Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods

Hedy ATTOUCH* Jérôme BOLTE† Benar Fux SVAITER ‡

December, 15, 2010

Abstract In view of the minimization of a nonsmooth nonconvex function f, we prove an abstract convergence result for descent methods satisfying a sufficient-decrease assumption, and allowing a relative error tolerance. Our result guarantees the convergence of bounded sequences, under the assumption that the function f satisfies the Kurdyka-Łojasiewicz inequality. This assumption allows to cover a wide range of problems, including nonsmooth semi-algebraic (or more generally tame) minimization. The specialization of our result to different kinds of structured problems provides several new convergence results for inexact versions of the gradient method, the proximal method, the forward-backward splitting algorithm, the gradient projection and some proximal regularization of the Gauss-Seidel method in a nonconvex setting. Our results are illustrated through feasibility problems, or iterative thresholding procedures for compressive sensing.

2010 Mathematics Subject Classification: 34G25, 47J25, 47J30, 47J35, 49M15, 49M37, 65K15, 90C25, 90C53.

Keywords: Nonconvex nonsmooth optimization, semi-algebraic optimization, tame optimization, Kurdyka-Łojasiewicz inequality, descent methods, relative error, sufficient decrease, forward-backward splitting, alternating minimization, proximal algorithms, iterative thresholding, block-coordinate methods, o-minimal structures.

1 Introduction

Being given a proper lower semicontinuous function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, we consider *descent methods* that generate sequences $(x^k)_{k\in\mathbb{N}}$ complying with the following conditions:

*I3M UMR CNRS 5149, Université Montpellier II, Place Eugène Bataillon, 34095 Montpellier, France (attouch@math.univ-montp2.fr) Partially supported by ANR-08-BLAN-0294-03.

†TSE (GREMAQ, Université Toulouse I), Manufacture des Tabacs, 21 allée de Brienne, Toulouse, France (jerome.bolte@tse-eu.fr) Partially supported by ANR-08-BLAN-0294-03.

‡IMPA, Estrada Dona Castorina 110, 22460 - 320 Rio de Janeiro, Brazil (benar@impa.br) Partially supported by CNPq grants 480101/2008-6, 303583/2008-8, FAPERJ grant E-26/102.821/2008 and PRONEX-Optimization.

– for each $k \in \mathbb{N}$, $f(x^{k+1}) + a\|x^{k+1} - x^k\|^2 \leq f(x^k)$;

– for each $k \in \mathbb{N}$, there exists $w^{k+1} \in \partial f(x^{k+1})$ such that

$$\|w^{k+1}\| \leq b\|x^{k+1} - x^k\|;$$

where a, b are positive constants and $\partial f(x^{k+1})$ denotes the set of limiting subgradients of f at x^{k+1} (see Section 2.1 for a definition). The first condition is intended to model a descent property: since it involves a measure of the quality of the descent, we call it *a sufficient-decrease condition* (see [7] for an interpretation of this condition in decision sciences). The second condition originates from the well-known fact that most algorithms in optimization are generated by an infinite sequence of subproblems which involve exact or inexact minimization processes. This is the case of gradient methods, Newton's method, forward-backward algorithm, Gauss-Seidel method, proximal methods *etc.*... The second set of conditions precisely reflects *relative inexact optimality conditions* for such minimization subproblems.

When dealing with descent methods for convex functions, it became natural to expect that the algorithm will provide globally convergent sequences. The standard recipe to obtain the convergence is to prove that the sequence is (quasi-)Fejér monotone relative to the set of minimizers of f. This fact has also been used intensively in the study of algorithms for nonexpansive mappings (see *e.g.* [23]). When the functions under consideration are not convex (or quasiconvex), the monotonicity properties are in general "broken", and descent methods may provide sequences that exhibit highly oscillatory behaviors. Apparently this phenomenon was first observed by Curry (see [26]); in the framework of differential equations similar behaviors occur, in [27] a nonconverging bounded curve of a 2-dimensional gradient system of a C^∞ function is provided, this example was adapted in [1] to gradient methods.

In order to circumvent such behaviors, it seems necessary to work with functions that present a certain structure. This structure can be of an algebraic nature, e.g. quadratic functions, polynomial functions, real analytic functions, but it can also be captured by adequate analytic assumptions, e.g. metric regularity [2, 40, 41], cohypomonotonicity [48, 35], self-concordance [47], partial smoothness [39, 56]. In this paper, our central assumption for the study of such algorithms is that the function f satisfies the (nonsmooth) Kurdyka-Łojasiewicz inequality, which means, roughly speaking, that the functions under consideration *are sharp up to a reparametrization* (see Section 2.2). The reader is referred to [42, 43, 37] for the smooth cases, and to [15, 17] for nonsmooth inequalities. Kurdyka-Łojasiewicz inequalities have been successfully used to analyze various types of asymptotic behavior: gradient-like systems [15, 33, 34, 38], PDE [52, 21], gradient methods [1, 46], proximal methods [3], projection methods or alternating methods [5, 14].

In the context of optimization, the importance of Kurdyka-Łojasiewicz inequality is due to the fact that *many problems* involve functions satisfying such inequalities, and it is often *elementary* to check that such an inequality is satisfied; real semi-algebraic functions provide a very rich class of functions satisfying the Kurdyka-Łojasiewicz, see [5] for a thorough discussion on these aspects, and also Section 2.2 for a simple illustration.

Many other functions, that are met in real world problems, and which are not semi-algebraic, satisfy very often the Kurdyka-Łojasiewicz inequality. An important class is given by functions definable in an o-minimal structure. The monographs [25, 29] are good references on o-minimal structures; concerning Kurdyka-Łojasiewicz inequalities in this context the reader is referred to [37, 17]. Functions definable in o-minimal structures or functions

whose graphs are locally definable are often called *tame functions*. We do not give a precise definition of definability in this work, but the flexibility of this concept is briefly illustrated in Example 5.4(b). Functions that are not necessarily tame but that satisfy Łojasiewicz inequality are given in [5], basic assumptions involve metric-regularity and transversality (see also [40, 41] and Example 5.5).

From a technical viewpoint, our work blends the approach to nonconvex problems provided in [1, 15, 3, 5] with the relative error philosophy developed in [53, 54, 55, 35]. A valuable guideline for the error aspects is the development of an inexact proximal algorithm for equations governed by a monotone operator, and which is based on an estimation of the relative error, see [53, 54, 55]. Related results without monotonicity (with a control on the lack of monotonicity) have been obtained in [35].

Thus, in summary, this article aims at:

– providing a unified framework for the analysis of classical descent methods,

– relaxing exact descent conditions,

– extending convergence results obtained in [1, 3, 5, 53, 54, 55, 35] to richer and more flexible algorithms,

– providing theorems which cover general nonsmooth problems under easily verifiable assumptions (e.g. semi-algebraicity).

Let us proceed with a more precise description of the contents of this article.

In Section 2, we consider functions satisfying the Kurdyka-Łojasiewicz inequality. We first give the definition and a brief analysis of this basic property. Then in subsection 2.3, we provide an abstract convergence result for sequences satisfying the sufficient-decrease condition and the relative inexact optimality condition mentioned above.

This result is then applied to the analysis of several descent methods with relative error tolerance.

We recover and improve previous works on the question of gradient methods (Section 3) and proximal algorithms (Section 4). Our results are illustrated through semi-algebraic feasibility problems by means of an inexact version of the averaged projection method.

We also provide, in Section 5, an in-depth analysis of forward-backward splitting algorithms in a nonsmooth nonconvex setting. Setting aside the convex case, we did not find any general convergence results for this kind of algorithm, also, the results we present here seem to be new. These results can be applied to general semi-algebraic problems (or tame problems) and to nonconvex problems presenting a well-conditioned structure. An important and enlightening consequence of our study is that the bounded sequences $(x^k)_{k\in\mathbb{N}}$ generated by the nonconvex gradient projection algorithm

$$x^{k+1} \in P_C\left(x^k - \frac{1}{2L}\nabla h(x^k)\right)$$

are convergent sequences so long as C is a closed semi-algebraic subset of $\mathbb{R}^n$ and $h : \mathbb{R}^n \to \mathbb{R}$ is C^1 semi-algebraic with L-Lipschitz gradient (see [9] for some applications in signal

processing). As an application of our general results on forward-backward splitting, we consider the following type of problem

$$(P) \quad \min \left\{ \lambda \|x\|_0 + \frac{1}{2}\|Ax - b\|^2 : \ x \in \mathbb{R}^n \right\}$$

where $\lambda > 0$ and $\|\cdot\|_0$ is the counting norm (or the ℓ^0 norm), A is an $m \times n$ real matrix and $b \in \mathbb{R}^m$. We recall that for x in $\mathbb{R}^n$, $\|x\|_0$ is the number of nonzero components of x. This kind of problem is central in compressive sensing [28]. In [11, 12] this problem is tackled by using a "hard iterative thresholding" algorithm

$$x^{k+1} \in \operatorname{prox}_{\gamma_k \lambda \|\cdot\|_0} \left(x^k - \gamma_k (A^T A x^k - A^T b) \right),$$

where $(\gamma_k)_{k\in\mathbb{N}}$ is a sequence of stepsizes evolving in a convenient interval. The convergence results the authors obtained involve different assumptions on the linear operator A: they either assume that $\|A\| < 1$ [11, Theorem 3] or that A satisfies the restricted isometry property [12, Theorem 4]. Our results show that convergence actually occurs for any linear map so long as the sequence $(x^k)_{k\in\mathbb{N}}$ is bounded. We also consider iterative thresholding with ℓ^p "norms" for sparse approximation (in the spirit of [20]) and hard-constrained feasibility problems; in both cases convergence of the bounded sequences is established.

In a last section, we study the proximal regularization of a p blocks alternating method (with $p \geq 2$). This method has been introduced by Auslender [8] for convex minimization; see also [31] in a nonconvex setting. Convergence results for such methods are usually stated in terms of cluster points. To our knowledge, the first convergence result in a nonconvex setting, under fairly general assumptions, was obtained in [5] for a two-blocks exact version. Our generalization is twofolds: we consider methods involving an arbitrary numbers of blocks, and we provide a proper convergence result.

2 An abstract convergence result for inexact descent methods

The Euclidean scalar product of $\mathbb{R}^n$ and its corresponding norm are respectively denoted by $\langle \cdot, \cdot \rangle$ and $\|\cdot\|$.

2.1 Some definitions from variational analysis

Standard references are [22, 51, 45].
If $F : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ is a point-to-set mapping its *graph* is defined by

$$\operatorname{Graph} F := \{(x, y) \in \mathbb{R}^n \times \mathbb{R}^m : y \in F(x)\},$$

while its domain is given by $\operatorname{dom} F := \{x \in \mathbb{R}^n : F(x) \neq \emptyset\}$. Similarly, the graph of a real-extended-valued function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is defined by

$$\operatorname{Graph} f := \{(x, s) \in \mathbb{R}^n \times \mathbb{R} : s = f(x)\},$$

and its domain by $\operatorname{dom} f := \{f(x) < +\infty\}$. The epigraph of f is defined as usual as

$$\operatorname{epi} f := \{(x, \lambda) \in \mathbb{R}^n \times \mathbb{R} : f(x) \leq \lambda\}.$$

When f is a proper function, *i.e.* when $\operatorname{dom} f \neq \emptyset$, the set of its *global* minimizers, possibly empty, is denoted by

$$\operatorname{argmin} f := \{x \in \mathbb{R}^n : f(x) = \inf f\}.$$

The notion of subdifferential plays a central role in the following theoretical and algorithm developments.

For each $x \in \operatorname{dom} f$, the *Fréchet subdifferential* of f at x, written $\hat{\partial} f(x)$, is the set of vectors $x^* \in \mathbb{R}^n$ which satisfy

$$\liminf_{\substack{y \neq x \\ y \to x}} \frac{1}{\|x-y\|}[f(y) - f(x) - \langle x^*, y - x\rangle] \geq 0.$$

When $x \notin \operatorname{dom} f$, we set $\hat{\partial} f(x) = \emptyset$.

The limiting processes used in an algorithmic context necessitate the introduction of the more stable notion of *limiting-subdifferential* ([45]) (or simply subdifferential) of f. The subdifferential of f at $x \in \operatorname{dom} f$, written $\partial f(x)$, is defined as follows

$$\partial f(x) := \{x^* \in \mathbb{R}^n : \exists x_n \to x,\ f(x_n) \to f(x),\ x_n^* \in \hat{\partial} f(x_n) \to x^*\}.$$

It is straightforward to check from the definition the following closedness property of ∂f: Let $(x^k, v^k)_{k\in\mathbb{N}}$ be a sequence in $\mathbb{R}^n \times \mathbb{R}^n$ such that $(x^k, v^k) \in \operatorname{Graph} \partial f$ for all $k \in \mathbb{N}$. If (x^k, v^k) converges to (x, v), and $f(x^k)$ converges to $f(x)$ then $(x, v) \in \operatorname{Graph} \partial f$.

These generalized notions of differentiation give birth to generalized notions of critical point. A necessary (but not sufficient) condition for $x \in \mathbb{R}^n$ to be a minimizer of f is

$$\partial f(x) \ni 0. \tag{1}$$

A point that satisfies (1) is called *limiting-critical* or simply critical.

We end this section by some words on an important class of functions which are intimately linked to projection mappings: the indicator functions. Recall that if C is a closed subset of $\mathbb{R}^n$, its *indicator function* i_C is defined by

$$i_C(x) = \begin{cases} 0 & \text{if } x \in C, \\ +\infty & \text{otherwise,} \end{cases}$$

where x ranges over $\mathbb{R}^n$. Being given x in C, the limiting subdifferential of i_C at x is called the normal cone to C at x, it is denoted by $N_C(x)$.

The *projection on* C, written P_C, is the following *point-to-set* mapping:

$$P_C : \begin{cases} \mathbb{R}^n & \rightrightarrows & \mathbb{R}^n \\ x & \mapsto & P_C(x) := \operatorname{argmin}\{\|x - z\| : z \in C\}. \end{cases}$$

When C is nonempty, the closedness of C and the compactness of the closed unit ball imply that $P_C(x)$ is *nonempty* for all x in $\mathbb{R}^n$.

2.2 Kurdyka-Łojasiewicz inequality: the nonsmooth case

We begin this section by a brief discussion on real semi-algebraic sets and functions which will provide a very rich class of functions satisfying the Kurdyka-Łojasiewicz.

Definition 2.1. *(a) A subset S of $\mathbb{R}^n$ is a* real semi-algebraic set *if there exists a finite number of real polynomial functions $P_{ij}, Q_{ij} : \mathbb{R}^n \to \mathbb{R}$ such that*

$$S = \bigcup_{j=1}^{p} \bigcap_{i=1}^{q} \{ x \in \mathbb{R}^n : P_{ij}(x) = 0,\, Q_{ij}(x) < 0\}.$$

(b) A function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ (resp. a point-to-set mapping $F : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$) is called semi-algebraic if its graph $\{(x, \lambda) \in \mathbb{R}^{n+1} : f(x) = \lambda\}$ (resp. $\{(x, y) \in \mathbb{R}^{n+m} : y \in F(x)\}$) is a semi-algebraic subset of $\mathbb{R}^{n+1}$ (resp. $\mathbb{R}^{n+m}$).

One easily sees that the class of semi-algebraic sets is stable under the operation of finite union, finite intersection, Cartesian product or complementation and that polynomial functions are, of course, semi-algebraic functions.

The high flexibility of the concept of semi-algebraic sets is captured by the following fundamental theorem known as Tarski-Seidenberg principle.

Theorem 2.2 (Tarski-Seidenberg)**.** *Let A be a semi-algebraic subset of $\mathbb{R}^{n+1}$, then its canonical projection on $\mathbb{R}^n$, namely*

$$\{(x_1, \ldots, x_n) \in \mathbb{R}^n : \exists z \in \mathbb{R}, (x_1, \ldots, x_n, z) \in A\}$$

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

Let us illustrate the power of this theorem by proving that max functions associated to polynomial functions are semi-algebraic. Let S be a nonempty semi-algebraic subset of $\mathbb{R}^m$ and $g : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ a real polynomial function. Set $f(x) = \sup\{g(x, y) : y \in S\}$ (note that f can assume infinite values). Let us prove that f is semi-algebraic.

Using the definition and the stability with respect to finite intersection, we see that the set

$$\begin{aligned} & \{(x, \lambda, y) \in \mathbb{R}^n \times \mathbb{R} \times S : g(x, y) > \lambda\} \\ = \; & \{(x, \lambda, y) \in \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^m : g(x, y) > \lambda\} \bigcap (\mathbb{R}^n \times \mathbb{R} \times S)\,, \end{aligned}$$

is semi-algebraic. For (x, λ, y) in $\mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^m$, define the projection $\Pi(x, \lambda, y) = (x, \lambda)$ and use Π to project the above set on $\mathbb{R}^n \times \mathbb{R}$. One obtains the following semi-algebraic set

$$\{(x, \lambda) \in \mathbb{R}^n \times \mathbb{R} : \exists y \in S,\, g(x, y) > \lambda\}.$$

The complement of this set is

$$\{(x, \lambda) \in \mathbb{R}^n \times \mathbb{R} : \forall y \in S,\, g(x, y) \leq \lambda\} = \text{epi } f.$$

Hence epi f is semi-algebraic. Similarly hypo $f := \{(x, \mu) : f(x) \geq \mu\}$ is semi-algebraic hence Graph $f = \text{epi } f \cap \text{hypo } f$ is semi-algebraic. Of course, this result also holds when replacing sup by inf.

As a byproduct of these stability results, we recover the following standard result which will be useful for further developments.

Lemma 2.3. *Let S be a nonempty semi-algebraic subset of $\mathbb{R}^m$, then the function*

$$\mathbb{R}^m \ni x \to \operatorname{dist}(x, S)^2$$

is semi-algebraic.

Proof. It suffices to consider the polynomial function $g(x, y) = \|x - y\|^2$ for x, y in $\mathbb{R}^m$ and to use the definition of the distance function. □

The facts that the composition of semi-algebraic mappings gives a semi-algebraic mapping or that the image (resp. the preimage) of a semi-algebraic set by a semi-algebraic mapping is a semi-algebraic set are also consequences of the Tarski-Seidenberg principle. The reader is referred to [10, 13] for those and many other consequences of this principle.

As already mentioned in the introduction, a prominent feature of semi-algebraic functions is that they admit locally a sharp reparametrization, leading to what we call here Kurdyka-Łojasiewicz inequality. The most fundamental works on this subject are of course due to Łojasiewicz [42] (1963) and Kurdyka [37] (1998).

We proceed now to a formal definition of this inequality.
Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a proper lower semicontinuous function. For η_1, η_2 such that $-\infty < \eta_1 < \eta_2 \leq +\infty$, we set

$$[\eta_1 < f < \eta_2] = \{x \in \mathbb{R}^n : \eta_1 < f(x) < \eta_2\}.$$

The following definition is taken from [5] (see also [18]).

Definition 2.4 (Kurdyka-Łojasiewicz property)**.** *(a) The function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is said to have the Kurdyka-Łojasiewicz property at $x^* \in \operatorname{dom} \partial f$ if there exist $\eta \in (0, +\infty]$, a neighborhood U of x^* and a continuous concave function $\varphi : [0, \eta) \to \mathbb{R}_+$ such that:*

(i) *$\varphi(0) = 0$,*

(ii) *φ is C^1 on $(0, \eta)$,*

(iii) *for all $s \in (0, \eta)$, $\varphi'(s) > 0$,*

(iv) *for all x in $U \cap [f(x^*) < f < f(x^*) + \eta]$, the Kurdyka-Łojasiewicz inequality holds*

$$\varphi'(f(x) - f(x^*)) \operatorname{dist}(0, \partial f(x)) \geq 1. \tag{2}$$

(b) Proper lower semicontinuous functions which satisfy the Kurdyka-Łojasiewicz inequality at each point of $\operatorname{dom} \partial f$ *are called* KL functions*.*

Remark 2.5. (a) One can easily check that the Kurdyka-Łojasiewicz property is automatically satisfied at any non critical point $x^* \in \operatorname{dom} \partial f$, see for example Lemma 2.1 and Remark 3.2 (b) of [5].
(b) When f is smooth, finite-valued, and $f(x^*) = 0$, inequality (2) can be rewritten as

$$\|\nabla(\varphi \circ f)(x)\| \geq 1,$$

for each convenient x in $\mathbb{R}^n$. This inequality may be interpreted as follows: up to the reparametrization of the values of f via φ, we face a *sharp function*. Since the function φ is used here to turn a singular region –a region in which the gradients are arbitrarily small– into

a regular region, *i.e.* a place where the gradients are bounded away from zero, it is called a *desingularizing* function for f. For theoretical and geometrical developments concerning this inequality, see [18].
(c) The concavity assumption imposed on the function φ does not explicitly belong to the usual formulation of the Kurdyka-Łojasiewicz inequality. However this inequality holds in many instances with a concave function φ, see [5] for illuminating examples.
(d) It is important to observe that the KL inequality implies that the critical points lying in $U \cap [f(x^*) < f < f(x^*) + \eta]$ have the same critical value $f(x^*)$.

Among real-extended-valued lower-semicontinuous functions, typical KL functions are semi-algebraic functions or more generally functions definable in an o-minimal structure, see [15, 16, 17]. References on functions definable in an o-minimal structure are [25, 29]. Such examples are abundantly commented in [5], and they strongly motivate the present study. Other types of examples based on more analytical assumptions like uniform convexity, transversality or metric regularity can be found in [5], inequality (8.7) of [41], and Remark 3.6.

2.3 An inexact descent convergence result for KL functions

In this section, a and b are fixed positive constants. Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a proper lower semicontinuous function. In the sequel, we consider sequences $(x^k)_{k\in\mathbb{N}}$ which satisfy the following conditions, which we will subsequently refer to as H1, H2, H3:

H1. (*Sufficient decrease condition*). For each $k \in \mathbb{N}$,

$$f(x^{k+1}) + a\|x^{k+1} - x^k\|^2 \leq f(x^k);$$

H2. (*Relative error condition*). For each $k \in \mathbb{N}$, there exists $w^{k+1} \in \partial f(x^{k+1})$ such that

$$\|w^{k+1}\| \leq b\|x^{k+1} - x^k\|;$$

H3. (*Continuity condition*). There exists a subsequence $(x^{k_j})_{j\in\mathbb{N}}$ and $\tilde{x}$ such that

$$x^{k_j} \to \tilde{x} \text{ and } f(x^{k_j}) \to f(\tilde{x}), \qquad \text{as } j \to \infty.$$

Conditions H1 and H2 have been commented in the introduction; concerning condition H3, it is important to note that f itself is not required, in general, to be continuous or even continuous on its domain. Indeed, as we will see in the next sections, the nature of some algorithms (*e.g.* forward-backward splitting, Gauss-Seidel methods) forces the sequences to comply with condition H3 under a simple lower semicontinuity assumption.

The following abstract result is at the core of our convergence analysis.

Lemma 2.6. *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a proper lower semicontinuous function which satisfies the Kurdyka-Łojasiewicz property at some $x^* \in \mathbb{R}^n$. Denote by U, η and $\varphi : [0, \eta) \to \mathbb{R}_+$ the objects appearing in the Definition 2.4 of the KL property at x^*. Let $\delta, \rho > 0$ be such that $B(x^*, \delta) \subset U$ with $\rho \in (0, \delta)$.*

Consider a sequence $(x^k)_{k\in\mathbb{N}}$ *which satisfies conditions* H1, H2. *Assume moreover that*

$$f(x^*) \leq f(x^0) < f(x^*) + \eta, \tag{3}$$

$$\|x^* - x^0\| + 2\sqrt{\frac{f(x^0) - f(x^*)}{a}} + \frac{b}{a}\varphi(f(x^0) - f(x^*)) < \rho, \tag{4}$$

and

$$\forall k \in \mathbb{N},\ x^k \in B(x^*, \rho) \Rightarrow x^{k+1} \in B(x^*, \delta) \text{ with } f(x^{k+1}) \geq f(x^*). \tag{5}$$

Then, the sequence $(x^k)_{k\in\mathbb{N}}$ *satisfies*

$$\begin{aligned}
&\forall k \in \mathbb{N},\ x^k \in B(x^*, \rho),\\
&\sum_{k=0}^{+\infty} \|x^{k+1} - x^k\| < +\infty,\\
&f(x^k) \to f(x^*), \qquad \text{as } k \to \infty,
\end{aligned}$$

and converges to a point $\bar{x} \in B(x^*, \delta)$ *such that* $f(\bar{x}) \leq f(x^*)$.

If the sequence $(x^k)_{k\in\mathbb{N}}$ *also satisfies condition* H3, *then* $\bar{x}$ *is a critical point of* f, *and* $f(\bar{x}) = f(x^*)$.

Proof. The key point is to establish the following claim: for $j = 1, 2, \ldots$

$$x^j \in B(x^*, \rho), \tag{6}$$

and

$$\sum_{i=1}^{j} \|x^{i+1} - x^i\| + \|x^{j+1} - x^j\| \leq \|x^1 - x^0\| + \frac{b}{a}\big[\varphi(f(x^1) - f(x^*)) - \varphi(f(x^{j+1}) - f(x^*))\big]. \tag{7}$$

Concerning the above claim, first note that condition H1 implies that the sequence $(f(x^k))_{k\in\mathbb{N}}$ is nonincreasing, which by (3) gives $f(x^{j+1}) \leq f(x^0) < f(x^*) + \eta$. On the other hand, by assumption (5), the property $x^j \in B(x^*, \rho)$ implies $f(x^{j+1}) \geq f(x^*)$. Hence, the quantity $\varphi(f(x^{j+1}) - f(x^*))$ appearing in (7) makes sense.

Let us observe beforehand that, for all $k \geq 1$, the set $\partial f(x^k)$ is nonempty, and therefore x^k belongs to $\operatorname{dom} f$. As we already noticed, condition H1 implies that the sequence $(f(x^k))_{k\in\mathbb{N}}$ is nonincreasing, and it immediately yields

$$\|x^{k+1} - x^k\| \leq \sqrt{\frac{f(x^k) - f(x^{k+1})}{a}},\ \forall k \in \mathbb{N}. \tag{8}$$

Fix $k \geq 1$. We claim that if $f(x^k) < f(x^*) + \eta$ and $x^k \in B(x^*, \rho)$, then

$$2\|x^{k+1} - x^k\| \leq \|x^k - x^{k-1}\| + \frac{b}{a}\big[\varphi(f(x^k) - f(x^*)) - \varphi(f(x^{k+1}) - f(x^*))\big]. \tag{9}$$

If $x^{k+1} = x^k$ this inequality holds trivially. So, we assume that $x^{k+1} \neq x^k$. In this case, using (5) and (8), we conclude that $f(x^k) > f(x^{k+1}) \geq f(x^*)$ which, combined with KL inequality

and H2 shows that $w^k \neq 0$ and $x^{k-1} \neq x^k$. Since $w^k \in \partial f(x^k)$, using (again) KL inequality and H2, we obtain

$$\varphi'(f(x^k) - f(x^*)) \geq \frac{1}{\|w^k\|} \geq \frac{1}{b\|x^k - x^{k-1}\|}.$$

The concavity assumption on φ, $\varphi' > 0$, (5), and H1 imply

$$\begin{aligned}\varphi(f(x^k) - f(x^*)) - \varphi(f(x^{k+1}) - f(x^*)) \geq& \varphi'(f(x^k) - f(x^*))(f(x^k) - f(x^{k+1})) \\ \geq& \varphi'(f(x^k) - f(x^*))a\|x^{k+1} - x^k\|^2.\end{aligned}$$

Direct combination of the two above inequalities yields

$$\frac{b}{a}\big[\varphi(f(x^k) - f(x^*)) - \varphi(f(x^{k+1}) - f(x^*))\big] \geq \frac{\|x^{k+1} - x^k\|^2}{\|x^k - x^{k-1}\|}.$$

Multiplying this inequality by $\|x^k - x^{k-1}\|$, taking the square root on both sides and using inequality $2\sqrt{\alpha\beta} \leq \alpha + \beta$, we conclude that inequality (9) is satisfied.

Let us prove claims (6), (7) by induction on j. From (5) with $k = 0$, we obtain that $x^1 \in B(x^*, \delta)$ and $f(x^1) \geq f(x^*)$. Using now (8) with $k = 0$, we have

$$\|x^1 - x^0\| \leq \sqrt{\frac{f(x^0) - f(x^1)}{a}} \leq \sqrt{\frac{f(x^0) - f(x^*)}{a}}. \tag{10}$$

Combining the above equation with assumption (4), and using the triangle inequality we obtain

$$\|x^* - x^1\| \leq \|x^* - x^0\| + \|x^0 - x^1\| \leq \|x^* - x^0\| + \sqrt{\frac{f(x^0) - f(x^*)}{a}} < \rho,$$

which expresses that x^1 belongs to $B(x^*, \rho)$. Direct use of (9) with $k = 1$ shows that (7) holds with $j = 1$.
Suppose that (6) and (7) hold for some $j \geq 1$. Then, using the triangle inequality and (7) we have

$$\begin{aligned}\|x^* - x^{j+1}\| \leq& \|x^* - x^0\| + \|x^0 - x^1\| + \sum_{i=1}^{j} \|x^{i+1} - x^i\| \\ \leq& \|x^* - x^0\| + 2\|x^0 - x^1\| \\ &+ \frac{b}{a}\big[\varphi(f(x^1) - f(x^*)) - \varphi(f(x^{j+1}) - f(x^*))\big].\end{aligned}$$

Using the above inequality, (10) and assumption (4) we conclude that $x^{j+1} \in B(x^*, \rho)$. Hence, (9) holds with $k = j + 1$, *i.e.*

$$2\|x^{(j+1)+1} - x^{j+1}\| \leq \|x^{j+1} - x^j\| + \frac{b}{a}\big[\varphi(f(x^{j+1}) - f(x^*)) - \varphi(f(x^{(j+1)+1}) - f(x^*))\big].$$

Adding the above inequality with (7) (with $k = j$) yields (7) with $k = j + 1$, which completes the induction proof.

Direct use of (7) shows that

$$\sum_{i=1}^{j} \|x^{i+1} - x^i\| \leq \|x^1 - x^0\| + \frac{b}{a}\varphi(f(x^1) - f(x^*)).$$

Therefore,

$$\sum_{i=1}^{\infty} \|x^{i+1} - x^i\| < +\infty$$

which implies that the sequence $(x^k)_{k\in\mathbb{N}}$ converges to some $\bar{x}$. From H2 and (5) (note that φ concave yields φ' decreasing) we infer that $w^k \to 0$ and $f(x^k) \to \beta \geq f(x^*)$. If $\beta > f(x^*)$ then using Definition 2.4, (2) we have

$$\varphi'(\beta - f(x^*))\|w_k\| \geq 1, \qquad k = 0, 1, \ldots$$

which is absurd, because $w_k \to 0$. Therefore $\beta = f(x^*)$ and, since f is lower semicontinuous, $f(\bar{x}) \leq \beta = f(x^*)$.

To end the proof, note that if the sequence $(x^k)_{k\in\mathbb{N}}$ satisfies H3, then $\tilde{x} = \bar{x}$, $\bar{x}$ is critical and $f(\bar{x}) = \lim_{k\to\infty} f(x^k) = f(x^*)$.

□

Corollary 2.7. Let f, x^*, ρ, δ be as in the previous Lemma. For $q \geq 1$, consider a *finite* family $x^0, \ldots, x^q$ which satisfies H1 and H2, conditions (3), (4) and

$$\forall k \in \{0, \ldots, q\},\ \left(x^k \in B(x^*, \rho)\right) \Rightarrow \left(x^{k+1} \in B(x^*, \delta) \text{ with } f(x^{k+1}) \geq f(x^*)\right).$$

Then $x^j \in B(x^*, \rho)$ for all $j = 0, \ldots, q$.

Proof. Simply reproduce the beginning of the proof of the previous lemma. □

Corollary 2.8. *If we replace the assumption* (5) *in Lemma 2.6 by the set of assumptions,*

$$\eta < a(\delta - \rho)^2, \tag{11}$$

$$f(x^k) \geq f(x^*),\ \forall k \in \mathbb{N}, \tag{12}$$

the conclusion remains unchanged.

Proof. It suffices to prove that (11) and (12) implies (5). Let $x^k \in B(x^*, \rho)$. By H2, we have

$$\|x^{k+1} - x^k\| \leq \sqrt{\frac{f(x^k) - f(x^{k+1})}{a}} \leq \sqrt{\frac{\eta}{a}} < \delta - \rho.$$

Hence $\|x^{k+1} - x^*\| \leq \|x^{k+1} - x^k\| + \|x^k - x^*\| < \delta$. □

Lemma 2.6 and its corollaries have several important consequences that we now proceed to discuss.

Theorem 2.9 (Convergence to a critical point)**.** *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a proper lower semicontinuous function. Consider a sequence $(x^k)_{k\in\mathbb{N}}$ that satisfies* H1, H2*, and* H3*.*

If f has the Kurdyka-Łojasiewicz property at the cluster point $\tilde{x}$ specified in H3 *then the sequence $(x^k)_{k\in\mathbb{N}}$ converges to $\bar{x} = \tilde{x}$ as k goes to infinity, and $\bar{x}$ is a critical point of f. Moreover the sequence $(x^k)_{k\in\mathbb{N}}$ has a finite length, i.e.*

$$\sum_{k=0}^{+\infty} \|x^{k+1} - x^k\| < +\infty.$$

Proof. Let $\bar{x} = \tilde{x}$ be a cluster point of $(x^k)_{k\in\mathbb{N}}$ as given by H3 (i.e., $x^{k_j} \to \bar{x}$ and $f(x^{k_j}) \to f(\bar{x})$). Since $(f(x^k))_{k\in\mathbb{N}}$ is a nonincreasing sequence (a direct consequence of H1), we deduce that $f(x^k) \to f(\bar{x})$ and $f(x^k) \geq f(\bar{x})$ for all integers k. The function f has the KL property around $\bar{x}$, hence there exists φ, U, η as in Definition 2.4. Let $\delta > 0$ be such $B(\bar{x}, \delta) \subset U$, $\rho \in (0, \delta)$. If necessary, shrink η so that $\eta < a(\delta - \rho)^2$. Use the continuity property of φ to obtain the existence of an integer k_0 such that: $f(x^k) \in [f(\bar{x}), f(\bar{x}) + \eta)$ for all $k \geq k_0$ and

$$\|\bar{x} - x^{k_0}\| + 2\sqrt{\frac{f(x^{k_0}) - f(\bar{x})}{a}} + \frac{b}{a}\varphi(f(x^{k_0}) - f(\bar{x})) < \rho.$$

Since $f(x^k) \geq f(\bar{x})$ for all integers k, the conclusion follows by applying Corollary 2.8 to the sequence $(y^k)_{k\in\mathbb{N}}$ defined by $y^k = x^{k_0+k}$ for all integers k. □

As it will be shown later on, sequences complying with conditions H1, H2 and H3 are not necessarily generated by a local model (see Section 6) and therefore the proximity of the starting point x^0 with a local minimizer x^* does not imply, in general, that the limit point of the sequence lies in a neighbourhood of x^*.

However, under the following specific assumption, we can establish a convergence result to a local minimizer.

H4: For any $\delta > 0$ there exist $0 < \rho < \delta$ and $\nu > 0$ such that

$$\left.\begin{array}{c} x \in B(x^*, \rho),\ f(x) < f(x^*) + \nu \\ y \notin B(x^*, \delta) \end{array}\right\} \Rightarrow f(x) < f(y) + a\|y - x\|^2.$$

Theorem 2.10 (Local convergence to local minima)**.** *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a proper lower semicontinuous function which has the KL property at some local minimizer x^*. Assume that* H4 *holds at x^*.*

Then, for any $r > 0$, there exist $u \in (0, r)$ and $\mu > 0$ such that the inequalities

$$\|x^0 - x^*\| < u, \quad f(x^*) < f(x^0) < f(x^*) + \mu,$$

imply that any sequence $(x^k)_{k\in\mathbb{N}}$ starting from x^0, that satisfies H1, H2 *has the finite length property, remains in $B(x^*, r)$ and converges to some $\bar{x} \in B(x^*, r)$ critical point of f with $f(\bar{x}) = f(x^*)$.*

Proof. Take $r > 0$. Since x^* is a local minimum (hence critical) and f satisfies the Kurdyka-Łojasiewicz property, there exist $\eta_0 \in (0, +\infty]$, $\delta \in (0, r)$, and a continuous concave function $\varphi : [0, \eta_0) \to \mathbb{R}_+$ such that:

- $\varphi(0) = 0$,
- φ is C^1 on $(0, \eta_0)$,
- for all $s \in (0, \eta_0)$, $\varphi'(s) > 0$.
- for all x in $B(x^*, \delta) \cap [f(x^*) < f < f(x^*) + \eta_0]$, the Kurdyka-Łojasiewicz inequality holds

$$\varphi'(f(x) - f(x^*)) \operatorname{dist}(0, \partial f(x)) \geq 1. \tag{13}$$

- for all x in $B(x^*, \delta)$,

$$f(x) \geq f(x^*). \tag{14}$$

We infer from assumption H4 that there exist $\rho \in (0, \delta)$ and $\nu > 0$ such that

$$\left.\begin{array}{c} x \in B(x^*, \rho), f(x) < f(x^*) + \nu \\ y \notin B(x^*, \delta) \end{array}\right\} \Rightarrow f(x) < f(y) + a\|y - x\|^2. \tag{15}$$

Set $\eta = \min\{\eta_0, \nu\}$ and let $k \in \mathbb{N}$. If x^k is such that $f(x^k) < f(x^*) + \eta$ and $\|x^k - x^*\| < \rho$ then H4, together with H1, implies that $x^{k+1} \in B(x^*, \delta)$, and thus that $f(x^{k+1}) \geq f(x^*)$ (recall that x^* is a local minimizer on $B(x^*, \delta)$). That's precisely property (5) of Lemma 2.6.

Choose $u, \mu > 0$ such that

$$u < \rho/3,\ \mu < \eta,\ 2\sqrt{\frac{\mu}{a}} + \frac{b}{a}\varphi(\mu) < \frac{2\rho}{3}.$$

If x^0 satisfies the set of inequalities $\|x^0 - x^*\| < u$ and $f(x^*) < f(x^0) < f(x^*) + \mu$ we therefore have

$$\|x^* - x^0\| + 2\sqrt{\frac{f(x^0) - f(x^*)}{a}} + \frac{b}{a}\varphi(f(x^0) - f(x^*)) < \rho.$$

which is precisely property (4) of Lemma 2.6. Using Lemma 2.6 we conclude that the sequence $(x^k)_{k\in\mathbb{N}}$ has the finite length property, remains in $B(x^*, \rho)$, converges to some $\bar{x} \in B(x^*, \delta)$, $f(x^k) \to f(x^*)$ and $f(\bar{x}) \leq f(x^*)$. Since $f(x^*)$ is the minimum value of f in $B(x^*, \delta)$, $f(\bar{x}) = f(x^*)$ and the sequence $(x^k)_{k\in\mathbb{N}}$ has also property H3. So, $\bar{x}$ is a critical point of f. □

Remark 2.11. Let us verify that Condition H4 is satisfied when $x^* \in \operatorname{dom} f$ is a local minimum and the function f satisfies the following global growth condition:

$$f(y) \geq f(x^*) - \frac{a}{4}\|y - x^*\|^2 \text{ for all } y \in \mathbb{R}^n. \tag{16}$$

Let $\delta > \rho$ and ν be positive real numbers. Take $y \in \mathbb{R}^n$ such that $\|y - x^*\| \geq \delta$ and $x \in \mathbb{R}^n$ such that $\|x - x^*\| \leq \rho$ and $f(x) < f(x^*) + \nu$. From (16), and the triangle inequality we infer

$$\begin{aligned} f(y) &\geq f(x) - \nu - \frac{a}{4}\|y - x^*\|^2 \\ &\geq f(x) - \nu - \frac{a}{2}\|y - x^*\|^2 + \frac{a}{4}\|y - x^*\|^2 \\ &\geq f(x) - \nu - a\|y - x\|^2 - a\|x - x^*\|^2 + \frac{a}{4}\|y - x^*\|^2 \\ &\geq f(x) - \nu - a\|y - x\|^2 - a\rho^2 + \frac{a}{4}\delta^2. \end{aligned}$$

Hence

$$f(y) + a\|y - x\|^2 \geq f(x) + \left(-\nu - a\rho^2 + \frac{a}{4}\delta^2\right) \text{ for all } y \in \mathbb{R}^n. \tag{17}$$

We conclude by noticing that $-\nu - a\rho^2 + \frac{a}{4}\delta^2$ is nonnegative for ρ and ν sufficiently small.

We end this section by a result on the convergence toward a global minimum similar to [5], Theorem 3.3. Observe that, in this context, the set of global minimizers may be a continuum.

Theorem 2.12 (Local convergence to global minima)**.** *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a lower semicontinuous function which has the KL property at some x^*, a global minimum point of f.*

For each $r > 0$, there exist $u \in (0, r)$, $\mu > 0$ such that the inequalities

$$\|x^0 - x^*\| < u, \quad \min f < f(x^0) < \min f + \mu$$

imply that any sequence $(x^k)_{k\in\mathbb{N}}$ that satisfies H1, H2 *and which starts from x^0 satisfies*

(i) $x^k \in B(x^*, r)$, $\forall k \in \mathbb{N}$,

(ii) x^k *converges to some* $\bar{x}$ *and* $\sum_{k=1}^{+\infty} \|x^{k+1} - x^k\| < +\infty$,

(iii) $f(\bar{x}) = \min f$.

Proof. It is a straightforward variant of Theorems 2.9 and 2.10. □

3 Inexact gradient methods

The first natural domain of application of our previous results concerns the simplest first-order methods, namely the gradient methods. As we shall see, our abstract framework (Theorem 2.9) allows to recover some of the results of [1]. In order to illustrate the versatility of our algorithmic framework, we also consider a fairly general semi-algebraic feasibility problem, and we provide, in the line of [41], a local convergence proof for an inexact averaged projection method.

3.1 General convergence result

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a C^1 function whose gradient is Lipschitz continuous with constant L (or L-Lipschitz continuous). We consider the following algorithm.

Algorithm 1 Take some positive parameters a, b with $a > L$.
Fix x^0 in $\mathbb{R}^n$. For $k = 0, 1, \ldots$ consider:

$$\langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{a}{2}\|x^{k+1} - x^k\|^2 \leq 0, \tag{18}$$

$$\|\nabla f(x^k)\| \leq b\|x^{k+1} - x^k\|. \tag{19}$$

To illustrate the variety of dynamics covered by Algorithm 1, let us show how variable metric gradient algorithms can be cast in this framework. Consider a sequence $(A^k)_{k\in\mathbb{N}}$ of symmetric positive definite matrices in $\mathbb{R}^{n\times n}$ such that for each $k \in \mathbb{N}$ the eigenvalues λ_i^k of A^k satisfy

$$0 < \underline{\lambda} \leq \lambda_i^k \leq \overline{\lambda},$$

where $\underline{\lambda}$ and $\overline{\lambda}$ are given thresholds. For each integer k, consider the following subproblem built on a second-order model of f around the point x^k:

$$\text{minimize } \left\{ \langle \nabla f(x^k), u - x^k \rangle + \frac{1}{2} \langle A^k(u - x^k), u - x^k \rangle : u \in \mathbb{R}^n \right\}.$$

This type of quadratic models arises, for instance, in trust-region methods (see [1] which is also connected to Łojasiewicz inequality). When solving the above problem exactly, we obtain the following method

$$x^{k+1} = x^k - (A^k)^{-1} \nabla f(x^k),$$

which satisfies

$$\langle \nabla f(x^k), x^{k+1} - x^k \rangle + \underline{\lambda} \|x^{k+1} - x^k\|^2 \leq 0, \tag{20}$$

$$\|\nabla f(x^k)\| \leq \overline{\lambda} \|x^{k+1} - x^k\|. \tag{21}$$

So long as $\underline{\lambda} > \frac{L}{2}$, the sequence $(x^k)_{k \in \mathbb{N}}$ falls into the general category delineated by Algorithm 1.

For the convergence analysis of Algorithm 1, we shall of course use the elementary but important descent lemma –its elementary proof is left to the reader.

Lemma 3.1 (Descent lemma)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function and C a convex subset of $\mathbb{R}^n$ with nonempty interior. Assume that f is C^1 on a neighborhood of each point in C and that ∇f is L-Lipschitz continuous on C. Then, for any two points x, u in C,*

$$f(u) \leq f(x) + \langle \nabla f(x), u - x \rangle + \frac{L}{2} \|u - x\|^2. \tag{22}$$

We then have the following result:

Theorem 3.2. *Assume that $f : \mathbb{R}^n \to \mathbb{R}$ is a C^1 function with L-Lipschitz continuous gradient, and that f is bounded from below. If f is a KL function, then each bounded sequence $(x^k)_{k \in \mathbb{N}}$ generated by* Algorithm 1 *converges to some critical point $\bar{x}$ of f.*
Moreover, the sequence $(x^k)_{k \in \mathbb{N}}$ has a finite length, i.e. $\sum_k \|x^{k+1} - x^k\| < +\infty$.

Proof. Applying the descent lemma at points $u = x^{k+1}$ and $x = x^k$, inequality (18) becomes

$$f(x^{k+1}) - f(x^k) + \frac{a - L}{2} \|x^{k+1} - x^k\|^2 \leq 0.$$

Since $a > L$, condition H1 of Theorem 2.9 is satisfied. To see that H2 is satisfied, use the Lipschitz continuity property of ∇f and (19) to obtain

$$\|\nabla f(x^{k+1})\| \leq \|\nabla f(x^{k+1}) - \nabla f(x^k)\| + \|\nabla f(x^k)\| \leq (L + b) \|x^{k+1} - x^k\|.$$

The sequence $(x^k)_{k \in \mathbb{N}}$ has been assumed to be bounded. Thus it admits a converging subsequence, and, by continuity of f, H3 is trivially fulfilled. We can therefore apply theorem 2.9 to conclude. □

Remark 3.3. The conclusion of Theorem 3.2 remains unchanged if the assumption that ∇f is Lipschitz continuous on $\mathbb{R}^n$ and f is a KL function are replaced by the assumptions: *There exists a closed subset S of $\mathbb{R}^n$ such that*

(i) ∇f is L-Lipschitz continuous on $\operatorname{co} S$;

(ii) $x^k \in S$ for all $k \in \mathbb{N}$;

(iii) f satisfies the KL inequality at each point of S,

where $\operatorname{co} S$ denotes the convex envelope of S. The result is evident from the proof. Just notice that the L-Lipschitz continuity of ∇f on $\operatorname{co} S$ is needed in order to apply the descent lemma.

3.2 Prox-regularity

When considering nonconvex feasibility problems, we are led to consider squared distance functions to nonconvex sets. Contrary to what happens in the standard convex setting, such functions may fail to be differentiable. If we want to handle feasibility problems through gradient methods (e.g. Algorithm 1), this lack of regularity causes serious trouble. The key concept of *prox-regularity* provides a characterization of the local differentiability of these functions and, as we will see in the next section, it allows in turn to design averaged projection methods with interesting converging properties.

A closed subset F of $\mathbb{R}^n$ is *prox-regular* if its projection operator P_F is single-valued around each point x in F (see [50, Theorem 1.3, $(a) \Leftrightarrow (f)$]). Prominent examples of prox-regular sets are closed convex sets and C^2 submanifolds of $\mathbb{R}^n$ (see [50] and references therein).

Set $g(x) = \frac{1}{2}\operatorname{dist}(x, F)^2$ and assume that F is prox-regular. Let us gather the following definition/properties concerning F that are fundamental for our purpose.

Theorem 3.4 ([50])**.** *Let F be a closed prox-regular set. Then for each $\bar{x}$ in F there exists $r > 0$ such that:*

(a) The projection P_F is single-valued on $B(\bar{x}, r)$,

(b) the function g is C^1 on $B(\bar{x}, r)$ and $\nabla g(x) = x - P_F(x)$,

(c) the gradient mapping ∇g is 1-Lipschitz continuous on $B(\bar{x}, r)$.

Item (c) is not explicitly developed in [50], a proper proof can be found in [41, Proposition 8.1].

3.3 Averaged projections for feasibility problems

Let $F_1, \ldots, F_p$ be nonempty closed semi-algebraic, prox-regular subsets of $\mathbb{R}^n$ such that

$$\bigcap_{i=1}^{p} F_i \neq \emptyset.$$

A classical approach to the problem of finding a common point to the sets $F_1, \ldots, F_p$ is to find a global minimizer of the function $f : \mathbb{R}^n \to [0, +\infty)$

$$f(x) := \frac{1}{2} \sum_{i=1}^{p} \operatorname{dist}(x, F_i)^2, \tag{23}$$

where dist $(\cdot, F_i)$ is the distance function to the set F_i.

As it is well known in the convex case, the averaged projection method corresponds exactly to an explicit gradient method applied to the function f. In a nonconvex setting, we are thus led to study the following algorithm:

Inexact averaged projection algorithm Take $\theta \in (0,1)$, $\alpha < \frac{1}{2}$ and $M > 0$ such that

$$\frac{1-\alpha}{\theta} > \frac{1}{2}. \tag{24}$$

Given a starting point x^0 in $\mathbb{R}^n$, consider the following algorithm

$$x^{k+1} \in (1-\theta)\, x^k + \theta \left(\frac{1}{p} \sum_{i=1}^{p} P_{F_i}(x^k) \right) + \epsilon^k, \tag{25}$$

where $(\epsilon^k)_{k\in\mathbb{N}}$ is a sequence of errors which satisfies

$$\langle \epsilon^k, x^{k+1} - x^k \rangle \leq \alpha \|x^{k+1} - x^k\|^2 \tag{26}$$

$$\|\epsilon^k\| \leq M \|x^{k+1} - x^k\| \tag{27}$$

for all $k \in \mathbb{N}$.

We then have the following result.

Theorem 3.5 (Inexact averaged projection method)**.** *Let $F_1, \ldots, F_p$ be semi-algebraic, and prox-regular subsets of $\mathbb{R}^n$ which satisfy*

$$\bigcap_{i=1}^{p} F_i \neq \emptyset.$$

If x^0 is sufficiently close to $\bigcap_{i=1}^{p} F_i$, then the inexact averaged projection algorithm reduces to the gradient method

$$x^{k+1} = x^k - \frac{\theta}{p} \nabla f(x^k) + \epsilon^k,$$

with f being given by (23), which therefore defines a unique sequence. Moreover, this sequence has a finite length and converges to a feasible point $\bar{x}$, i.e. *such that*

$$\bar{x} \in \bigcap_{i=1}^{p} F_i.$$

Proof. Let us first observe that the function f (given by (23)) is semi-algebraic, because the distance function to any nonempty semi-algebraic set is semi-algebraic (see Lemma 2.3 or [29, 15]). This implies in particular that f is a KL function (see the end of Section 2.2).

Take a point x^* in $\cap_1^p F_i$ and use Theorem 3.4 to obtain $\delta > 0$ such that, for each $i = 1, \ldots, p$,

(a) the projection P_{F_i} is single-valued on $B(x^*, \delta)$,

(b) the function $g_i := \frac{1}{2}\text{dist}\,(\cdot, F_i)^2$ is C^1 on $B(x^*, \delta)$ and $\nabla g_i(x) = x - P_{F_i}(x)$,

(c) the gradient mapping ∇g_i is 1-Lipschitz continuous on $B(x^*, \delta)$.

Since the function f has the KL property around x^*, there exist φ, U, η as in Definition 2.4. Shrinking δ if necessary, we may assume that $B(x^*, \delta) \subset U$. Take $\rho \in (0, \delta)$ and shrink η so that

$$\eta < \frac{1-2\alpha}{2p}(\delta - \rho)^2. \tag{28}$$

Choose a starting point x^0 such that: $0 = f(x^*) \leq f(x^0) < \eta$ and

$$\|x^* - x^0\| + 2\sqrt{\frac{f(x^0)}{a}} + \frac{b}{a}\varphi(f(x^0)) < \rho. \tag{29}$$

Introduce $a = p(\frac{1-\alpha}{\theta} - \frac{1}{2}) > 0$ (cf (24)) and $b = p\left(1 + \frac{1+M}{\theta}\right)$.

Let us prove by induction that the averaged projection algorithm defines a unique sequence that satisfies:

- Conditions H1 and H2 of Section 2.3 with respect to the function f and the constants a, b,
- $x^k \in B(x^*, \rho)$ for all integers $k \geq 0$.

The case $k = 0$ follows from (29). Before proceeding, note that, if a point x belongs to $B(x^*, \delta)$, we have

$$\nabla f(x) = \sum_{i=1}^{p}(x - P_{F_i}(x)).$$

Using Cauchy-Schwarz inequality (one may as well use the convexity of $\|\cdot\|^2$), we obtain

$$\begin{aligned} \|\nabla f(x)\|^2 &\leq \left(\sum_{i=1}^{p}\|x - P_{F_i}(x)\|\right)^2 \\ &\leq p\sum_{i=1}^{p}\|x - P_{F_i}(x)\|^2 \\ &= 2pf(x). \end{aligned} \tag{30}$$

Let $k \geq 0$. Assume now that $x^k \in B(x^*, \rho)$ and properties H1, H2 hold for the $k+1$-uple $x^0, \ldots, x^k$. Using Theorem 3.4, the inclusion (25) defining x^{k+1} may be rewritten as follows

$$x^{k+1} = x^k - \frac{\theta}{p}\nabla f(x^k) + \epsilon^k,$$

hence x^{k+1} is uniquely defined. The above equality yields (note that $\theta \in (0,1)$ and $p \geq 1$)

$$\|x^{k+1} - x^k\|^2 - 2\langle x^{k+1} - x^k, \epsilon^k\rangle + \|\epsilon^k\|^2 \leq \|\nabla f(x^k)\|^2,$$

thus, in view of (26), (28) and (30),

$$\|x^{k+1} - x^k\|^2 \leq \frac{2p}{1-2\alpha}f(x^k) \leq (\delta - \rho)^2. \tag{31}$$

Since $\|x^{k+1} - x^*\| \leq \|x^{k+1} - x^k\| + \|x^k - x^*\|$, this implies that $x^{k+1} \in B(x^*, \delta)$. Using (26) and (27), let us verify that property H1 is satisfied for $x^0, x^1, \ldots, x^{k+1}$. We have

$$\begin{aligned}\langle \nabla f(x^k), x^{k+1} - x^k \rangle &= \frac{p}{\theta}\langle (x^k - x^{k+1}) + \epsilon^k, x^{k+1} - x^k \rangle \\ &\leq -\frac{p}{\theta}(1-\alpha)\|x^{k+1} - x^k\|^2.\end{aligned}$$

By Theorem 3.4, we know that ∇f is p-Lipschitz on $B(x^*, \delta)$; we can thus combine the above inequality with the descent lemma to obtain

$$f(x^{k+1}) + \frac{2\frac{p}{\theta}(1-\alpha) - p}{2}\|x^{k+1} - x^k\|^2 \leq f(x^k),$$

that is

$$f(x^{k+1}) + a\|x^{k+1} - x^k\|^2 \leq f(x^k),$$

which is exactly property H1. On the other hand we have

$$\begin{aligned}\|\nabla f(x^{k+1})\| &\leq \|\nabla f(x^{k+1}) - \nabla f(x^k)\| + \|\nabla f(x^k)\| \\ &\leq p\|x^{k+1} - x^k\| + \frac{p}{\theta}(\|x^{k+1} - x^k\| + \|\epsilon^k\|) \\ &\leq p\left(1 + \frac{1+M}{\theta}\right)\|x^{k+1} - x^k\|, \\ &= b\|x^{k+1} - x^k\|\end{aligned}$$

where the second inequality comes from the Lipschitz property of ∇f and the definition of the sequence, while the last one follows from the error stepsize inequality, namely (27). Property H2 is therefore satisfied.

Applying now Corollary 2.7, we get $x^{k+1} \in B(x^*, \rho)$ and our induction proof is complete.

As a consequence, the algorithm defines a unique sequence that satisfies the assumption of Lemma 2.6 (or Theorem 3.2), hence it generates a finite length sequence which converges to a point $\bar{x}$ such that $f(\bar{x}) = 0$. □

Remark 3.6. (a) In [41], a paper that inspired the above development, the authors establish similar results for sets F_i having a *linearly regular intersection at some point* $\bar{x}$, an important concept that originates from [45, Theorem 2.8]. A linearly regular intersection at $\bar{x}$ means that the equation

$$\sum_{i=1}^{p} y_i = 0, \text{ with } y_i \in N_{F_i}(\bar{x})$$

admits $y_i = 0$, $\forall i = 1, \ldots, p$ as a *unique* solution.

An important fact, tightly linked to the convergence result for averaged projections given in [41, Theorem 7.3], is that the objective $f(x) := \frac{1}{2}\sum_i \operatorname{dist}(x, F_i)^2$ satisfies the inequality

$$\|\nabla f(x)\|^2 \geq c f(x),$$

where x is in a neighborhood of $\bar{x}$ and with c being a positive constant (see [41, Proposition 8.6]). One recognizes the Łojasiewicz inequality with a desingularizing function of the form $\varphi(s) = \frac{2}{\sqrt{c}}\sqrt{s}$ with $s \geq 0$.

(b) The above proof does not rely directly on Theorem 3.2 or Lemma 2.6, because we do not know a priori that the sequence enters the abstract framework of descent methods defined in Section 2.3.

4 Inexact proximal algorithm

Let us first recall the exact version of the proximal algorithm for nonconvex functions [35, 3].

Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a proper lower semicontinuous function which is bounded from below, and λ a positive parameter. It is convenient to introduce formally the proximal correspondence $\mathrm{prox}_{\lambda f} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$, which is defined through the formula

$$\mathrm{prox}_{\lambda f} x := \mathrm{argmin}\ \left\{ f(y) + \frac{1}{2\lambda}\|y - x\|^2 : y \in \mathbb{R}^n \right\}.$$

Note that for any $\mu > 0$, we have $\mathrm{prox}_{\lambda(\mu f)} = \mathrm{prox}_{(\lambda\mu) f}$, so that these objects may be simply denoted by $\mathrm{prox}_{\lambda\mu f}$.

In view of the assumption $\inf f > -\infty$, the lower semicontinuity of f and the coercivity of the squared norm imply that $\mathrm{prox}_{\lambda f}$ has *nonempty* values. Observe finally that, contrary to the case when f is convex, we generally do not face here a single-valued operator.

The classical proximal algorithm writes

$$x^{k+1} \in \mathrm{prox}_{\lambda_k f}(x^k), \tag{32}$$

where λ_k is a sequence of stepsize parameters lying in an interval $[\underline{\lambda}, \overline{\lambda}] \subset (0, +\infty)$, and $x_0 \in \mathbb{R}^n$. Writing successively the definition of the proximal operator and the associated first optimality condition (use the sum rule [51]), we obtain

$$f(x^{k+1}) + \frac{1}{2\lambda_k}\|x^{k+1} - x^k\|^2 \leq f(x^k) \tag{33}$$

$$w^{k+1} \in \partial f(x^{k+1}); \tag{34}$$

$$\lambda_k w^{k+1} + x^{k+1} - x^k = 0. \tag{35}$$

4.1 Convergence of an inexact proximal algorithm for KL functions

Let us introduce an inexact version of the proximal point method. Consider the sequence $(x^k)_{k\in\mathbb{N}}$ generated by the following algorithm:

Algorithm 2: Take $x_0 \in \mathbb{R}^n$, $\quad 0 < \underline{\lambda} \leq \overline{\lambda} < \infty, \quad 0 \leq \sigma < 1, \quad 0 < \theta \leq 1$.
For $k = 0, 1, \ldots$, choose $\lambda_k \in [\underline{\lambda}, \overline{\lambda}]$, and find $x^{k+1} \in \mathbb{R}^n$, $w^{k+1} \in \mathbb{R}^n$ such that

$$f(x^{k+1}) + \frac{\theta}{2\lambda_k}\|x^{k+1} - x^k\|^2 \leq f(x^k) \tag{36}$$

$$w^{k+1} \in \partial f(x^{k+1}); \tag{37}$$

$$\|\lambda_k w^{k+1} + x^{k+1} - x^k\|^2 \leq \sigma(\|\lambda_k w^{k+1}\|^2 + \|x^{k+1} - x^k\|^2). \tag{38}$$

The error criterion (38) is a particular case of the error criterion considered in [54], but here, contrary to [54], we are not dealing with a maximal monotone operator and no extra-gradient step is performed. In our setting, condition (38) can be replaced by a weaker condition: *for some positive* $b > 0$

$$\|\lambda_k w^{k+1}\| \leq b\|x^{k+1} - x^k\|. \tag{39}$$

The fact that Algorithm 2 is an inexact version of the proximal algorithm is transparent: the first inequality (36) reflects the fact that a sufficient decrease of the value must be achieved, while the last lines (38), (39) both correspond to an inexact optimality condition.

The following elementary Lemma is useful for the convergence analysis of the algorithm.

Lemma 4.1. *Let $\sigma \in (0, 1]$. If $x, y \in R^n$, and*

$$\|x + y\|^2 \leq \sigma(\|x\|^2 + \|y\|^2), \tag{40}$$

then

$$\frac{1-\sigma}{2}(\|x\|^2 + \|y\|^2) \leq -\langle x, y\rangle.$$

Assuming moreover $\sigma \in (0, 1)$

$$\frac{1-\sqrt{1-(1-\sigma)^2}}{1-\sigma}\|y\| \leq \|x\| \leq \frac{1+\sqrt{1-(1-\sigma)^2}}{1-\sigma}\|y\|.$$

Proof. Note that (40) is equivalent to

$$\|x\|^2 + 2\langle x, y\rangle + \|y\|^2 \leq \sigma(\|x\|^2 + \|y\|^2).$$

Direct algebraic manipulation of the above inequality yields the first inequality. For proving the second and third inequalities, combine the above inequality with Cauchy-Schwarz inequality to obtain

$$(1-\sigma)\|x\|^2 - 2\|x\|\|y\| + (1-\sigma)\|y\|^2 \leq 0$$

Viewing the left-hand side of the above inequality as a quadratic function of $\frac{\|x\|}{\|y\|}$ yields the conclusion. □

The main result of this section is the following theorem.

Theorem 4.2 (Inexact proximal algorithm)**.** *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a proper lower semicontinuous KL function which is bounded from below. Assume that the restriction of f to its domain is a continuous function. If a sequence $(x^k)_{k\in\mathbb{N}}$ generated by* Algorithm 2 *(or by* (36)*,* (37) *and* (39)*) is bounded, then it converges to some critical point $\bar{x}$ of f.*
Moreover the sequence $(x^k)_{k\in\mathbb{N}}$ has a finite length, i.e. $\sum_k \|x^{k+1} - x^k\| < +\infty$.

Proof. First use Lemma 4.1 to conclude that condition (38) implies (39). Therefore, we assume that (36), (37) and (39) holds. If $(x^k)_{k\in\mathbb{N}}$ is bounded, there exists a subsequence (x^{k_j}) and $\bar{x}$ such that

$$x^{k_j} \to \bar{x} \quad \text{as } j \to \infty.$$

Since f is continuous on its effective domain and $f(x^{k_j}) \leq f(x_0) < +\infty$ for all j, we conclude that

$$f(x^{k_j}) \to f(\bar{x}) \text{ as } j \to \infty.$$

We can now apply Theorem 2.9, and thus obtain the convergence of the sequence $(x^k)_{k\in\mathbb{N}}$ to a critical point of f. □

4.2 A variant for convex functions

When the function under consideration is convex and satisfies the Kurdyka-Łojasiewicz property, Algorithm 2 can be simplified while its convergence properties are maintained.

Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a proper lower semicontinuous convex function. Consider the sequence $(x^k)_{k\in\mathbb{N}}$ generated by the following algorithm.

Algorithm 2bis: Take $0 < \underline{\lambda} \leq \overline{\lambda} < \infty, \quad 0 \leq \sigma < 1$.
For $k = 0, 1, \ldots,$ choose $\lambda_k \in [\underline{\lambda}, \overline{\lambda}]$ and find $x^{k+1} \in \mathbb{R}^n$, $w^{k+1} \in \mathbb{R}^n$ such that

$$w^{k+1} \in \partial f(x^{k+1}), \tag{41}$$

$$\|\lambda_k w^{k+1} + x^{k+1} - x^k\|^2 \leq \sigma(\|\lambda_k w^{k+1}\|^2 + \|x^{k+1} - x^k\|^2). \tag{42}$$

Before stating our main results, let us establish some elementary inequalities. We claim that for each k

$$f(x^{k+1}) + \frac{1-\sigma}{2\overline{\lambda}}\|x^{k+1} - x^k\|^2 + \frac{(1-\sigma)\underline{\lambda}}{2}\|w^{k+1}\|^2 \leq f(x^k), \tag{43}$$

and

$$\|w^{k+1}\| \leq \frac{1 + \sqrt{1-(1-\sigma)^2}}{\underline{\lambda}(1-\sigma)}\|x^{k+1} - x^k\|. \tag{44}$$

For proving (43), use the convexity of f and inclusion (41) to obtain

$$f(x^k) \geq f(x^{k+1}) + \langle x^k - x^{k+1}, w^{k+1}\rangle.$$

Using the above inequality, the algebraic identity

$$\langle x^k - x^{k+1}, w^{k+1}\rangle = \frac{1}{2\lambda_k}[\|\lambda_k w^{k+1}\|^2 + \|x^{k+1} - x^k\|^2 - \|\lambda_k w^{k+1} + x^{k+1} - x^k\|^2]$$

and (42), we obtain

$$f(x^k) \geq f(x^{k+1}) + \frac{1-\sigma}{2}\left[\lambda_k\|w^{k+1}\|^2 + \frac{1}{\lambda_k}\|x^{k+1} - x^k\|^2\right]. \tag{45}$$

Combining this inequality with the assumption $\underline{\lambda} \leq \lambda_k \leq \overline{\lambda}$ yields (43).
Equation (44) follows from Lemma 4.1, inequality (42) and assumption $\underline{\lambda} \leq \lambda_k \leq \overline{\lambda}$.

Theorem 4.3. *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a proper convex lower semicontinuous. Assume that f is a KL function which is bounded from below and let $(x^k)_{k\in\mathbb{N}}$ be a sequence generated by* Algorithm 2bis.

If $(x^k)_{k\in\mathbb{N}}$ is bounded, then it converges to a minimizer of f and the sequence of values $f(x^k)$ converges to the program value $\min f$. *Moreover, the sequence has a finite length, i.e. $\sum_k \|x^{k+1} - x^k\| < +\infty$.*

Proof. Since f is bounded from below, it follows from (43) and $\sigma < 1$ that

$$\sum_{k=1}^{+\infty} \|x^{k+1} - x^k\|^2 < +\infty, \qquad \sum_{k=1}^{+\infty} \|w^k\|^2 < +\infty.$$

Therefore

$$w_k \to 0 \quad \text{as } k \to +\infty.$$

Since (x^k) has been assumed to be bounded, there exists a subsequence (x^{k_j}) which converges to some $\bar{x}$. By (43) and $\sigma < 1$, we also see that the sequence $(f(x^k))$ is decreasing. From this property and the lower semicontinuity of f, we deduce that

$$f(\bar{x}) \leq \liminf_{j\to\infty} f(x^{k_j}) = \lim_{k\to\infty} f(x^k).$$

Using the convexity of f and the inclusion $w^k \in \partial f(x^k)$ for $k \geq 1$, we obtain

$$f(\bar{x}) \geq f(x^{k_j}) + \langle \bar{x} - x^{k_j}, w^{k_j} \rangle, \qquad j = 2, 3, \ldots.$$

Passing to the limit, as $j \to \infty$, in the above inequality we conclude that

$$f(\bar{x}) \geq \lim_{k \to \infty} f(x^k).$$

Therefore

$$f(\bar{x}) = \lim_{k \to \infty} f(x^k) = \lim_{j \to \infty} f(x^{k_j}).$$

Then use (43), (44) and Theorem 2.9 to obtain the convergence of the sequence (x^k) to some $\bar{x}$. From $w^k \in \partial f(x^k)$, $w^k \to 0$ as $k \to +\infty$, and the closedness property of ∂f, we deduce that $\partial f(\bar{x}) \ni 0$, which expresses that $\bar{x}$ is a minimizer of f.

□

Remark 4.4. (a) As mentioned in the introduction, many functions encountered in finite-dimensional applications are of semi-algebraic (or tame) nature and are thus KL functions. So are in particular convex functions: this fact was a strong motivation for the above result. (b) Building a convex function that does not satisfy the Kurdyka-Łojasiewicz property is not easy. It is however possible to do so in dimension 2 (see [18]), but such functions must somehow have an highly oscillatory collection of sublevel sets (a behavior which is unlikely as far as applications are concerned).

5 Inexact forward-backward algorithm

Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a proper lower semicontinuous function which is bounded from below, and which satisfies the Kurdyka-Łojasiewicz property.
We assume that f is a structured function that can be split as

$$f = g + h \tag{46}$$

where $h : \mathbb{R}^n \to \mathbb{R}$ is a C^1 function whose gradient ∇h is Lipschitz continuous. The Lipschitz constant of ∇h is denoted by L. This kind of structured problem occurs frequently, see for instance [24, 6] and Example 5.4.

We consider sequences generated according to the following algorithm:

Algorithm 3: Take $a, b > 0$ with $a > L$. Take $x^0 \in \operatorname{dom} g$.
For $k = 0, 1, \ldots$, find $x^{k+1} \in \mathbb{R}^n$, $w^{k+1} \in \mathbb{R}^n$ such that

$$g(x^{k+1}) + \langle x^{k+1} - x^k, \nabla h(x^k) \rangle + \frac{a}{2}\|x^{k+1} - x^k\|^2 \leq g(x^k); \tag{47}$$

$$v^{k+1} \in \partial g(x^{k+1}); \tag{48}$$

$$\|v^{k+1} + \nabla h(x^k)\| \leq b\|x^{k+1} - x^k\|. \tag{49}$$

This section is divided into three distinct parts. In a first part, we recall what is the classical forward-backward algorithm and explain how Algorithm 3 provides an inexact version of the latter; the special case of projection methods is also discussed. In a second part, we provide a general convergence result for KL functions. We end this section by providing illustrations of our results through problems coming from compressive sensing, and hard-constrained feasibility problems.

5.1 The forward-backward splitting algorithm for nonconvex functions

Let us further assume that g is bounded from below. Being given a sequence of positive parameters γ_k that satisfies

$$0 < \underline{\gamma} < \gamma_k < \overline{\gamma} < \frac{1}{L}$$

where $\underline{\gamma}$ and $\overline{\lambda}$ are given thresholds, the forward-backward splitting algorithm reads

$$x^{k+1} \in \mathrm{prox}_{\gamma_k g}(x^k - \gamma_k \nabla h(x^k)). \tag{50}$$

An important observation here is that the sequence is not uniquely defined since $\mathrm{prox}_{\gamma_k g}$ may be multivalued; a surprising fact is that this freedom in the choice of the sequence does not impact the convergence properties of the algorithm (see Theorem 5.1).

Let us show how this algorithm fits into the general framework of Algorithm 3. By definition of the proximal operator we have

$$\gamma_k g(x^{k+1}) + \frac{1}{2}\|x^{k+1} - x^k + \gamma_k \nabla h(x^k)\|^2 \leq \gamma_k g(x^k) + \frac{1}{2}\|\gamma_k \nabla h(x^k)\|^2,$$

which after simplification gives

$$\gamma_k g(x^{k+1}) + \frac{1}{2}\|x^{k+1} - x^k\|^2 + \gamma_k \langle \nabla h(x^k), x^{k+1} - x^k \rangle \leq \gamma_k g(x^k).$$

Thus

$$g(x^{k+1}) + \langle \nabla h(x^k), x^{k+1} - x^k \rangle + \frac{1}{2\overline{\gamma}}\|x^{k+1} - x^k\|^2 \leq g(x^k),$$

so that the sufficient-decrease condition (47) holds (that's where we precisely use $\overline{\gamma} < \frac{1}{L}$). Writing down the optimality condition yields

$$\gamma_k v^{k+1} + \gamma_k \nabla h(x^k) + x^{k+1} - x^k = 0$$

where $v^{k+1} \in \partial g(x^{k+1})$. Dividing by γ_k, we end up with

$$\begin{aligned} \|v^{k+1} + \nabla h(x^k)\| &= \frac{1}{\gamma_k}\|x^{k+1} - x^k\| \\ &\leq \frac{1}{\underline{\gamma}}\|x^{k+1} - x^k\|, \end{aligned}$$

which is the inexact optimality conditions announced in (49).

As for the proximal algorithm, the inexact version offers some flexibility in the choice of x^{k+1} by relaxing both the descent condition and the optimality conditions.

Gradient projection algorithm Let us specialize the forward-backward splitting algorithm to functions of the form $i_C + h$ (where C is a nonempty closed subset of $\mathbb{R}^n$). For all positive λ, we have the elementary equality

$$\mathrm{prox}_{\lambda i_C} x = P_C(x).$$

We thus find the nonconvex nonsmooth gradient-projection method

$$x^{k+1} \in P_C(x^k - \gamma_k \nabla h(x^k)). \tag{51}$$

5.2 Convergence of an inexact forward-backward splitting algorithm

Let us now return to the general inexact forward-backward splitting Algorithm 3, and show the following convergence result.

Theorem 5.1 (Nonconvex nonsmooth forward-backward splitting)**.** *Let $f = g + h : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a proper lower semicontinuous KL function which is bounded from below. Assume further that $h : \mathbb{R}^n \to \mathbb{R}$ is finite valued, differentiable, has a L-Lipschitz continuous gradient, and that the restriction of g to its domain is continuous.*
If $(x^k)_{k\in\mathbb{N}}$ is a bounded sequence generated by Algorithm 3*, then it converges to some critical point of $f = g + h$.*
Moreover, the sequence $(x^k)_{k\in\mathbb{N}}$ has a finite length, i.e. $\sum_k \|x^{k+1} - x^k\| < +\infty$.

Proof. Using the descent lemma for the C^1 function h at x^{k+1} and x^k, and the sufficient decrease property (47) of Algorithm 3, we obtain

$$\begin{aligned} g(x^{k+1}) + h(x^{k+1}) + \frac{a-L}{2}\|x^{k+1} - x^k\|^2 \leq & g(x^{k+1}) + h(x^k) \\ & + \langle x^{k+1} - x^k, \nabla h(x^k)\rangle + \frac{a}{2}\|x^{k+1} - x^k\|^2 \\ \leq & g(x^k) + h(x^k). \end{aligned}$$

Therefore, setting

$$\tilde{a} = a - L > 0$$

we have

$$f(x^{k+1}) + \frac{\tilde{a}}{2}\|x^{k+1} - x^k\|^2 \leq f(x^k). \tag{52}$$

Define

$$w^{k+1} = v^{k+1} + \nabla h(x^{k+1}).$$

The classical derivation rule for the sum, see [51], and property (48) of Algorithm 3 yield

$$w^{k+1} \in \partial f(x^{k+1}). \tag{53}$$

Moreover, by property (49) of Algorithm 3, and the triangle inequality, we obtain

$$\begin{aligned} \|w^{k+1}\| \leq & \|v^{k+1} + \nabla h(x^k)\| + \|\nabla h(x^{k+1}) - \nabla h(x^k)\| \\ \leq & b\|x^{k+1} - x^k\| + L\|x^{k+1} - x^k\|. \end{aligned}$$

We are precisely in the case which has been examined in Theorem 4.2 (continuous functions on their domain). □

Remark 5.2. (a) For the exact forward-backward splitting algorithm the continuity assumption concerning g is useless. Indeed in that case, we have for all u in $\mathbb{R}^n$

$$\gamma_k g(x^{k+1}) + \frac{1}{2}\|x^{k+1} - x^k + \gamma_k \nabla h(x^k)\|^2 \leq \gamma_k g(u) + \frac{1}{2}\|u - x^k + \gamma_k \nabla h(x^k)\|^2,$$

so that

$$g(x^{k+1}) + \frac{1}{2\gamma_k}\|x^{k+1} - x^k\|^2 + \langle x^{k+1} - x^k, \nabla h(x^k)\rangle \leq g(u) + \frac{1}{2\gamma_k}\|u - x^k\|^2 + \langle u - x^k, \nabla h(x^k)\rangle. \tag{54}$$

Let x^{k_j} be a subsequence of x^k which converges to $\bar{x}$. Take $u = \bar{x}$, $k = k_j$ in (54) and let $j \to +\infty$. Since $x^{k+1} - x^k \to 0$, we obtain

$$\limsup_{k \to +\infty} g(x^{k+1}) \leq g(\bar{x}),$$

and, since g is lower semicontinuous, $\lim g(x^{k_j}) = g(\bar{x})$. The end of the proof is the same that the one of Theorem 5.1.

(b) Forward-backward splitting algorithms have many applications to parallel splitting of coupled systems. For applications involving monotone operators one may consult [6].

An important consequence of the above result is a general convergence result for gradient projection methods.

Theorem 5.3 (Nonconvex gradient projection method)**.** *Let $h : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function whose gradient is L-Lipschitz continuous, and C a nonempty closed subset of $\mathbb{R}^n$. Being given $\epsilon \in (0, \frac{1}{2L})$ and a sequence of stepsizes γ_k such that $\epsilon < \gamma_k < \frac{1}{L} - \epsilon$, we consider a sequence $(x^k)_{k \in \mathbb{N}}$ that complies with*

$$x^{k+1} \in P_C(x^k - \gamma_k \nabla h(x^k)), \text{ with } x^0 \in C.$$

If the function $h + i_C$ is a KL function and if $(x^k)_{k \in \mathbb{N}}$ is bounded, then the sequence $(x^k)_{k \in \mathbb{N}}$ converges to a point x^* in C such that

$$\nabla h(x^*) + N_C(x^*) \ni 0.$$

Proof. It is a direct consequence of Remark 5.2 (a). □

As mentioned in the Introduction and in Section 2.2, the assumption that $h + i_C$ is a KL function is very general. For instance, when h is C^1 semi-algebraic and C is a nonempty closed semi-algebraic set, $h + i_C$ is a KL function and the above result applies. Let us also emphasize here that our convergence result, contrary to those of Theorem 3.5 and [41], *do not rely* on any regularity properties of the set C (in the sense of variational analysis). In particular, C does not need to be prox-regular so that the projection mapping may be multi-valued in any neighborhood of C.

5.3 Examples

Example 5.4 (Forward-backward splitting for compressive sensing)**.** (a) The central issue in compressive sensing is to recover sparse solutions of under-determined linear systems (see [28]). The model problem is the following

$$(P) \quad \min\{\|x\|_0 : Ax = b\}$$

where $\|\cdot\|_0$ is the counting norm (or the ℓ^0 norm), $A \neq 0$ is an $m \times n$ real matrix and $b \in \mathbb{R}^m$. We recall that for x in $\mathbb{R}^n$, $\|x\|_0$ is the number of nonzero components of x.

As in [11], we proceed in the spirit of Tikhonov regularization for the least squares method. Fix a parameter $\lambda > 0$. We aim at solving the nonsmooth nonconvex problem:

$$(P') \quad \min\{\lambda\|x\|_0 + \frac{1}{2}\|Ax - b\|^2\}.$$

If we set $g(x) = \lambda\|x\|_0$ and $h(x) = \frac{1}{2}\|Ax - b\|^2$, it is straightforward to check that the topological assumptions of Remark 5.2(a) are satisfied (observe indeed that $\|\cdot\|_0$ is lower semicontinuous). To see that $g + h$ is a KL function, we simply note that h is a polynomial function and that $\|\cdot\|_0$ has a piecewise linear graph, hence the sum $g + h$ is semi-algebraic. Consider now the proximal operator $\mathrm{prox}_{\gamma\lambda\|\cdot\|_0}$ ([1]). When $n = 1$, the counting norm is denoted by $|\cdot|_0$; in that case one easily establishes that

$$\mathrm{prox}_{\gamma\lambda|\cdot|_0} u = \begin{cases} u & \text{if } |u| > \sqrt{2\gamma\lambda} \\ \{0, u\} & \text{if } |u| = \sqrt{2\gamma\lambda} \\ 0 & \text{otherwise.} \end{cases}$$

When n is arbitrary, trivial algebraic manipulations yield, with $u = (u_1, \ldots, u_n) \in \mathbb{R}^n$,

$$\mathrm{prox}_{\gamma\lambda\|\cdot\|_0} u = (\mathrm{prox}_{\lambda\gamma|\cdot|_0} u_1, \ldots, \mathrm{prox}_{\gamma\lambda|\cdot|_0} u_n),$$

and thus $\mathrm{prox}_{\gamma\lambda\|\cdot\|_0}$ is a perfectly known object.

Let $\|\cdot\|_F$ denote the Frobenius norm in $\mathbb{R}^{m\times n}$. Applying the previous result (Remark 5.2(a)) to the bounded sequences generated by the thresholding process

$$x^{k+1} \in \mathrm{prox}_{\gamma_k\lambda\|\cdot\|_0}\left(x^k - \gamma_k(A^TAx^k - A^Tb)\right)$$

where $0 < \underline{\gamma} < \gamma_k < \overline{\gamma} < \frac{1}{||A^TA||_F}$, shows that the sequence $(x^k)_{k\in\mathbb{N}}$ converges to a critical point of $\lambda\|x\|_0 + \frac{1}{2}\|Ax - b\|^2$, *i.e.* towards a point x^* that satisfies

$$(A^TAx^*)_i = (A^Tb)_i,$$

for all i such that $\bar{x}_i^* \neq 0$.

As mentioned in the introduction, these results offer a complementary view to the theoretical develoments of [11, 12]. They also provide at the same time a very general convergence result which can be immediately generalized to compressive sensing problems involving semi-algebraic or real-analytic nonlinear measurements.

(b) Alternative approaches to (P) are based on the following approximation

$$(P'') \quad \min\{\lambda\|x\|_p + \frac{1}{2}\|Ax - b\|^2\},$$

where p is in $(0, 1)$ and $\|x\|_p = \sum_1^n |x_i|^p$ (see [20]). Some encouraging numerical results have been reported in [20]. In [19] some theoretical results in the framework of Hilbert spaces are announced but, even when the space is finite dimensional, no convergence result and no estimate result are provided.

Using the separable structure of $\|\cdot\|_p$, the computation of the proximal operator $\mathrm{prox}_{\gamma\lambda\|\cdot\|_p}$ can be reduced to the one dimensional minimization problem: for $u \in \mathbb{R}$, find x solution of

$$\min\left\{2\gamma\lambda|x|^p + (x - u)^2 : x \in \mathbb{R}\right\},$$

that can be solved numerically by standard methods. Thus the forward-backward splitting algorithm may be run in a simple way. To obtain convergence, the only nontrivial fact that

[1] Recall that $\mathrm{prox}_{\gamma\lambda\|\cdot\|_0} = \mathrm{prox}_{(\gamma\lambda)\|\cdot\|_0} = \mathrm{prox}_{\gamma(\lambda\|\cdot\|_0)}$

has to be checked is that $f = \lambda\|x\|_p + \frac{1}{2}\|Ax - b\|^2$ is a KL function. For this, we recall that there exists a (polynomially bounded) o-minimal structure that contains the family of functions $\{x^\alpha : x > 0,\ \alpha \in \mathbb{R}\}$ and restricted analytic functions (see [30, Example (5), p. 505 and Property 5.2 p. 513]). As a consequence, the results of [17] apply and f is a KL function with a desingularizing function of the form $\varphi(s) = cs^\theta$ where $c > 0,\ \theta \in [0, 1)$. Hence the previous convergence and estimate results apply to the algorithm

$$x^{k+1} \in \text{prox}_{\lambda\gamma_k\|\cdot\|_p}(x^k - \gamma_k(A^TAx^k - A^Tb)),$$

and to its inexact counterparts (note that $g(\cdot) = \|\cdot\|_p$ is continuous and that γ_k is taken as in remark (a) above).

Example 5.5 (Hard-constrained feasibility problems)**.** Let $F, F_1, \ldots, F_p$ be a finite collection of nonempty closed subsets of $\mathbb{R}^n$, and assume that $F_1, \ldots, F_p$ are convex sets. *The hard constraint F is not supposed to be convex.* We consider the following minimization problem

$$\min\left\{\frac{1}{2}\sum_{i=1}^{p}\omega_i\text{dist}\,(x, F_i)^2 : x \in F\right\},$$

where ω_i are positive constants such that $\sum_i \omega_i = 1$. By applying the forward-backward splitting algorithm to this problem, we aim at finding a point which satisfies the hard constraints modelled by F, while the other constraints are satisfied in a possibly weaker sense (see [24] and references therein). Set

$$h(x) = \frac{1}{2}\sum_{i=1}^{p}\omega_i\text{dist}\,(x, F_i)^2.$$

By a standard convex analysis result, each function $h_i(x) = \frac{1}{2}\text{dist}\,(x, F_i)^2$ is C^1 convex, and its gradient, equal to $\nabla h_i(x) = x - P_{F_i}(x)$, is Lipschitz continuous with Lipschitz constant equal to 1. By convex combination, the same property holds true for h, and we can take $L = 1$ as a Lipschitz constant of ∇h.
Thus the forward-backward splitting algorithm (gradient-projection) (50) reads:

Take $0 < \underline{\theta} < \overline{\theta} < 1$. Take $x^0 \in \mathbb{R}^n$. For $k = 0, 1, ...,$

$$x^{k+1} \in P_F\left((1 - \theta_k)x^k + \theta_k\sum_{i=1}^{p}P_{F_i}(x^k)\right), \tag{55}$$

where $\theta_k \in [\underline{\theta}, \overline{\theta}]$.

Let us consider successively the convergence properties of this algorithm in two different situations, which are based respectively on the concepts of semi-algebraic sets, and linear regular intersection.

Theorem 5.6. *Assume that the sets $F, F_1, \ldots, F_p$ are semi-algebraic. Let $(x^k)_{k\in\mathbb{N}}$ be a sequence generated by the forward-backward splitting algorithm (55). If $(x^k)_{k\in\mathbb{N}}$ is bounded, and x^0 is sufficiently close to the intersection of the sets $F, F_1, \ldots, F_p$, then the sequence $(x^k)_{k\in\mathbb{N}}$ converges to a point which lies in the intersection of the sets $F, F_1, \ldots, F_p$.*

Proof. The proof relies on the fact that the underlying function

$$f(x) = i_F(x) + \frac{1}{2}\sum_{i=1}^{p} \omega_i \text{dist}\,(x, F_i)^2,\ x \in \mathbb{R}^n, \tag{56}$$

is a KL function. This follows immediately from the fact that the distance function to a semi-algebraic set is semi-algebraic (see Lemma 2.3) and hence satisfies KL. Then apply Theorem 5.1 to obtain the finite length property and the convergence of the sequence $(x^k)_{k\in\mathbb{N}}$ to a critical point of f. Then by direct application of the local convergence to global minima Theorem 2.12, we obtain the convergence of the sequence $(x^k)_{k\in\mathbb{N}}$ to a point which lies in the intersection of the sets $F, F_1, \ldots, F_p$. □

Let us now consider the KL analysis in the regular intersection case (see definition in Remark 3.6). To this end, we will use the following result [41, Proposition 8.5] (based itself on a characterization given in [36]).

Lemma 5.7. *Let $C_1, \ldots, C_m$ be closed subsets of $\mathbb{R}^n$ whose intersection is nonempty. Let $\bar{x} \in \cap_i C_i$. Assume that the intersection of $C_1, \ldots, C_m$ is linearly regular at $\bar{x}$. Then, there exists a positive constant α such that for each x sufficiently close to $\bar{x}$, we have:*

$$\alpha\sqrt{\sum_{i=1}^{m} \|y_i\|^2} \leq \|\sum_{i=1}^{m} y_i\|,\ \ \forall (y_1, \ldots, y_m) \in N_{F_1}(x) \times \ldots \times N_{F_m}(x). \tag{57}$$

We shall see below that this property entails that the function

$$\tilde{f}(x) = i_F(x) + \frac{1}{2}\sum_{i=1}^{p} \text{dist}\,(x, F_i)^2,\ x \in \mathbb{R}^n, \tag{58}$$

satisfies the KL inequality. Thus, in that case, we are led to consider the algorithm obtained by applying to $\tilde{f}$ the forward-backward splitting algorithm (50), which gives:

Take $0 < \underline{\theta} < \overline{\theta} < \frac{1}{p}$. Take $x^0 \in \mathbb{R}^n$. For $k = 0, 1, ...,$

$$x^{k+1} \in P_F\left((1-\theta_k)x^k + \theta_k \sum_{i=1}^{p} P_{F_i}(x^k)\right), \tag{59}$$

where $\theta_k \in [\underline{\theta}, \overline{\theta}]$.

Theorem 5.8. *Assume that the sets $F, F_1, \ldots, F_p$ have a linearly regular intersection around a point $\bar{x}$ and that one of them is a compact set. If x^0 is sufficiently close to $\bar{x}$, then the sequence $(x^k)_{k\in\mathbb{N}}$ generated by the algorithm (59) converges to a point which lies in the intersection of the sets $F, F_1, \ldots, F_p$.*

Proof. The convergence proof can be obtained like in Theorem 5.1 by using Theorem 2.12. We simply need to verify that the function $\tilde{f}$, as defined by (58), satisfies the KL inequality. Let K be a compact neighborhood of $\bar{x}$ on which (57) holds. Take x in K; we have

$$\partial\tilde{f}(x) = N_F(x) + \sum_{i=1}^{p}(x - P_{F_i}(x)).$$

For each $i = 1, \ldots, p$, set $y_i = (x - P_{Fi}(x))$ and observe that $y_i \in N_{F_i}(x)$. If x is in $\operatorname{dom} \partial \tilde{f}$, use Lemma 5.7 and inequality (57) to obtain

$$\begin{aligned}
\operatorname{dist}(0, \partial \tilde{f}(x)) &= \min\{\|z + \sum_{i=1}^{m} y_i\| : z \in N_F(x)\} \\
&\geq \alpha \min\{\sqrt{\|z\|^2 + \sum_{i=1}^{m} \|y_i\|^2} : z \in N_F(x)\} \\
&\geq \alpha \sqrt{\sum_{i=1}^{m} \|y_i\|^2} \\
&\geq c \tilde{f}(x)^{\frac{1}{2}}
\end{aligned}$$

where c is a positive constant. This shows that $\tilde{f}$ is a KL function. □

6 An inexact regularized Gauss-Seidel method

Fix an integer $p \geq 2$, and let $n_1, \ldots, n_p$ be positive integers. The current vector x belongs to the product space $\mathbb{R}^{n_1} \times \ldots \times \mathbb{R}^{n_p}$, it is denoted by $x = (x_1, \ldots, x_p)$, where each x_i belongs to $\mathbb{R}^{n_i}$. We are concerned with the minimization of functions $f : \mathbb{R}^{n_1} \times \ldots \times \mathbb{R}^{n_p} \to \mathbb{R} \cup \{+\infty\}$ having the following structure

$$f(x) = Q(x_1, \ldots, x_p) + \sum_{i=1}^{p} f_i(x_i), \tag{60}$$

where Q is a C^1 function with locally Lipschitz continuous gradient, and $f_i : \mathbb{R}^{n_i} \to \mathbb{R} \cup \{+\infty\}$ is a proper lower semicontinuous function, $i = 1, 2, ..., p$.

For each i in $\{1, \ldots, p\}$, we consider a bounded sequence of symmetric positive definite matrices $(B_i^k)_{k \in \mathbb{N}}$ of size n_i. We assume that the eigenvalues of the matrices $\{B_i^k : k \in \mathbb{N},\ i \in \{1, \ldots, p\}\}$ are bounded away from zero.
Our model algorithm is the following.

A proximal modification of the Gauss-Seidel method (see [8])
Take a starting point $x^0 = (x_1^0, \ldots, x_p^0)$ in $\mathbb{R}^{n_1} \times \ldots \times \mathbb{R}^{n_p}$ and consider the alternating minimizing procedure.
For x^k being given in $\mathbb{R}^{n_1} \times \ldots \times \mathbb{R}^{n_p}$ construct x^{k+1} as follows

$$x_1^{k+1} \in \operatorname{argmin}\{f(\mathbf{u}_1, x_2^k, \ldots, x_p^k) + \frac{1}{2}\langle B_1^k(\mathbf{u}_1 - x_1^k), u_1 - x_1^k\rangle : \mathbf{u}_1 \in \mathbb{R}^{n_1}\}. \tag{61}$$

Successively for $i = 2, \ldots, p-1$:

$$x_i^{k+1} \in \operatorname{argmin}\{f(x_1^{k+1}, \ldots, x_{i-1}^{k+1}, \mathbf{u}_i, x_{i+1}^k, \ldots) + \frac{1}{2}\langle B_i^k(\mathbf{u}_i - x_i^k), \mathbf{u}_i - x_i^k\rangle : \mathbf{u}_i \in \mathbb{R}^{n_i}\}; \tag{62}$$

$$x_p^{k+1} \in \operatorname{argmin}\{f(x_1^{k+1}, \ldots, x_{p-1}^{k+1}, \mathbf{u}_p) + \frac{1}{2}\langle B_p^k(\mathbf{u}_p - x_p^k), \mathbf{u}_p - x_p^k\rangle : \mathbf{u}_p \in \mathbb{R}^{n_p}\}. \tag{63}$$

Set $x^{k+1} = (x_1^{k+1}, \ldots, x_p^{k+1})$.

Remark 6.1. When $B_i^k = 0$ for all integers i and k, which is not allowed in our framework, one recovers the classical Gauss-Seidel model. When $B_i^k = \alpha_k I$ where α_k is a positive real number and I is the identity matrix, we recover the exact methods studied in [8, 31, 5].

Let us now introduce an inexact version of the above alternating method.

Algorithm 4 Take $0 < \underline{\lambda} < \overline{\lambda} < \infty$.
For each i in $\{1, \ldots, p\}$, take a sequence of symmetric positive definite matrices $(A_i^k)_{k\in\mathbb{N}}$ of size n_i such that the eigenvalues of each A_i^k ($k \in \mathbb{N}$, $i \in \{1, \ldots, p\}$) lie in $[\underline{\lambda}, \overline{\lambda}]$.
Take some positive parameters b_i ($i = 1, \ldots, p$).
Take a starting point $x^0 = (x_1^0, \ldots, x_p^0)$ in $\mathbb{R}^{n_1} \times \ldots \times \mathbb{R}^{n_p}$,.
For $k = 0, 1, \ldots$, find x^{k+1} and $v^{k+1} \in \mathbb{R}^{n_1} \times \ldots \times \mathbb{R}^{n_p}$ such that

$$f_i(x_i^{k+1}) + Q(x_1^{k+1}, \ldots, x_{i-1}^{k+1}, x_i^{k+1}, \ldots, x_p^k) + \frac{1}{2}\langle A_i^k(x_i^{k+1} - x_i^k), x_i^{k+1} - x_i^k\rangle$$
$$\leq f_i(x_i^k) + Q(x_1^{k+1}, \ldots, x_{i-1}^{k+1}, x_i^k, \ldots, x_p^k); \tag{64}$$
$$v_i^{k+1} \in \partial f_i(x_i^{k+1}); \tag{65}$$
$$\|v_i^{k+1} + \nabla_{x_i} Q(x_1^{k+1}, \ldots, x_i^{k+1}, x_{i+1}^k, \ldots, x_p^k)\| \leq b_i \|x_i^{k+1} - x_i^k\|, \tag{66}$$

where i ranges over $\{1, \ldots, p\}$.

Elementary computations show that the model algorithm (61)-(62)-(63) is a special instance of Algorithm 4.

Convergence analysis of the regularized Gauss-Seidel method
Define

$$w^{k+1} = (v_i^{k+1} + \nabla_{x_i} Q(x_1^{k+1}, \ldots, x_p^{k+1}))_{i=1,\ldots,p} \in \mathbb{R}^{n_1} \times \ldots \times \mathbb{R}^{n_p},$$

Using the differentiation rules for separable functions, we obtain

$$w^{k+1} \in \partial f(x^{k+1}). \tag{67}$$

Assume that the sequence $(x^k)_{k\in\mathbb{N}}$ is bounded, and denote by L the Lipschitz constant of ∇Q on a product of balls $\bar{B}_1 \times \ldots \times \bar{B}_p$ containing the sequence $(x^k)_{k\in\mathbb{N}}$. For all $i = 1, \ldots, p$, we have

$$\begin{aligned}
&\|v_i^{k+1} + \nabla_{x_i} Q(x_1^{k+1}, \ldots, x_p^{k+1})\| \\
&\leq \|v_i^{k+1} + \nabla_{x_i} Q(x_1^{k+1}, \ldots, x_i^{k+1}, x_{i+1}^k, \ldots, x_p^k)\| \\
&+ \|\nabla_{x_i} Q(x_1^{k+1}, \ldots, x_i^{k+1}, x_{i+1}^k, \ldots, x_p^k) - \nabla_{x_i} Q(x_1^{k+1}, \ldots, x_p^{k+1})\| \\
&\leq b_i \|x_i^{k+1} - x_i^k\| + L\|x^{k+1} - x^k\|.
\end{aligned}$$

Therefore, for some $M > 0$,

$$\|w^{k+1}\| \leq M\|x^{k+1} - x^k\|. \tag{68}$$

Summing inequalities of the type (64) from $i = 1$ to $i = p$, and using the inequalities $\underline{\lambda}\,\|u\|^2 \leq \langle A_i^k u, u\rangle$ for all integers i and k, we conclude that

$$f(x^{k+1}) + \underline{\lambda}\|x^{k+1} - x^k\|^2 \leq f(x^k).$$

We are in position to apply Theorem 2.9 to obtain

Theorem 6.2 (Proximal regularization of Gauss-Seidel method)**.** *Assume that f defined in (60) is a KL function which is bounded from below. Let $(x^k)_{k\in\mathbb{N}}$ be a sequence generated by* Algorithm 4. *If $(x^k)_{k\in\mathbb{N}}$ is bounded, then it converges to some critical point $\bar{x}$ of f. Moreover the sequence $(x^k)_{k\in\mathbb{N}}$ has a finite length, i.e. $\sum_k \|x^{k+1} - x^k\| < +\infty$.*

Remark 6.3 (Convex minimization)**.** Observe that this result is new even in the context of convex optimization where this problem was considered first (see the seminal work [8] and the recent study [4]). Indeed it allows both to choose a general smooth convex coupling term Q and to adapt the geometry of the proximal operators (through the choice of a metric A_i^k) to the geometry of the problem. Due to the fact that a convex function has at most one critical value, the bounded sequences generated by the above algorithms converge to a global minimizer.

7 Conclusion

Very often, iterative minimization algorithms rely on inexact solution of minimization subproblems, whose exact solution may be almost as difficult to obtain as the solution of the original minimization problem.

Even when the minimization subproblem can be solved with high accuracy, its solutions are mere approximations of the solution of the original problems. In these cases, over-solving the minimization subproblems would increase the computational burden of the method, and may slow down the final computation of a good approximation of the solution. On the other hand, under-solving the minimization subproblems may result in a breakdown of the algorithm, and convergence to a solution may be lost.

In this paper we gave theoretical basis for the application of numerical methods for minimizing a class of functions (which satisfies the KL inequality). In particular our abstract scheme was designed to handle relative errors because practical methods always involve numerical approximation, *e.g.*, the representation of a real number in floating points numbers with a fixed byte-length. We provided practical examples where the approximated solution of the minimization subproblems within the proposed error tolerance is feasible in a single step. Moreover, we also supplied stopping criteria for the solution of the minimization subproblems in general.

The computational implementation of the methods analyzed in this paper, as well as these stopping rules are topics for future research.

References

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

[2] Aragon, A., Dontchev, A. , Geoffroy, M., Convergence of the proximal point method for metrically regular mappings, ESAIM Proc., **17**, EDP Sci., Les Ulis, (2007), 1–8.

[3] Attouch, H., Bolte, J., On the convergence of the proximal algorithm for nonsmooth functions involving analytic features, Math. Program., Ser. B, **116** (2009), 5-16.

[4] Attouch, H., Bolte, J., Redont, P., Soubeyran, A., Alternating Proximal Algorithms for Weakly Coupled Convex Minimization Problems. Applications to Dynamical Games and PDE's, Journal of Convex Analysis **15** (2008), 485–506

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

[6] Attouch, H., Briceño-Arias, L.M., Combettes, P.L. A parallel splitting method for coupled monotone inclusions, SIAM J. Control Optim., **48**, no. 5, (2010), 3246-3270.

[7] Attouch, H., Soubeyran, A. Local search proximal algorithms as decision dynamics with costs to move, Set Valued and Variational Analysis, Online First, 12 May 2010.

[8] Auslender, A., Asymptotic properties of the Fenchel dual functional and applications to decomposition problems, J. Optim. Theory Appl., **73** (1992), 427–449.

[9] Beck, A., Teboulle, M.,, A Linearly Convergent Algorithm for Solving a Class of Nonconvex/Affine Feasibility Problems, July 2010, to appear in the book Fixed-Point Algorithms for Inverse Problems in Science and Engineering, part of the Springer Verlag series Optimization and Its Applications. Available online http://ie.technion.ac.il/Home/Users/becka.html

[10] Benedetti, R., Risler, J.-J., *Real Algebraic and Semialgebraic Sets*, Hermann, Éditeur des Sciences et des Arts, (Paris, 1990).

[11] Blumensath T., Davis, M. E., Iterative Thresholding for Sparse Approximations, J. of Fourier Anal. App. **14** (2008), 629–654.

[12] Blumensath T., Davis, M. E., Iterative hard thresholding for compressed sensing, App. Comput. Harmon. Anal., **27** (2009), 265–274.

[13] Bochnak, J., Coste, M., Roy, M.-F., *Real Algebraic Geometry*, (Springer, 1998).

[14] Bolte, J., Combettes, P.L., Pesquet, J.-C., Alternating proximal algorithm for blind image recovery, Proceedings of the IEEE International Conference on Image Processing. Hong-Kong, September 26-29, 2010.

[15] Bolte, J., Daniilidis, A. , Lewis, A., The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems, SIAM J. Optim., **17** , no. 4, (2006), 1205–1223.

[16] Bolte, J., Daniilidis, A., Lewis, A., A nonsmooth Morse-Sard theorem for subanalytic functions, J. Math. Anal. Appl., **321**, no. 2, (2006), 729–740.

[17] Bolte, J., Daniilidis, A., Lewis, A., Shiota, M., Clarke subgradients of stratifiable functions, SIAM J. Optim., **18**, no. 2, (2007), 556–572.

[18] Bolte, J., Daniilidis, A., Ley, O., Mazet, L., Characterizations of Łojasiewicz inequalities: Subgradient flows, talweg, convexity, Trans. Amer. Math. Soc., **362**, (2010), 3319-3363.

[19] Bredies, K., Lorenz, D.A., Minimization of non-smooth, non-convex functionals by iterative thresholding, preprint available at http://www.uni-graz.at/ bredies/publications.html

[20] Chartrand, R., Exact Reconstruction of Sparse Signals via Nonconvex Minimization, Signal Processing Letters IEEE, **14** (2007), 707–710.

[21] Chill, R., Jendoubi, M.A. Convergence to steady states in asymptotically autonomous semilinear evolution equations, Nonlinear Analysis, **53**, (2003), 1017–1039.

[22] Clarke, F.H., Ledyaev, Yu., Stern, R.I. , Wolenski, P.R., *Nonsmooth analysis and control theory*, Graduate texts in Mathematics **178**, (Springer-Verlag, New-York, 1998).

[23] Combettes, P.L., Quasi-Fejerian analysis of some optimization algorithms, in Inherently Parallel Algorithms in Feasibility and Optimization and Their Applications, (D. Butnariu, Y. Censor, and S. Reich, Eds.), New York: Elsevier, 2001, 115-152.

[24] Combettes, P.L., Wajs, V.R., Signal recovery by proximal forward-backward splitting., Multiscale Model. Simul., **4** (2005), 1168–1200.

[25] Coste, M., *An introduction to o-minimal geometry*, RAAG Notes, 81 p., Institut de Recherche Mathématiques de Rennes, November 1999.

[26] Curry, H.B., The method of steepest descent for non-linear minimization problems, Quart. Appl. Math., **2** (1944), 258–261.

[27] Palis, J.,& De Melo, W., *Geometric theory of dynamical systems. An introduction,* (Translated from the Portuguese by A. K. Manning), Springer-Verlag, New York-Berlin, 1982.

[28] Donoho, D. L., Compressed Sensing, IEEE Trans. Inform. Theory **4** (2006), 1289–1306.

[29] van den Dries, L., Tame topology and o-minimal structures. London Mathematical Society Lecture Note Series, **248**, Cambridge University Press, Cambridge, (1998) x+180 pp.

[30] van den Dries, L., & Miller, C., Geometric categories and o-minimal structures, *Duke Math. J.* **84** (1996), 497-540.

[31] Grippo, L., Sciandrone, M., Globally convergent block-coordinate techniques for unconstrained optimization, Optimization Methods and Software, **10** (4), (1999), 587–637.

[32] Hare, W., Sagastizábal, C. Computing proximal points of nonconvex functions, Math. Program., **116** (2009), **1-2**, Ser. B, 221–258.

[33] Haraux, A., Jendoubi, M.A. Convergence of solutions of second-order gradient-like systems with analytic nonlinearities, J. Differential Equations, **144** (2), (1999), 313–320.

[34] Huang, S.-Z., Takač, P. Convergence in gradient-like systems which are asymptotically autonomous and analytic, Nonlinear Anal., Ser. A, Theory Methods, **46**, (2001), 675–698.

[35] Iusem A.N., Pennanen T., Svaiter, B.F. Inexact variants of the proximal point algorithm without monotonicity, SIAM Journal on Optimization, **13**, no. 4 (2003), 1894–1097.

[36] Kruger, A.Y., About regularity of collections of sets, Set Valued Analysis, **14**, (2006), 187–206.

[37] Kurdyka, K., On gradients of functions definable in o-minimal structures, Ann. Inst. Fourier, **48**, (1998), 769-783.

[38] Lageman, C., Pointwise convergence of gradient-like systems, Math. Nachr., **280**, (2007), no. 13-14, 1543-1558.

[39] Lewis, A.S., Active sets, nonsmoothness and sensitivity, SIAM Journal on Optimization, **13**, (2003), 702–725.

[40] Lewis, A.S., Malick, J., Alternating projection on manifolds, Mathematics of Operations Research, **33**, no. 1, (2008), 216-234.

[41] Lewis, A.S., Luke, D.R., Malick, J., Local linear convergence for alternating and averaged nonconvex projections., Found. Comput. Math. **9**, (2009), 485–513.

[42] Łojasiewicz, S., Une propriété topologique des sous-ensembles analytiques réels, in: *Les Équations aux Dérivées Partielles*, pp. 87–89, Éditions du centre National de la Recherche Scientifique, Paris 1963.

[43] Łojasiewicz, S., Sur la géométrie semi- et sous-analytique, Ann. Inst. Fourier **43**, (1993), 1575-1595.

[44] Mordukhovich, B., Maximum principle in the problem of time optimal response with nonsmooth constraints, J. Appl. Math. Mech., **40** (1976), 960–969 ; [translated from Prikl. Mat. Meh. 40 (1976), 1014–1023].

[45] Mordukhovich, B., *Variational analysis and generalized differentiation. I. Basic theory*, Grundlehren der Mathematischen Wissenschaften, **330**, Springer-Verlag, Berlin, 2006.

[46] Nesterov, Yu., Accelerating the cubic regularization of Newton's method on convex problems, Math. Program., **112** (2008), no. 1, Ser. B, 159–181.

[47] Nesterov, Yu., Nemirovskii, A., *Interior-point polynomial algorithms in convex programming*, SIAM Studies in Applied Mathematics, **13**, Philadelphia, PA, 1994.

[48] Pennanen, T., Local convergence of the proximal point algorithm and multiplier methods without monotonicity, Math. Oper. Res. **27**, (2002), 170–191 .

[49] Peypouquet, J., Sorin, S., Evolution equations for maximal monotone operators: asymptotic analysis in continuous and discrete time, J. Convex Analysis, **17**, (2010), 1113–1163.

[50] Poliquin, R.A., Rockafellar, R.T., Thibault, L., Local differentiability of distance functions, Trans. AMS, **352**, (2000), 5231–5249.

[51] Rockafellar, R.T. , Wets, R., *Variational Analysis*, Grundlehren der Mathematischen Wissenschaften, **317**, Springer, 1998.

[52] Simon, L., Asymptotics for a class of non-linear evolution equations, with applications to geometric problems, Ann. of Math., **118** (1983), 525–571.

[53] Solodov, M.V., Svaiter, B.F., A hybrid projection-proximal point algorithm, Journal of Convex Analysis, **6**, no. 1, (1999), 59–70.

[54] Solodov, M.V., Svaiter, B.F., A hybrid approximate extragradient-proximal point algorithm using the enlargement of a maximal monotone operator, Set-Valued Analysis, **7**, (1999), 323–345.

[55] Solodov, M.V., Svaiter, B.F., A unified framework for some inexact proximal point algorithms, Numerical Functional Analysis and Optimization, **22**, (2001), 1013-1035.

[56] Wright, S.J., Identifiable surfaces in constrained optimization. SIAM Journal on Control and Optimization, **31**, (1993), 1063-1079.