A forward-backward penalty scheme with inertial effects for montone inclusions. Applications to convex bilevel programming

Radu Ioan Boţ* Dang-Khoa Nguyen†

November 24, 2018

Abstract. We investigate forward-backward splitting algorithm of penalty type with inertial effects for finding the zeros of the sum of a maximally monotone operator and a cocoercive one and the convex normal cone to the set of zeroes of an another cocoercive operator. Weak ergodic convergence is obtained for the iterates, provided that a condition expressed via the Fitzpatrick function of the operator describing the underlying set of the normal cone is verified. Under strong monotonicity assumptions, strong convergence for the sequence of generated iterates can be proved. As a particular instance we consider a convex bilevel minimization problem including the sum of a nonsmooth and a smooth function in the upper level and another smooth function in the lower level. We show that in this context weak nonergodic and strong convergence can be also achieved under inf-compactness assumptions for the involved functions.

Keywords. maximally monotone operator, Fitzpatrick function, forward-backward splitting algorithm, convex bilevel optimization

AMS subject classification. 47H05, 65K05, 90C25

1 Introduction and preliminaries

1.1 Motivation and problems formulation

During the last couple years one can observe in the optimization community an increasing interest in numerical schemes for solving variational inequalities expressed as monotone inclusion problems of the form

$$0 \in Ax + N_M(x), \tag{1.1}$$

where $\mathcal{H}$ is a real Hilbert space, $A\colon \mathcal{H} \rightrightarrows \mathcal{H}$ is a maximally monotone operator, $M := \arg\min h$ is the set of global minima of the proper, convex and lower semicontinuous function $h\colon \mathbb{R} \to \bar{\mathbb{R}} := \mathbb{R} \cup \{\pm\infty\}$ and $N_M\colon \mathcal{H} \rightrightarrows \mathcal{H}$ is the normal cone of the set M. The article [7] was starting point for a series of papers [6, 9, 10, 12, 18, 19, 24, 25, 33, 37, 38] addressing this topic or related ones. All these papers share the common feature that the proposed iterative schemes use penalization strategies, namely, by evaluating the penalized h by its gradient, in case the function is smooth (see, for instance, [9]), and by its proximal operator, in case it is nonsmooth (see, for instance,[10]).

Weak ergodic convergence has been obtained in [9, 10] under the hypothesis:

$$\text{For all } p \in \operatorname{Ran}N_M, \sum_{n\geqslant 1} \lambda_n\beta_n \left[h^*\left(\frac{p}{\beta_n}\right) - \sigma_M\left(\frac{p}{\beta_n}\right)\right] < +\infty, \tag{1.2}$$

*Faculty of Mathematics, University of Vienna, Oskar-Morgenstern-Platz 1, 1090 Vienna, Austria, e-mail: `radu.bot@univie.ac.at`. Research partially supported by FWF (Austrian Science Fund), project I 2419-N32.

†Faculty of Mathematics, University of Vienna, Oskar-Morgenstern-Platz 1, 1090 Vienna, Austria, e-mail: `dang-khoa.nguyen@univie.ac.at`. The author gratefully acknowledges the financial support of the Doctoral Programme *Vienna Graduate School on Computational Optimization (VGSCO)* which is funded by Austrian Science Fund (FWF, project W1260-N35).

with $(\lambda_n)_{n\geqslant 1}$, the sequence of step sizes, $(\beta_n)_{n\geqslant 1}$, the sequence of penalty parameters, $h^*\colon \mathcal{H} \to \bar{\mathbb{R}}$, the Fenchel conjugate function of h, and $\text{Ran}N_M$ the range of the normal cone operator $N_M\colon \mathcal{H} \rightrightarrows \mathcal{H}$. Let us mention that (1.2) is the discretized counterpart of a condition introduced in [7] for continuous-time nonautonomous differential inclusions.

One motivation for studying numerical algorithms for monotone inclusions of type (1.1) comes from the fact that, when $A \equiv \partial f$ is the convex subdifferential of a proper, convex and lower semicontinuous function $f\colon \mathcal{H} \to \bar{\mathbb{R}}$, they furnish iterative methods for solving bilevel optimization problems of the form

$$\min_{x\in\mathcal{H}} \left\{ f(x) : x \in \arg\min h \right\}. \tag{1.3}$$

Among the applications where bilevel programming problems play an important role we mention ithe modelling of Stackelberg games, the determination of Wardrop equilibria for network flows, convex feasibility problems [5], domain decomposition methods for PDEs [4], image processing problems [18], and optimal control problems [10].

Later on, in [19], the following monotone inclusion problem, which turned out to be more suitable for applications, has been addressed in the same spirit of penalty algorithms

$$0 \in Ax + Dx + N_M(x), \tag{1.4}$$

where $A\colon \mathcal{H} \rightrightarrows \mathcal{H}$ is a maximally monotone operator, $D\colon \mathcal{H} \to \mathcal{H}$ is cocoercive operator and the constraint set M is the set of zeros of another cocoercive operator $B\colon \mathcal{H} \to \mathcal{H}$. The provided algorithm of forward-backward type evaluates the operator A by a backward step and the two single-valued operators by forward steps. For the convergence analysis, (1.2) has been replaced by a condition formulated in terms of the Fitzpatrick function associated with the operator B, which we will also use in this paper. In [12], several particular situations for which this new condition is fulfilled have been provided.

The aim of this work is to endow the forward-backward penalty scheme for solving (1.4) from [19] with inertial effects, which means that the new iterate is defined in terms of the previous two iterates. Inertial algorithms have their roots in the time discretization of second order differential systems [3]. They can accelerate the convergence of iterates when minimizing a differentiable function [39] and the convergence of the objective function values when minimizing the sum of a convex nonsmooth and a convex smooth function [15, 28]. Moreover, as emphasized in [16], see also [23], algorithms with inertial effects may detect optimal solutions of minimization problems which cannot be found by their noninertial variants. In the last years, a huge interest in inertial algorithms can be notices (see, for instance, [1, 2, 3, 8, 11, 15, 20, 21, 22, 23, 24, 25, 29, 30, 34, 35, 36]).

We prove weak ergodic convergence of the sequence generated by the inertial forward-backward penalty algorithm to a solution of the monotone inclusion problem (1.4), under reasonable assumptions for the sequences of step sizes, penalty and inertial parameters. When the operator A is assumed to be strongly monotone, we also prove strong convergence of the generated iterates to the unique solution of (1.4).

In Section 3, we address the minimization of the sum of a convex nonsmooth and a convex smooth function with respect to the set of minimizes of another convex and smooth function. Besides the convergence results obtained from the general case, we achieve weak nonergodic and strong convergence statements under inf-compactness assumptions for the involved functions. The weak nonergodic theorem is an useful alternative to the one in [25], where a similar statement has been obtained for the inertial forward-bacward penalty algorithm with constant inertial parameter under assumptions which are quite complicated and hard to verify (see also [37, 38]).

1.2 Notations and preliminaries

In this subsection we introduce some notions and basic results which we will use throughout this paper (see [13, 17, 40]). Let $\mathcal{H}$ be a real Hilbert space with *inner product* $\langle \cdot, \cdot \rangle$ and associated *norm* $\|\cdot\| = \sqrt{\langle \cdot, \cdot \rangle}$.

For a function $\Psi \colon \mathcal{H} \to \bar{\mathbb{R}} := \mathbb{R} \cup \{\pm\infty\}$, we denote $\mathrm{Dom}\Psi = \{x \in \mathcal{H} \colon \Psi(x) < +\infty\}$ its *effective domain* and say that Ψ is *proper*, if $\mathrm{Dom}\Psi \neq \varnothing$ and $\Psi(x) > -\infty$ for all $x \in \mathcal{H}$. The *conjugate function* of Ψ is $\Psi^* \colon \mathcal{H} \to \bar{\mathbb{R}}$, $\Psi^*(u) = \sup_{x \in \mathcal{H}} \{\langle x, u \rangle - \Psi(x)\}$. The *convex subdifferential* of Ψ at the point $x \in \mathcal{H}$ is the set $\partial\Psi(x) = \{p \in \mathcal{H} \colon \langle y - x, p \rangle \leqslant \Psi(y) - \Psi(x) \ \forall y \in \mathcal{H}\}$, whenever $\Psi(x) \in \mathbb{R}$. We take by convention $\partial\Psi(x) = \varnothing$, if $\Psi(x) \in \{\pm\infty\}$.

Let M be a nonempty subset of $\mathcal{H}$. The *indicator function* of M, which is denoted by $\delta_M \colon \mathcal{H} \to \bar{\mathbb{R}}$, takes the value 0 on M and $+\infty$ otherwise. The *convex subdifferential* of the *indicator function* is the *normal cone* of M, that is $N_M(x) = \{p \in \mathcal{H} \colon \langle y - x, p \rangle \leqslant 0 \ \forall y \in \mathcal{H}\}$, if $x \in M$, and $N_M(x) = \varnothing$ otherwise. Notice that for $x \in M$ we have $p \in N_M(x)$ if and only if $\sigma_M(x) = \langle x, p \rangle$, where $\sigma_M = \delta_M^*$ is the *support function* of M.

For an arbitrary set-value operator $A \colon \mathcal{H} \rightrightarrows \mathcal{H}$ we denote by $\mathrm{Gr}A = \{(x, v) \in \mathcal{H} \times \mathcal{H} \colon v \in Ax\}$ its *graph*, by $\mathrm{Dom}A = \{x \in \mathcal{H} \colon Ax \neq \varnothing\}$ its *domain*, by $\mathrm{Ran}A = \{v \in \mathcal{H} \colon \exists x \in \mathcal{H} \text{ with } v \in Ax\}$ its *range* and by $A^{-1} \colon \mathcal{H} \rightrightarrows \mathcal{H}$ its *inverse operator*, defined by $(v, x) \in \mathrm{Gr}A^{-1}$ if and only if $(x, v) \in \mathrm{Gr}A$. We use also the notation $\mathrm{Zer}A = \{x \in \mathcal{H} \colon 0 \in Ax\}$ for the *set of zeros* of the operator A. We say that A is *monotone*, if $\langle x - y, v - w \rangle \geqslant 0$ for all $(x, v), (y, w) \in \mathrm{Gr}A$. A monotone operator A is said to be *maximally monotone*, if there exists no proper monotone extension of the graph of A on $\mathcal{H} \times \mathcal{H}$. Let us mention that if A is maximally monotone, then $\mathrm{Zer}A$ is a convex and closed set, [13, Proposition 23.39]. We refer to [13, Section 23.4] for conditions ensuring that $\mathrm{Zer}A$ is nonempty. If A is maximally monotone, then one has the following characterization for the set of its zeros

$$z \in \mathrm{Zer}A \text{ if and only if } \langle u - z, y \rangle \geqslant 0 \text{ for all } (u, y) \in \mathrm{Gr}A. \tag{1.5}$$

The operator A is said to be $\gamma-$*strongly monotone* with $\gamma > 0$, if $\langle x - y, v - w \rangle \geqslant \|x - y\|^2$ for all $(x, v), (y, w) \in \mathrm{Gr}A$. If A is maximally monotone and strongly monotone, then $\mathrm{Zer}A$ is a singleton, thus nonempty, [13, Corollary 23.27].

The *resolvent* of A, $J_A \colon \mathcal{H} \rightrightarrows \mathcal{H}$, is defined by $J_A := (\mathrm{Id} + A)^{-1}$, where $\mathrm{Id} \colon \mathcal{H} \to \mathcal{H}$ denotes the *identity operator* on $\mathcal{H}$. If A is maximally monotone, then $J_A \colon \mathcal{H} \to \mathcal{H}$ is single-value and maximally monotone, [13, Proposition 23.7, Corollary 23.10]. For an arbitrary $\gamma > 0$, we have the following identity ([13, Proposition 23.18])

$$J_{\gamma A} + \gamma J_{\gamma^{-1} A^{-1}} \circ \gamma^{-1}\mathrm{Id} = \mathrm{Id}.$$

We denote $\Gamma(\mathcal{H})$ the family of proper, convex and lower semicontinuous extended real-valued functions defined on $\mathcal{H}$. When $\Psi \in \Gamma(\mathcal{H})$ and $\gamma > 0$, we denote by $\mathrm{prox}_{\gamma\Psi}(x)$ the *proximal point* with parameter γ of function Ψ at point $x \in \mathcal{H}$, which is the unique optimal solution of the optimization problem

$$\inf_{y \in \mathcal{H}} \left\{ \Psi(y) + \frac{1}{2\gamma} \|y - x\|^2 \right\}.$$

Notice that $J_{\gamma\partial\Psi} = (\mathrm{Id} + \gamma\partial\Psi)^{-1} = \mathrm{prox}_{\gamma\Psi}$, thus $\mathrm{prox}_{\gamma\Psi} \colon \mathcal{H} \to \mathcal{H}$ is a single-valued operator fulfilling the so-called *Moreau's decomposition formula*:

$$\mathrm{prox}_{\gamma\Psi} + \gamma \mathrm{prox}_{\gamma^{-1}\Psi^*} \circ \gamma^{-1}\mathrm{Id} = \mathrm{Id}.$$

The function $\Psi \colon \mathcal{H} \to \bar{\mathbb{R}}$ is said to be $\gamma-$*strongly convex* with $\gamma > 0$, if $\Psi - \dfrac{\gamma}{2}\|\cdot\|^2$ is a convex function. This property implies that $\partial\Psi$ is $\gamma-$strongly monotone.

The *Fitzpatrick function* ([32]) associated to a monotone operator A is defined as

$$\varphi_A : \mathcal{H} \times \mathcal{H} \to \bar{\mathbb{R}}, \quad \varphi_A(x,u) := \sup_{(y,v)\in \mathrm{Gr}A} \{\langle x, v\rangle + \langle y, u\rangle - \langle y, v\rangle\}$$

and it is a convex and lower semicontinuous function. For insights in the outstanding role played by the Fitzpatrick function in relatin the convex analysis with the theory of monotone operators we refer to [13, 14, 17, 26, 27] and the references therein. If A is maximally monotone, then φ_A is proper and it fulfills

$$\varphi_A(x,u) \geqslant \langle x, u\rangle \quad \forall (x,u) \in \mathcal{H} \times \mathcal{H},$$

with equality if and only if $(x,u) \in \mathrm{Gr}A$. Notice that if $\Psi \in \Gamma(\mathcal{H})$, then $\partial\Psi$ is a maximally monotone operator and it holds $(\partial\Psi)^{-1} = \partial\Psi^*$. Furthermore, the following inequality is true (see [14]):

$$\varphi_{\partial\Psi}(x,u) \leqslant \Psi(x) + \Psi^*(u) \quad \forall (x,v) \in \mathcal{H} \times \mathcal{H}. \tag{1.6}$$

We present as follows some statements that will be essential when carrying out the convergence analysis. Let $(x_n)_{n\geqslant 0}$ be a sequence in $\mathcal{H}$ and $(\lambda_n)_{n\geqslant 1}$ be a sequenceof positive real numbers. The *sequence of weighted averages* $(z_n)_{n\geqslant 1}$ is defined for every $n \geqslant 1$ as

$$z_n := \frac{1}{\tau_n}\sum_{k=1}^{n} \lambda_k x_k, \text{ where } \tau_n := \sum_{k=1}^{n} \lambda_k. \tag{1.7}$$

Lemma 1.1 (**Opial-Passty**). *Let Z be a nonempty subset of $\mathcal{H}$ and assume that the limit $\lim_{n\to+\infty} \|x_n - u\|$ exists for every element $u \in Z$. If every sequential weak cluster point of $(x_n)_{n\geqslant 0}$, respectively $(z_n)_{n\geqslant 1}$, lies in Z, then the sequence $(x_n)_{n\geqslant 0}$, respectively $(z_n)_{n\geqslant 1}$, converges weakly to an element in Z as $n \to +\infty$.*

Two following result can be found in [12, 19].

Lemma 1.2. *Let $(\theta_n)_{n\geqslant 0}, (\xi_n)_{n\geqslant 1}$ and $(\delta_n)_{n\geqslant 1}$ be sequences in $\mathbb{R}_+$ with $(\delta_n)_{n\geqslant 1} \in \ell^1$. If there exists $n_0 \geqslant 1$ such that*

$$\theta_{n+1} - \theta_n \leqslant \alpha_n(\theta_n - \theta_{n-1}) - \xi_n + \delta_n \quad \forall n \geqslant n_0$$

and α such that

$$0 \leqslant \alpha_n \leqslant \alpha < 1 \quad \forall n \geqslant 1,$$

then the following statements are true:

(i) $\sum_{n\geqslant 1} [\theta_n - \theta_{n-1}]_+ < +\infty$, *where* $[s]_+ := \max\{s, 0\}$;

(ii) *the limit* $\lim_{n\to\infty} \theta_n$ *exists.*

(iii) *the sequence* $(\xi_n)_{n\geqslant 1}$ *belongs to* ℓ^1.

The following result follows from Lemma 1.2, applied in case $\alpha_n := 0$ and $\theta_n := \rho_n - \rho$ for all $n \geqslant 1$, where ρ is a lower bound for $(\rho_n)_{n\geqslant 1}$.

Lemma 1.3. *Let $(\rho_n)_{n\geqslant 1}$ be a sequence in $\mathbb{R}$, which is bounded from below, and $(\xi_n)_{n\geqslant 1}$, $(\delta_n)_{n\geqslant 1}$ be sequences in $\mathbb{R}_+$ with $(\delta_n)_{n\geqslant 1} \in \ell^1$. If there exists $n_0 \geqslant 1$ such that*

$$\rho_{n+1} \leqslant \rho_n - \xi_n + \delta_n \quad \forall n \geqslant n_0,$$

then the following statements are true:

(i) *the sequence* $(\rho_n)_{n\geqslant 1}$ *is convergent.*

(ii) *the sequence* $(\xi_n)_{n\geqslant 1}$ *belongs to* ℓ^1.

The following result, which will be useful in this work, shows that statement (ii) in Lemma 1.3 can be obtained also when $(\rho_n)_{n\geqslant 1}$ is not bounded by below, but it has a particular form.

Lemma 1.4. *Let* $(\rho_n)_{n\geqslant 1}$ *be a sequence in* $\mathbb{R}$ *and* $(\xi_n)_{n\geqslant 1}$, $(\delta_n)_{n\geqslant 1}$ *be sequences in* $\mathbb{R}_+$ *with* $(\delta_n)_{n\geqslant 1} \in \ell^1$ *and*

$$\rho_n := \theta_n - \alpha_n\theta_{n-1} + \chi_n \quad \forall n \geqslant 1,$$

where $(\theta_n)_{n\geqslant 0}, (\chi_n)_{n\geqslant 1}$ *are sequences in* $\mathbb{R}_+$ *and there exists* α *such that*

$$0 \leqslant \alpha_n \leqslant \alpha < 1 \quad \forall n \geqslant 1.$$

If there exists $n_0 \geqslant 1$ *such that*

$$\rho_{n+1} - \rho_n \leqslant -\xi_n + \delta_n \quad \forall n \geqslant n_0, \tag{1.8}$$

then the sequence $(\xi_n)_{n\geqslant 1}$ *belongs to* ℓ^1.

Proof. We fix an integer $\bar{N} \geqslant n_0$, sum up the inequalities in (1.8) for $n = n_0, n_0+1, \cdots, \bar{N}$ and obtain

$$\rho_{\bar{N}+1} - \rho_{n_0} \leqslant -\sum_{n=n_0}^{\bar{N}} \xi_n + \sum_{n=n_0}^{\bar{N}} \delta_n \leqslant \sum_{n\geqslant 1} \delta_n < +\infty. \tag{1.9}$$

Hence the sequence $\{\rho_n\}_{n\geqslant 1}$ is bounded from above. Let $\bar{\rho} > 0$ be an upper bound of this sequence. For all $n \geqslant 1$ it holds

$$\theta_n - \alpha\theta_{n-1} \leqslant \theta_n - \alpha_n\theta_{n-1} + \chi_n = \rho_n \leqslant \bar{\rho},$$

from which we deduce that

$$-\rho_n \leqslant -\theta_n + \alpha\theta_{n-1} \leqslant \alpha\theta_{n-1}. \tag{1.10}$$

By induction we obtain for all $n \geqslant n_0 + 1$

$$\theta_n \leqslant \alpha\theta_{n-1} + \bar{\rho} \leqslant \cdots \leqslant \alpha^{n-n_0}\theta_{n_0} + \bar{\rho}\sum_{k=1}^{n-n_0} \alpha^{k-1} \leqslant \alpha^{n-n_0}\theta_{n_0} + \frac{\bar{\rho}}{1-\alpha}. \tag{1.11}$$

Then inequality (1.9) combined with (1.10) and (1.11) leads to

$$\begin{aligned} \sum_{n=n_0}^{\bar{N}} \xi_n &\leqslant \rho_{n_0} - \rho_{\bar{N}+1} + \sum_{n=n_0}^{\bar{N}} \delta_n \leqslant \rho_{n_0} + \alpha\theta_{\bar{N}} + \sum_{n\geqslant 1} \delta_n \\ &\leqslant \rho_{n_0} + \alpha^{\bar{N}-n_0+1}\theta_{n_0} + \frac{\alpha\bar{\rho}}{1-\alpha} + \sum_{n\geqslant 1} \delta_n < +\infty \end{aligned} \tag{1.12}$$

We let $\bar{N}$ converge to $+\infty$ and obtain that $\sum_{n\geqslant 1} \xi_n < +\infty$. □

2 The general monotone inclusion problem

In this section we address the following monotone inclusion problem.

Problem 2.1. *Let* $\mathcal{H}$ *be a real Hilbert space,* $A\colon \mathcal{H} \rightrightarrows \mathcal{H}$ *a maximally monotone operator,* $D\colon \mathcal{H} \to \mathcal{H}$ *an* $\eta-$*cocoercive with* $\eta > 0$, $B\colon \mathcal{H} \to \mathcal{H}$ *a* $\mu-$*cocoercive with* $\mu > 0$ *and assume that* $M := \operatorname{Zer} B \neq \varnothing$. *The monotone inclusion problem to solve reads*

$$0 \in Ax + Dx + N_M(x).$$

The following forward-backward penalty algorithm with inertial effects for solving Problem 2.1 will be in the focus of our investigations in this paper.

Algorithm 2.2. *Let $(\alpha_n)_{n\geqslant 1}, (\lambda_n)_{n\geqslant 1}$ and $(\beta_n)_{n\geqslant 1}$ be sequences of positive real numbers such that*

(C$_1$) $\{\lambda_n\}_{n\geqslant 1} \in \ell^2 \setminus \ell^1$;

(C$_2$) $\{\alpha_n\}_{n\geqslant 1}$ *is nondecreasing;*

(C$_3$) $0 \leqslant \alpha_n \leqslant \alpha < +\infty$ *for all* $n \geqslant 1$.

Let $x_0, x_1 \in \mathcal{H}$. For all $n \geqslant 1$ we set

$$x_{n+1} := J_{\lambda_n A}\left(x_n - \lambda_n D x_n - \lambda_n \beta_n B x_n + \alpha_n\left(x_n - x_{n-1}\right)\right).$$

When $D = 0$ and $B = \nabla h$, where $h : \mathcal{H} \to \mathbb{R}$ is a convex and differentiable function with $\mu^{-1}-$Lipschitz continuous gradient with $\mu > 0$ fulfilling $\min h = 0$, then Problem 2.1 recovers the monotone inclusion problem addressed in [9, Section 3] and Algorithm 2.2 can be seen as an inertial version of the iterative scheme considered in this paper. When $B = 0$, we have that $N_M = \{\mathbf{0}\}$ and Algorithm 2.2 is nothing else than the inertial version of the classical forward-backward algorithm (see for instance [13, 31]).

Hypotheses 2.3. *The convergence analysis will be carry out in the following hypotheses (see also [19]):*

(H$_1^{\text{fitz}}$) $A + N_M$ *is maximally monotone and* $\operatorname{Zer}\left(A + D + N_M\right) \neq \varnothing$;

(H$_2^{\text{fitz}}$) *for every* $p \in \operatorname{Ran}N_M$, $\sum_{n\geqslant 1} \lambda_n \beta_n \left[\sup_{u\in M} \varphi_B\left(u, \frac{p}{\beta_n}\right) - \sigma_M\left(\frac{p}{\beta_n}\right)\right] < +\infty.$

Since A and N_M are maximally monotone operators, the sum $A + N_M$ is maximally monotone, provided some specific regularity conditions are fulfilled (see [13, 17, 26, 40]). Furthermore, since D is also maximally monotone [13, Example 20.28] and $\operatorname{Dom}D \equiv \mathcal{H}$, if $A + N_M$ is maximally monotone, then $A + D + N_M$ is also maximally monotone.

Let us also notice that for $p \in \operatorname{Ran}N_M$ there exists $\widehat{u} \in M$ such that $p \in N_M\left(\widehat{u}\right)$, hence, for every $\beta > 0$ it holds

$$\sup_{u\in M} \varphi_B\left(u, \frac{p}{\beta}\right) - \sigma_M\left(\frac{p}{\beta}\right) \geqslant \left\langle \widehat{u}, \frac{p}{\beta}\right\rangle - \sigma_M\left(\frac{p}{\beta}\right) = 0.$$

For situations where (H$_2^{\text{fitz}}$) is satisfied we refer the reader [12, 24, 25, 37].

Before formulating the main theorem of this section we will prove some useful technical results.

Lemma 2.4. *Let $(x_n)_{n\geqslant 0}$ be the sequence generated by* Algorithm 2.2 *and (u, y) be an element in $\operatorname{Gr}\left(A + D + N_M\right)$ such that $y = v + Du + p$ with $v \in Au$ and $p \in N_M\left(u\right)$. Further, let $\varepsilon_1, \varepsilon_2, \varepsilon_3 > 0$ be such that $1 - \varepsilon_3 > 0$. Then the following inequality holds for all $n \geqslant 1$*

$$\begin{aligned}
&\left\|x_{n+1} - u\right\|^2 - \left\|x_n - u\right\|^2 \\
\leqslant{}& \alpha_n \left\|x_n - u\right\|^2 - \alpha_n \left\|x_{n-1} - u\right\|^2 - \left(1 - 4\varepsilon_1 - \varepsilon_2\right)\left\|x_{n+1} - x_n\right\|^2 \\
&+ \left(\alpha_n + \frac{\alpha_n^2}{4\varepsilon_1}\right)\left\|x_n - x_{n-1}\right\|^2 + \left(\frac{2}{\varepsilon_2}\lambda_n^2\beta_n^2 - 2\mu\left(1 - \varepsilon_3\right)\lambda_n\beta_n\right)\left\|Bx_n\right\|^2 \\
&+ \left(\frac{4}{\varepsilon_2}\lambda_n^2 - 2\eta\lambda_n\right)\left\|Dx_n - Du\right\|^2 + \frac{4}{\varepsilon_2}\lambda_n^2\left\|Du + v\right\|^2 \\
&+ 2\varepsilon_3\lambda_n\beta_n\left[\sup_{u\in M} \varphi_B\left(u, \frac{p}{\varepsilon_3\beta_n}\right) - \sigma_M\left(\frac{p}{\varepsilon_3\beta_n}\right)\right] + 2\lambda_n\left\langle u - x_n, y\right\rangle. \qquad (2.1)
\end{aligned}$$

Proof. Let $n \geqslant 1$ be fixed. According to definition of the resolvent of the operator A we have

$$x_n - x_{n+1} - \lambda_n \left(Dx_n + \beta_n Bx_n \right) + \alpha_n \left(x_n - x_{n-1} \right) \in \lambda_n A x_{n+1} \tag{2.2}$$

and, since $\lambda_n v \in \lambda_n A u$, the monotonicity of A guarantees

$$\left\langle x_{n+1} - u, x_n - x_{n+1} - \lambda_n \left(Dx_n + \beta_n Bx_n + v \right) + \alpha_n \left(x_n - x_{n-1} \right) \right\rangle \geqslant 0 \tag{2.3}$$

or, equivalently,

$$2 \left\langle u - x_{n+1}, x_n - x_{n+1} \right\rangle \leqslant 2\lambda_n \left\langle u - x_{n+1}, \beta_n Bx_n + Dx_n + v \right\rangle - 2\alpha_n \left\langle u - x_{n+1}, x_n - x_{n-1} \right\rangle . \tag{2.4}$$

For the term in the left-hand side of (2.4) we have

$$2 \left\langle u - x_{n+1}, x_n - x_{n+1} \right\rangle = \left\| x_{n+1} - u \right\|^2 + \left\| x_{n+1} - x_n \right\|^2 - \left\| x_n - u \right\|^2 . \tag{2.5}$$

Since

$$-2\alpha_n \left\langle u - x_n, x_n - x_{n-1} \right\rangle = -\alpha_n \left\| u - x_{n-1} \right\|^2 + \alpha_n \left\| u - x_n \right\|^2 + \alpha_n \left\| x_n - x_{n-1} \right\|^2$$

and

$$2 \left\langle x_{n+1} - x_n, \alpha_n \left(x_n - x_{n-1} \right) \right\rangle \leqslant 4\varepsilon_1 \left\| x_{n+1} - x_n \right\|^2 + \frac{\alpha_n^2}{4\varepsilon_1} \left\| x_n - x_{n-1} \right\|^2 ,$$

by adding the two inequalities, we obtain the following estimation for the second term in the right-hand side of (2.4)

$$\begin{aligned} & - 2\alpha_n \left\langle u - x_{n+1}, x_n - x_{n-1} \right\rangle \\ & \leqslant \alpha_n \left\| x_n - u \right\|^2 - \alpha_n \left\| x_{n-1} - u \right\|^2 + 4\varepsilon_1 \left\| x_{n+1} - x_n \right\|^2 + \left(\alpha_n + \frac{\alpha_n^2}{4\varepsilon_1} \right) \left\| x_n - x_{n-1} \right\|^2 . \end{aligned} \tag{2.6}$$

We turn now our attention to the first term in the right-hand side of (2.4), which can be written as

$$\begin{aligned} & 2\lambda_n \left\langle u - x_{n+1}, \beta_n Bx_n + Dx_n + v \right\rangle \\ = {} & 2\lambda_n \left\langle u - x_n, \beta_n Bx_n + Dx_n + v \right\rangle + 2\lambda_n \beta_n \left\langle x_n - x_{n+1}, Bx_n \right\rangle + 2\lambda_n \left\langle x_n - x_{n+1}, Dx_n + v \right\rangle . \end{aligned} \tag{2.7}$$

We have

$$2\lambda_n \beta_n \left\langle x_n - x_{n+1}, Bx_n \right\rangle \leqslant \frac{\varepsilon_2}{2} \left\| x_{n+1} - x_n \right\|^2 + \frac{2}{\varepsilon_2} \lambda_n^2 \beta_n^2 \left\| Bx_n \right\|^2 \tag{2.8}$$

and

$$\begin{aligned} 2\lambda_n \left\langle x_n - x_{n+1}, Dx_n + v \right\rangle & \leqslant \frac{\varepsilon_2}{2} \left\| x_{n+1} - x_n \right\|^2 + \frac{2}{\varepsilon_2} \lambda_n^2 \left\| Dx_n + v \right\|^2 \\ & \leqslant \frac{\varepsilon_2}{2} \left\| x_{n+1} - x_n \right\|^2 + \frac{4}{\varepsilon_2} \lambda_n^2 \left\| Dx_n - Du \right\|^2 + \frac{4}{\varepsilon_2} \lambda_n^2 \left\| Du + v \right\|^2 . \end{aligned} \tag{2.9}$$

On the other hand, we have

$$\begin{aligned} & 2\lambda_n \left\langle u - x_n, \beta_n Bx_n + Dx_n + v \right\rangle \\ = {} & 2\lambda_n \beta_n \left\langle u - x_n, Bx_n \right\rangle + 2\lambda_n \left\langle u - x_n, Dx_n - Du \right\rangle + 2\lambda_n \left\langle u - x_n, Du + v \right\rangle . \end{aligned} \tag{2.10}$$

Since $0 < \varepsilon_3 < 1$ and $Bu = 0$, the cocoercivity of B gives us

$$2\lambda_n \beta_n \left\langle u - x_n, Bx_n \right\rangle \leqslant -2\mu \left(1 - \varepsilon_3 \right) \lambda_n \beta_n \left\| Bx_n \right\|^2 + 2\varepsilon_3 \lambda_n \beta_n \left\langle u - x_n, Bx_n \right\rangle . \tag{2.11}$$

Similarly, the cocoercivity of D gives us

$$2\lambda_n \langle u - x_n, Dx_n - Du \rangle \leqslant -2\eta\lambda_n \|Dx_n - Du\|^2 . \tag{2.12}$$

Combining (2.11) - (2.12) with (2.10) and by using the definition Fitzpatrick function and the fact that $\sigma_M\left(\frac{p}{\varepsilon_3\beta_n}\right) = \left\langle u, \frac{p}{\varepsilon_3\beta_n}\right\rangle$, we obtain

$$\begin{aligned}
& 2\lambda_n \langle u - x_n, \beta_n Bx_n + Dx_n + v \rangle \\
\leqslant & -2\mu(1-\varepsilon_3)\lambda_n\beta_n \|Bx_n\|^2 + 2\varepsilon_3\lambda_n\beta_n \langle u - x_n, Bx_n\rangle - 2\eta\lambda_n \|Dx_n - Du\|^2 \\
& + 2\lambda_n \langle u - x_n, Du + v\rangle \\
= & -2\mu(1-\varepsilon_3)\lambda_n\beta_n \|Bx_n\|^2 + 2\varepsilon_3\lambda_n\beta_n \langle u - x_n, Bx_n\rangle - 2\eta\lambda_n \|Dx_n - Du\|^2 \\
& + 2\lambda_n \langle u - x_n, y - p\rangle \\
= & -2\mu(1-\varepsilon_3)\lambda_n\beta_n \|Bx_n\|^2 - 2\eta\lambda_n \|Dx_n - Du\|^2 + 2\lambda_n \langle u - x_n, y\rangle \\
& + 2\varepsilon_3\lambda_n\beta_n \left(\langle u, Bx_n\rangle + \left\langle x_n, \frac{p}{\varepsilon_3\beta_n}\right\rangle - \langle x_n, Bx_n\rangle - \left\langle u, \frac{p}{\varepsilon_3\beta_n}\right\rangle\right) \\
\leqslant & -2\mu(1-\varepsilon_3)\lambda_n\beta_n \|Bx_n\|^2 - 2\eta\lambda_n \|Dx_n - Du\|^2 + 2\lambda_n \langle u - x_n, y\rangle \\
& + 2\varepsilon_3\lambda_n\beta_n \left[\sup_{u\in M} \varphi_B\left(u, \frac{p}{\varepsilon_3\beta_n}\right) - \sigma_M\left(\frac{p}{\varepsilon_3\beta_n}\right)\right].
\end{aligned} \tag{2.13}$$

The inequalities (2.8), (2.9) and (2.13) lead to

$$\begin{aligned}
& 2\lambda_n \langle u - x_{n+1}, \beta_n Bx_n + Dx_n + v\rangle \\
\leqslant & \left(\frac{2}{\varepsilon_2}\lambda_n^2\beta_n^2 - 2\mu(1-\varepsilon_3)\lambda_n\beta_n\right)\|Bx_n\|^2 + \left(\frac{4}{\varepsilon_2}\lambda_n^2 - 2\eta\lambda_n\right)\|Dx_n - Du\|^2 + \varepsilon_2\|x_{n+1} - x_n\|^2 \\
& + \frac{4}{\varepsilon_2}\lambda_n^2\|Du + v\|^2 + 2\varepsilon_3\lambda_n\beta_n\left[\sup_{u\in M}\varphi_B\left(u, \frac{p}{\varepsilon_3\beta_n}\right) - \sigma_M\left(\frac{p}{\varepsilon_3\beta_n}\right)\right] + 2\lambda_n\langle u - x_n, y\rangle .
\end{aligned} \tag{2.14}$$

Finally, by combining (2.5), (2.6) and (2.14), we obtain (2.1). □

From now on we will assume that for $0 < \alpha < \frac{1}{3}$ the constants $\varepsilon_1, \varepsilon_2, \varepsilon_3 > 0$ and the sequences $(\lambda_n)_{n\geqslant 1}$ and $(\beta_n)_{n\geqslant 1}$ are chosen such that

$$(\mathrm{C}_4) \quad 1 - \varepsilon_3 > 0, \quad \varepsilon_2 < 1 - 4\varepsilon_1 - \alpha - \frac{\alpha^2}{4\varepsilon_1} \quad \text{and} \quad \sup_{n\geqslant 1}\lambda_n\beta_n < \mu\varepsilon_2(1-\varepsilon_3).$$

As a consequence, there exists $0 < s \leqslant 1 - \dfrac{\varepsilon_1}{1 - 3\varepsilon_1 - \varepsilon_2}\left(1 + \dfrac{\alpha}{2\varepsilon_1}\right)^2$, which means that for all $n \geqslant 1$ it holds

$$\alpha_{n+1} + \frac{\alpha_{n+1}^2}{4\varepsilon_1} - (1 - 4\varepsilon_1 - \varepsilon_3) \leqslant \alpha + \frac{\alpha^2}{4\varepsilon_1} - (1 - 4\varepsilon_1 - \varepsilon_3) < -s, \tag{2.15}$$

On the other hand, there exists $0 < t \leqslant \mu(1-\varepsilon_2) - \dfrac{1}{\varepsilon_3}\sup_{n\geqslant 0}\lambda_n\beta_n$, which means that for all $n \geqslant 1$ it holds

$$\frac{1}{\varepsilon_3}\lambda_n\beta_n - \mu(1-\varepsilon_2) \leqslant -t. \tag{2.16}$$

Remark 2.5. (i) Since $0 < \alpha < \frac{1}{3}$, one can always find $\varepsilon_1, \varepsilon_2 > 0$ such that $\varepsilon_2 < 1 - 4\varepsilon_1 - \alpha - \dfrac{\alpha^2}{4\varepsilon_1}$. One possible choice is $\varepsilon_1 = \frac{\alpha}{4}$ and $0 < \varepsilon_2 < 1 - 3\alpha$. From the second inequality in (C_4) it follows that $1 - 3\varepsilon_1 - \varepsilon_2 > \varepsilon_1 + \alpha + \dfrac{\alpha^2}{4\varepsilon_1} > 0$.

(ii) As

$$1-\frac{\varepsilon_1}{1-3\varepsilon_1-\varepsilon_2}\left(1+\frac{\alpha}{2\varepsilon_1}\right)^2=\frac{1}{1-3\varepsilon_1-\varepsilon_2}\left(1-4\varepsilon_1-\varepsilon_2-\alpha-\frac{\alpha^2}{4\varepsilon_1}\right)>0,$$

it is always possible to choose s such that $0<s\leqslant 1-\dfrac{\varepsilon_1}{1-3\varepsilon_1-\varepsilon}\left(1+\dfrac{\alpha}{2\varepsilon_1}\right)^2$. Since in this case $s<1-4\varepsilon_1-\varepsilon_2-\alpha-\dfrac{\alpha^2}{4\varepsilon_1}$, one has (2.15).

The following proposition brings us closer to the convergence result.

Proposition 2.6. *Let $0<\alpha<\frac{1}{3}$, $\varepsilon_1,\varepsilon_2,\varepsilon_3>0$ and the sequences $(\lambda_n)_{n\geqslant 1}$ and $(\beta_n)_{n\geqslant 1}$ satisfy condition* (C_4). *Let $(x_n)_{n\geqslant 0}$ be the sequence generated by Algorithm 2.2 and assume that the Hypotheses 2.3 are verified. Then the following statements are true:*

(i) *the sequence $(\|x_{n+1}-x_n\|)_{n\geqslant 0}$ belongs to ℓ^2 and the sequence $\left(\lambda_n\beta_n\|Bx_n\|^2\right)_{n\geqslant 1}$ belongs to ℓ^1;*

(ii) *if, moreover, $\liminf_{n\to+\infty}\lambda_n\beta_n>0$, then $\lim_{n\to+\infty}\|Bx_n\|=0$ and thus every cluster point of the sequence $(x_n)_{n\geqslant 0}$ lies in M.*

(iii) *for every $u\in\operatorname{Zer}(A+D+N_M)$, the limit $\lim_{n\to+\infty}\|x_n-u\|$ exists.*

Proof. Since $\lim_{n\to+\infty}\lambda_n=0$, there exists a integer $n_1\geqslant 1$ such that $\lambda_n\leqslant\dfrac{2}{\varepsilon_2}\eta$ for all $n\geqslant n_0$. According to Lemma 2.4, for every $(u,y)\in\operatorname{Gr}(A+D+N_M)$ such that $y=v+Du+p$, with $v\in Au$ and $p\in N_M(u)$, and all $n\geqslant n_0$ the following inequality holds

$$\begin{aligned}
&\|x_{n+1}-u\|^2-\|x_n-u\|^2\\
&\leqslant\alpha_n\|x_n-u\|^2-\alpha_n\|x_{n-1}-u\|^2-(1-4\varepsilon_1-\varepsilon_2)\|x_{n+1}-x_n\|^2\\
&\quad+\left(\alpha_n+\frac{\alpha_n^2}{4\varepsilon_1}\right)\|x_n-x_{n-1}\|^2+\left(\frac{2}{\varepsilon_2}\lambda_n\beta_n-2\mu(1-\varepsilon_3)\right)\lambda_n\beta_n\|Bx_n\|^2\\
&\quad+\frac{4}{\varepsilon_2}\lambda_n^2\|Du+v\|^2+2\varepsilon_3\lambda_n\beta_n\left[\sup_{u\in M}\varphi_B\left(u,\frac{p}{\varepsilon_3\beta_n}\right)-\sigma_M\left(\frac{p}{\varepsilon_3\beta_n}\right)\right]+2\lambda_n\langle u-x_n,y\rangle.
\end{aligned}\tag{2.17}$$

We consider $u\in\operatorname{Zer}(A+D+N_M)$, which means that we can take $y=0$ in (2.17). For all $n\geqslant 1$ we denote

$$\theta_n:=\|x_n-u\|^2,\quad\rho_n:=\theta_n-\alpha_n\theta_{n-1}+\left(\alpha_n+\frac{\alpha_n^2}{4\varepsilon_1}\right)\|x_n-x_{n-1}\|^2\tag{2.18}$$

and

$$\delta_n:=\frac{4}{\varepsilon_2}\lambda_n^2\|Du+v\|^2+2\varepsilon_3\lambda_n\beta_n\left[\sup_{u\in M}\varphi_B\left(u,\frac{p}{\varepsilon_3\beta_n}\right)-\sigma_M\left(\frac{p}{\varepsilon_3\beta_n}\right)\right].\tag{2.19}$$

Using that $(\alpha_n)_{n\geqslant 1}$ is nondecreasing, for all $n\geqslant n_0$ it yields

$$\begin{aligned}
\rho_{n+1}-\rho_n&\leqslant\left(\alpha_{n+1}+\frac{\alpha_{n+1}^2}{4\varepsilon_1}-(1-4\varepsilon_1-\varepsilon_2)\right)\|x_{n+1}-x_n\|^2\\
&\quad+\left(\frac{2}{\varepsilon_3}\lambda_n\beta_n-2\mu(1-\varepsilon_2)\right)\lambda_n\beta_n\|Bx_n\|^2+\delta_n\\
&\leqslant -s\|x_{n+1}-x_n\|^2-2t\lambda_n\beta_n\|Bx_n\|^2+\delta_n,
\end{aligned}\tag{2.20}$$

where $s,t>0$ are chosen according to (2.15) and (2.16), respectively.

Thanks to (H_2^{fitz}) and (C_1) it holds

$$\sum_{n\geqslant 1}\delta_n=\frac{4}{\varepsilon_2}\left\|Du+v\right\|^2\sum_{n\geqslant 1}\lambda_n^2+2\sum_{n\geqslant 1}\varepsilon_3\lambda_n\beta_n\left[\sup_{u\in M}\varphi_B\left(u,\frac{p}{\varepsilon_3\beta_n}\right)-\sigma_M\left(\frac{p}{\varepsilon_3\beta_n}\right)\right]<+\infty. \tag{2.21}$$

Hence, according to Lemma 1.4, we obtain

$$\sum_{n\geqslant 0}\left\|x_{n+1}-x_n\right\|^2<+\infty \text{ and } \sum_{n\geqslant 1}\lambda_n\beta_n\left\|Bx_n\right\|^2<+\infty, \tag{2.22}$$

which proves (i). If, in addtion $\liminf_{n\to\infty}\lambda_n\beta_n>0$, then $\lim_{n\to+\infty}\left\|Bx_n\right\|=0$, which means every cluster point of the sequence $(x_n)_{n\geqslant 0}$ lies in Zer $B=M$.

In order to prove (iii), we consider again the inequality (2.17) for an arbitrary element $u\in \operatorname{Zer}(A+D+N_M)$ and $y=0$. With the notations in (2.18) and (2.19), we get for all $n\geqslant n_0$

$$\theta_{n+1}-\theta_n\leqslant\alpha_n\left(\theta_n-\theta_{n-1}\right)+\left(\alpha_n+\frac{\alpha_n^2}{4\varepsilon_1}\right)\left\|x_n-x_{n-1}\right\|^2+\delta_n. \tag{2.23}$$

According to (2.21) and (2.22) we have

$$\sum_{n\geqslant 1}\left(\alpha_n+\frac{\alpha_n^2}{4\varepsilon_1}\right)\left\|x_n-x_{n-1}\right\|^2+\sum_{n\geqslant 1}\delta_n\leqslant\left(\alpha+\frac{\alpha^2}{4\varepsilon_1}\right)\sum_{n\geqslant 1}\left\|x_n-x_{n-1}\right\|^2+\sum_{n\geqslant 1}\delta_n<+\infty, \tag{2.24}$$

therefore, by Lemma 1.2, the limit $\lim_{n\to+\infty}\theta_n=\lim_{n\to+\infty}\left\|x_n-u\right\|^2$ exists, which means that the limit $\lim_{n\to+\infty}\left\|x_n-u\right\|$ exists, too. $\square$

Remark 2.7. The condition (C_3) that we imposed in combination with $0<\alpha<\frac{1}{3}$ on the sequence of inertial parameters $(\alpha_n)_{n\geqslant 1}$ is the one proposed in [3, Proposition 2.1] when addressing the convergence of the inertial proximal point algorithm. However, the statements in proposition above and in the following convergence theorem remain valid if one alternatively assumes that there exists α' such that $0\leqslant\alpha_n\leqslant\alpha'<1$ for all $n\geqslant 1$ and

$$\sum_{n\geqslant 1}\left(\alpha_n+\frac{\alpha_n^2}{4\varepsilon_1}\right)\left\|x_n-x_{n-1}\right\|^2<+\infty.$$

This can be realized if one chooses for a fixed $p>1$

$$\alpha_n\leqslant\min\left\{\alpha',2\varepsilon_1\left(-1+\sqrt{1+n^{-p}\left\|x_n-x_{n-1}\right\|^{-2}}\right)\right\}\ \forall n\geqslant 1.$$

Indeed, in this situation we have that $\frac{\alpha_n^2}{4\varepsilon_1}+\alpha_n-\frac{1}{n^p\left\|x_n-x_{n-1}\right\|^2}\leqslant 0$ for all $n\geqslant 1$, which gives

$$\sum_{n\geqslant 1}\left(\alpha_n+\frac{\alpha_n^2}{4\varepsilon_1}\right)\left\|x_n-x_{n-1}\right\|^2\leqslant\sum_{n\geqslant 1}\frac{1}{n^p}<+\infty.$$

Now we are ready to prove the main theorem of this section, which addresses the convergence of the sequence generated by Algorithm 2.2.

Theorem 2.8. *Let $0<\alpha<\frac{1}{3}$, $\varepsilon_1,\varepsilon_2,\varepsilon_3>0$ and the sequences $(\lambda_n)_{n\geqslant 1}$ and $(\beta_n)_{n\geqslant 1}$ satisfy condition (C_4). Let $(x_n)_{n\geqslant 0}$ be the sequence generated by Algorithm 2.2, $(z_n)_{n\geqslant 1}$ be the sequence defined in (1.7) and assume that the Hypotheses 2.3 are verified. Then the following statements are true:*

(i) *the sequence* $(z_n)_{n\geqslant 1}$ *converges weakly to an element in* $\operatorname{Zer}(A+D+N_M)$ *as* $n\to+\infty$.

(ii) *if* A *is* $\gamma-$*strongly monotone with* $\gamma>0$, *then* $(x_n)_{n\geqslant 0}$ *converges strongly to the unique element in* $\operatorname{Zer}(A+D+N_M)$ *as* $n\to+\infty$.

Proof. (i) According to Proposition 2.6 (iii), the limit $\lim_{n\to+\infty}\|x_n-u\|$ exisits for every $u\in\operatorname{Zer}(A+D+N_M)$. Let z be a sequential weak cluster point of $(z_n)_{n\geqslant 1}$. We will show that $z\in\operatorname{Zer}(A+D+N_M)$, by using the characterization (1.5) of the maximal monotonicity, and the conclusion will follow by Lemma 1.1.

To this end we consider an arbitrary $(u,y)\in\operatorname{Gr}(A+D+N_M)$ such that $y=v+Du+p$, where $v\in Au$ and $p\in N_M(u)$. From (2.17), with the notations (2.18) and (2.19), we have for all $n\geqslant n_0$

$$\begin{aligned}&\rho_{n+1}-\rho_n\\&\leqslant -s\|x_{n+1}-x_n\|^2-2t\lambda_n\beta_n\|Bx_n\|^2+\delta_n+2\lambda_n\langle u-x_n,y\rangle\leqslant\delta_n+2\lambda_n\langle u-x_n,y\rangle.\end{aligned}\tag{2.25}$$

Recall that from (2.21) that $\sum_{n\geqslant 1}\delta_n<+\infty$. Since $(x_n)_{n\geqslant 0}$ is bounded, the sequence $(\rho_n)_{n\geqslant 1}$ is also bounded.

We fix an arbitrary integer $\bar{N}\geqslant n_0$ and sum up the inequalities in (2.25) for $n=n_0+1,n_0+2,\cdots,\bar{N}$. This yields

$$\rho_{\bar{N}+1}-\rho_{n_0+1}\leqslant\sum_{n\geqslant 1}\delta_n+2\left\langle-\sum_{n=1}^{n_0}\lambda_n u+\sum_{n=1}^{n_0}\lambda_n x_n,y\right\rangle+2\left\langle\tau_{\bar{N}}u-\sum_{n=1}^{\bar{N}}\lambda_n x_n,y\right\rangle.$$

After dividing this last inequality by $2\tau_{\bar{N}}=2\sum_{n=1}^{\bar{N}}\lambda_n$, we obtain

$$\frac{1}{2\tau_{\bar{N}}}\left(\rho_{\bar{N}+1}-\rho_{n_0+1}\right)\leqslant\frac{1}{2\tau_{\bar{N}}}T+2\langle u-z_{\bar{N}},y\rangle,\tag{2.26}$$

where $T:=\sum_{n\geqslant 1}\delta_n+2\left\langle-\sum_{n=1}^{n_0}\lambda_n u+\sum_{n=1}^{n_0}\lambda_n x_n,y\right\rangle\in\mathbb{R}$. By passing in (2.26) to the limit and by using that $\lim_{N\to\infty}\tau_{\bar{N}}=\lim_{\bar{N}\to\infty}\sum_{n=1}^{\bar{N}}\lambda_n=+\infty$, we get

$$\liminf_{\bar{N}\to\infty}\langle u-z_{\bar{N}},y\rangle\geqslant 0.$$

As z is a sequential weak cluster point of $(z_n)_{n\geqslant 1}$, the above inequality gives us $\langle u-z,y\rangle\geqslant 0$, which finally means that $z\in\operatorname{Zer}(A+D+N_M)$.

(ii) Let $u\in\mathcal{H}$ be the unique element in $\operatorname{Zer}(A+D+N_M)$. Since A is $\gamma-$strongly monotone with $\gamma>0$, the formula in (2.3) reads for all $n\geqslant 1$

$$\langle x_{n+1}-u,x_n-x_{n+1}-\lambda_n(Dx_n+\beta_n Bx_n+v)+\alpha_n(x_n-x_{n-1})\rangle\geqslant\gamma\lambda_n\|x_{n+1}-u\|^2$$

or, equivalently,

$$\begin{aligned}&2\gamma\lambda_n\|x_{n+1}-u\|^2+2\langle u-x_{n+1},x_n-x_{n+1}\rangle\\&\leqslant 2\lambda_n\langle u-x_{n+1},\beta_n Bx_n+Dx_n+v\rangle-2\alpha_n\langle u-x_{n+1},x_n-x_{n-1}\rangle.\end{aligned}$$

By using again (2.5), (2.6) and (2.14) we obtain for all $n \geqslant 1$

$$\begin{aligned}
&2\gamma\lambda_n \left\|x_{n+1}-u\right\|^2 + \left\|x_{n+1}-u\right\|^2 - \left\|x_n-u\right\|^2 \\
&\leqslant \alpha_n \left\|x_n-u\right\|^2 - \alpha_n \left\|x_{n-1}-u\right\|^2 - \left(1-4\varepsilon_1-\varepsilon_2\right)\left\|x_{n+1}-x_n\right\|^2 \\
&\quad + \left(\alpha_n + \frac{\alpha_n^2}{4\varepsilon_1}\right)\left\|x_n-x_{n-1}\right\|^2 + \left(\frac{2}{\varepsilon_2}\lambda_n^2\beta_n^2 - 2\mu\left(1-\varepsilon_3\right)\lambda_n\beta_n\right)\left\|Bx_n\right\|^2 \\
&\quad + \left(\frac{4}{\varepsilon_2}\lambda_n^2 - 2\eta\lambda_n\right)\left\|Dx_n-Du\right\|^2 + \frac{4}{\varepsilon_2}\lambda_n^2\left\|Du+v\right\|^2 \\
&\quad + 2\varepsilon_3\lambda_n\beta_n\left[\sup_{u\in M}\varphi_B\left(u,\frac{p}{\varepsilon_3\beta_n}\right) - \sigma_M\left(\frac{p}{\varepsilon_3\beta_n}\right)\right] + 2\lambda_n\left\langle u-x_n, y\right\rangle.
\end{aligned}$$

By using the notations in (2.18) and (2.19), this yields for all $n \geqslant 1$

$$2\gamma\lambda_n \left\|x_{n+1}-u\right\|^2 + \theta_{n+1} - \theta_n \leqslant \alpha_n\left(\theta_n-\theta_{n-1}\right) + \left(\alpha_n + \frac{\alpha_n^2}{4\varepsilon_1}\right)\left\|x_n-x_{n-1}\right\|^2 + \delta_n$$

By taking into account (2.24), from Lemma 1.2 we get

$$2\gamma\sum_{n\geqslant 1}\lambda_n\left\|x_n-u\right\|^2 < +\infty.$$

According to (C_1) we have $\sum_{n\geqslant 1}\lambda_n = +\infty$, which implies that the limit $\lim_{n\to\infty}\left\|x_n-u\right\|$ must be equal to zero. This provides the desired conclusion. □

3 Applications to convex bilevel programming

We will employ the results obtained in the previous section, in the context of monotone inclusions, to the solving of convex bilevel programming problems.

Problem 3.1. *Let $\mathcal{H}$ be a real Hilbert space, $f\colon \mathcal{H}\to\bar{\mathbb{R}}$ a proper, convex and lower semicontinuous function and $g,h\colon \mathcal{H}\to\mathbb{R}$ differentiable functions with L_g-Lipschitz continuous and, respectively, L_h-Lipschitz continuous gradients. Suppose that* $\arg\min h \neq \varnothing$ *and* $\min h = 0$. *The bilevel programming problem to solve reads*

$$\min_{x\in\arg\min h} f(x) + g(x).$$

The assumption $\min h = 0$ is not a restricttive as, otherwise, one can replace h with $h-\min h$.

Hypotheses 3.2. *The convergence analysis will be carry out in the following hypotheses:*

(H_1^{prog}) $\partial f + N_{\arg\min h}$ *is maximally monotone and* $\mathcal{S} := \arg\min_{x\in\arg\min h}\{f(x)+g(x)\} \neq \varnothing$*;*

(H_2^{prog}) *for every* $p \in \operatorname{Ran} N_{\arg\min h}$, $\sum_{n\geqslant 1}\lambda_n\beta_n\left[h^*\left(\frac{p}{\beta_n}\right) - \sigma_{\arg\min h}\left(\frac{p}{\beta_n}\right)\right] < +\infty$.

In the above hypotheses, we have that $\partial f + \nabla g + N_{\arg\min h} = \partial\left(f+g+\delta_{\arg\min h}\right)$ and hence $\mathcal{S} = \operatorname{Zer}\left(\partial f + \nabla g + N_{\arg\min h}\right) \neq \varnothing$. Since according to the Theorem of Baillon-Haddad (see, for example, [13, Corollary 18.16]), ∇g and ∇h are L_g^{-1}-cocoercive and, respectively, L_h^{-1}-cocoercive, and $\arg\min h = \operatorname{Zer}\nabla h$ solving the bilevel programming problem in Problem 3.1 reduces to solving the monotone inclusion

$$0 \in \partial f(x) + \nabla g(x) + N_{\arg\min h}(x).$$

By using to this end Algorithm 2.2, we recieve the following iterative scheme.

Algorithm 3.3. *Let $(\alpha_n)_{n\geqslant 1}, (\lambda_n)_{n\geqslant 1}$ and $(\beta_n)_{n\geqslant 1}$ be sequences of positive real numbers such that*

(C_1) $\{\lambda_n\}_{n\geqslant 1} \in \ell^2 \setminus \ell^1$;

(C_2) $\{\alpha_n\}_{n\geqslant 1}$ *is nondecreasing;*

(C_3) *there exists α with $0 \leqslant \alpha_n \leqslant \alpha < 1/3$ for all $n \geqslant 1$.*

Let $x_0, x_1 \in \mathcal{H}$. For all $n \geqslant 1$ we set

$$x_{n+1} := \operatorname{prox}_{\lambda_n f}\left(x_n - \lambda_n \nabla g\left(x_n\right) - \lambda_n \beta_n \nabla h\left(x_n\right) + \alpha_n\left(x_n - x_{n-1}\right)\right).$$

By using the inequality (1.6), one can easily notice, that (H_2^{prog}) implies (H_2^{fitz}), which means that the convergence statements for Algorithm 3.3 can be derived as particular instances of the ones derived in the previous section.

Alternatively, one can use to this end the following lemma and employ the same ideas and techniques as in Section 2. Lemma 3.4 is similar to Lemma 2.4, however, it will allow us to provide convergence statements also for the sequence of function values $(h(x_n))_{n\geqslant 0}$.

Lemma 3.4. *Let $(x_n)_{n\geqslant 0}$ be the sequence generated by* Algorithm 3.3 *and (u, y) be an element in* $\operatorname{Gr}\left(\partial f + \nabla g + N_{\arg\min h}\right)$ *such that $y = v + \nabla g(u) + p$ with $v \in \partial f(u)$ and $p \in N_{\arg minh}(u)$. Further, let $\varepsilon_1, \varepsilon_2, \varepsilon_3 > 0$ be such that $1 - \varepsilon_3 > 0$. Then the following inequality holds for all $n \geqslant 1$*

$$\begin{aligned}
&\left\|x_{n+1} - u\right\|^2 - \left\|x_n - u\right\|^2 \\
\leqslant\ & \alpha_n\left\|x_n - u\right\|^2 - \alpha_n\left\|x_{n-1} - u\right\|^2 - \left(1 - 4\varepsilon_1 - \varepsilon_2\right)\left\|x_{n+1} - x_n\right\|^2 + \left(\alpha_n + \frac{\alpha_n^2}{4\varepsilon_1}\right)\left\|x_n - x_{n-1}\right\|^2 \\
&\left(\frac{2}{\varepsilon_2}\lambda_n^2\beta_n^2 - 2\mu\left(1 - \varepsilon_3\right)\lambda_n\beta_n\right)\left\|\nabla h\left(x_n\right)\right\|^2 + \left(\frac{4}{\varepsilon_2}\lambda_n^2 - 2\eta\lambda_n\right)\left\|\nabla g\left(x_n\right) - \nabla g\left(u\right)\right\|^2 \\
&+ \lambda_n\beta_n\left[h\left(u\right) - h\left(x_n\right)\right] + \frac{4}{\varepsilon_2}\lambda_n^2\left\|v + \nabla g\left(u\right)\right\|^2 \\
&+ \varepsilon_3\lambda_n\beta_n\left[h^*\left(\frac{2p}{\varepsilon_3\beta_n}\right) - \sigma_{\arg\min h}\left(\frac{2p}{\varepsilon_3\beta_n}\right)\right] + 2\lambda_n\left\langle u - x_n, y\right\rangle.
\end{aligned}$$

Proof. Let be $n \geqslant 1$ fixed. The proof follows by combining the estimates used in the proof of Lemma 2.4 with some inequalities which better exploits the convexity of h. From (2.11) we have

$$2\lambda_n\beta_n\left\langle u - x_n, \nabla h\left(x_n\right)\right\rangle \leqslant -2\mu\left(1 - \varepsilon_3\right)\lambda_n\beta_n\left\|\nabla h\left(x_n\right)\right\|^2 + 2\varepsilon_3\lambda_n\beta_n\left\langle u - x_n, \nabla h\left(x_n\right)\right\rangle.$$

Since h is convex, the following relation also hold

$$2\lambda_n\beta_n\left\langle u - x_n, \nabla h\left(x_n\right)\right\rangle \leqslant 2\lambda_n\beta_n\left[h\left(u\right) - h\left(x_n\right)\right].$$

Summing up the two inequalities above give us

$$\begin{aligned}
2\lambda_n\beta_n\left\langle u - x_n, \nabla h\left(x_n\right)\right\rangle \leqslant & -\mu\left(1 - \varepsilon_3\right)\lambda_n\beta_n\left\|\nabla h\left(x_n\right)\right\|^2 + \varepsilon_3\lambda_n\beta_n\left\langle u - x_n, \nabla h\left(x_n\right)\right\rangle \\
& + \lambda_n\beta_n\left[h\left(u\right) - h\left(x_n\right)\right].
\end{aligned}$$

Using the same techniques as in the derivation of (2.13), we get

$$\begin{aligned}
&2\lambda_n\left\langle u - x_n, v + \nabla g\left(x_n\right) + \beta_n\nabla h\left(x_n\right)\right\rangle \\
\leqslant & -\mu\left(1 - \varepsilon_3\right)\lambda_n\beta_n\left\|\nabla h\left(x_n\right)\right\|^2 - 2\eta\lambda_n\left\|\nabla g\left(x_n\right) - \nabla g\left(u\right)\right\|^2 + \lambda_n\beta_n\left[h\left(u\right) - h\left(x_n\right)\right] \\
&+ 2\lambda_n\left\langle u - x_n, y\right\rangle + \varepsilon_3\lambda_n\beta_n\left[h^*\left(u, \frac{2p}{\varepsilon_3\beta_n}\right) - \sigma_{\arg\min h}\left(\frac{2p}{\varepsilon_3\beta_n}\right)\right].
\end{aligned}$$

With this improved estimates, the conclusion follows as in the proof of Lemma 2.4. □

By using now Lemma 3.4, one obains, after slightly adapting the proof of Proposition 2.6, the following result.

Proposition 3.5. *Let $0 < \alpha < \frac{1}{3}$, $\varepsilon_1, \varepsilon_2, \varepsilon_3 > 0$ and the sequences $(\lambda_n)_{n\geqslant 1}$ and $(\beta_n)_{n\geqslant 1}$ satisfy condition* (C$_4$)*. Let $(x_n)_{n\geqslant 0}$ be the sequence generated by Algorithm 3.3 and assume that the Hypotheses 3.2 are verified. Then the following statements are true:*

(i) *the sequence $(\|x_{n+1} - x_n\|)_{n\geqslant 0}$ belongs to ℓ^2 and the sequences $\left(\lambda_n\beta_n \|\nabla h(x_n)\|^2\right)_{n\geqslant 1}$ and $(\lambda_n\beta_n h(x_n))_{n\geqslant 1}$ belong to ℓ^1;*

(ii) *if, moreover, $\liminf_{n\to+\infty} \lambda_n\beta_n > 0$, then $\lim_{n\to+\infty} \|\nabla h(x_n)\| = \lim_{n\to+\infty} h(x_n) = 0$ and thus every cluster point of the sequence $(x_n)_{n\geqslant 0}$ lies in* $\arg\min h$*.*

(iii) *for every $u \in \mathcal{S}$, the limit $\lim_{n\to+\infty} \|x_n - u\|$ exists.*

Finally, the above proposition leads to the following convergence result.

Theorem 3.6. *Let $0 < \alpha < \frac{1}{3}$, $\varepsilon_1, \varepsilon_2, \varepsilon_3 > 0$ and the sequences $(\lambda_n)_{n\geqslant 1}$ and $(\beta_n)_{n\geqslant 1}$ satisfy condition* (C$_4$)*. Let $(x_n)_{n\geqslant 0}$ be the sequence generated by Algorithm 3.3, $(z_n)_{n\geqslant 1}$ be the sequence defined in* (1.7) *and assume that the Hypotheses 3.2 are verified. Then the following statements are true:*

(i) *the sequence $(z_n)_{n\geqslant 1}$ converges weakly to an element in $\mathcal{S}$ as $n \to +\infty$.*

(ii) *if f is $\gamma-$strongly convex with $\gamma > 0$, then $(x_n)_{n\geqslant 0}$ converges strongly to the unique element in $\mathcal{S}$ as $n \to +\infty$.*

As follows we will show that under inf-compactness assumptions one can achieve weak nonergodic convergence for the sequence $(x_n)_{n\geqslant 0}$. Weak nonergodic convergence has been obtained for Algorithm 3.3 in [25] when $\alpha_n = \alpha$ for all $n \geqslant 1$ and for restrictive choices for both the sequence of step sizes and penalty parameters.

We denote by $(f+g)_* = \min_{x\in\arg\min h} (f(x) + g(x))$. For every element x in $\mathcal{H}$, we denote by $\operatorname{dist}(x, \mathcal{S}) = \inf_{u\in\mathcal{S}} \|x - u\|$ the distance from x to $\mathcal{S}$. In particular, $\operatorname{dist}(x, \mathcal{S}) = \|x - \mathbf{Pr}_{\mathcal{S}}x\|$, where $\mathbf{Pr}_{\mathcal{S}}x$ denotes the projection of x onto $\mathcal{S}$. The projection operator $\mathbf{Pr}_{\mathcal{S}}$ is *firmly nonexpansive* ([13, Proposition 4.8]), this means

$$\|\mathbf{Pr}_{\mathcal{S}}(x) - \mathbf{Pr}_{\mathcal{S}}(y)\|^2 + \|[\mathrm{Id} - \mathbf{Pr}_{\mathcal{S}}](x) - [\mathrm{Id} - \mathbf{Pr}_{\mathcal{S}}](y)\|^2 \leqslant \|x - y\|^2 \quad \forall x, y \in \mathcal{H}. \tag{3.1}$$

Denoting $d(x) = \dfrac{1}{2}\operatorname{dist}(x, \mathcal{S})^2 = \dfrac{1}{2}\|x - \mathbf{Pr}_{\mathcal{S}}x\|^2$ for all $x \in \mathcal{H}$, one has that $x \mapsto d(x)$ is differentiable and it holds $\nabla d(x) = x - \mathbf{Pr}_{\mathcal{S}}x$ for all $x \in \mathcal{H}$.

Lemma 3.7. *Let $(x_n)_{n\geqslant 0}$ be the sequence generated by Algorithm 3.3 and assume that the Hypotheses 3.2 are verified. Then the following inequality holds for all $n \geqslant 1$*

$$\begin{aligned} & d(x_{n+1}) - d(x_n) - \alpha_n (d(x_n) - d(x_{n-1})) + \lambda_n [(f+g)(x_{n+1}) - (f+g)_*] \\ & \leqslant \left(\frac{L_g}{2}\lambda_n + \frac{L_h}{4}\lambda_n\beta_n + \frac{\alpha_n}{2}\right)\|x_{n+1} - x_n\|^2 + \alpha_n \|x_n - x_{n-1}\|^2 . \end{aligned} \tag{3.2}$$

Proof. Let $n \geqslant 1$ be fixed. Since d is convex, we have

$$d(x_{n+1}) - d(x_n) \leqslant \langle x_{n+1} - \mathbf{Pr}_{\mathcal{S}}(x_{n+1}), x_{n+1} - x_n\rangle . \tag{3.3}$$

Then there exists $v_{n+1} \in \partial f(x_{n+1})$ such that (see (2.2))

$$x_n - x_{n+1} - \lambda_n(\nabla g(x_n) + \beta_n\nabla h(x_n)) + \alpha_n(x_n - x_{n-1}) = \lambda_n v_{n+1}$$

and, so,

$$\begin{aligned}
&\left\langle x_{n+1}-\mathbf{Pr}_{\mathcal{S}}\left(x_{n+1}\right), x_{n+1}-x_{n}\right\rangle \\
=&\left\langle x_{n+1}-\mathbf{Pr}_{\mathcal{S}}\left(x_{n+1}\right), -\lambda_n v_{n+1}-\lambda_n \nabla g\left(x_n\right)-\lambda_n\beta_n\nabla h\left(x_n\right)+\alpha_n\left(x_n-x_{n-1}\right)\right\rangle \\
&-\lambda_n\beta_n\left\langle x_{n+1}-\mathbf{Pr}_{\mathcal{S}}\left(x_{n+1}\right), \nabla h\left(x_n\right)\right\rangle+\alpha_n\left\langle x_{n+1}-\mathbf{Pr}_{\mathcal{S}}\left(x_{n+1}\right), x_n-x_{n-1}\right\rangle. \qquad (3.4)
\end{aligned}$$

Since $v_{n+1}\in\partial f\left(x_{n+1}\right)$, we get

$$-\lambda_n\left\langle x_{n+1}-\mathbf{Pr}_{\mathcal{S}}\left(x_{n+1}\right), v_{n+1}\right\rangle \leqslant \lambda_n\left[f\left(\mathbf{Pr}_{\mathcal{S}}\left(x_{n+1}\right)\right)-f\left(x_{n+1}\right)\right]. \qquad (3.5)$$

Using the convexity of g it follows

$$g\left(x_n\right)-g\left(\mathbf{Pr}_{\mathcal{S}}\left(x_{n+1}\right)\right)\leqslant\left\langle \nabla g\left(x_n\right), x_n-\mathbf{Pr}_{\mathcal{S}}\left(x_{n+1}\right)\right\rangle. \qquad (3.6)$$

On the other hand, the Descent Lemma gives

$$g\left(x_{n+1}\right)\leqslant g\left(x_n\right)+\left\langle\nabla g\left(x_n\right), x_{n+1}-x_n\right\rangle+\frac{L_g}{2}\left\|x_{n+1}-x_n\right\|^2. \qquad (3.7)$$

By adding (3.6) and (3.7), it yields

$$-\lambda_n\left\langle x_{n+1}-\mathbf{Pr}_{\mathcal{S}}\left(x_{n+1}\right), \nabla g\left(x_n\right)\right\rangle\leqslant \lambda_n\left[g\left(\mathbf{Pr}_{\mathcal{S}}\left(x_{n+1}\right)\right)-g\left(x_{n+1}\right)\right]+\frac{L_g\lambda_n}{2}\left\|x_{n+1}-x_n\right\|^2. \qquad (3.8)$$

Using the $\frac{1}{L_h}-$cocoercivity of ∇h combined with the fact that $\nabla h\left(\mathbf{Pr}_{\mathcal{S}}\left(x_{n+1}\right)\right)=0$ (as $\mathbf{Pr}_{\mathcal{S}}\left(x_{n+1}\right)$ belongs to $\mathcal{S}$), it yields

$$-\left\langle x_n-\mathbf{Pr}_{\mathcal{S}}\left(x_{n+1}\right), \nabla h\left(x_n\right)\right\rangle\leqslant-\frac{1}{L_h}\left\|\nabla h\left(x_n\right)\right\|^2.$$

Therefore

$$\begin{aligned}
-\lambda_n\beta_n\left\langle x_{n+1}-\mathbf{Pr}_{\mathcal{S}}\left(x_{n+1}\right), \nabla h\left(x_n\right)\right\rangle &\leqslant \lambda_n\beta_n\left(\left\langle x_n-x_{n+1}, \nabla h\left(x_n\right)\right\rangle-\frac{1}{L_h}\left\|\nabla h\left(x_n\right)\right\|^2\right)\\
&\leqslant \lambda_n\beta_n\frac{L_h}{4}\left\|x_{n+1}-x_n\right\|^2. \qquad (3.9)
\end{aligned}$$

Further, we have

$$\begin{aligned}
&\alpha_n\left\langle x_{n+1}-\mathbf{Pr}_{\mathcal{S}}\left(x_{n+1}\right)-\left(x_n-\mathbf{Pr}_{\mathcal{S}}\left(x_n\right)\right), x_n-x_{n-1}\right\rangle\\
\leqslant&\frac{\alpha_n}{2}\left\|\left[\mathrm{Id}-\mathbf{Pr}_{\mathcal{S}}\right]\left(x_{n+1}\right)-\left[\mathrm{Id}-\mathbf{Pr}_{\mathcal{S}}\right]\left(x_n\right)\right\|^2+\frac{\alpha_n}{2}\left\|x_n-x_{n-1}\right\|^2\\
\leqslant&\frac{\alpha_n}{2}\left\|x_{n+1}-x_n\right\|^2+\frac{\alpha_n}{2}\left\|x_n-x_{n-1}\right\|^2,
\end{aligned}$$

and

$$\begin{aligned}
&\alpha_n\left\langle x_n-\mathbf{Pr}_{\mathcal{S}}\left(x_n\right), x_n-x_{n-1}\right\rangle\\
=&\alpha_n d\left(x_n\right)+\frac{\alpha_n}{2}\left\|x_n-x_{n-1}\right\|^2-\frac{\alpha_n}{2}\left\|x_{n-1}-\mathbf{Pr}_{\mathcal{S}}\left(x_n\right)\right\|^2\\
\leqslant&\alpha_n d\left(x_n\right)+\frac{\alpha_n}{2}\left\|x_n-x_{n-1}\right\|^2-\alpha_n d\left(x_{n-1}\right).
\end{aligned}$$

By adding two relations above, we obtain

$$\begin{aligned}
&\alpha_n\left\langle x_{n+1}-\mathbf{Pr}_{\mathcal{S}}\left(x_{n+1}\right), x_n-x_{n-1}\right\rangle\\
=&\alpha_n\left\langle x_{n+1}-\mathbf{Pr}_{\mathcal{S}}\left(x_{n+1}\right)-\left(x_n-\mathbf{Pr}_{\mathcal{S}}\left(x_n\right)\right), x_n-x_{n-1}\right\rangle+\alpha_n\left\langle x_n-\mathbf{Pr}_{\mathcal{S}}\left(x_n\right), x_n-x_{n-1}\right\rangle\\
\leqslant&\frac{\alpha_n}{2}\left\|x_{n+1}-x_n\right\|^2+\alpha_n\left\|x_n-x_{n-1}\right\|^2+\alpha_n\left(d\left(x_n\right)-d\left(x_{n-1}\right)\right). \qquad (3.10)
\end{aligned}$$

By combining (3.5) , (3.8) , (3.9) and (3.10) with (3.4) we obtain the desired conclusion. □

Definition 3.8. A function $\Psi\colon \mathcal{H} \to \bar{\mathbb{R}}$ is sad to be inf-compact if for every $r > 0$ and $\kappa \in \mathbb{R}$ the set

$$\mathrm{Lev}_\kappa^r(\Psi) := \{x \in \mathcal{H}\colon \|x\| \leqslant r, \Psi(x) \leqslant \kappa\}$$

is relatively compact in $\mathcal{H}$.

An useful property of inf-compact functions follows.

Lemma 3.9. *Let $\Psi\colon \mathcal{H} \to \bar{\mathbb{R}}$ be inf-compact and $(x_n)_{n\geqslant 0}$ be a bounded sequence in $\mathcal{H}$ such that $(\Psi(x_n))_{n\geqslant 0}$ is bounded as well. If the sequence $(x_n)_{n\geqslant 0}$ converges weakly to an element in $\widehat{x}$ as $n \to +\infty$, then it converges strongly to this element.*

Proof. Let be $\bar{r} > 0$ and $\bar{\kappa} \in \mathbb{R}$ such that for all $n \geqslant 1$

$$\|x_n\| \leqslant \bar{r} \quad \text{and} \quad \Psi(x_n) \leqslant \bar{\kappa}.$$

Hence, $(x_n)_{n\geqslant 0}$ belongs to the set $\mathrm{Lev}_{\bar{\kappa}}^{\bar{r}}(\Psi)$, which is relatively compact. Then $(x_n)_{n\geqslant 0}$ has at least one strongly convergente subsequence. Since every strongly convergent subsequence $(x_{n_l})_{l\geqslant 0}$ of $(x_n)_{n\geqslant 0}$ has as limit $\widehat{x}$, the desired conclusion follows. $\square$

We can formulate now the weak nonergodic convergence result.

Theorem 3.10. *Let $0 < \alpha < \frac{1}{3}$, $\varepsilon_1, \varepsilon_2, \varepsilon_3 > 0$, the sequences $(\lambda_n)_{n\geqslant 1}$ and $(\beta_n)_{n\geqslant 1}$ satisfy the condition $0 < \liminf_{n\to\infty} \lambda_n\beta_n \leqslant \sup_{n\geqslant 0} \lambda_n\beta_n \leqslant \mu$, $(x_n)_{n\geqslant 0}$ be the sequence generated by Algorithm 3.3, and assume that the Hypotheses 3.2 are verified and that either $f+g$ or h is inf-compact. Then the following statements are true:*

(i) $\lim_{n\to+\infty} d(x_n) = 0$;

(ii) *the sequence $(x_n)_{n\geqslant 0}$ converges weakly to an element in $\mathcal{S}$ as $n \to +\infty$;*

(iii) *if h is inf-compact, then the sequence $(x_n)_{n\geqslant 0}$ converges strongly to an element in $\mathcal{S}$ as $n \to +\infty$.*

Proof. (i) Thanks to Lemma 3.7 , for all $n \geqslant 1$ we have

$$d(x_{n+1}) - d(x_n) + \lambda_n [(f+g)(x_{n+1}) - (f+g)_*] \leqslant \alpha_n (d(x_n) - d(x_{n-1})) + \zeta_n, \tag{3.11}$$

where

$$\zeta_n := \left(\frac{L_g}{2}\lambda_n + \frac{L_h}{4}\lambda_n\beta_n + \frac{\alpha_n}{2}\right)\|x_{n+1} - x_n\|^2 + \alpha_n \|x_n - x_{n-1}\|^2.$$

From Proposition 3.5 (i), combined with the fact that both sequences $(\lambda_n)_{n\geqslant 1}$ and $(\beta_n)_{n\geqslant 1}$ are bounded, it follows that $\sum_{n\geqslant 1} \zeta_n < +\infty$.

In general, since $(x_n)_{n\geqslant 0}$ is not necessarily included in $\arg\min h$, we have to treat two different cases.

Case 1: There exists an integer $n_1 \geqslant 1$ such that $(f+g)(x_n) \geqslant (f+g)_*$ for all $n \geqslant n_1$. In this case, we obtain from Lemma 1.2 that:

- the limit $\lim_{n\to+\infty} d(x_n)$ exists.
- $\sum_{n\geqslant n_2} \lambda_n [(f+g)(x_{n+1}) - (f+g)_*] < +\infty$. Moreover, since $(\lambda_n)_{n\geqslant 1} \notin \ell^1$, we must have

$$\liminf_{n\to+\infty} (f+g)(x_n) \leqslant (f+g)_*. \tag{3.12}$$

Consider a subsequence $(x_{n_k})_{k\geqslant 0}$ of $(x_n)_{n\geqslant 0}$ such that

$$\lim_{k\to+\infty}(f+g)(x_{n_k}) = \liminf_{n\to+\infty}(f+g)(x_n)$$

and note that, thanks to (3.12), the sequence $((f+g)(x_{n_k}))_{k\geqslant 0}$ is bounded. From Proposition 3.5 (ii)-(iii) we get that also $(x_{n_k})_{k\geqslant 0}$ and $(h(x_{n_k}))_{k\geqslant 0}$ are bounded. Thus, since either $f+g$ or h is inf-compact, there exists a subsequence $(x_{n_l})_{l\geqslant 0}$ of $(x_{n_k})_{k\geqslant 0}$, which converges strongly to an element $\widehat{x}$ as $l\to+\infty$. According to Proposition 3.5 (ii)-(iii), $\widehat{x}$ belongs to $\arg\min h$. On the other hand,

$$\lim_{l\to+\infty}(f+g)(x_{n_l}) = \liminf_{n\to+\infty}(f+g)(x_n) \geqslant (f+g)(\widehat{x}) \geqslant (f+g)_*. \tag{3.13}$$

We deduce from (3.12) - (3.13) that $(f+g)(\widehat{x}) = (f+g)_*$, or in other words, that $\widehat{x}\in\mathcal{S}$. In conclusion, thanks to the continuity of d,

$$\lim_{n\to+\infty} d(x_n) = \lim_{l\to\infty} d(x_{n_l}) = d(\widehat{x}) = 0.$$

Case 2: For all $n\geqslant 1$ there exists some $n'>n$ such that $(f+g)(x_{n'}) < (f+g)_*$. We define the set

$$V = \left\{n'\geqslant 1\colon (f+g)(x_{n'}) < (f+g)_*\right\}.$$

There exist an integer $n_2\geqslant 2$ such that for all $n\geqslant n_2$ the set $\{k\leqslant n\colon k\in V\}$ is nonempty. Hence, for all $n\geqslant n_2$ the number

$$t_n := \max\{k\leqslant n\colon k\in V\}$$

is well-defined. By definition $t_n\leqslant n$ for all $n\geqslant n_3$ and moreover the sequence $\{t_n\}_{n\geqslant n_2}$ is nondecreasing and $\lim\limits_{n\to+\infty} t_n = \infty$. Indeed, if $\lim\limits_{n\to\infty} t_n = t\in\mathbb{R}$, then for all $n'>t$ it holds $(f+g)(x_{n'})\geqslant (f+g)_*$, contradiction. Choose an integer $N\geqslant n_2$.

- If $t_N < N$, then, for all $n = t_N,\cdots,N-1$, since $(f+g)(x_n)\geqslant (f+g)_*$, the inequality (3.11) gives

$$\begin{aligned} d(x_{n+1}) - d(x_n) &\leqslant d(x_{n+1}) - d(x_n) + \lambda_n\left[F(x_{n+1}) - F_*\right] \\ &\leqslant \alpha_n\left(d(x_n) - d(x_{n-1})\right) + \zeta_n. \end{aligned} \tag{3.14}$$

  Summing (3.14) for $n = t_N,\cdots,N-1$ and using tht $\{\alpha_n\}_{n\geqslant 1}$ is nondecreasing, it yields

$$\begin{aligned} d(x_N) - d(x_{t_N}) &\leqslant \sum_{n=t_N}^{N-1}\left(\alpha_n d(x_n) - \alpha_{n-1} d(x_{n-1})\right) + \sum_{n=t_N}^{N-1}\zeta_n \\ &\leqslant \alpha d(x_{N-1}) + \sum_{n\geqslant t_N}\zeta_n. \end{aligned} \tag{3.15}$$

- If $t_N = N$, then $d(x_N) = d(x_{t_N})$ and we have

$$d(x_N) - \alpha d(x_{N-1}) \leqslant d(x_{t_N}) + \sum_{n\geqslant t_N}\zeta_n. \tag{3.16}$$

For all $n\geqslant 1$ we define $a_n := d(x_n) - \alpha d(x_{n-1})$. In both cases it yields

$$a_N \leqslant d(x_{t_N}) + \sum_{n=t_N}^{N}\zeta_n \leqslant d(x_{t_N}) + \sum_{n\geqslant t_N}\zeta_n. \tag{3.17}$$

Passing in (3.17) to limit as $N \to +\infty$ we obtain that

$$\limsup_{n\to+\infty} a_n \leqslant \limsup_{n\to+\infty} d\left(x_{t_n}\right). \tag{3.18}$$

Let be $u \in \mathcal{S}$. For all $n \geqslant 1$ we have

$$d\left(x_n\right) = \frac{1}{2}\operatorname{dist}\left(x_n, \mathcal{S}\right)^2 \leqslant \frac{1}{2}\left\|x_n - u\right\|^2,$$

which shows that $(d\left(x_n\right))_{n\geqslant 0}$ is bounded, as $\lim_{n\to+\infty} \|x_n - u\|$ exists. We obtain

$$\limsup_{n\to\infty} a_n = \limsup_{n\to\infty}\left[d\left(x_n\right) - \alpha d\left(x_{n-1}\right)\right] \geqslant (1-\alpha)\limsup_{n\to\infty} d\left(x_n\right) \geqslant 0. \tag{3.19}$$

Further, for all $n \geqslant 1$ we have $(f+g)\left(x_{t_n}\right) < (f+g)_*$, which gives

$$\limsup_{n\to+\infty}(f+g)\left(x_{t_n}\right) \leqslant (f+g)_*. \tag{3.20}$$

This means that the sequence $\left((f+g)\left(x_{t_n}\right)\right)_{n\geqslant 0}$ is bounded from above. Consider a subsequence $\left(x_{t_k}\right)_{k\geqslant 0}$ of $\left(x_{t_n}\right)_{n\geqslant 0}$ such that

$$\lim_{k\to+\infty} d\left(x_{t_k}\right) = \limsup_{n\to+\infty} d\left(x_{t_n}\right).$$

From Proposition 3.5 (ii)-(iii) we get that also $\left(x_{t_k}\right)_{k\geqslant 0}$ and $\left(h(x_{t_k})\right)_{k\geqslant 0}$ are bounded. Thus, since either $f+g$ or h is inf-compact, there exists a subsequence $\left(x_{t_l}\right)_{l\geqslant 0}$ of $\left(x_{t_k}\right)_{k\geqslant 0}$, which converges strongly to an element $\widehat{x}$ as $l \to +\infty$. According to Proposition 3.5 (ii)-(iii), $\widehat{x}$ belongs to $\arg\min h$. Furthermore, it holds

$$\liminf_{l\to+\infty}(f+g)\left(x_{t_l}\right) \geqslant (f+g)\left(\widehat{x}\right) \geqslant (f+g)_*. \tag{3.21}$$

We deduce from (3.20) and (3.21) that

$$(f+g)_* \leqslant (f+g)\left(\widehat{x}\right) \leqslant \limsup_{n\to+\infty}(f+g)\left(x_{t_l}\right) \leqslant \limsup_{n\to+\infty}(f+g)\left(x_{t_n}\right) \leqslant (f+g)_*,$$

which gives $\widehat{x} \in \mathcal{S}$. Thanks to the continuity of d we get

$$\limsup_{n\to+\infty} d\left(x_{t_n}\right) = \lim_{l\to+\infty} d\left(x_{t_l}\right) = d\left(\widehat{x}\right) = 0. \tag{3.22}$$

By combining (3.18), (3.19) and (3.22), it yields

$$0 \leqslant (1-\alpha)\limsup_{n\to+\infty} d\left(x_n\right) \leqslant \limsup_{n\to+\infty} a_n \leqslant \limsup_{n\to+\infty} d\left(x_{t_n}\right) = 0,$$

which implies $\limsup_{n\to+\infty} d\left(x_n\right) = 0$ and thus

$$\lim_{n\to+\infty} d\left(x_n\right) = \liminf_{n\to+\infty} d\left(x_n\right) = \limsup_{n\to+\infty} d\left(x_n\right) = 0.$$

(ii) According to (i) we have $\lim_{n\to\infty} d\left(x_n\right) = 0$, thus every weak cluster point of the sequence $\left(x_n\right)_{n\geqslant 0}$ belongs to $\mathcal{S}$. From Lemma 1.1 it follows that $\left(x_n\right)_{n\geqslant 0}$ converges weakly to a point in $\mathcal{S}$ as $n \to +\infty$.

(iii) Since $\liminf_{n\to\infty} \lambda_n \beta_n > 0$, from Proposition 3.5(ii) we have that

$$\lim_{n\to+\infty} \left\|\nabla h\left(x_n\right)\right\| = \lim_{n\to+\infty} h\left(x_n\right) = 0.$$

Since $\left(x_n\right)_{n\geqslant 0}$ is bounded, there exist $\overline{r} > 0$ and $\overline{\kappa} \in \mathbb{R}$ such that for all $n \geqslant 1$

$$\left\|x_n\right\| \leqslant \overline{r} \quad \text{ and } \quad h\left(x_n\right) \leqslant \overline{\kappa}.$$

Thanks to (ii) the sequence $\left(x_n\right)_{n\geqslant 0}$ converges weakly to an element in $\mathcal{S}$. Therefore, according to Lemma 3.9, it converges strongly to this element in $\mathcal{S}$. □

References

[1] **F. Alvarez**. *On the minimizing property of a second order dissipative system in Hilbert spaces.* SIAM Journal on Control and Optimization, Vol. 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, Vol. 14(3): 773–782 (2004)

[3] **F. Alvarez and H. Attouch**. *An inertial proximal method for maximal monotone operators via discretization of a nonlinear oscillator with damping.* Set-Valued Analysis, Vol. 9(1): 3–11 (2001)

[4] **H. Attouch, J. Bolte, P. Redont and A. Soubeyran**. *Alternating proximal algorithms for weakly coupled convex minimization problems, applications to dynamical games and PDE's.* Journal of Convex Analysis, Vol. 15(3): 485–506 (2008)

[5] **H. Attouch, L. M. Briceno-Arias and P. L. Combettes**. *A parallel splitting method for coupled monotone inclusions.* SIAM Journal on Control and Optimization, Vol. 48(5): 3246–3270 (2010)

[6] **H. Attouch, A. Cabot and M.-O. Czarnecki** *Asymptotic behavior of nonautonomous monotone and subgradient evolution equations.* Transactions of the American Mathematical Society, Vol. 370(2): 755–790 (2018)

[7] **H. Attouch and M.-O. Czarnecki**. *Asymptotic behavior of coupled dynamical systems with multiscale aspects.* Journal of Differential Equations, Vol. 248(6): 1315–1344 (2010)

[8] **H. Attouch and M.-O. Czarnecki**. *Asymptotic behavior of gradient-like dynamical systems involving inertia and multiscale aspects.* Journal of Differential Equations, Vol. 262(3): 2745–2770 (2017)

[9] **H. Attouch, M.-O. Czarnecki and J. Peypouquet**. *Coupling forward-backward with penalty schemes and parallel splitting for constrained variational inequalities.* SIAM Journal on Optimization, Vol. 21(4): 1251–1274 (2011)

[10] **H. Attouch, M.-O. Czarnecki and J. Peypouquet**. *Prox-penalization and splitting methods for constrained variational problems.* SIAM Journal on Optimization, Vol. 21(1): 149–173 (2011)

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

[12] **S. Banert and R. I. Boţ**. *Backward penalty schemes for monotone inclusion problems.* Journal of Optimization Theory and Applications, Vol. 166(3): 930–948 (2015)

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

[14] **H. H. Bauschke, D.A. McLaren and H.S. Sendov**. *Fitzpatrick functions: inequalities, examples and remarks on a problem by S. Fitzpatrick.* Journal of Convex Analysis, Vol. 13(3): 499–523 (2006)

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

[16] **D. P. Bertsekas**. *Nonlinear programming.* Athena Scientific, Cambridge (MA) (1999)

[17] **R. I. Boţ**. *Conjugate Duality in Convex Optimization.* Lecture Notes in Economics and Mathematical Systems, Vol. 637, Springer, Berlin Heidelberg (2010)

[18] **R. I. Boţ and E. R. Csetnek**. *A Tseng's type penalty scheme for solving inclusion problems involving linearly composed and parallel-sum type monotone operators.* Vietnam Journal of Mathematics, Vol. 42(4): 451–465 (2014)

[19] **R. I. Boţ and E. R. Csetnek**. *Forward-backward and Tseng's type penalty schemes for monotone inclusion problems.* Set-Valued Variational Analysis, Vol. 22(2): 313–331 (2014)

[20] **R. I. Boţ and E. R. Csetnek**. *An inertial forward-backward-forward primal-dual splitting algorithm for solving monotone inclusion problems.* Numerical Algorithms, Vol. 71(3): 519–540 (2016)

[21] **R. I. Boţ and E. R. Csetnek**. *Penalty schemes with inertial effects for monotone inclusion problems.* Optimization, Vol. 66(6): 313–331 (2017)

[22] **R. I. Boţ, E. R. Csetnek and C. Hendrich**. *Inertial Douglas-Rachford splitting for monotone inclusion problems.* Applied Mathematics and Computation, Vol. 256(1): 472–487 (2015)

[23] **R. I. Boţ, E. R. Csetnek and S. C. Laszlo**. *An inertial forward-backward algorithm for the minimization of the sum of two nonconvex functions.* EURO Journal on Computational Optimization, Vol.4(1): 3–25 (2016).

[24] **R. I. Boţ, E. R. Csetnek and N. Nimana**. *Gradient-type penalty method with inertial effects for solving constrained convex optimization problems with smooth data.* Optimization Letters, Vol.12(1): 17–33 (2018).

[25] **R. I. Boţ, E. R. Csetnek and N. Nimana**. *An inertial proximal-gradient penalization scheme for constrained convex optimization problems.* Vietnam Journal of Mathematics, Vol.46(1): 53–71 (2018).

[26] **J. M. Borwein**. *Maximal monotonicity via convex analysis.* Journal of Convex Analysis, Vol. 13(3): 561–586 (2006)

[27] **R. S. Burachik and B. F. Svaiter**. *Maximal monotone operators, convex functions and a special family of enlargements.* Set-Valued Analysis, Vol. 10(4): 29–316 (2002)

[28] **A. Chambolle and Ch. Dossal**. *On the convergence of the iterates of the "Fast Iterative Shrinkage/Thresholding Algorithm".* Journal of Optimization Theory and Applications, Vol. 166(3): 968–982 (2015)

[29] **C. Chen, R. H. Chan, S. Ma and J. Yang**. *Inertial proximal ADMM for linearly constrained separable convex optimization.* SIAM Journal on Imaging Science, Vol. 8(4): 2239–2267 (2015)

[30] **C. Chen, S. Ma and J. Yang**. *A general inertial proximal point algorithm for mixed variational inequality problem.* SIAM Journal on Optimization, Vol. 25(4): 2120–2142 (2015)

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

[32] **S. Fitzpatrick**. *Representing monotone operators by convex functions.* In *Proceedings of the Centre for Mathematical Analysis.* Workshop/Miniconference on Functional Analysis and Optimization, Vol.20, Australian National University, Canberra (1988)

[33] **P. Frankel and J. Peypouquet**. *Lagrangian-Penalization Algorithm for Constrained Optimization and Variational Inequalities.* Set-Valued and Variational Analysis, Vol. 20(2): 169–185 (2012)

[34] **P.-E. Maingé**. *Convergence theorems for inertial Krasnosel'skiĭ-Mann type algorithms.* Journal of Computational and Applied Mathematics, Vol. 219(1): 223–236 (2008)

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

[36] **A. Moudafi and M. Oliny**. *Convergence of a splitting inertial proximal method for monotone operators.* Journal of Computational and Applied Mathematics, Vol. 155(2): 447–454 (2003)

[37] **N. Noun and J. Peypouquet**. *Forward-backward penalty scheme for constrained convex minimization without inf-compactness.* Journal of Optimization Theory and Applications, Vol. 158(3): 787–795 (2013)

[38] **J. Peypouquet**. *Coupling the Gradient Method with a General Exterior Penalization Scheme for Convex Minimization.* Journal of Optimization Theory and Applications, Vol. 153(1): 123–138 (2012)

[39] **B. T. Polyak**. *Introduction to Optimization.* Translations Series in Mathematics and Engineering. Publications Division, Optimization Software Inc., New York (1987)

[40] **C. Zălinescu**. *Convex Analysis in General Vector Spaces.* World Scientific, Singapore (2002)