Majorization-minimization procedures and convergence of SQP methods for semi-algebraic and tame programs

Jérôme Bolte[*] and Edouard Pauwels[†]

March 31, 2015

Abstract

In view of solving nonsmooth and nonconvex problems involving complex constraints (like standard NLP problems), we study general maximization-minimization procedures produced by families of strongly convex sub-problems. Using techniques from semi-algebraic geometry and variational analysis –in particular Łojasiewicz inequality– we establish the convergence of sequences generated by this type of schemes to critical points.

The broad applicability of this process is illustrated in the context of NLP. In that case critical points coincide with KKT points. When the data are semi-algebraic or real analytic our method applies (for instance) to the study of various SQP methods: the moving balls method, $S\ell^1QP$, ESQP. Under standard qualification conditions, this provides –to the best of our knowledge– the first general convergence results for general nonlinear programming problems. We emphasize the fact that, unlike most works on this subject, no second-order conditions and/or convexity assumptions whatsoever are made. Rate of convergence are shown to be of the same form as those commonly encountered with first order methods.

Keywords. SQP methods, $S\ell^1QP$, Moving balls method, Extended Sequential Quadratic Method, KKT points, KL inequality, Nonlinear programming, Converging methods, Tame optimization.

[*]TSE (GREMAQ, Université Toulouse I), Manufacture des Tabacs, 21 allée de Brienne, 31015 Toulouse, France. E-mail: `jerome.bolte@tse-fr.eu`. Effort sponsored by the Air Force Office of Scientific Research, Air Force Material Command, USAF, under grant number FA9550-14-1-0056. This research also benefited from the support of the "FMJH Program Gaspard Monge in optimization and operations research".

[†]Faculty of Industrial Engineering and Management, Technion, Haifa, Israel. E-mail: `epauwels@ie.technion.ac.il`

Contents

1 Introduction

Many optimization methods consist in approximating a given problem by a sequence of simpler problems that can be solved in closed form or computed fast, and which eventually provide a solution or some acceptable improvement. From a mathematical viewpoint some central questions are the convergence of the sequence to a desirable point (e.g. a KKT point), complexity estimates, rates of convergence. For these theoretical purposes it is often assumed that the constraints are simple, in the sense that their projection is easy to compute (i.e. known through a closed formula), or that the objective involve nonsmooth terms whose proximal operators are available (see e.g. Combettes and Pesquet (2011) [18], Attouch et al. (2010) [3], Attouch et al. (2013) [4]). An important challenge is to go beyond this *prox friendly* ([1].) setting and to address mathematically the issue of nonconvex nonsmooth problems presenting complex geometries.

    The richest field in which these problems are met, and which was the principal motivation to this research, is probably "standard nonlinear programming" in which KKT points are generally sought through the resolution of quadratic programs of various sorts. We shall refer here to these methods under the general vocable of *SQP methods*. The bibliography on the subject is vast, we refer the readers to Bertsekas (1995) [8], Nocedal and Wright (2006) [38], Gill and

---

[1]A term we borrow from Cox et al. recent developments (2013) [19]

Wong (2012) [28], Fletcher et al. (2002) [25] (for instance) and references therein for an insight. Although these methods are quite old now –the pioneering work seems to originate in the PhD thesis of Wilson [47] in 1963– and massively used in practice, very few general convergence or complexity results are available. Most of them are local and are instances of the classical case of convergence of Newton's method ([2]) Fletcher et al. (2002) [24], Bonnans et al. (2003) [13], Nocedal and Wright (2006) [38]. Surprisingly the mere question "are the limit points KKT points?" necessitates rather strong assumptions and/or long developments – see Bonnans et al. (2003) [13, Theorem 17.2] or Burke and Han (1989) [14], Byrd et al. (2005) [15], Solodov (2009) [45] for the drawbacks of "raw SQP" in this respect, see also Bertsekas (1995) [8], Bonnans et al. (2003) [13] for some of the standard conditions/corrections/recipes ensuring that limit points are KKT.

Works in which actual convergence (or even limit point convergence) are obtained under minimal assumptions seem to be pretty scarce. In [26] (2003), Fukushima et al. provided a general SQCQP method ([3]) together with a convergence result in terms of limit points, the results were further improved and simplified by Solodov (2004) [44]. More related to the present work is the contribution of Solodov (2009) [45], in which a genuinely non-trivial proof for a SQP method to eventually provide KKT limit points is given. More recently, Auslender (2013) [5] addressed the issue of the actual convergence in the convex case by modifying and somehow reversing the basic SQP protocol: "the merit function" (see Han (1977) [29], Powell (1973) [41]) is directly used to devise descent directions as in Fletcher's pioneering $S\ell^1QP$ method (1985) [23]. In this line of research one can also quote the works of Auslender et al. (2010) [6] on the "moving balls" method – another instance of the class of SQCQP methods.

Apart from Auslender (2013) [5], Auslender et al. (2010) [6], we are not aware of other results providing actual convergence for general smooth convex functions ([4]). After our own unfruitful tries, we think this is essentially due to the fact that the dynamics of active/inactive constraints is not well understood – despite some recent breakthroughs Lewis (2002) [33], Wright (2003) [48], Hare and Lewis (2004) [30] to quote a few. In any cases "usual" methods for convergence or complexity fail and to our knowledge there are very few other works on the topic. In the nonconvex world, the recent advances of Cartis et al. (2014) [16] are first steps towards a complexity theory for NLP. Since we focus here on convergence our approach is pretty different but obviously connections and complementarities must be investigated.

Let us describe our method for addressing these convergence issues. Our approach is three-fold:

– We consider nonconvex, possibly nonsmooth, semi-algebraic/real analytic data; we actually provide results for definable sets. These model many, if not most, applications.

– Secondly, we delineate and study a wide class of *majorization-minimization methods* for nonconvex nonsmooth constrained problems. Our main assumption being that the procedures involve *locally Lipschitz continuous, strongly convex upper approximations*.
Under a general qualification assumption, we establish the convergence of the process. Once more, nonsmooth Kurdyka-Łojasiewicz (KL) inequality (Łojasiewicz (1963) [34], Kurdyka (1998) [32]) appears as an essential tool.

– Previous results are applied to derive convergence of SQP methods (Fletcher's $S\ell^1QP$ (1985) [23], Auslender (2013) [5]) and SQCQP methods (moving balls method Auslender

---

[2]Under second-order conditions and assuming that no Maratos effect [36] troubles the process.

[3]Sequential quadratically constrained quadratic programming.

[4]We focus here on SQP methods but alternative methods for treating complex constraints are available, see e.g. Cox et al. (2013) [19] and references therein.

et al. (2010) [6]). To the best of our knowledge, these are the first general nonconvex results dealing with possibly large problems with complex geometries – which are not "prox-friendly". Convergence rates have the form $O\left(\frac{1}{k^{\gamma}}\right)$ with $\gamma > 0$.

We describe now these results into more details which will also give an insight at the main results obtained in this paper.

Majorization-minimization procedures (MMP). These methods consist in devising at each point of the objective a simple upper model (e.g. quadratic forms) and to minimize/update these models dynamically in order to produce minimizing/descent sequences. This principle can be traced back, at least, to Ortega (1970) [40, section 8.3.(d)] and have found many applications since then, mostly in the statistics literature Dempster et al. (1977) [20], but also in other branches like recently in imaging sciences Chouzenoux et al. (2013) [17]. In the context of optimization, many iterative methods follow this principle, see Beck and Teboulle (2010) [7], Mairal (2013) [35] for numerous examples –and also Noll (2014) [39] where KL inequality is used to solve nonsmooth problems using a specific class of models. These procedures, which we have studied as tools, appeared to have an interest for their own sake. Our main results in this respect are self-contained and can be found in Sections 3 and 4. Let us briefly sketch a description of the MM models we use.

Being given a problem of the form

$$(\mathscr{P}) \qquad \min\left\{f(x) : x \in \mathscr{D}\right\}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a semi-algebraic continuous function and $\mathscr{D}$ is a nonempty closed semi-algebraic set, we define at each feasible point x, local semi-algebraic convex models for f and $\mathscr{D}$, respectively $h(x, \cdot) : \mathbb{R}^n \to \mathbb{R}$ – which is actually strongly convex– and $D(x) \subset \mathbb{R}^n$. We then iteratively solve problems of the form

$$x_{k+1} = p(x_k) := \operatorname{argmin}\left\{h(x_k, y) : y \in D(x_k)\right\}, \, k \in \mathbb{N}.$$

An essential assumption is that of using *upper approximations* ([5]): $D(x) \subset \mathscr{D}$ and $h(x, \cdot) \geq f(\cdot)$ on $D(x)$. When assuming semi-algebraicity of the various ingredients, convergence cannot definitely be seen as a consequence of the results in Attouch et al. (2013) [4], Bolte et al. (2013) [12]. This comes from several reasons. First, we do not have a "proper (sub)gradient method" for $(\mathscr{P})$ as required in the general protocol described in Attouch et al. (2013) [4]. A flavor of these difficulty is easily felt when considering SQP. For these methods there is, at least apparently, a sort of an unpredictability of future active/inactive constraints: the descent direction does not allow to forecast future activity and thus does not necessarily mimic an adequate subgradient of $f + i_D$ or of similar aggregate costs ([6]). Besides, even when a better candidate for being the descent function is identified, explicit features inherent to the method still remain to be dealt with.

The cornerstone of our analysis is the introduction and the study of the *value (improvement) function* $F(x) = h(x, p(x))$. It helps circumventing the possible anarchic behavior of active/inactive constraints by an implicit inclusion of future progress within the cost. We establish that the sequence x_k has a behavior very close to a subgradient method for F, see Section 4.3.

Our main result is an asymptotic alternative, a phenomena already guessed in Attouch et al. (2010) [3]: either the sequence x_k tends to infinity, or it converges to a critical point. As a

---

[5]Hence the wording of majorization-minimization method
[6]i_D denotes the indicator function as defined in Section 4.1.1

consequence, we have convergence of the sequence to *a single point* whenever the problem $(\mathscr{P})$ is coercive.

Convergence of SQP type methods. The previous results can be applied to many algorithms (see e.g. Attouch et al. (2010) [3], Bolte et al. (2013) [12], Chouzenoux et al. (2013) [17]), but we concentrate on some SQP methods for which such results are novel. In order to avoid a too important concentration of hardships, we do not discuss here computational issues of the sub-steps, the prominent role of step sizes, the difficult question of the feasibility of sub-problems, we refer the reader to Fletcher (2000) [24], Bertsekas (1995) [8], Gill and Wong (2012) [28] and references therein. We would like also to emphasize that, by construction, the methods we investigate may or may not involve hessians in their second-order term but they must systematically include a fraction of the identity as a regularization parameter, à la Levenberg-Marquardt (see e.g. Nocedal and Wright (2006) [38]). Replacing Hessian terms by their regularization or by fractions of the identity is a common approach to regularize ill-posed problems; it is also absolutely crucial when facing large scale problems see e.g. Gill et al. (2005) [27], Svanberg (2002) [46].

The aspects we just evoked above have motivated our choice of Auslender SQP method and of the moving balls method which are both relatively "simple" SQP/SQCQP methods. To show the flexibility of our approach, we also study a slight variant of $S\ell^1QP$, Fletcher (1985) [23]. This method, also known as "elastic SQP", is a modification of SQP making the sub-problems feasible by the addition of slack variables. In Gill et al. (2005) [27] the method has been adapted and redesigned to solve large scale problems (SNOPT); a solver based on this methodology is available.

For these methods, we show that a bounded sequence must converge to a single KKT point, our results rely only on semi-algebraic techniques and do not use convexity nor second order conditions. The semi-algebraic assumption can be relaxed to definability or local definability (tameness, see Ioffe (2009) [31] for an overview). We also establish that these methods come with convergence rates similar to those observed in classical first-order method (Attouch and Bolte (2009) [2]). Finally, we would like to stress that the analysis relies on geometrical tools which have a long history in the study of convergence of dynamical systems of gradient type, see for e.g. Łojasiewicz (1963) [34], Kurdyka (1998) [32].

Organization of the paper. Section 2 presents our main results concerning SQP methods. In Section 3, we describe an abstract framework for majorization-minimization methods that is used to analyze the algorithms presented in Section 2. We give in particular a general result on the convergence of MM methods. Definitions, proofs and technical aspects can be found in Section 4. Our results on MM procedures and SQP are actually valid for the broader class of real analytic or definable data, this is explained in Section 5. The Appendix (Section 6) is devoted to the technical study of SQP methods, it is shown in particular how they can be interpreted as MM processes.

2 Sequential quadratic programming for semi-algebraic and tame problems

We consider in this section problems of the form:

$$
(1) \qquad
\begin{array}{ll}
(\mathscr{P}_{\mathrm{NLP}}) & \min_{x \in \mathbb{R}^n} \quad f(x) \\
& \text{s.t.} \qquad f_i(x) \leq 0,\ i = 1, \ldots, m \\
& \qquad\qquad x \in Q
\end{array}
$$

where each f_i is twice continuously differentiable and Q is a nonempty closed convex set. Q should be thought as a "simple" set, i.e., a set whose projection is known in closed form (or "easily" computed), like for instance one of the archetypal self dual cone $\mathbb{R}^n_+$, second order cone, positive semi-definite symmetric cone (^{7}), but also an affine space, an ℓ^1 ball, the unit simplex, or a box. Contrary to Q, the set

$$\mathscr{F} = \{x, \ f_i(x) \leq 0, \ i = 1, \ldots, m\},$$

has, in general, a complex geometry and its treatment necessitates local approximations in the spirit of SQP methods. Specific assumptions regarding coercivity, regularity and constraint qualification are usually required in order to ensure correct behavior of numerical schemes, we shall make them precise for each method we present here. Let us simply recall that under these assumptions, any minimizer x of $(\mathscr{P}_{\text{NLP}})$ must satisfy the famous KKT conditions:

(2) $\qquad\qquad\qquad x \in Q, \ f_1(x) \leq 0, \ldots, f_m(x) \leq 0,$

(3) $\qquad\qquad\qquad \exists\, \lambda_1 \geq 0, \ldots, \lambda_m \geq 0,$

(4) $\qquad\qquad\qquad \nabla f(x) + \sum \lambda_i \nabla f_i(x) + N_Q(x) \ni 0,$

(5) $\qquad\qquad\qquad \lambda_i f_i(x) = 0, \forall i = 1, \ldots, m,$

where $N_Q(x)$ is the normal cone to Q at x (see Section 4.1).

SQP methods assume very different forms, we pertain here to three "simple models" with the intention of illustrating the versatility of our approach (but other studies could be led):

– Moving balls method: an SQCQP method,

– ESQP method: a merit function approach with ℓ^∞ penalty,

– Sℓ^1QP method: a merit function approach with ℓ^1 penalty.

2.1 A sequentially constrained quadratic method: the moving balls method

This method was introduced in Auslender et al. (2010) [6] for solving problems of the form of (1) with $Q = \mathbb{R}^n$. The method enters the framework of sequentially constrained quadratic problems. It consists in approximating the original problem by a sequence of quadratic problems over an intersection of balls. Strategies for simplifying constraints approximation, computations of the intermediate problems are described in Auslender et al. (2010) [6], we only focus here on the convergence properties and rate estimates. The following assumptions are necessary.

Regularity: The functions

(6) $\qquad\qquad\qquad\qquad f, f_1, \ldots, f_m : \mathbb{R}^n \to \mathbb{R}$

are C^2, with Lipschitz continuous gradients. For each $i = 1, \ldots, m$, we denote by $L_i > 0$ some Lipschitz constants of ∇f_i and by $L > 0$ a Lipschitz constant of ∇f.

Mangasarian-Fromovitz Qualification Condition (MFQC): For x in $\mathscr{F}$, set $I(x) = \{i = 1, \ldots, m : f_i(x) = 0\}$. MFQC writes

(7) $\qquad\qquad \forall x \in \mathscr{F}, \ \exists d \in \mathbb{R}^n \text{ such that } \langle \nabla f_i(x), d \rangle < 0, \forall i \in I(x).$

---

^{7}Computing the projection in that case requires to compute eigenvalues, which may be very hard for large size problems

Compactness: There exists a feasible $x_0 \in \mathscr{F}$ such that

$$(8) \qquad \{x \in \mathbb{R}^n : f(x) \leq f(x_0)\} \text{ is bounded.}$$

Remark 1 *As presented in Auslender et al. (2010) [6], the moving ball method is applicable to functions that are only C^1 with Lipschitz continuous gradient. The assumption made in (6) is therefore slightly more restrictive than the original presentation of Auslender et al. (2010) [6].*

The moving balls method is obtained by solving a sequence of quadratically constrained problems.

Moving balls method

Step 1 $x_0 \in \mathscr{F}$.

Step 2   Compute x_{k+1} solution of

$$\min_{y \in \mathbb{R}^n} \quad f(x_k) + \langle \nabla f(x_k), y - x_k \rangle + \frac{L}{2}\|y - x_k\|^2$$
$$\text{s.t.} \quad f_i(x_k) + \langle \nabla f_i(x_k), y - x_k \rangle + \frac{L_i}{2}\|y - x_k\|^2 \leq 0, \, i = 1 \ldots m$$

The algorithm can be proven to be well defined and to produce a feasible method provided that x_0 is feasible, i.e.,

$$x_k \in \mathscr{F}, \forall k \geq 0.$$

These aspects are thoroughly described in Auslender et al. (2010) [6].

Theorem 2.1 (Convergence of the moving balls method) *Recall that $Q = \mathbb{R}^n$ and assume that the following conditions hold*

– *The functions $f, f_1, \ldots, f_m$ are semi-algebraic,*

– *Lipschitz continuity conditions (6),*

– *Mangasarian-Fromovitz qualification condition (7),*

– *boundedness condition (8),*

– *feasibility of the starting point $x_0 \in \mathscr{F}$.*

Then,

(i) The sequence $\{x_k\}_{k \in \mathbb{N}}$ defined by the moving balls method converges to a feasible point x_∞ satisfying the KKT conditions for the nonlinear programming problem $(\mathscr{P}_{\text{NLP}})$.

(ii) Either convergence occurs in a finite number of steps or the rate is of the form:

    (a) $\|x_k - x_\infty\| = O(q^k)$, with $q \in (0, 1)$,

    (b) $\|x_k - x_\infty\| = O\left(\frac{1}{k^\gamma}\right)$, with $\gamma > 0$.

2.2   Extended sequential quadratic method

ESQM method (and Sℓ^1QP) grounds on the well known observation that an NLP problem can be reformulated as an "unconstrained problem" involving an exact penalization. Set $f_0 = \mathbf{0}$ and consider

$$(9) \qquad \min_{x \in Q} \left\{ f(x) + \beta \max_{i=0,\ldots,m} f_i(x) \right\}$$

where β is positive parameter. Under mild qualification assumptions and for β sufficiently large, critical points of the above are KKT points of the initial nonlinear programming $(\mathscr{P}_{\text{NLP}})$. Building on this fact, ESQM is devised as follows:

- At a fixed point x (non necessarily feasible), form a model of (9) such that:
  - complex terms $f, f_1, \ldots, f_m$ are linearized,
  - a quadratic term $\frac{\beta}{2}||y - x||^2$ is added both for conditioning and local control,
- minimize the model to find a descent direction and perform a step of size $\lambda > 0$,
- both terms λ, β are adjusted online:
  - λ is progressively made smaller to ensure a descent condition,
  - β is increased to eventually reach a threshold for exact penalization.

We propose here a variation of this method which consists in modifying the quadratic penalty term instead of relying on a line search procedure to ensure some sufficient decrease. For a fixed x in $\mathbb{R}^n$, we consider a local model of the form:

$$h_\beta(x, y) = f(x) + \langle \nabla f(x), y - x \rangle + \beta \max_{i=0,\ldots,p} \{f_i(x) + \langle \nabla f_i(x), y - x \rangle\}$$
$$+ \frac{(\lambda + \beta \lambda')}{2} ||y - x||^2 + i_Q(y),$$

where β is a parameter and $\lambda, \lambda' > 0$ are fixed.

As we shall see this model is to be iteratively used to provide descent directions and ultimately KKT points. Before describing into depth the algorithm, let us state our main assumptions (recall that $\mathscr{F} = \{x \in \mathbb{R}^n : f_i(x) \leq 0, \forall i = 1, \ldots, m\}$).

Regularity: The functions

(10) $$f, f_1, \ldots, f_m : \mathbb{R}^n \to \mathbb{R}$$

are C^2, with Lipschitz continuous gradients. For each $i = 1, \ldots, m$, we denote by $L_i > 0$ some Lipschitz constants of ∇f_i and by $L > 0$ a Lipschitz constant of ∇f. We also assume that the step size parameters satisfy

(11) $$\lambda \geq L \text{ and } \lambda' \geq \max_{i=1,\ldots,m} L_i.$$

Compactness: For all real numbers $\mu_1, \ldots, \mu_m$, the set

(12) $$\{x \in Q, \ f_i(x) \leq \mu_i, \ i = 1, \ldots, m\} \text{ is compact.}$$

Boundedness:

(13) $$\inf_{x \in Q} f(x) > -\infty.$$

Qualification condition: The function

$$\max_{i=1,\ldots,m} f_i + i_Q$$

has *no critical points* on the set $\left\{x \in Q : \exists i = 1, \ldots, m, f_i(x) \geq 0\right\}$.

Equivalently, $\forall x \in \{x \in Q : \exists i = 1, \ldots, m, f_i(x) \geq 0\}$, there cannot exist $\{u_i\}_{i \in I}$ such that

(14) $$u_i \geq 0, \ \sum_{i \in I} u_i = 1, \ \sum_{i \in I} \langle u_i \nabla f_i(x), z - x \rangle \geq 0, \ \forall z \in Q,$$

where $I = \left\{j > 0, \ f_j(x) = \max_{i=1,\ldots,m}\{f_i(x)\}\right\}$.

Remark 2 (a) Set $J = \{1, \ldots, m\}$. The intuition behind this condition is simple: $\max_{i \in J} f + i_Q$ is assumed to be (locally) "sharp" and thus $\beta \max_{i \in J}(\mathbf{0}, f_i) + i_Q$ resembles $i_{Q \cap \mathscr{F}}$ for big β.

(b) The condition (14) implies the generalized Mangasarian-Fromovitz condition (sometimes called Robinson condition):

$$\forall x \in Q \cap \mathscr{F}, \exists y \in Q \setminus \{x\}, \ \langle \nabla f_i(x), y - x \rangle < 0, \forall i = 1, \ldots, m, \ \text{ such that } f_i(x) = 0.$$

(c) Assume $Q = \mathbb{R}^n$. The qualification condition (14) implies that the feasible set is connected, which is a natural extension of the more usual convexity assumption. [Proof. Argue by contradiction and assume that the feasible set has at least two connected components. Take two points a, b in each of these components. The function $g = \max\{f_i : i = 1, \ldots, m\}$ satisfies $g(a) = g(b) = 0$. Using the compactness assumption (12), the conditions of the mountain pass theorem (Shuzhong (1985) [43, Theorem 1]) are thus satisfied. Hence, there exists a critical point c such that $g(c) > 0$ (strictly speaking, this is a Clarke critical point, but in this specific setting, this corresponds to the notion of crtical point we use un this paper see Rockafellar and Wets (1998) [42, Theorem 10.31]). Thence c is non feasible and the criticality of c contradicts our qualification assumption.]

Let us finally introduce *feasibility test functions*

(15) $$\text{test}_i(x, y) = f_i(x) + \langle \nabla f_i(x), y - x \rangle$$

for all $i = 1, \ldots, m$ and x, y in $\mathbb{R}^n$.

Remark 3 (Online feasibility test) We shall use the above functions for measuring the quality of β_k. These tests function will also be applied to the analysis of $S\ell^1 QP$. Depending on the information provided by the algorithm, other choices could be done, as for instance $\text{test}_i(x, y) = f_i(x) + \langle \nabla f_i(x), y - x \rangle + \frac{L_{f_i}}{2}\|y - x\|$ or simply $\text{test}_i(x, y) = f_i(y)$.

We proceed now to the description of the algorithm.

Extended Sequential Quadratic Method (ESQM)

Step 1   Choose $x_0 \in Q$, β_0, $\delta > 0$

Step 2   Compute the unique solution x_{k+1} of $\min_{y \in \mathbb{R}^n} h_{\beta_k}(x_k, y)$,
i.e. solve for y (and s) in:
$$\min_{y,s} \quad f(x_k) + \langle \nabla f(x_k), y - x_k \rangle + \beta_k s + \frac{(\lambda + \beta_k \lambda')}{2}\|y - x\|^2$$
$$\text{s.t.} \quad f_i(x_k) + \langle \nabla f_i(x_k), y - x_k \rangle \leq s, i = 1 \ldots m,$$
$$s \geq 0$$
$$y \in Q.$$

Step 3   If $\text{test}_i(x_k, x_{k+1}) \leq 0$ for all $i = 1, \ldots, m$, then $\beta_{k+1} = \beta_k$, otherwise $\beta_{k+1} = \beta_k + \delta$

Remark 4 (a) Working with quadratic terms involving Hessians in h_{β_k} is possible provided that local models are upper approximations (one can work for instance with approximate functions à la Levenberg-Marquardt Nocedal and Wright (2006) [38]).

(b) The algorithm presented in Auslender (2013) [5] is actually slightly different from the one above. Indeed, the quadratic penalty term was there simply proportional to β and the step sizes were chosen by line search. Original ESQP could thus be seen as a kind of backtracking

version of the above method.

(c) Let us also mention that many updates rules are possible, in particular rules involving upper bounds of local Lagrange multipliers. The essential aspect is that exact penalization is reached in a finite number of iterations.

(d) Observe that the set Q of simple constraints is kept as is in the sub-problems.

The convergence analysis carried out in Auslender (2013) [5] can be extended to our setting, leading to the following theorem (note we do not use the semi-algebraicity assumptions).

Theorem 2.2 (Auslender (2013) [5]) *Assume that the following properties hold*

– *Lipschitz continuity conditions* (10),

– *steplength conditions* (11),

– *qualification assumption* (14),

– *boundedness assumptions* (12), (13),

then the sequence of parameters β_k stabilizes after a finite number of iterations k_0 and all cluster points of the sequence $\{x_k\}_{k\in\mathbb{N}}$ are KKT points of the nonlinear programming problem $(\mathscr{P}_{\mathrm{NLP}})$.

The application of the techniques developed in this paper allow to prove a much stronger result:

Theorem 2.3 (Convergence of ESQM) *Assume that the following conditions hold*

– *The functions $f, f_1, \ldots, f_m$ and the set Q are real semi-algebraic,*

– *Lipschitz continuity condition* (10),

– *steplength condition* (11),

– *qualification assumption* (14),

– *boundedness assumptions* (12), (13),

Then,

(i) *The sequence $\{x_k\}_{k\in\mathbb{N}}$ generated by (ESQM) converges to a feasible point x_∞ satisfying the KKT conditions for the nonlinear programming problem $(\mathscr{P}_{\mathrm{NLP}})$.*

(ii) *Either convergence occurs in a finite number of steps or the rate is of the form:*
   (a) $\|x_k - x_\infty\| = O(q^k)$, *with $q \in (0,1)$,*
   (b) $\|x_k - x_\infty\| = O\left(\frac{1}{k^\gamma}\right)$, *with $\gamma > 0$.*

This result gives a positive answer to the "Open problem 3" in Auslender (2013) [5, Section 6] (with a slightly modified algorithm).

2.3 $S\ell^1 QP$ aka "elastic sequential quadratic method"

The $S\ell^1 QP$ is an ℓ^1 version of the previous method. It seems to have been introduced in the eighties by Fletcher[23]. Several aspects of this method are discussed in Fletcher (2000) [24]; see also Gill et al. (2005) [27] for its use in the resolution of large size problems (SNOPT algorithm). The idea is based this time on the minimization of the ℓ^1 penalty function:

$$(16) \qquad \min_{x \in Q} f(x) + \beta \sum_{i=0}^{m} f_i^+(x)$$

where β is a positive parameter and where we have set $a^+ = \max(0, a)$ for any real number a. Local models are of the form:

$$h_\beta(x, y)$$
$$= f(x) + \langle \nabla f(x), y - x \rangle + \beta \sum_{i=1}^{m} (f_i(x) + \langle \nabla f_i(x), y - x \rangle)^+$$
$$+ \frac{(\lambda + \beta \lambda')}{2} ||y - x||^2 + i_Q(y), \quad \forall x, y \in \mathbb{R}^n,$$

where as previously $\lambda, \lambda' > 0$ are fixed parameters. Using slack variables the minimization of $h_\beta(x, .)$ amounts to solve the problem

$$
\begin{array}{ll}
\min & f(x) + \langle \nabla f(x), y - x \rangle + \beta \sum_{i=1}^{m} s_i + \frac{(\lambda + \beta \lambda')}{2} ||y - x||^2 \\
\text{s.t.} & f_i(x) + \langle \nabla f_i(x), y - x \rangle \leq s_i, \, i = 1 \ldots m \\
& s_1, \ldots, s_m \geq 0 \\
& y \in Q.
\end{array}
$$

Once again, the above is very close to the "usual" SQP step, the only difference being the elasticity conferred to the constraints by the penalty term.

The main requirements needed for this method are quasi-identical to those we used for ESQP: we indeed assume (10), (14), (12), (13), while (11) is replaced by:

$$(17) \qquad \lambda \geq L \text{ and } \lambda' \geq \sum_{i=1}^{m} L_i.$$

The latter is more restrictive in the sense that smaller step lengths are required, but on the other hand this restriction comes with more flexibility in the relaxation of the constraints.

In the description of the algorithm below, we make use the test functions (15) described in the previous section.

$S\ell^1 QP$

Step 1   Choose $x_0 \in Q$, $\beta_0, \delta > 0$

Step 2   Compute the unique solution x_{k+1} of $\min_{y \in \mathbb{R}^n} h_{\beta_k}(x_k, y)$,
i.e. solve for y (and s) in:
$$
\begin{array}{ll}
\min & f(x) + \langle \nabla f(x), y - x \rangle + \beta_k \sum_{i=1}^{m} s_i + \frac{(\lambda + \beta_k \lambda')}{2} ||y - x||^2 \\
\text{s.t.} & f_i(x) + \langle \nabla f_i(x), y - x \rangle \leq s_i, \, i = 1 \ldots m \\
& s_1, \ldots, s_m \geq 0 \\
& y \in Q.
\end{array}
$$

Step 3   If $\text{test}_i(x_k, x_{k+1}) \leq 0$ for all $i = 1, \ldots, m$, then $\beta_{k+1} = \beta_k$,
otherwise $\beta_{k+1} = \beta_k + \delta$

The convergence in terms of limit points and for the sequence β_k is similar to that of previous section. In this theorem semi-algebraicity is not necessary.

Theorem 2.4 *Assume that the following properties hold*

  – *Lipschitz continuity conditions* (10),

  – *steplength conditions* (17),

  – *qualification assumption* (14),

  – *boundedness assumptions* (12), (13),

then the sequence of parameters β_k stabilizes after a finite number of iterations k_0 and all cluster points of the sequence $\{x_k\}_{k\in\mathbb{N}}$ are KKT points of the nonlinear programming problem $(\mathscr{P}_{\mathrm{NLP}})$.

We obtain finally the following result:

Theorem 2.5 (Convergence of $\mathbf{S\ell^1QP}$) *Assume that the following conditions hold*

  – *The functions $f, f_1, \ldots, f_m$ and the set Q are semi-algebraic,*

  – *Lipschitz continuity condition* (10),

  – *steplength condition* (17),

  – *qualification assumption* (14),

  – *boundedness assumptions* (12), (13).

Then,

 (i) *The sequence $\{x_k\}_{k\in\mathbb{N}}$ generated by (ESQM) converges to a feasible point x_∞ satisfying the KKT conditions for the nonlinear programming problem $(\mathscr{P}_{\mathrm{NLP}})$.*

 (ii) *Either convergence occurs in a finite number of steps or the rate is of the form:*

  (a) $\|x_k - x_\infty\| = O(q^k)$, *with $q \in (0,1)$,*

  (b) $\|x_k - x_\infty\| = O\left(\frac{1}{k^\gamma}\right)$, *with $\gamma > 0$.*

3   Majorization-minimization procedures

3.1   Sequential model minimization

We consider a general problem of the form

$$(18) \qquad\qquad (\mathscr{P}) \qquad \min\left\{ f(x) : x \in \mathscr{D} \right\}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a continuous function and $\mathscr{D}$ is a nonempty closed set.

In what follows we study the properties of majorization-minimization methods. At each feasible point, x, local convex models for f and $\mathscr{D}$ are available, say $h(x, \cdot) : \mathbb{R}^n \to \mathbb{R}$ and $D(x) \subset \mathbb{R}^n$; we then iteratively solve problems of the form

$$x_{k+1} \in \operatorname{argmin}\left\{ h(x_k, y) : y \in D(x_k) \right\}.$$

In order to describe the majorization-minimization method we study, some elementary notions from variational analysis and semi-algebraic geometry are required. However, since the concepts and notations we use are quite standard, we have postponed their formal introduction in Section 4.1 page 17. We believe this contributes to a smoother presentation of our results.

3.2 Majorization-minimization procedures

For the central problem at stake

$$(\mathscr{P}) \quad \min\left\{ f(x) : x \in \mathscr{D} \right\}$$

we make the following standing assumptions

$$(\mathscr{S}) \quad \begin{cases} f : \mathbb{R}^n \to \mathbb{R} \text{ is locally Lipschitz continuous, subdifferentially regular and semi-algebraic,} \\ \inf\left\{ f(x) : x \in \mathscr{D} \right\} > -\infty, \\ \mathscr{D} \subset \mathbb{R}^n \text{ is nonempty, closed, regular and semi-algebraic.} \end{cases}$$

Remark 5 (Role of regularity) The meaning of the terms subdifferential regularity/regularity is recalled in the next section. It is important to mention that these two assumptions are only used to guarantee the good behavior of the sum rule (and thus of KKT conditions)

$$\partial\left(f + i_{\mathscr{D}} \right)(x) = \partial f(x) + N_{\mathscr{D}}(x), \ x \in \mathscr{D}.$$

One could thus use alternative sets of assumptions, like: f is C^1 and D is closed (not necessarily regular).

A critical point $x \in \mathbb{R}^n$ for $\left(\mathscr{P}\right)$ is characterized by the relation $\partial(f + i_{\mathscr{D}})(x) \ni 0$, i.e. using the sum rule:

$$\partial f(x) + N_{\mathscr{D}}(x) \ni 0 \quad \text{(Fermat's rule for constrained optimization)}.$$

When $\mathscr{D}$ is a nonempty intersection of sublevel sets, as in Section 2, it necessarily satisfies the assumptions $\left(\mathscr{S}\right)$ (see Appendix). Besides, by using the generalized Mangasarian-Fromovitz qualification condition at x, one sees that Fermat's rule exactly amounts to KKT conditions (see Proposition 4.1).

Inner convex constraints approximation

Constraints are locally modeled at a point $x \in \mathbb{R}^n$ by a subset $D(x)$ of $\mathbb{R}^n$. One assumes that $D : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ satisfies

$$(19) \quad \begin{cases} \operatorname{dom} D \supset \mathscr{D}, \\ D(x) \subset \mathscr{D} \text{ and } N_{D(x)}(x) \subset N_{\mathscr{D}}(x), \forall x \in \mathscr{D}, \\ D \text{ has closed convex values,} \\ D \text{ is continuous (in the sense of multivalued mappings).} \end{cases}$$

Local strongly convex upper models for f

Fix $\mu > 0$.

$$
(20) \quad
\begin{cases}
\text{The family of local models} \\[2mm]
\qquad h : \begin{cases} \mathbb{R}^n \times \mathbb{R}^n & \longrightarrow & \mathbb{R} \\ (x,\ y) & \longrightarrow & h(x,y) \end{cases} \\[4mm]
\text{satisfies:} \\[3mm]
\quad \textit{(i)}\ \ h(x,x) = f(x) \text{ for all } x \text{ in } \mathscr{D}, \\[2mm]
\quad \textit{(ii)}\ \ \partial_y h(x,y)|_{y=x} \subset \partial f(x) \text{ for all } x \text{ in } \mathscr{D}, \\[2mm]
\quad \textit{(iii)}\ \text{For all } x \text{ in } \mathscr{D},\ h(x,y) \geq f(y), \forall y \in D(x), \\[2mm]
\quad \textit{(iv)}\ \ h \text{ is continuous. For each fixed } x \text{ in } \mathscr{D}, \text{ the function } h(x,\cdot) \text{ is } \mu \text{ strongly} \\
\qquad\quad \text{convex.}
\end{cases}
$$

Example 1 (a) A typical, but important, example of upper approximations that satisfy these properties comes from the descent lemma (Lemma 4.4). Given a C^1 function f with L_f-Lipschitz continuous gradient we set

$$
h(x,y) = f(x) + \langle \nabla f(x), y - x \rangle + \frac{L_f}{2}\|x - y\|^2.
$$

Then h satisfies all of the above (with $\mathscr{D} = D(x) = \mathbb{R}^n$ for all x).
(b) SQP methods of the previous section provide more complex examples.

A qualification condition for the surrogate problem

We require a relation between the minimizers of $y \to h(x,y)$ and the general variations of h. Set
$$
\hat{h}(x,y) = h(x,y) + i_{D(x)}(y),
$$
for all x,y in $\mathbb{R}^n$, and $\hat{h}(x,y) = +\infty$ whenever $D(x)$ is empty.

For any compact $C \subset \mathbb{R}^n$, there is a constant $K(C)$, such that,

$$
(21) \qquad x \in \mathscr{D} \cap C,\ y \in D(x) \cap C \text{ and } (v,0) \in \partial \hat{h}(x,y) \implies \|v\| \leq K(C)\|x - y\|.
$$

The iteration mapping and the value function

For any fixed x in $\mathscr{D}$, we define the iteration mapping as the solution of the sub-problem

$$
(22) \qquad \Big(\mathscr{P}(x) \Big) \quad \min\Big\{ h(x,y) : y \in D(x) \Big\}
$$

that is

$$
(23) \qquad p(x) := \operatorname{argmin}\Big\{ h(x,y) : y \in D(x) \Big\}.
$$

We set for all x in $\mathscr{D}$,

$$
(24) \qquad \operatorname{val}(x) = \text{value of } \mathscr{P}(x) = h(x,p(x)),
$$

and $\operatorname{val}(x) = +\infty$ otherwise.

Remark 6 (a) The restriction "x belongs to $\mathscr{D}$" is due to the fact that our process is based on upper approximations and thus it is a *feasible model* (i.e. generating sequences in $\mathscr{D}$). Note however that this does not mean that non feasible methods cannot be studied with this process (see ESQP and Sℓ^1QP in the previous section) .

(b) Recalling Example 1, assuming further that $D(x) = \mathscr{D}$ for all x, and denoting by $P_{\mathscr{D}}$ the projection onto $\mathscr{D}$, the above writes

$$p(x) = P_{\mathscr{D}}\left(x - \frac{1}{L_f}\nabla f(x)\right).$$

With these simple instances for h and D, we recover the gradient projection iteration mapping. Note also that for this example $\partial \hat{h}(x, y) = (v, 0)$ implies that

$$v = (LI_n - \nabla^2 f(x))(x - p(x)).$$

Thus the qualification assumption is trivially satisfied whenever f is C^2.

Our general procedure can be summarized as:

Majorization-minimization procedure (MMP)

Assume the local approximations satisfy the assumptions:

- inner constraints approximation (19),

- upper objective approximation (20),

- qualification conditions (21),

Let x_0 be in $\mathscr{D}$ and define iteratively

$$x_{k+1} = p(x_k),$$

where p is the iteration mapping (23).

Example 2 Coming back to our model example, (MMP) reduces simply to the gradient projection method

$$x_{k+1} = P_{\mathscr{D}}\left(x_k - \frac{1}{L_f}\nabla f(x_k)\right), \ x_0 \in \mathscr{D}.$$

3.3 Main convergence result

Recall the standing assumptions $(\mathscr{S})$ on $(\mathscr{P})$, our main "abstract" contribution is the following theorem.

Theorem 3.1 (Convergence of MMP for semi-algebraic problems) *Assume that the local model pair $\left(h, D(\cdot)\right)$ satisfies:*

- *the inner convex constraints assumptions* (19),

- *the upper local model assumptions* (20),

- *the qualification assumptions* (21),

– *the tameness assumptions: f, h and D are real semi-algebraic.*

Let $x_0 \in \mathscr{D}$ be a feasible starting point and consider the sequence $\{x_k\}_{k=1,2,\ldots}$ defined by $x_{k+1} = p(x_k)$. Then,

(I) *The following asymptotic alternative holds*

  (i) *either the sequence $\{x_k\}_{k=1,2,\ldots}$ diverges, i.e. $\|x_k\| \to +\infty$,*

  (ii) *or it converges to a single point x_∞ such that*

$$\partial f(x_\infty) + N_{\mathscr{D}}(x_\infty) \ni 0.$$

(II) *In addition, when x_k converges, either it converges in a finite number of steps or the rate of convergence is of the form:*

  (a) $\|x_k - x_\infty\| = O(q^k)$, *with $q \in (0,1)$,*

  (b) $\|x_k - x_\infty\| = O\left(\frac{1}{k^\gamma}\right)$, *with $\gamma > 0$.*

Remark 7 (Coercivity/Divergence) (a) If in addition $[f \leq f(x_0)] \cap D$ is bounded, the sequence x_k cannot diverge and converges thus to a critical point.
(b) The divergence property $(I) - (i)$ is a positive result, a *convergence result*, which does not correspond to a failure of the method but rather *to the absence of minimizers in a given zone.*

Theorem 3.1 draws its strength from the fact that majorization-minimization schemes are ubiquitous in continuous optimization (see Beck and Teboulle (2010) [7]). This is illustrated with SQP methods but other applications can be considered.

The proof (to be developed in the next section) is not trivial but the ideas can be briefly sketched as follows:

- Study the auxiliary function, the "value improvement function":

$$F = \text{val} : \begin{cases} \mathscr{D} \to & \mathbb{R} \\ x \to & h(x, p(x)). \end{cases}$$

- Show that there is a non-negative constants K_1 such that sequence of iterates satisfies:

$$F(x_k) + K_1 \|x_k - x_{k+1}\|^2 \leq f(x_k) \leq F(x_{k-1})$$

- Show that for any compact C, there is a constant $K_2(C)$ such that if $x_k \in C$, we have:

$$\text{dist}\left(0, \partial F(x_k)\right) \leq K_2(C)\|x_{k+1} - x_k\|.$$

- Despite the *explicit type* of the second inequality, one may use KL property (see Section 4.1) and techniques akin to those presented in Bolte et al. (2013) [12], Attouch et al. (2013) [4] to obtain convergence of the iterative process.

4 Convergence analysis of majorization-minimization procedures

This section is entirely devoted to the exposition of the technical details related to the proof of Theorem 3.1.

16

4.1 Some concepts for nonsmooth and semi-algebraic optimization

We hereby recall a few definitions and concepts that structure our main results. In particular, we introduce the notion of a subdifferential and of a KL function, which are the most crucial tools used in our analysis.

4.1.1 Nonsmooth functions and subdifferentiation

A detailed exposition of these notions can be found in Rockafellar and Wets (1998) [42]. In what follows, g denotes a proper lower semi-continuous function from $\mathbb{R}^n$ to $(-\infty, +\infty]$ whose domain is denoted and defined by $\operatorname{dom} g = \{x \in \mathbb{R}^n : g(x) < +\infty\}$. Recall that g is called proper if $\operatorname{dom} g \neq \emptyset$.

Definition 1 (Subdifferentials) Let g be a proper lower semicontinuous function from $\mathbb{R}^n$ to $(-\infty, +\infty]$.

1. Let $x \in \operatorname{dom} g$, the *Fréchet subdifferential of g at x* is the subset of vectors v in $\mathbb{R}^n$ that satisfy

$$\liminf_{y \to x, \, y \neq x} \frac{g(y) - g(x) - \langle v, y - x \rangle}{\|x - y\|} \geq 0.$$

When $x \notin \operatorname{dom} g$, the Fréchet subdifferential is empty by definition. The Fréchet subdifferential of g at x is denoted by $\hat{\partial} g(x)$.

2. The *limiting subdifferential*, or simply *the subdifferential of g at x*, is defined by the following closure process:

$$\partial g(x) = \{v \in \mathbb{R}^n : \exists x_j \to x, g(x_j) \to g(x), u_k \in \hat{\partial} g(x_j), u_j \to v \text{ as } j \to \infty\}.$$

3. Assume g is finite valued and locally Lipschitz continuous. The function g is said to be *subdifferentially regular*, if $\hat{\partial} g(x) = \partial g(x)$ for all x in $\mathbb{R}^n$.

Being given a closed subset C of $\mathbb{R}^n$, its *indicator function $i_C : \mathbb{R}^n \to (-\infty, +\infty]$* is defined as follows

$$i_C(x) = 0 \text{ if } x \in C, \; i_C(x) = +\infty \text{ otherwise.}$$

C is said to be *regular* if $\hat{\partial} i_C(x) = \partial i_C(x)$ on C. In this case, the *normal cone to C* is defined by the identity

$$N_C(x) = \partial i_C(x), \forall x \in \mathbb{R}^n.$$

The *distance function to C* is defined as

$$\operatorname{dist}(x, C) = \min \{\|x - y\| : y \in C\}.$$

We recall the two following fundamental results.

Proposition 4.1 (Fermat's rule, critical points, KKT points) *We have the following extensions of the classical Fermat's rule:*

(i) When x is a local minimizer of g, then $0 \in \partial g(x)$.

(ii) If x is a local minimizer of $(\mathscr{P})$, under assumption $(\mathscr{S})$, then:

$$\partial f(x) + N_{\mathscr{D}}(x) \ni 0.$$

(iii) Assume further that $\mathscr{D}$ is of the form

$$\mathscr{D} = \{x \in Q : f_1(x) \leq 0, \ldots, f_m(x) \leq 0\},$$

where Q is closed, convex and nonempty and $f_1, \ldots, f_m : \mathbb{R}^n \to \mathbb{R}$ are C^1 functions. For x in $\mathscr{D}$, set $I(x) = \{i : f_i(x) = 0\}$ and assume that there exists $y \in Q$ such that,

$$(\text{Robinson QC}) \qquad \langle \nabla f_i(x), y - x \rangle < 0, \forall i \in I(x).$$

Then $\mathscr{D}$ is regular,

$$N_{\mathscr{D}}(x) = \left\{ \sum_{i \in I(x)} \lambda_i \nabla f_i(x) : \lambda_i \geq 0, i \in I(x) \right\} + N_Q(x),$$

and critical points for $(\mathscr{P})$ are exactly KKT points of $(\mathscr{P})$.

Proof. (i) is Rockafellar and Wets (1998) [42, Theorem 10.1]). (ii) is obtained by using the sum rule Rockafellar and Wets (1998) [42, Corollary 10.9]. For (iii), regularity and normal cone expression follow from Rockafellar and Wets (1998) [42, Theorem 6.14] (Robinson condition appears there in a generalized form). $\qquad\square$

Recall that a convex cone $L \subset \mathbb{R}_+^n$ is a nonempty convex set such that $\mathbb{R}_+ L \subset L$. Being given a subset S of $\mathbb{R}^n$, the *conic hull of S*, denoted $\operatorname{cone} S$ is defined as the smallest convex cone containing S. Since a cone always contains 0, $\operatorname{cone} \emptyset = \{0\}$.

Proposition 4.2 (Subdifferential of set-parameterized indicator functions) *Let n_1, n_2, m be positive integers and $g_1, \ldots, g_m : \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \to \mathbb{R}$ continuously differentiable functions. Set*

$$C(x) = \{y \in \mathbb{R}^{n_2} : g_i(x, y) \leq 0, \ \forall i = 1, \ldots, m\} \subset \mathbb{R}^{n_2}, \quad \forall x \in \mathbb{R}^{n_1},$$

and for any $y \in C(x)$ put $I(x, y) = \{i = 1, \ldots, m : g_i(x, y) = 0\}$, the set $I(x, y)$ is empty otherwise. Assume that the following parametric Mangasarian-Fromovitz qualification condition holds:

$$\forall (x, y) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}, \exists d \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}, \langle \nabla g_i(x, y), d \rangle < 0, \forall i \in I(x, y).$$

Consider the real extended-valued function $H : \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \to (-\infty, +\infty]$ defined through

$$H(x, y) = \begin{cases} i_{C(x)}(y) \text{ whenever } C(x) \text{ is nonempty} \\ \\ +\infty \text{ otherwise.} \end{cases}$$

Then the subdifferential of H is given by

$$(25) \qquad \partial H(x, y) = \operatorname{cone}\{\nabla g_i(x, y) : i \in I(x, y)\}.$$

Proof. For any (x, y) in $\operatorname{dom} H$, set $G(x, y) = (g_1(x, y), \ldots, g_m(x, y))$. Then $H(x, y) = i_{\mathbb{R}_-^m}(G(x, y))$ and H is the indicator of the set $C = \{(x, y) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} : G(x, y) \in \mathbb{R}_-^m\}$. We will justify the application of the last equality of Rockafellar and Wets (1998) [42, Theorem 6.14]. We fix (x, y) in $\operatorname{dom} H$ such that $G(x, y) \leq 0$ and we set $I = I(x, y)$. The abstract qualification constraint required in Rockafellar and Wets (1998) [42, Theorem 6.14] is equivalent to $\lambda_i \geq 0, \sum_{i \in I} \lambda_i \nabla g_i(x, y) = 0 \Rightarrow \lambda_i = 0$. Using Hahn-Banach separation theorem this appears

to be equivalent to the parametric MFQC condition. The set $\mathbb{R}^m_-$ is regular and we can apply Rockafellar and Wets (1998) [42, Theorem 6.14] which assesses that C is regular at (x, y). In this case, the normal cone of C and the sub-differential of H coincide and are given by

$$\partial H(x,y) = \left\{ \sum_{i=1}^m \lambda_i \nabla g_i(x,y) : \lambda \in N_{\mathbb{R}^m_-}(G(x,y)) \right\} = \left\{ \sum_{i \in I} \lambda_i \nabla g_i(x,y) : \lambda_i \geq 0 \right\}.$$

$\square$

4.1.2 Multivalued mappings

A *multivalued mapping* $F : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ maps a point x in $\mathbb{R}^n$ to a subset $F(x)$ of $\mathbb{R}^m$. The set

$$\operatorname{dom} F := \left\{ x \in \mathbb{R}^n : F(x) \neq \emptyset \right\}$$

is called the *domain* of F. For instance the subdifferential of a lsc function defines a multivalued mapping $\partial f : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$.

Several regularity properties for such mappings are useful in optimization; we focus here on one of the most natural concept: set-valued continuity (see e.g. Dontchev and Rockafellar (2009) [21, Section 3.B, p. 142]).

Definition 2 (Continuity of point-to-set mappings) Let $F : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ and x in $\operatorname{dom} F$.
(i) F is called *outer semi-continuous at x*, if for each sequence $x_j \to x$ and each sequence $y_j \to y$ with $y_j \in F(x_j)$, we have $y \in F(x)$.
(ii) F is called *inner semi-continuous at x*, if for all $x_j \to x$ and $y \in F(x)$ there exists a sequence $y_j \in F(x_j)$, after a given term, such that $y_j \to y$.
(iii) F is called *continuous at x* if it is both outer and inner semi-continuous.

4.1.3 The KL property and some facts from real semi-algebraic geometry

KL is a shorthand here for Kurdyka-Łojasiewicz. This property constitutes a crucial tool in our convergence analysis. We consider the nonsmooth version of this property which is given in Bolte et al. (2007) [11, Theorem 11] – precisions regarding concavity of the desingularizing function are given in Attouch et al. (2010) [3, Theorem 14].

Being given real numbers a and b, we set $[a \leq g \leq b] = \{x \in \mathbb{R}^n : a \leq g(x) \leq b\}$. The sets $[a < g < b]$, $[g < a]$... are defined similarly.

For $\alpha \in (0, +\infty]$, we denote by Φ_α the class of functions $\varphi : [0, \alpha) \to \mathbb{R}$ that satisfy the following conditions

(a) $\varphi(0) = 0$;

(b) φ is positive, concave and continuous;

(c) φ is continuously differentiable on $(0, \alpha)$, with $\varphi' > 0$.

Definition 3 (KL property) Let g be a proper lower semi-continuous function from $\mathbb{R}^n$ to $(-\infty, +\infty]$.

(i) The function g is said to have the *Kurdyka-Łojaziewicz (KL) property* at $\bar{x} \in \operatorname{dom} \partial g$, if there exist $\alpha \in (0, +\infty]$, a neighborhood V of $\bar{x}$ and a function $\varphi \in \Phi_\alpha$ such that

(26) $$\varphi'(g(x) - g(\bar{x})) \operatorname{dist}(0, \partial g(x)) \geq 1$$

for all $x \in V \cap [g(\bar{x}) < g(x) < \alpha]$.

(ii) The function g is said to be a *KL function* if it has the KL property at each point of $\operatorname{dom} \partial g$.

KL property basically asserts that a function can be made sharp by a reparameterization of its values. This appears clearly when g is differentiable and $g(\bar{x}) = 0$, since in this case (26) writes:

$$\|\nabla(\varphi \circ g)(x)\| \geq 1, \quad \forall x \in V \cap [0 < g(x) < \alpha].$$

The function φ used in this parameterization is called a *desingularizing function*. As we shall see such functions are ubiquitous in practice, see Attouch et al. (2010) [3], Attouch et al. (2013) [4].

When φ is of the form $\varphi(s) = cs^{1-\theta}$ with $c > 0$ and $\theta \in [0, 1)$, the number θ is called *a Łojasiewicz exponent*.

Definition 4 (Semi-algebraic sets and functions)

(i) A set $A \subset \mathbb{R}^n$ is said to be semi-algebraic if there exist a finite number of real polynomial functions $g_{ij}, h_{ij} \colon \mathbb{R}^n \to \mathbb{R}$ such that

$$A = \bigcup_{i=1}^{p} \bigcap_{j=1}^{q} \{y \in \mathbb{R}^n : g_{ij}(y) = 0, h_{ij}(y) > 0\}$$

(ii) A mapping $G : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ is a said to be semi-algebraic if its graph

$$\operatorname{graph} G = \Big\{(x, y) \in \mathbb{R}^{n+m} : y \in G(x)\Big\}$$

is a semi-algebraic subset of $\mathbb{R}^{n+m}$.

Similarly, a real extended-valued function $g : \mathbb{R}^n \to (-\infty, +\infty]$ is semi-algebraic if its graph $\big\{(x, y) \in \mathbb{R}^{n+1} : y = g(x)\big\}$ is semi-algebraic.

For this class of functions, we have the following result which provides a vast field of applications for our method – see also Section 5.

Theorem 4.3 (Bolte et al. (2007) [11], Bolte et al. (2007) [10]) *Let g be a proper lower semi-continuous function from $\mathbb{R}^n$ to $(-\infty, +\infty]$. If g is semi-algebraic, then g is a KL function.*

4.2 An auxiliary Lyapunov function: the value function

Basic estimations

Lemma 4.4 (Descent lemma) *Let $g : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function with L-Lipschitz continuous gradient. Then for all x and y in $\mathbb{R}^n$,*

$$|g(y) - g(x) - \langle \nabla g(x), y - x \rangle| \leq \frac{L}{2}\|x - y\|^2$$

The proof is elementary, see e.g. Nesterov (2004) [37, Lemma 1.2.3].

Lemma 4.5 (Quadratic growth of the local models)

$$\text{Fix } x \text{ in } \mathscr{D}. \text{ Then: } \quad h(x, y) - h(x, p(x)) \geq \frac{\mu}{2}\|y - p(x)\|^2, \quad \forall y \in D(x).$$

Proof. Since $y \to h(x, y)$ is μ strongly convex, the function

$$D(x) \ni y \to h(x, y) - \frac{\mu}{2}\|y - p(x)\|^2$$

is convex. Since $p(x)$ minimizes $y \to h(x, y)$ over $D(x)$, it also minimizes $y \to h(x, y) - \frac{\mu}{2}\|y - p(x)\|^2$. This follows by writing down the first order optimality condition for $p(x)$ (convex setting) and by using the convexity of $y \to h(x, y) - \frac{\mu}{2}\|y - p(x)\|^2$. The inequality follows readily. $\square$

Lemma 4.6 (Descent property) *For all x in $\mathscr{D}$,*

(27) $$f(x) = h(x, x) \geq h(x, p(x)) + \frac{\mu}{2}\|x - p(x)\|^2 \geq f(p(x)) + \frac{\mu}{2}\|x - p(x)\|^2.$$

Proof. From Lemma 4.5, we have for all x in $\mathscr{D}$ that

$$h(x, x) - h(x, p(x)) \geq \frac{\mu}{2}\|x - p(x)\|^2.$$

Therefore from the fact that $h(x, \cdot)$ is an upper model for f on $D(x)$, we infer

(28) $$f(x) = h(x, x) \geq h(x, p(x)) + \frac{\mu}{2}\|x - p(x)\|^2 \geq f(p(x)) + \frac{\mu}{2}\|x - p(x)\|^2.$$

$\square$

Iteration mapping

For any fixed x in $\mathscr{D}$, we recall that the iteration mapping is defined through:

$$p(x) := \operatorname{argmin}\{h(x, y) : y \in D(x)\}.$$

Lemma 4.7 (Continuity of the iteration mapping) *The iteration function p is continuous (on $\mathscr{D}$).*

Proof. Let x be a point in $\mathscr{D}$ and let $x_j \in \mathscr{D}$ be a sequence converging to x. Fix $y \in D(x)$ and let y_j be a sequence of points such that $y_j \in D(x_j)$ and $y_j \to y$ (use the inner semi-continuity of D). We prove first that $p(x_j)$ is bounded. To this end, observe that

(29) $$h(x_j, p(x_j)) + \frac{\mu}{2}\|y_j - p(x_j)\|^2 \leq h(x_j, y_j).$$

Recall that $h(x_j, p(x_j)) \geq f(p(x_j)) \geq \inf_{\mathscr{D}} f > -\infty$. Thus

$$\frac{\mu}{2}\|y_j - p(x_j)\|^2 \leq h(x_j, y_j) - \inf_{\mathscr{D}} f$$

and $p(x_j)$ is bounded by continuity of h. Denote by π a cluster point of $p(x_j)$. Observe that since $p(x_j) \in D(x_j)$, the outer semi-continuity of D implies that $\pi \in D(x)$. Passing to the limit in (29) above, one obtains

$$h(x, \pi) + \frac{\mu}{2}\|y - \pi\|^2 \leq h(x, y).$$

Since this holds for arbitrary y in $D(x)$, we have established that π minimizes $h(x, \cdot)$ on $D(x)$, that is $\pi = p(x)$. This proves that p is continuous. $\square$

Lemma 4.8 (Fixed points of the iteration mapping) *Let x be in $\mathscr{D}$ such that $p(x) = x$. Then x is critical for $(\mathscr{P})$ that is*

$$\partial f(x) + N_{\mathscr{D}}(x) \ni 0.$$

Proof. Using the optimality condition and the sum rule for subdifferential of convex functions one has

$$(30) \qquad \partial_y h(x, p(x)) + N_{D(x)} \ni 0.$$

By assumption (20) (ii), we have $\partial_y h(x, p(x)) = \partial_y h(x, x) \subset \partial f(x)$. On the other hand $D(x) \subset \mathscr{D}$ and $N_{D(x)}(x) \subset N_{\mathscr{D}}(x)$, by (19). Using these inclusions in (30) yields the result. $\qquad\square$

Value function

The value function is defined through

$$F = \mathrm{val} : \begin{cases} \mathbb{R}^n & \longrightarrow & (-\infty, +\infty] \\[2mm] x & \longrightarrow & h(x, p(x)). \end{cases}$$

Being given x in $\mathbb{R}^n$, and a value $f(x) = h(x, x)$, it measures the progress made not on the objective f, but on the value of the model.

Tarski-Seidenberg theorem asserts a linear projection of a semi-algebraic set is semi-algebraic set. This implies that the class of semi-algebraic functions is closed under many operations, such as addition, multiplication, composition, inverse, projection and partial minimization (see Bochnak et al. (2003) [9] and Attouch et al. (2013) [4, Theorem 2.2] for an illustration in optimization). Applying standard techniques of semi-algebraic geometry, we obtain therefore:

Lemma 4.9 (Semi-algebraicity of the value function) *If f, h, $\mathscr{D}$ are semi-algebraic then F is semi-algebraic.*

Let $\mathscr{D}'$ denote the domain where F is differentiable. By standard stratification results, this set contains a dense *finite* union of open sets (a family of strata of maximal dimension, see e.g. Van Den Dries and Milller (1996) [22, 4.8], see also Ioffe (2009) [31, Theorem 2.3] for a self contained exposition). Thus we have:

$$(31) \qquad \mathrm{int}\, \mathscr{D}' \text{ is dense in } \mathscr{D}.$$

We now have the following estimates

Lemma 4.10 (Subgradient bounds) *Let $C \subset \mathscr{D}$ be a bounded set. Then there exists $K \geq 0$ such that $\forall x \in \mathscr{D}' \cap C$*

$$(32) \qquad ||\nabla F(x)|| \leq K ||p(x) - x||.$$

As a consequence

$$(33) \qquad \mathrm{dist}\,(0, \partial F(x)) \leq K ||p(x) - x||, \, \forall x \in \mathscr{D} \cap C.$$

Proof. Fix $\bar{x}$ in int $\mathscr{D}' \cap C$ and let δ and μ be in $\mathbb{R}^n$. Then

$$
\begin{aligned}
\hat{h}(\bar{x} + \delta, p(\bar{x}) + \mu)) &= h(\bar{x} + \delta, p(\bar{x}) + \mu) + i_{D(\bar{x}+\delta)}(p(\bar{x}) + \mu) \\
&\geq h(\bar{x} + \delta, p(\bar{x} + \delta)) \\
&= h(\bar{x}, p(\bar{x})) + \langle \nabla F(\bar{x}), \delta \rangle + o(||\delta||) \\
&= \hat{h}(\bar{x}, p(\bar{x})) + \langle \nabla F(\bar{x}), \delta \rangle + o(||\delta||).
\end{aligned}
$$

This implies that $(\nabla F(\bar{x}), 0) \in \partial \hat{h}(\bar{x}, p(\bar{x}))$. Since C is bounded, the qualification assumption of Section 3.2 yields (32).

To obtain (33), it suffices to use the definition of the subdifferential, the continuity of p and the fact that int $\mathscr{D}'$ is dense in $\mathscr{D}$. $\qquad\square$

We have the following property for the sequence generated by the method

Proposition 4.11 (Hidden gradient steps) *Let $\{x_k\}_{k=1,2,\dots}$ be the sequence defined through $x_{k+1} = p(x_k)$ with $x_0 \in \mathscr{D}$. Then x_k lies in $\mathscr{D}$ and*

$$
(34) \qquad F(x_k) + \frac{\mu}{2}||x_k - x_{k+1}||^2 \leq f(x_k) \leq F(x_{k-1}), \; \forall k \geq 1.
$$

Moreover, for all compact subset C of $\mathbb{R}^n$, there exists $K_2(C) >$ such that

$$
\mathrm{dist}\,(0, \partial F(x_k)) \leq K_2(C)||x_{k+1} - x_k||, \;\; \text{whenever } x_k \in C.
$$

Proof. The sequence x_k lies in $\mathscr{D}$ since $p(x_k) \in D(x_k) \subset \mathscr{D}$ by (19). We only need to prove the second item (34) since the third one immediately follows from (33). Using inequality (27) and the fact that $h(x, y) \geq f(y)$ for all y in $D(x)$, we have

$$
\begin{aligned}
F(x) &= h(x, p(x)) \\
&\geq f(p(x)) \\
&= h(p(x), p(x)) \\
&\geq F(p(x)) + \frac{\mu}{2}||p(x) - p(p(x))||^2,
\end{aligned}
$$

therefore

$$
F(x_{k-1}) \geq f(x_k) \geq F(x_k) + \frac{\mu}{2}||x_k - x_{k+1}||^2
$$

which proves (34). $\qquad\square$

4.3 An abstract convergence result

The following abstract result is similar in spirit to Attouch et al. (2013) [4] and to recent variations Bolte et al. (2013) [12]. However, contrary to previous works it deals with conditions on a triplet $\{x_{k-1}, x_k, x_{k+1}\}$ and the subgradient estimate is of explicit type (like in Absil et al. (2005) [1] and even more closely Noll (2014) [39]).

Proposition 4.12 (Gradient sequences converge) *Let $\bar{G}: \mathbb{R}^n \to (-\infty, +\infty]$ be proper, lower semi-continuous, semi-algebraic function. Suppose that there exists a sequence $\{x_k\}_{k \in \mathbb{N}}$ such that,*

(a) *$\exists K_1 > 0$ such that $\bar{G}(x_k) + K_1||x_{k+1} - x_k||^2 \leq \bar{G}(x_{k-1})$*

23

(b) *For all compact subset C of $\mathbb{R}^n$, there exists $K_2(C) >$ such that*

$$\text{dist}\,(0, \partial \bar{G}(x_k)) \le K_2(C) \|x_{k+1} - x_k\|, \quad \text{whenever } x_k \in C.$$

(c) *If there exists $x_{k_j} \to \bar{x}$ as $j \to +\infty$, then $\bar{G}(x_{k_j}) \to \bar{G}(\bar{x})$.*

Then,

(I) *The following asymptotic alternative holds:*

   (i) *Either the sequence $\{x_k\}_{k\in\mathbb{N}}$ satisfies $\|x_k\| \to +\infty$,*

   (ii) *or it converges to a critical point of $\bar{G}$.*

As a consequence each bounded sequence is a converging sequence.

(II) *When x_k converges, we denote by x_∞ its limit and we take $\theta \in [0,1)$ a Łojasiewicz exponent of $\bar{G}$ at x_∞. Then,*

   (i) *If $\theta = 0$, the sequence $(x_k)_{k\in\mathbb{N}}$ converges in a finite number of steps,*

   (ii) *If $\theta \in (0, \frac{1}{2}]$ then there exist $c > 0$ and $q \in [0,1)$ such that*

$$\|x_k - x_\infty\| \le c\, q^k, \forall k \ge 1.$$

   (iii) *If $\theta \in (\frac{1}{2}, 1)$ then there exists $c > 0$ such that*

$$\|x_k - x_\infty\| \le c\, k^{-\frac{1-\theta}{2\theta-1}}, \forall k \ge 1.$$

Proof. We first deal with (I). Suppose that there exists $k_0 \ge 0$ such that $x_{k_0+1} = x_{k_0}$. This implies by (a), that $x_{k_0+l} = x_{k_0}$ for all $l > 0$. Thus the sequence converges and the second inequality (b) implies that we have a critical point of $\bar{G}$. We now suppose that $\|x_{k+1} - x_k\| > 0$ for all $k \ge 0$.

Definition of a KL neighborhood. Suppose that $(I)(i)$ does not hold. There exists therefore a cluster point $\bar{x}$ of x_k. Combining (a) and (c), we obtain that

$$(35) \qquad\qquad \lim_{k\to+\infty} \bar{G}(x_k) = \bar{G}(\bar{x}).$$

With no loss of generality, we assume that $\bar{G}(\bar{x}) = 0$. Since $\bar{G}$ is semi-algebraic, it is a KL function (Theorem 4.3). There exist $\delta > 0$, $\alpha > 0$ and $\varphi \in \varphi_\alpha$ such that

$$\varphi'(\bar{G}(x))\,\text{dist}\,(0, \partial \bar{G}(x)) \ge 1,$$

for all x such that $\|x - \bar{x}\| \le \delta$ and $x \in [0 < \bar{G} < \alpha]$. In view of assumption (b), set

$$K_2 = K_2\left(\bar{B}(\bar{x}, \delta)\right).$$

Estimates within the neighborhood. Let $r \ge s > 1$ be some integers and assume that the points $x_{s-1}, x_s \ldots, x_{r-1}$ belong to $B(\bar{x}, \delta)$ with $\bar{G}(x_{s-1}) < \alpha$. Take $k \in \{s, \ldots, r\}$, using (a), we have

$$\begin{aligned}
\bar{G}(x_k) &\le \bar{G}(x_{k-1}) - K_1 \|x_{k+1} - x_k\|^2 \\
&= \bar{G}(x_{k-1}) - K_1 \frac{\|x_{k+1} - x_k\|^2}{\|x_k - x_{k-1}\|} \|x_k - x_{k-1}\| \\
&\le \bar{G}(x_{k-1}) - \frac{K_1}{K_2} \frac{\|x_{k+1} - x_k\|^2}{\|x_k - x_{k-1}\|} \text{dist}\,(0, \partial \bar{G}(x_{k-1})).
\end{aligned}$$

From the monotonicity and concavity of φ, we derive

$$\varphi \circ \bar{G}(x_k) \leq \varphi \circ \bar{G}(x_{k-1}) - \varphi' \circ \bar{G}(x_{k-1}) \frac{K_1}{K_2} \frac{||x_{k+1} - x_k||^2}{||x_k - x_{k-1}||} \mathrm{dist}\,(0, \partial \bar{G}(x_{k-1})),$$

thus by using KL property, for $k \in \{s, \ldots, r\}$,

$$(36) \qquad \varphi \circ \bar{G}(x_k) \leq \varphi \circ \bar{G}(x_{k-1}) - \frac{K_1}{K_2} \frac{||x_{k+1} - x_k||^2}{||x_k - x_{k-1}||}.$$

We now use the following simple fact: for $a > 0$ and $b \in \mathbb{R}$,

$$2(a - b) - \frac{a^2 - b^2}{a} = \frac{a^2 - 2ab + b^2}{a} = \frac{(a - b)^2}{a} \geq 0,$$

thus for $a > 0$ and $b \in \mathbb{R}$

$$(37) \qquad 2(a - b) \geq \frac{a^2 - b^2}{a}.$$

We have therefore, for k in $\{s, \ldots, r\}$,

$$
\begin{aligned}
||x_k - x_{k-1}|| &= \frac{||x_k - x_{k-1}||^2}{||x_k - x_{k-1}||} \\
&= \frac{||x_{k+1} - x_k||^2}{||x_k - x_{k-1}||} + \frac{||x_k - x_{k-1}||^2 - ||x_{k+1} - x_k||^2}{||x_k - x_{k-1}||} \\
&\overset{(37)}{\leq} \frac{||x_{k+1} - x_k||^2}{||x_k - x_{k-1}||} + 2(||x_k - x_{k-1}|| - ||x_{k+1} - x_k||) \\
&\overset{(36)}{\leq} \frac{K_2}{K_1}\left(\varphi \circ \bar{G}(x_{k-1}) - \varphi \circ \bar{G}(x_k)\right) + 2(||x_k - x_{k-1}|| - ||x_{k+1} - x_k||).
\end{aligned}
$$

Hence by summation

$$(38) \qquad \sum_{k=s}^{r} ||x_k - x_{k-1}|| \leq \frac{K_2}{K_1}\left(\varphi \circ \bar{G}(x_{s-1}) - \varphi \circ \bar{G}(x_r)\right) + 2\left(||x_s - x_{s-1}|| - ||x_{r+1} - x_r||\right).$$

The sequence remains in the neighborhood and converges. Assume that for N sufficiently large one has

$$(39) \qquad ||x_N - \bar{x}|| \leq \frac{\delta}{4},$$

$$(40) \qquad \frac{K_2}{K_1}\left(\varphi \circ \bar{G}\right)(x_N) \leq \frac{\delta}{4},$$

$$(41) \qquad \sqrt{K_1^{-1} \bar{G}(x_{N-1})} < \min\left(\frac{\delta}{4}, \sqrt{K_1^{-1}\alpha}\right).$$

One can require (40) and (41) because φ is continuous and $\bar{G}(x_k) \downarrow 0$. By (a), one has

$$(42) \qquad ||x_{N+1} - x_N|| \leq \sqrt{K_1^{-1} \bar{G}(x_{N-1})} < \frac{\delta}{4}.$$

Let us prove that $x_r \in B(\bar{x}, \delta)$ for $r \geq N + 1$. We proceed by induction on r. By (39), $x_N \in B(\bar{x}, \delta)$ thus the induction assumption is valid for $r = N + 1$. Since by (41) one has

$\bar{G}(x_N) < \alpha$, estimation (38) can be applied with $s = N + 1$. Suppose that $r \geq N + 1$ and $x_N, \ldots, x_{r-1} \in B(\bar{x}, \delta)$, then we have the following

$$
\begin{aligned}
\|x_r - \bar{x}\| \quad &\leq \quad \|x_r - x_N\| + \|x_N - \bar{x}\| \\
&\overset{(39)}{\leq} \quad \sum_{k=N+1}^{r} \|x_k - x_{k-1}\| + \frac{\delta}{4} \\
&\overset{(38)}{\leq} \quad \frac{K_2}{K_1} \varphi \circ \bar{G}(x_N) + 2\|x_{N+1} - x_N\| + \frac{\delta}{4} \\
&\overset{(40),(42)}{<} \quad \delta.
\end{aligned}
$$

Hence $x_N, \ldots, x_r \in B(\bar{x}, \delta)$ and the induction proof is complete. Therefore, $x_r \in B(\bar{x}, \delta)$ for any $r \geq N$. Using (38) again, we obtain that the series $\sum \|x_{k+1} - x_k\|$ converges, hence x_k also converges by Cauchy criterion.

The second part *(II)* is proved as in Attouch and Bolte (2009) [2, Theorem 2]. First, because of the semi-algebraicity of the data, φ can be chosen of the form $\varphi(s) = c.s^{1-\theta}$ with $c > 0$ and $\theta \in [0, 1)$. In this case, (38) combined with KL property and (b) yields a similar result as formula (11) in Attouch and Bolte (2009) [2], which therefore leads to the same estimates. □

Remark 8 (1) **(Coercivity implies convergence)** Quite often in practice $\bar{G}$ has bounded level sets. In that case the alternative reduces to convergence because of assumption (a).
(2) **(Assumption (c))** Assumption (c) is very often satisfied in practice: for instance when $\bar{G}$ has a closed domain and is continuous on its domain or when $\bar{G}$ is locally convex up to a square (locally semi-convex).

At last, Propositions 4.11 and 4.12 can be combined to prove Theorem 3.1. First, we can consider the restriction of F to the closed semi-algebraic set $\mathscr{D}$, since the sequence of Proposition 4.11 stays in $\mathscr{D}$. F is semi-algebraic by Lemma 4.9 and F is continous on $\mathscr{D}$ by continuity of h and p. Propositions 4.11 shows that F satisfies assumptions (a) and (b) of Proposition 4.12, and assumption (c) follows by the previous remark. Hence Proposition 4.12 applies to F and the result follows.

5 Beyond semi-algebraicity: MMP and NLP with real analytic data

Many concrete and essential problems involve objectives and constraints defined through real analytic functions –which are not in general semi-algebraic functions– and this raises the question of the actual scope of the results described previously. We would thus like to address here the following question: *Can we deal with nonlinear programming problems involving real analytic data?*

A convenient framework to capture most of what is needed to handle real analytic problems, and of an even larger class of problems, is the use of o-minimal structures. These are classes of sets and functions whose stability properties and topological behavior are the same as those encountered in the semi-algebraic world.

We give below some elements necessary to understand what is at stake and how our results enter this larger framework.

Definition 5 (O-minimal structures, see Van Den Dries and Miller (1996) [22]**)** An *o-minimal* structure on $(\mathbb{R}, +, .)$ is a sequence of families $\mathscr{O} = (\mathscr{O}_p)_{p \in \mathbb{N}}$ with $\mathscr{O}_p \subset \mathscr{P}(\mathbb{R}^p)$ (the collection of subsets of $\mathbb{R}^p$), such that for each $p \in \mathbb{N}$:

(i) Each $\mathscr{O}_p$ contains $\mathbb{R}^p$ and is stable by finite union, finite intersection and complementation;

(ii) if A belongs to $\mathscr{O}_p$, then $A \times \mathbb{R}$ and $\mathbb{R} \times A$ belong to $\mathscr{O}_{p+1}$;

(iii) if $\Pi : \mathbb{R}^{p+1} \to \mathbb{R}^p$ is the canonical projection onto $\mathbb{R}^p$ then for any A in $\mathscr{O}_{p+1}$, the set $\Pi(A)$ belongs to $\mathscr{O}_p$;

(iv) $\mathscr{O}_p$ contains the family of real algebraic subsets of $\mathbb{R}^p$, that is, every set of the form

$$\{x \in \mathbb{R}^p : g(x) = 0\},$$

where $g : \mathbb{R}^p \to \mathbb{R}$ is a real polynomial function ;

(v) the elements of $\mathscr{O}_1$ are exactly the finite unions of intervals.

Examples of such structures are given in Van Den Dries and Milller (1996) [22]. We focus here on the class of globally subanalytic sets which allows us to deal with real analytic NLP in a simple manner. Thanks to Gabrielov's theorem of the complement, the *class of globally subanalytic subsets* can be seen as the smallest o-minimal structure containing semi-algebraic sets and the graphs of all real analytic functions of the form: $f : [-1, 1]^n \to \mathbb{R}$, see e.g. Van Den Dries and Milller (1996) [22]. As a consequence any real analytic function defined on an open neighborhood of a box is globally subanalytic.

Note that a real analytic function might not be globally subanalytic (take sin whose graph intersects the x-axis infinitely many times, and thus (iv) is not fulfilled for (graph sin) $\cap$ (Ox), however it follows from the definition that the restriction of a real analytic function to a compact set included in its (open) domain is globally subanalytic.

We come now to the results we need for our purpose. For any o-minimal structure, one can assert that:

(a) The KL property holds – i.e. one can replace the term "semi-algebraic" by "definable" in Theorem 4.3, see Bolte et al. (2007) [11].

(b) The stratification properties (31) used to derive the abstract qualification condition hold, see Van Den Dries and Milller (1996) [22].

> *As a consequence, and at the exception of convergence rates, all the results announced in the paper are actually valid for an arbitrary o-minimal structure instead of the specific choice of the class of semi-algebraic sets.*

To deal with the case of real analytic problems, we combine the use of compactness and of the properties of globally subanalytic sets. This leads to the following results.

Theorem 5.1 (Convergence of ESQM/Sℓ^1QP for analytic functions) *Assume that the following properties hold*

– *The functions $f, f_1, \ldots, f_m$ are real analytic and Q is globally subanalytic ([8]),*

---

[8]Q subanalytic is actually enough, see Van Den Dries and Milller (1996) [22]

– *Lipschitz continuity assumptions* (10),

– *steplength condition* (11),

– *qualification assumptions* (14),

– *boundedness assumptions* (12), (13).

Then,

(i) *the sequence $\{x_k\}_{k\in\mathbb{N}}$ generated by* (ESQM) *(resp.* $S\ell^1QP$*) converges to a feasible point x_∞ satisfying the KKT conditions for the nonlinear programming problem $(\mathscr{P}_{NLP})$.*

(ii) *Either convergence occurs in a finite number of steps or the rate is of the form:*

(a) $\|x_k - x_\infty\| = O(q^k)$, *with $q \in (0,1)$,*

(b) $\|x_k - x_\infty\| = O\left(\frac{1}{k^\gamma}\right)$, *with $\gamma > 0$.*

Theorem 5.2 (Convergence of the moving balls method) *Recall that $Q = \mathbb{R}^n$ and assume that the following properties hold*

– *The functions $f, f_1, \ldots, f_m$ are real analytic,*

– *Lipschitz continuity assumptions* (6),

– *Mangasarian-Fromovitz qualification condition* (7),

– *boundedness condition* (8),

– *feasibility of the starting point $x_0 \in \mathscr{F}$.*

Then,

(i) *The sequence $\{x_k\}_{k\in\mathbb{N}}$ defined by the moving balls method converges to a feasible point x_∞ satisfying the KKT conditions for the nonlinear programming problem $(\mathscr{P}_{NLP})$.*

(ii) *Either convergence occurs in a finite number of steps or the rate is of the form:*

(a) $\|x_k - x_\infty\| = O(q^k)$, *with $q \in (0,1)$,*

(b) $\|x_k - x_\infty\| = O\left(\frac{1}{k^\gamma}\right)$, *with $\gamma > 0$.*

Proof. The "proofs" of both theorems are the same. We observe first that in both cases the sequences are bounded. Let thus $a > 0$ be such that $x_k \in [-a, a]^n$ for all nonnegative k. Now the initial problem can be artificially transformed to a definable problem by including the constraints $x_i \leq a$ and $-x_i \leq a$ without inducing any change for the sequences. This imposes restrictions to real analytic function making them globally subanalytic hence definable.

The fact that the rate of convergence are of the same nature is well known and comes from the fact that Puiseux Lemma holds for subanalytic functions (see Van Den Dries and Milller (1996) [22, 5.2] and the discussion in Kurdyka (1998) [32, Theorem LI]). $\qquad\square$

6 Appendix: convergence proofs for SQP methods

6.1 Convergence of the moving balls method

The local model of f is given at a feasible x by

$$h_{\mathrm{MB}}(x,y) = f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2, \ x, y \in \mathbb{R}^n,$$

while the constraint approximation is given by

$$D(x) = \left\{ y \in \mathbb{R}^n : f_i(x) + \langle \nabla f_i(x), y - x \rangle + \frac{L_i}{2}\|y - x\|^2 \leq 0 \right\}.$$

The fact that for all x in $\mathscr{F}$, $D(x) \subset \mathscr{F}$ is ensured by Lemma 4.4. As an intersection of a finite number of balls containing x the set $D(x)$ is a nonempty compact (hence closed) convex set. The proof of the continuity of D is as in Auslender et al. (2010) [6, Proposition A1 & A2].

Let us also recall that Mangasarian-Fromovitz condition implies that

Lemma 6.1 (Slater condition for $\mathscr{P}(x)$) *Auslender et al. (2010) [6, Proposition 2.1] The set $D(x)$ satisfies the Slater condition for each x in $\mathscr{F}$.*

Corollary 6.2 *For a given feasible x, set*

$$g_i(y) = f(x) + \langle \nabla f_i(x), y - x \rangle + \frac{L_i}{2}\|y - x\|^2, \ y \in \mathbb{R}^n, \ i = 1, \dots, m.$$

Suppose that (x, y) is such that $g_i(y) \leq 0$, $i = 1 \dots m$. Then the only solution $u = (u_1, \dots, u_m)$ to

$$\sum_{i=1}^{m} u_i \, \nabla g_i(y) = 0, \ u_i \geq 0 \ and \ u_i \, g_i(y) = 0 \ for \ 1 \leq i \leq m$$

is the trivial solution $u = 0$.

Proof. When $J = \{i = 1, \dots, m : \ g_i(y) = 0\}$ is empty, the result is trivial. Suppose that J is not empty and argue by contradiction. This means that 0 is in the convex envelope of $\{\nabla g_j(y), j \in J\}$ and thus one cannot have Mangasarian-Fromovitz condition for $\mathscr{P}(x)$ at y (recall that $\mathscr{P}(x)$ involves constraints of the form $g_i \leq 0$). This contradicts the fact that Slater condition holds for $\mathscr{P}(x)$, since Slater condition classically implies Mangasarian-Fromovitz condition at each point. $\qquad\square$

Corollary 6.3 (Lagrange multipliers of the subproblems are bounded) *For each x in $\mathscr{F}$, we denote by $\Lambda(x) \subset \mathbb{R}_+^m$ the set of Lagrange multipliers associated to $\mathscr{P}(x)$. Then for any compact subset B of $\mathscr{F}$,*

$$(43) \qquad \sup\left\{ \max_{i=1,\dots,m} \lambda_i(x) : (\lambda_1(x), \dots, \lambda_m(x)) \in \Lambda(x), x \in B \right\} < +\infty.$$

Proof. Observe that, at this stage, we know that p is continuous. We argue by contradiction and assume that the supremum is not finite. One can thus assume, by compactness, that there

exists a point $\bar{x}$ in $\mathscr{F}$ together with a sequence $z_j \to \bar{x}$ such that at least one of the $\lambda_j(z_j)$ tends to infinity. Writing down the optimality conditions, one derives Lagrange relations

$$\frac{1}{\sum_{i=1}^m \lambda_i(x)} \left(\nabla f(z_j) + L(p(z_j) - z_j) \right) + \sum_{i=1}^m \frac{\lambda_i(z_j)}{\sum_{i=1}^m \lambda_i(z_j)} \left(\nabla f_i(z_j) + L_i(p(z_j) - z_j) \right) = 0$$

and complementary slackness

$$\lambda_j(z_j) \left(f_i(z_j) + \langle \nabla f_i(z_j), p(z_j) - z_j \rangle + \frac{L_i}{2} \| p(z_j) - z_j \|^2 \right) = 0.$$

Up to an extraction one can assume that the sequence of p-uplet $\left\{ \left(\frac{\lambda_i(z_j)}{\sum_{i=1}^m \lambda_i(z_j)} \right)_{i=1,\dots,m} \right\}_j$ converges to $u = (u_1, \dots, u_m)$ in the unit simplex and that, for all i, the limit of $\lambda_i(z_j)$ exists and is either finite or infinite. Passing to the limit, one obtains that

$$(44) \qquad \sum_{i=1}^m u_i \, \nabla g_i(p(\bar{x})) = 0, \ g_i(p(\bar{x})) \le 0 \text{ and } u_i \, g_i(p(\bar{x})) = 0 \text{ for } 1 \le i \le m,$$

where $g_1, \dots, g_m$ are as defined in Lemma 6.2. But Lemma 6.2 asserts that the unique solution to such a set of equations is $u = 0$, which contradicts the fact that u is a point of the unit simplex. $\qquad \square$

Recall that for all $x, y \in \mathbb{R}^n$, we set $\hat{h}_{\mathrm{MB}}(x, y) = h_{\mathrm{MB}}(x, y) + i_{D(x)}(y)$. Fix $x \in \mathscr{F}$ and y in $D(x)$ set

$$I(x, y) = \left\{ i \in \{1, \dots, m\} : \ f_i(x) + \langle \nabla f_i(x), y - x \rangle + \frac{L_i}{2} \| y - x \|^2 = 0 \right\}.$$

Combining Proposition 4.2 with Corollary 6.2, one has that the subdifferential of $\hat{h}_{\mathrm{MB}}$ is given by

(45)

$$\partial \hat{h}_{\mathrm{MB}}(x, y)$$
$$= \begin{pmatrix} L(x - y) - \nabla^2 f(x)(x - y) \\ \nabla f(x) + L(y - x) \end{pmatrix} + \mathrm{cone} \left\{ \begin{pmatrix} L_i(x - y) - \nabla^2 f_i(x)(x - y) \\ \nabla f_i(x) + L_i(y - x) \end{pmatrix}, \ i \in I(x, y) \right\}.$$

The only assumption of Section 3.2 that needs to remain established is the qualification assumption (21).

Lemma 6.4 *The qualification assumption* (21) *holds for* h_{MB}.

Proof. $(v, 0) \in \partial \hat{h}_{\mathrm{MB}}(x, y)$ implies that

$$y = \mathrm{argmin}_z \{ \hat{h}_{\mathrm{MB}}(x, z) : z \in D(x) \},$$

in other words that $y = p(x)$. In view of (45), one has the existence of non-negative $\lambda_i(x), i = 1, \dots, m$ such that

$$\left(L.I_n - \nabla^2 f(x) + \sum_{i=1}^p \lambda_i(x) \left(L_i I_n - \nabla^2 f_i(x) \right) \right) (x - p(x)) \ = v,$$

$$(46) \qquad \nabla f(x) + L(y - x) + \sum_{i=1}^p \lambda_i(x) \left(\nabla f_i(x) + L_i(p(x) - x) \right) \ = 0.$$

The desired bound on v follows from the bound on the the Lagrange multipliers in (46) obtained in Corollary 6.3 $\qquad\square$

The assumptions for applying Theorem 3.1 are now gathered and Theorem 5.2 follows. The fact that we eventually obtain a KKT point is a consequence of the qualification condition and Proposition 4.1.

6.2 Convergence of Extended SQP and Sℓ^1QP

6.2.1 Sketch of proof of Theorem 2.2

The proof arguments are adapted from Auslender (2013) [5, Theorem 3.1, Proposition 3.2]. Set $l = \inf_Q f$ and recall that $l > -\infty$ by (13). Use first regularity assumptions (10), (11) in combination with Lemma 4.4, to derive that

$$\frac{1}{\beta_{k+1}}(f(x_{k+1}) - l) + \max_{i=0,\dots,m} f_i(x_{k+1}) \leq \frac{1}{\beta_k}(f(x_{k+1}) - l) + \max_{i=0,\dots,m} f_i(x_{k+1})$$
$$\leq \frac{1}{\beta_k}(h_{\beta_k}(x_{k+1}, x_k) - l)$$
$$\leq \frac{1}{\beta_k}\left(h_{\beta_k}(x_k, x_k) - l - \frac{\lambda + \beta_k \lambda'}{2}||x_{k+1} - x_k||^2\right)$$
$$\leq \frac{1}{\beta_k}(f(x_k) - l) + \max_{i=0,\dots,m} f_i(x_k) - \frac{\lambda'}{2}||x_{k+1} - x_k||^2,$$

where the first inequality follows from the monotonicity of β_k and the fact that $f(x_{k+1}) - l \geq 0$, the second inequality is due to the descent lemma while the third one is a consequence of the strong convexity of the local model.

The above implies that

$$\frac{1}{\beta_{k+1}}(f(x_{k+1}) - l) + \max_{i=0,\dots,m} f_i(x_{k+1}) \leq \frac{1}{\beta_0}(f(x_0) - l) + \max_{i=0,\dots,m} f_i(x_0),$$

thus $\max_{i=0,\dots,m} f_i(x_{k+1})$ is bounded for all k and the compactness assumption (12) ensures the boundedness of x_k.

Since $\frac{1}{\beta_k}(f(x_k) - l) + \max_{i=0,\dots,m} f_i(x_k) \geq 0$, a standard telescopic sum argument gives that $||x_{k+1} - x_k|| \to 0$. Set

$$J_k = \{i = 0, \dots, m : \text{test}_i(x_k, x_{k+1}) = \max_{j=0\dots m} \text{test}_j(x_k, x_{k+1})\},$$

and suppose that $\beta_k \to \infty$. This means that, up to a subsequence, there exists a nonempty set $I \subset \{1, \dots, m\}$ such that

(47) $$J_k = I$$
$$\forall t \in \mathbb{N}, \forall i \in I, f_i(x_k) + \langle \nabla f_i(x_k), x_{k+1} - x_k \rangle > 0.$$

Recall that the optimality condition for the local model minimization ensures that, for all k, there exists dual variables $u_i \geq 0, i \in J_k$ such that $\sum_{i \in J_k} u_i = 1$ and

(48) $$\left\langle \frac{1}{\beta_k}\left(\nabla f(x_k) + (\lambda + \lambda'\beta_k)(x_{k+1} - x_k)\right) + \sum_{i \in J_k} u_i \nabla f_i(x_k), z - x_{k+1}\right\rangle \geq 0,$$

for any $z \in Q$. Using the boundedness properties of x_k and u_i, up to another subsequence, we can pass to the limit in equations (47), (48) to find $\bar{x} \in Q$, $\bar{u}_i, i \in I$ such that

$$\bar{u}_i \geq 0$$

$$\sum_{i \in I} \bar{u}_i = 1$$

$$f_i(\bar{x}) \geq 0, i \in I$$

$$\left\langle \sum_{i \in I} \bar{u}_i \nabla f_i(\bar{x}), z - \bar{x} \right\rangle \geq 0, z \in Q,$$

which contradicts qualification assumption (14) ($\lim \|x_{k+1} - x_k\| = 0$). Therefore, for k sufficiently large, we have

$$\beta_k = \beta > 0,$$
$$f_i(x_k) + \langle \nabla f_i(x_k), x_{k+1} - x_k \rangle \leq 0,$$
$$0 \in J_k.$$

Given that $x_{k+1} - x_k \to 0$, any accumulation point is feasible. Furthermore, given an accumulation point $\bar{x}$, set $\bar{I} = \{0 \leq i \leq m, f_i(\bar{x}) = 0\}$. It must holds that (up to a subsequence) $J_k = \bar{I}$ for a sufficiently large k. The fact that $\bar{x}$ is a stationary point follows by passing to the limit in (48).

6.2.2 Proof of convergence of ESQM

As granted by Theorem 2.2, there exists k_0 such that $\beta_k = \beta$ for all integer $k \geq k_0$. Since our interest goes to the convergence of the sequence, we may assume with no loss of generality that β_k is equal to β. Therefore, we only need to consider the behavior of the sequence $\{x_k\}$ with respect to the function

$$\Psi_\beta(x) = f(x) + \beta \max_{i=0,\dots,m} (f_i(x)) + i_Q(x),$$

whose minimization defines problem $(\mathscr{P})$. Set $\mu = \lambda + \beta\lambda'$, the local model we shall use to study (ESQM) is given by

$$h_{\text{ESQM}}(x, y)$$
$$= f(x) + \langle \nabla f(x), y - x \rangle + \beta \max_{i=0,\dots,m} (f_i(x) + \langle \nabla f_i(x), y - x \rangle) + \frac{\mu}{2} \|y - x\|^2,$$

while the constraints inner approximations reduce to a constant multivalued mapping

$$D(x) = Q.$$

The assumptions (19) for D are obviously fulfilled. Let us establish (20). From assumptions (10), (11), we have for any x and y in Q,

$$f_i(y) \leq f_i(x) + \langle \nabla f_i(x), y - x \rangle + \frac{\lambda'}{2} \|x - y\|^2, \quad 0 \leq i \leq m,$$

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\lambda}{2} \|x - y\|^2.$$

Multiplying the first inequalities by β, taking the maximum over i and adding the the last inequality gives $\Psi_\beta(y) \leq h_{\text{ESQM}}(x, y)$ for any x and y in Q which yields (i), (iii) (20). Item (iv)

is obvious while item (iii) (20) follows from the formula of the subdifferential of a max function Rockafellar and Wets (1998) [42]. Assumption $(\mathscr{S})$ is also fulfilled (Q is convex, hence regular and so is Ψ_β).

Once more the only point that needs to be checked more carefully is the qualification assumption (21). For all $x, y \in Q$, let $I(x, y)$ be the active indices in the definition of $h_{\mathrm{ESQM}}(x, y)$. The subdifferential of $\hat{h}_{ESQM}$ is given by

$$\partial \hat{h}_{\mathrm{ESQM}}(x, y)$$
$$= \begin{pmatrix} \mu(x - y) - \nabla^2 f(x)(x - y) \\ \nabla f(x) + \mu(y - x) \end{pmatrix} + \beta \mathrm{co} \left\{ \begin{pmatrix} -\nabla^2 f_i(x)(x - y) \\ \nabla f_i(x) \end{pmatrix} : i \in I(x, y) \right\} + \begin{pmatrix} 0 \\ N_Q(y) \end{pmatrix},$$

where co denotes the convex hull. The result follows from the fact that the f_i is C^2 and that the hessian are bounded on bounded sets.

Theorem 3.1 applies and gives the desired conclusion. The fact that we eventually obtain a KKT point of $(\mathscr{P})$ is a consequence of Theorem 2.2.

6.2.3 Convergence of Sℓ^1QP

The proof is quasi-identical to that of ESQP, it is left to the reader.

Acknowledgments.

References

[1] P. A. Absil, R. Mahony, and B. Andrews, *Convergence of the iterates of descent methods for analytic cost functions*, SIAM Journal on Optimization **16** (2005), no. 2, 531–547.

[2] H. Attouch and J. Bolte, *On the convergence of the proximal algorithm for nonsmooth functions involving analytic features*, Mathematical Programming **116** (2009), no. 1-2, 5–16.

[3] H. Attouch, J. Bolte, P. Redont, and A. Soubeyran, *Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-Łojasiewicz inequality*, Mathematics of Operations Research **35** (2010), no. 2, 438–457.

[4] H. Attouch, J. Bolte, and B. F. Svaiter, *Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized Gauss–Seidel methods*, Mathematical Programming **137** (2013), no. 1-2, 91–129.

[5] A. Auslender, *An extended sequential quadratically constrained quadratic programming algorithm for nonlinear, semidefinite, and second-order cone programming*, Journal of Optimization Theory and Applications **156** (2013), no. 2, 183–212.

[6] A. Auslender, R. Shefi, and M. Teboulle, *A moving balls approximation method for a class of smooth constrained minimization problems*, SIAM Journal on Optimization **20** (2010), no. 6, 3232–3259.

[7] A. Beck and M. Teboulle, *Gradient-based algorithms with applications to signal recovery problems*, Convex Optimization in Signal Processing and Communications (D. Palomar and Y. Eldar, eds.), Cambribge University Press, Cambridge, 2010, pp. 42–88.

[8] D. Bertsekas, *Nonlinear programming*, Athena Scientific, Belmont, MA, 1995.

[9] J. Bochnak, M. Coste, and Roy M.-F., *Real algebraic geometry*, Springer, 1998.

[10] J. Bolte, A. Daniilidis, and A. S. Lewis, *The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems*, SIAM Journal on Optimization **17** (2007), no. 4, 1205–1223.

[11] J. Bolte, A. Daniilidis, A. S. Lewis, and M. Shiota, *Clarke subgradients of stratifiable functions*, SIAM Journal on Optimization **18** (2007), no. 2, 556–572.

[12] J. Bolte, S. Sabach, and M. Teboulle, *Proximal alternating linearized minimization for nonconvex and nonsmooth problems*, Mathematical Programming **146** (2013), no. 1-2, 459–494.

[13] J. F. Bonnans, J. Ch. Gilbert, C. Lemaréchal, and C. Sagastizábal, *Numerical optimization: theoretical and practical aspects*, Springer-Verlag, Berlin, Germany, 2003.

[14] J. V. Burke and S. P. Han, *A robust sequential quadratic programming method*, Mathematical Programming **43** (1989), no. 1-3, 277–303.

[15] R. H. Byrd, N. Gould, J. Nocedal, and R. Waltz, *On the convergence of successive linear-quadratic programming algorithms*, SIAM Journal on Optimization **16** (2005), no. 2, 471–489.

[16] C. Cartis, N. Gould, and P. Toint, *On the complexity of finding first-order critical points in constrained nonlinear optimization*, Mathematical Programming A **144** (2014), no. 1, 93–106.

[17] E. Chouzenoux, A. Jezierska, J. Pesquet, and H. Talbot, *A majorize-minimize subspace approach for $\ell_2 - \ell_0$ image regularization*, SIAM Journal on Imaging Sciences **6** (2013), no. 1, 563–591.

[18] P. L. Combettes and J.-C. Pesquet, *Proximal splitting methods in signal processing*, Fixed-Point Algorithms for Inverse Problems in Science and Engineering (H.H. Bauschke, R. S. Burachik, P. L. Combettes, V. Elser, R. Luke, and H. Wolkowicz, eds.), Springer Optimization and Its Applications, Springer New York, 2011, pp. 185–212.

[19] B. Cox, A. Juditsky, and A. Nemirovski, *Dual subgradient algorithms for large-scale nonsmooth learning problems*, Mathematical Programming **148** (2013), no. 1–2, 1–38.

[20] A. P. Dempster, N. M. Laird, and D. B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm*, Journal of the Royal Statistical Society, Series B **39** (1977), no. 1, 1–38.

[21] A. Dontchev and R. T. Rockafellar, *Implicit functions and solution mappings*, Springer Monograph Series, New York, 2009.

[22] L. van den Dries and C. Miller, *Geometric categories and o-minimal structures*, Duke Mathematical Journal **84** (1996), no. 2, 497–540.

[23] R. Fletcher, *An ℓ^1 penalty method for nonlinear constraints*, Numerical optimization (P. T. Boggs, R. H. Byrd, and Schnabel R. B., eds.), SIAM, 1985, pp. 26–40.

[24] ⎯⎯⎯, *Practical methods of optimization, 2nd Edition*, Wiley, 2000.

[25] R. Fletcher, N. Gould, S. Leyffer, P. Toint, and A. Wächter, *Global convergence of a trust-region SQP-filter algorithm for general nonlinear programming*, SIAM Journal on Optimization **13** (2002), no. 3, 635–659.

[26] M. Fukushima, Z. Luo, and P. Tseng, *A sequential quadratically constrained quadratic programming method for differentiable convex minimization*, SIAM Journal on Optimization **13** (2003), no. 4, 1098–1119.

[27] P. E. Gill, W. Murray, and M. Saunders, *SNOPT: An SQP algorithm for large-scale constrained optimization*, SIAM Review **47** (2005), no. 1, 99–131.

[28] P. E. Gill and E. Wong, *Sequential quadratic programming methods*, Mixed Integer Nonlinear Programming, The IMA volumes in mathematics and its applications (J. Lee and S. Leyffer, eds.), vol. 154, Springer New York, 2012, pp. 147–224.

[29] S.P. Han, *A globally convergent method for nonlinear programming*, Journal of Optimization Theory and Applications **22** (1977), no. 3, 297–309.

[30] W. L. Hare and A. S. Lewis, *Identifying active constraints via partial smoothness and prox-regularity*, Journal of Convex Analysis **11** (2004), no. 2, 251–266.

[31] A. Ioffe, *An invitation to tame optimization*, SIAM Journal on Optimization **19** (2009), no. 4, 1894–1917.

[32] K. Kurdyka, *On gradients of functions definable in o-minimal structures*, Annales de l'institut Fourier **48** (1998), no. 3, 769–783.

[33] A. S. Lewis, *Active sets, nonsmoothness, and sensitivity*, SIAM Journal on Optimization **13** (2002), no. 3, 702–725.

[34] S. Łojasiewicz, *Une propriété topologique des sous-ensembles analytiques réels*, Les Équations aux Dérivées Partielles, vol. 117, Éditions du Centre National de la Recherche Scientifique, 1963, pp. 87–89.

[35] J. Mairal, *Optimization with first-order surrogate functions*, ICML 2013-International Conference on Machine Learning, vol. 28, 2013, pp. 783–791.

[36] N. Maratos, *Exact penalty function algorithms for finite dimensional and control optimization problems*, Ph.D. thesis, Imperial College, University of London, London, U.K, 1978.

[37] Y. Nesterov, *Introductory lectures on convex optimization: A basic course*, vol. 87, Springer, 2004.

[38] J. Nocedal and S. Wright, *Numerical optimization*, Springer Series in Operations Research and Financial Engineering, Springer New York, 2006.

[39] D. Noll, *Convergence of non-smooth descent methods using the Kurdyka-Łojasiewicz inequality*, Journal of Optimization Theory and Applications **160** (2014), no. 2, 553–572.

[40] J. M. Ortega and W. C. Rheinboldt, *Iterative solution of nonlinear equations in several variables*, vol. 30, SIAM, 1970.

[41] M. J. D. Powell, *On search directions for minimization algorithms*, Mathematical Programming **4** (1973), 193–201.

[42] R. T. Rockafellar and R. Wets, *Variational analysis*, vol. 317, Springer, 1998.

[43] S. Shuzhong, *Ekeland's variational principle and the mountain pass lemma*, Acta Mathematica Sinica **1** (1985), no. 4, 348–355.

[44] M. V. Solodov, *On the sequential quadratically constrained quadratic programming methods*, Mathematics of Operations Research **29** (2004), no. 1, 64–79.

[45] _____, *Global convergence of an SQP method without boundedness assumptions on any of the iterative sequences*, Mathematical Programming **118** (2009), no. 1, 1–12.

[46] K. Svanberg, *A class of globally convergent optimization methods based on conservative convex separable approximations*, SIAM Journal on Optimization **12** (2002), no. 2, 555–573.

[47] A. Wilson, *Simplicial method for convex programming*, Ph.D. thesis, Harvard University, 1963.

[48] S. Wright, *Constraint identification and algorithm stabilization for degenerate nonlinear programs*, Mathematical Programming **95** (2003), no. 1, 137–160.