Primal-dual extrapolation methods for monotone inclusions under local Lipschitz continuity with applications to variational inequality, conic constrained saddle point, and convex conic optimization problems

Zhaosong Lu * Sanyou Mei *

June 1, 2022 (Revised: June 24, 2022)

Abstract

In this paper we consider a class of structured monotone inclusion (MI) problems that consist of finding a zero in the sum of two monotone operators, in which one is maximal monotone while another is *locally Lipschitz* continuous. In particular, we first propose a primal-dual extrapolation (PDE) method for solving a structured strongly MI problem by modifying the classical forward-backward splitting method by using a point and operator extrapolation technique, in which the parameters are adaptively updated by a backtracking line search scheme. The proposed PDE method is *almost parameter-free*, equipped with a *verifiable* termination criterion, and enjoys an operation complexity of $\mathcal{O}\left(\log\varepsilon^{-1}\right)$, measured by the amount of fundamental operations consisting only of evaluations of one operator and resolvent of another operator, for finding an ε-residual solution of the structured strongly MI problem. We then propose another PDE method for solving a structured non-strongly MI problem by applying the above PDE method to approximately solve a sequence of structured strongly MI problems. The resulting PDE method is *parameter-free*, equipped with a *verifiable* termination criterion, and enjoys an operation complexity of $\mathcal{O}\left(\varepsilon^{-1}\log\varepsilon^{-1}\right)$ for finding an ε-residual solution of the structured non-strongly MI problem. As a consequence, we apply the latter PDE method to convex conic optimization, conic constrained saddle point, and variational inequality problems, and obtain complexity results for finding an ε-KKT or ε-residual solution of them under local Lipschitz continuity. To the best of our knowledge, no prior studies were conducted to investigate methods with complexity guarantees for solving the aforementioned problems under local Lipschitz continuity. All the complexity results obtained in this paper are entirely new.

Keywords: Local Lipschitz continuity, primal-dual extrapolation, operator splitting, monotone inclusion, convex conic optimization, saddle point, variational inequality, iteration complexity, operation complexity

Mathematics Subject Classification: 47H05, 47J20, 49M29, 65K15, 90C25

1 Introduction

A broad range of optimization, saddle point (SP), and variational inequality (VI) problems can be solved as a *monotone inclusion* (MI) problem, namely, finding a point x such that $0 \in \mathscr{T}(x)$, where $\mathscr{T}:\mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is a maximal monotone set-valued (i.e., point-to-set) operator (see Section

*Department of Industrial and Systems Engineering, University of Minnesota, USA (email: zhaosong@umn.edu, mei00035@umn.edu).

1.1 for the definition of monotone and maximal monotone operators). In this paper we consider a class of MI problems as follows:

$$\text{find } x \in \mathbb{R}^n \text{ such that } 0 \in (F+B)(x), \tag{1}$$

where $B : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is a maximal monotone set-valued operator with a nonempty convex domain denoted by $\mathrm{dom}(B)$, and $F : \mathrm{dom}(F) \to \mathbb{R}^n$ is a monotone point-valued (i.e., point-to-point) operator with a domain $\mathrm{dom}(F) \supseteq \mathrm{cl}(\mathrm{dom}(B))$. It shall be mentioned that $\mathrm{dom}(B)$ is possibly *unbounded*. We make the following additional assumptions throughout this paper.

Assumption 1. *(a) Problem* (1) *has at least one solution.*

(b) $F+B$ *is monotone on* $\mathrm{dom}(B)$ *with a monotonicity parameter* $\mu \geq 0$ *such that*

$$\langle u - v, x - y \rangle \geq \mu \|x - y\|^2 \qquad \forall x, y \in \mathrm{dom}(B), u \in (F+B)(x), v \in (F+B)(y). \tag{2}$$

(c) F *is locally Lipschitz continuous on* $\mathrm{cl}(\mathrm{dom}(B))$*.*

(d) The resolvent of γB *can be exactly evaluated for any* $\gamma > 0$*.*

The *local Lipschitz continuity* of F on $\mathrm{cl}(\mathrm{dom}(B))$ is generally weaker than the (global) Lipschitz continuity of F on $\mathrm{cl}(\mathrm{dom}(B))$ usually imposed in the literature. It can sometimes be easily verified. For example, if F is differentiable on $\mathrm{cl}(\mathrm{dom}(B))$, it is clearly locally Lipschitz continuous there. In addition, by the maximal monotonicity of B and Assumptions 1(b) and 1(c), it can be observed that $F+B$ is maximal monotone (e.g., see [12, Proposition A.1]) and it is also strongly monotone when $\mu > 0$.

Some *special cases* of problem (1) have been considerably studied and can be solved by some existing methods. For example, when F is *cocoercive*[1], problem (1) can be suitably solved by the newly developed Halpern fixed-point splitting (HFPS) method [18], and also the classical forward-backward splitting (FBS) method [7, 15] that generates a solution sequence $\{x^k\}$ according to

$$x^{k+1} = (I + \gamma_k B)^{-1} \left(x^k - \gamma_k F(x^k) \right) \quad \forall k \geq 1.$$

In addition, a modified FBS (MFBS) method [19] and its variant [11] were proposed for (1) *with* F *being Lipschitz continuous* on $\mathrm{cl}(\mathrm{dom}(B))$. An *operation complexity* of $\mathcal{O}\left(\varepsilon^{-1}\right)$ and $\mathcal{O}\left(\varepsilon^{-2}\right)$, measured by the amount of fundamental operations consisting of evaluations of F and resolvent of B, was respectively established for the HFPS method [18, Theorem 3.1] and the variant of Tseng's MFBS method [11, Theorem 4.6] for finding an *ε-residual solution*[2] of these special cases of problem (1).

There has however been little algorithmic development for solving problem (1) with F being locally Lipschitz continuous. The forward-reflected-backward splitting (FRBS) method [9] appears to be the only existing method in the literature for solving problem (1). It modifies the forward term in the FBS method by using an operator extrapolation technique that has been popularly used to design algorithms for solving optimization, SP, and VI problems (e.g., [16, 10, 5, 2, 3]). Specifically, the FRBS method generates a solution sequence $\{x^k\}$ according to

$$x^{k+1} = (I + \gamma_k B)^{-1} \left(x^k - \gamma_k F(x^k) - \gamma_{k-1}(F(x^k) - F(x^{k-1})) \right) \quad \forall k \geq 1 \tag{3}$$

for a suitable choice of stepsizes $\{\gamma_k\}$. Interestingly, it is established in [9, Theorem 3.4] that $\{x^k\}$ converges weakly to a solution of problem (1), only assuming that F is monotone and

[1] F is cocoercive if there exists some $\sigma > 0$ such that $\langle F(x) - F(y), x - y \rangle \geq \sigma \|F(x) - F(y)\|^2$ for all $x, y \in \mathrm{dom}(F)$. It can be observed that if F is cocoercive, then it is monotone and Lipschitz continuous on $\mathrm{dom}(F)$.

[2] An ε-residual solution of problem (1) is a point $x \in \mathrm{dom}(B)$ satisfying $\mathrm{res}_{F+B}(x) \leq \varepsilon$, where $\mathrm{res}_{F+B}(x) = \inf\{\|v\| : v \in (F+B)(x)\}$.

locally Lipschitz continuous. Despite this nice convergence result, there is a lack of complexity guarantees for the FRBS method for finding an approximate solution of (1). Indeed, it is generally impossible to find an exact solution of problem (1) by this method. Consequently, when applying this method to solve (1), one has to terminate it once a certain approximate solution is found. Two natural questions then arise: (i) what approximate solution of (1) shall be found by this method in practice? (ii) what is the (worst-case) operation complexity of this method for finding such an approximate solution? The answer to these questions yet remains unknown.

To the best of our knowledge, there has been no method yet *with complexity guarantees* for finding an approximate solution of problem (1). In this paper we propose new variants of FBS method, called *primal-dual extrapolation* (PDE) methods, for finding an *ε-residual solution* of (1) with complexity guarantees. In particular, we first propose a PDE method for solving a strongly MI problem, namely, problem (1) with $\mu > 0$, by modifying the forward term in the FBS method by using a *point and operator extrapolation technique* that has recently been proposed to design algorithms for solving stochastic VI problems in [2]. Specifically, this PDE method generates a solution sequence $\{x^k\}$ according to

$$x^{k+1} = (I + \gamma_k B)^{-1} \left(x^k + \alpha_k (x^k - x^{k-1}) - \gamma_k [F(x^k) + \beta_k (F(x^k) - F(x^{k-1}))] \right) \quad \forall k \geq 1,$$

where the sequences $\{\alpha_k\}$, $\{\beta_k\}$ and $\{\gamma_k\}$ are adaptively updated by a backtracking line search scheme (see Algorithm 1). We show that this PDE method enjoys an operation complexity of $\mathcal{O}\left(\log \varepsilon^{-1}\right)$ for finding an ε-residual solution of (1) with $\mu > 0$. We then propose another PDE method for solving a non-strongly MI problem, namely, problem (1) with $\mu = 0$ by applying the above PDE method to approximately solve a sequence of strongly MI problems $0 \in (F_k + B)(x)$ with F_k being a perturbation of F (see Algorithm 2). We show that the resulting PDE method enjoys an operation complexity of $\mathcal{O}\left(\varepsilon^{-1} \log \varepsilon^{-1}\right)$ for finding an ε-residual solution of problem (1) with $\mu = 0$.

The main contributions of our paper are summarized as follows.

- Primal-dual extrapolation methods are proposed for the MI problem (1). These methods enjoy several attractive features: (i) they are applicable to a broad range of problems since only local rather than global Lipschitz continuity of F is required; (ii) they are computationally cheap and their fundamental operations consist only of evaluations of F and resolvent of B;[3] (iii) they are (almost) parameter-free, equipped with a verifiable termination criterion, and guaranteed to output an ε-residual solution of problem (1) with complexity guarantees.

- Our proposed methods are the first ones in the literature with complexity guarantees for solving problem (1) under local Lipschitz continuity of F. In particular, they enjoy an operation complexity of $\mathcal{O}\left(\log \varepsilon^{-1}\right)$ and $\mathcal{O}\left(\varepsilon^{-1} \log \varepsilon^{-1}\right)$ for finding an ε-residual solution of problem (1) with $\mu > 0$ and $\mu = 0$, respectively. These complexity results are entirely new and fill the research gap in the relevant topics. Even for the case where F is globally Lipschitz continuous, our operation complexity of $\mathcal{O}\left(\varepsilon^{-1} \log \varepsilon^{-1}\right)$ is significantly better than the previously best-known operation complexity of $\mathcal{O}\left(\varepsilon^{-2}\right)$ obtained in [11, Theorem 4.6] for finding an ε-residual solution of a non-strongly MI problem, namely, problem (1) with $\mu = 0$.

- The applications of our proposed methods to convex conic optimization, conic constrained SP, and VI problems are studied. Complexity results for finding an ε-KKT or ε-residual solution of these problems under local Lipschitz continuity are obtained for the first time.

[3] Since the evaluation of $(I + \gamma B)^{-1}(x)$ is often as cheap as that of $(I + B)^{-1}(x)$, we count the evaluation of $(I + \gamma B)^{-1}(x)$ as *one evaluation of resolvent* of B for any $\gamma > 0$ and x.

The rest of this paper is organized as follows. In Section 1.1 we introduce some notation and terminology. In Sections 2 and 3, we propose PDE methods for problem (1) with $\mu > 0$ and $\mu = 0$, respectively, and study their worst-case complexity. In Section 4 we study the applications of the PDE methods for solving convex conic optimization, conic constrained saddle point, and variational inequality problems. Finally, we present the proofs of the main results in Section 5.

1.1 Notation and terminology

The following notations will be used throughout this paper. Let $\mathbb{R}^n$ denote the Euclidean space of dimension n, $\langle \cdot, \cdot \rangle$ denote the standard inner product, and $\|\cdot\|$ stand for the Euclidean norm. For any $\omega \in \mathbb{R}$, let $\omega_+ = \max\{\omega, 0\}$ and $\lceil \omega \rceil$ denote the least integer number greater than or equal to ω.

Given a proper closed convex function $h : \mathbb{R}^n \to (-\infty, \infty]$, ∂h denotes its subdifferential. The proximal operator associated with h is denoted by prox_h, which is defined as

$$\mathrm{prox}_h(z) = \arg\min_{x \in \mathbb{R}^n} \left\{ \frac{1}{2}\|x - z\|^2 + h(x) \right\} \qquad \forall z \in \mathbb{R}^n.$$

Given an operator $\mathscr{T}$, $\mathrm{dom}(\mathscr{T})$ and $\mathrm{cl}(\mathrm{dom}(\mathscr{T}))$ denote its domain and the closure of its domain, respectively. For a mapping $g : \mathbb{R}^n \to \mathbb{R}^m$, ∇g denotes the transpose of the Jacobian of g. The mapping g is called L-Lipschitz continuous on a set Ω for some constant $L > 0$ if $\|g(x) - g(y)\| \le L\|x - y\|$ for all $x, y \in \Omega$. Let I stand for the identity operator. For a maximal monotone operator $\mathscr{T} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$, the resolvent of $\mathscr{T}$ is denoted by $(I + \mathscr{T})^{-1}$, which is a mapping defined everywhere in $\mathbb{R}^n$. In particular, $z = (I + \mathscr{T})^{-1}(x)$ if and only if $x \in (I + \mathscr{T})(z)$. Since the evaluation of $(I + \gamma\mathscr{T})^{-1}(x)$ is often as cheap as that of $(I + \mathscr{T})^{-1}(x)$, we count the evaluation of $(I + \gamma\mathscr{T})^{-1}(x)$ as *one evaluation of resolvent* of $\mathscr{T}$ for any $\gamma > 0$ and x. The *residual* of $\mathscr{T}$ at a point $x \in \mathrm{dom}(\mathscr{T})$ is defined as $\mathrm{res}_{\mathscr{T}}(x) = \inf\{\|v\| : v \in \mathscr{T}(x)\}$. For any given $\varepsilon > 0$, a point x is called an *ε-residual solution* of problem (1) if $x \in \mathrm{dom}(B)$ and $\mathrm{res}_{F+B}(x) \le \varepsilon$.

Given a nonempty closed convex set $C \subseteq \mathbb{R}^n$, $\mathrm{dist}(z, C)$ stands for the Euclidean distance from z to C, and $\Pi_C(z)$ denotes the Euclidean projection of z onto C, namely,

$$\Pi_C(z) = \arg\min\{\|z - x\| : x \in C\}, \quad \mathrm{dist}(z, C) = \|z - \Pi_C(z)\|, \quad \forall z \in \mathbb{R}^n.$$

The normal cone of C at any $z \in C$ is denoted by $\mathcal{N}_C(z)$. For the closed convex cone $\mathcal{K}$, we use $\mathcal{K}^*$ to denote the dual cone of $\mathcal{K}$, that is, $\mathcal{K}^* = \{y \in \mathbb{R}^m : \langle y, x \rangle \ge 0, \ \forall x \in \mathcal{K}\}$.

2 A primal-dual extrapolation method for problem (1) with $\mu > 0$

In this section we propose a primal-dual extrapolation method for solving a strongly MI problem, namely, problem (1) with $\mu > 0$. Our method is a variant of the classical forward-backward splitting (FBS) method [7, 15]. It modifies the forward term in (3) by using a *point and operator extrapolation technique*[4] that has recently been proposed to design algorithms for solving stochastic VI problems in [2]. Note that the choice of the parameters for extrapolations in [2] requires Lipschitz continuity of F. Since F is only assumed to be locally Lipschitz continuous in this paper, the choice of them in [2] is not applicable to our method. To resolve

[4]In the context of optimization, the operator F is typically the gradient of a function and $F(x)$ can be viewed as a point in the dual space. As a result, $\{x^t\}$ and $\{F(x^t)\}$ generated by this method can be respectively viewed as a primal and dual sequence and thus the extrapolations on them are alternatively called primal and dual extrapolations just for simplicity. Accordingly, we refer to our method as a *primal-dual extrapolation method.*

this issue, we propose an *adaptive scheme* to decide on parameters for extrapolations and splitting by using a backtracking technique. In addition, we propose a *verifiable termination criterion*, which guarantees that our method *outputs an ϵ-residual solution* of problem (1) with $\mu > 0$ for any given tolerance ϵ. The proposed method is presented in Algorithm 1 below.

Algorithm 1 A primal-dual extrapolation method for problem (1) with $\mu > 0$

Input: $\epsilon > 0$, $\gamma_0 > 0$, $\delta \in (0,1)$, $0 \le \xi < \nu \le 1/2$, $\{\kappa_t\} \subset [0, \xi/(1+\xi)]$, and $x^0 = x^1 \in \mathrm{dom}(B)$.

1: **for** $t = 1, 2, \ldots$ **do**

2: Compute

$$x^{t+1} = (I + \gamma_t B)^{-1} \left(x^t + \alpha_t (x^t - x^{t-1}) - \gamma_t (F(x^t) + \beta_t (F(x^t) - F(x^{t-1}))) \right), \qquad (4)$$

where

$$\gamma_t = \gamma_0 \delta^{n_t}, \ \beta_t = \frac{\gamma_{t-1}(1-\kappa_t)}{\gamma_t(1-\kappa_{t-1})} \left(1 + \frac{2\mu\gamma_{t-1}}{1-\kappa_{t-1}} \right)^{-1}, \ \alpha_t = \frac{\kappa_{t-1}\gamma_t\beta_t}{\gamma_{t-1}}, \qquad (5)$$

and n_t is the smallest nonnegative integer such that

$$\nu(1-\kappa_t)\|x^{t+1} - x^t\| \ge \gamma_t \|F(x^{t+1}) - F(x^t) - \kappa_t \gamma_t^{-1}(x^{t+1} - x^t)\|. \qquad (6)$$

3: Terminate the algorithm and output x^{t+1} if

$$\|x^t - x^{t+1} + \alpha_t(x^t - x^{t-1}) + \gamma_t(F(x^{t+1}) - F(x^t)) - \gamma_t\beta_t(F(x^t) - F(x^{t-1}))\| \le \gamma_t \epsilon. \qquad (7)$$

4: **end for**

Remark 1. *(i) Algorithm 1 is almost parameter-free except that the monotonicity parameter μ of $F + B$ is required.*

(ii) As will be established below, Algorithm 1 is well-defined at each iteration. One can observe that the fundamental operations of Algorithm 1 consist only of evaluations of F and resolvent of B. Specifically, at iteration t, Algorithm 1 requires $n_t + 1$ evaluations of F and resolvent of B for finding x^{t+1} satisfying (6)*.*

We next establish that Algorithm 1 *well-defined* and *outputs an ϵ-residual solution* of problem (1). We also study its complexity including: (i) *iteration complexity* measured by the number of iterations; (ii) *operation complexity* measured by the number of evaluations of F and resolvent of B.

To proceed, we assume throughout this section that problem (1) is a strongly MI problem (namely, $\mu > 0$) and that x^* is the solution of (1). Let $\{x^t\}_{t \in \mathbb{T}}$ denote all the iterates generated by Algorithm 1, where $\mathbb{T}$ is a subset of consecutive nonnegative integers starting from 0. We also define

$$\mathbb{T} - 1 = \{t - 1 : t \in \mathbb{T}\}, \quad r_0 = \|x^0 - x^*\|, \quad \mathcal{S} = \left\{ x \in \mathrm{dom}(B) : \ \|x - x^*\| \le \frac{r_0}{\sqrt{1 - 2\nu^2}} \right\}, \ (8)$$

where x^0 is the initial point and $\nu \in (0, 1/2]$ is an input parameter of Algorithm 1.

The following lemma establishes that F is *Lipschitz* continuous on $\mathcal{S}$ and also on an enlarged set induced by γ_0, r_0, ν, x^*, F and $\mathcal{S}$, albeit F is *locally Lipschitz* continuous on $\mathrm{cl}(\mathrm{dom}(B))$. This result will play an important role in this section.

Lemma 1. *Let $\mathcal{S}$ be defined in* (8)*. Then the following statements hold.*

(i) F is $L_{\mathcal{S}}$-Lipschitz continuous on $\mathcal{S}$ for some constant $L_{\mathcal{S}} > 0$.

(ii) F is $L_{\widehat{\mathcal{S}}}$-Lipschitz continuous on $\widehat{\mathcal{S}}$ for some constant $L_{\widehat{\mathcal{S}}} > 0$, where

$$\widehat{\mathcal{S}} = \left\{ x \in \mathrm{dom}(B) : \ \|x - x^*\| \leq \frac{(2 + 4\gamma_0 L_{\mathcal{S}}) r_0}{\sqrt{1 - 2\nu^2}} \right\}, \tag{9}$$

r_0 is defined in (8), and $\gamma_0 > 0$ and $\nu \in (0, 1/2]$ are the input parameters of Algorithm 1.

Proof. Notice that $\mathcal{S}$ is a convex and bounded subset in $\mathrm{dom}(B)$. By this and the local Lipschitz continuity of F on $\mathrm{cl}(\mathrm{dom}(B))$, it is not hard to observe that there exists some constant $L_{\mathcal{S}} > 0$ such that F is $L_{\mathcal{S}}$-Lipschitz continuous on $\mathcal{S}$. Hence, statement (i) holds and moreover the set $\widehat{\mathcal{S}}$ is well-defined. By a similar argument, one can see that statement (ii) also holds. □

The following theorem shows that Algorithm 1 is well-defined at each iteration. Its proof is deferred to Section 5.

Theorem 1. *Let $\{x^t\}_{t \in \mathbb{T}}$ be generated by Algorithm 1. Then the following statements hold.*

(i) Algorithm 1 is well-defined at each iteration.

(ii) $x^t \in \mathcal{S}$ for all $t \in \mathbb{T}$, and moreover, $n_t \leq N_t$ for all $1 \leq t \in \mathbb{T} - 1$, where $\mathcal{S}$ is defined in (8) and

$$N_t = \left\lceil \log\left(\frac{\nu(1 - \kappa_t) - \kappa_t}{\gamma_0 L_{\widehat{\mathcal{S}}}} \right) / \log \delta \right\rceil_+. \tag{10}$$

The next theorem presents iteration and operation complexity of Algorithm 1 for finding an ϵ-residual solution of problem (1) with $\mu > 0$, whose proof is deferred to Section 5.

Theorem 2. *Suppose that $\mu > 0$, i.e., $F + B$ is strongly monotone on $\mathrm{dom}(B)$. Then Algorithm 1 terminates and outputs an ϵ-residual solution of problem (1) in at most T iterations. Moreover, the total number of evaluations of F and resolvent of B performed in Algorithm 1 is no more than N, respectively, where*

$$T = 3 + \left\lceil 2 \log\left(\frac{r_0 \left(3 + 5\gamma_0 L_{\mathcal{S}}\right)}{\epsilon\sqrt{1 - 2\nu^2} \min\left\{ \frac{2\delta(\nu - \xi)}{3L_{\widehat{\mathcal{S}}}}, \gamma_0 \right\}} \right) \Bigg/ \log\left(\min\left\{ \frac{1}{L_{\widehat{\mathcal{S}}}} \left(L_{\widehat{\mathcal{S}}} + 2\mu\delta\nu - 2\mu\delta\xi \right), 1 + 2\mu\gamma_0 \right\} \right) \right\rceil_+, \tag{11}$$

$$N = \left(3 + \left\lceil \frac{2 \log \frac{r_0 (3 + 5\gamma_0 L_{\mathcal{S}})}{\epsilon\sqrt{1 - 2\nu^2} \min\left\{ \frac{2\delta(\nu - \xi)}{3L_{\widehat{\mathcal{S}}}}, \gamma_0 \right\}}}{\log\left(\min\left\{ \frac{1}{L_{\widehat{\mathcal{S}}}} \left(L_{\widehat{\mathcal{S}}} + 2\mu\delta\nu - 2\mu\delta\xi \right), 1 + 2\mu\gamma_0 \right\} \right)} \right\rceil_+ \right) \left(1 + \left\lceil \frac{\log\left(\frac{2(\nu - \xi)}{3\gamma_0 L_{\widehat{\mathcal{S}}}} \right)}{\log \delta} \right\rceil_+ \right). \tag{12}$$

Remark 2. *It can be seen from Theorem 2 that Algorithm 1 enjoys an iteration and operation complexity of $\mathcal{O}\left(\log \epsilon^{-1}\right)$ for finding an ϵ-residual solution of problem (1) with $\mu > 0$ under the assumption that F is locally Lipschitz continuous on $\mathrm{cl}(\mathrm{dom}(B))$.*

3 A primal-dual extrapolation method for problem (1) with $\mu = 0$

In this section we propose a primal-dual extrapolation method for solving a non-strongly MI problem, namely, problem (1) with $\mu = 0$. Our method consists of applying Algorithm 1 to approximately solving a sequence of strongly MI problems $0 \in (F_k + B)(x)$, where F_k is a perturbation of F given in (13). The proposed method is presented in Algorithm 2.

Algorithm 2 A primal-dual extrapolation method for problem (1) with $\mu = 0$

Input: $\varepsilon > 0$, $\gamma_0 > 0$, $z^0 \in \mathrm{dom}(B)$, $0 < \delta < 1$, $0 \le \xi < \nu \le 1/2$, $\{\kappa_t\} \subset [0, \xi/(1+\xi)]$, $\rho_0 \ge 1$, $0 < \tau_0 \le 1$, $\zeta > 1$, $0 < \sigma < 1/\zeta$, $\rho_k = \rho_0 \zeta^k$, $\tau_k = \tau_0 \sigma^k$ for all $k \ge 0$.

1: **for** $k = 0, 1, \ldots$ **do**

2: Call Algorithm 1 with $F \leftarrow F_k$, $\mu \leftarrow \rho_k^{-1}$, $\epsilon \leftarrow \tau_k$, $x^0 = x^1 \leftarrow z^k$ and the parameters γ_0, $\{\kappa_t\}$, δ and ν, and output z^{k+1}, where

$$F_k(x) = F(x) + \frac{1}{\rho_k}(x - z^k) \quad \forall x \in \mathrm{dom}(F). \tag{13}$$

3: Terminate this algorithm and output z^{k+1} if

$$\frac{1}{\rho_k}\|z^{k+1} - z^k\| \le \frac{\varepsilon}{2}, \ \tau_k \le \frac{\varepsilon}{2}. \tag{14}$$

4: **end for**

Remark 3. *It is easy to see that Algorithm 2 is well-defined at each iteration and equipped with a verifiable termination criterion. In addition, it is parameter-free and shares the same fundamental operations as Algorithm 1, which consist only of evaluations of F and resolvent of B.*

We next show that Algorithm 2 *outputs an ε-residual solution* of problem (1). We also study its complexity including: (i) *iteration complexity* measured by the number of iterations; (ii) *operation complexity* measured by the total number of evaluations of F and resolvent of B.

To proceed, we assume that x^* is an arbitrary solution of problem (1) and fixed throughout this section. Let $\{z^k\}_{k\in\mathbb{K}}$ denote all the iterates generated by Algorithm 2, where $\mathbb{K}$ is a subset of consecutive nonnegative integers starting from 0. We also define $\mathbb{K} - 1 = \{k - 1 : k \in \mathbb{K}\}$, and

$$\bar{r}_0 = \|z^0 - x^*\|, \quad \mathcal{Q} = \left\{x \in \mathrm{dom}(B) : \ \|x - x^*\| \le \left(\frac{1}{\sqrt{1-2\nu^2}} + 1\right)\left(\bar{r}_0 + \frac{\rho_0\tau_0}{1-\sigma\zeta}\right)\right\}, \tag{15}$$

where z^0 is the initial point and ρ_0, τ_0, ν, ζ, σ are the input parameters of Algorithm 2.

The following lemma establishes that F_k is Lipschitz continuous on $\mathcal{Q}$ and also on an enlarged set induced by F_k and $\mathcal{Q}$ with a Lipschitz constant independent on k. This result will play an important role in this section.

Lemma 2. *Let F_k and $\mathcal{Q}$ be defined in* (13) *and* (15)*. Then the following statements hold.*

(i) F_k is $L_{\mathcal{Q}}$-Lipschitz continuous on $\mathcal{Q}$ for some constant $L_{\mathcal{Q}} > 0$ independent on k.

(ii) F_k is $L_{\widehat{\mathcal{Q}}}$-Lipschitz continuous on $\widehat{\mathcal{Q}}$ for some constant $L_{\widehat{\mathcal{Q}}} > 0$ independent on k, where

$$\widehat{\mathcal{Q}} = \left\{x \in \mathrm{dom}(B) : \ \|x - x^*\| \le \left(\frac{2 + 4\gamma_0 L_{\mathcal{Q}}}{\sqrt{1-2\nu^2}} + 1\right)\left(\bar{r}_0 + \frac{\rho_0\tau_0}{1-\sigma\zeta}\right)\right\}. \tag{16}$$

Proof. Notice that $\mathcal{Q}$ is a convex and bounded subset in $\mathrm{dom}(B)$. By this and the local Lipschitz continuity of F on $\mathrm{cl}(\mathrm{dom}(B))$, it is not hard to observe that there exists some constant $\tilde{L}_{\mathcal{Q}} > 0$ such that F is $\tilde{L}_{\mathcal{Q}}$-Lipschitz continuous on $\mathcal{Q}$. In addition, notice from Algorithm 2 that $\rho_k \ge \rho_0$ for all $k \ge 0$. Using these and (13), we can easily see that F_k is $L_{\mathcal{Q}}$-Lipschitz continuous on $\mathcal{Q}$ with $L_{\mathcal{Q}} = \tilde{L}_{\mathcal{Q}} + 1/\rho_0$. Hence, statement (i) holds and moreover the set $\widehat{\mathcal{Q}}$ is well-defined. By a similar argument, one can see that statement (ii) also holds. □

The next theorem presents iteration and operation complexity of Algorithm 2 for finding an ε-residual solution of problem (1) with $\mu = 0$, whose proof is deferred to Section 5.

Theorem 3. *Let*

$$\Lambda=\frac{\rho_0\tau_0}{1-\sigma\zeta},\quad C_1=\log\left(\frac{(\bar{r}_0+\Lambda)\,(3+5\gamma_0 L_{\mathcal{Q}})}{\tau_0\sqrt{1-2\nu^2}\min\left\{\frac{2\delta(\nu-\xi)}{3L_{\widehat{\mathcal{Q}}}},\gamma_0\right\}}\right), \tag{17}$$

$$C_2=1+\left\lceil\log\left(\frac{2(\nu-\xi)}{3\gamma_0 L_{\widehat{\mathcal{Q}}}}\right)/\log\delta\right\rceil_+,\quad C_3=\frac{1}{(\zeta-1)\log\left(\min\left\{1+\frac{2\delta\nu-2\delta\xi}{\rho_0 L_{\widehat{\mathcal{Q}}}},1+\frac{2\gamma_0}{\rho_0}\right\}\right)}. \tag{18}$$

Suppose that $\mu=0$*, i.e.,* $F+B$ *is monotone but not strongly monotone on* $\mathrm{dom}(B)$*. Then Algorithm 2 terminates and outputs an* ε*-residual solution in at most* $K+1$ *iterations. Moreover, the total number of evaluations of* F *and resolvent of* B *performed in Algorithm 2 is no more than* M*, respectively, where*

$$K=\left\lceil\max\left\{\log_\zeta\left(\frac{2\bar{r}_0+2\Lambda}{\varepsilon\rho_0}\right),\frac{\log(2\tau_0/\varepsilon)}{\log(1/\sigma)}\right\}\right\rceil_+, \tag{19}$$

and

$$\begin{aligned}M=\;&4C_2+4C_2\left\lceil\max\left\{\log_\zeta\left(\frac{2\bar{r}_0+2\Lambda}{\varepsilon\rho_0}\right),\frac{\log(2\tau_0/\varepsilon)}{\log(1/\sigma)}\right\}\right\rceil_+\\&+2\zeta(C_1)_+C_2C_3\max\left\{\frac{2\zeta(\bar{r}_0+\Lambda)}{\varepsilon\rho_0},\zeta\left(\frac{2\tau_0}{\varepsilon}\right)^{\frac{\log\zeta}{\log(1/\sigma)}},1\right\}\\&-2\zeta C_2C_3(\log\sigma)\max\left\{\frac{2\zeta(\bar{r}_0+\Lambda)}{\varepsilon\rho_0},\zeta\left(\frac{2\tau_0}{\varepsilon}\right)^{\frac{\log\zeta}{\log(1/\sigma)}},1\right\}\\&\times\left\lceil\max\left\{\log_\zeta\left(\frac{2\bar{r}_0+2\Lambda}{\varepsilon\rho_0}\right),\frac{\log(2\tau_0/\varepsilon)}{\log(1/\sigma)}\right\}\right\rceil_+.\end{aligned} \tag{20}$$

Remark 4. *Since* $1<\zeta<1/\sigma$*, it can be seen from Theorem 3 that Algorithm 2 enjoys an iteration complexity of* $\mathcal{O}\left(\log\varepsilon^{-1}\right)$ *and an operation complexity of* $\mathcal{O}\left(\varepsilon^{-1}\log\varepsilon^{-1}\right)$ *for finding an* ε*-residual solution of problem* (1) *with* $\mu=0$ *under the assumption that* F *is locally Lipschitz continuous on* $\mathrm{cl}(\mathrm{dom}(B))$*. These complexity results are entirely new. Even for the case where* F *is globally Lipschitz continuous, our operation complexity of* $\mathcal{O}\left(\varepsilon^{-1}\log\varepsilon^{-1}\right)$ *is significantly better than the previously best-known operation complexity of* $\mathcal{O}\left(\varepsilon^{-2}\right)$ *obtained in [11, Theorem 4.6] for finding an* ε*-residual solution of a non-strongly MI problem, namely, problem (1) with* $\mu=0$*.*

4 Applications

In this section we study applications of our PDE method, particularly Algorithm 2, for solving several important classes of problems, particularly, convex conic optimization, conic constrained saddle point, and variational inequality problems. As a consequence, complexity results are obtained for finding an ε-KKT or ε-residual solution of these problems under local Lipschitz continuity for the first time.

4.1 Convex conic optimization

In this subsection we consider convex conic optimization

$$\begin{array}{rl}\min & f(x)+P(x)\\ \text{s.t.} & -g(x)\in\mathcal{K},\end{array} \tag{21}$$

where $f, P : \mathbb{R}^n \to (-\infty, \infty]$ are proper closed convex functions, $\mathcal{K}$ is a closed convex cone in $\mathbb{R}^m$, and the mapping $g : \mathbb{R}^n \to \mathbb{R}^m$ is $\mathcal{K}$-convex, that is,

$$\vartheta g(x) + (1-\vartheta)g(y) - g(\vartheta x + (1-\vartheta)y) \in \mathcal{K} \qquad \forall x, y \in \mathbb{R}^n,\ \vartheta \in [0,1]. \tag{22}$$

It shall be mentioned that $\mathrm{dom}(P)$ is possibly *unbounded*.

Problem (21) includes a rich class of problems as special cases. For example, when $\mathcal{K} = \mathbb{R}^{m_1}_+ \times \{0\}^{m_2}$ for some m_1 and m_2, $g(x) = (g_1(x), \ldots, g_{m_1}(x), h_1(x), \ldots, h_{m_2}(x))^T$ with convex g_i's and affine h_j's, and $P(x)$ is the indicator function of a simple convex set $\mathcal{X} \subseteq \mathbb{R}^n$, problem (21) reduces to an ordinary convex optimization problem

$$\min_{x \in \mathcal{X}} \{f(x) : g_i(x) \le 0,\ i = 1, \ldots, m_1; h_j(x) = 0,\ j = 1, \ldots, m_2\}.$$

We make the following additional assumptions for problem (21).

Assumption 2. *(a) The proximal operator associated with P and also the projection onto $\mathcal{K}^*$ can be exactly evaluated.*

(b) The function f and the mapping g are differentiable on $\mathrm{cl}(\mathrm{dom}(\partial P))$. Moreover, ∇f and ∇g are locally Lipschitz continuous on $\mathrm{cl}(\mathrm{dom}(\partial P))$.

(c) Both problem (21) and its Lagrangian dual problem

$$\sup_{\lambda \in \mathcal{K}^*} \inf_x \{f(x) + P(x) + \langle \lambda, g(x) \rangle\} \tag{23}$$

have optimal solutions, and moreover, they share the same optimal value.

Under the above assumptions, it can be shown that (x, λ) is a pair of optimal solutions of (21) and (23) if and only if it satisfies the Karush-Kuhn-Tucker (KKT) condition

$$0 \in \begin{pmatrix} \nabla f(x) + \nabla g(x)\lambda + \partial P(x) \\ -g(x) + \mathcal{N}_{\mathcal{K}^*}(\lambda) \end{pmatrix}. \tag{24}$$

In general, it is difficult to find an exact optimal solution of (21) and (23). Instead, for any given $\varepsilon > 0$, we are interested in finding a pair of ε-KKT solutions (x, λ) of (21) and (23) that satisfies

$$\mathrm{dist}(0, \nabla f(x) + \nabla g(x)\lambda + \partial P(x)) \le \varepsilon, \quad \mathrm{dist}(0, -g(x) + \mathcal{N}_{\mathcal{K}^*}(\lambda)) \le \varepsilon. \tag{25}$$

Observe from (24) that problems (21) and (23) can be solved as the MI problem

$$0 \in F(x, \lambda) + B(x, \lambda), \tag{26}$$

where

$$F(x, \lambda) = \begin{pmatrix} \nabla f(x) + \nabla g(x)\lambda \\ -g(x) \end{pmatrix}, \quad B(x, \lambda) = \begin{pmatrix} \partial P(x) \\ \mathcal{N}_{\mathcal{K}^*}(\lambda) \end{pmatrix}. \tag{27}$$

One can also observe that F is monotone and locally Lipschitz continuous on $\mathrm{cl}(\mathrm{dom}(B))$ and B is maximal monotone. As a result, Algorithm 2 can be suitably applied to the MI problem (26). It then follows from Theorem 3 that Algorithm 2, when applied to problem (26), finds an ε-residual solution (x, λ) of (26) within $\mathcal{O}\left(\varepsilon^{-1} \log \varepsilon^{-1}\right)$ evaluations of F and resolvent of B. Notice from (25) and (27) that such (x, λ) is also a pair of ε-KKT solutions of (21) and (23). In addition, the evaluation of F requires that of ∇f and ∇g, and also the resolvent of B can be computed as

$$(I + \gamma B)^{-1} \begin{pmatrix} x \\ \lambda \end{pmatrix} = \begin{pmatrix} \mathrm{prox}_{\gamma P}(x) \\ \Pi_{\mathcal{K}^*}(\lambda) \end{pmatrix} \qquad \forall (x, \lambda) \in \mathbb{R}^n \times \mathbb{R}^m, \gamma > 0.$$

The above discussion leads to the following result regarding Algorithm 2 for finding a pair of ε-KKT solutions of problems (21) and (23).

Theorem 4. *For any $\varepsilon > 0$, Algorithm 2, when applied to the MI problem* (26)*, outputs a pair of ε-KKT solutions of problems (21) and (23) within $\mathcal{O}\left(\varepsilon^{-1}\log\varepsilon^{-1}\right)$ evaluations of ∇f, ∇g,* $\mathrm{prox}_{\gamma P}$ *and $\Pi_{\mathcal{K}^*}$ for some $\gamma > 0$.*

Remark 5. *(i) This is the first time to propose an algorithm for finding an ε-KKT solution of problem (21) without the usual assumption that ∇f and ∇g are Lipschitz continuous and/or the domain of P is bounded. Moreover, the proposed algorithm is equipped with a verifiable termination criterion and enjoys an operation complexity of $\mathcal{O}\left(\varepsilon^{-1}\log\varepsilon^{-1}\right)$.*

(ii) A first-order augmented Lagrangian method was recently proposed in [8] for finding a pair of ε-KKT solutions of a subclass of problems (21) and (23), which also requires $\mathcal{O}\left(\varepsilon^{-1}\log\varepsilon^{-1}\right)$ evaluations of ∇f, ∇g, $\mathrm{prox}_{\gamma P}$ *and $\Pi_{\mathcal{K}^*}$. However, this method and its complexity analysis require that ∇f and ∇g be Lipschitz continuous on an open set containing* $\mathrm{dom}(P)$ *and also that* $\mathrm{dom}(P)$ *be bounded. As a result, it is generally not applicable to problem (21).*

(iii) A variant of Tseng's MFBS method was proposed in [12, Section 6] for finding a pair of ε-KKT solutions of a special class of problems (21) and (23), in which g is an affine mapping , $\mathcal{K} = \{0\}^m$, and ∇f is Lipschitz continuous on $\mathrm{cl}(\mathrm{dom}(P))$*. Due to the latter assumption, this method is generally not applicable to problem (21). In addition, this method enjoys an operation complexity of $\mathcal{O}\left(\varepsilon^{-2}\right)$ (see [12, Theorem 6.3]). In contrast, the operation complexity of $\mathcal{O}\left(\varepsilon^{-1}\log\varepsilon^{-1}\right)$ achieved by our method is significantly better.*

4.2 Conic constrained saddle point problems

In this subsection we consider the following conic constrained saddle point (CCSP) problem:

$$\min_{-g(x)\in\mathcal{K}} \max_{-\tilde{g}(y)\in\tilde{\mathcal{K}}} \{\Psi(x,y) := f(x,y) + P(x) - \tilde{P}(y)\}, \tag{28}$$

where $f : \mathbb{R}^n \times \mathbb{R}^m \to [-\infty,\infty]$ is convex in x and concave in y, $P : \mathbb{R}^n \to (-\infty,\infty]$ and $\tilde{P} : \mathbb{R}^m \to (-\infty,\infty]$ are proper closed convex functions, $\mathcal{K} \subseteq \mathbb{R}^p$ and $\tilde{\mathcal{K}} \subseteq \mathbb{R}^{\tilde{p}}$ are closed convex cones, and g and $\tilde{g}$ are $\mathcal{K}$- and $\tilde{\mathcal{K}}$-convex in the sense of (22), respectively. It shall be mentioned that $\mathrm{dom}(P)$ and $\mathrm{dom}(\tilde{P})$ are possibly *unbounded*.

We make the following additional assumptions for problem (28).

Assumption 3. *(a) The proximal operator associated with P and $\tilde{P}$ and also the projection onto $\mathcal{K}^*$ and $\tilde{\mathcal{K}}^*$ can be exactly evaluated.*

(b) The function f is differentiable on $\mathrm{cl}(\mathrm{dom}(\partial P)) \times \mathrm{cl}(\mathrm{dom}(\partial\tilde{P}))$*. Moreover, ∇f is locally Lipschitz continuous on* $\mathrm{cl}(\mathrm{dom}(\partial P)) \times \mathrm{cl}(\mathrm{dom}(\partial\tilde{P}))$*.*

(c) The mapping s g and $\tilde{g}$ are respectively differentiable on $\mathrm{cl}(\mathrm{dom}(\partial P))$ *and* $\mathrm{cl}(\mathrm{dom}(\partial\tilde{P}))$*. Moreover, ∇g and $\nabla\tilde{g}$ are locally Lipschitz continuous on* $\mathrm{cl}(\mathrm{dom}(\partial P))$ *and* $\mathrm{cl}(\mathrm{dom}(\partial\tilde{P}))$*, respectively.*

(d) There exists a pair $(x^*, y^*) \in \mathrm{dom}(P) \times \mathrm{dom}(\tilde{P})$ satisfying $-g(x^*) \in \mathcal{K}$ and $-\tilde{g}(y^*) \in \tilde{\mathcal{K}}$ such that

$$\Psi(x^*, y) \le \Psi(x^*, y^*) \le \Psi(x, y^*)$$

holds for any $(x, y) \in \mathrm{dom}(P) \times \mathrm{dom}(\tilde{P})$ satisfying $-g(x) \in \mathcal{K}$ and $-\tilde{g}(y) \in \tilde{\mathcal{K}}$.

Problem (28) includes a rich class of saddle point problems as special cases. Several of them have been studied in the literature. For example, extragradient method [4], mirror-prox method [13], dual extrapolation method [14], and accelerated proximal point method [6] were developed for solving the special CCSP problem

$$\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}} \tilde{f}(x,y), \tag{29}$$

where $\tilde{f}$ is convex in x and concave in y with *Lipschitz continuous* gradient on $\mathcal{X}\times\mathcal{Y}$, and $\mathcal{X}$ and $\mathcal{Y}$ are simple convex sets. Also, optimistic gradient method [10] and extra anchored gradient method [23] were proposed for solving problem (29) with $\mathcal{X}=\mathbb{R}^n$ and $\mathcal{Y}=\mathbb{R}^m$. In addition, accelerated proximal gradient method [20], a variant of MFBS method [12], and also generalized extragradient method [12] were proposed for solving the special CCSP problem

$$\min_{x}\max_{y}\left\{\tilde{f}(x,y)+P(x)-\tilde{P}(y)\right\}, \tag{30}$$

where $\tilde{f}$ is convex in x and concave in y with *Lipschitz* continuous gradient on $\mathrm{dom}(P)\times\mathrm{dom}(\tilde{P})$. Besides, several optimal or nearly optimal first-order methods were developed for solving problem (29) or (30) with a strongly-convex-(strongly)-concave $\tilde{f}$ (e.g., see [6, 22, 21]). Recently, extra-gradient method of multipliers [24] was proposed for solving a special case of problem (28) with g and $\tilde{g}$ being an affine mapping , $\mathrm{dom}(P)$ and $\mathrm{dom}(\tilde{P})$ being compact, and ∇f being Lipschitz continuous on $\mathrm{dom}(P)\times\mathrm{dom}(\tilde{P})$. Iteration complexity of these methods except [23] was established based on the duality gap on the ergodic (i.e., weight-averaged) solution sequence. Yet, the duality gap can often be difficult to measure. In practice, one may use a computable upper bound on the duality gap to terminate these methods, which however typically requires the knowledge of an upper bound on the distance between the initial point and the solution set. Besides, there is a lack of complexity guarantees for these methods in terms of the original solution sequence.

Due to the sophistication of the constraints $-g(x)\in\mathcal{K}$ and $-\tilde{g}(y)\in\tilde{\mathcal{K}}$ and also the local Lipschitz continuity of ∇f, ∇g and $\nabla\tilde{g}$, the aforementioned methods [4, 14, 13, 20, 12, 10, 6, 22, 21, 23] are generally not suitable for solving the CCSP problem (28). We next apply our Algorithm 2 to find an ε-KKT solution of (28) and also study its operation complexity for finding such an approximate solution under the local Lipschitz continuity of ∇f, ∇g and $\nabla\tilde{g}$.

Under the above assumptions, it can be shown that (x,y) is a pair of optimal minmax solutions of problem (28) if and only if it together with some $(\lambda,\tilde{\lambda})$ satisfies the KKT condition

$$0\in\begin{pmatrix}\nabla_x f(x,y)+\nabla g(x)\lambda+\partial P(x)\\ -\nabla_y f(x,y)+\nabla\tilde{g}(y)\tilde{\lambda}+\partial\tilde{P}(y)\\ -g(x)+\mathcal{N}_{\mathcal{K}^*}(\lambda)\\ -\tilde{g}(y)+\mathcal{N}_{\tilde{\mathcal{K}}^*}(\tilde{\lambda})\end{pmatrix}. \tag{31}$$

Generally, it is difficult to find a pair of exact optimal minimax solutions of (28). Instead, for any given $\varepsilon>0$, we are interested in finding an ε-KKT solution $(x,y,\lambda,\tilde{\lambda})$ of (28) that satisfies

$$\mathrm{dist}(0,\nabla_x f(x,y)+\nabla g(x)\lambda+\partial P(x))\le\varepsilon,\quad \mathrm{dist}(0,-\nabla_y f(x,y)+\nabla\tilde{g}(y)\tilde{\lambda}+\partial\tilde{P}(y))\le\varepsilon, \tag{32}$$

$$\mathrm{dist}(0,-g(x)+\mathcal{N}_{\mathcal{K}^*}(\lambda))\le\varepsilon,\quad \mathrm{dist}(0,-\tilde{g}(y)+\mathcal{N}_{\tilde{\mathcal{K}}^*}(\tilde{\lambda}))\le\varepsilon. \tag{33}$$

Observe from (31) that problem (28) can be solved as the MI problem

$$0\in F(x,y,\lambda,\tilde{\lambda})+B(x,y,\lambda,\tilde{\lambda}), \tag{34}$$

where

$$F(x,y,\lambda,\tilde{\lambda})=\begin{pmatrix}\nabla_x f(x,y)+\nabla g(x)\lambda\\ -\nabla_y f(x,y)+\nabla\tilde{g}(y)\tilde{\lambda}\\ -g(x)\\ -\tilde{g}(y)\end{pmatrix},\quad B(x,y,\lambda,\tilde{\lambda})=\begin{pmatrix}\partial P(x)\\ \partial\tilde{P}(y)\\ \mathcal{N}_{\mathcal{K}^*}(\lambda)\\ \mathcal{N}_{\tilde{\mathcal{K}}^*}(\tilde{\lambda})\end{pmatrix}. \tag{35}$$

One can also observe that F is monotone and locally Lipschitz continuous on $\mathrm{cl}(\mathrm{dom}(B))$ and B is maximal monotone. As a result, Algorithm 2 can be suitably applied to the MI problem (34). It then follows from Theorem 3 that Algorithm 2, when applied to problem (34), finds an ε-residual solution $(x, y, \lambda, \tilde{\lambda})$ of (34) within $\mathcal{O}\left(\varepsilon^{-1}\log\varepsilon^{-1}\right)$ evaluations of F and resolvent of B. Notice from (32), (33) and (35) that such $(x, y, \lambda, \tilde{\lambda})$ is also an ε-KKT solution of problem (28). In addition, the evaluation of F requires that of ∇f, ∇g and $\nabla \tilde{g}$, and also the resolvent of B can be computed as

$$(I+\gamma B)^{-1}\begin{pmatrix}x\\ y\\ \lambda\\ \tilde{\lambda}\end{pmatrix}=\begin{pmatrix}\mathrm{prox}_{\gamma P}(x)\\ \mathrm{prox}_{\gamma \tilde{P}}(y)\\ \Pi_{\mathcal{K}^*}(\lambda)\\ \Pi_{\tilde{\mathcal{K}}^*}(\tilde{\lambda})\end{pmatrix}\qquad \forall (x,y,\lambda,\tilde{\lambda})\in\mathbb{R}^n\times\mathbb{R}^m\times\mathbb{R}^p\times\mathbb{R}^{\tilde{p}}, \gamma>0.$$

The above discussion leads to the following result regarding Algorithm 2 for finding an ε-KKT solution of problem (28).

Theorem 5. *For any $\varepsilon > 0$, Algorithm 2, when applied to the MI problem* (34)*, outputs an ε-KKT solution of problem (28) within $\mathcal{O}\left(\varepsilon^{-1}\log\varepsilon^{-1}\right)$ evaluations of ∇f, ∇g, $\nabla \tilde{g}$, $\mathrm{prox}_{\gamma P}$, $\mathrm{prox}_{\gamma \tilde{P}}$, $\Pi_{\mathcal{K}^*}$ and $\Pi_{\tilde{\mathcal{K}}^*}$ for some $\gamma > 0$.*

Remark 6. *This is the first time to propose an algorithm for finding an ε-KKT solution of problem (28). Moreover, the proposed algorithm is equipped with a verifiable termination criterion and enjoys an operation complexity of $\mathcal{O}\left(\varepsilon^{-1}\log\varepsilon^{-1}\right)$ without the usual assumption that ∇f is Lipschitz continuous and/or the domains P and $\tilde{P}$ are bounded.*

4.3 Variational inequality

In this subsection we consider the following variational inequality (VI) problem:

$$\text{find } x\in\mathbb{R}^n \text{ such that } g(y)-g(x)+\langle y-x, F(x)\rangle \geq 0 \ \forall y\in\mathbb{R}^n, \tag{36}$$

where $g : \mathbb{R}^n \to (-\infty, \infty]$ is a proper closed convex function, and $F : \mathrm{dom}(F) \to \mathbb{R}^n$ is monotone and *locally Lipschitz continuous* on $\mathrm{cl}(\mathrm{dom}(\partial g)) \subseteq \mathrm{dom}(F)$. It shall be mentioned that $\mathrm{dom}(g)$ is possibly *unbounded*. Assume that problem (36) has at least one solution. For the details of VI and its applications, we refer the reader to [1] and the references therein.

Some special cases of (36) have been well studied in the literature. For example, projection method [17], extragradient method [4], mirror-prox method [13], dual extrapolation method [14], operator extrapolation method [5], extra-point method [2, 3], and extra-momentum method [2] were developed for solving problem (36) with g being the indicator function of a closed convex set and F being *Lipschitz continuous* on it or the entire space. In addition, a variant of Tseng's MFBS method [12], and generalized extragradient method [12] were proposed for solving problem (36) with F being *Lipschitz continuous*. Iteration complexity of these methods except [5] was established based on the weak gap or its variant on the ergodic (i.e., weight-averaged) solution sequence. Yet, the weak gap can often be difficult to measure. In practice, one may use a computable upper bound on the weak gap to terminate these methods, which however typically requires the knowledge of an upper bound on the distance between the initial point and the solution set. Besides, there is a lack of complexity guarantees for these methods in terms of the original solution sequence. In addition, since F is only assumed to be *locally Lipschitz continuous* on $\mathrm{cl}(\mathrm{dom}(g))$ in our paper, these methods are generally not suitable for solving problem (36).

Generally, it is difficult to find an exact solution of problem (36). Instead, for any given $\varepsilon > 0$, we are interested in finding an ε-residual solution of (36), which is a point x satisfying $\mathrm{res}_{F+\partial g}(x) \leq \varepsilon$. To this end, we first observe that problem (36) is equivalent to the MI problem

$$0 \in (F + \partial g)(x). \tag{37}$$

Since F is monotone and locally Lipschitz continuous on $\mathrm{cl}(\mathrm{dom}(\partial g))$ and ∂g is maximal monotone, Algorithm 2 can be suitably applied to the MI problem (37). It then follows from Theorem 3 that Algorithm 2, when applied to problem (37), finds an ε-residual solution x of (37), which is indeed also an ε-residual solution of (36), within $\mathcal{O}\left(\varepsilon^{-1}\log\varepsilon^{-1}\right)$ evaluations of F and resolvent of ∂g. Notice that the resolvent of ∂g can be computed as

$$(I+\gamma\partial g)^{-1}(x) = \mathrm{prox}_{\gamma g}(x), \quad \forall x\in\mathbb{R}^n, \gamma>0.$$

The above discussion leads to the following result regarding Algorithm 2 for finding an ε-residual solution of problem (36).

Theorem 6. *For any $\varepsilon>0$, Algorithm 2, when applied to the MI problem (37), outputs an ε-residual solution of problem (36) within $\mathcal{O}\left(\varepsilon^{-1}\log\varepsilon^{-1}\right)$ evaluations of F and $\mathrm{prox}_{\gamma g}$ for some $\gamma>0$.*

Remark 7. *(i) This is the first time to propose an algorithm for finding an ε-residual solution of problem (36) without the usual assumption that F is Lipschitz continuous and/or the domain g is bounded. Moreover, the proposed algorithm is equipped with a verifiable termination criterion and enjoys an operation complexity of $\mathcal{O}\left(\varepsilon^{-1}\log\varepsilon^{-1}\right)$.*

(ii) An operator extrapolation method [5] was recently proposed for finding an ε-residual solution of a special case of problem (36) with g being the indicator function of a closed convex set and F being Lipschitz continuous *on it. Due to the latter assumption, this method is generally not applicable to problem (21). In addition, this method enjoys an operation complexity of $\mathcal{O}\left(\varepsilon^{-2}\right)$ (see [5, Theorem 2.5]). In contrast, the operation complexity of $\mathcal{O}\left(\varepsilon^{-1}\log\varepsilon^{-1}\right)$ achieved by our method is significantly better.*

5 Proof of the main results

In this section we provide a proof of our main results presented in Sections 2 and 3, which are particularly Theorems 1, 2, and 3.

5.1 Proof of the main results in Section 2

In this subsection we first establish several technical lemmas and then use them to prove Theorems 1 and 2.

Before proceeding, we introduce some notation that will be used shortly. Recall from Section 2 that $\{x^t\}_{t\in\mathbb{T}}$ denotes all the iterates generated by Algorithm 1, where $\mathbb{T}$ is a subset of consecutive nonnegative integers starting from 0. For any $1\le t\in\mathbb{T}$, we define

$$\Delta^t = F(x^t) - F(x^{t-1}), \tag{38}$$

$$\tilde{\Delta}^t = \Delta^t - \kappa_{t-1}\gamma_{t-1}^{-1}(x^t - x^{t-1}). \tag{39}$$

In addition, we define

$$v^t = \frac{1}{\gamma_t}(x^t - x^{t+1} + \alpha_t(x^t - x^{t-1}) + \gamma_t\Delta^{t+1} - \gamma_t\beta_t\Delta^t) \qquad \forall 1\le t\in\mathbb{T}-1, \tag{40}$$

$$\tilde{\gamma}_t = \frac{\gamma_t}{1-\kappa_t}, \qquad \theta_t = \prod_{i=1}^{t-1}(1+2\mu\tilde{\gamma}_i) = \prod_{i=1}^{t-1}\left(1+\frac{2\mu\gamma_i}{1-\kappa_i}\right) \quad \forall 1\le t\in\mathbb{T}-1.^5 \tag{41}$$

The following lemma establishes some properties of $\{v^t\}_{1\le t\in\mathbb{T}-1}$.

[5]We set $\theta_1 = 1$.

Lemma 3. *Let $\{x^t\}_{t\in\mathbb{T}}$ be generated by Algorithm 1. Then for all $1 \le t \in \mathbb{T}-1$, the following relations hold.*

$$v^t \in (F+B)(x^{t+1}), \tag{42}$$

$$v^t = \frac{1}{\tilde{\gamma}_t}(x^t - x^{t+1} + \tilde{\gamma}_t\tilde{\Delta}^{t+1} - \tilde{\gamma}_t\beta_t\tilde{\Delta}^t). \tag{43}$$

Proof. By (4), one has

$$x^t + \alpha_t(x^t - x^{t-1}) - \gamma_t(F(x^t) + \beta_t(F(x^t) - F(x^{t-1}))) \in x^{t+1} + \gamma_t B(x^{t+1}).$$

Adding $\gamma_t F(x^{t+1})$ to both sides of this relation, we obtain

$$x^t + \alpha_t(x^t - x^{t-1}) + \gamma_t(F(x^{t+1}) - F(x^t)) - \gamma_t\beta_t(F(x^t) - F(x^{t-1})) \in x^{t+1} + \gamma_t(F+B)(x^{t+1}),$$

which together with (38) and (40) yields

$$v^t = \frac{1}{\gamma_t}(x^t - x^{t+1} + \alpha_t(x^t - x^{t-1}) + \gamma_t\Delta^{t+1} - \gamma_t\beta_t\Delta^t) \in (F+B)(x^{t+1}),$$

and hence (42) holds. In addition, recall from (5) that $\alpha_t = \kappa_{t-1}\gamma_t\beta_t/\gamma_{t-1}$. By this, (39) and (41), one has

$$\begin{aligned}
v^t &= \frac{1}{\gamma_t}(x^t - x^{t+1} + \alpha_t(x^t - x^{t-1}) + \gamma_t\Delta^{t+1} - \gamma_t\beta_t\Delta^t) \\
&= \frac{1}{\gamma_t}\left((1-\kappa_t)(x^t - x^{t+1}) + \gamma_t(\Delta^{t+1} - \frac{\kappa_t}{\gamma_t}(x^{t+1} - x^t)) - \gamma_t\beta_t(\Delta^t - \frac{\alpha_t}{\gamma_t\beta_t}(x^t - x^{t-1}))\right) \\
&= \frac{1}{\tilde{\gamma}_t}(x^t - x^{t+1} + \tilde{\gamma}_t\tilde{\Delta}^{t+1} - \tilde{\gamma}_t\beta_t\tilde{\Delta}^t).
\end{aligned}$$

Hence, (43) holds as desired. □

The next two lemmas establish some properties of $\{x^t\}_{t\in\mathbb{T}}$.

Lemma 4. *Let $\{x^t\}_{t\in\mathbb{T}}$ be generated by Algorithm 1. Then for all $1 \le k \in \mathbb{T}-1$, we have*

$$\frac{1}{2}\theta_1\|x^0 - x^*\|^2 - \frac{1}{2}(1+2\mu\tilde{\gamma}_k)\theta_k\|x^{k+1} - x^*\|^2 \ge -\tilde{\gamma}_k\theta_k\langle\tilde{\Delta}^{k+1}, x^{k+1} - x^*\rangle + R_k, \tag{44}$$

where

$$R_k = \sum_{t=1}^{k}\left(\tilde{\gamma}_t\beta_t\theta_t\langle\tilde{\Delta}^t, x^{t+1} - x^t\rangle + \frac{1}{2}\theta_t\|x^{t+1} - x^t\|^2\right). \tag{45}$$

Proof. By (2), (42) and $0 \in (F+B)(x^*)$, one has

$$\langle v^t, x^{t+1} - x^*\rangle \ge \mu\|x^{t+1} - x^*\|^2,$$

which along with (43) implies that

$$\begin{aligned}
\tilde{\gamma}_t\mu\|x^{t+1} - x^*\|^2 &\le \langle x^t - x^{t+1} + \tilde{\gamma}_t\tilde{\Delta}^{t+1} - \tilde{\gamma}_t\beta_t\tilde{\Delta}^t, x^{t+1} - x^*\rangle \\
&= \langle x^t - x^{t+1}, x^{t+1} - x^*\rangle + \tilde{\gamma}_t\langle\tilde{\Delta}^{t+1}, x^{t+1} - x^*\rangle - \tilde{\gamma}_t\beta_t\langle\tilde{\Delta}^t, x^{t+1} - x^*\rangle \\
&= \frac{1}{2}\left(\|x^t - x^*\|^2 - \|x^{t+1} - x^*\|^2 - \|x^t - x^{t+1}\|^2\right) + \tilde{\gamma}_t\langle\tilde{\Delta}^{t+1}, x^{t+1} - x^*\rangle \\
&\quad - \tilde{\gamma}_t\beta_t\langle\tilde{\Delta}^t, x^t - x^*\rangle - \tilde{\gamma}_t\beta_t\langle\tilde{\Delta}^t, x^{t+1} - x^t\rangle.
\end{aligned}$$

Rearranging the terms in the above inequality yields

$$\frac{1}{2}\|x^t - x^*\|^2 - \frac{1}{2}(1+2\tilde{\gamma}_t\mu)\|x^{t+1} - x^*\|^2 \ge \tilde{\gamma}_t\beta_t\langle\tilde{\Delta}^t, x^t - x^*\rangle - \tilde{\gamma}_t\langle\tilde{\Delta}^{t+1}, x^{t+1} - x^*\rangle \\ + \tilde{\gamma}_t\beta_t\langle\tilde{\Delta}^t, x^{t+1} - x^t\rangle + \frac{1}{2}\|x^{t+1} - x^t\|^2.$$

Multiplying both sides of this inequality by θ_t and summing it up for $t = 1, \dots, k$, we have

$$\begin{aligned}
&\sum_{t=1}^{k}\left(\frac{1}{2}\theta_t\|x^t - x^*\|^2 - \frac{1}{2}(1+2\tilde{\gamma}_t\mu)\theta_t\|x^{t+1} - x^*\|^2\right) \\
&\ge \sum_{t=1}^{k}\theta_t\tilde{\gamma}_t\beta_t\langle\tilde{\Delta}^t, x^t - x^*\rangle - \sum_{t=1}^{k}\theta_t\tilde{\gamma}_t\langle\tilde{\Delta}^{t+1}, x^{t+1} - x^*\rangle + R_k \\
&= \theta_1\tilde{\gamma}_1\beta_1\langle\tilde{\Delta}^1, x^1 - x^*\rangle + \sum_{t=1}^{k-1}(\theta_{t+1}\tilde{\gamma}_{t+1}\beta_{t+1} - \theta_t\tilde{\gamma}_t)\langle\tilde{\Delta}^{t+1}, x^{t+1} - x^*\rangle - \tilde{\gamma}_k\theta_k\langle\tilde{\Delta}^{k+1}, x^{k+1} - x^*\rangle + R_k.
\end{aligned} \tag{46}$$

In addition, it follows from $x^0 = x^1$ and (39) that $\tilde{\Delta}^1 = 0$. Also, by the definition of $\tilde{\gamma}_t$, θ_t and β_t in (5) and (41), one has

$$\begin{aligned}
\theta_{t+1}\tilde{\gamma}_{t+1}\beta_{t+1} - \theta_t\tilde{\gamma}_t &= \theta_t\left(1 + \frac{2\mu\gamma_t}{1-\kappa_t}\right)\cdot\frac{\gamma_{t+1}}{1-\kappa_{t+1}}\cdot\frac{\gamma_t(1-\kappa_{t+1})}{\gamma_{t+1}(1-\kappa_t)}\left(1 + \frac{2\mu\gamma_t}{1-\kappa_t}\right)^{-1} - \theta_t\frac{\gamma_t}{1-\kappa_t} = 0, \\
\theta_{t+1} &= (1+2\tilde{\gamma}_t\mu)\theta_t.
\end{aligned} \tag{47}$$

Using these, $\tilde{\Delta}^1 = 0$ and (46), we obtain

$$\sum_{t=1}^{k}\left(\frac{1}{2}\theta_t\|x^t - x^*\|^2 - \frac{1}{2}\theta_{t+1}\|x^{t+1} - x^*\|^2\right) \ge -\tilde{\gamma}_k\theta_k\langle\tilde{\Delta}^{k+1}, x^{k+1} - x^*\rangle + R_k,$$

which yields

$$\frac{1}{2}\theta_1\|x^0 - x^*\|^2 - \frac{1}{2}\theta_{k+1}\|x^{k+1} - x^*\|^2 \ge -\tilde{\gamma}_k\theta_k\langle\tilde{\Delta}^{k+1}, x^{k+1} - x^*\rangle + R_k.$$

The conclusion then follows from this and (47) with $t = k$. □

Lemma 5. *Let $\{x^t\}_{t\in\mathbb{T}}$ be generated by Algorithm 1. Then we have*

$$\|x^{k+1} - x^*\|^2 \le \frac{1}{(1-2\nu^2)\theta_k}\|x^0 - x^*\|^2 \qquad \forall 1 \le k \in \mathbb{T} - 1. \tag{48}$$

Proof. By the definition of β_t and $\tilde{\gamma}_t$ in (5) and (41), one has $\tilde{\gamma}_{t-1}^{-1}\tilde{\gamma}_t\beta_t = (1 + 2\mu\gamma_{t-1}/(1-\kappa_{t-1}))^{-1}$. Using this and the definition of θ_t in (41), we obtain

$$\begin{aligned}
\theta_{t-1} - 4\nu^2\tilde{\gamma}_{t-1}^{-2}\tilde{\gamma}_t^2\beta_t^2\theta_t &= \theta_{t-1}\left(1 - 4\nu^2\left(1 + \frac{2\mu\gamma_{t-1}}{1-\kappa_{t-1}}\right)^{-2}\frac{\theta_t}{\theta_{t-1}}\right) \\
&\overset{(41)}{=} \theta_{t-1}\left(1 - 4\nu^2\left(1 + \frac{2\mu\gamma_{t-1}}{1-\kappa_{t-1}}\right)^{-1}\right) \ge 0,
\end{aligned} \tag{49}$$

where the last inequality follows from the fact that $0 < \nu \le 1/2$. In addition, it follows from (6), (38), (39), and the definition of $\tilde{\gamma}_t$ in (41) that

$$\|\tilde{\Delta}^t\| \le \nu\tilde{\gamma}_{t-1}^{-1}\|x^t - x^{t-1}\| \qquad \forall 2 \le t \in \mathbb{T}. \tag{50}$$

Recall that R_k is defined in (45). Letting $\theta_0 = 0$, and using (45), (49), (50) and $x^0 = x^1$, we have

$$
\begin{aligned}
R_k &\overset{(45)}{\geq} \sum_{t=1}^{k} \left(-\tilde{\gamma}_t \beta_t \theta_t \|\tilde{\Delta}^t\| \|x^{t+1} - x^t\| + \frac{1}{2}\theta_t \|x^{t+1} - x^t\|^2 \right) \\
&\overset{(50)}{\geq} \sum_{t=1}^{k} \left(-\nu \tilde{\gamma}_{t-1}^{-1} \tilde{\gamma}_t \beta_t \theta_t \|x^t - x^{t-1}\| \|x^{t+1} - x^t\| + \frac{1}{2}\theta_t \|x^{t+1} - x^t\|^2 \right) \\
&= \sum_{t=1}^{k} \left(-\nu \tilde{\gamma}_{t-1}^{-1} \tilde{\gamma}_t \beta_t \theta_t \|x^t - x^{t-1}\| \|x^{t+1} - x^t\| + \frac{1}{4}\theta_t \|x^{t+1} - x^t\|^2 + \frac{1}{4}\theta_{t-1} \|x^t - x^{t-1}\| \right) \\
&\quad + \frac{1}{4}\theta_k \|x^{k+1} - x^k\|^2 \\
&\geq \sum_{t=1}^{k} \left(\left(\sqrt{\theta_t \theta_{t-1}}/2 - \nu \tilde{\gamma}_{t-1}^{-1} \tilde{\gamma}_t \beta_t \theta_t \right) \|x^t - x^{t-1}\| \|x^{t+1} - x^t\| \right) + \frac{1}{4}\theta_k \|x^{k+1} - x^k\|^2 \\
&\overset{(49)}{\geq} \frac{1}{4}\theta_k \|x^{k+1} - x^k\|^2.
\end{aligned}
$$

Using this, (44) and (50), we further obtain

$$
\begin{aligned}
\frac{1}{2}\theta_1 \|x^0 - x^*\|^2 - \frac{1}{2}(1 + 2\mu\tilde{\gamma}_k)\theta_k \|x^{k+1} - x^*\|^2 &\geq -\tilde{\gamma}_k \theta_k \langle \tilde{\Delta}^{k+1}, x^{k+1} - x^* \rangle + \frac{1}{4}\theta_k \|x^{k+1} - x^k\|^2 \\
&\geq -\tilde{\gamma}_k \theta_k \|\tilde{\Delta}^{k+1}\| \|x^{k+1} - x^*\| + \frac{1}{4}\theta_k \|x^{k+1} - x^k\|^2 \\
&\overset{(50)}{\geq} -\nu\theta_k \|x^{k+1} - x^k\| \|x^{k+1} - x^*\| + \frac{1}{4}\theta_k \|x^{k+1} - x^k\|^2 \\
&\geq -\nu^2 \theta_k \|x^{k+1} - x^*\|^2.
\end{aligned}
$$

It then follows from this, $\theta_1 = 1$, and $0 < \nu \leq 1/2$ that

$$
\|x^{k+1} - x^*\|^2 \leq \frac{\theta_1}{(1 + 2\mu\tilde{\gamma}_k - 2\nu^2)\theta_k} \|x^0 - x^*\|^2 \leq \frac{1}{(1 - 2\nu^2)\theta_k} \|x^0 - x^*\|^2.
$$

□

In what follows, we will show that $\{n_t\}_{1 \leq t \in \mathbb{T}-1}$ is bounded, that is, the number of evaluations of F and resolvent of B is bounded above by a constant for all iterations $t \in \mathbb{T} - 1$. To this end, we define

$$
x^{t+1}(\gamma) = (I + \gamma B)^{-1} \left(x^t + \alpha(\gamma)(x^t - x^{t-1}) - \gamma \left(F(x^t) + \beta(\gamma)(F(x^t) - F(x^{t-1})) \right) \right) \quad \forall \gamma > 0, \tag{51}
$$

$$
L_{t+1}(\gamma) = \|F(x^{t+1}(\gamma)) - F(x^t) - \kappa_t \gamma^{-1}(x^{t+1}(\gamma) - x^t)\| / \|x^{t+1}(\gamma) - x^t\| \quad \forall \gamma > 0, \tag{52}
$$

where

$$
\beta(\gamma) = \frac{\gamma_{t-1}(1 - \kappa_t)}{\gamma(1 - \kappa_{t-1})} \left(1 + \frac{2\mu\gamma_{t-1}}{1 - \kappa_{t-1}} \right)^{-1}, \quad \alpha(\gamma) = \frac{\kappa_{t-1}\gamma\beta(\gamma)}{\gamma_{t-1}}. \tag{53}
$$

The following lemma establishes some property of $x^{t+1}(\gamma)$, which will be used shortly.

Lemma 6. *Let $\mathcal{S}$ and $\widehat{\mathcal{S}}$ be defined in (8) and (9). Assume that $x^t, x^{t-1} \in \mathcal{S}$ for some $1 \leq t \in \mathbb{T} - 1$. Then $x^{t+1}(\gamma) \in \widehat{\mathcal{S}}$ for any $0 < \gamma \leq \gamma_0$.*

Proof. Fix any $\gamma \in (0, \gamma_0]$. It follows from (51) that

$$
x^t - x^{t+1}(\gamma) + \alpha(\gamma)(x^t - x^{t-1}) - \gamma(F(x^t) + \beta(\gamma)(F(x^t) - F(x^{t-1}))) \in \gamma B(x^{t+1}(\gamma)).
$$

Also, by the definition of x^*, one has $-\gamma F(x^*) \in \gamma B(x^*)$. These along with the monotonicity of B imply that

$$\langle x^t - x^{t+1}(\gamma) + w, x^{t+1}(\gamma) - x^* \rangle \geq 0, \tag{54}$$

where

$$w = \alpha(\gamma)(x^t - x^{t-1}) - \gamma(F(x^t) - F(x^*)) - \gamma\beta(\gamma)(F(x^t) - F(x^{t-1})). \tag{55}$$

It follows from (54) that

$$\|x^{t+1}(\gamma) - x^*\|^2 \leq \langle x^t - x^* + w, x^{t+1}(\gamma) - x^* \rangle \leq \|x^t - x^* + w\| \|x^{t+1}(\gamma) - x^*\|,$$

which implies that

$$\|x^{t+1}(\gamma) - x^*\| \leq \|x^t - x^* + w\| \leq \|x^t - x^*\| + \|w\|. \tag{56}$$

Notice from Algorithm 1 that $0 < \gamma_{t-1} \leq \gamma_0$ and $0 \leq \kappa_t, \kappa_{t-1} < 1/3$. Using these and (53), we have

$$\gamma\beta(\gamma) = \frac{\gamma_{t-1}(1-\kappa_t)}{(1-\kappa_{t-1})}\left(1 + \frac{2\mu\gamma_{t-1}}{1-\kappa_{t-1}}\right)^{-1} \leq \frac{3}{2}\gamma_0, \tag{57}$$

$$\alpha(\gamma) = \frac{\kappa_{t-1}\gamma\beta(\gamma)}{\gamma_{t-1}} = \frac{\kappa_{t-1}(1-\kappa_t)}{1-\kappa_{t-1}}\left(1 + \frac{2\mu\gamma_{t-1}}{1-\kappa_{t-1}}\right)^{-1} \leq \frac{1}{2}. \tag{58}$$

Recall that $\mathcal{S}$, r_0 and w are given in (8) and (55), respectively. Using $x^t, x^{t-1}, x^* \in \mathcal{S}$, $0 < \gamma \leq \gamma_0$, (8), (55), (57) and (58), we have

$$\begin{aligned}
\|w\| &\leq \alpha(\gamma)\|x^t - x^{t-1}\| + \gamma L_{\mathcal{S}}\|x^t - x^*\| + \gamma\beta(\gamma)L_{\mathcal{S}}\|x^t - x^{t-1}\| \\
&\leq \left(\frac{1}{2} + \frac{3}{2}\gamma_0 L_{\mathcal{S}}\right)\|x^t - x^{t-1}\| + \gamma_0 L_{\mathcal{S}}\|x^t - x^*\| \\
&\leq \left(\frac{1}{2} + \frac{3}{2}\gamma_0 L_{\mathcal{S}}\right)(\|x^t - x^*\| + \|x^{t-1} - x^*\|) + \gamma_0 L_{\mathcal{S}}\|x^t - x^*\| \\
&\leq \frac{2}{\sqrt{1-2\nu^2}}\left(\frac{1}{2} + \frac{3}{2}\gamma_0 L_{\mathcal{S}}\right)r_0 + \frac{1}{\sqrt{1-2\nu^2}}\gamma_0 L_{\mathcal{S}} r_0 = \frac{(1+4\gamma_0 L_{\mathcal{S}})r_0}{\sqrt{1-2\nu^2}}.
\end{aligned}$$

This together with $x^t \in \mathcal{S}$, (8) and (56) yields

$$\|x^{t+1}(\gamma) - x^*\| \leq \|x^t - x^*\| + \|w\| \leq \frac{(2+4\gamma_0 L_{\mathcal{S}})r_0}{\sqrt{1-2\nu^2}}.$$

The conclusion then follows from this and the definition of $\widehat{\mathcal{S}}$ in (9). □

The next lemma provides an upper bound on n_t, which will be used to prove Theorem 1.

Lemma 7. *Assume that $x^{t-1}, x^t \in \mathcal{S}$ for some $t \geq 1$ and Algorithm 1 has not yet terminated at iteration $t-1$. Then x^{t+1} is successfully generated by Algorithm 1 at iteration t with $n_t \leq N_t$, that is, the number of evaluations of F and resolvent of B at the iteration t is at most $N_t + 1$, where N_t is defined in* (10).

Proof. Recall that $x^{t+1}(\gamma)$ and $L_{t+1}(\gamma)$ are defined in (51) and (52), respectively. It follows from Lemma 6 that $x^{t+1}(\gamma) \in \widehat{\mathcal{S}}$ for any $0 < \gamma \leq \gamma_0$. Also, notice that $x^t \in \mathcal{S} \subset \widehat{\mathcal{S}}$. By these and Lemma 1(ii), one has

$$\|F(x^{t+1}(\gamma)) - F(x^t)\| \leq L_{\widehat{\mathcal{S}}}\|x^{t+1}(\gamma) - x^t\| \quad \forall 0 < \gamma \leq \gamma_0.$$

Using this and (52), we obtain that for any $0 < \gamma \le \gamma_0$,

$$\begin{aligned} L_{t+1}(\gamma) &\le (\|F(x^{t+1}(\gamma)) - F(x^t)\| + \kappa_t\gamma^{-1}\|(x^{t+1}(\gamma) - x^t)\|)/\|x^{t+1}(\gamma) - x^t\| \\ &\le \|F(x^{t+1}(\gamma)) - F(x^t)\|/\|x^{t+1}(\gamma) - x^t\| + \kappa_t\gamma^{-1} \le L_{\widehat{\mathcal{S}}} + \kappa_t\gamma^{-1}. \end{aligned} \tag{59}$$

Let $\gamma = \gamma_0\delta^{N_t}$. One can verify that $0 < \gamma \le \min\{(\nu(1-\kappa_t) - \kappa_t)/L_{\widehat{\mathcal{S}}}, \gamma_0\}$. It then follows from this and (59) that

$$L_{t+1}(\gamma)\gamma \le L_{\widehat{\mathcal{S}}}\gamma + \kappa_t \le \nu(1-\kappa_t),$$

which, together with (6), (52), $\gamma = \gamma_0\delta^{N_t}$ and the definition of n_t (see step 2 of Algorithm 1), implies that $n_t \le N_t$ and hence x^{t+1} is successfully generated. □

We are now ready to prove the main results presented in Section 2, namely, Theorems 1 and 2.

Proof of Theorem 1. We prove this theorem by induction. Indeed, notice from Algorithm 1 that $x^0 = x^1 \in \mathcal{S}$. It then follows from Lemma 7 that x^2 is successfully generated and $n_1 \le N_1$. Hence, Algorithm 1 is well-defined at iteration 1. By this, (41), and (48) with $k = 1$, one has

$$\|x^2 - x^*\|^2 \overset{(48)}{\le} \frac{1}{(1-2\nu^2)\theta_1}\|x^0 - x^*\|^2 \overset{(41)}{=} \frac{1}{1-2\nu^2}\|x^0 - x^*\|^2,$$

which together with (8) implies that $x^2 \in \mathcal{S}$. Now, suppose for induction that Algorithm 1 is well-defined at iteration 1 to $t-1$ and $x^i \in \mathcal{S}$ for all $0 \le i \le t$ for some $2 \le t \in \mathbb{T} - 1$. It then follows from Lemma 7 that x^{t+1} is successfully generated and $n_t \le N_t$. Hence, Algorithm 1 is well-defined at iteration t. By this, (41), and (48) with $k = t$, one has

$$\|x^{t+1} - x^*\|^2 \overset{(48)}{\le} \frac{1}{(1-2\nu^2)\theta_t}\|x^0 - x^*\|^2 \overset{(41)}{\le} \frac{1}{1-2\nu^2}\|x^0 - x^*\|^2,$$

which together with (8) implies that $x^{t+1} \in \mathcal{S}$. Hence, the induction is completed and the conclusion of this theorem holds. □

Proof of Theorem 2. Notice from Algorithm 1 that $0 < \gamma_t \le \gamma_0$ and $0 \le \kappa_t < 1/3$ for all $0 \le t \in \mathbb{T} - 1$. Using these and (5), we have that for all $1 \le t \in \mathbb{T} - 1$,

$$\gamma_t\beta_t = \frac{\gamma_{t-1}(1-\kappa_t)}{(1-\kappa_{t-1})}\left(1 + \frac{2\mu\gamma_{t-1}}{1-\kappa_{t-1}}\right)^{-1} \le \frac{3}{2}\gamma_0, \quad \alpha_t = \frac{\kappa_{t-1}\gamma_t\beta_t}{\gamma_{t-1}} = \frac{\kappa_{t-1}(1-\kappa_t)}{1-\kappa_{t-1}}\left(1 + \frac{2\mu\gamma_{t-1}}{1-\kappa_{t-1}}\right)^{-1} \le \frac{1}{2}. \tag{60}$$

In addition, it follows from Theorem 1 that $n_t \le N_t$ for all $1 \le t \in \mathbb{T} - 1$, where N_t is defined in (10). By this and (5), one has

$$\gamma_t = \gamma_0\delta^{n_t} \ge \min\left\{\frac{\delta\left(\nu(1-\kappa_t) - \kappa_t\right)}{L_{\widehat{\mathcal{S}}}}, \gamma_0\right\} \quad \forall 1 \le t \in \mathbb{T} - 1. \tag{61}$$

Notice from Algorithm 1 that $0 \le \kappa_t \le \xi/(1+\xi)$. It then follows that $\kappa_t/(1-\kappa_t) \le \xi$. Using this and (61), we obtain that for all $1 \le t \in \mathbb{T} - 1$,

$$\gamma_t \overset{(61)}{\ge} \min\left\{\frac{\delta(1-\kappa_t)}{L_{\widehat{\mathcal{S}}}}\left(\nu - \frac{\kappa_t}{1-\kappa_t}\right), \gamma_0\right\} \ge \min\left\{\frac{2\delta(\nu - \xi)}{3L_{\widehat{\mathcal{S}}}}, \gamma_0\right\}, \tag{62}$$

$$\begin{aligned} 1 + \frac{2\mu\gamma_t}{1-\kappa_t} &\ge \min\left\{1 + 2\mu\delta\left(\nu - \frac{\kappa_t}{1-\kappa_t}\right)/L_{\widehat{\mathcal{S}}}, 1 + \frac{2\mu\gamma_0}{1-\kappa_t}\right\} \\ &\ge \min\left\{\frac{1}{L_{\widehat{\mathcal{S}}}}\left(L_{\widehat{\mathcal{S}}} + 2\mu\delta\nu - 2\mu\delta\xi\right), 1 + 2\mu\gamma_0\right\}. \end{aligned} \tag{63}$$

By (8), (41), (48), and (63), one has that for all $1 \le t \in \mathbb{T} - 1$,

$$\|x^{t+1} - x^*\|^2 \le \frac{1}{(1-2\nu^2)\prod_{i=1}^{t-1}\left(1 + \frac{2\mu\gamma_i}{1-\kappa_i}\right)}\|x^0 - x^*\|^2$$
$$\le \frac{r_0^2}{1-2\nu^2}\left(\max\left\{\frac{L_{\widehat{\mathcal{S}}}}{L_{\widehat{\mathcal{S}}} + 2\mu\delta\nu - 2\mu\delta\xi}, \frac{1}{1+2\mu\gamma_0}\right\}\right)^{t-1}.$$

It then follows that for all $3 \le t \in \mathbb{T} - 1$,

$$\max\{\|x^t - x^{t-1}\|, \|x^{t+1} - x^t\|\} \le \max\{\|x^t - x^*\| + \|x^{t-1} - x^*\|, \|x^{t+1} - x^*\| + \|x^t - x^*\|\}$$
$$\le \frac{2r_0}{\sqrt{1-2\nu^2}}\left(\max\left\{\frac{L_{\widehat{\mathcal{S}}}}{L_{\widehat{\mathcal{S}}} + 2\mu\delta\nu - 2\mu\delta\xi}, \frac{1}{1+2\mu\gamma_0}\right\}\right)^{\frac{t-3}{2}}. \tag{64}$$

Suppose for contradiction that Algorithm 1 runs for at least $T+1$ iterations. It then follows that (7) fails for $t = T$, which along with (40) implies that $\|v^T\| > \epsilon$. In addition, recall from Theorem 1(ii) that $x^t \in \mathcal{S}$ for all $t \in \mathbb{T}$. By this, (38) and Lemma 1(i), one has

$$\|\Delta^t\| = \|F(x^t) - F(x^{t-1})\| \le L_{\mathcal{S}}\|x^t - x^{t-1}\| \quad \forall 1 \le t \in \mathbb{T}. \tag{65}$$

Also, notice from (11) that $T \ge 3$. By this, $\gamma_T \le \gamma_0$, (11), (40), (60), (62), (64), and (65), one has

$$\|v^T\| \overset{(40)}{\le} \frac{1}{\gamma_T}\left(\|x^{T+1} - x^T\| + \alpha_T\|x^T - x^{T-1}\| + \gamma_T\|\Delta^{T+1}\| + \gamma_T\beta_T\|\Delta^T\|\right)$$
$$\overset{(65)}{\le} \frac{1}{\gamma_T}\left(\|x^{T+1} - x^T\| + \alpha_T\|x^T - x^{T-1}\| + \gamma_T L_{\mathcal{S}}\|x^{T+1} - x^T\| + \gamma_T\beta_T L_{\mathcal{S}}\|x^T - x^{T-1}\|\right)$$
$$\overset{(64)}{\le} \frac{2r_0}{\gamma_T\sqrt{1-2\nu^2}}(1 + \alpha_T + \gamma_T L_{\mathcal{S}} + \gamma_T\beta_T L_{\mathcal{S}})\left(\max\left\{\frac{L_{\widehat{\mathcal{S}}}}{L_{\widehat{\mathcal{S}}} + 2\mu\delta\nu - 2\mu\delta\xi}, \frac{1}{1+2\mu\gamma_0}\right\}\right)^{\frac{T-3}{2}}$$
$$\overset{(60)}{\le} \frac{2r_0}{\gamma_T\sqrt{1-2\nu^2}}\left(1 + \frac{1}{2} + \gamma_0 L_{\mathcal{S}} + \frac{3}{2}\gamma_0 L_{\mathcal{S}}\right)\left(\max\left\{\frac{L_{\widehat{\mathcal{S}}}}{L_{\widehat{\mathcal{S}}} + 2\mu\delta\nu - 2\mu\delta\xi}, \frac{1}{1+2\mu\gamma_0}\right\}\right)^{\frac{T-3}{2}}$$
$$\overset{(62)}{\le} \frac{r_0(3 + 5\gamma_0 L_{\mathcal{S}})}{\sqrt{1-2\nu^2}\min\left\{\frac{2\delta(\nu-\xi)}{3L_{\widehat{\mathcal{S}}}}, \gamma_0\right\}}\left(\max\left\{\frac{L_{\widehat{\mathcal{S}}}}{L_{\widehat{\mathcal{S}}} + 2\mu\delta\nu - 2\mu\delta\xi}, \frac{1}{1+2\mu\gamma_0}\right\}\right)^{\frac{T-3}{2}} \overset{(11)}{\le} \epsilon,$$

which leads to a contradiction. Hence, Algorithm 1 terminates in at most T iterations. Suppose that Algorithm 1 terminates at iteration t and outputs x^{t+1} for some $t \le T$. It then follows that (7) holds for such t. By this and (40), one can see that $\|v^t\| \le \epsilon$, which together with (42) implies that $\mathrm{res}_{F+B}(x^{t+1}) \le \epsilon$.

Notice from Algorithm 1 that $0 < \delta < 1$, $0 \le \kappa_t < 1/3$ and $\kappa_t/(1-\kappa_t) \le \xi$. By these, (10) and Theorem 1(ii), one has that

$$n_t \le N_t = \left\lceil \log\left(\frac{(1-\kappa_t)\left(\nu - \frac{\kappa_t}{1-\kappa_t}\right)}{\gamma_0 L_{\widehat{\mathcal{S}}}}\right)/\log\delta\right\rceil_+ \le \left\lceil \log\left(\frac{2(\nu-\xi)}{3\gamma_0 L_{\widehat{\mathcal{S}}}}\right)/\log\delta\right\rceil_+ \quad \forall 1 \le t \le |\mathbb{T}| - 2. \tag{66}$$

Observe that $|\mathbb{T}| \le T + 2$ and also the total number of inner iterations of Algorithm 1 is $\sum_{t=1}^{|\mathbb{T}|-2}(n_t + 1)$. It then follows from (66) that

$$\sum_{t=1}^{|\mathbb{T}|-2}(n_t + 1) \le T\left(1 + \left\lceil \log\left(\frac{2(\nu-\xi)}{3\gamma_0 L_{\widehat{\mathcal{S}}}}\right)/\log\delta\right\rceil_+\right),$$

which together with (11) implies that the conclusion holds. □

5.2 Proof of the main result in Section 3

In this subsection we first establish several technical lemmas and then use them to prove Theorem 3.

Recall from Section 3 that $\{z^k\}_{k\in\mathbb{K}}$ denotes all the iterates generated by Algorithm 2, where $\mathbb{K}$ is a subset of consecutive nonnegative integers starting from 0. Notice that at iteration $0 \le k \in \mathbb{K}-1$ of Algorithm 2, Algorithm 1 is called to find an approximate solution of the following strongly MI problem

$$0 \in (F_k + B)(x) = (F+B)(x) + \frac{1}{\rho_k}(x - z^k). \tag{67}$$

Since $F+B$ is maximal monotone, it follows that the domain of the resolvent of $F+B$ is $\mathbb{R}^n$. As a result, there exists some $z_*^k \in \mathbb{R}^n$ such that

$$z_*^k = (I + \rho_k(F+B))^{-1}(z^k). \tag{68}$$

Moreover, z_*^k is the unique solution of problem (67) and thus

$$0 \in (F_k + B)(z_*^k) = (F+B)(z_*^k) + \frac{1}{\rho_k}(z_*^k - z^k). \tag{69}$$

Lemma 8. *Let $\{z^k\}_{k\in\mathbb{K}}$ be generated by Algorithm 2. Then for all $0 \le k \in \mathbb{K}-1$, we have*

$$\|z^{k+1} - z_*^k\| \le \rho_k \tau_k, \tag{70}$$

where z_*^k is defined in (68).

Proof. By the definition of z^{k+1} (see step 2 of Algorithm 2) and Theorem 2, there exists some $v \in (F_k + B)(z^{k+1})$ with $\|v\| \le \tau_k$. It follows from this and (69) that

$$v - \frac{1}{\rho_k}(z^{k+1} - z^k) \in (F+B)(z^{k+1}), \quad -\frac{1}{\rho_k}(z_*^k - z^k) \in (F+B)(z_*^k).$$

By the monotonicity of $F+B$, one has

$$\langle v - \frac{1}{\rho_k}(z^{k+1} - z^k) + \frac{1}{\rho_k}(z_*^k - z^k), z^{k+1} - z_*^k\rangle \ge 0,$$

which yields

$$\|z^{k+1} - z_*^k\|^2 \le \rho_k\langle v, z^{k+1} - z_*^k\rangle \le \rho_k\|v\|\|z^{k+1} - z_*^k\|.$$

It then follows from this and $\|v\| \le \tau_k$ that $\|z^{k+1} - z_*^k\| \le \rho_k\|v\| \le \rho_k\tau_k$. □

Lemma 9. *Let $\{z^k\}_{k\in\mathbb{K}}$ be generated by Algorithm 2. Then we have*

$$\|z^s - x^*\| \le \|z^0 - x^*\| + \sum_{k=0}^{s-1} \rho_k\tau_k \qquad \forall 1 \le s \in \mathbb{K}, \tag{71}$$

$$\|z^{s+1} - z^s\| \le \|z^0 - x^*\| + \sum_{k=0}^{s} \rho_k\tau_k \qquad \forall 1 \le s \in \mathbb{K}-1. \tag{72}$$

Proof. By (69) and the definition of x^*, one has

$$z^k - z_*^k \in \rho_k(F+B)(z_*^k), \quad 0 \in \rho_k(F+B)(x^*),$$

which together with the monotonicity of $F+B$ yield

$$0 \le 2\langle z^k - z_*^k, z_*^k - x^*\rangle = \|z^k - x^*\|^2 - \|z^k - z_*^k\|^2 - \|z_*^k - x^*\|^2.$$

It follows that

$$\|z^k - z_*^k\|^2 + \|z_*^k - x^*\|^2 \le \|z^k - x^*\|^2,$$

which implies that

$$\|z_*^k - x^*\| \le \|z^k - x^*\|, \quad \|z^k - z_*^k\| \le \|z^k - x^*\|. \tag{73}$$

By the first relation in (73), one has

$$\|z^{k+1} - x^*\| \le \|z^{k+1} - z_*^k\| + \|z_*^k - x^*\| \le \|z^{k+1} - z_*^k\| + \|z^k - x^*\|.$$

Summing up the above inequalities for $k = 0, \ldots, s-1$ yields

$$\|z^s - x^*\| \le \|z^0 - x^*\| + \sum_{k=0}^{s-1} \|z^{k+1} - z_*^k\|,$$

which along with (70) implies that (71) holds. In addition, using (70) with $k = s$, (71) and (73), we have

$$\|z^{s+1} - z^s\| \le \|z^s - z_*^s\| + \|z^{s+1} - z_*^s\| \overset{(70),(73)}{\le} \|z^s - x^*\| + \rho_s \tau_s \overset{(71)}{\le} \|z^0 - x^*\| + \sum_{k=0}^{s} \rho_k \tau_k.$$

Hence, (72) holds as desired. □

Define

$$\mathcal{S}_k = \left\{ x \in \mathrm{dom}(B) : \ \|x - z_*^k\| \le \frac{1}{\sqrt{1-2\nu^2}} \|z^k - z_*^k\| \right\} \qquad \forall 0 \le k \in \mathbb{K} - 1, \tag{74}$$

$$\widehat{\mathcal{S}}_k = \left\{ x \in \mathrm{dom}(B) : \ \|x - z_*^k\| \le \frac{2 + 4\gamma_0 L_{\mathcal{Q}}}{\sqrt{1-2\nu^2}} \|z^k - z_*^k\| \right\} \qquad \forall 0 \le k \in \mathbb{K} - 1, \tag{75}$$

where z_*^k is defined in (68), $L_{\mathcal{Q}}$ is given in Lemma 2, and ν and γ_0 are the input parameters of Algorithm 2.

Lemma 10. *Let $\mathcal{S}_k$ and $\widehat{\mathcal{S}}_k$ be defined in (74) and (75). Then for all $0 \le k \in \mathbb{K} - 1$, we have*

$$\mathcal{S}_k \subseteq \mathcal{Q}, \quad \widehat{\mathcal{S}}_k \subseteq \widehat{\mathcal{Q}},$$

where $\mathcal{Q}$ and $\widehat{\mathcal{Q}}$ are defined in (15) and (16). Consequently, for all $0 \le k \in \mathbb{K} - 1$, F_k is $L_{\mathcal{Q}}$- and $L_{\widehat{\mathcal{Q}}}$-Lipschitz continuous on $\mathcal{S}_k$ and $\widehat{\mathcal{S}}_k$, respectively, where $L_{\mathcal{Q}}$ and $L_{\widehat{\mathcal{Q}}}$ are given in Lemma 2.

Proof. Fix any $0 \le k \in \mathbb{K} - 1$ and $x \in \mathcal{S}_k$. By this, (71), (73), (74), and $\sum_{i=0}^{\infty} \rho_i \tau_i = \rho_0 \tau_0 / (1 - \sigma\zeta)$, we have

$$\|x - x^*\| \le \|x - z_*^k\| + \|z_*^k - x^*\| \overset{(74)}{\le} \frac{1}{\sqrt{1-2\nu^2}} \|z^k - z_*^k\| + \|z_*^k - x^*\| \overset{(73)}{\le} \left(\frac{1}{\sqrt{1-2\nu^2}} + 1 \right) \|z^k - x^*\|$$
$$\overset{(71)}{\le} \left(\frac{1}{\sqrt{1-2\nu^2}} + 1 \right) \left(\|z^0 - x^*\| + \sum_{i=0}^{k-1} \rho_i \tau_i \right) \le \left(\frac{1}{\sqrt{1-2\nu^2}} + 1 \right) \left(\|z^0 - x^*\| + \frac{\rho_0 \tau_0}{1 - \sigma\zeta} \right),$$

which together with (15) implies that $x \in \mathcal{Q}$. It then follows that $\mathcal{S}_k \subseteq \mathcal{Q}$. Similarly, one can show that $\widehat{\mathcal{S}}_k \subseteq \widehat{\mathcal{Q}}$. By these and the definition of $L_{\mathcal{Q}}$ and $L_{\widehat{\mathcal{Q}}}$ in Lemma 2, one can see that F_k is $L_{\mathcal{Q}}$- and $L_{\widehat{\mathcal{Q}}}$-Lipschitz continuous on $\mathcal{S}_k$ and $\widehat{\mathcal{S}}_k$, respectively. □

Lemma 11. *Let γ_0, δ, ν, ξ, $\{\rho_k\}$ and $\{\tau_k\}$ be given in Algorithm 2, and let $\bar{r}_0$ and Λ be defined in* (15) *and* (17)*. Then for any $0 \le k \in \mathbb{K}-1$, the number of evaluations of F and resolvent of B performed in the kth iteration of Algorithm 2 is at most M_k, where*

$$M_k = \left(3 + \left\lceil \frac{2\log\frac{(\bar{r}_0+\Lambda)(3+5\gamma_0 L_{\mathcal{Q}})}{\tau_k\sqrt{1-2\nu^2}\min\left\{\frac{2\delta(\nu-\xi)}{3L_{\widehat{\mathcal{Q}}}},\gamma_0\right\}}}{\log\left(\min\left\{1+\frac{2\delta(\nu-\xi)}{\rho_k L_{\widehat{\mathcal{Q}}}}, 1+\frac{2\gamma_0}{\rho_k}\right\}\right)} \right\rceil_+ \right)\left(1+\left\lceil \frac{\log\left(\frac{2(\nu-\xi)}{3\gamma_0 L_{\widehat{\mathcal{Q}}}}\right)}{\log\delta} \right\rceil_+\right). \tag{76}$$

Proof. Recall that F_k+B is $1/\rho_k$-strongly monotone and $(F_k+B)^{-1}(0) \neq \emptyset$. In addition, it follows from Lemma 10 that F_k is $L_{\mathcal{Q}}$- and $L_{\widehat{\mathcal{Q}}}$-Lipschitz continuous on $\mathcal{S}_k$ and $\widehat{\mathcal{S}}_k$, respectively. Using (71), (73), and the fact that $\Lambda = \sum_{k=0}^{\infty}\rho_k\tau_k$, we have

$$\|z^k - z_*^k\| \le \|z^k - x^*\| \le \|z^0 - x^*\| + \sum_{t=0}^{k-1}\rho_t\tau_t \le \bar{r}_0 + \Lambda,$$

where z_*^k is given in (68). The conclusion then follows from applying Theorem 2 to the subproblem (67) with ϵ, μ, $L_{\mathcal{S}}$, $L_{\widehat{\mathcal{S}}}$, and r_0 in (12) being replaced by τ_k, $1/\rho_k$, $L_{\mathcal{Q}}$, $L_{\widehat{\mathcal{Q}}}$, and $\bar{r}_0+\Lambda$, respectively. □

Proof of Theorem 3. Suppose for contradiction that Algorithm 2 runs for more than $K+1$ outer iterations. By this and Algorithm 2, one can assert that (14) does not hold for $k=K$. On the other hand, by (72), $\rho_k = \rho_0\zeta^k$ and $\tau_k = \tau_0\sigma^k$, one has

$$\frac{\|z^{K+1}-z^K\|}{\rho_K} \le \frac{\|z^0-x^*\| + \sum_{k=0}^{K}\rho_k\tau_k}{\rho_K} \le \frac{\|z^0-x^*\| + \sum_{k=0}^{\infty}\rho_k\tau_k}{\rho_K} = \frac{\bar{r}_0+\Lambda}{\rho_0\zeta^K}.$$

This together with the definition of K in (19) implies that

$$\frac{\|z^{K+1}-z^K\|}{\rho_K} \le \frac{\bar{r}_0+\Lambda}{\rho_0\zeta^K} \le \frac{\varepsilon}{2}, \quad \tau_K = \tau_0\sigma^K \le \frac{\varepsilon}{2},$$

and thus (14) holds for $k=K$, which contradicts the above assertion. Hence, Algorithm 2 must terminate in at most $K+1$ outer iterations.

Suppose that Algorithm 2 terminates at some iteration k. Then we have

$$\frac{1}{\rho_k}\|z^{k+1}-z^k\| \le \frac{\varepsilon}{2}, \quad \tau_k \le \frac{\varepsilon}{2}. \tag{77}$$

In addition, by the definition of z^{k+1} (see step 2 of Algorithm 2), Theorem 2, and (13), there exists some v such that

$$v \in (F+B)(z^{k+1}) + \frac{1}{\rho_k}(z^{k+1}-z^k), \quad \|v\| \le \tau_k. \tag{78}$$

Observe that $v - (z^{k+1}-z^k)/\rho_k \in (F+B)(z^{k+1})$. It follows from this, (77), (78), and the definition of res_{F+B} that

$$\mathrm{res}_{F+B}(z^{k+1}) \le \left\|v - \frac{1}{\rho_k}(z^{k+1}-z^k)\right\| \le \|v\| + \frac{1}{\rho_k}\|z^{k+1}-z^k\| \overset{(78)}{\le} \tau_k + \frac{1}{\rho_k}\|z^{k+1}-z^k\| \overset{(77)}{\le} \varepsilon.$$

Recall from Lemma 11 that the number of evaluations of F and resolvent of B performed in the kth iteration of Algorithm 2 is at most M_k, where M_k is given in (76). In addition, using (17), (18) and (76), we have

$$M_k = C_2\left(3 + \left\lceil \frac{2C_1 - 2k\log\sigma}{\log\left(\min\left\{1+\frac{2\delta(\nu-\xi)}{\rho_0\zeta^k L_{\widehat{\mathcal{Q}}}}, 1+\frac{2\gamma_0}{\rho_0\zeta^k}\right\}\right)} \right\rceil_+ \right). \tag{79}$$

By the concavity of $\log(1+y)$, one has $\log(1+\vartheta y) \geq \vartheta \log(1+y)$ for all $y > -1$ and $\vartheta \in [0,1]$. Using this, we obtain that

$$\log\left(\min\left\{1+\frac{2\delta(\nu-\xi)}{\rho_0\zeta^k L_{\widehat{Q}}}, 1+\frac{2\gamma_0}{\rho_0\zeta^k}\right\}\right) = \min\left\{\log\left(1+\zeta^{-k}\frac{2\delta(\nu-\xi)}{\rho_0 L_{\widehat{Q}}}\right), \log\left(1+\zeta^{-k}\frac{2\gamma_0}{\rho_0}\right)\right\}$$
$$\geq \min\left\{\zeta^{-k}\log\left(1+\frac{2\delta(\nu-\xi)}{\rho_0 L_{\widehat{Q}}}\right), \zeta^{-k}\log\left(1+\frac{2\gamma_0}{\rho_0}\right)\right\}. \tag{80}$$

By (79), (80), $\sigma \in (0,1)$ and $\zeta > 1$, one has that for all $k \geq 0$,

$$M_k \overset{(79)}{\leq} C_2\left(4+\frac{(2C_1-2k\log\sigma)_+}{\log\left(\min\left\{1+\frac{2\delta(\nu-\xi)}{\rho_0\zeta^k L_{\widehat{Q}}}, 1+\frac{2\gamma_0}{\rho_0\zeta^k}\right\}\right)}\right) \leq C_2\left(4+\frac{2\,(C_1)_+-2k\log\sigma}{\log\left(\min\left\{1+\frac{2\delta(\nu-\xi)}{\rho_0\zeta^k L_{\widehat{Q}}}, 1+\frac{2\gamma_0}{\rho_0\zeta^k}\right\}\right)}\right)$$
$$\overset{(80)}{\leq} C_2\left(4+\frac{2\zeta^k\,(C_1)_+-2k\zeta^k\log\sigma}{\log\left(\min\left\{1+\frac{2\delta\nu-2\delta\xi}{\rho_0 L_{\widehat{Q}}}, 1+\frac{2\gamma_0}{\rho_0}\right\}\right)}\right). \tag{81}$$

Observe that $|\mathbb{K}| \leq K+2$ and also the total number of inner iterations of Algorithm 2 is at most $\sum_{t=0}^{|\mathbb{K}|-2} M_t$. It then follows from (18), (79) and (81) that the total number of evaluations of F and resolvent of B performed in Algorithm 2 is at most

$$\sum_{k=0}^{|\mathbb{K}|-2} M_k \leq C_2\sum_{k=0}^{K}\left(4+\frac{2\zeta^k\,(C_1)_+-2k\zeta^k\log\sigma}{\log\left(\min\left\{1+\frac{2\delta\nu-2\delta\xi}{\rho_0 L_{\widehat{Q}}}, 1+\frac{2\gamma_0}{\rho_0}\right\}\right)}\right)$$
$$\leq C_2\left(4K+4+2C_3\,(C_1)_+\zeta^{K+1}-2C_3(\log\sigma)K\zeta^{K+1}\right), \tag{82}$$

where the second inequality is due to $\sum_{k=0}^{K}\zeta^k \leq \zeta^{K+1}/(\zeta-1)$ and $\sum_{k=0}^{K}k\zeta^k \leq K\zeta^{K+1}/(\zeta-1)$. By the definition of K in (19), one has

$$K \leq \max\left\{\log_\zeta\left(\frac{2\bar{r}_0+2\Lambda}{\varepsilon\rho_0}\right)+1, \frac{\log(2\tau_0/\varepsilon)}{\log(1/\sigma)}+1, 0\right\},$$

which together with $\zeta > 1$ implies that

$$\zeta^K \leq \max\left\{\frac{2\zeta(\bar{r}_0+\Lambda)}{\varepsilon\rho_0}, \zeta\left(\frac{2\tau_0}{\varepsilon}\right)^{\frac{\log\zeta}{\log(1/\sigma)}}, 1\right\}. \tag{83}$$

Using (19), (20), (82) and (83), we can see that $\sum_{k=0}^{|\mathbb{K}|-2} M_k \leq M$. □

References

[1] F. Facchinei and J. S. Pang. *Finite-Dimensional Variational Inequalities and Complementarity Problems.* Springer Science & Business Media, 2007.

[2] K. Huang and S. Zhang. New first-order algorithms for stochastic variational inequalities. *arXiv preprint arXiv:2107.08341*, 2021.

[3] K. Huang and S. Zhang. A unifying framework of accelerated first-order approach to strongly monotone variational inequalities. *arXiv preprint arXiv:2103.15270*, 2021.

[4] G. M. Korpelevich. Extragradient method for finding saddle points and other problems. *Ekonomika i Matem. Metody*, 12:747–756, 1976.

[5] G. Kotsalis, G. Lan, and T. Li. Simple and optimal methods for stochastic variational inequalities, I: operator extrapolation. *arXiv preprint arXiv:2011.02987*, 2020.

[6] T. Lin, C. Jin, and M. I. Jordan. Near-optimal algorithms for minimax optimization. In *Proceedings of Machine Learning Research*, pages 1–42, 2020.

[7] P.-L. Lions and B. Mercier. Splitting algorithms for the sum of two nonlinear operators. *SIAM Journal on Numerical Analysis*, 16(6):964–979, 1979.

[8] Z. Lu and Z. Zhou. Iteration-complexity of first-order augmented Lagrangian methods for convex conic programming. *arXiv preprint arXiv:1803.09941*, 2018.

[9] Y. Malitsky and M. K. Tam. A forward-backward splitting method for monotone inclusions without cocoercivity. *SIAM Journal on Optimization*, 30(2):1451–1472, 2020.

[10] A. Mokhtari, A. E. Ozdaglar, and S. Pattathil. Convergence rate of $O(1/k)$ for optimistic gradient and extragradient methods in smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 30:3230–3251, 2020.

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

[12] R. D. Monteiro and B. F. Svaiter. Complexity of variants of Tseng's modified FB splitting and Korpelevich's methods for hemivariational inequalities with applications to saddle-point and convex optimization problems. *SIAM Journal on Optimization*, 21(4):1688–1720, 2011.

[13] A. Nemirovski. Prox-method with rate of convergence $O(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle-point problems. *SIAM Journal on Optimization*, pages 229–251, 2005.

[14] Y. E. Nesterov. Dual extrapolation and its applications to solving variational inequalities and related problems. *Mathematical Programming*, 109:319–344, 2003.

[15] G. B. Passty. Ergodic convergence to a zero of the sum of monotone operators in Hilbert space. *Journal of Mathematical Analysis and Applications*, 72(2):383–390, 1979.

[16] L. D. Popov. A modification of the Arrow-Hurwicz method for search of saddle points. *Mathematical notes of the Academy of Sciences of the USSR*, 28(5):845–848, 1980.

[17] M. Sibony. Méthodes itératives pour les équations et inéquations aux dérivées partielles non linéaires de type monotone. *CALCOLO*, 7:65–183, 1970.

[18] Q. Tran-Dinh. The connection between Nesterov's accelerated methods and Halpern fixed-point iterations. *arXiv preprint arXiv:2203.04869*, 2022.

[19] P. Tseng. A modified forward-backward splitting method for maximal monotone mappings. *SIAM Journal on Control and Optimization*, 38(2):431–446, 2000.

[20] P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. Manuscript, May 2008.

[21] T. Vladislav, T. Yaroslav, B. Ekaterina, K. Dmitry, A. Gasnikov, and P. Dvurechensky. On accelerated methods for saddle-point problems with composite structure. *arXiv preprint arXiv:2103.09344*, 2021.

[22] J. Yang, S. Zhang, N. Kiyavash, and N. He. A catalyst framework for minimax optimization. In *Advances in Neural Information Processing Systems*, pages 5667–5678, 2020.

[23] T. Yoon and E. K. Ryu. Accelerated algorithms for smooth convex-concave minimax problems with $O(1/k^2)$ rate on squared gradient norm. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 12098–12109, 2021.

[24] J. Zhang, M. Wang, M. Hong, and S. Zhang. Primal-dual first-order methods for affinely constrained multi-block saddle point problems. *arXiv preprint arXiv:2109.14212*, 2021.