Convergence rate of inexact proximal point methods with relative error criteria for convex optimization

Renato D. C. Monteiro[*] B. F. Svaiter[†]

August 22, 2010 (Revised: December 21, 2011)

Abstract

In this paper, we consider a framework of inexact proximal point methods for convex optimization that allows a relative error tolerance in the approximate solution of each proximal subproblem and establish its convergence rate. We then show that the well-known forward-backward splitting algorithm for convex optimization belongs to this framework. Finally, we propose and establish the iteration-complexity of an inexact forward-backward splitting algorithm for solving optimization problems whose objective functions are obtained by maximizing convex-concave saddle functions.

1 Introduction

In this paper, we consider a framework of inexact proximal point (IPP) methods for convex optimization (CO) which allows a relative error tolerance in the approximate solution of each proximal subproblem. This framework, which we refer to as the IPP-CO framework, is a subset of the hybrid proximal extragradient (HPE) method introduced by Solodov and Svaiter in [19] (see also [20, 21, 22]) for solving monotone inclusion problems. Global convergence rate results for the HPE method have been derived in [11] (see also [10]), and hence apply to the IPP-CO framework. However, by exploiting the special structure of convex optimization, convergence rate results stronger than those obtained for the HPE method are derived for the IPP-CO framework.

We show, as illustration, that the well-known forward-backward splitting method for convex optimization (see for example [5]) belongs to the IPP-CO framework and, as a consequence, we derive iteration-complexity bounds similar to, but under more general assumptions than, those of Theorem 4 of [13]. More specifically, [13] assumes that the sublevel subsets of the objective function are bounded and express the complexity bounds in terms of the diameter of the sublevel set corresponding to the initial iterate. On the other hand, our results do not assume boundedness of the sublevel sets and express the bounds in items of the distance of the initial iterate to the optimal solution set.

We also consider convex optimization problems whose objective functions are obtained by maximizing convex-concave saddle functions and propose an inexact forward-backward splitting algorithm

[*] School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, 30332-0205. (email: monteiro@isye.gatech.edu). The work of this author was partially supported by NSF Grants CCF-0808863 and CMMI-0900094 and ONR Grant ONR N00014-08-1-0033.

[†] IMPA, Estrada Dona Castorina 110, 22460-320 Rio de Janeiro, Brazil (email: benar@impa.br). The work of this author was partially supported by CNPq grants 480101/2008-6, 303583/2008-8, FAPERJ grant E-26/102.821/2008 and PRONEX-Optimization

for solving them. The inexactness of the proposed method originates from the assumption that the objective function and its gradient are approximately evaluated in the sense that the corresponding saddle function maximization subproblem is solved inexactly. Iteration-complexity bounds are obtained for the inexact forward-backward splitting algorithm by showing that it also belongs to the IPP-CO framework.

This paper is organized as follows. Subsection 1.1 describes the notation and basic concepts about convex analysis used in our presentation. Section 2 describes the IPP-CO framework and derives general convergence rate results for it. Section 3 obtains iteration-complexity results for the forward-backward splitting method by showing that it belongs to the IPP-CO framework. Section 4 proposes and establishes the iteration-complexity of an inexact forward-backward splitting method for saddle-based convex optimization problems. Finally, Section 5 gives some concluding remarks.

1.1 Notation

Throughout this paper, $\mathcal{X}$ denotes a finite dimensional inner product real vector space with inner product and induced norm denoted by $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$, respectively. We let $\mathbb{N}$ denote the set of all positive integers and $\mathbb{R}$ denote the set of real numbers. We let $\mathbb{R}_+$ and $\mathbb{R}_{++}$ denote the set of non-negative and positive real numbers, respectively. For a nonempty closed convex set $\Omega \subseteq \mathcal{X}$, we denote the projection operator onto Ω (with respect to $\langle \cdot, \cdot \rangle$) by P_Ω. The identity operator from $\mathcal{X}$ onto $\mathcal{X}$ is denoted by I. The domain of definition of a point-to-point function F is denoted by $\operatorname{Dom} F$.

For a scalar $\varepsilon \geq 0$, the ε-subdifferential of a function $f : \mathcal{X} \to \bar{\mathbb{R}}$ is the point-to-set operator $\partial_\varepsilon f : \mathcal{X} \rightrightarrows \mathcal{X}$ defined as

$$\partial_\varepsilon f(x) = \{v \mid f(\tilde{x}) \geq f(x) + \langle \tilde{x} - x, v \rangle - \varepsilon, \ \forall \tilde{x} \in \mathcal{X}\}, \quad \forall x \in \mathcal{X}. \tag{1}$$

When $\varepsilon = 0$, the operator $\partial_\varepsilon f$ is simply denoted by ∂f and is referred to as the subdifferential of f. The operator ∂f is trivially monotone if f is proper. If f is a proper lower semi-continuous convex function, then ∂f is maximal monotone [17].

The *indicator function* of a closed convex set $X \subseteq \mathcal{X}$ is the function $\delta_X : \mathcal{X} \to \bar{\mathbb{R}}$ defined as

$$\delta_X(x) = \begin{cases} 0, & x \in X; \\ \infty, & \text{otherwise.} \end{cases}$$

2 A framework of inexact proximal point methods

In this section, we describe the IPP-CO framework of IPP methods for convex optimization. We mention that the IPP-CO framework does not specify how its steps are implemented, and hence the overall cost of an iteration of a specific instance of the IPP-CO framework. However, global convergence rate results, and hence potential complexity bounds on the number of iterations, are derived for the IPP-CO framework. Sections 3 and 4 describe two specific instances of the IPP-CO framework for which the actual implementation of the steps are illustrated.

The optimization problem we consider in this section is

$$f^* = \inf\{f(x) : x \in \mathcal{X}\}, \tag{2}$$

where:

O.1) $f : \mathcal{X} \to \bar{\mathbb{R}}$ is a proper closed convex function;

O.2) the set of optimal solutions $X^* := \{x : f(x) = f^*\}$ is nonempty.

We now state the IPP-CO framework for solving (2).

IPP-CO Framework:

0) Let $x_0 \in \mathcal{X}$ and $0 \leq \sigma < 1$ be given and set $k = 1$;

1) choose $\lambda_k > 0$ and find $x_k \in \mathcal{X}$, $\sigma_k \in [0, \sigma]$ and $\varepsilon_k \geq 0$ such that

$$v_k := \frac{1}{\lambda_k}(x_{k-1} - x_k) \in \partial_{\varepsilon_k} f(x_k), \qquad 2\lambda_k \varepsilon_k \leq \sigma_k \|x_k - x_{k-1}\|^2 \ ; \tag{3}$$

2) set $k \leftarrow k + 1$ and go to step 1.

end

We now make a few observations regarding the IPP-CO framework. First, the inclusion in step 1) can also be described as

$$x_k \in (I + \lambda_k \partial_{\varepsilon_k} f)^{-1}(x_{k-1}).$$

Second, if $\varepsilon_k = 0$ for every $k \in \mathbb{N}$, the IPP-CO framework reduces to the *exact* proximal point method [9, 18], which is known to be an important tool for the design and analysis of algorithms. Hence, the IPP-CO framework may be regarded as an IPP method. Third, as opposed to the inexactness allowed in the classical IPP method of [18], namely:

$$r_k + x_k \in (I + \lambda_k \partial f)^{-1}(x_{k-1}), \qquad \sum_{k=1}^{\infty} \|r_k\| < \infty,$$

the IPP-CO framework allows an error in the subgradient inclusion by using the ε-subdifferential and a relative error criterion, i.e., the second inequality in (3), instead of the summable error tolerance criterion as above. Fourth, the advantage of allowing the error $\varepsilon_k \geq 0$ and the relative error criterion in the IPP-CO framework is that it contains important instances of convex optimization algorithms (see Sections 3 and 4), including those for which only approximate gradients of the objective function can be computed (see for example Section 4 and reference [4]). Fifth, the HPE algorithm for solving (2) (see [19]) requires, instead of step 1) of the IPP-CO framework, that we find $\lambda_k > 0$, $\sigma_k \in [0, \sigma]$, $\tilde{x}_k \in \mathcal{X}$, $v_k \in \mathcal{X}$ and $\varepsilon_k \geq 0$ such that

$$v_k \in (\partial f)^{\varepsilon_k}(\tilde{x}_k), \quad \|\lambda_k v_k + \tilde{x}_k - x_{k-1}\|^2 + 2\lambda_k \varepsilon_k \leq \sigma_k \|\tilde{x}_k - x_{k-1}\|^2,$$

and then set

$$x_k = x_{k-1} - \lambda_k v_k,$$

where $(\partial f)^{\varepsilon_k}$ is the ε-enlargement (see for example [2]) of the maximal monotone operator ∂f. Since $\partial_\varepsilon f(x) \subseteq (\partial f)^\varepsilon(x)$ for every $x \in \mathcal{X}$ (see for example Lemma 3.4 and Theorem 3.5 in [23]), we easily see that the IPP-CO framework is a special case of the HPE method by letting $\tilde{x}_k = x_k$.

In what follows, we will derive a convergence rate result (Theorem 2.5) for the IPP-CO framework. First, we need to establish a few technical lemmas.

Lemma 2.1. *For every $k \in \mathbb{N}$,*

$$f(x) \geq f(x_k) + \langle x - x_k, v_k \rangle - \frac{\sigma}{2}\lambda_k \|v_k\|^2, \quad \forall x \in \mathcal{X}. \tag{4}$$

Proof. Using the fact that $v_k \in \partial_{\varepsilon_k} f(x_k)$ by (3), and definition (1), we have

$$f(x) \geq f(x_k) + \langle x - x_k, v_k \rangle - \varepsilon_k \geq f(x_k) + \langle x - x_k, v_k \rangle - \frac{\sigma}{2\lambda_k}\|x_k - x_{k-1}\|^2, \qquad \forall x \in \mathcal{X},$$

where the last inequality is due to the assumption that $\sigma_k \in [0, \sigma]$ and the inequality in (3). Now, (4) follows from the above inequality and the definition of v_k in (3). □

Lemma 2.2. *For every $k \in \mathbb{N}$ and $x^* \in X^*$, we have:*

$$f(x_{k-1}) \geq f(x_k) + \left(1 - \frac{\sigma}{2}\right) \lambda_k \|v_k\|^2, \tag{5}$$

and

$$\frac{1}{2}\|x^* - x_{k-1}\|^2 \geq \frac{1}{2}\|x^* - x_k\|^2 + \frac{1-\sigma}{2}\lambda_k^2 \|v_k\|^2 + \lambda_k (f(x_k) - f(x^*)). \tag{6}$$

Proof. In view of the definition of v_k in (3), inequality (5) follows immediately from (4) with $x = x_{k-1}$. Now, using again the definition of v_k, we have

$$\begin{aligned} \frac{1}{2}\|x^* - x_{k-1}\|^2 &= \frac{1}{2}\|x^* - x_k\|^2 + \frac{1}{2}\|x_k - x_{k-1}\|^2 + \langle x^* - x_k, x_k - x_{k-1} \rangle \\ &= \frac{1}{2}\|x^* - x_k\|^2 + \frac{1}{2}\lambda_k^2 \|v_k\|^2 + \lambda_k \langle x_k - x^*, v_k \rangle. \end{aligned}$$

Inequality (6) now follows by combining the latter inequality and (4) with $x = x^*$. □

Lemma 2.3. *Define*

$$\Lambda_0 = 0, \qquad \Lambda_k = \sum_{i=1}^{k} \lambda_i, \quad \forall k \in \mathbb{N}. \tag{7}$$

Then, for any $k \in \mathbb{N}$ and $x^* \in X^*$,

$$\begin{aligned} &\frac{1}{2}\|x^* - x_{k-1}\|^2 + \Lambda_{k-1}(f(x_{k-1}) - f^*) \\ &\qquad \geq \frac{1}{2}\|x^* - x_k\|^2 + \Lambda_k (f(x_k) - f^*) + \frac{1-\sigma}{2}\Lambda_k \lambda_k \|v_k\|^2. \end{aligned} \tag{8}$$

Proof. By (5), we have

$$\Lambda_{k-1}(f(x_{k-1}) - f^*) \geq \Lambda_{k-1}\left(f(x_k) - f^* + \left(1 - \frac{\sigma}{2}\right)\lambda_k \|v_k\|^2\right).$$

The result now follows by adding this inequality to (6), using the fact that $\Lambda_k = \Lambda_{k-1} + \lambda_k$ and $1 - \sigma/2 \geq (1-\sigma)/2$. □

The following result follows by adding inequality (8) from 1 to k.

Lemma 2.4. *For any $k \in \mathbb{N}$ and $x^* \in X^*$,*

$$\frac{1}{2}\|x^* - x_0\|^2 \geq \frac{1}{2}\|x^* - x_k\|^2 + \Lambda_k(f(x_k) - f^*) + \frac{1-\sigma}{2}\sum_{j=1}^{k} \Lambda_j \lambda_j \|v_j\|^2.$$

We are now ready to state the convergence rate result for the IPP-CO framework. Throughout this paper, we denote the distance of x_0 to X^* by d_0.

Theorem 2.5. *For every $k \in \mathbb{N}$, the following statements hold:*

a) $f(x_k) - f^* \leq d_0^2/(2\Lambda_k)$;

b) $v_k \in \partial_{\varepsilon_k} f(x_k)$ and there exists $i \leq k$ such that

$$\|v_i\| \leq \frac{d_0}{(1-\sigma)^{1/2}\Theta_k^{1/2}}, \qquad \varepsilon_i \leq \frac{\sigma d_0^2 \lambda_i}{2(1-\sigma)\Theta_k}, \tag{9}$$

where Λ_k is defined in (7) and $\Theta_k := \sum_{j=1}^{k} \lambda_j \Lambda_j$.

Proof. Fix $k \in \mathbb{N}$. Let $x^* \in X^*$ be such that $\|x^* = x_0\| = d_0$. By Lemma 2.4 with such $x^* \in X^*$, we have

$$\frac{d_0^2}{2} \geq \Lambda_k(f(x_k) - f^*) + \frac{1-\sigma}{2}\sum_{j=1}^{k} \Lambda_j \lambda_j \|v_j\|^2,$$

which immediately implies a). To prove item b), let i be such that

$$i \in \operatorname{Argmin}\{\|v_j\| \mid j = 1, \ldots, k\}.$$

Using the previous inequality, the above definition and the definition of Θ_k, we conclude that

$$\frac{d_0^2}{2} \geq \frac{1-\sigma}{2}\left(\sum_{j=1}^{k} \Lambda_j \lambda_j\right)\|v_i\|^2 = \frac{1-\sigma}{2}\Theta_k \|v_i\|^2,$$

which clearly implies the first inequality in b). Moreover, by (3) and the assumption that $\sigma_i \in [0, \sigma]$, we have $2\lambda_i \varepsilon_i \leq \sigma\|\lambda_i v_i\|^2$. Hence, $\varepsilon_i \leq \lambda_i \sigma \|v_i\|^2/2$ and the second inequality in b) follows from the first one in b). Since the inclusion in b) follows immediately from (3), the result follows. □

In the remaining part of this section, we focus our attention on those instances of the IPP-CO framework in which the sequence of stepsizes $\{\lambda_k\}$ is constant. The first result below is a specialization of Theorem 2.5 to these instances.

Corollary 2.6. *Consider an instance of the IPP-CO framework with $\lambda_k = \lambda > 0$ for every $k \in \mathbb{N}$. Then, for every $k \in \mathbb{N}$, the following statements hold:*

a) $f(x_k) - f^* \leq d_0^2/(2k\lambda)$;

b) $v_k \in \partial_{\varepsilon_k} f(x_k)$ and there exists $i \leq k$ such that

$$\|v_i\| \leq \frac{\sqrt{2}d_0}{(1-\sigma)^{1/2}\lambda k}, \qquad \varepsilon_i \leq \frac{\sigma d_0^2}{(1-\sigma)\lambda k^2}.$$

Proof. This result follows immediately from Theorem 2.5 and the fact that $\Lambda_k = k\lambda$ and $\Theta_k = \lambda^2 k(k+1)/2$, which are due to the assumption that $\lambda_k = \lambda$ for every $k \in \mathbb{N}$. □

We observe that the bounds on $\|v_i\|$ and ε_i implied by the analysis of the HPE method in [11] are $\mathcal{O}(1/\sqrt{k})$ and $\mathcal{O}(1/k)$, respectively. Hence, the bounds obtained in Corollary 2.6 for those instances of the IPP-CO framework with constant stepsizes improve the ones implied by the analysis of [11].

Consider the natural goal of obtaining the following notion of approximate solution.

Definition 2.7. *For a given tolerance pair $(\bar\rho, \bar\varepsilon) \in \mathbb{R}^2_{++}$, $\bar x \in \mathcal{X}$ is called a $(\bar\rho, \bar\varepsilon)$-solution of* (2) *if there exists a pair $(v, \varepsilon) \in \mathcal{X} \times \mathbb{R}_+$ such that*

$$v \in \partial_\varepsilon f(\bar x), \quad \|v\| \le \bar\rho, \quad \varepsilon \le \bar\varepsilon,$$

in which case (v, ε) is said to be a $(\bar\rho, \bar\varepsilon)$-residual for $\bar x$.

Observe that a stopping condition based on the above notion of approximate solution has the nice feature that it can be used for instances of (2) in which the effective domain of f is unbounded.

The following iteration-complexity result follows as an immediate consequence of Corollary 2.6(b).

Corollary 2.8. *For a given tolerance pair $(\bar\rho, \bar\varepsilon) \in \mathbb{R}^2_{++}$, an instance of the IPP-CO framework with $\lambda_k = \lambda > 0$ for every $k \in \mathbb{N}$, finds a $(\bar\rho, \bar\varepsilon)$-solution of* (2)*, together with a corresponding $(\bar\rho, \bar\varepsilon)$-residual, in at most*

$$\mathcal{O}\left(\max\left\{\left\lceil \frac{d_0}{\lambda\bar\rho}\right\rceil, \left\lceil \frac{d_0}{\sqrt{\lambda\bar\varepsilon}}\right\rceil\right\}\right)$$

iterations.

We now discuss the complexity of computing and detecting an $\bar\varepsilon$-solution of (2), i.e., a solution $\bar x$ such that $f(\bar x) - f^* \le \bar\varepsilon$. Note that $\bar x$ is $\bar\varepsilon$-solution of (2) if, and only if, $\bar x$ is a $(0, \bar\varepsilon)$-solution of (2) in the sense of Definition 2.7. Clearly, verification that an iterate is an $\bar\varepsilon$-solution directly from its definition is only possible for those instances of (2) in which f^* is known.

Consider now those instances of (2) for which f^* is not known. Corollary 2.6(a) provides one trivial way of detecting an $\bar\varepsilon$-solution based on the stopping criterion $D_0^2/(2k\lambda) \le \bar\varepsilon$, where D_0 is a known upper bound on d_0. For example, when $\operatorname{dom} f$ is bounded and $x_0 \in \operatorname{dom} f$, then D_0 can be chosen to be a known upper bound on the diameter of $\operatorname{dom} f$. Another possibility for detecting an $\bar\varepsilon$-solution is to use the stopping criterion

$$\max\left\{\langle v_k, x_k - x\rangle + \varepsilon_k : x \in C\right\} \le \bar\varepsilon, \tag{10}$$

where C is a "simple" compact convex set containing the effective domain of f. Note that the validity of (10) implies that x_k is an $\bar\varepsilon$-solution. Indeed, assuming (10) and using the fact that $v_k \in \partial_{\varepsilon_k} f(x_k)$, we conclude that

$$f(x_k) - f^* = f(x_k) - f(x^*) \le \langle x_k - x^*, v_k\rangle + \varepsilon_k \le \bar\varepsilon,$$

where x^* is an arbitrary optimal solution of (2). The following result describes the iteration-complexity for finding an iterate satisfying (10), which is then a provably $\bar\varepsilon$-solution of (2).

Corollary 2.9. *Consider an instance of the IPP-CO framework with $\lambda_k = \lambda > 0$ for every $k \in \mathbb{N}$, applied to an instance of* (2) *in which* $\operatorname{dom} f$ *is bounded. Assume that a compact convex set C containing* $\operatorname{dom} f$ *is given and let D_C denote the diameter of C. Then, there exists an index*

$$i = \mathcal{O}\left(\left\lceil \frac{d_0 D_C}{\lambda \bar{\varepsilon}} \right\rceil\right)$$

such that the iterate x_i satisfies (10). As a consequence, for any $k \geq i$, x_k is an $\bar{\varepsilon}$-solution of (2).

Proof. Let $\bar{k}$ be the smallest $k \in \mathbb{N}$ satisfying

$$\frac{\sqrt{2} D_C d_0}{(1-\sigma)^{1/2} \lambda k} + \frac{\sigma d_0^2}{(1-\sigma)\lambda k^2} \leq \bar{\varepsilon}.$$

In view of the definition of $\bar{k}$, the fact that $d_0 \leq D_C$ and Corollary 2.6, there exists

$$i \leq \bar{k} = \mathcal{O}\left(\max\left\{\left\lceil \frac{d_0 D_C}{\lambda \bar{\varepsilon}} \right\rceil, \left\lceil \frac{d_0}{\sqrt{\lambda \bar{\varepsilon}}} \right\rceil\right\}\right) = \mathcal{O}\left(\left\lceil \frac{d_0 D_C}{\lambda \bar{\varepsilon}} \right\rceil\right)$$

such that

$$\max\left\{\langle v_i, x_i - x\rangle + \varepsilon_i : x \in C\right\} \leq D_C \|v_i\| + \varepsilon_i \leq \frac{\sqrt{2} D_C d_0}{(1-\sigma)^{1/2} \lambda \bar{k}} + \frac{\sigma d_0^2}{(1-\sigma)\lambda \bar{k}^2} \leq \bar{\varepsilon}.$$

Thus, x_i satisfies (10). Moreover, in view of (5) and the observation preceding Corollary 2.9, we conclude that $f(x_k) - f^* \leq f(x_i) - f^* \leq \bar{\varepsilon}$ for every $k \geq i$. □

Note that, under the assumption of Corollary 2.9, it is also possible to use the first stopping criterion discussed above with $D_0 = D_C$, namely $D_C^2/(2k\lambda) \leq \bar{\varepsilon}$. Clearly, the iteration-complexity bound for IPP-CO framework based on this stopping criterion would be $\mathcal{O}(\lceil D_C^2/(\lambda\bar{\varepsilon})\rceil)$, which is substantially worse than the one stated in Corollary 2.9 when $d_0 << D_C$.

Finally, observe also that the above discussion would also hold had we only made the weaker assumption that the compact convex set C is such that $x_0 \in C$ and $C \cap X^* \neq \emptyset$.

3 Application I: Forward-backward splitting method

In this section, we show that the well-known forward-backward splitting method (see for example [5]) is a special case of the IPP-CO framework described in the previous section.

In this section, we assume that

S.1) $h : \mathcal{X} \to \bar{\mathbb{R}}$ is a proper closed convex function and $p : \mathcal{X} \to \bar{\mathbb{R}}$ is a proper function such that $\operatorname{dom} p \supseteq \operatorname{dom} h$;

S.2) p is convex on $\operatorname{dom} h$ and there exists an L-Lipschitz function $g : \operatorname{Dom} g \subset \mathcal{X} \to \mathcal{X}$ such that $\operatorname{Dom} g \supseteq \operatorname{dom} h$ and

$$0 \leq p(\tilde{x}) - p(x) - \langle g(x), \tilde{x} - x\rangle \leq \frac{L}{2}\|\tilde{x} - x\|^2, \quad \forall x, \tilde{x} \in \operatorname{dom} h; \tag{11}$$

and consider the optimization problem (2) in which the objective function f is assumed to have the following structure:

$$f(x) := \begin{cases} p(x) + h(x), & x \in \operatorname{dom} h, \\ +\infty, & x \notin \operatorname{dom} h. \end{cases} \tag{12}$$

We now discuss Assumption S.2. If p is differentiable and convex on $\operatorname{dom} h$ and ∇p is L-Lipschitz continuous on $\operatorname{dom} h$, then $g = \nabla p$ satisfies S.2. However, the weaker assumption S.2 do not require p to be differentiable on $\operatorname{dom} h$, and hence to be defined in a neighborhood of $\operatorname{dom} h$. This generality will be particularly useful when dealing with the primal function of a convex-concave saddle function (see Section 4). It can also be shown that the second inequality in (11) is implied by the other conditions assumed in S.2, and hence can be dropped.

The following simple result, whose proof is given in the appendix, establishes the lower semicontinuity of f.

Proposition 3.1. *Under Assumptions S.1 and S.2, the function f defined in* (12) *is proper closed convex.*

We now state the algorithm we are interested in studying in this section.

Algorithm I (Forward-backward splitting algorithm for (2)-(12)):

0) Let $x_0 \in \mathcal{X}$ and $0 \leq \sigma < 1$ be given and set $\lambda = \sigma/L$ and $k = 1$;

1) compute $x_k \in \mathcal{X}$ as
$$x_k = (I + \lambda \partial h)^{-1}(x_{k-1} - \lambda g(x_{k-1})); \tag{13}$$

2) set $k \leftarrow k + 1$ and go to step 1.

end

We now explain the terminology "forward-backward splitting" used by Algorithm I. If p is a differentiable convex function with L-Lipschitz continuous gradient, then (13) becomes

$$x_k = (I + \lambda \partial h)^{-1}(x_{k-1} - \lambda \nabla p(x_{k-1})).$$

Note that in this case, this algorithm is a particular case of a more general method which iterates as

$$x_{k+1} = (I + \lambda A)^{-1}(x_{k-1} - \lambda B(x_{k-1})) = (I + \lambda A)^{-1}(I - \lambda B)(x_{k-1}),$$

where A is a point-to-set maximal monotone operator and B is a point-to-point monotone map. According to [3, 26, 5], the above method is called the forward-backward splitting method, and converges to a solution of the inclusion $0 \in (A+B)(x)$, whenever B is L-co-coercive and $0 < \lambda < 1/L$. Its origin dates back to [7, 1, 8, 15] or even earlier (see [26] and the references therein). In addition to convex optimization, this method has also been used to solve variational and complementarity problems as far back as 1982 (see [14, 6, 25, 24]). According to [3, 5], the easily computable step

$$y_k = (I - \lambda B)(x_{k-1}) = x_{k-1} - \lambda B(x_{k-1}),$$

is the "forward" step in the direction $-B(x_{k-1})$, while the backward step is the evaluation of the resolvent

$$x_k = (I + \lambda A)^{-1}(y_k).$$

In the remaining part of this section, we study the iteration-complexity of Algorithm I. Our first goal is show that Algorithm I is a special instance of the IPP-CO framework. We start by stating the following well-known transportation formula for the subgradient of a proper convex function.

Lemma 3.2. *If $\phi : \mathcal{X} \to \bar{\mathbb{R}}$ be a proper convex function and $x, \tilde{x}, v \in \mathcal{X}$ are such that $v \in \partial\phi(x)$ and $\phi(\tilde{x}) < \infty$, then $v \in \partial_\varepsilon \phi(\tilde{x})$ for every $\varepsilon \geq \phi(\tilde{x}) - [\phi(x) + \langle \tilde{x} - x, v\rangle]$.*

We are now ready to show that Algorithm I is a special case of the IPP-CO framework.

Theorem 3.3. *Consider the sequence $\{x_k\}$ generated by Algorithm I and the sequences $\{h_k\}$, $\{\varepsilon_k\}$, $\{\sigma_k\}$ and $\{\lambda_k\}$ defined for every $k \in \mathbb{N}$ as $\sigma_k = \sigma$, $\lambda_k = \lambda$,*

$$h_k := \frac{1}{\lambda}(x_{k-1} - x_k) - g(x_{k-1}), \qquad \varepsilon_k := p(x_k) - p(x_{k-1}) - \langle g(x_{k-1}), x_k - x_{k-1}\rangle.$$

Then, for every $k \in \mathbb{N}$, (3) holds with $f = p + h$. As a consequence, Algorithm I is a special case of the IPP-CO framework.

Proof. First note that (13) and the definition of h_k imply that $h_k \in \partial h(x_k)$. Now, consider the function $\tilde{p} : \mathcal{X} \to \bar{\mathbb{R}}$ defined as

$$\tilde{p}(x) = \begin{cases} p(x), & x \in \operatorname{dom} h, \\ \infty, & x \notin \operatorname{dom} h. \end{cases} \tag{14}$$

Clearly, $\tilde{p}$ is a proper convex function, $f = \tilde{p} + h$ and, in view of the first inequality in (11), $g(x) \in \partial\tilde{p}(x)$ for every $x \in \operatorname{dom} h$. Moreover, using the definition of ε_k and Lemma 3.2 with $(\phi, x, \tilde{x}, v) = (\tilde{p}, x_{k-1}, x_k, g(x_{k-1}))$, we conclude that $g(x_{k-1}) \in \partial_{\varepsilon_k}\tilde{p}(x_k)$, and hence that

$$\frac{x_{k-1} - x_k}{\lambda_k} = g(x_{k-1}) + h_k \in \partial_{\varepsilon_k}\tilde{p}(x_k) + \partial h(x_k) \subseteq \partial_{\varepsilon_k}(\tilde{p} + h)(x_k) = \partial_{\varepsilon_k} f(x_k), \tag{15}$$

where the first identity is due to the definition of h_k and λ_k, and the last inclusion follows from a well-known property about subgradients. Also, the second inequality in (11), the definition of ε_k, λ_k and σ_k, and the assumption that $\lambda = \sigma/L$, imply that

$$2\lambda_k\varepsilon_k = 2\lambda\varepsilon_k \leq \lambda L\|x_k - x_{k-1}\|^2 = \sigma_k\|x_k - x_{k-1}\|^2.$$ □

The complexity result below follows as a consequence of the above result and Corollary 2.6. A similar result was obtained in Theorem 4 of [13] under the assumption that the sublevel sets of f are bounded, and the bounds are expressed in terms of the diameter of the sublevel set determined by the initial iterate. On the hand, the result below gives bounds in terms of d_0 and does not assume that the sublevel sets of f are bounded.

Corollary 3.4. *Consider the sequence $\{x_k\}$ generated by Algorithm I and the sequences $\{h_k\}$ and $\{\varepsilon_k\}$ defined as in Theorem 3.3. Then, for every $k \in \mathbb{N}$, the following statements hold:*

a) $f(x_k) - f^* \leq Ld_0^2/(2k\sigma)$;

b) $g(x_{k-1}) \in \partial_{\varepsilon_k} p(x_k)$ and $h_k \in \partial h(x_k)$, and hence $g(x_{k-1}) + h_k \in \partial_{\varepsilon_k} f(x_k)$; moreover, there exists $i \leq k$ such that

$$\|g(x_{i-1}) + h_i\| \leq \frac{\sqrt{2}}{(1-\sigma)^{1/2}\sigma}\,\frac{Ld_0}{k}, \qquad \varepsilon_i \leq \frac{Ld_0^2}{(1-\sigma)k^2},$$

and

$$\|g(x_i)+h_i\| \le \frac{\sqrt{2}(1+\sigma)}{(1-\sigma)^{1/2}\sigma}\,\frac{Ld_0}{k}.$$

Proof. By Theorem 3.3, Algorithm I is a special case of the IPP-CO framework with $\lambda_k=\lambda$ for every $k\in\mathbb{N}$. This observation together with Corollary 2.6(a), and the assumption that $\lambda_k=\lambda=\sigma/L$ for every $k\in\mathbb{N}$, immediately imply a). We now prove b). First note that the inclusions in b) have already been shown in the proof of Theorem 3.3. The first two estimates in b) follow from Corollary 2.6(b), the assumption that $\lambda_k=\lambda=\sigma/L$ for every $k\in\mathbb{N}$, and the first identity in (15). Moreover, the last estimate follows from the first one, the triangle inequality for norms, and the fact that, by Assumption S.2 and the definition of h_k and λ, we have

$$\|g(x_i)-g(x_{i-1})\| \le L\|x_i-x_{i-1}\| = \lambda L\|g(x_{i-1})+h_i\| = \sigma\|g(x_{i-1})+h_i\|. \qquad \square$$

We end this section by making some remarks. First, when $h=\delta_X$ for some nonempty closed convex set $X\subseteq\mathcal{X}$, Algorithm I reduces to the classical projected gradient method, which is based on the following recursive formula

$$x_k = P_X(x_{k-1}-\lambda g(x_{k-1})). \tag{16}$$

This is due to the fact that the resolvent $(I+\lambda\partial h)^{-1}$ of ∂h in expression (13) is exactly the projection operator P_X onto X. Second, all the analysis of this section holds for any $L\ge L_g$, where L_g is the smallest Lipschitz constant for the map g. If a constant $L\ge L_g$ is not known a priori, then we can use the ideas of [13] for estimating such a constant. Third, in fact Algorithm I does not need to work with a fixed stepsize $\lambda=\sigma/L$ for some $L\ge L_g$, but only with an adaptive stepsize $\lambda_k>0$ such that

$$2\lambda_k\left[p(x_k)-p(x_{k-1})-\langle x_k-x_{k-1},g(x_{k-1})\rangle\right] \le \sigma\|x_k-x_{k-1}\|^2.$$

where $x_k:=(I+\lambda_k\partial h)^{-1}(I-\lambda_k g)(x_{k-1})$.

4 Application II: A saddle point problem

In this section, we consider an optimization problem of the form (2)-(12), where the function p is assumed to be the primal function associated with a convex-concave saddle function. We then develop an inexact forward-backward splitting method for solving it in which the gradient of p is computed only in an approximate sense.

We will now describe the structure of the function p in detail. Let $\mathcal{Y}$ denote another finite dimensional inner product space with inner product and associated norm also denoted by $\langle\cdot,\cdot\rangle$ and $\|\cdot\|$, respectively. Let $\Psi:\operatorname{Dom}\Psi\subseteq\mathcal{X}\times\mathcal{Y}\to\mathbb{R}$ and convex sets $X\subseteq\mathcal{X}$ and $Y\subseteq\mathcal{Y}$ such that

$$X\times Y\subseteq\operatorname{Dom}\Psi, \qquad \operatorname{dom}h\subseteq X \tag{17}$$

be given. Assume that:

C.1) for every $y\in Y$, the function $\Psi(\cdot,y)$ is differentiable and convex on X;

C.2) there exist $L_{xx},L_{xy}\ge 0$ such that

$$\|\nabla_x\Psi(x',y')-\nabla_x\Psi(x,y)\| \le L_{xx}\|x'-x\|+L_{xy}\|y'-y\|, \qquad \forall (x,y),(x',y')\in X\times Y;$$

C.3) there exists $\beta > 0$ such that, for every $x \in X$, the function $\Psi_Y(x, \cdot) : \mathcal{X} \to (-\infty, \infty]$ defined as

$$\Psi_Y(x,y) = \begin{cases} \Psi(x,y), & y \in Y, \\ -\infty, & y \notin Y, \end{cases}$$

is an upper semi-continuous β-strongly concave function.

The function $p : \mathcal{X} \to \bar{\mathbb{R}}$ is then defined as

$$p(x) := \left\{ \begin{array}{cl} \sup_{y \in Y} \Psi(x,y), & \text{if } x \in X; \\ +\infty, & \text{otherwise.} \end{array} \right. \tag{18}$$

In view of conditions C.1 and C.3 and assumption (17), p is a proper convex function such that $\operatorname{dom} p = X \supseteq \operatorname{dom} h$. This observation together with Proposition 4.1 below imply that the functions p and h, and the map $g : X \to \mathcal{X}$ defined as

$$g(x) = \nabla_x \Psi(x, y(x)), \quad \forall x \in X,$$

where

$$y(x) := \arg\max_{y \in Y} \Psi(x,y), \tag{19}$$

satisfy conditions S.1 and S.2 of Section 3. Hence, direct application of Algorithm I to problem (2)-(12), with p of the form (18), requires the computation of the exact solution of an optimization problem of the form (18) at every iteration in order to evaluate $g(x_{k-1})$ in (13). However, such an approach is only possible for those instances of (2)-(12)-(18) for which it is possible to compute $y(x)$ as in (19) for every $x \in X$. An natural idea to circumvent this drawback is to instead work with approximate solutions of (18) which, in view of the result below, yield approximate subgradients of p.

Proposition 4.1. *Let $\eta \geq 0$ and $\bar{x} \in X$ be given. Assume that $\bar{y} \in Y$ is such that*

$$p(\bar{x}) - \Psi(\bar{x}, \bar{y}) \leq \eta. \tag{20}$$

and define $\bar{g} := \nabla_x \Psi(\bar{x}, \bar{y})$ and

$$L := 2\left(L_{xx} + \frac{L_{xy}^2}{\beta}\right). \tag{21}$$

Then, the following statements hold:

a) $\bar{g} \in \partial_\eta p(\bar{x})$;

b) for every $x \in X$,

$$p(x) \leq p(\bar{x}) + \langle \bar{g}, x - \bar{x} \rangle + \frac{L}{2}\|x - \bar{x}\|^2 + \eta; \tag{22}$$

c) for every $\tilde{x} \in X$, we have $\bar{g} \in \partial_{\tilde{\eta}} p(\tilde{x})$, where

$$\tilde{\eta} = \tilde{\eta}(\tilde{x}) := 2\eta + \frac{L}{2}\|\tilde{x} - \bar{x}\|^2.$$

Proof. a) Using (20), the definition of p and Assumption C.1, we conclude that

$$\begin{aligned} p(x) - p(\bar{x}) &\geq p(x) - \Psi(\bar{x}, \bar{y}) - \eta \geq \Psi(x, \bar{y}) - \Psi(\bar{x}, \bar{y}) - \eta \\ &\geq \langle \nabla_x \Psi(\bar{x}, \bar{y}), x - \bar{x} \rangle - \eta, \quad \forall x \in X, \end{aligned}$$

and hence that that $\bar{g} \in \partial_\eta p(\bar{x})$.

b) Using Assumption C.3, Proposition B.2 with $h = -\Psi_Y(\bar{x}, \cdot)$, and the fact that $-p(\bar{x}) = \min\{-\Psi_Y(\bar{x}, y) : y \in \mathcal{Y}\}$, we conclude that

$$-\Psi(\bar{x}, y) \geq -p(\bar{x}) + \frac{\beta}{2}\left(\|y - \bar{y}\| - \sqrt{\frac{2\eta}{\beta}}\right)^2, \qquad \forall y \in Y.$$

Also, by Assumptions C.1 and C.2 and the definition of $\bar{g}$, we have

$$\begin{aligned} \Psi(\bar{x}, y) - \Psi(x, y) - \langle \bar{g}, \bar{x} - x \rangle &\geq \langle \nabla_x \Psi(x, y) - \nabla_x \Psi(\bar{x}, \bar{y}), \bar{x} - x \rangle \\ &\geq -\|\nabla_x \Psi(x, y) - \nabla_x \Psi(\bar{x}, \bar{y})\| \, \|x - \bar{x}\| \\ &\geq -\left(L_{xx}\|x - \bar{x}\| + L_{xy}\|y - \bar{y}\|\right) \|x - \bar{x}\|, \qquad \forall (x, y) \in X \times Y. \end{aligned}$$

Adding these two inequalities, we then conclude that for every $(x, y) \in X \times Y$,

$$\begin{aligned} \Psi(x, y) &\leq p(\bar{x}) + \langle \bar{g}, x - \bar{x} \rangle + L_{xx}\|x - \bar{x}\|^2 + L_{xy}\|y - \bar{y}\| \, \|x - \bar{x}\| - \frac{\beta}{2}\left(\|y - \bar{y}\| - \sqrt{\frac{2\eta}{\beta}}\right)^2 \\ &\leq p(\bar{x}) + \langle \bar{g}, x - \bar{x} \rangle + L_{xx}\|x - \bar{x}\|^2 + \max_{t \in \mathbb{R}}\left\{ L_{xy}\|x - \bar{x}\| t - \frac{\beta}{2}\left(t - \sqrt{\frac{2\eta}{\beta}}\right)^2 \right\} \\ &= p(\bar{x}) + \langle \bar{g}, x - \bar{x} \rangle + \left(L_{xx} + \frac{L_{xy}^2}{2\beta}\right)\|x - \bar{x}\|^2 + \sqrt{\frac{2\eta}{\beta}} L_{xy}\|x - \bar{x}\| \\ &\leq p(\bar{x}) + \langle \bar{g}, x - \bar{x} \rangle + \left(L_{xx} + \frac{L_{xy}^2}{\beta}\right)\|x - \bar{x}\|^2 + \eta. \end{aligned}$$

Inequality (22) now follows from the definition of p and the previous relation.

c) Since $\bar{g} \in \partial_\eta p(\bar{x})$, it follows that for any $x \in X$,

$$\begin{aligned} p(x) &\geq p(\bar{x}) + \langle \bar{g}, x - \bar{x} \rangle - \eta = p(\tilde{x}) + \langle \bar{g}, x - \tilde{x} \rangle - [\eta + p(\tilde{x}) - p(\bar{x}) - \langle \bar{g}, \tilde{x} - \bar{x} \rangle] \\ &\geq p(\tilde{x}) + \langle \bar{g}, x - \tilde{x} \rangle - \left[2\eta + \frac{L}{2}\|\tilde{x} - \bar{x}\|^2\right] \geq p(\tilde{x}) + \langle \bar{g}, x - \tilde{x} \rangle - \tilde{\eta}. \end{aligned}$$

Hence, $\bar{g} \in \partial_{\tilde{\eta}} p(\tilde{x})$. □

Our main goal now is to state an inexact version of Algorithm I for solving problem (2)-(12), with function p given by (18). The method, which we refer to as the inexact forward-backward splitting algorithm, is as follows.

Algorithm II (An inexact forward-backward splitting algorithm for (2)**-(12)-**(18)**):**

0) define L as in (21) and let $x_0 \in \mathcal{X}$, $0 \le \sigma < 1$, $\lambda \in (0, \sigma/L)$ and a tolerance pair $(\bar{\rho}, \bar{\varepsilon}) \in \mathbb{R}^2_{++}$ be given; set $k = 1$ and

$$\eta := \min\left\{ \frac{\bar{\rho}^2 \lambda(\sigma - \lambda L)}{4} \,,\, \frac{\bar{\varepsilon}(\sigma - \lambda L)}{2\sigma} \right\}; \tag{23}$$

1) use the black-box to compute $y_k \in Y$ such that

$$p(x_{k-1}) - \Psi(x_{k-1}, y_k) \le \max\left\{ \eta \,,\, \frac{(\sigma - \lambda L)}{4\lambda} \|x_k(y_k) - x_{k-1}\|^2 \right\}, \tag{24}$$

and set $x_k = x_k(y_k)$, where

$$x_k(y) := (I + \lambda \partial h)^{-1}(x_{k-1} - \lambda \nabla_x \Psi(x_{k-1}, y)), \qquad \forall y \in \mathcal{Y}; \tag{25}$$

2) if

$$\frac{\sigma - \lambda L}{4\lambda} \|x_k - x_{k-1}\|^2 \le \eta, \tag{26}$$

then **stop** and output $(x_k, v_k, \varepsilon_k)$, where

$$\varepsilon_k = 2 \max\left\{ \eta \,,\, \frac{(\sigma - \lambda L)}{4\lambda} \|x_k - x_{k-1}\|^2 \right\} + \frac{L}{2} \|x_k - x_{k-1}\|^2, \qquad v_k := \frac{x_{k-1} - x_k}{\lambda};$$

otherwise, set $k \leftarrow k + 1$ and go to step 1.

end

Note that step 2 of Algorithm II requires a *subroutine* that is able to obtain an approximate solution x_k of the problem $\min\{-\Psi(x_{k-1}, y) : y \in Y\}$ in the sense that its functional error $p(x_{k-1}) - \Psi(x_{k-1}, y_k)$ is bounded by an adaptive tolerance, i.e., the right hand side of (24), that has the following properties: i) it is bounded below by η; ii) it is larger than η in those iterations for which the stopping criterion (26) is not satisfied.

In what follows, we establish iteration-complexity bounds for Algorithm II by using the fact that it is a special instance of the IPP-CO framework.

Lemma 4.2. *The following statements hold:*

a) $v_k \in \partial_{\varepsilon_k} p(x_k) + \partial h(x_k) \subseteq \partial_{\varepsilon_k} f(x_k)$*;*

b) inequality (26) *holds if, and only if,* $\|v_k\| \le \bar{\rho}$ *and* $\varepsilon_k \le \bar{\varepsilon}$*.*

As a consequence, if Algorithm II stops at step 2, then x_k *is a* $(\bar{\rho}, \bar{\varepsilon})$*-solution of* (2)-(12)-(18) *and* (v_k, ε_k) *is a* $(\bar{\rho}, \bar{\varepsilon})$*-residual at* x_k*.*

Proof. We first prove a). First, note that (24), Proposition 4.1(c) and the definition of ε_k imply that

$$\nabla_x \Psi(x_{k-1}, y_k) \in \partial_{\varepsilon_k} p(x_k).$$

This inclusion, the definition of v_k, and (25) with $y = y_k$, then imply that

$$\begin{aligned} v_k = \frac{x_{k-1} - x_k}{\lambda} &\in \nabla_x \Psi(x_{k-1}, y_k) + \partial h(x_k) \\ &\in \partial_{\varepsilon_k} p(x_k) + \partial h(x_k) = \partial_{\varepsilon_k} [p + h](x_k) = \partial_{\varepsilon_k} f(x_k). \end{aligned} \tag{27}$$

We now prove b). Using the definition of v_k, we easily see that $\|v_k\| \le \bar{\rho}$ if, and only if,

$$\frac{\sigma - \lambda L}{4\lambda} \|x_k - x_{k-1}\|^2 \le \frac{\bar{\rho}^2 \lambda(\sigma - \lambda L)}{4}.$$

Moreover, using the definition of ε_k, we easily see that $\varepsilon_k \le \bar{\varepsilon}$ if, and only if,

$$\frac{\sigma - \lambda L}{4\lambda} \|x_k - x_{k-1}\|^2 \le \frac{(\sigma - \lambda L)}{2} \min\left\{ \frac{\bar{\varepsilon}}{\sigma}, \frac{\bar{\varepsilon} - 2\eta}{L\lambda} \right\} = \frac{\bar{\varepsilon}(\sigma - \lambda L)}{2\sigma},$$

where the last equality follows from the fact that $\eta \le \bar{\varepsilon}(\sigma - \lambda L)/(2\sigma)$, due to (23). In view of the definition of η in (23), we have thus shown that (b) holds. □

Lemma 4.3. *If Algorithm II does not stop at the k-th iteration, then (3) holds with $f = p + h$, $\sigma_k = \sigma$ and $\lambda_k = \lambda$.*

Proof. The assumption of the lemma implies that (26) does not hold. This together with the definition of ε_k then imply that

$$\varepsilon_k = \frac{(\sigma - \lambda L)}{2\lambda} \|x_k - x_{k-1}\|^2 + \frac{L}{2} \|x_k - x_{k-1}\|^2 = \frac{\sigma}{2\lambda} \|x_k - x_{k-1}\|^2,$$

which, together with (27), shows that (3) holds with $f = p + h$, $\sigma_k = \sigma$ and $\lambda_k = \lambda$. □

Theorem 4.4. *Algorithm II terminates in at most*

$$\mathcal{O}\left(\max\left\{ \left\lceil \frac{d_0}{\lambda \bar{\rho}} \right\rceil, \left\lceil \frac{d_0}{\sqrt{\lambda \bar{\varepsilon}}} \right\rceil \right\} \right) \tag{28}$$

iterations with a $(\bar{\rho}, \bar{\varepsilon})$-solution of (2)-(12)-(18) *together with a corresponding $(\bar{\rho}, \bar{\varepsilon})$-residual.*

Proof. Assume that Algorithm II has not stopped at the k-th iteration. In view of Lemma 4.3, it follows that Algorithm II (until the k-th iteration) is a special case of the IPP-CO framework in which $\lambda_i = \lambda$ for every $i = 1, \ldots, k$. Hence, in view of Corollary 2.8, we conclude that k is bounded above by (28). Thus, the conclusion of the theorem follows. □

Note that checking whether (24) holds requires that $p(\cdot)$ be evaluated at x_{k-1}, which is exactly what the approach described in this section is trying to avoid. For the sake of shortness, let η_k denote the right hand side of (24). Clearly, (24) is equivalent to the inclusion

$$0 \in \partial_{\eta_k} [-\Psi_Y(x_{k-1}, \cdot)](y_k),$$

It turns out that we may instead use the checking criterion:

$$w_k \in \partial_{\tau_k} [-\Psi_Y(x_{k-1}, \cdot)](y_k),$$

where $(w_k, \tau_k) \in \mathcal{Y} \times \mathbb{R}_+$ is a (small) residual pair. The result below shows that, as long as (w_k, τ_k) is sufficiently small, we can still guarantee that the first condition above holds. Moreover, it is worth noting that, in view of Theorem 2.5(b) and/or Corollary 2.6(b), any instance of the IPP-CO framework, and in particular Algorithm I, applied to the problem

$$\min_{y \in \mathcal{Y}} (-\Psi_Y)(x_{k-1}, y) = \max_{y \in Y} \Psi(x_{k-1}, y)$$

will eventually generate a pair as above, without any need to evaluate p.

Proposition 4.5. *Let $x \in X$ be given and assume that $(y, \varepsilon, w) \in Y \times \mathbb{R}_+ \times \mathcal{Y}$ satisfies*

$$w \in \partial_\varepsilon[-\Psi_Y(x, \cdot)](y). \tag{29}$$

Then,

$$p(x) - \Psi(x, y) \le \left(\frac{\|w\|}{\sqrt{2\beta}} + \sqrt{\varepsilon} \right)^2. \tag{30}$$

Proof. Define the function $\phi := -\Psi_Y(x, \cdot) - \langle w, \cdot \rangle$. Note that condition C.3 implies that ϕ is a proper lower semi-continuous β-strongly convex function. Moreover, assumption (29) is equivalent to the condition that $0 \in \partial_\varepsilon \phi(y)$, or equivalently

$$\phi(y) - \phi^* \le \varepsilon, \tag{31}$$

where $\phi^* := \inf\{\phi(y') : y' \in \mathcal{Y}\}$. Hence, it follows from Proposition B.2 that for every $\tilde{y} \in \mathcal{Y}$:

$$\phi(\tilde{y}) \ge \phi^* + \frac{\beta}{2} \left(\|\tilde{y} - y\| - \sqrt{\frac{2\varepsilon}{\beta}} \right)^2 \ge \phi(y) - \varepsilon + \frac{\beta}{2} \left(\|\tilde{y} - y\| - \sqrt{\frac{2\varepsilon}{\beta}} \right)^2,$$

where the last inequality is due to (31). Noting the definition of ϕ and $\Psi_Y(x, \cdot)$, we easily see that the above inequality implies that

$$\begin{aligned} \Psi(x, \tilde{y}) - \Psi(x, y) - \varepsilon &\le \langle w, \tilde{y} - y \rangle - \frac{\beta}{2} \left(\|\tilde{y} - y\| - \sqrt{\frac{2\varepsilon}{\beta}} \right)^2 \\ &\le \max \left\{ \|w\| \|y' - y\| - \frac{\beta}{2} \left(\|y' - y\| - \sqrt{\frac{2\varepsilon}{\beta}} \right)^2 \right\} \\ &= \sqrt{\frac{2\varepsilon}{\beta}} \|w\| + \frac{\|w\|^2}{2\beta}, \qquad \forall \tilde{y} \in Y. \end{aligned}$$

This inequality together with (18) then imply that (30) holds. □

In view of the above result, Algorithm II with step 1 replaced by the following alternative step would still possess all the convergence properties of its original version.

Step 1': Compute $(y_k, \varepsilon_k, w_k) \in Y \times \mathbb{R}_+ \times \mathcal{Y}$ such that

$$w_k \in \partial_{\varepsilon_k}[-\Psi_Y(x_{k-1}, \cdot)](y_k), \qquad \left(\frac{\|w_k\|}{\sqrt{2\beta}} + \sqrt{\varepsilon_k} \right)^2 \le \max \left\{ \eta \, , \, \frac{(\sigma - \lambda L)}{4\lambda} \|x_k(y_k) - x_{k-1}\|^2 \right\}.$$

5 Concluding Remarks

After the release of the first version of this work, Devolder at al. released the paper [4], where unaccelerated and/or accelerated inexact first-order (gradient) methods for convex optimization are proposed. In particular, they discuss how their methods can be used to minimize primal functions associated with convex-concave saddle-point problems, where the inner subproblems (needed to evaluate the primal functions) are solved inexactly. Motivated by their work, we have added a new section, namely Section 4, to the present version, dealing with the same type of saddle-based convex minimization problems. However, we note that the class of saddle functions considered here, i.e., those satisfying conditions C.1-C.3 of Section 4, are more general than those considered in Section 3.2 of [4].

Observe that condition C.3 requires that the saddle function $\Psi_Y(x,\cdot)$ be β-strongly concave for every $x \in X$. For saddle functions which satisfies C.1, C.2 and C.3 with $\beta = 0$, it is possible to add, for some small $\mu > 0$, a μ-strongly concave function on Y to Ψ to obtain a perturbed saddle function satisfying C.1, C.2 and C.3 with $\beta = \mu > 0$, to which Algorithm II can be applied. Under the assumption that Y is a compact convex set and by properly choosing $\mu > 0$, it is possible to present an unaccelerated smoothing minimization scheme where the primal function of the perturbed saddle function is minimized by an instance of the IPP-CO framework and the inner subproblems solved inexactly instead of exactly as in the accelerated smoothing minimization scheme of [12]. For the sake of shortness, we have omitted the details of the aforementioned smoothing minimization scheme.

A Proof of Proposition 3.1

Our goal in this section is to establish Proposition 3.1.

Lemma A.1. *Let $\phi : \mathcal{X} \to \bar{\mathbb{R}}$ be a proper convex function such that ϕ restricted to its domain is lower semi-continuous. Then,* $\mathrm{cl}\,\phi(x) = \phi(x)$ *for every $x \in \mathrm{dom}\,\phi$.*

Proof. We know that $(\mathrm{cl}\, f)(x) = \liminf_{y \to x} f(y)$ for every $x \in \mathcal{X}$. Since, by assumption,

$$\liminf_{\substack{y \to x \\ y \in \mathrm{dom}\, f}} f(y) = f(x), \quad \forall x \in X,$$

and $f(y) = \infty$ for every $y \notin \mathrm{dom}\,\phi$, the result follows. □

Proof of Proposition 3.1. Consider the function $\tilde{p}$ defined according to (14). We know that $\tilde{p}$ is a proper convex function and $f = \tilde{p} + h$. Since $\mathrm{dom}\,\tilde{p} = \mathrm{dom}\, h \neq \emptyset$, it follows from Theorem 9.3 of [16] and Assumption S.1 that $\mathrm{cl}\, f = \mathrm{cl}\,\tilde{p} + h$. Moreover, since $\tilde{p}$ is continuous on its domain due to (11), we conclude from Lemma A.1 that $\mathrm{cl}\,\tilde{p}$ and $\tilde{p}$ coincide on $\mathrm{dom}\,\tilde{p} = \mathrm{dom}\, g$. Based on these observations, we can now easily see that $\mathrm{cl}\, f = f$. □

B A technical result on convex optimization

In this section, we establish a technical result, namely Proposition B.2, needed in the proof of Proposition 4.1.

Proposition B.1. *Let $\phi : \mathcal{X} \to (-\infty, \infty]$ be a proper lower semi-continuous β-strongly convex function. Then, the problem*

$$\phi^* := \inf\{\phi(x) : x \in \mathcal{X}\} \tag{32}$$

has a unique optimal solution $x^* \in \mathcal{X}$ and

$$\phi(x) \geq \phi^* + \frac{\beta}{2}\|x - x^*\|^2, \qquad \forall x \in \mathcal{X}. \tag{33}$$

Using the above proposition, we can now establish the following variant of the above result.

Proposition B.2. *Let $\phi : \mathcal{X} \to (-\infty, \infty]$ be a proper lower semi-continuous β-strongly convex function and assume that $\bar{x}$ is a η-approximate solution of* (32)*, i.e., it satisfies*

$$\phi(\bar{x}) - \phi^* \leq \eta. \tag{34}$$

Then,

$$\phi(x) \geq \phi^* + \frac{\beta}{2}\left(\|x - \bar{x}\| - \sqrt{\frac{2\eta}{\beta}}\right)^2, \qquad \forall x \in \mathcal{X}.$$

Proof. By (33) with $x = \bar{x}$ and (34), we have

$$\frac{\beta}{2}\|\bar{x} - x^*\|^2 \leq \phi(\bar{x}) - \phi^* \leq \eta.$$

This inequality together with (33) then imply that

$$\phi(x) - \phi^* \geq \frac{\beta}{2}\|x - x^*\|^2 \geq \frac{\beta}{2}\left(\|x - \bar{x}\| - \|\bar{x} - x^*\|\right)^2 \geq \frac{\beta}{2}\left(\|x - \bar{x}\| - \sqrt{\frac{2\eta}{\beta}}\right)^2. \qquad \square$$

References

[1] R. E. Bruck, Jr. An iterative solution of a variational inequality for certain monotone operators in Hilbert space. *Bull. Amer. Math. Soc.*, 81(5):890–892, 1975.

[2] R. S. Burachik, Alfredo N. Iusem, and B. F. Svaiter. Enlargement of monotone operators with applications to variational inequalities. *Set-Valued Anal.*, 5(2):159–180, 1997.

[3] G. H.-G. Chen and R. T. Rockafellar. Convergence rates in forward-backward splitting. *SIAM J. Optim.*, 7(2):421–444, 1997.

[4] O. Devolder, F. Glineur, and Y. E. Nesterov. First-order methods of smooth convex optimization with inexact oracle. Core discussion paper 2011/02, Center for Operations Research and Econometrics (CORE), Catholic University of Louvain, December 2010.

[5] F. Facchinei and J.-S. Pang. *Finite-dimensional variational inequalities and complementarity problems, Volume II.* Springer-Verlag, New York, 2003.

[6] D. Gabay. Applications of the method of multipliers to variational inequalities. In M. Fortin and R. Glowinski, editors, *Augmented Lagrangian Methods: Applications to the Numerical Solution of Boundary-value Problems*, Amsterdam, 1983. North-Holland Publishing Company.

[7] A. A. Goldstein. Convex programming in Hilbert space. *Bull. Amer. Math. Soc.*, 70:709–710, 1964.

[8] P.-L. Lions and B. Mercier. Splitting algorithms for the sum of two nonlinear operators. *SIAM J. Numer. Anal.*, 16(6):964–979, 1979.

[9] B. Martinet. Régularisation d'inéquations variationnelles par approximations successives. *Rev. Française Informat. Recherche Opérationnelle*, 4(Ser. R-3):154–158, 1970.

[10] R. D. C. Monteiro and B. F. Svaiter. Complexity of variants of Tseng's modified F-B splitting and Korpelevich's methods for generalized variational inequalities with applications to saddle point and convex optimization problems. Manuscript, School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0205, USA, July 2010. To appear in *SIAM Journal on Optimization*.

[11] R. D. C. Monteiro and B. F. Svaiter. On the complexity of the hybrid proximal extragradient method for the iterates and the ergodic mean. *SIAM J. Optim.*, 20(6):2755–2787, 2010.

[12] Y. E. Nesterov. Smooth minimization of nonsmooth functions. *Mathematical Programming*, 103:127–152, 2005.

[13] Y. E. Nesterov. Gradient methods for minimizing composite objective function. Core discussion paper 2007/96, Center for Operations Research and Econometrics (CORE), Catholic University of Louvain, September 2007.

[14] J. S. Pang and D. Chan. Iterative methods for variational and complementarity problems. *Math. Programming*, 24(3):284–313, 1982.

[15] G. B. Passty. Ergodic convergence to a zero of the sum of monotone operators in Hilbert space. *J. Math. Anal. Appl.*, 72(2):383–390, 1979.

[16] R. T. Rockafellar. *Convex Analysis.* Princeton University Press, Princeton, NJ, 1970.

[17] R. T. Rockafellar. On the maximal monotonicity of subdifferential mappings. *Pacific J. Math.*, 33:209–216, 1970.

[18] R. T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM J. Control Optimization*, 14(5):877–898, 1976.

[19] M. V. Solodov and B. F. Svaiter. A hybrid approximate extragradient-proximal point algorithm using the enlargement of a maximal monotone operator. *Set-Valued Anal.*, 7(4):323–345, 1999.

[20] M. V. Solodov and B. F. Svaiter. A hybrid projection-proximal point algorithm. *J. Convex Anal.*, 6(1):59–70, 1999.

[21] M. V. Solodov and B. F. Svaiter. An inexact hybrid generalized proximal point algorithm and some new results on the theory of Bregman functions. *Math. Oper. Res.*, 25(2):214–230, 2000.

[22] M. V. Solodov and B. F. Svaiter. A unified framework for some inexact proximal point algorithms. *Numer. Funct. Anal. Optim.*, 22(7-8):1013–1035, 2001.

[23] B. F. Svaiter. A family of enlargements of maximal monotone operators. *Set-Valued Anal.*, 8(4):311–328, 2000.

[24] P. Tseng. Further applications of a splitting algorithm to decomposition in variational inequalities and convex programming. *Math. Programming*, 48(2, (Ser. B)):249–263, 1990.

[25] P. Tseng. Applications of a splitting algorithm to decomposition in convex programming and variational inequalities. *SIAM J. Control Optim.*, 29(1):119–138, 1991.

[26] C. Y. Zhu. Asymptotic convergence analysis of the forward-backward splitting algorithm. *Math. Oper. Res.*, 20(2):449–464, 1995.