An inertial forward-backward algorithm for the minimization of the sum of two nonconvex functions

Radu Ioan Boţ [*] Ernö Robert Csetnek [†] Szilárd Csaba László [‡]

July 15, 2015

Abstract. We propose a forward-backward proximal-type algorithm with inertial/memory effects for minimizing the sum of a nonsmooth function with a smooth one in the nonconvex setting. Every sequence of iterates generated by the algorithm converges to a critical point of the objective function provided an appropriate regularization of the objective satisfies the Kurdyka-Łojasiewicz inequality, which is for instance fulfilled for semi-algebraic functions. We illustrate the theoretical results by considering two numerical experiments: the first one concerns the ability of recovering the local optimal solutions of nonconvex optimization problems, while the second one refers to the restoration of a noisy blurred image.

Key Words. nonsmooth optimization, limiting subdifferential, Kurdyka-Łojasiewicz inequality, Bregman distance, inertial proximal algorithm

AMS subject classification. 90C26, 90C30, 65K10

1 Introduction

Proximal-gradient splitting methods are powerful techniques used in order to solve optimization problems where the objective to be minimized is the sum of a finite collection of smooth and/or nonsmooth functions. The main feature of this class of algorithmic schemes is the fact that they access each function separately, either by a gradient step if this is smooth or by a proximal step if it is nonsmooth.

In the convex case (when all the functions involved are convex), these methods are well understood, see for example [8], where the reader can find a presentation of the most prominent methods, like the forward-backward, forward-backward-forward and the Douglas-Rachford splitting algorithms.

On the other hand, the nonconvex case is less understood, one of the main difficulties coming from the fact that the proximal point operator is in general not anymore single-valued. However, one can observe a considerable progress in this direction when the functions in the objective have the *Kurdyka-Łojasiewicz property* (so-called *KL functions*), as it

[*] University of Vienna, Faculty of Mathematics, Oskar-Morgenstern-Platz 1, A-1090 Vienna, Austria, email: radu.bot@univie.ac.at. Research partially supported by DFG (German Research Foundation), project BO 2516/4-1.

[†] University of Vienna, Faculty of Mathematics, Oskar-Morgenstern-Platz 1, A-1090 Vienna, Austria, email: ernoe.robert.csetnek@univie.ac.at. Research supported by DFG (German Research Foundation), project BO 2516/4-1.

[‡] Technical University of Cluj-Napoca, Department of Mathematics, 400027 Cluj-Napoca, Romania, e-mail: szilard.laszlo@math.utcluj.ro

is the case for the ones with different analytic features. This applies for both the forward-backward algorithm (see [14], [6]) and the forward-backward-forward algorithm (see [18]). We refer the reader also to [4, 5, 23, 25, 26, 34] for literature concerning proximal-gradient splitting methods in the nonconvex case relying on the *Kurdyka-Lojasiewicz property*.

A particular class of the proximal-gradient splitting methods are the ones with inertial/memory effects. These iterative schemes have their origins in the time discretization of some differential inclusions of second order type (see [1, 3]) and share the feature that the new iterate is defined by using the previous two iterates. The increasing interest in this class of algorithms is emphasized by a considerable number of papers written in the last fifteen years on this topic, see [1–3, 7, 15–22, 29, 30, 32, 35].

Recently, an inertial forward-backward type algorithm has been proposed and analyzed in [34] in the nonconvex setting, by assuming that the nonsmooth part of the objective function is convex, while the smooth counterpart is allowed to be nonconvex. It is the aim of this paper to introduce an inertial forward-backward algorithm in the full nonconvex setting and to study its convergence properties. The techniques for proving the convergence of the numerical scheme use the same three main ingredients, as other algorithms for nonconvex optimization problems involving KL functions. More precisely, we show a sufficient decrease property for the iterates, the existence of a subgradient lower bound for the iterates gap and, finally, we use the analytic features of the objective function in order to obtain convergence, see [6,14]. The *limiting (Mordukhovich) subdifferential* and its properties play an important role in the analysis. The main result of this paper shows that, provided an appropriate regularization of the objective satisfies the Kurdyka-Łojasiewicz property, the convergence of the inertial forward-backward algorithm is guaranteed. As a particular instance, we also treat the case when the objective function is semi-algebraic and present the convergence properties of the algorithm.

In the last section of the paper we consider two numerical experiments. The first one has an academic character and shows the ability of algorithms with inertial/memory effects to detect optimal solutions which are not found by the non-inertial versions (similar allegations can be found also in [34, Section 5.1] and [10, Example 1.3.9]). The second one concerns the restoration of a noisy blurred image by using a nonconvex misfit functional with nonconvex regularization.

2 Preliminaries

In this section we recall some notions and results which are needed throughout this paper. Let $\mathbb{N} = \{0, 1, 2, ...\}$ be the set of nonnegative integers. For $m \geq 1$, the Euclidean scalar product and the induced norm on $\mathbb{R}^m$ are denoted by $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$, respectively. Notice that all the finite-dimensional spaces considered in the manuscript are endowed with the topology induced by the Euclidean norm.

The *domain* of the function $f : \mathbb{R}^m \to (-\infty, +\infty]$ is defined by $\operatorname{dom} f = \{x \in \mathbb{R}^m : f(x) < +\infty\}$. We say that f is *proper* if $\operatorname{dom} f \neq \emptyset$. For the following generalized subdifferential notions and their basic properties we refer to [31, 36]. Let $f : \mathbb{R}^m \to (-\infty, +\infty]$ be a proper and lower semicontinuous function. If $x \in \operatorname{dom} f$, we consider the *Fréchet (viscosity) subdifferential* of f at x as the set

$$\hat{\partial} f(x) = \left\{ v \in \mathbb{R}^m : \liminf_{y \to x} \frac{f(y) - f(x) - \langle v, y - x \rangle}{\|y - x\|} \geq 0 \right\}.$$

For $x \notin \operatorname{dom} f$ we set $\hat{\partial} f(x) := \emptyset$. The *limiting (Mordukhovich) subdifferential* is defined at $x \in \operatorname{dom} f$ by

$$\partial f(x) = \{v \in \mathbb{R}^m : \exists x_n \to x, f(x_n) \to f(x) \text{ and } \exists v_n \in \hat{\partial} f(x_n), v_n \to v \text{ as } n \to +\infty\},$$

while for $x \notin \operatorname{dom} f$, one takes $\partial f(x) := \emptyset$.

Notice that in case f is convex, these notions coincide with the *convex subdifferential*, which means that $\hat{\partial} f(x) = \partial f(x) = \{v \in \mathbb{R}^m : f(y) \geq f(x) + \langle v, y - x \rangle \ \forall y \in \mathbb{R}^m\}$ for all $x \in \operatorname{dom} f$.

Notice the inclusion $\hat{\partial} f(x) \subseteq \partial f(x)$ for each $x \in \mathbb{R}^m$. We will use the following closedness criteria concerning the graph of the limiting subdifferential: if $(x_n)_{n\in\mathbb{N}}$ and $(v_n)_{n\in\mathbb{N}}$ are sequences in $\mathbb{R}^m$ such that $v_n \in \partial f(x_n)$ for all $n \in \mathbb{N}$, $(x_n, v_n) \to (x, v)$ and $f(x_n) \to f(x)$ as $n \to +\infty$, then $v \in \partial f(x)$.

The Fermat rule reads in this nonsmooth setting as: if $x \in \mathbb{R}^m$ is a local minimizer of f, then $0 \in \partial f(x)$. Notice that in case f is continuously differentiable around $x \in \mathbb{R}^m$ we have $\partial f(x) = \{\nabla f(x)\}$. Let us denote by

$$\operatorname{crit}(f) = \{x \in \mathbb{R}^m : 0 \in \partial f(x)\}$$

the set of *(limiting)-critical points* of f. Let us mention also the following subdifferential rule: if $f : \mathbb{R}^m \to (-\infty, +\infty]$ is proper and lower semicontinuous and $h : \mathbb{R}^m \to \mathbb{R}$ is a continuously differentiable function, then $\partial(f + h)(x) = \partial f(x) + \nabla h(x)$ for all $x \in \mathbb{R}^m$.

We turn now our attention to functions satisfying the *Kurdyka-Łojasiewicz property*. This class of functions will play a crucial role when proving the convergence of the proposed inertial algorithm. For $\eta \in (0, +\infty]$, we denote by Θ_η the class of concave and continuous functions $\varphi : [0, \eta) \to [0, +\infty)$ such that $\varphi(0) = 0$, φ is continuously differentiable on $(0, \eta)$, continuous at 0 and $\varphi'(s) > 0$ for all $s \in (0, \eta)$. In the following definition (see [5, 14]) we use also the *distance function* to a set, defined for $A \subseteq \mathbb{R}^m$ as $\operatorname{dist}(x, A) = \inf_{y\in A} \|x - y\|$ for all $x \in \mathbb{R}^m$.

Definition 1 (*Kurdyka-Łojasiewicz property*) Let $f : \mathbb{R}^m \to (-\infty, +\infty]$ be a proper and lower semicontinuous function. We say that f satisfies the *Kurdyka-Łojasiewicz (KL) property* at $\overline{x} \in \operatorname{dom} \partial f = \{x \in \mathbb{R}^m : \partial f(x) \neq \emptyset\}$ if there exist $\eta \in (0, +\infty]$, a neighborhood U of $\overline{x}$ and a function $\varphi \in \Theta_\eta$ such that for all x in the intersection

$$U \cap \{x \in \mathbb{R}^m : f(\overline{x}) < f(x) < f(\overline{x}) + \eta\}$$

the following inequality holds

$$\varphi'(f(x) - f(\overline{x})) \operatorname{dist}(0, \partial f(x)) \geq 1.$$

If f satisfies the KL property at each point in $\operatorname{dom} \partial f$, then f is called a *KL function*.

The origins of this notion go back to the pioneering work of Łojasiewicz [28], where it is proved that for a real-analytic function $f : \mathbb{R}^m \to \mathbb{R}$ and a critical point $\overline{x} \in \mathbb{R}^m$ (that is $\nabla f(\overline{x}) = 0$), there exists $\theta \in [1/2, 1)$ such that the function $|f - f(\overline{x})|\|\nabla f\|^{-1}$ is bounded around $\overline{x}$. This corresponds to the situation when $\varphi(s) = s^{1-\theta}$. The result of Łojasiewicz allows the interpretation of the KL property as a re-parametrization of the function values in order to avoid flatness around the critical points. Kurdyka [27] extended this property to differentiable functions definable in an o-minimal structure. Further extensions to the nonsmooth setting can be found in [5, 11–13].

One of the remarkable properties of the KL functions is their ubiquity in applications, according to [14]. To the class of KL functions belong semi-algebraic, real sub-analytic, semiconvex, uniformly convex and convex functions satisfying a growth condition. We refer the reader to [4–6, 11–14] and the references therein for more details regarding all the classes mentioned above and illustrating examples.

An important role in our convergence analysis will be played by the following uniformized KL property given in [14, Lemma 6].

Lemma 1 *Let $\Omega \subseteq \mathbb{R}^m$ be a compact set and let $f : \mathbb{R}^m \to (-\infty, +\infty]$ be a proper and lower semicontinuous function. Assume that f is constant on Ω and f satisfies the KL property at each point of Ω. Then there exist $\varepsilon, \eta > 0$ and $\varphi \in \Theta_\eta$ such that for all $\overline{x} \in \Omega$ and for all x in the intersection*

$$\{x \in \mathbb{R}^m : \operatorname{dist}(x, \Omega) < \varepsilon\} \cap \{x \in \mathbb{R}^m : f(\overline{x}) < f(x) < f(\overline{x}) + \eta\} \tag{1}$$

the following inequality holds

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

We close this section by presenting two convergence results which will play a determined role in the proof of the results we provide in the next section. The first one was often used in the literature in the context of Fejér monotonicity techniques for proving convergence results of classical algorithms for convex optimization problems or more generally for monotone inclusion problems (see [8]). The second one is probably also known, see for example [18].

Lemma 2 *Let $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$ be real sequences such that $b_n \geq 0$ for all $n \in \mathbb{N}$, $(a_n)_{n\in\mathbb{N}}$ is bounded below and $a_{n+1} + b_n \leq a_n$ for all $n \in \mathbb{N}$. Then $(a_n)_{n\in\mathbb{N}}$ is a monotonically decreasing and convergent sequence and $\sum_{n\in\mathbb{N}} b_n < +\infty$.*

Lemma 3 *Let $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$ be nonnegative real sequences, such that $\sum_{n\in\mathbb{N}} b_n < +\infty$ and $a_{n+1} \leq a \cdot a_n + b \cdot a_{n-1} + b_n$ for all $n \geq 1$, where $a \in \mathbb{R}$, $b \geq 0$ and $a + b < 1$. Then $\sum_{n\in\mathbb{N}} a_n < +\infty$.*

3 A forward-backward algorithm

In this section we present an inertial forward-backward algorithm for a fully nonconvex optimization problem and study its convergence properties. The problem under investigation has the following formulation.

Problem 1. Let $f : \mathbb{R}^m \to (-\infty, +\infty]$ be a proper, lower semicontinuous function which is bounded below and let $g : \mathbb{R}^m \to \mathbb{R}$ be a Fréchet differentiable function with Lipschitz continuous gradient, i.e. there exists $L_{\nabla g} \geq 0$ such that $\|\nabla g(x) - \nabla g(y)\| \leq L_{\nabla g}\|x - y\|$ for all $x, y \in \mathbb{R}^m$. We deal with the optimization problem

$$(P) \ \inf_{x\in\mathbb{R}^m} [f(x) + g(x)]. \tag{3}$$

In the iterative scheme we propose below, we use also the function $F : \mathbb{R}^m \to \mathbb{R}$, assumed to be $\sigma-$strongly convex, i.e. $F - \frac{\sigma}{2}\|\cdot\|^2$ is convex, Fréchet differentiable and such that ∇F is $L_{\nabla F}$-Lipschitz continuous, where $\sigma, L_{\nabla F} > 0$. The *Bregman distance* to F, denoted by $D_F : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$, is defined as

$$D_F(x,y) = F(x) - F(y) - \langle \nabla F(y), x - y\rangle \ \forall (x,y) \in \mathbb{R}^m \times \mathbb{R}^m.$$

Notice that the properties of the function F ensure the following inequalities

$$\frac{\sigma}{2}\|x-y\|^2 \leq D_F(x,y) \leq \frac{L_{\nabla F}}{2}\|x-y\|^2 \ \forall x,y \in \mathbb{R}^m. \tag{4}$$

We propose the following iterative scheme.

Algorithm 1. Choose $x_0, x_1 \in \mathbb{R}^m$, $\underline{\alpha}, \overline{\alpha} > 0$, $\beta \geq 0$ and the sequences $(\alpha_n)_{n\geq 1}, (\beta_n)_{n\geq 1}$ fulfilling

$$0 < \underline{\alpha} \leq \alpha_n \leq \overline{\alpha} \ \forall n \geq 1$$

and

$$0 \leq \beta_n \leq \beta \ \forall n \geq 1.$$

Consider the iterative scheme

$$(\forall n \geq 1) \ \ x_{n+1} \in \operatorname*{argmin}_{u\in\mathbb{R}^m} \left\{D_F(u,x_n) + \alpha_n\langle u, \nabla g(x_n)\rangle + \beta_n\langle u, x_{n-1} - x_n\rangle + \alpha_n f(u)\right\}. \tag{5}$$

Due to the subdifferential sum formula mentioned in the previous section, one can see that any sequence generated by this algorithm satisfies the relation

$$x_{n+1} \in (\nabla F + \alpha_n \partial f)^{-1}(\nabla F(x_n) - \alpha_n \nabla g(x_n) + \beta_n(x_n - x_{n-1})) \ \forall n \geq 1. \tag{6}$$

Further, since f is proper, lower semicontinuous and bounded from below and D_F is coercive in its first argument (that is $\lim_{\|x\|\to+\infty} D_F(x,y) = +\infty$ for all $y \in \mathbb{R}^m$), the iterative scheme is well-defined, meaning that the existence of x_n is guaranteed for each $n \geq 2$, since the objective function in the minimization problem to be solved at each iteration is coercive.

Remark 4 The condition that f should be bounded below is imposed in order to ensure that in each iteration one can choose at least one x_n (that is the argmin in (5) is nonempty). One can replace this requirement by asking that the objective function in the minimization problem considered in (5) is coercive and the theory presented below still remains valid. This observation is useful when dealing with optimization problems as the ones considered in Subsection 4.1.

Before proceeding with the convergence analysis, we discuss the relation of our scheme to other algorithms from the literature. Let us take first $F(x) = \frac{1}{2}\|x\|^2$ for all $x \in \mathbb{R}^m$. In this case $D_F(x,y) = \frac{1}{2}\|x-y\|^2$ for all $(x,y) \in \mathbb{R}^m \times \mathbb{R}^m$ and $\sigma = L_{\nabla F} = 1$. The iterative scheme becomes

$$(\forall n \geq 1) \ \ x_{n+1} \in \operatorname*{argmin}_{u\in\mathbb{R}^m} \left\{\frac{\|u - (x_n - \alpha_n \nabla g(x_n) + \beta_n(x_n - x_{n-1}))\|^2}{2\alpha_n} + f(u)\right\}. \tag{7}$$

A similar inertial type algorithm has been analyzed in [34], however in the restrictive case when f is convex. If we take in addition $\beta = 0$, which enforces $\beta_n = 0$ for all $n \geq 1$, then (7) becomes

$$(\forall n \geq 1) \;\; x_{n+1} \in \underset{u \in \mathbb{R}^m}{\operatorname{argmin}} \left\{ \frac{\|u - (x_n - \alpha_n \nabla g(x_n))\|^2}{2\alpha_n} + f(u) \right\}, \tag{8}$$

the convergence of which has been investigated in [14] in the full nonconvex setting. Notice that forward-backward algorithms with variable metrics for KL functions have been proposed in [23, 25].

On the other hand, if we take $g(x) = 0$ for all $x \in \mathbb{R}^m$, the iterative scheme in (7) becomes

$$(\forall n \geq 1) \;\; x_{n+1} \in \underset{u \in \mathbb{R}^m}{\operatorname{argmin}} \left\{ \frac{\|u - (x_n + \beta_n(x_n - x_{n-1}))\|^2}{2\alpha_n} + f(u) \right\}, \tag{9}$$

which is a proximal point algorithm with inertial/memory effects formulated in the nonconvex setting designed for finding the critical points of f. The iterative scheme without the inertial term, that is when $\beta = 0$ and, so, $\beta_n = 0$ for all $n \geq 1$, has been considered in the context of KL functions in [4].

Let us mention that in the full convex setting, which means that f and g are convex functions, in which case for all $n \geq 2$, x_n is uniquely determined and can be expressed via the *proximal operator* of f, (7) can be derived from the iterative scheme proposed in [32], (8) is the classical forward-backward algorithm (see for example [8] or [24]) and (9) has been analyzed in [3] in the more general context of monotone inclusion problems.

In the convergence analysis of the algorithm the following result will be useful (see for example [33, Lemma 1.2.3]).

Lemma 5 *Let $g : \mathbb{R}^m \to \mathbb{R}$ be Fréchet differentiable with $L_{\nabla g}$-Lipschitz continuous gradient. Then*

$$g(y) \leq g(x) + \langle \nabla g(x), y - x \rangle + \frac{L_{\nabla g}}{2}\|y - x\|^2, \; \forall x, y \in \mathbb{R}^m.$$

Let us start now with the investigation of the convergence of the proposed algorithm.

Lemma 6 *In the setting of Problem 1, let $(x_n)_{n \in \mathbb{N}}$ be a sequence generated by Algorithm 1. Then one has*

$$(f + g)(x_{n+1}) + M_1\|x_n - x_{n+1}\|^2 \leq (f + g)(x_n) + M_2\|x_{n-1} - x_n\|^2 \;\; \forall n \geq 1,$$

where

$$M_1 = \frac{\sigma - \overline{\alpha} L_{\nabla g}}{2\overline{\alpha}} - \frac{\beta}{2\underline{\alpha}} \text{ and } M_2 = \frac{\beta}{2\underline{\alpha}}. \tag{10}$$

Moreover, for $0 < \underline{\alpha} \leq \overline{\alpha}$ and $\beta > 0$ satisfying

$$\sigma > \overline{\alpha} L_{\nabla g} + 2\beta \frac{\overline{\alpha}}{\underline{\alpha}}, \tag{11}$$

one has $M_1 > M_2$.

Proof. Let be $n \geq 1$ fixed. Due to (5) we have

$$\begin{aligned} &D_F(x_{n+1}, x_n) + \alpha_n \langle x_{n+1}, \nabla g(x_n) \rangle + \beta_n \langle x_{n+1}, x_{n-1} - x_n \rangle + \alpha_n f(x_{n+1}) \\ &\leq D_F(x_n, x_n) + \alpha_n \langle x_n, \nabla g(x_n) \rangle + \beta_n \langle x_n, x_{n-1} - x_n \rangle + \alpha_n f(x_n) \end{aligned}$$

or, equivalently,

$$D_F(x_{n+1}, x_n) + \langle x_{n+1} - x_n, \alpha_n \nabla g(x_n) - \beta_n (x_n - x_{n-1}) \rangle + \alpha_n f(x_{n+1}) \leq \alpha_n f(x_n). \quad (12)$$

On the other hand, by Lemma 5 we have

$$\langle \nabla g(x_n), x_{n+1} - x_n \rangle \geq g(x_{n+1}) - g(x_n) - \frac{L_{\nabla g}}{2} \|x_n - x_{n+1}\|^2.$$

At the same time

$$\langle x_{n+1} - x_n, x_{n-1} - x_n \rangle \geq -\left(\frac{1}{2} \|x_n - x_{n+1}\|^2 + \frac{1}{2} \|x_{n-1} - x_n\|^2 \right),$$

and from (4) we have

$$\frac{\sigma}{2} \|x_{n+1} - x_n\|^2 \leq D_F(x_{n+1}, x_n).$$

Hence, (12) leads to

$$(f+g)(x_{n+1}) + \frac{\sigma - L_{\nabla g}\alpha_n - \beta_n}{2\alpha_n} \|x_{n+1} - x_n\|^2 \leq (f+g)(x_n) + \frac{\beta_n}{2\alpha_n} \|x_{n-1} - x_n\|^2. \quad (13)$$

Obviously $M_1 = \frac{\sigma - L_{\nabla g}\overline{\alpha}}{2\overline{\alpha}} - \frac{\beta}{2\underline{\alpha}} \leq \frac{\sigma - L_{\nabla g}\alpha_n - \beta_n}{2\alpha_n}$ and $M_2 = \frac{\beta}{2\underline{\alpha}} \geq \frac{\beta_n}{2\alpha_n}$ thus,

$$(f+g)(x_{n+1}) + M_1 \|x_n - x_{n+1}\|^2 \leq (f+g)(x_n) + M_2 \|x_{n-1} - x_n\|^2$$

and the first part of the lemma is proved.

Finally, for $0 < \underline{\alpha} \leq \overline{\alpha}$ and $\beta > 0$ satisfying $\sigma > \overline{\alpha} L_{\nabla g} + 2\beta \frac{\overline{\alpha}}{\underline{\alpha}}$, one has that $M_1 > M_2$ and the proof is complete. ■

Remark 7 If $\underline{\alpha}$ and β are positive numbers such that $\sigma > \underline{\alpha} L_{\nabla g} + 2\beta$, then

$$\underline{\alpha} < \frac{\underline{\alpha}\sigma}{\underline{\alpha} L_{\nabla g} + 2\beta}.$$

By choosing

$$\underline{\alpha} \leq \overline{\alpha} < \frac{\underline{\alpha}\sigma}{\underline{\alpha} L_{\nabla g} + 2\beta},$$

relation (11) is satisfied.

On the other hand, if $\overline{\alpha}$ and β are positive numbers such that $\sigma > \overline{\alpha} L_{\nabla g} + 2\beta$, then

$$\frac{2\beta\overline{\alpha}}{\sigma - \overline{\alpha} L_{\nabla g}} < \overline{\alpha}.$$

By choosing

$$\frac{2\beta\overline{\alpha}}{\sigma - \overline{\alpha} L_{\nabla g}} < \underline{\alpha} \leq \overline{\alpha},$$

relation (11) is again satisfied.

Proposition 8 *In the setting of Problem 1, choose $\underline{\alpha}, \overline{\alpha}, \beta$ satisfying* (11) *and M_1, M_2 satisfying* (10). *Assume that $f + g$ is bounded from below. Then the following statements hold:*

(a) $\sum_{n\geq 1} \|x_n - x_{n-1}\|^2 < +\infty$*;*

(b) the sequence $((f + g)(x_n) + M_2\|x_{n-1} - x_n\|^2)_{n\geq 1}$ *is monotonically decreasing and convergent;*

(c) the sequence $((f + g)(x_n))_{n\in\mathbb{N}}$ *is convergent.*

Proof. For every $n \geq 1$, set $a_n = (f+g)(x_n)+M_2\|x_{n-1}-x_n\|^2$ and $b_n = (M_1-M_2)\|x_n - x_{n+1}\|^2$. Then obviously from Lemma 6 one has for every $n \geq 1$

$$a_{n+1} + b_n = (f + g)(x_{n+1}) + M_1\|x_n - x_{n+1}\|^2 \leq (f + g)(x_n) + M_2\|x_{n-1} - x_n\|^2 = a_n.$$

The conclusion follows now from Lemma 2. ■

Lemma 9 *In the setting of Problem 1, consider the sequences generated by Algorithm 1. For every $n \geq 1$ we have*

$$y_{n+1} \in \partial(f + g)(x_{n+1}), \tag{14}$$

where

$$y_{n+1} = \frac{\nabla F(x_n) - \nabla F(x_{n+1})}{\alpha_n} + \nabla g(x_{n+1}) - \nabla g(x_n) + \frac{\beta_n}{\alpha_n}(x_n - x_{n-1}).$$

Moreover,

$$\|y_{n+1}\| \leq \frac{L_{\nabla F} + \alpha_n L_{\nabla g}}{\alpha_n}\|x_n - x_{n+1}\| + \frac{\beta_n}{\alpha_n}\|x_n - x_{n-1}\| \ \forall n \geq 1 \tag{15}$$

Proof. Let us fix $n \geq 1$. From (6) we have that

$$\frac{\nabla F(x_n) - \nabla F(x_{n+1})}{\alpha_n} - \nabla g(x_n) + \frac{\beta_n}{\alpha_n}(x_n - x_{n-1}) \in \partial f(x_{n+1}),$$

or, equivalently,

$$y_{n+1} - \nabla g(x_{n+1}) \in \partial f(x_{n+1}),$$

which shows that $y_{n+1} \in \partial(f + g)(x_{n+1})$.

The inequality (15) follows now from the definition of y_{n+1} and the triangle inequality. ■

Lemma 10 *In the setting of Problem 1, choose $\underline{\alpha}, \overline{\alpha}, \beta$ satisfying* (11) *and M_1, M_2 satisfying* (10). *Assume that $f + g$ is coercive, i.e.*

$$\lim_{\|x\|\to+\infty} (f + g)(x) = +\infty.$$

Then any sequence $(x_n)_{n\in\mathbb{N}}$ generated by Algorithm 1 has a subsequence convergent to a critical point of $f + g$. Actually every cluster point of $(x_n)_{n\in\mathbb{N}}$ is a critical point of $f + g$.

Proof. Since $f+g$ is a proper, lower semicontinuous and coercive function, it follows that $\inf_{x\in\mathbb{R}^m}[f(x)+g(x)]$ is finite and the infimum is attained. Hence $f+g$ is bounded from below.

Let $(x_n)_{n\in\mathbb{N}}$ be a sequence generated by Algorithm 1. According to Proposition 8(b), we have

$$(f+g)(x_n) \leq (f+g)(x_n) + M_2\|x_n - x_{n-1}\|^2 \leq (f+g)(x_1) + M_2\|x_1 - x_0\|^2 \ \forall n \geq 1.$$

Since the function $f+g$ is coercive, its lower level sets are bounded, thus the sequence $(x_n)_{n\in\mathbb{N}}$ is bounded.

Let x be a cluster point of $(x_n)_{n\in\mathbb{N}}$. Then there exists a subsequence $(x_{n_k})_{k\in\mathbb{N}}$ such that $x_{n_k} \to x$ as $k \to +\infty$. We show that $(f+g)(x_{n_k}) \to (f+g)(x)$ as $k \to +\infty$ and that x is a critical point of $f+g$, that is $0 \in \partial(f+g)(x)$.

We show first that $f(x_{n_k}) \to f(x)$ as $k \to +\infty$. Since f is lower semicontinuous one has

$$\liminf_{k\to+\infty} f(x_{n_k}) \geq f(x).$$

On the other hand, from (5) we have for every $n \geq 1$

$$\begin{aligned} D_F(x_{n+1},x_n) + \alpha_n\langle x_{n+1}, \nabla g(x_n)\rangle + \beta_n\langle x_{n+1}, x_{n-1}-x_n\rangle + \alpha_n f(x_{n+1}) \leq \\ D_F(x,x_n) + \alpha_n\langle x, \nabla g(x_n)\rangle + \beta_n\langle x, x_{n-1}-x_n\rangle + \alpha_n f(x), \end{aligned}$$

which leads to

$$\begin{aligned} \frac{1}{\alpha_{n_k-1}}\left(D_F(x_{n_k},x_{n_k-1}) - D_F(x,x_{n_k-1})\right) + \\ \frac{1}{\alpha_{n_k-1}}\left(\langle x_{n_k}-x, \alpha_{n_k-1}\nabla g(x_{n_k-1}) - \beta_{n_k-1}(x_{n_k-1}-x_{n_k-2})\rangle\right) + \\ f(x_{n_k}) \leq f(x) \ \forall k \geq 2. \end{aligned}$$

The latter combined with Proposition 8(a) and (4) shows that $\limsup_{k\to+\infty} f(x_{n_k}) \leq f(x)$, hence $\lim_{k\to+\infty} f(x_{n_k}) = f(x)$. Since g is continuous, obviously $g(x_{n_k}) \to g(x)$ as $k \to +\infty$, thus $(f+g)(x_{n_k}) \to (f+g)(x)$ as $k \to +\infty$.

Further, by using the notations from Lemma 9, we have $y_{n_k} \in \partial(f+g)(x_{n_k})$ for every $k \geq 2$. By Proposition 8(a) and Lemma 9 we get $y_{n_k} \to 0$ as $k \to +\infty$.

Concluding, we have:

$$y_{n_k} \in \partial(f+g)(x_{n_k}) \ \forall k \geq 2,$$
$$(x_{n_k}, y_{n_k}) \to (x,0),\ k \to +\infty$$
$$(f+g)(x_{n_k}) \to (f+g)(x),\ k \to +\infty.$$

Hence $0 \in \partial(f+g)(x)$, that is, x is a critical point of $f+g$. ■

Lemma 11 *In the setting of Problem 1, choose* $\underline{\alpha}, \overline{\alpha}, \beta$ *satisfying* (11) *and* M_1, M_2 *satisfying* (10). *Assume that* $f+g$ *is coercive and consider the function*

$$H : \mathbb{R}^m \times \mathbb{R}^m \to (-\infty,+\infty],\ H(x,y) = (f+g)(x) + M_2\|x-y\|^2 \ \forall (x,y) \in \mathbb{R}^m \times \mathbb{R}^m.$$

Let $(x_n)_{n\in\mathbb{N}}$ *be a sequence generated by Algorithm 1. Then there exist* $M, N > 0$ *such that the following statements hold:*

(H_1) $H(x_{n+1},x_n)+M\|x_{n+1}-x_n\|^2 \le H(x_n,x_{n-1})$ *for all* $n\ge 1$*;*

(H_2) for all $n\ge 1$*, there exists* $w_{n+1}\in \partial H(x_{n+1},x_n)$ *such that* $\|w_{n+1}\|\le N(\|x_{n+1}-x_n\|+\|x_n-x_{n-1}\|)$*;*

(H_3) if $(x_{n_k})_{k\in\mathbb{N}}$ *is a subsequence such that* $x_{n_k}\to x$ *as* $k\to+\infty$*, then* $H(x_{n_k},x_{n_k-1})\to H(x,x)$ *as* $k\to+\infty$ *(there exists at least one subsequence with this property).*

Proof. For (H_1) just take $M=M_1-M_2$ and the conclusion follows from Lemma 6.

Let us prove (H_2). For every $n\ge 1$ we define

$$w_{n+1}=(y_{n+1}+2M_2(x_{n+1}-x_n),2M_2(x_n-x_{n+1})),$$

where $(y_n)_{n\ge 2}$ is the sequence introduced in Lemma 9. The fact that $w_{n+1}\in\partial H(x_{n+1},x_n)$ follows from Lemma 9 and the relation

$$\partial H(x,y)=\big(\partial(f+h)(x)+2M_2(x-y)\big)\times\{2M_2(y-x)\}\ \forall(x,y)\in\mathbb{R}^m\times\mathbb{R}^m. \tag{16}$$

Further, one has (see also Lemma 9)

$$\|w_{n+1}\|\le\|y_{n+1}+2M_2(x_{n+1}-x_n)\|+\|2M_2(x_n-x_{n+1})\|\le$$
$$\left(\frac{L_{\nabla F}}{\alpha_n}+L_{\nabla g}+4M_2\right)\|x_{n+1}-x_n\|+\frac{\beta_n}{\alpha_n}\|x_n-x_{n-1}\|.$$

Since $0<\underline{\alpha}\le\alpha_n\le\overline{\alpha}$ and $0\le\beta_n\le\beta$ for all $n\ge 1$, one can choose

$$N=\sup_{n\ge 1}\left\{\frac{L_{\nabla F}}{\alpha_n}+L_{\nabla g}+4M_2,\frac{\beta_n}{\alpha_n}\right\}<+\infty$$

and the conclusion follows.

For (H_3), consider $(x_{n_k})_{k\in\mathbb{N}}$ a subsequence such that $x_{n_k}\to x$ as $k\to+\infty$. We have shown in the proof of Lemma 10 that $(f+g)(x_{n_k})\to(f+g)(x)$ as $k\to+\infty$. From Proposition 8(a) and the definition of H we easily derive that $H(x_{n_k},x_{n_k-1})\to H(x,x)=(f+g)(x)$ as $k\to+\infty$. The existence of such a sequence follows from Lemma 10. ■

In the following we denote by $\omega((x_n)_{n\in\mathbb{N}})$ the set of cluster points of the sequence $(x_n)_{n\in\mathbb{N}}$.

Lemma 12 *In the setting of Problem 1, choose* $\underline{\alpha},\overline{\alpha},\beta$ *satisfying* (11) *and* M_1,M_2 *satisfying* (10)*. Assume that* $f+g$ *is coercive and consider the function*

$$H:\mathbb{R}^m\times\mathbb{R}^m\to(-\infty,+\infty],\ H(x,y)=(f+g)(x)+M_2\|x-y\|^2\ \forall(x,y)\in\mathbb{R}^m\times\mathbb{R}^m.$$

Let $(x_n)_{n\in\mathbb{N}}$ *be a sequence generated by Algorithm 1. Then the following statements are true:*

(a) $\omega((x_n,x_{n-1})_{n\ge1})\subseteq\operatorname{crit}(H)=\{(x,x)\in\mathbb{R}^m\times\mathbb{R}^m : x\in\operatorname{crit}(f+g)\}$*;*

(b) $\lim_{n\to\infty}\operatorname{dist}((x_n,x_{n-1}),\omega((x_n,x_{n-1}))_{n\ge1})=0$*;*

(c) $\omega((x_n,x_{n-1})_{n\ge1})$ *is nonempty, compact and connected;*

(d) H is finite and constant on $\omega((x_n,x_{n-1})_{n\geq 1})$.

Proof. (a) According to Lemma 10 and Proposition 8(a) we have $\omega((x_n,x_{n-1})_{n\geq 1}) \subseteq \{(x,x) \in \mathbb{R}^m \times \mathbb{R}^m : x \in \operatorname{crit}(f+g)\}$. The equality $\operatorname{crit}(H) = \{(x,x) \in \mathbb{R}^m \times \mathbb{R}^m : x \in \operatorname{crit}(f+g)\}$ follows from (16).

(b) and (c) can be shown as in [14, Lemma 5], by also taking into consideration [14, Remark 5], where it is noticed that the properties (b) and (c) are generic for sequences satisfying $x_{n+1} - x_n \to 0$ as $n \to +\infty$.

(d) According to Proposition 8, the sequence $((f+g)(x_n))_{n\in\mathbb{N}}$ is convergent, i.e. $\lim_{n\to+\infty}(f+g)(x_n) = l \in \mathbb{R}$. Take an arbitrary $(x,x) \in \omega((x_n,x_{n-1})_{n\geq 1})$, where $x \in \operatorname{crit}(f+g)$ (we took statement (a) into consideration). From Lemma 11(H_3) it follows that there exists a subsequence $(x_{n_k})_{k\in\mathbb{N}}$ such that $x_{n_k} \to x$ as $k \to +\infty$ and $H(x_{n_k},x_{n_k-1}) \to H(x,x)$ as $k \to +\infty$. Moreover, from Proposition 8 one has $H(x,x) = \lim_{k\to+\infty} H(x_{n_k},x_{n_k-1}) = \lim_{k\to+\infty}(f+g)(x_{n_k}) + M_2\|x_{n_k} - x_{n_k-1}\|^2 = l$ and the conclusion follows. ■

We give now the main result concerning the convergence of the whole sequence $(x_n)_{n\in\mathbb{N}}$.

Theorem 13 *In the setting of Problem 1, choose $\underline{\alpha}, \overline{\alpha}, \beta$ satisfying* (11) *and M_1, M_2 satisfying* (10)*. Assume that $f+g$ is coercive and that*

$$H : \mathbb{R}^m \times \mathbb{R}^m \to (-\infty,+\infty],\ H(x,y) = (f+g)(x) + M_2\|x-y\|^2\ \forall (x,y) \in \mathbb{R}^m \times \mathbb{R}^m$$

is a KL function. Let $(x_n)_{n\in\mathbb{N}}$ be a sequence generated by Algorithm 1. Then the following statements are true:

(a) $\sum_{n\in\mathbb{N}} \|x_{n+1} - x_n\| < +\infty$;

(b) there exists $x \in \operatorname{crit}(f+g)$ such that $\lim_{n\to+\infty} x_n = x$.

Proof. (a) Let $(x_n)_{n\in\mathbb{N}}$ be a sequence generated by Algorithm 1. According to Lemma 12 we can consider an element $\overline{x} \in \operatorname{crit}(f+g)$ such that $(\overline{x},\overline{x}) \in \omega((x_n,x_{n-1})_{n\geq 1})$. In analogy to the proof of Lemma 11 (by taking into account also the decrease property (H1)) one can easily show that $\lim_{n\to+\infty} H(x_n,x_{n-1}) = H(\overline{x},\overline{x})$. We separately treat the following two cases.

I. There exists $\overline{n} \in \mathbb{N}$ such that $H(x_{\overline{n}},x_{\overline{n}-1}) = H(\overline{x},\overline{x})$. The decrease property $(H1)$ in Lemma 11 implies $H(x_n,x_{n-1}) = H(\overline{x},\overline{x})$ for every $n \geq \overline{n}$. By using again property $(H1)$ in Lemma 11, one can show inductively that the sequence $(x_n,x_{n-1})_{n\geq\overline{n}}$ is constant. From here the conclusion follows automatically.

II. For all $n \geq 1$ we have $H(x_n,x_{n-1}) > H(\overline{x},\overline{x})$. Take $\Omega := \omega((x_n,x_{n-1})_{n\geq 1})$.

In virtue of Lemma 12(c) and (d) and Lemma 1, the KL property of H leads to the existence of positive numbers ϵ and η and a concave function $\varphi \in \Phi_\eta$ such that for all

$$\begin{aligned}(x,y) \in &\{(u,v) \in \mathbb{R}^m \times \mathbb{R}^m : \operatorname{dist}((u,v),\Omega) < \epsilon\} \\ &\cap \{(u,v) \in \mathbb{R}^m \times \mathbb{R}^m : H(\overline{x},\overline{x}) < H(u,v) < H(\overline{x},\overline{x}) + \eta\} \qquad (17)\end{aligned}$$

one has

$$\varphi'(H(x,y) - H(\overline{x},\overline{x}))\operatorname{dist}((0,0),\partial H(x,y)) \geq 1. \qquad (18)$$

Let $n_1 \in \mathbb{N}$ such that $H(x_n, x_{n-1}) < H(\overline{x}, \overline{x}) + \eta$ for all $n \geq n_1$. According to Lemma 12(b), there exists $n_2 \in \mathbb{N}$ such that $\operatorname{dist}((x_n, x_{n-1}), \Omega) < \epsilon$ for all $n \geq n_2$.

Hence the sequence $(x_n, x_{n-1})_{n \geq \overline{n}}$ where $\overline{n} = \max\{n_1, n_2\}$, belongs to the intersection (17). So we have (see (18))

$$\varphi'(H(x_n, x_{n-1}) - H(\overline{x}, \overline{x})) \operatorname{dist}((0,0), \partial H(x_n, x_{n-1})) \geq 1 \ \forall n \geq \overline{n}.$$

Since φ is concave, it holds

$$\begin{aligned} \varphi(H(x_n, x_{n-1}) - H(\overline{x}, \overline{x})) - \varphi(H(x_{n+1}, x_n) - H(\overline{x}, \overline{x})) \geq \\ \varphi'(H(x_n, x_{n-1}) - H(\overline{x}, \overline{x})) \cdot (H(x_n, x_{n-1}) - H(x_{n+1}, x_n)) \geq \\ \frac{H(x_n, x_{n-1}) - H(x_{n+1}, x_n)}{\operatorname{dist}((0,0), \partial H(x_n, x_{n-1}))} \ \forall n \geq \overline{n}. \end{aligned}$$

Let $M, N > 0$ be the real numbers furnished by Lemma 11. According to Lemma 11(H_2) there exists $w_n \in \partial H(x_n, x_{n-1})$ such that $\|w_n\| \leq N(\|x_n - x_{n-1}\| + \|x_{n-1} - x_{n-2}\|)$ for all $n \geq 2$. Then obviously $\operatorname{dist}((0,0), \partial H(x_n, x_{n-1})) \leq \|w_n\|$, hence

$$\begin{aligned} \varphi(H(x_n, x_{n-1}) - H(\overline{x}, \overline{x})) - \varphi(H(x_{n+1}, x_n) - H(\overline{x}, \overline{x})) \geq \\ \frac{H(x_n, x_{n-1}) - H(x_{n+1}, x_n)}{\|w_n\|} \geq \\ \frac{H(x_n, x_{n-1}) - H(x_{n+1}, x_n)}{N(\|x_n - x_{n-1}\| + \|x_{n-1} - x_{n-2}\|)} \ \forall n \geq \overline{n}. \end{aligned}$$

On the other hand, from Lemma 11(H_1) we obtain that

$$H(x_n, x_{n-1}) - H(x_{n+1}, x_n) \geq M\|x_{n+1} - x_n\|^2 \ \forall n \geq 1.$$

Hence, one has

$$\begin{aligned} \varphi(H(x_n, x_{n-1}) - H(\overline{x}, \overline{x})) - \varphi(H(x_{n+1}, x_n) - H(\overline{x}, \overline{x})) \geq \\ \frac{M\|x_{n+1} - x_n\|^2}{N(\|x_n - x_{n-1}\| + \|x_{n-1} - x_{n-2}\|)} \ \forall n \geq \overline{n}. \end{aligned}$$

For all $n \geq 1$, let us denote $\frac{N}{M}(\varphi(H(x_n, x_{n-1}) - H(\overline{x}, \overline{x})) - \varphi(H(x_{n+1}, x_n) - H(\overline{x}, \overline{x}))) = \epsilon_n$ and $\|x_n - x_{n-1}\| = a_n$. Then the last inequality becomes

$$\epsilon_n \geq \frac{a_{n+1}^2}{a_n + a_{n-1}} \ \forall n \geq \overline{n}. \tag{19}$$

Obviously, since $\varphi \geq 0$, for $S \geq 1$ we have

$$\begin{aligned} \sum_{n=1}^{S} \epsilon_n =& \frac{N}{M}(\varphi(H(x_1, x_0) - H(\overline{x}, \overline{x})) - \varphi(H(x_{S+1}, x_S) - H(\overline{x}, \overline{x}))) \\ \leq& \frac{N}{M}(\varphi(H(x_1, x_0) - H(\overline{x}, \overline{x}))), \end{aligned}$$

hence $\sum_{n \geq 1} \epsilon_n < +\infty$.

On the other hand, from (19) we derive

$$a_{n+1} = \sqrt{\epsilon_n(a_n + a_{n-1})} \leq \frac{1}{4}(a_n + a_{n-1}) + \epsilon_n \ \forall n \geq \overline{n}.$$

Hence, according to Lemma 3, $\sum_{n\geq 1} a_n < +\infty$, that is $\sum_{n\in\mathbb{N}} \|x_n - x_{n+1}\| < +\infty$.

(b) It follows from (a) that $(x_n)_{n\in\mathbb{N}}$ is a Cauchy sequence, hence it is convergent. Applying Lemma 10, there exists $x \in \operatorname{crit}(f+g)$ such that $\lim_{n\to+\infty} x_n = x$. ■

Remark 14 As kindly pointed out by an anonymous reviewer, a similar conclusion to the one of Theorem 13 can be obtained by applying [6, Theorem 2.9] in $\mathbb{R}^m \times \mathbb{R}^m$ endowed with the Euclidean product topology for the function $\widetilde{H} : \mathbb{R}^m \times \mathbb{R}^m \to \overline{\mathbb{R}}, \widetilde{H}(x,y) = (f+g)(x) + \frac{1}{2}(M_1 + M_2)\|x-y\|^2$. Indeed, from Lemma 6 it yields

$$\widetilde{H}(x_{n+1}, x_n) + \frac{1}{2}(M_1 - M_2)(\|x_{n+1} - x_n\|^2 + \|x_n - x_{n-1}\|^2) \leq \widetilde{H}(x_n, x_{n-1}) \ \forall n \geq 1,$$

which shows that H1 in [6] is fulfilled. The assumptions H2 and H3 in the above-named article are direct consequences of (H_2) and, respectively, (H_3) in Lemma 11. Under these premises, provided that $\widetilde{H}$ is a KL function, one obtains via [6, Theorem 2.9] the same conclusion as in Theorem 13.

However, the hypothesis that H is a KL function, as assumed in Theorem 13, is in our opinion in this context the most natural one, at least in what concerns the way in which it approaches the non-inertial case. Indeed, if β is equal to zero, then M_2 is equal to zero, too, and the conclusion of Theorem 13 follows by only assuming that $f+g$ is a KL function. On the other hand, in order to apply [6, Theorem 2.9], one would ask that $(x,y) \mapsto (f+g)(x) + \frac{1}{2}M_1\|x-y\|^2$ is a KL function, which is in general a stronger assumption.

Since the class of semi-algebraic functions is closed under addition (see for example [14]) and $(x,y) \mapsto c\|x-y\|^2$ is semi-algebraic for $c > 0$, we obtain also the following direct consequence.

Corollary 15 *In the setting of Problem 1, choose* $\underline{\alpha}, \overline{\alpha}, \beta$ *satisfying* (11) *and* M_1, M_2 *satisfying* (10). *Assume that* $f+g$ *is coercive and semi-algebraic. Let* $(x_n)_{n\in\mathbb{N}}$ *be a sequence generated by Algorithm 1. Then the following statements are true:*

(a) $\sum_{n\in\mathbb{N}} \|x_{n+1} - x_n\| < +\infty$*;*

(b) there exists $x \in \operatorname{crit}(f+g)$ *such that* $\lim_{n\to+\infty} x_n = x$.

Remark 16 As one can notice by taking a closer look at the proof of Lemma 10, the conclusion of this statement as the ones of Lemma 11, Lemma 12, Theorem 13 and Corollary 15 remain true, if instead of imposing that $f+g$ is coercive, we assume that $f+g$ is bounded from below and the sequence $(x_n)_{n\in\mathbb{N}}$ generated by Algorithm 1 is bounded. This observation is useful when dealing with optimization problems as the ones considered in Subsection 4.2.

4 Numerical experiments

This section is devoted to the presentation of two numerical experiments which illustrate the applicability of the algorithm proposed in this work. In both numerical experiments we considered $F = \frac{1}{2}\|\cdot\|^2$ and set $\sigma = 1$.

4.1 Detecting minimizers of nonconvex optimization problems

As emphasized in [34, Section 5.1] and [10, Exercise 1.3.9] one of the aspects which makes algorithms with inertial/memory effects useful is given by the fact that they are able to detect optimal solutions of minimization problems which cannot be found by their non-inertial variants. In this subsection we show that this phenomenon arises even when solving problems of type (20), where the nonsmooth function f is nonconvex. A similar situation has been addressed in [34], however, by assuming that f is convex.

Consider the optimization problem

$$\inf_{(x_1,x_2)\in\mathbb{R}^2} |x_1| - |x_2| + x_1^2 - \log(1+x_1^2) + x_2^2. \tag{20}$$

The function

$$f : \mathbb{R}^2 \to \mathbb{R}, f(x_1,x_2) = |x_1| - |x_2|,$$

is nonconvex and continuous, the function

$$g : \mathbb{R}^2 \to \mathbb{R}, g(x_1,x_2) = x_1^2 - \log(1+x_1^2) + x_2^2,$$

is continuously differentiable with Lipschitz continuous gradient with Lipschitz constant $L_{\nabla g} = 9/4$ and one can easily prove that $f+g$ is coercive. Furthermore, combining [5, the remarks after Definition 4.1], [12, Remark 5(iii)] and [14, Section 5: Example 4 and Theorem 3], one can easily conclude that H in Theorem 13 is a KL function. By considering the first order optimality conditions

$$-\nabla g(x_1,x_2) \in \partial f(x_1,x_2) = \partial(|\cdot|)(x_1) \times \partial(-|\cdot|)(x_2)$$

and by noticing that for all $x \in \mathbb{R}$ we have

$$\partial(|\cdot|)(x) = \begin{cases} 1, & \text{if } x > 0 \\ -1, & \text{if } x < 0 \\ \text{[-1,1]}, & \text{if } x = 0 \end{cases} \text{ and } \partial(-|\cdot|)(x) = \begin{cases} -1, & \text{if } x > 0, \\ 1, & \text{if } x < 0, \\ \{-1,1\}, & \text{if } x = 0, \end{cases}$$

(for the latter, see for example [31]), one can easily determine the two critical points $(0,1/2)$ and $(0,-1/2)$ of (20), which are actually both optimal solutions of this minimization problem. In Figure 2 the level sets and the graph of the objective function in (20) are represented.

For $\gamma > 0$ and $x = (x_1,x_2) \in \mathbb{R}^2$ we have (see Remark 4)

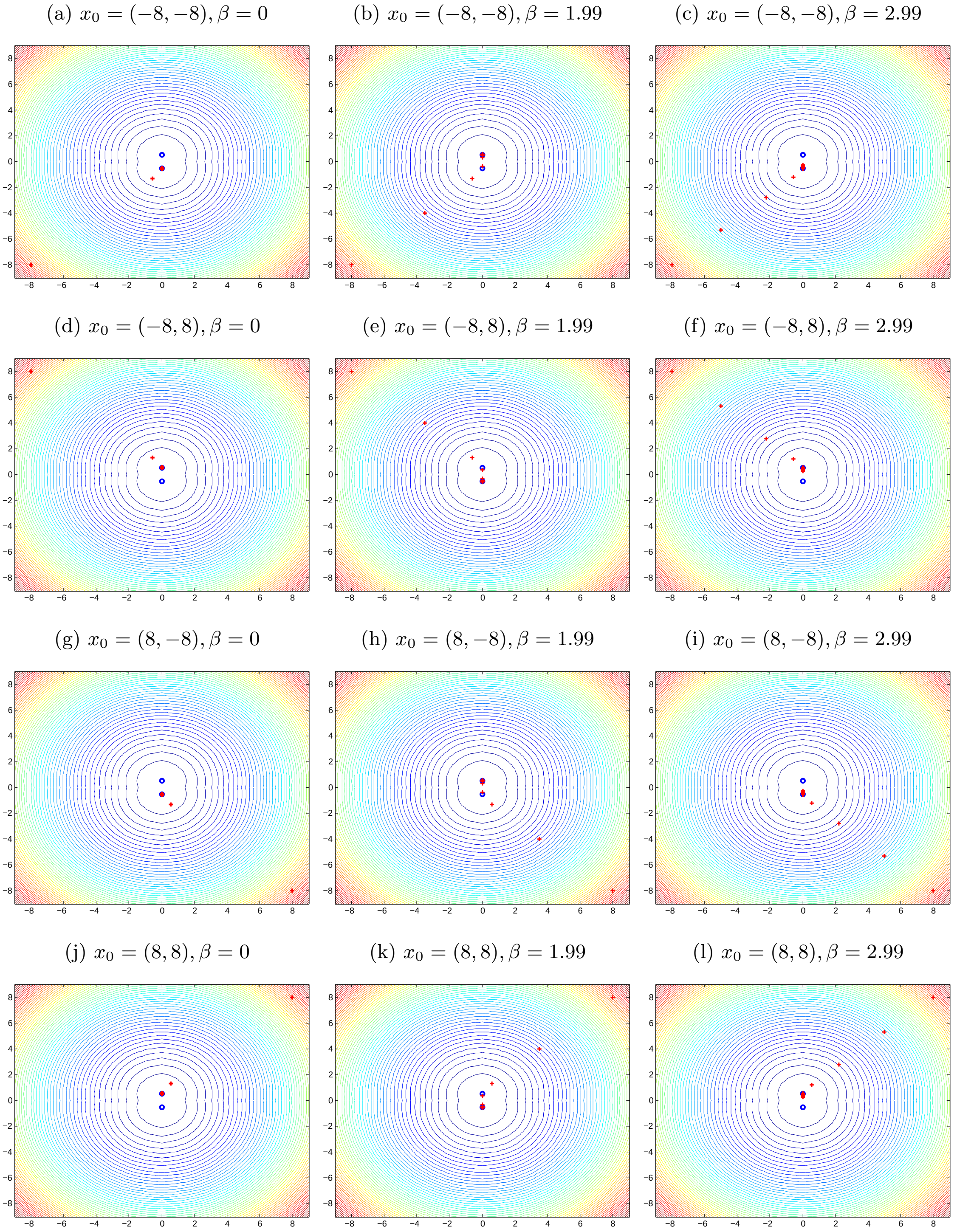
$$\operatorname{prox}_{\gamma f}(x) = \underset{u\in\mathbb{R}^2}{\operatorname{argmin}} \left\{ \frac{\|u-x\|^2}{2\gamma} + f(u) \right\} = \operatorname{prox}_{\gamma|\cdot|}(x_1) \times \operatorname{prox}_{\gamma(-|\cdot|)}(x_2),$$

where in the first component one has the well-known shrinkage operator

$$\operatorname{prox}_{\gamma|\cdot|}(x_1) = x_1 - \operatorname{sgn}(x_1) \cdot \min\{|x_1|, \gamma\},$$

while for the proximal operator in the second component the following formula can be proven

$$\operatorname{prox}_{\gamma(-|\cdot|)}(x_2) = \begin{cases} x_2 + \gamma, & \text{if } x_2 > 0 \\ x_2 - \gamma, & \text{if } x_2 < 0 \\ \{-\gamma, \gamma\}, & \text{if } x_2 = 0. \end{cases}$$



(a) $x_0 = (-8, -8), \beta = 0$

(b) $x_0 = (-8, -8), \beta = 1.99$

(c) $x_0 = (-8, -8), \beta = 2.99$

(d) $x_0 = (-8, 8), \beta = 0$

(e) $x_0 = (-8, 8), \beta = 1.99$

(f) $x_0 = (-8, 8), \beta = 2.99$

(g) $x_0 = (8, -8), \beta = 0$

(h) $x_0 = (8, -8), \beta = 1.99$

(i) $x_0 = (8, -8), \beta = 2.99$

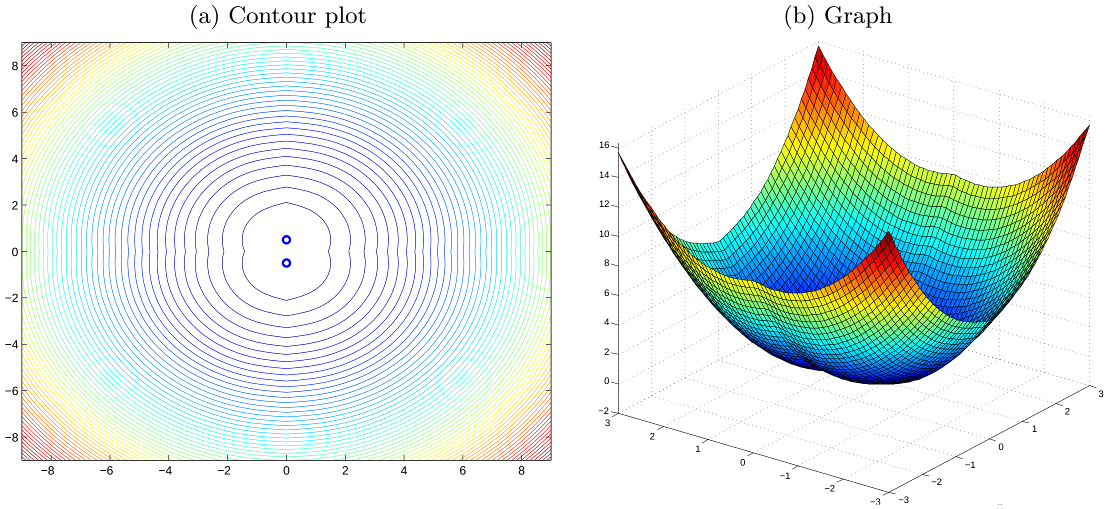
(j) $x_0 = (8, 8), \beta = 0$

(k) $x_0 = (8, 8), \beta = 1.99$

(l) $x_0 = (8, 8), \beta = 2.99$

Figure 1: Algorithm 1 after 100 iterations and with starting points $(-8, -8), (-8, 8), (8, -8)$ and $(8, 8)$, respectively: the first column shows the iterates of the non-inertial version ($\beta_n = \beta = 0$ for all $n \geq 1$), the second column the ones of the inertial version with $\beta_n = \beta = 1.99$ for all $n \geq 1$ and the third column the ones of the inertial version with $\beta_n = \beta = 2.99$ for all $n \geq 1$.

We implemented Algorithm 1 by choosing $\beta_n = \beta = 0$ for all $n \geq 1$ (which corresponds to the non-inertial version), $\beta_n = \beta = 0.199$ for all $n \geq 1$ and $\beta_n = \beta = 0.299$ for all $n \geq 1$,



(a) Contour plot

(b) Graph

Figure 2: Contour plot and graph of the objective function in (20). The two global optimal solutions $(0, 0.5)$ and $(0, -0.5)$ are marked on the first image.

respectively, and by setting $\alpha_n = (0.99999 - 2\beta_n)/L_{\nabla g}$ for all $n \geq 1$. As starting points we considered the corners of the box generated by the points $(\pm 8, \pm 8)$. Figure 1 shows that independently of the four starting points we have the following phenomenon: the non-inertial version recovers only one of the two optimal solutions, situation which persists even when changing the value of α_n; on the other hand, the inertial version is capable to find both optimal solutions, namely, one for $\beta = 0.199$ and the other one for $\beta = 0.299$.

4.2 Restoration of noisy blurred images

The following numerical experiment concerns the restoration of a noisy blurred image by using a nonconvex misfit functional with nonconvex regularization. For a given matrix $A \in \mathbb{R}^{m \times m}$ describing a blur operator and a given vector $b \in \mathbb{R}^m$ representing the blurred and noisy image, the task is to estimate the unknown original image $\overline{x} \in \mathbb{R}^m$ fulfilling

$$A\overline{x} = b.$$

To this end we solve the following regularized nonconvex minimization problem

$$\inf_{x \in \mathbb{R}^m} \left\{ \sum_{k=1}^{M} \sum_{l=1}^{N} \varphi\big((Ax - b)_{kl}\big) + \lambda \|Wx\|_0 \right\}, \tag{21}$$

where $\varphi : \mathbb{R} \to \mathbb{R}$, $\varphi(t) = \log(1 + t^2)$, is derived form the Student t distribution, $\lambda > 0$ is a regularization parameter, $W : \mathbb{R}^m \to \mathbb{R}^m$ is a discrete Haar wavelet transform with four levels and $\|y\|_0 = \sum_{i=1}^m |y_i|_0$ ($|\cdot|_0 = |\operatorname{sgn}(\cdot)|$) furnishes the number of nonzero entries of the vector $y = (y_1, ..., y_m) \in \mathbb{R}^m$. In this context, $x \in \mathbb{R}^m$ represents the vectorized image $X \in \mathbb{R}^{M \times N}$, where $m = M \cdot N$ and $x_{i,j}$ denotes the normalized value of the pixel located in the i-th row and the j-th column, for $i = 1, \dots, M$ and $j = 1, \dots, N$. Again, by combining [5, the remarks after Definition 4.1], [12, Remark 5(iii)] and [14, Section 5: Example 3, Example 4 and Theorem 3], one can conclude that H in Theorem 13 is a KL function.

It is immediate that (21) can be written in the form (3), by defining $f(x) = \lambda\|Wx\|_0$ and $g(x) = \sum_{k=1}^{M}\sum_{l=1}^{N}\varphi\big((Ax-b)_{kl}\big)$ for all $x \in \mathbb{R}^m$. By using that $WW^* = W^*W = I_m$, one can prove the following formula concerning the proximal operator of f

$$\operatorname{prox}_{\gamma f}(x) = W^* \operatorname{prox}_{\lambda\gamma\|\cdot\|_0}(Wx) \;\forall x \in \mathbb{R}^m \;\forall \gamma > 0,$$

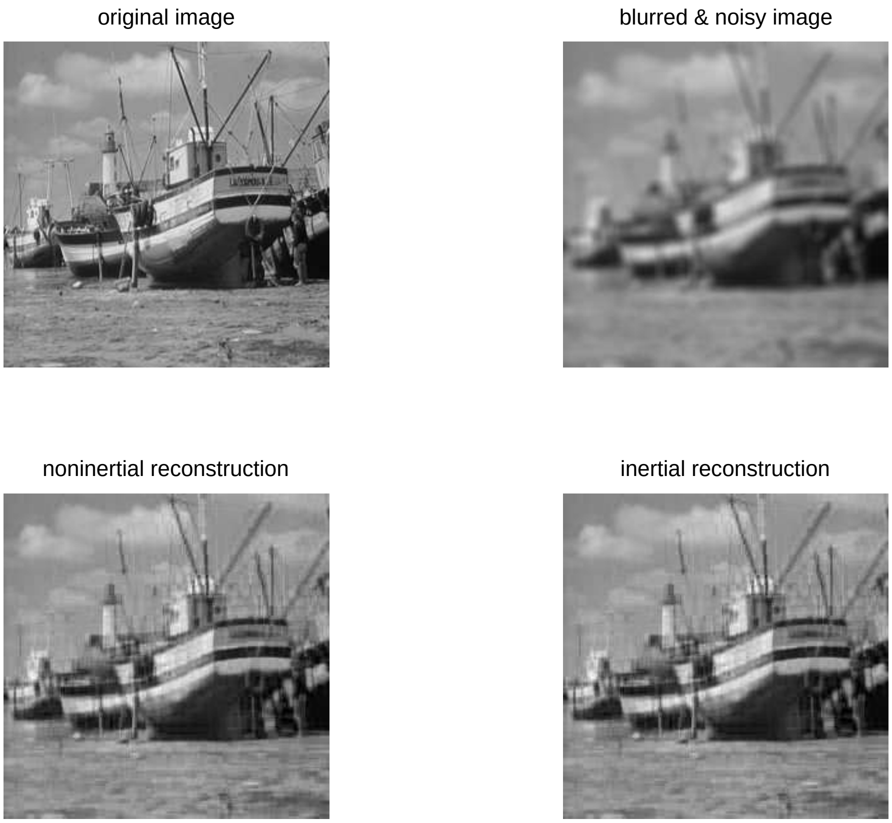
where for all $u = (u_1, ..., u_m)$ we have (see [6, Example 5.4(a)])

$$\operatorname{prox}_{\lambda\gamma\|\cdot\|_0}(u) = (\operatorname{prox}_{\lambda\gamma|\cdot|_0}(u_1), ..., \operatorname{prox}_{\lambda\gamma|\cdot|_0}(u_m))$$

and for all $t \in \mathbb{R}$

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

For the experiments we used the 256×256 boat test image which we first blurred by using a Gaussian blur operator of size 9×9 and standard deviation 4 and to which we afterward added a zero-mean white Gaussian noise with standard deviation 10^{-6}. In the first row of Figure 3 the original boat test image and the blurred and noisy one are represented, while in the second row one has the reconstructed images by means of the non-inertial (for $\beta_n = \beta = 0$ for all $n \geq 1$) and inertial versions (for $\beta_n = \beta = 10^{-7}$ for all $n \geq 1$) of Algorithm 1, respectively. We took as regularization parameter $\lambda = 10^{-5}$ and set $\alpha_n = (0.999999 - 2\beta_n)/L_{\nabla g}$ for all $n \geq 1$, whereby the Lipschitz constant of the gradient of the smooth misfit function is $L_{\nabla g} = 2$.



original image

blurred & noisy image

noninertial reconstruction

inertial reconstruction

Figure 3: The first row shows the original 256×256 boat test image and the blurred and noisy one and the second row the reconstructed images after 300 iterations.

β	0.4	0.2	0.01	0.0001	10^{-7}	0
ISNR(300)	2.081946	3.101028	3.492989	3.499428	3.511135	3.511134

Table 1: The ISNR values after 300 iterations for different choices of β.

We compared the quality of the recovered images for $\beta_n = \beta$ for all $n \geq 1$ and different values of β by making use of the improvement in signal-to-noise ratio (ISNR), which is defined as

$$\mathrm{ISNR}(n) = 10 \log_{10} \left(\frac{\|x - b\|^2}{\|x - x_n\|^2} \right),$$

where x, b and x_n denote the original, observed and estimated image at iteration n, respectively.

In Table 1 we list the values of the ISNR-function after 300 iterations, whereby the case $\beta = 0$ corresponds to the non-inertial version of the algorithm. One can notice that for β taking very small values, the inertial version is competitive with the non-inertial one.

5 Concluding remarks

In this paper we propose a forward-backward proximal-type algorithm with inertial/memory effects for minimizing the sum of a nonsmooth with a smooth function in the nonconvex setting. Every sequence of iterates generated by the algorithm is proved to converge to a critical point of the objective function, whenever an appropriate regularization of the latter satisfies the Kurdyka-Łojasiewicz inequality. In this way we extend to the full nonconvex setting the inertial forward-backward type algorithm proposed in [34] for minimizing the sum of a nonsmooth convex with a smooth (not necessarily convex) function.

As it is the case for the particular instances considered in Section 4, very tight bounds for the parameters used in the iterative scheme are needed. More than that, for these particular instances, there is a minimal difference between the inertial and non-inertial schemes.

In the context of proving convergence for algorithms designed to solve nonsmooth optimization problems with KL functions two approaches can be found in the literature. One of them is the approach proposed in [14], which we also follow in our manuscript, while the second one was used in [6]. As explained in Remark 14, the two approaches mainly differ in the way the regularization of the objective is constructed. Opting for the approach in [6], one could come to the conclusion by using in a straightforward way the statements of Lemma 11. However, different to [6], the inertial and non-inertial schemes are treated with our choice of H in an unitary way. Furthermore, in the inertial case, working with H does not assume to have any information about $L_{\nabla g}$, a constant which explicitly appears in the definition of M_1 (see Remark 14). On the other hand, for the choice of $\underline{\alpha}, \overline{\alpha}$ and β and, consequently, for the defintion of M_2, the Lipschitz constant $L_{\nabla g}$ can be unknown if an upper bound $L > L_{\nabla g}$ is available.

Acknowledgements. The authors are thankful to two anonymous reviewers for pertinent comments and remarks which improved the quality of the paper.

References

[1] F. Alvarez, *On the minimizing property of a second order dissipative system in Hilbert spaces*, SIAM Journal on Control and Optimization 38(4), 1102–1119, 2000

[2] F. Alvarez, *Weak convergence of a relaxed and inertial hybrid projection-proximal point algorithm for maximal monotone operators in Hilbert space*, SIAM Journal on Optimization 14(3), 773–782, 2004

[3] F. Alvarez, H. Attouch, *An inertial proximal method for maximal monotone operators via discretization of a nonlinear oscillator with damping*, Set-Valued Analysis 9, 3–11, 2001

[4] H. Attouch, J. Bolte, *On the convergence of the proximal algorithm for nonsmooth functions involving analytic features*, Mathematical Programming 116(1-2) Series B, 5–16, 2009

[5] H. Attouch, J. Bolte, P. Redont, A. Soubeyran, *Proximal alternating minimization and projection methods for nonconvex problems: an approach based on the Kurdyka-Łojasiewicz inequality*, Mathematics of Operations Research 35(2), 438–457, 2010

[6] H. Attouch, J. Bolte, B.F. Svaiter, *Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods*, Mathematical Programming 137(1-2) Series A, 91–129, 2013

[7] H. Attouch, J. Peypouquet, P. Redont, *A dynamical approach to an inertial forward-backward algorithm for convex minimization*, SIAM Journal on Optimization 24(1), 232–256, 2014

[8] H.H. Bauschke P.L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, CMS Books in Mathematics, Springer, New York, 2011

[9] A. Beck and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM Journal of Imaging Sciences 2(1), 183–202, 2009

[10] D.P. Bertsekas, *Nonlinear Programming*, 2nd ed., Athena Scientific, Cambridge, MA, 1999

[11] J. Bolte, A. Daniilidis, A. Lewis, *The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems*, SIAM Journal on Optimization 17(4), 1205–1223, 2006

[12] J. Bolte, A. Daniilidis, A. Lewis, M. Shiota, *Clarke subgradients of stratifiable functions*, SIAM Journal on Optimization 18(2), 556–572, 2007

[13] J. Bolte, A. Daniilidis, O. Ley, L. Mazet, *Characterizations of Łojasiewicz inequalities: subgradient flows, talweg, convexity*, Transactions of the American Mathematical Society 362(6), 3319–3363, 2010

[14] J. Bolte, S. Sabach, M. Teboulle, *Proximal alternating linearized minimization for nonconvex and nonsmooth problems*, Mathematical Programming Series A (146)(1–2), 459–494, 2014

[15] R.I. Boţ, E.R. Csetnek, *An inertial forward-backward-forward primal-dual splitting algorithm for solving monotone inclusion problems*, Numerical Algorithms, DOI: 10.1007/s11075-015-0007-5, 2015

[16] R.I. Boţ, E.R. Csetnek, *An inertial alternating direction method of multipliers*, Minimax Theory and its Applications 1(1), 2015

[17] R.I. Boţ, E.R. Csetnek, *A hybrid proximal-extragradient algorithm with inertial effects*, Numerical Functional Analysis and Optimization, DOI:10.1080/01630563.2015.1042113, 2015

[18] R.I. Boţ, E.R. Csetnek, *An inertial Tseng's type proximal algorithm for nonsmooth and nonconvex optimization problems*, Journal of Optimization Theory and Applications, DOI 10.1007/s10957-015-0730-z, 2015

[19] R.I. Boţ, E.R. Csetnek, C. Hendrich, *Inertial Douglas-Rachford splitting for monotone inclusion problems*, Applied Mathematics and Computation 256, 472–487, 2015

[20] A. Cabot, P. Frankel, *Asymptotics for some proximal-like method involving inertia and memory aspects*, Set-Valued and Variational Analysis 19, 59–74, 2011

[21] R.H. Chan, S. MA, J. Yang, *Inertial primal-dual algorithms for structured convex optimization*, arXiv:1409.2992v1, 2014

[22] C. Chen, S. MA, J. Yang, *A general inertial proximal point method for mixed variational inequality problem*, arXiv:1407.8238v2, 2014

[23] E. Chouzenoux, J.-C. Pesquet, A. Repetti, *Variable metric forward-backward algorithm for minimizing the sum of a differentiable function and a convex function*, Journal of Optimization Theory and its Applications 162(1), 107–132, 2014

[24] P.L. Combettes, *Solving monotone inclusions via compositions of nonexpansive averaged operators*, Optimization 53(5-6), 475–504, 2004

[25] P. Frankel, G. Garrigos, J. Peypouquet, *Splitting methods with variable metric for Kurdyka-Łojasiewicz functions and general convergence rates*, Journal of Optimization Theory and its Applications 165(3), 874–900, 2015

[26] R. Hesse, D.R. Luke, S. Sabach, M.K. Tam, *Proximal heterogeneous block input-output method and application to blind ptychographic diffraction imaging*, arXiv:1408.1887v1, 2014

[27] K. Kurdyka, *On gradients of functions definable in o-minimal structures*, Annales de l'institut Fourier (Grenoble) 48(3), 769–783, 1998

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

[29] P.-E. Maingé, *Convergence theorems for inertial KM-type algorithms*, Journal of Computational and Applied Mathematics 219, 223–236, 2008

[30] P.-E. Maingé, A. Moudafi, *Convergence of new inertial proximal methods for dc programming*, SIAM Journal on Optimization 19(1), 397–413, 2008

[31] B. Mordukhovich, *Variational Analysis and Generalized Differentiation, I: Basic Theory, II: Applications*, Springer-Verlag, Berlin, 2006.

[32] A. Moudafi, M. Oliny, *Convergence of a splitting inertial proximal method for monotone operators*, Journal of Computational and Applied Mathematics 155, 447–454, 2003

[33] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, Kluwer Academic Publishers, Dordrecht, 2004

[34] P. Ochs, Y. Chen, T. Brox, T. Pock, *iPiano: Inertial proximal algorithm for nonconvex optimization*, SIAM Journal on Imaging Sciences 7(2), 1388–1419, 2014

[35] J.-C. Pesquet, N. Pustelnik, *A parallel inertial proximal optimization method*, Pacific Journal of Optimization 8(2), 273–306, 2012

[36] R.T. Rockafellar, R.J.-B. Wets, *Variational Analysis*, Fundamental Principles of Mathematical Sciences 317, Springer-Verlag, Berlin, 1998