tag{3.21}$$

or

$$\rho_k \geq \eta \quad \text{and} \quad \delta f_{i,k} > 0 \quad \text{and} \quad \delta f_{i,k} > \delta m_{i,k} + \kappa_{\rm big}|\delta f_k| \tag{3.22}$$

then set

$$\sigma_{i,k+1} \in [\max[\sigma_{\min}, \gamma_0\sigma_{i,k}], \sigma_{i,k}], \tag{3.23}$$

else set

$$\sigma_{i,k+1} = \sigma_{i,k}. \tag{3.24}$$

Increment k by one and go to Step 1.

AS.3 The p-th derivative of each f_i ($i \in \mathcal{N}$) is Lipschitz continuous on $\mathcal{F}$ with associated Lipschitz constant L_i (in the sense of (3.1)).

AS.4 There exists a constant $f_{\rm low}$ such that $f_{\mathcal{N}}(x) \geq f_{\rm low}$ for all $x \in \mathcal{F}$.

AS.5 There exists a constant $\kappa_{\mathcal{N}}^0 \geq 0$ such that

$$\|\nabla_x^j f_{\mathcal{N}}(0)\| \leq \kappa_{\mathcal{N}}^0$$

for all $j \in \{1, \ldots, p\}$.

Note that AS.4 is necessary for problem (1.1) to be well-defined. Also observe that AS.5 guarantees the existence of a constant $\kappa_{\mathcal{N}} \geq \kappa_{\mathcal{N}}^0$ such that

$$\|\nabla_x^1 f_{\mathcal{N}}(x))\| \leq \kappa_{\mathcal{N}} \ \text{ for all } \ x \in \{x \in \mathcal{F} \mid \|x\| \leq 1\}. \tag{4.1}$$

Obviously, AS.2 alone implies (4.1) (without the need of assuming AS.5) if $\mathcal{F}$ is bounded.

We first observe that our assumptions on the partially separable nature of the objective function imply the following useful bounds.

Lemma 4.1 There exist constants $0 < \varsigma_{\min} \leq \varsigma_{\max}$ such that, for all $s \in \mathbb{R}^m$ and all $v \geq 1$ and for any subset $\mathcal{X} \subseteq \mathcal{M}$,

$$\varsigma_{\min}^v \|s_{\mathcal{X}}\|^v \leq \sum_{i \in \mathcal{X}} \|s_i\|^v \leq |\mathcal{X}| \, \varsigma_{\max}^v \, \|s_{\mathcal{X}}\|^v, \tag{4.2}$$

where $s_{\mathcal{X}} = P_{\mathrm{span}_{i \in \mathcal{X}}\{U_i^T\}}(s)$.

Proof. Assume that, for every $\varsigma > 0$ there exists a vector s_ς in $\mathrm{span}_{i \in \mathcal{X}}\{U_i^T\}$ of norm 1 such that $\max_{i \in \mathcal{X}} \|U_i s_\varsigma\| < \varsigma \|s_\varsigma\| = \varsigma$. Then taking a sequence of $\{\varsigma_i\}$ converging to zero and using the compactness of the unit sphere, we obtain that the sequence $\{s_{\varsigma_i}\}$ has at least one limit point s_0 with $\|s_0\| = 1$ such that $\max_{i \in \mathcal{X}} \|U_i s_0\| = 0$, which is impossible since we assumed that the intersection of the nullspaces of the U_i is reduced to the origin. Thus our assumption is false and there is constant $\varsigma_{\min} > 0$ such that, for every $s \in \mathrm{span}_{i \in \mathcal{X}}\{U_i^T\}$,

$$\max_{i \in \mathcal{X}} \|s_i\| = \max_{i \in \mathcal{X}} \|U_i s\| \geq \varsigma_{\min} \|s\|.$$

The first inequality of (4.2) then follows from the fact that

$$\sum_{i \in \mathcal{X}} \|s_i\|^v \geq \max_{i \in \mathcal{X}} \|s_i\|^v \geq \varsigma_{\min}^v \|s\|^v.$$

We have also that

$$\sum_{i \in \mathcal{X}} \|s_i\|^v \leq |\mathcal{X}| \max_{i \in \mathcal{X}} \|U_i s\|^v \leq |\mathcal{X}| \max_{i \in \mathcal{X}} \left(\|U_i\| \|s\|\right)^v,$$

which yields the second inequality of (4.2) with $\varsigma_{\max} = \max_{i \in \mathcal{X}} \|U_i\|$. □

Taken for $v = 1$ and $\mathcal{X} = \mathcal{N}$, this lemma states that $\sum_{i \in \mathcal{N}} \|\cdot\|$ is a norm on $\mathbb{R}^n$ whose equivalence constants with respect to the Euclidean one are $\varsigma_{\min}$ and $|\mathcal{N}|\,\varsigma_{\max}$. In most applications, these constants are very moderate numbers.
We now turn to the consequence of the Lipschitz continuity of $\nabla_x^p f_i$ and define, for a given $k \geq 0$ and a given constant $\phi > 0$ independent of ϵ,

$$\mathcal{O}_{k,\phi} \stackrel{\rm def}{=} \{ i \in \mathcal{W}_k^+ \cap \mathcal{H} \mid \min[\,|x_{i,k}|,\, |x_{i,k} + s_{i,k}|\,] \geq \phi \}. \tag{4.3}$$

Note that

$$\mathcal{O}_{k,\phi} = \mathcal{H} \setminus \Big[\mathcal{C}(x_k, \phi) \cup \mathcal{C}(x_k + s_k, \phi) \Big].$$

Lemma 4.2 Suppose that AS.2 and AS.3 hold. Then, for $k \geq 0$ and $L_{\max} \stackrel{\rm def}{=} \max_{i \in \mathcal{N}} L_i$,

$$f_i(x_{i,k} + s_{i,k}) = m_i(x_{i,k}, s_{i,k}) + \frac{1}{(p+1)!} \Big[\tau_{i,k}(p+1) L_{\max} - \sigma_{i,k} \Big] \|s_{i,k}\|^{p+1} \quad \text{with} \quad |\tau_{i,k}| \leq 1, \tag{4.4}$$

for all $i \in \mathcal{N}$. If, in addition, $\phi > 0$ is given and independent of ϵ, then there exists a constant $L(\phi)$ independent of ϵ such that

$$\|\nabla_x^1 f_{\mathcal{N} \cup \mathcal{O}_{k,\phi}}(x_k + s_k) - \nabla_s^1 m_{\mathcal{N} \cup \mathcal{O}_{k,\phi}}(x_k, s_k)\| \leq L(\phi) \|s_k\|^p. \tag{4.5}$$

Proof. First note that, if f_i has a Lipschitz continuous p-th derivative as a function of $U_i x$, then (1.7) shows that it also has a Lipschitz continuous p-th derivative as a function of x. It is therefore enough to consider the element functions as functions of $x_i = U_i x$.

AS.3 and (3.1) imply that

$$f_i(x_{i,k} + s_{i,k}) = T_{f_i,p}(x_{i,k}, s_{i,k}) + \frac{\tau_{i,k}}{p!} L_{\max} \|s_{i,k}\|^{p+1} \quad \text{with} \quad |\tau_{i,k}| \leq 1, \tag{4.6}$$

for each $i \in \mathcal{N}$ (see [3] or [9, Section 2.2]), and (4.4) then follows from (3.3).

Consider now $i \in \mathcal{O}_{k,\phi}$ and assume first that $x_{i,k} > \phi$ and $x_{i,k} + s_{i,k} > \phi$. Then $f_i(x_i) = x_i^q$ is infinitely differentiable on the interval $[x_{i,k}, x_{i,k} + s_{i,k}] \subset [\phi, \infty)$ and the norm of its $(p+1)$-st derivative tensor is bounded above on this interval by

$$L_{\mathcal{H}}(\phi) \stackrel{\rm def}{=} \left| \prod_{\ell=0}^{p+1} (q - \ell) \right| \phi^{q-p-1}. \tag{4.7}$$

We then apply the same reasoning as above using the Taylor series expansion of x_i^q at $x_{i,k}$ and, because of the first line of (3.11), deduce that

$$f_i(x_{i,k} + s_{i,k}) = m_i(x_{i,k}, s_{i,k}) + \frac{1}{(p+1)!} \tau_{i,k}(p+1) L_{\mathcal{H}}(\phi) |s_{i,k}|^{p+1} \quad \text{with} \quad |\tau_{i,k}| \leq 1, \tag{4.8}$$

and

$$\|\nabla_x^1 f_i(x_{i,k} + s_{i,k}) - \nabla_s^1 m_i(x_{i,k}, s_{i,k})\| \leq L_{\mathcal{H}}(\phi) |s_{i,k}|^p, \tag{4.9}$$

hold in this case (see [3]). The argument is obviously similar if $x_{i,k} < -\phi$ and $x_{i,k}+s_{i,k} < -\phi$, using symmetry and the second line of (3.11). Let us now consider the case where $x_{i,k} > \phi$ and $x_{i,k}+s_{i,k} < -\phi$. The expansion (3.4) then shows that we may reason as for $x_{i,k} < -\phi$ and $x_{i,k}+s_{i,k} < -\phi$ using a Taylor expansion at $-x_i$ (which we know by symmetry) and the third line of (3.11). The case where $x_{i,k} < -\phi$ and $x_{i,k}+s_{i,k} > \phi$ is similar, using the fourth line of (3.11). As a consequence, (4.8) and (4.9) hold for every $i \in \mathcal{O}_{k,\phi}$ with Lipschitz constant $L_{\mathcal{H}}(\phi)$. Moreover, using (4.2) and the definitions (4.7),

$$\sum_{i\in\mathcal{N}\cup\mathcal{O}_{k,\phi}} L_i\|s_i\|^{p+1} \leq \max\left[L_{\max}, L_{\mathcal{H}}(\phi)\right] \sum_{i\in\mathcal{N}\cup\mathcal{O}_{k,\phi}} \|s_i\|^{p+1}$$

from which (4.5) may in turn be derived from (4.9) and (4.2) with

$$L(\phi) \stackrel{\rm def}{=} |\mathcal{M}|\,\varsigma_{\max}^{p+1}\max\left[L_{\max}, L_{\mathcal{H}}(\phi)\right]. \tag{4.10}$$

□

Note that there is no dependence on ϕ in L if $\mathcal{H} = \emptyset$.

We now return to our statement that

$$\ker(U_i)\cap\mathcal{F} \neq \emptyset \tag{4.11}$$

may be assumed without loss of generality for all $i \in \mathcal{H}$. Indeed, assume that (4.11) fails for $j \in \mathcal{H}$. Then $j \in \mathcal{O}_{k,\xi_j}$ for all $k \geq 0$, where $\xi_j > 0$ is the distance between $\ker(U_j)$ and $\mathcal{F}$, and we may transfer j from $\mathcal{H}$ to $\mathcal{N}$ (possibly modifying $L_{\max}$).

The definition of the model in (3.13) also implies a simple lower bound on model decrease.

Lemma 4.3 For all $k \geq 0$,

$$\delta T_k \geq \frac{1}{(p+1)!}\,\sigma_{\min}\sum_{i\in\mathcal{N}}\|s_{i,k}\|^{p+1}, \tag{4.12}$$

$s_k \neq 0$ and (3.18) is well-defined.

Proof. The bound directly follows from (3.17), the observation that the algorithm enforces $\delta m_k > 0$ and (3.23). Moreover, $\chi_m(x_k,0,\epsilon) = \chi_f(x_k,\epsilon) > \epsilon$. As a consequence, (3.15) cannot hold for $s_k = 0$ since termination would have then occured in Step 1 of Algorithm 3.1. Hence at least one $\|s_{i,k}\|$ is strictly positive because of (4.2) and (4.12) therefore implies that (3.18) is well-defined. □

We now verify that the two-sided model (3.12) is an overestimate of the function $|x|^q$ for all relevant x_i and s_i.

Lemma 4.4 Suppose that AS.2 holds. Then, for $i \in \mathcal{H}$ and all $x_i, s_i \in \mathbb{R}^n$ with $x_i \neq 0 \neq x_i + s_i$, we have that

$$|x_i+s_i|^q \leq m_i(x_i,s_i). \tag{4.13}$$

Proof. Since $i \in \mathcal{H}$ by assumption, this implies that $\mathcal{H} \neq \emptyset$, and thus, by AS.2, that p is odd. From the mean-value theorem, we obtain that

$$\begin{aligned} |x_i + s_i|^q &= |x_i|^q + q \sum_{j=1}^{p} \frac{1}{j!} \left[\prod_{\ell=1}^{j-1} (q-\ell) \right] |x_i|^{q-j} \mu(x_i, s_i)^j \\ &\quad + \frac{1}{(p+1)!} \left[\prod_{\ell=1}^{p} (q-\ell) \right] |U_i z|^{q-p-1} \mu(x_i, s_i)^{p+1} \end{aligned} \tag{4.14}$$

for some z such that, using symmetry, $z \in [x, x+s]$ if $(U_i x)(U_i(x+s)) > 0$ or $z \in [-x, x+s]$ otherwise. As a consequence, we have that

$$|U_i z| \geq \min[\, |x_i|,\, |x_i + s_i|\,] > 0.$$

Remember now that p is odd. Then, using that $q \in (0,1)$, we have that

$$\mu(x_i, s_i)^{p+1} \geq 0 \quad \text{and} \quad \prod_{\ell=1}^{p} (q-\ell) < 0.$$

The inequality

$$|x_i + s_i|^q \leq |x_i|^q + q \sum_{j=1}^{p} \frac{1}{j!} \left[\prod_{\ell=1}^{j-1} (q-\ell) \right] |x_i|^{q-j} \mu(x_i, s_i)^j \tag{4.15}$$

therefore immediately follows from (4.14), proving (4.13). □

We next investigate the consequences of the model's definition (3.12) when the singularity at the origin is approached and show that the two-sided model has to remain large along the steps when $x_{i,k}$ is not too far from the singularity.

Lemma 4.5 Suppose that $p \geq 1$ is odd, $q \in (0,1)$, $i \in \mathcal{H}$, $|x_i| \in (\epsilon, 1]$, and $|x_i + s_i| \geq \epsilon$. Then

$$|\nabla^1_{s_i} m_i(x_i, s_i)| > \tfrac{1}{2} q\, |\nabla^1_{s_i} m_i(x_i, 0)|. \tag{4.16}$$

Proof. Following the argument in the proof of Lemma 4.2, it is sufficient to consider that $x_i > 0$ and $x_i + s_i > 0$. From (3.11) (where $\mu(x_i, s_i) = s_i$), we have that

$$\nabla^1_{s_i} T_{x^q,p}(x_i, s_i) = q \sum_{j=1}^{p} \frac{1}{(j-1)!} \left[\prod_{\ell=1}^{j-1} (q-\ell) \right] x_i^{q-j} s_i^{j-1}. \tag{4.17}$$

Define $s_i = \beta x_i$. This gives that (4.17) now reads

$$\nabla^1_{s_i} T_{x^q,p}(x_i, \beta x_i) = q \sum_{j=1}^{p} \frac{1}{(j-1)!} \left[\prod_{\ell=1}^{j-1} (q-\ell) \right] x_i^{q-1} \beta^{j-1}, \tag{4.18}$$

from which we deduce that

$$\nabla^1_{s_i} m_i(x_i,0) = \nabla^1_{s_i} T_{x^q,p}(x_i,0) = q x_i^{q-1}. \tag{4.19}$$

Suppose first that $s_i < 0$, i.e. $\beta \in (-1,0)$, and observe that $s_i^{j-1} < 0$ exactly whenever $\prod_{\ell=1}^{j-1}(q-\ell) < 0$, and thus, using $x_i \leq 1$ and (4.19), that

$$\nabla^1_{s_i} m_i(x_i,s_i) > q x_i^{q-1} = \nabla^1_{s_i} m_i(x_i,0) \quad \text{for} \quad \beta \in (-1,0). \tag{4.20}$$

Suppose now that $\beta \in (0,\frac{1}{3})$. Then (4.18) implies that

$$\begin{aligned} \nabla^1_{s_i} T_{x^q,p}(x_i,\beta x_i) &\geq q x_i^{q-1} - q \sum_{j=2}^{p} \left| \frac{1}{(j-1)!} \left[\prod_{\ell=1}^{j-1}(q-\ell) \right] \right| x_i^{q-1} (\tfrac{1}{3})^{j-1} \\ &= q x_i^{q-1} \left(1 - \sum_{j=2}^{p} \left| \frac{1}{(j-1)!} \left[\prod_{\ell=1}^{j-1}(q-\ell) \right] \right| (\tfrac{1}{3})^{j-1} \right). \end{aligned}$$

Observe now that

$$\left| \frac{1}{(j-1)!} \left[\prod_{\ell=1}^{j-1}(q-\ell) \right] \right| = \left| \prod_{\ell=1}^{j-1} \frac{q-\ell}{\ell} \right| \leq 1, \tag{4.21}$$

and therefore

$$\begin{aligned} \nabla^1_{s_i} T_{x^q,p}(x_i,\beta x_i) &\geq q x_i^{q-1} \left(1 - \sum_{j=2}^{p} (\tfrac{1}{3})^{j-1} \right) \\ &> q x_i^{q-1} \left(1 - \sum_{j=2}^{\infty} (\tfrac{1}{3})^{j-1} \right) \\ &= q x_i^{q-1} \left(1 - \tfrac{\frac{1}{3}}{1-\frac{1}{3}} \right). \end{aligned}$$

Using (4.19), this implies that

$$\nabla^1_{s_i} T_{x^q,p}(x_i,\beta x_i) \geq \frac{1}{2} \nabla^1_{s_i} T_{x^q,p}(x_i,0) \quad \text{for} \quad \beta \in [0,\tfrac{1}{3}]. \tag{4.22}$$

Suppose therefore that

$$\beta > \tfrac{1}{3}. \tag{4.23}$$

We note that (4.18) gives that

$$\nabla^1_{s_i} T_{x^q,1}(x_i,s_i) = q x_i^{q-1} \quad \text{and} \quad \nabla^1_{s_i} T_{x^q,t+2}(x_i,s_i) = \nabla^1_{s_i} T_{x^q,t}(x_i,s_i) + q x_i^{q-1} h_t(\beta)$$

for $t \in \{1,\ldots,p-2\}$ odd, where

$$\begin{aligned} h_t(\beta) &\stackrel{\rm def}{=} \frac{1}{t!} \left[\prod_{\ell=1}^{t}(q-\ell) \right] \beta^t + \frac{1}{(t+1)!} \left[\prod_{\ell=1}^{t+1}(q-\ell) \right] \beta^{t+1} \\ &= \frac{1}{t!} \left[\prod_{\ell=1}^{t}(q-\ell) \right] \beta^t \left(1 + \frac{q-(t+1)}{t+1} \beta \right). \end{aligned} \tag{4.24}$$

It is easy to verify that $h_t(\beta)$ has a root of multiplicity t at zero and another root

$$\beta_{0,t} = \frac{t+1}{t+1-q} \in \left(1\,,\,\frac{2}{2-q}\right),$$

where the last inclusion follows from the fact that $q \in (0,1)$. We also observe that $h_t(\beta)$ is a polynomial of even degree (since t is odd). Thus

$$h_t(\beta) \geq 0 \quad \text{for all} \quad \beta \geq \frac{t+1}{t+1-q} \quad \text{and} \quad t \in \{1,\ldots,p\} \text{ odd.} \tag{4.25}$$

Now

$$\begin{aligned} \frac{\nabla^1_{s_i} T_{x^q,p}(x_i,\beta x_i)}{q x_i^{q-1}} &= \frac{\nabla^1_{s_i} T_{x^q,p-2}(x_i,\beta x_i)}{q x_i^{q-1}} + h_{p-2}(\beta) \\ &= \frac{\nabla^1_{s_i} T_{x^q,1}(x_i,\beta x_i)}{q x_i^{q-1}} + \sum_{j=1,\, j \text{ odd}}^{p-2} h_j(\beta) \\ &= 1 + \sum_{\substack{j=1,\, j \text{ odd} \\ h_j(\beta)<0}}^{p-2} h_j(\beta) + \sum_{\substack{j=1,\, j \text{ odd} \\ h_j(\beta)\geq 0}}^{p-2} h_j(\beta) \\ &\geq 1 + \sum_{\substack{j=1,\, j \text{ odd} \\ h_j(\beta)<0}}^{p-2} h_j(\beta) \end{aligned} \tag{4.26}$$

where we used (4.18) to derive the third equality. Observe now that, because of (4.25),

$$\begin{aligned} \{j \in \{1,\ldots,p-2\} \text{ odd} \mid h_j(\beta) < 0\} &= \left\{ j \in \{1,\ldots,p-2\} \text{ odd} \mid \beta < \tfrac{t+1}{t+1-q} \right\} \\ &\stackrel{\text{def}}{=} \{j \in \{1,\ldots,t_0\}| \mid j \text{ odd}\} \end{aligned} \tag{4.27}$$

for some odd integer $t_0 \in \{1,\ldots,p-2\}$. Hence we deduce from (4.24) and (4.26) that

$$\frac{\nabla^1_{s_i} T_{x^q,p}(x_i,\beta x_i)}{q x_i^{q-1}} \geq 1 + \sum_{j=1}^{t_0+1} \frac{1}{j!} \left[\prod_{\ell=1}^{j} (q-\ell) \right] \beta^j. \tag{4.28}$$

Moreover, since $h_t(\beta) < 0$ for $t \in \{1,\ldots,t_0\}$ odd and observing that the second term in the first right-hand side of (4.24) is always positive for t odd, we deduce that the terms in the summation of (4.28) alternate in sign. We also note that they are decreasing in absolute value since

$$\frac{1}{(t+1)!} \left| \prod_{\ell=1}^{t+1} (q-\ell) \right| \beta^{t+1} - \frac{1}{t!} \left| \prod_{\ell=1}^{t} (q-\ell) \right| \beta^t = \frac{1}{t!} \left| \prod_{\ell=1}^{t} (q-\ell) \right| \beta^t \left(\frac{t+1-q}{t+1}\, \beta - 1 \right)$$

and (4.25) ensures that the term in brackets in the right-hand side is always negative for $q \in (0,1)$ and $t \in \{1,\ldots,t_0\}$ odd. Thus, keeping the first (most negative) term in (4.28), we obtain that

$$\nabla^1_{s_i} T_{x^q,p}(x_i,\beta x_i) \geq q x_i^{q-1}(1+(q-1)\beta) \geq \frac{q}{2-q} \nabla^1_{s_i} T_{x^q,p}(x_i,0) > \frac{q}{2} \nabla^1_{s_i} T_{x^q,p}(x_i,0). \tag{4.29}$$

where we used (4.18) to deduce the second inequality. Combining (4.20), (4.22) and (4.29) then yields that (4.16) holds for all $\beta \in (-1, \infty)$, which completes the proof since $s_i = \beta x_i$. $\Box$

Our next step is to verify that the regularization parameters $\{\sigma_{i,k}\}_{i \in \mathcal{N}}$ cannot grow unbounded.

Lemma 4.6 Suppose that AS.2 and AS.3 hold. Then, for all $i \in \mathcal{N}$ and all $k \geq 0$,

$$\sigma_{i,k} \in [\sigma_{\min}, \sigma_{\max}], \tag{4.30}$$

where $\sigma_{\max} \stackrel{\text{def}}{=} \gamma_2(p+1)L_{\max}$.

Proof. Assume that, for some $i \in \mathcal{N}$ and $k \geq 0$, $\sigma_{i,k} \geq (p+1)L_i$. Then (4.4) gives that (3.19) must fail, ensuring (4.30) because of the mechanism of the algorithm. $\Box$

We next investigate the consequences of the model's definition (3.12) when the singularity at the origin is approached.

Lemma 4.7 Suppose that AS.2 and AS.5 (and thus (4.1)) hold and that $\mathcal{H} \neq \emptyset$. Let

$$\omega \stackrel{\text{def}}{=} \min\left[1, \left(\frac{4\left[p\,\kappa_{\mathcal{N}} + \frac{|\mathcal{N}|}{p!}\varsigma^p\sigma_{\max}\right]}{q^2}\right)^{\frac{1}{q-1}}\right], \tag{4.31}$$

and suppose, in addition, that

$$\|s_k\| \leq 1 \tag{4.32}$$

and that, for some $i \in \mathcal{H}$,

$$|x_{i,k}| \in (0, \omega). \tag{4.33}$$

Then

$$\|P_{\mathcal{R}\{i\}}[\nabla_s^1 m(x_k, s_k)]\| \geq \tfrac{1}{4}q^2\omega^{q-1} \quad \text{and} \quad \operatorname{sgn}\left(P_{\mathcal{R}\{i\}}[\nabla_s^1 m(x_k, s_k)]\right) = \operatorname{sgn}(x_{i,k} + s_{i,k}) \tag{4.34}$$

where $\mathcal{R}_{\{i\}} \stackrel{\text{def}}{=} \operatorname{span}\{U_i^T\}$.

Proof. Consider $i \in \mathcal{H}$. Suppose, for the sake of simplicity, that

$$x_{i,k} > 0 \quad \text{and} \quad x_{i,k} + s_{i,k} > 0. \tag{4.35}$$

We first observe that Lemma 4.5 implies that

$$\nabla_{s_i}^1 m_i(x_{i,k}, s_i) \geq \tfrac{1}{2}q\nabla_{s_i}^1 m_i(x_{i,k}, 0) \quad \text{for all} \quad s_i \neq -x_{i,k}. \tag{4.36}$$

Moreover,

$$\nabla_s^1 m_{\mathcal{N}}(x_k,s_k) = \nabla_x^1 f_{\mathcal{N}}(x_k) + \sum_{j=2}^{p} \frac{1}{(j-1)!}\nabla_x^j f_{\mathcal{N}}(x_k)[s_k]^{j-1} + \frac{1}{p!}\sum_{\ell\in\mathcal{N}} \sigma_{\ell,k} s_{\ell,k}\|s_{\ell,k}\|^{p-1}$$

and thus, using the contractive property of orthogonal projections,(4.32), (4.1) and (4.2), that

$$\begin{aligned}\|P_{\mathcal{R}\{i\}}[\nabla_s^1 m_{\mathcal{N}}(x_k,s_k)]\| &\le \|\nabla_s^1 m_{\mathcal{N}}(x_k,s_k)\|\\ &\le \kappa_{\mathcal{N}}[1+(p-1)] + \frac{|\mathcal{N}|}{p!}\varsigma^p\sigma_{\max}\\ &= p\,\kappa_{\mathcal{N}} + \frac{|\mathcal{N}|}{p!}\varsigma^p\sigma_{\max}.\end{aligned} \tag{4.37}$$

We next successively use the linearity of $P_{\mathcal{R}\{i\}}[\cdot]$, the triangle inequality, (4.36), the facts that

$$\|U_i^T\| = 1 \quad\text{and}\quad |\nabla_{s_i}^1 m_i(x_k,s_k)| = q|x_{i,k}|^{q-1} \ge q\omega^{q-1},$$

the bound (4.37), and (4.31) to deduce that

$$\begin{aligned}\|P_{\mathcal{R}\{i\}}[\nabla_s^1 m(x_k,s_k)]\| &= \|P_{\mathcal{R}\{i\}}[\nabla_s^1 m_{\mathcal{N}}(x_k,s_k) + \nabla_s^1 \sum_{j\in\mathcal{H}} m_j(x_k,s_k)]\|\\ &= \|P_{\mathcal{R}\{i\}}[\nabla_s^1 m_{\mathcal{N}}(x_k,s_k) + \sum_{j\in\mathcal{H}} U_j^T\nabla_{s_j}^1 m_j(x_k,s_k)]\|\\ &= \|P_{\mathcal{R}\{i\}}[\nabla_s^1 m_{\mathcal{N}}(x_k,s_k)] + U_i^T\nabla_{s_i}^1 m_i(x_k,s_k)\|\\ &\ge \Big|\|U_i^T\nabla_{s_i}^1 m_i(x_k,s_k)\| - \|P_{\mathcal{R}\{i\}}[\nabla_s^1 m_{\mathcal{N}}(x_k,s_k)]\|\Big|\\ &\ge \tfrac{1}{2}q^2\omega^{q-1} - \left[p\,\kappa_{\mathcal{N}} + \frac{|\mathcal{N}|}{p!}\varsigma^p\sigma_{\max}\right]\\ &\ge \tfrac{1}{4}q^2\omega^{q-1},\end{aligned}$$

which proves the first part of (4.34) and, because of (4.36), implies the second, for the case where (4.35) holds. The proof for the cases where

$$\Big[x_{i,k} < 0 \quad\text{and}\quad x_{i,k}+s_{i,k} < 0\Big] \quad\text{or}\quad x_{i,k}(x_{i,k}+s_{i,k}) < 0$$

are identical when making use of the symmetry $m_i(x_i)$ with respect to the origin. □

Note that, like $\sigma_{\max}$, ω and β only depend on problem data. In particular, they are independent of ϵ. Lemma 4.7 has the following crucial consequence.

Lemma 4.8 Suppose that AS.2, AS.5 and the assumptions (4.32)–(4.33) of Lemma 4.7 hold and that $\mathcal{H} \neq \emptyset$. Suppose in addition that (3.15) holds at x_k, s_k. Then, either

$$|x_{i,k}+s_{i,k}| \le \epsilon \quad\text{or}\quad |x_{i,k}+s_{i,k}| \ge \omega \quad (i\in\mathcal{H}). \tag{4.38}$$

Proof. If $j \in \mathcal{H}\cap\mathcal{C}(x_k+s_k,\epsilon)$, then clearly $|x_{j,k}+s_{j,k}| \le \epsilon$, and there is nothing more to prove. Consider therefore any $j \in \mathcal{H}\setminus\mathcal{C}_k^+ \subseteq \mathcal{W}_k^+$ and observe that the separable nature of the linear optimization problem in (3.16) implies that

$$\begin{array}{rcl}
\left|\displaystyle\min_{\substack{x_k+s_k+d\in\mathcal{F}\\ d\in\mathcal{R}_{\{j\}},\|d\|\le 1}} P_{\mathcal{R}_{\{j\}}}[\nabla_s^1 m(x_k,s_k)]^T d\right| & = & \left|\displaystyle\min_{\substack{x_k+s_k+d\in\mathcal{F}\\ d\in\mathcal{R}_{\{j\}},\|d\|\le 1}} \nabla_s^1 m_{\mathcal{W}_k^+}(x_k,s_k)^T d\right| \\
& \le & \left|\displaystyle\min_{\substack{x_k+s_k+d\in\mathcal{F}\\ d\in\mathcal{R}_k^+,\|d\|\le 1}} \nabla_s^1 m_{\mathcal{W}_k^+}(x_k,s_k)^T d\right| \\
& = & \chi_m(x_k,s_k,\epsilon) \\
& \le & \frac{1}{4}q^2|x_{j,k}+s_{j,k}|^r.
\end{array} \tag{4.39}$$

Observe now that, because of the second part of (4.34) and the fact that $n_j = 1$ because of (1.2), the optimal value for the convex optimization problem in the left-hand side of this relation is given by

$$|P_{\mathcal{R}_{\{j\}}}[\nabla_s^1 m(x_k,s_k)]|\,|d_*|$$

where d_* is the problem solution and d_* has the opposite sign of $P_{\mathcal{R}_{\{j\}}}[\nabla_s^1 m(x_k,s_k)]$. Moreover, the facts that $j\in\mathcal{H}$ and (1.3) ensure that $x_{j,k}+s_{j,k}+d_j = 0$ is feasible for the optimization problem on the left-hand side of (4.39), and hence that $|d_*|\ge|x_{j,k}+s_{j,k}|$. Hence, we obtain that

$$\tfrac{1}{4}q^2\omega^{q-1}|x_{j,k}+s_{j,k}| \le \tfrac{1}{4}q^2|x_{j,k}+s_{j,k}|^r,$$

and thus, since $\omega\le 1$, that

$$|x_{j,k}+s_{j,k}| \ge \omega^{\frac{q-1}{r-1}} \ge \omega,$$

and the second alternative in (4.38) holds. □

The rest of our complexity analysis depends on the following partitioning of the set of iterations. Let the index set of the “successful” and “unsuccessful” iterations be given by

$$\mathcal{S} \stackrel{\rm def}{=} \{k\ge 0 \mid \rho_k\ge\eta\} \;\text{ and }\; \mathcal{U} \stackrel{\rm def}{=} \{k\ge 0 \mid \rho_k<\eta\}.$$

We next focus on the case where $\mathcal{H}\neq\emptyset$ and partition $\mathcal{S}$ into subsets depending on $|x_{i,k}|$ and $|x_{i,k}+s_{i,k}|$ for $i\in\mathcal{H}$. We first isolate the set of sucessful iterations which “deactivate” some variable, that is

$$\mathcal{S}_\epsilon \stackrel{\rm def}{=} \{k\in\mathcal{S} \mid |x_{i,k}+s_{i,k}|\le\epsilon \;\text{ for some }\; i\in\mathcal{H}\},$$

as well as the set of successful iterations with large steps

$$\mathcal{S}_{\|s\|} \stackrel{\rm def}{=} \{k\in\mathcal{S}\setminus\mathcal{S}_\epsilon \mid \|s_k\|>1\}. \tag{4.40}$$

Let us now choose a constant $\alpha\ge 0$ such that

$$\alpha = \left\{\begin{array}{ll} \frac{3}{4}\omega & \text{if } \;\mathcal{H}\neq\emptyset, \\ 0 & \text{otherwise.}\end{array}\right. \tag{4.41}$$

Then, at iteration $k \in \mathcal{S} \setminus (\mathcal{S}_\epsilon \cup \mathcal{S}_{\|s\|})$, we distinguish

$$\begin{aligned}
\mathcal{I}_{\heartsuit,k} &\stackrel{\rm def}{=} \Big\{ i \in \mathcal{H} \setminus \mathcal{C}_k \mid \quad |x_{i,k}| \in [\alpha, +\infty) \text{ and } |x_{i,k} + s_{i,k}| \in [\alpha, +\infty) \Big\},\\
\mathcal{I}_{\diamondsuit,k} &\stackrel{\rm def}{=} \Big\{ i \in \mathcal{H} \setminus \mathcal{C}_k \mid \quad \Big(|x_{i,k}| \in [\omega, +\infty) \text{ and } |x_{i,k} + s_{i,k}| \in (\epsilon, \alpha) \Big)\\
&\qquad\qquad \text{or } \Big(|x_{i,k}| \in (\epsilon, \alpha) \text{ and } |x_{i,k} + s_{i,k}| \in [\omega, +\infty) \Big) \Big\},\\
\mathcal{I}_{\clubsuit,k} &\stackrel{\rm def}{=} \Big\{ i \in \mathcal{H} \setminus \mathcal{C}_k \mid \quad |x_{i,k}| \in (\epsilon, \omega) \text{ and } |x_{i,k} + s_{i,k}| \in (\epsilon, \omega) \Big\}.
\end{aligned}$$

Using these notations, we further define

$$\begin{aligned}
&\mathcal{S}_\heartsuit \stackrel{\rm def}{=} \{ k \in \mathcal{S} \setminus (\mathcal{S}_\epsilon \cup \mathcal{S}_{\|s\|}) \mid \mathcal{I}_{\heartsuit,k} = \mathcal{H} \setminus \mathcal{C}_k \}, \quad \mathcal{S}_\diamondsuit \stackrel{\rm def}{=} \{ k \in \mathcal{S} \setminus (\mathcal{S}_\epsilon \cup \mathcal{S}_{\|s\|}) \mid \mathcal{I}_{\diamondsuit,k} \neq \emptyset \},\\
&\mathcal{S}_\clubsuit \stackrel{\rm def}{=} \{ k \in \mathcal{S} \setminus (\mathcal{S}_\epsilon \cup \mathcal{S}_{\|s\|}) \mid \mathcal{I}_{\clubsuit,k} \neq \emptyset \}.
\end{aligned}$$

Figure (4.2) displays the various kinds of steps in $\mathcal{S}_{\heartsuit,k}$, $\mathcal{S}_{\diamondsuit,k}$, $\mathcal{S}_{\clubsuit,k}$ and $\mathcal{S}_{\epsilon,k}$.

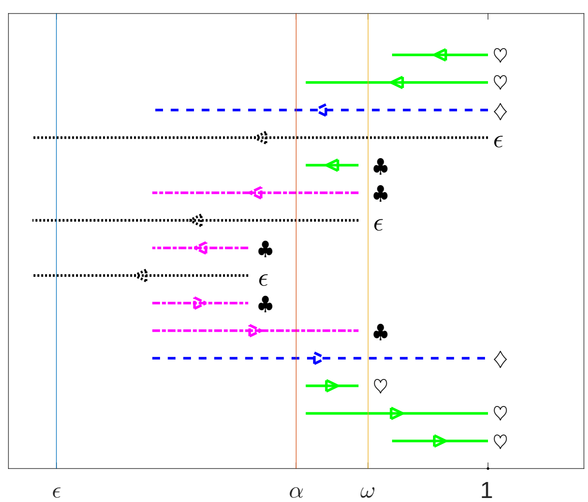

Figure 4.2: The various steps in $\mathcal{S} \setminus \mathcal{S}_{\|s\|}$ depending on intervals containing their origin $|x_{i,k}|$ and end $|x_{i,k} + s_{i,k}|$ points. The vertical lines show, in increasing order, ϵ, α and ω. The line type of the represented step indicates that it belongs to $\mathcal{S}_{\epsilon,k}$ (dotted), $\mathcal{S}_{\heartsuit,k}$ (solid), $\mathcal{S}_{\diamondsuit,k}$ (dashed) and $\mathcal{S}_{\clubsuit,k}$ (dash-dotted).The vertical axis is meaningless.

It is important to observe that the mechanism of the algorithm ensures that, once an x_i falls in the interval $[-\epsilon, \epsilon]$ at iteration k, it never leaves it (and essentially "drops out" of the calculation). Thus there are no right-oriented dotted steps in Figure 4.2 and also

$$|\mathcal{S}_\epsilon| \leq |\mathcal{H}|. \tag{4.42}$$

Moreover Lemma 4.8 ensures that $\mathcal{I}_{\clubsuit,k} = \emptyset$ for all $k \in \mathcal{S}$, and hence that

$$|\mathcal{S}_\clubsuit| = 0. \tag{4.43}$$

As a consequence, one has that $\mathcal{S}_\epsilon$, $\mathcal{S}_{\|s\|}$, $\mathcal{S}_\heartsuit$, and $\mathcal{S}_\diamondsuit$ form a partition of $\mathcal{S}$. It is also easy to verify that, if $k \in \mathcal{S}_\diamondsuit$ and $i \in \mathcal{I}_{\diamondsuit,k}$, then

$$\|s_k\| \geq \|P_{\mathcal{R}_{\{i\}}}(s_k)\| = |s_{i,k}| \geq \omega - \alpha = \tfrac{1}{4}\omega > 0, \tag{4.44}$$

where we have used the contractive property of orthogonal projections.

We now show that the steps at iterations whose index is in $\mathcal{S}_\heartsuit$ are not too short.

Lemma 4.9 Suppose that AS.1-AS.3 and AS.5 hold, that

$$\epsilon < \alpha \tag{4.45}$$

and consider $k \in \mathcal{S}_\heartsuit$ before termination. Then

$$\|s_k\| \geq (\kappa_\heartsuit\, \epsilon)^{\frac{1}{p}}, \tag{4.46}$$

where

$$\kappa_\heartsuit \stackrel{\rm def}{=} \left[2(L(\alpha) + \theta + \frac{|\mathcal{N}|}{p!}\, \varsigma_{\max}^{p+1}\, \sigma_{\max})\right]^{-1}. \tag{4.47}$$

Proof. Observe first that, since $k \in \mathcal{S}_\heartsuit \subseteq \mathcal{S}$, we have that $x_{k+1} = x_k + s_k$ and, because $\epsilon \leq \alpha$ and $\mathcal{C}_k^+ \subseteq \mathcal{C}_k$, we deduce that $\mathcal{C}_k = \mathcal{C}_k^+ = \mathcal{C}_{k+1}$ and $\mathcal{R}_k = \mathcal{R}_k^+ = \mathcal{R}_{k+1}$. Moreover the definition of $\mathcal{S}_\heartsuit$ ensures that, for all $i \in \mathcal{H} \setminus \mathcal{C}_k$,

$$\min\Big[|x_{i,k}|, |x_{i,k} + s_{i,k}|\Big] \geq \alpha. \tag{4.48}$$

Hence

$$\mathcal{O}_* \stackrel{\rm def}{=} \mathcal{O}_{k,\alpha} = \mathcal{H} \setminus \mathcal{C}_k = \mathcal{H} \setminus \mathcal{C}_k^+,$$

and thus

$$\mathcal{R}_* \stackrel{\rm def}{=} \mathcal{R}_k = \mathcal{R}_k^+ \quad \text{and} \quad \mathcal{W}_* \stackrel{\rm def}{=} \mathcal{W}_k = \mathcal{W}_k^+ = \mathcal{N} \cup \mathcal{O}_*. \tag{4.49}$$

As a consequence the step computation must have been completed because (3.15) holds, which implies that

$$\chi_m(x_k, s_k, \epsilon) = \chi_{m_{\mathcal{W}_*}}(x_k, s_k, \epsilon) = \left| \min_{\substack{x_k+s_k+d\in\mathcal{F}\\ d\in\mathcal{R}_*, \|d\|\leq 1}} \nabla_s m_{\mathcal{W}_*}(x_k, s_k)^T d \right| \leq \theta \|s_k\|^p. \tag{4.50}$$

Observe also that (4.49), (4.5) with $\phi = \alpha$ (because $k \in \mathcal{S}_\heartsuit$) , (4.30) and (4.2) then imply that

$$\begin{aligned}
\|\nabla_x^1 f_{\mathcal{W}_*}(x_{k+1}) - \nabla_s^1 m_{\mathcal{W}_*}(x_k, s_k)\| &= \|\nabla_x^1 f_{\mathcal{N}\cup\mathcal{O}_*}(x_{k+1}) - \nabla_s^1 m_{\mathcal{N}\cup\mathcal{O}_*}(x_k, s_k)\| \\
&\leq L(\alpha)\|s_k\|^p + \frac{1}{(p+1)!}\, \sigma_{\max} \sum_{i\in\mathcal{N}} \|\, \nabla_s^1 \|s_{i,k}\|^{p+1} \,\| \\
&\leq L(\alpha)\|s_k\|^p + \frac{1}{p!}\, \sigma_{\max} \sum_{i\in\mathcal{N}} \|s_{i,k}\|^p \\
&\leq L(\alpha)\|s_k\|^p + \frac{|\mathcal{N}|}{p!}\, \varsigma_{\max}^{p+1}\, \sigma_{\max} \|s_k\|^p \\
&= \left[L(\alpha) + \frac{|\mathcal{N}|}{p!}\, \varsigma_{\max}^{p+1}\, \sigma_{\max}\right] \|s_k\|^p,
\end{aligned} \tag{4.51}$$

and also that

$$\begin{aligned} \chi_f(x_{k+1},\epsilon) &= |\nabla_x^1 f_{\mathcal{W}_*}(x_{k+1})[d_{k+1}]| \\ &\le |\nabla_x^1 f_{\mathcal{W}_*}(x_{k+1})[d_{k+1}] - \nabla_s^1 m_{\mathcal{W}_*}(x_k,s_k)[d_{k+1}]| \\ &\qquad +|\nabla_s^1 m_{\mathcal{W}_*}(x_k,s_k)[d_{k+1}]|, \end{aligned} \tag{4.52}$$

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

$$\|d_{k+1}\| \le 1. \tag{4.53}$$

Assume now, for the purpose of deriving a contradiction, that

$$\|s_k\| < \left[\frac{\chi_f(x_{k+1},\epsilon)}{2(L(\alpha)+\theta+\frac{|\mathcal{N}|}{p!}\varsigma_{\max}^{p+1}\sigma_{\max})}\right]^{\frac{1}{p}} \tag{4.54}$$

at iteration $k \in \mathcal{S}_\heartsuit$. Using (4.53) and (4.51), we then obtain that

$$\begin{aligned} -\nabla_x^1 f_{\mathcal{W}_*}(x_{k+1})[d_{k+1}] &+ \nabla_s^1 m_{\mathcal{W}_*}(x_k,s_k)[d_{k+1}] \\ &\le |\nabla_x^1 f_{\mathcal{W}_*}(x_{k+1})[d_{k+1}] - \nabla_s^1 m_{\mathcal{W}_*}(x_k,s_k)[d_{k+1}]| \\ &= |(\nabla_x^1 f_{\mathcal{W}_*}(x_{k+1}) - \nabla_s^1 m_{\mathcal{W}_*}(x_k,s_k))[d_{k+1}]| \\ &\le \|\nabla_x^1 f_{\mathcal{W}_*}(x_{k+1}) - \nabla_s^1 m_{\mathcal{W}_*}(x_k,s_k)\| \, \|d_{k+1}\| \\ &< (L(\alpha)+\frac{|\mathcal{N}|}{p!}\varsigma_{\max}^{p+1}\sigma_{\max})\|s_k\|^p. \end{aligned} \tag{4.55}$$

From (4.54) and the first part of (4.52), we have that

$$\begin{aligned} -\nabla_x^1 f_{\mathcal{W}_*}(x_{k+1})[d_{k+1}] + \nabla_s^1 m_{\mathcal{W}_*}(x_k,s_k))[d_{k+1}] &< \tfrac{1}{2}\chi_f(x_{k+1},\epsilon) \\ &= -\tfrac{1}{2}\nabla_x^1 f_{\mathcal{W}_*}(x_{k+1})[d_{k+1}], \end{aligned}$$

which in turn ensures that

$$\nabla_s^1 m_{\mathcal{W}_*}(x_k,s_k)[d_{k+1}] < \tfrac{1}{2}\nabla_x^1 f_{\mathcal{W}_*}(x_{k+1})[d_{k+1}] < 0.$$

Moreover, by definition of $\chi_f(x_{k+1},\epsilon)$,

$$x_{k+1}+d_{k+1} \in \mathcal{F} \;\text{ and }\; d_{k+1} \in \mathcal{R}_{k+1} = \mathcal{R}_k^+.$$

Hence, using (3.16) and (4.53),

$$|\nabla_s^1 m_{\mathcal{W}_*}(x_k,s_k)[d_{k+1}]| \le \chi_{m_{\mathcal{W}_*}}(x_k,s_k,\epsilon). \tag{4.56}$$

We may then substitute this inequality in (4.52) to deduce as above that

$$\begin{aligned} \chi_f(x_{k+1}) &\le |\nabla_x^1 f_{\mathcal{W}_*}(x_{k+1})[d_{k+1}] - \nabla_s^1 m_{\mathcal{W}_*}(x_k,s_k)[d_{k+1}]| + \chi_{m_{\mathcal{W}_*}}(x_k,s_k,\epsilon) \\ &\le (L(\alpha)+\theta+\frac{|\mathcal{N}|}{p!}\varsigma_{\max}^{p+1}\sigma_{\max})\|s_k\|^p \end{aligned} \tag{4.57}$$

where the last inequality results from (4.55), the identity $x_{k+1} = x_k + s_k$ and (4.50). But this contradicts our assumption that (4.54) holds. Hence (4.54) must fail. The inequality (4.46) then follows by combining this conclusion with the fact that $\chi_f(x_{k+1},\epsilon) > \epsilon$ before termination. $\square$

We are now ready to consider our first complexity result, whose proof uses restrictions of the successful and unsuccessful iteration index sets defined above to $\{0,\ldots,k\}$, which are given by

$$\mathcal{S}_k \stackrel{\rm def}{=} \{0,\ldots,k\}\cap\mathcal{S}, \quad \mathcal{U}_k \stackrel{\rm def}{=} \{0,\ldots,k\}\setminus\mathcal{S}_k, \tag{4.58}$$

respectively.

Theorem 4.10 Suppose that AS.1-AS.5 hold and that

$$\epsilon \le \left[\alpha, \left(\tfrac{1}{4}\omega\kappa_{\heartsuit}^{-\frac{1}{p+1}}\right)^p\right] \quad \text{if} \quad \mathcal{H}\neq\emptyset. \tag{4.59}$$

Then Algorithm 3.1 requires at most

$$\kappa_{\mathcal{S}}(f(x_0)-f_{\rm low})\epsilon^{-\frac{p+1}{p}} + |\mathcal{H}| \tag{4.60}$$

successful iterations to return a point $x_\epsilon\in\mathcal{F}$ such that $\chi_f(x_\epsilon,\epsilon)\le\epsilon$, for

$$\kappa_{\mathcal{S}} = \frac{(p+1)!}{\eta\,\sigma_{\min}\,\varsigma_{\min}^{p+1}}\left[2(L(\alpha)+\theta+\frac{|\mathcal{N}|}{p!}\,\varsigma_{\max}^{p+1}\gamma_2)\right]^{\frac{p+1}{p}}. \tag{4.61}$$

Proof. Let $k\in\mathcal{S}$ be index of a successful iteration before termination, and suppose first that $\mathcal{H}\neq\emptyset$. Because the iteration is successful, we obtain, using AS.4 and Lemma 4.3, that

$$f(x_0)-f_{\rm low} \ge f(x_0)-f(x_{k+1}) \ge \sum_{\ell\in\mathcal{S}_k}\Big[f(x_\ell)-f(x_\ell+s_\ell)\Big] \ge \eta\sum_{\ell\in\mathcal{S}_k}\Big[f(x_\ell)-T_{f,p}(x_\ell,s_\ell)\Big]. \tag{4.62}$$

In addition to (4.58), let us define

$$\mathcal{S}_{\epsilon,k} \stackrel{\rm def}{=} \{0,\ldots,k\}\cap\mathcal{S}_\epsilon, \quad \mathcal{S}_{\|s\|,k} \stackrel{\rm def}{=} \{0,\ldots,k\}\cap\mathcal{S}_{\|s\|}, \tag{4.63}$$

$$\mathcal{S}_{\heartsuit,k} \stackrel{\rm def}{=} \{0,\ldots,k\}\cap\mathcal{S}_{\heartsuit}, \quad \mathcal{S}_{\diamondsuit,k} \stackrel{\rm def}{=} \{0,\ldots,k\}\cap\mathcal{S}_{\diamondsuit}.$$

We now use the fact that $\mathcal{S}_{\|s\|,k} \cup \mathcal{S}_{\heartsuit,k} \cup \mathcal{S}_{\diamondsuit,k} = \mathcal{S}_k \setminus \mathcal{S}_{\epsilon,k} \subseteq \mathcal{S}_k$, and (4.2) to deduce from (4.62) that

$$\begin{aligned}
f(x_0) - f_{\rm low} \;\; &\geq \;\; \eta \Bigg\{ \sum_{\ell \in \mathcal{S}_{\|s\|,k}} \Big[f(x_\ell) - T_{f,p}(x_\ell, s_\ell) \Big] + \sum_{\ell \in \mathcal{S}_{\heartsuit,k}} \Big[f(x_\ell) - T_{f,p}(x_\ell, s_\ell) \Big] \\
&\qquad + \sum_{\ell \in \mathcal{S}_{\diamondsuit,k}} \Big[f(x_\ell) - T_{f,p}(x_\ell, s_\ell) \Big] \Bigg\} \\
&\geq \;\; \frac{\eta \sigma_{\min}}{(p+1)!} \Bigg\{ |\mathcal{S}_{\|s\|,k}| \min_{\ell \in \mathcal{S}_{\|s\|,k}} \left[\sum_{i \in \mathcal{N}} \|s_{i,\ell}\|^{p+1} \right] + |\mathcal{S}_{\heartsuit,k}| \min_{\ell \in \mathcal{S}_{\heartsuit,k}} \left[\sum_{i \in \mathcal{N}} \|s_{i,\ell}\|^{p+1} \right] \\
&\qquad + |\mathcal{S}_{\diamondsuit,k}| \min_{\ell \in \mathcal{S}_{\diamondsuit,k}} \left[\sum_{i \in \mathcal{N}} \|s_{i,\ell}\|^{p+1} \right] \Bigg\} \\
&\geq \;\; \frac{\eta \sigma_{\min} \varsigma_{\min}^{p+1}}{(p+1)!} \Bigg\{ |\mathcal{S}_{\|s\|,k}| \min_{\ell \in \mathcal{S}_{\|s\|,k}} \|s_\ell\|^{p+1} + |\mathcal{S}_{\heartsuit,k}| \min_{\ell \in \mathcal{S}_{\heartsuit,k}} \|s_\ell\|^{p+1} \\
&\qquad + |\mathcal{S}_{\diamondsuit,k}| \min_{\ell \in \mathcal{S}_{\diamondsuit,k}} \|s_\ell\|^{p+1} \Bigg\}.
\end{aligned}$$

Because of of (4.40), (4.63), Lemma 4.9 and (4.44), this now yields that

$$\begin{aligned}
f(x_0) - f_{\rm low} \;\; &\geq \;\; \frac{\eta \sigma_{\min} \varsigma_{\min}^{p+1}}{(p+1)!} \Big\{ |\mathcal{S}_{\|s\|,k}| + |\mathcal{S}_{\heartsuit,k}| (\kappa_\heartsuit \epsilon)^{\frac{p+1}{p}} + |\mathcal{S}_{\diamondsuit|,k} (\omega - \alpha)^{p+1} \Big\} \\
&\geq \;\; \frac{\eta \sigma_{\min} \varsigma_{\min}^{p+1}}{(p+1)!} \Big\{ |\mathcal{S}_{\|s\|,k}| + |\mathcal{S}_{\heartsuit,k}| + |\mathcal{S}_{\diamondsuit,k}| \Big\} \min \Big[(\kappa_\heartsuit \epsilon)^{\frac{p+1}{p}}, (\tfrac{1}{4}\omega)^{p+1} \Big] \\
&\geq \;\; \frac{\eta \sigma_{\min} \varsigma_{\min}^{p+1}}{(p+1)!} \, |\mathcal{S}_k \setminus \mathcal{S}_\epsilon| \, (\kappa_\heartsuit \epsilon)^{\frac{p+1}{p}}
\end{aligned}$$

where we used (4.59), the partition of $\mathcal{S}_k \setminus \mathcal{S}_{\epsilon,k}$ in $\mathcal{S}_{\|s\|,k} \cup \mathcal{S}_{\heartsuit,k} \cup \mathcal{S}_{\diamondsuit,k}$ and the inequality $\frac{1}{4}\omega < 1$ to obtain the last inequality. Thus

$$|\mathcal{S}_k| \leq \kappa_{\mathcal{S}} (f(x_0) - f_{\rm low}) \epsilon^{-\frac{p+1}{p}} + |\mathcal{S}_{\epsilon,k}|, \tag{4.64}$$

where $\kappa_{\mathcal{S}}$ is given by (4.61). The desired iteration complexity (4.60) then follows from this bound, $|\mathcal{S}_{\epsilon,k}| \leq |\mathcal{S}_\epsilon|$ and (4.42). $\square$

To complete our analysis in terms of evaluations rather than successful iterations, we need to bound the total number of all (successful and unsuccessful) iterations.

Lemma 4.11 Assume that AS.2 and AS.3 hold. Then, for all $k \geq 0$,

$$k \leq \kappa^a |\mathcal{S}_k| + \kappa^b,$$

where

$$\kappa^a \stackrel{\rm def}{=} 1 + \frac{|\mathcal{N}| \, |\log \gamma_0|}{\log \gamma_1} \quad \text{and} \quad \kappa^b \stackrel{\rm def}{=} \frac{|\mathcal{N}|}{\log \gamma_1} \log \left(\frac{\sigma_{\max}}{\sigma_{\min}} \right).$$

Proof. For $i \in \mathcal{N}$, define

$$\mathcal{J}_{i,k} \stackrel{\rm def}{=} \{j \in \{0,\ldots,k\} \mid (3.20) \text{ holds with } k \leftarrow j\},$$

(the set of iterations where $\sigma_{i,j}$ is increased) and

$$\mathcal{D}_{i,k} \stackrel{\rm def}{=} \{j \in \{0,\ldots,k\} \mid (3.23) \text{ holds with } k \leftarrow j\} \subseteq \mathcal{S}_k$$

(the set of iterations where $\sigma_{i,j}$ in decreased), the final inclusion resulting from the condition that $\rho_k \geq \eta$ in both (3.21) and (3.22). Observe also that the mechanism of the algorithm, the fact that $\gamma_0 \in (0,1)$ and Lemma 4.6 impose that, for each $i \in \mathcal{N}$,

$$\sigma_{\min}\gamma_1^{|\mathcal{J}_{i,k}|}\gamma_0^{|\mathcal{S}_k|} \leq \sigma_{i,0}\gamma_1^{|\mathcal{J}_{i,k}|}\gamma_0^{|\mathcal{D}_{i,k}|} \leq \sigma_{i,k} \leq \sigma_{\max}.$$

Dividing by $\sigma_{\min} > 0$ and taking logarithms yields that, for all $i \in \mathcal{N}$ and all $k > 0$,

$$|\mathcal{J}_{i,k}| \log \gamma_1 + |\mathcal{S}_k| \log \gamma_0 \leq \log\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right). \tag{4.65}$$

Note now that, if (3.19) fails for all $i \in \mathcal{N}$ and given that Lemma 4.4 ensures that $f_i(x_i + s_i) \leq m_i(x_i, s_i)$ for $i \in \mathcal{H} \setminus \mathcal{C}_k^+$, then

$$\delta f_k = \sum_{i \in \mathcal{W}_k^+} \delta f_{i,k} \geq \sum_{i \in \mathcal{W}_k^+} \delta m_{i,k} = \delta m_k.$$

Thus, in view of (3.18), we have that $\rho_k \geq 1 > \eta$ and iteration k is successful. Thus, if iteration k is unsuccessful, $\sigma_{i,k}$ is increased with (3.20) for at least one $i \in \mathcal{N}$. Hence we deduce that

$$|\mathcal{U}_k| \leq \sum_{i \in \mathcal{N}} |\mathcal{J}_{i,k}| \leq |\mathcal{N}| \max_{i \in \mathcal{N}} |\mathcal{J}_{i,k}|. \tag{4.66}$$

The desired bound follows from (4.65) and (4.66) by using the fact that $k = |\mathcal{S}_k| + |\mathcal{U}_k| - 1 \leq |\mathcal{S}_k| + |\mathcal{U}_k|$, the term -1 in the equality accounting for iteration 0. □

We may now state our main evaluation complexity result.

Theorem 4.12 Suppose that AS.1, (1.3), AS.2-AS.5 and (4.59) hold. Then Algorithm 3.1 using models (3.12) for $i \in \mathcal{H}$ requires at most

$$\kappa^a \left[\kappa_{\mathcal{S}}(f(x_0) - f_{\rm low})\epsilon^{-\frac{p+1}{p}} + |\mathcal{H}|\right] + \kappa^b + 1 \tag{4.67}$$

iterations and evaluations of f and its first p derivatives to return a point $x_\epsilon \in \mathcal{F}$ such that $\chi_f(x_\epsilon, \epsilon) \leq \epsilon$.

Proof. If termination occurs at iteration 0, the theorem obviously holds. Assume therefore that termination occurs at iteration $k+1$, in which case there must be at least one successful iteration. We may therefore deduce the desired bound from Theorem 4.10, Lemma 4.11 and the fact that each successful iteration involves the evaluation of $f(x_k)$ and $\{\nabla_x^i f_{\mathcal{W}_k}(x_k)\}_{i=1}^p$, while each unsuccessful iteration only involves that of $f(x_k)$ and $\nabla_x^1 f_{\mathcal{W}_k}(x_k)$. $\Box$

Note that we may count derivatives' evaluations in Theorem 4.12 because only the derivatives of $f_{\mathcal{W}_k}$ are ever evaluated, and these are well-defined. For completeness, we state the complexity bound of the important purely Lipschitzian case.

Corollary 4.13 Suppose that AS.1-AS.4 hold and $\mathcal{H} = \emptyset$. Then Algorithm 3.1 requires at most

$$\kappa^a \left[\kappa_{\mathcal{S}}(f(x_0) - f_{\rm low})\epsilon^{-\frac{p+1}{p}}\right] + \kappa^b + 1$$

iterations and evaluations of f and its first p derivatives to return a point $x_\epsilon \in \mathcal{F}$ such that

$$\chi_f(x_\epsilon) \stackrel{\rm def}{=} \left| \min_{\substack{x+d\in\mathcal{F}\\ \|d\|\le 1}} \nabla_x^1 f_{\mathcal{W}(x)}(x)^T d \right| \le \epsilon.$$

Proof. Directly follows from Theorem 4.12, $\mathcal{H} = \emptyset$ and the obesrvation that $\mathcal{R}(x,\epsilon) = \mathbb{R}^n$ for all $x \in \mathcal{F}$ since $\mathcal{C}(x,\epsilon) = \emptyset$. $\Box$

5 Evaluation complexity for general convex $\mathcal{F}$

The two-sided model (3.12) has clear advantages, the main ones being that, except at the origin where it is non-smooth, it is polynomial and has finite gradients (and higher derivatives) over each of its two branches. It is not however without drawbacks. The first of these is that its prediction for the gradient (and higher derivatives) is arbitrarily inaccurate as the origin is approached, the second being its evaluation cost which is typically higher than evaluating $|x+s|^q$ or its derivative directly. In particular, it is the first drawback that required the careful analysis of Lemma 4.5, in turn leading, via Lemma 4.7, to the crucial Lemma 4.8. This is significant because this last lemma, in addition to the use of (3.12) and the requirement that p must be odd, also requires the 'kernel-centered' assumption (1.3), a sometimes undesirable restriction of the feasible domain geometry.

In the case where evaluating $f_{\mathcal{N}}$ is very expensive and the convex $\mathcal{F}$ is not 'kernel-centered', it may sometimes be acceptable to push the difficulty of handling the non-Lipschitzian nature of the ℓ_q norm regularization in the subproblem of computing s_k, if evaluations of $f_{\mathcal{N}}$ can be saved. In this context, a simple alternative is then to use

$$m_i(x_i, s_i) = |x_i + s_i|^q \text{ for } i \in \mathcal{H} \tag{5.1}$$

that is $m_i(x_i, s_i) = f_i(x_i + s_i)$ for $i \in \mathcal{H}$. The cost of finding a suitable step satisfying (3.15) may of course be increased, but, as we already noted, this cost is irrelevant for worst-case evaluation analysis as long as only the evaluation of $f_{\mathcal{N}}$ and its derivatives is taken into account. The choice (5.1) clearly maintains the overestimation property of Lemma 4.4.

Moreover, it is easy to verify (using AS.3 and (5.1)) that

$$\|\nabla_x f_{\mathcal{W}_k^+}(x_k+s_k)-\nabla_s^1 m_{\mathcal{W}_k^+}(x_k,s_k)\| = \|\nabla_x f_{\mathcal{N}}(x_k+s_k)-\nabla_s^1 m_{\mathcal{N}}(x_k,s_k)\| \le L_{\max}\|s_k\|^p. \tag{5.2}$$

This in turn implies that the proof of Lemma 4.9 can be extended without requiring (4.48) and using $\mathcal{O}_* = \mathcal{H}\setminus\mathcal{C}_k^+$. The derivation of (4.51) then simplifies because of (5.2) and holds for all $i \in \mathcal{H}\setminus\mathcal{C}_k^+$ with $L(\alpha) = L_{\max}$, so that (4.46) holds for all $k \in \mathcal{S}$, the assumption (4.45) being now irrelevant. This result then implies that the distinction made between $\mathcal{S}_\heartsuit$, $\mathcal{S}_\diamondsuit$, $S_\clubsuit$ and $\mathcal{S}_{\|s\|}$ is unecessary because (4.46) holds for all $k \in \mathcal{S} = \mathcal{S}_\heartsuit$. Moreover, since we no longer need Lemma 4.8 to prove that $\mathcal{S}_\clubsuit = \emptyset$, we no longer need the restrictions that p is odd and (1.3) either. As consequence, we deduce that Theorem 4.10 holds for arbitrary $p \ge 1$ and for arbitrary convex, closed non-empty $\mathcal{F}$, without the need to assume (4.59) and with $L(\alpha)$ replaced by $L_{\max}$ in (4.61). Without altering Lemma 4.11, we may therefore deduce the following complexity result.

Theorem 5.1 Suppose that AS.1, AS.2 (without the restriction that p must be odd), AS.3 and AS.4 hold. Then Algorithm 3.1 using the true models (5.1) for $i \in \mathcal{H}$ requires at most

$$\kappa^a\left[\kappa_{\mathcal{S}}^{\rm true}(f(x_0)-f_{\rm low})\epsilon^{-\frac{p+1}{p}} + |\mathcal{H}|\right] + \kappa^b + 1$$

iterations and evaluations of $f_{\mathcal{N}}$ and its first p derivatives to return a point $x_\epsilon \in \mathcal{F}$ such that $\chi_f(x_\epsilon,\epsilon) \le \epsilon$, where

$$\kappa_{\mathcal{S}}^{\rm true} = \frac{(p+1)!}{\eta\,\sigma_{\min}\,\varsigma_{\min}^{p+1}}\left[2|\mathcal{N}|\varsigma^{p+1}\left(L+\theta+\frac{\gamma_2}{p!}\right)\right]^{\frac{p+1}{p}}.$$

As indicated, the complexity is expressed in this theorem in terms of evaluations of $f_{\mathcal{N}}$ and its derivatives only. The evaluation count for the terms f_i ($i \in \mathcal{H}$) may be higher since these terms are evaluated in computing the step s_k using the models (5.1). Note that the difficulty of handling infinite derivatives is passed on to the subproblem solver in this approach.

Moreover, it also results from the analysis in this section that one may consider objective functions of the form

$$f(x) = f_{\mathcal{N}}(x) + f_{\mathcal{H}}(x)$$

and prove an $O(\epsilon^{-\frac{p+1}{p}})$ evaluation compexity bound if $f_{\mathcal{N}}$ has Lipschitz continuous derivatives of order p and if $m_{\mathcal{H}}(x_k,s) = f_{\mathcal{H}}(x_k+s)$, passing all difficulties associated with $f_{\mathcal{H}}$ to the subproblem of computing the step s_k.

As it turns out, an evaluation complexity bound may also be computed if one insist on using the Taylor's models (3.12) while allowing the feasible set to be an arbitrary convex, closed and non-empty set. Not surprisingly, the bound is (significantly) worse than that provided by Theorem 4.12, but has the merit of existing. Its derivation is based on the observation that (4.14) in Lemma 4.4 and (4.21) imply that, for $i \in \mathcal{H}\setminus\mathcal{C}_k^+$,

$$|\nabla_{s_i}^1|x_i+s_i|^q - \nabla_{s_i}^1 m_i(x_i,s_i)| \le q\left(\min\Big[|x_i|,|x_i+s_i|\Big]\right)^{q-p-1}|\mu(x_i,s_i)|^p \le q\epsilon^{q-p-1}|s_i|^p. \tag{5.3}$$

This bound can then be used in a variant of Lemma 4.9 just like (5.2) was in Section 5. In the updated version of Lemma 4.9, we replace $L(\alpha)$ by

$$L_* \stackrel{\rm def}{=} |\mathcal{N}|\,\varsigma_{\max}^p\,L_{\max} + |\mathcal{H}|\,\varsigma_{\max}^p\,q$$

and (4.51) now becomes

$$\|\nabla_x^1 f_{\mathcal{W}_k^+}(x_{k+1}) - \nabla_s^1 m_{\mathcal{W}_k^+}(x_k,s_k)\| \le \left[L_*\epsilon^{q-p-1} + \frac{|\mathcal{N}|}{p!}\varsigma^p\sigma_{\max}\right]\|s_k\|^p.$$

This results in replacing (4.57) by

$$\chi_f(x_{k+1}) \le (L_*\epsilon^{q-p-1}+\theta+\frac{|\mathcal{N}|}{p!}\,\varsigma_{\max}^{p+1}\,\sigma_{\max})\|s_k\|^p \le (L_*+\theta+\frac{|\mathcal{N}|}{p!}\,\varsigma_{\max}^{p+1}\,\sigma_{\max})\epsilon^{q-p-1}\|s_k\|^p \quad (5.4)$$

and therefore (4.46) is replaced by

$$\|s_k\| \ge \left[2\left(L_*+\theta+\frac{|\mathcal{N}|}{p!}\,\varsigma_{\max}^{p+1}\,\sigma_{\max}\right)\right]^{-\frac{1}{p}}\epsilon^{\frac{p+2-q}{p}}.$$

We may now follow the steps leading to Theorem 5.1 and deduce a new complexity bound.

Theorem 5.2 Suppose that AS.1–AS.4 hold. Then Algorithm 3.1 using the Taylor models (3.12) for $i \in \mathcal{H}$ requires at most

$$\kappa^a\left[\kappa_{\mathcal{S}}^*(f(x_0)-f_{\rm low})\epsilon^{-\frac{(p+2-q)(p+1)}{p}} + |\mathcal{H}|\right] + \kappa^b + 1$$

iterations and evaluations of f and its first p derivatives to return a point $x_\epsilon \in \mathcal{F}$ such that $\chi_f(x_\epsilon,\epsilon) \le \epsilon$, where

$$\kappa_{\mathcal{S}}^* = \frac{(p+1)!}{\eta\,\sigma_{\min}\,\varsigma_{\min}^{p+1}}\left[2\left(L_*+\theta+\frac{|\mathcal{N}|}{p!}\,\varsigma_{\max}^{p+1}\,\gamma_2\right)\right]^{\frac{p+1}{p}}.$$

Observe that, due to the second inequality of (5.4), θ can be replaced in (3.15) by $\theta_* = \theta\epsilon^{q-p-1}$, making the termination condition for the step computation very weak.

6 Further discussion

The above results suggest some additional comments.

- The complexity result in $O(\epsilon^{-(p+1)/p})$ evaluations obtained in Theorem 4.12 is identical in order to that presented in [3] for the unstructured unconstrained and in [9] for the unstructured convexly constrained cases. It is remarkable that incorporating non-Lipschitzian singularities in the objective function does not affect the worst-case evaluation complexity of finding an ϵ-approximate first-order critical point.

- Interestingly, Corollary 4.13 also shows that using partially separable structure does not affect the evaluation complexity either, therefore allowing cost-effective use of problem structure with high-order models.

- The algorithm[(6)] presented here is considerably simpler than that discussed in [16, 18] in the context of structured trust-regions. In addition, the present assumptions are also weaker. Indeed, an additional condition on long steps (see AA.1s in [18, p.364]) is no longer needed.

- Can one use even order models with Taylor models in the present framework? The main issue is that, when p is even, the two-sided model $T_{|\cdot|^q,p}(x_i, s_i)$ is no longer always an overestimate of $|x_i + s_i|^q$ when $|x_i + s_i| > |x_i|$, as can be verified from (4.14). While this can be taken care of by adding a regularization term to m_i, the necessary size of the regularization parameter may be unbounded when the iterates are sufficiently close from the singularity. This in turn destroys the good complexity because it forces the algorithm to take much too short steps.

  An alternative is to use mixed-orders models, that is models of even order (p, say) for the f_i whose index is in $\mathcal{N}$ and odd order models for those with index in $\mathcal{H}$. However, this last (odd) order has to be at least as large as p, because it is the lowest order which dominates in the crucial Lemma 4.9 where the length is bounded below away from the singularity. The choice of a $(p + 1)$-st order model for $i \in \mathcal{H}$ is then most natural.

- A variant of the algorithm can be stated where it is possible for a particular x_i to leave the ϵ-neighbourhood of zero, provided the associated step results in a significant (in view of Theorem 4.10) objective function decrease, such as a multiple of $\epsilon^{(p+1)/p}$ or some ϵ-independent constant. These decreases can then be counted separately in the argument of Theorem 4.10 and cycling is impossible since there can be only a finite number of such decreases.

7 Conclusions

We have considered the problem of minimizing a partially-separable nonconvex objective function f involving non-Lipschitzian q-norm regularization terms and subject to general convex constraints. Problems of this type are important in many areas, including data compression, image processing and bioinformatics. We have shown that the introduction of the non-Lipschitzian singularities and the exploitation of problem structure do not affect the worst-case evaluation complexity. More precisely, we have first defined ϵ-approximate first-order critical points for the considered class of problems in a way that make the obtained complexity bounds comparable to existing results for the purely Lipschitzian case. We have then shown that, if p is the (odd) degree of the models used by the algorithm, if the feasible set is 'kernel-centered' and if Taylor models are used for the q-norm regularization terms, the bound of $O(\epsilon^{-\frac{p+1}{p}})$ evaluations of f and its relevant derivatives (derived for the Lipschitzian case in [9]) is preserved in the presence of non-Lipschitzian singularities. In addition, we have shown that partially-separable structure present in the problem can be exploited (especially for high degree derivative tensors) without affecting the evaluation complexity either. We

(6) And theory, if one restricts one's attention to the case where $\mathcal{H} = \emptyset$.

have also shown that, if the difficulty of handling the non-Lipschitzian regularization terms is passed to the subproblem (which can be meaningfull if evaluating the other parts of the objective function is very expensive) in that non-Lipschitz models are used for these terms, then the same bounds hold in terms of evaluation of the expensive part of the objective function, without the restriction that the feasible set be 'kernel-centered'. A worse complexity bound has finally been provided in the case where one uses Taylor models for the q-norm regularization terms with a general convex feasible set.

These objectives have been attained by introducing a new first-order criticality measure as well as the new two-sided model of the singularity given by (3.11), which exploits the inherent symmetry and provides a useful overestimate of the $|x|^q$ if its order is chosen odd, without the need for smoothing functions.

Acknowledgements

Xiaojun Chen would like to thank Hong Kong Research Grant Council for grant PolyU153000/15p. Philippe Toint would like to thank the Belgian Fund for Scientific Research (FNRS), the University of Namur and the Hong Kong Polytechnic University for their support while this research was being conducted.

References

[1] W. Bian and X. Chen. Linearly constrained non-Lipschitzian optimization for image restoration. *SIAM Journal on Imaging Sciences*, 8:2294–2322, 2015.

[2] W. Bian and X. Chen. Optimality and complexity for constrained optimization problems with nonconvex regularization. *Mathematics of Operations Research*, (to appear), 2017.

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

[4] A.M. Bruckstein, D.L. Donoho, and M. Elad. From sparse solutions of systems of equations to sparse modeling of signals and images. *SIAM Review*, 51:34–81, 2009.

[5] 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 Journal on Optimization*, 20(6):2833–2852, 2010.

[6] C. Cartis, N. I. M. Gould, and Ph. L. Toint. Adaptive cubic overestimation methods for unconstrained optimization. Part I: motivation, convergence and numerical results. *Mathematical Programming, Series A*, 127(2):245–295, 2011.

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

[8] 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.

[9] C. Cartis, N. I. M. Gould, and Ph. L. Toint. Second-order optimality and beyond: characterization and evaluation complexity in convexly-constrained nonlinear optimization. *Foundations of Computational Mathematics*, (submitted), 2016.

[10] L. Chen, N. Deng, and J. Zhang. A trust region method with partial-update technique for unary optimization. In D. Z. Du, X. S. Zhang, and K. Cheng, editors, *Operations Research and Its Applications. Proceedings of the First International Symposium, ISORA '95*, pages 40–46, Beijing, 1995. Beijing World Publishing.

[11] L. Chen, N. Deng, and J. Zhang. Modified partial-update Newton-type algorithms for unary optimization. *Journal of Optimization Theory and Applications*, 97(2):385–406, 1998.

[12] X. Chen, D. Ge, and Y. Ye. Complexity of unconstrained ℓ_2-ℓ_p minimization. *Mathematical Programming, Series A*, 143:371–383, 2014.

[13] X. Chen, L. Guo, Z. Lu, and J.J. Ye. An augmented Lagrangian method for non-Lipschitz nonconvex programming. *SIAM Journal on Numerical Analysis*, 55:168–193, 2017.

[14] X. Chen, L. Niu, and Y. Yuan. Optimality conditions and smoothing trust region Newton method for non-Lipschitz optimization. *SIAM Journal on Optimization*, 23:1528–1552, 2013.

[15] X. Chen, F. Xu, and Y. Ye. Lower bound theory of nonzero entries in solutions of ℓ_2-ℓ_p minimization. *SIAM Journal on Scientific Computing*, 32:2832–2852, 2010.

[16] A. R. Conn, N. I. M. Gould, A. Sartenaer, and Ph. L. Toint. Convergence properties of minimization algorithms for convex constraints using a structured trust region. *SIAM Journal on Optimization*, 6(4):1059–1086, 1996.

[17] A. R. Conn, N. I. M. Gould, and Ph. L. Toint. LANCELOT*: a Fortran package for large-scale nonlinear optimization (Release A).* Number 17 in Springer Series in Computational Mathematics. Springer Verlag, Heidelberg, Berlin, New York, 1992.

[18] A. R. Conn, N. I. M. Gould, and Ph. L. Toint. *Trust-Region Methods*. MPS-SIAM Series on Optimization. SIAM, Philadelphia, USA, 2000.

[19] R. Fourer, D. M. Gay, and B. W. Kernighan. AMPL: A mathematical programming language. Computer science technical report, AT&T Bell Laboratories, Murray Hill, USA, 1987.

[20] D. M. Gay. Automatically finding and exploiting partially separable structure in nonlinear programming problems. Technical report, Bell Laboratories, Murray Hill, New Jersey, USA, 1996.

[21] D. Goldfarb and S. Wang. Partial-update Newton methods for unary, factorable and partially separable optimization. *SIAM Journal on Optimization*, 3(2):383–397, 1993.

[22] N. I. M. Gould, J. Hogg, T. Rees, and J. Scott. Solving nonlinear least-squares problems. Technical report, Rutherford Appleton Laboratory, Chilton, England, 2016.

[23] N. I. M. Gould, D. Orban, and Ph. L. Toint. CUTEst: a constrained and unconstrained testing environment with safe threads for mathematical optimization. *Computational Optimization and Applications*, 60(3):545–557, 2015.

[24] N. I. M. Gould and Ph. L. Toint. FILTRANE, a Fortran 95 filter-trust-region package for solving systems of nonlinear equalities, nonlinear inequalities and nonlinear least-squares problems. *ACM Transactions on Mathematical Software*, 33(1):3–25, 2007.

[25] A. Griewank. The modification of Newton's method for unconstrained optimization by bounding cubic terms. Technical Report NA/12, Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, United Kingdom, 1981.

[26] A. Griewank and Ph. L. Toint. On the unconstrained optimization of partially separable functions. In M. J. D. Powell, editor, *Nonlinear Optimization 1981*, pages 301–312, London, 1982. Academic Press.

[27] J. Huang, J.L. Horowitz, and S. Ma. Asymptotic properties of bridge estimators in sparse highdimensional regression models. *Annals of Statistics*, 36:587–613, 2018.

[28] Y.F. Liu, S. Ma, Y.H. Dai, and S. Zhang. A smoothing SQP framework for a class of composite l_q minimization over polyhedron. *Mathematical Programming, Series A*, 158:467–500, 2016.

[29] Z. Lu. Iterative reweighted minimization methods for l_p regularized unconstrained nonlinear programming. *Mathematical Programming, Series A*, 147:277–307, 2014.

[30] J. Mareček, P. Richtárik, and M. Takáč. Distributed block coordinate descent for minimizing partially separable functions. Technical report, Department of Mathematics and Statistics, University of Edinburgh, Edinburgh, Scotland, 2014.

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

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

[33] J. S. Shahabuddin. *Structured trust-region algorithms for the minimization of nonlinear functions*. PhD thesis, Department of Computer Science, Cornell University, Ithaca, New York, USA, 1996.