Primal-dual extrapolation methods for monotone inclusions under local Lipschitz continuity

Zhaosong Lu * Sanyou Mei *

June 1, 2022 (Revised: August 30, 2024)

Abstract

In this paper we consider a class of monotone inclusion (MI) problems of finding a zero of the sum of two monotone operators, in which one operator is maximal monotone while the other is *locally Lipschitz* continuous. We propose primal-dual extrapolation methods to solve them using a point and operator extrapolation technique, whose parameters are chosen by a backtracking line search scheme. The proposed methods enjoy an operation complexity of $\mathcal{O}\left(\log\varepsilon^{-1}\right)$ and $\mathcal{O}\left(\varepsilon^{-1}\log\varepsilon^{-1}\right)$, measured by the number of fundamental operations consisting only of evaluations of one operator and resolvent of the other operator, for finding an ε-residual solution of strongly and non-strongly MI problems, respectively. The latter complexity significantly improves the previously best operation complexity $\mathcal{O}\left(\varepsilon^{-2}\right)$. As a byproduct, complexity results of the primal-dual extrapolation methods are also obtained for finding an ε-KKT or ε-residual solution of convex conic optimization, conic constrained saddle point, and variational inequality problems under *local Lipschitz* continuity. We provide preliminary numerical results to demonstrate the performance of the proposed methods.

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 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 domain denoted by $\operatorname{dom} B$, and F is a monotone point-valued (i.e., point-to-point) operator on $\operatorname{cl}(\operatorname{dom} B)$. It shall be mentioned that $\operatorname{dom} B$ 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 $\operatorname{dom} B$ *with a monotonicity parameter* $\mu \geq 0$ *such that*

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

(c) F is locally Lipschitz continuous on $\operatorname{cl}(\operatorname{dom} B)$.[1]

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

*Department of Industrial and Systems Engineering, University of Minnesota, USA (email: zhaosong@umn.edu, mei00035@umn.edu). This work was partially supported by NSF Award IIS-2211491.

[1] See Section 1.1 for the definition of local Lipschitz continuity of a mapping on a closed set.

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. Moreover, it can sometimes be easily verified. For example, if F is continuously 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 [16, Proposition A.1]) and it is also strongly monotone when $\mu > 0$.

Several special cases of problem (1) have been considerably studied in the literature. For example, when F is *cocoercive*[2] problem (1) can be suitably solved by a splitting inertial proximal method [17], a Halpern fixed-point splitting method [24], and also the classical forward-backward splitting (FBS) method [10, 20] 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 [25], its variant [15], an inertial forward-backward-forward splitting method [1], and an extra anchored gradient method [7, Algorithm 3] were proposed for (1) with F being *Lipschitz continuous*. It shall be mentioned that operation complexity bounds of $\mathcal{O}\left(\varepsilon^{-2}\right)$ and $\mathcal{O}\left(\varepsilon^{-1}\right)$, measured by the number of fundamental operations consisting of evaluations of F and resolvent of B, were respectively established for the variant of MFBS method [15, Theorem 4.6] and the extra anchored gradient method [7, Theorem 2] for finding an ε-residual solution[3] of (1) with Lipschitz continuous F.

There has been little algorithmic development for solving problem (1) with locally Lipschitz continuous F. Indeed, the MFBS method [25] and the forward-reflected-backward splitting (FRBS) method [13, Algorithm 3.1] appear to be the only existing methods for solving this problem. The MFBS method modifies the classical FRBS method in the spirit of the extragradient method [5] for monotone variational inequalities, while the FRBS method modifies the forward term in the classical FBS method using an operator extrapolation technique that has been popularly used to design algorithms for solving optimization, SP, and VI problems (e.g., [3, 4, 6, 14, 21]). 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 $\{\gamma_k\}$. Global convergence to a solution of problem (1) are established for these methods in [13, 25], respectively. Moreover, it can be shown that the FRBS method enjoys an operation complexity of $\mathcal{O}(\varepsilon^{-2})$ for finding an ε-residual solution of (1) by using [13, equation (2.14), Theorem 3.4, and Lemmas 3.2 and 3.3], although this result is not established in [13]. In addition, when $B = \partial g$, where g is a proper closed convex function, an adaptive golden ratio algorithm was proposed in [12, Algorithm 1]. While [12] did not specifically study the operation complexity of this algorithm for finding an ε-residual solution of (1) with $B = \partial g$, it can be shown that the algorithm achieves an operation complexity of $\mathcal{O}(\varepsilon^{-2})$ for such a solution by using [12, equation (34) and Lemma 2].

As seen from the above discussion, there is a significant gap between the best operation complexities of $\mathcal{O}(\varepsilon^{-2})$ and $\mathcal{O}(\varepsilon^{-1})$ for finding an ε-residual solution of (1) and its special case with Lipschitz continuous F, which are achieved by the FRBS method [13] and the extra anchored gradient method [7], respectively. To significantly shorten this gap, in this paper we propose new variants of FBS method, called *primal-dual* (PD) *extrapolation* methods, for finding an ε-residual solution of (1) with complexity guarantees. In particular, we first propose a PD extrapolation method for solving a strongly MI problem, namely, problem (1) with $\mu > 0$, by modifying the forward term in the FBS method using a *point and operator extrapolation technique* that has recently been used to design algorithms for solving stochastic VI problems in [4] and problem (1) with Lipschitz continuous F in [13]. Specifically, this PD extrapolation method generates a solution sequence $\{x^k\}$ according to

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

[2] 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 $\mathrm{dom}\, F$.

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

where the sequences $\{\alpha_k\}$, $\{\beta_k\}$ and $\{\gamma_k\}$ are updated by a backtracking line search scheme (see Algorithm 1). We show that this PD extrapolation 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 PD extrapolation method for solving a non-strongly MI problem, namely, problem (1) with $\mu = 0$ by applying the above PD extrapolation 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 PD extrapolation 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 $\mu = 0$, which significantly improves the previously best operation complexity $\mathcal{O}(\varepsilon^{-2})$ achieved by the FRBS method [13].

The main contributions of our paper are summarized as follows.

- Primal-dual extrapolation methods are proposed for the MI problem (1) with locally Lipschitz continuous F, which 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 adopt a point and operator extrapolation technique with fundamental operations consisting only of evaluations of F and resolvent of B; (iii) they are equipped with a verifiable termination criterion and output an ε-residual solution of problem (1) with complexity guarantees.

- We show that an ε-residual solution of problem (1) with locally Lipschitz continuous F can be found by our methods with an operation complexity of $\mathcal{O}\left(\log\varepsilon^{-1}\right)$ and $\mathcal{O}\left(\varepsilon^{-1}\log\varepsilon^{-1}\right)$ for $\mu > 0$ and $\mu = 0$, respectively. The latter complexity significantly improves the previously best operation complexity $\mathcal{O}(\varepsilon^{-2})$ achieved by the FRBS method [13].

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

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 PD extrapolation methods for problem (1) with $\mu > 0$ and $\mu = 0$, respectively, and study their complexity. In Section 4, we study the applications of the PD extrapolation methods for solving convex conic optimization, conic constrained saddle point, and variational inequality problems. In addition, we present some preliminary numerical results and the proofs of the main results in Sections 5 and 6, respectively. Finally, we make some concluding remarks in Section 7.

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 $x, y \in \Omega$. Besides, g is called locally Lipschitz continuous on a closed set $\widehat{\Omega}$ if g is L_Ω-Lipschitz continuous on any compact set $\Omega \subseteq \widehat{\Omega}$ for some $L_\Omega > 0$. 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 a closed convex cone $\mathcal{K}$, we use $\mathcal{K}^*$ to denote the dual cone of $\mathcal{K}$, that is, $\mathcal{K}^* = \{y \in \mathbb{R}^m : \langle y, x\rangle \geq 0,\ \forall x \in \mathcal{K}\}$.

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 (1), namely, the case in which Assumption 1(b) holds with $\mu > 0$. Our method is a variant of the classical forward-backward splitting (FBS) method [10, 20]. It modifies the forward term in (3) by using a primal and dual extrapolation technique[4] that has recently been proposed to design algorithms for solving stochastic VI problems in [4]. Note that the choice of the parameters for extrapolations in [4] requires Lipschitz continuity of F. Since F is only assumed to be locally Lipschitz continuous in this paper, the choice of them in [4] is not applicable to our method. To resolve this issue, we propose a *backtracking line search scheme* to decide on parameters for extrapolations and splitting.[5] 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: $\epsilon > 0$, $\gamma_0 > 0$, $\delta \in (0, 1)$, $0 < \nu \leq 1/2$, $\eta \in [0, \nu/(1+\nu))$, and $x^0 = x^1 \in \mathrm{dom}\, B$.

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

2: Compute

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

where

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

and n_t is the smallest nonnegative integer such that

$$\|F(x^{t+1}) - F(x^t) - \eta\gamma_t^{-1}(x^{t+1} - x^t)\| \leq \nu(1-\eta)\gamma_t^{-1}\|x^{t+1} - x^t\|. \tag{6}$$

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

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

4: **end for**

Remark 1. *(i) If $\eta = 0$, Algorithm 1 is reduced to a dual extrapolation method. Besides, α_t and β_t are for primal-dual extrapolation and γ_t is the stepsize, while γ_0, δ and ν are used for backtracking line search. For the sake of generality, we provide a flexible choice for ν and η satisfying the conditions stated in the input line of Algorithm 1. Nevertheless, one can easily specify them, for example, letting $(\nu, \eta) = (0.5, 0.33)$, which appears to be the best choice for Algorithm 1 as observed in practice.*

(ii) As will be shown in Lemma 3, it holds that

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

[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 called primal and dual extrapolations just for simplicity. Accordingly, we refer to our method as a *primal-dual extrapolation method*.

[5] It shall be mentioned that backtracking line search schemes have been widely used for designing algorithms for solving MI problems (e.g., see [13, 25]).

As a result, x^{t+1} satisfying (7) implies that $\mathrm{res}_{F+B}(x^{t+1}) \le \epsilon$, namely, x^{t+1} is an ϵ-residual solution of problem (1). Thus, (7) provides a verifiable termination criterion for Algorithm 1 to find an ϵ-residual solution of (1).

(iii) As will be established below, Algorithm 1 is well-defined at each iteration. Moreover, 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.[6] We also define

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

$$\mathbb{T} - 1 = \{t - 1 : t \in \mathbb{T}\}, \quad \xi = \nu(1-\eta) - \eta, \tag{9}$$

where x^0 is the initial point, and ν and η are the input parameters 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

$$\widehat{\mathcal{S}} = \left\{x \in \mathrm{dom}\, B : \ \|x - x^*\| \le \frac{(5 + 9\gamma_0 L_{\mathcal{S}})r_0}{3\sqrt{1-2\nu^2}}\right\},^{7} \tag{10}$$

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 bounded subset in $\mathrm{dom}\, B$. By this and the local Lipschitz continuity of F on $\mathrm{cl}(\mathrm{dom}\, B)$, 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 6.

Theorem 1. *Let $\{x^t\}_{t\in\mathbb{T}}$ and $\{n_t\}_{1\le t\in\mathbb{T}-1}$ be generated by Algorithm 1 and ξ be defined in (9). 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, $\sum_{i=1}^{t} n_i \le M + t$ for all $1 \le t \in \mathbb{T} - 1$, where $\mathcal{S}$ is defined in (8) and

$$M = \left\lceil \log\left(\frac{\xi}{\gamma_0 L_{\widehat{\mathcal{S}}}}\right) / \log\delta \right\rceil_+. \tag{11}$$

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 6.

[6] For the time being, it is possible that $\mathbb{T} = \{0, 1, 2, \ldots, T\}$ or $\{0, 1, 2, \ldots\}$ for some $T \ge 0$. The reason for not presuming $\mathbb{T}$ to be a finite set here is that the finite termination of Algorithm 1 is not yet established. Nevertheless, it will be shown in Theorem 2 that $\mathbb{T}$ is a finite set.

[7] The specific choices of the radius associated with $\mathcal{S}$ and $\widehat{\mathcal{S}}$ will become clear from the proofs of Lemmas 5 and 6.

Theorem 2. *Let γ_0, δ, ν, η and ϵ be given in Algorithm 1, $L_{\mathcal{S}}$ and $L_{\widehat{\mathcal{S}}}$ be given in Lemma 1, and r_0 and ξ be defined in (8) and (9). Suppose that $\mu > 0$, i.e., $F + B$ is strongly monotone on $\operatorname{dom} B$. Then Algorithm 1 terminates and outputs an ϵ-residual solution of problem (1) in at most T iterations. Moreover, the number of evaluations of F and resolvent of B performed in Algorithm 1 is no more than N, respectively, where*

$$T = 3 + \left\lceil 2\log\left(\frac{r_0\left(8 + 12\gamma_0 L_{\mathcal{S}}\right)}{3\epsilon\sqrt{1-2\nu^2}\min\left\{L_{\widehat{\mathcal{S}}}^{-1}\delta\xi, \gamma_0\right\}}\right) \Big/ \log\left(1 + \frac{2\mu}{1-\eta}\min\left\{L_{\widehat{\mathcal{S}}}^{-1}\delta\xi, \gamma_0\right\}\right)\right\rceil_+, \tag{12}$$

$$N = 2T + \left\lceil \log\left(\frac{\xi}{\gamma_0 L_{\widehat{\mathcal{S}}}}\right) / \log\delta\right\rceil_+. \tag{13}$$

Remark 2. *(i) 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 $\operatorname{cl}(\operatorname{dom} B)$. In addition, notice that if $\gamma_0 \geq \delta\xi/L_{\widehat{\mathcal{S}}}$,*

$$\log\left(1 + \frac{2\mu}{1-\eta}\min\left\{L_{\widehat{\mathcal{S}}}^{-1}\delta\xi, \gamma_0\right\}\right) \approx \frac{2\delta\xi}{1-\eta}\cdot\frac{\mu}{L_{\widehat{\mathcal{S}}}}.$$

It then follows from (12) and (13) that if $\gamma_0 \geq \delta\xi/L_{\widehat{\mathcal{S}}}$, T and N are roughly proportional to $L_{\widehat{\mathcal{S}}}/\mu$. Hence, $L_{\widehat{\mathcal{S}}}/\mu$ can be viewed as the "condition number" of problem (1) with $\mu > 0$.

(ii) Algorithm 1 will become a linearly convergent method if setting $\epsilon = 0$. Indeed, one can observe from Lemma 5 that the sequence $\{x^k\}$ generated by Algorithm 1 with $\epsilon = 0$ satisfies $\|x^k - x^*\|^2 \leq (1-2\nu^2)^{-1}(1+2\mu\underline{\gamma})^{2-k}\|x^0 - x^*\|^2$ for all $k \geq 2$, where x^* is the solution of (1) and $\underline{\gamma} := \inf_k \gamma_k$ is a positive number due to Theorem 1.

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 (1), namely, the case in which Assumption 1(b) holds with $\mu = 0$. Our method consists of applying Algorithm 1 to approximately solve a sequence of strongly MI problems $0 \in (F_k + B)(x)$, where F_k is a perturbation of F given in (14). The proposed method is presented in Algorithm 2.

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

Input: $\varepsilon > 0$, $\gamma_0 > 0$, $z^0 \in \operatorname{dom} B$, $0 < \delta < 1$, $0 < \nu \leq 1/2$, $\eta \in [0, \nu/(1+\nu))$, $\rho_0 \geq 1$, $0 < \tau_0 \leq 1$, $\zeta > 1$, $0 < \sigma < 1/\zeta$, $\rho_k = \rho_0\zeta^k$, $\tau_k = \tau_0\sigma^k$ for all $k \geq 0$.

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, η, δ and ν, and output z^{k+1}, where

$$F_k(x) = F(x) + \rho_k^{-1}(x - z^k) \quad \forall x \in \operatorname{dom} F. \tag{14}$$

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

$$\rho_k^{-1}\|z^{k+1} - z^k\| + \tau_k \leq \varepsilon. \tag{15}$$

4: **end for**

Remark 3. *(i) In Algorithm 2, the parameters γ_0, δ, ν and η have the same meaning as those for Algorithm 1 (see Remark 1(i)). Besides, ρ_0, τ_0, ζ and σ are used for subproblem regularization and subproblem termination criterion.*

(ii) It is easy to see that Algorithm 2 is well-defined at each iteration and equipped with a verifiable termination criterion, while it shares the same fundamental operations as Algorithm 1, consisting 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.[8] We also define $\mathbb{K}-1=\{k-1: k\in\mathbb{K}\}$, and

$$\bar{r}_0=\|z^0-x^*\|,\quad \mathcal{Q}=\left\{x\in\operatorname{dom}B:\ \|x-x^*\|\le\left(\frac{1}{\sqrt{1-2\nu^2}}+1\right)\left(\bar{r}_0+\frac{\rho_0\tau_0}{1-\sigma\zeta}\right)\right\}, \tag{16}$$

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* (14) *and* (16)*. 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 of k.

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

$$\widehat{\mathcal{Q}}=\left\{x\in\operatorname{dom}B:\ \|x-x^*\|\le\left(\frac{5+9\gamma_0L_{\mathcal{Q}}}{3\sqrt{1-2\nu^2}}+1\right)\left(\bar{r}_0+\frac{\rho_0\tau_0}{1-\sigma\zeta}\right)\right\}.^{9} \tag{17}$$

Proof. Notice that $\mathcal{Q}$ is a bounded subset in $\operatorname{dom}B$. By this and the local Lipschitz continuity of F on $\operatorname{cl}(\operatorname{dom}B)$, 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 $k\ge0$. Using these and (14), 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 6.

Theorem 3. *Let γ_0, δ, ν, η, ζ, σ, ρ_0, τ_0 and ε be given in Algorithm 2, $L_{\mathcal{Q}}$ and $L_{\widehat{\mathcal{Q}}}$ be given in Lemma 2, ξ and $\bar{r}_0$ be defined in* (9) *and* (16)*, and*

$$\Lambda=\frac{\rho_0\tau_0}{1-\sigma\zeta},\quad C_1=\log\left(\frac{(\bar{r}_0+\Lambda)\,(8+12\gamma_0L_{\mathcal{Q}})}{3\tau_0\sqrt{1-2\nu^2}\min\left\{L_{\widehat{\mathcal{Q}}}^{-1}\delta\xi,\gamma_0\right\}}\right), \tag{18}$$

$$C_2=\left\lceil\log\left(\frac{\xi}{\gamma_0L_{\widehat{\mathcal{Q}}}}\right)/\log\delta\right\rceil_+,\quad C_3=\frac{1}{(\zeta-1)\log\left(1+\frac{2}{\rho_0(1-\eta)}\min\left\{L_{\widehat{\mathcal{Q}}}^{-1}\delta\xi,\gamma_0\right\}\right)}. \tag{19}$$

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

$$K=\left\lceil\max\left\{\log_\zeta\left(\frac{2\bar{r}_0+2\Lambda}{\varepsilon\rho_0}\right),\frac{\log(2\tau_0/\varepsilon)}{\log(1/\sigma)}\right\}\right\rceil_+, \tag{20}$$

and

$$\begin{aligned}\bar{M}=\ &8+C_2+(8+C_2)K+4\zeta(C_1)_+C_3\max\left\{\frac{2\zeta(\bar{r}_0+\Lambda)}{\varepsilon\rho_0},\zeta\left(\frac{2\tau_0}{\varepsilon}\right)^{\frac{\log\zeta}{\log(1/\sigma)}},1\right\}\\&+4\zeta C_3(\log\sigma^{-1})K\max\left\{\frac{2\zeta(\bar{r}_0+\Lambda)}{\varepsilon\rho_0},\zeta\left(\frac{2\tau_0}{\varepsilon}\right)^{\frac{\log\zeta}{\log(1/\sigma)}},1\right\}.\end{aligned} \tag{21}$$

[8] For the time being, it is possible that $\mathbb{K}=\{0,1,2,\ldots,K\}$ or $\{0,1,2,\ldots\}$ for some $K\ge0$. The reason for not presuming $\mathbb{K}$ to be a finite set is that the finite termination of Algorithm 2 is not yet established. Nevertheless, it will be shown in Theorem 3 that $\mathbb{K}$ is a finite set.

[9] The specific choices of the radius associated with $\mathcal{Q}$ and $\widehat{\mathcal{Q}}$ will become clear from the proof of Lemma 10.

Remark 4. *(i) Since $1 < \zeta < 1/\sigma$ and $K = \mathcal{O}\left(\log \varepsilon^{-1}\right)$, 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 $\mathrm{cl}(\mathrm{dom}\, B)$. The latter complexity significantly improves the previously best operation complexity $\mathcal{O}(\varepsilon^{-2})$ achieved by the FRBS method [13]. In addition, notice that if $\gamma_0 \geq \delta\xi/L_{\widehat{\mathcal{Q}}}$,*

$$\log\left(1 + \frac{2}{\rho_0(1-\eta)} \min\left\{L_{\widehat{\mathcal{Q}}}^{-1}\delta\xi, \gamma_0\right\}\right) \approx \frac{2\delta\xi}{\rho_0(1-\eta)} L_{\widehat{\mathcal{Q}}}^{-1}.$$

It then follows from (19) and (21) that if $\gamma_0 \geq \delta\xi/L_{\widehat{\mathcal{Q}}}$, $\bar{M}$ is roughly proportional to $L_{\widehat{\mathcal{Q}}}$. Hence, $L_{\widehat{\mathcal{Q}}}$ can be viewed as the "Lipschitz constant" of problem (1) with $\mu = 0$.

(ii) Algorithm 2 will become a globally convergent method if setting $\varepsilon = 0$. Indeed, one can observe from Lemma 8 that the sequence $\{z^k\}$ generated by Algorithm 2 with $\varepsilon = 0$ satisfies $\|z^k - (I + \rho_k(F+B))^{-1}(z^k)\| \leq \rho_k\tau_k$ for all $k \geq 0$, where $0 < \rho_k \to \infty$ and $\sum_k \rho_k\tau_k < \infty$. Besides, one can see from Lemma 9 that $\{z^k\}$ is bounded. It then follows from [22, Theorem 1] that the sequence $\{z^k\}$ converges to a solution of (1).

(iii) While Algorithm 2 is proposed to solve problem (1) with $\mu = 0$, it is also applicable to (1) with $\mu > 0$. Similar to the proof of Theorem 3, it can be shown that Algorithm 2 achieves an operation complexity of $\mathcal{O}((\log \varepsilon^{-1})^2)$ for finding and ε-residual solution of problem (1) with $\mu > 0$. This complexity is at most worse by a logarithmic factor compared to the complexity achieved by directly calling Algorithm 1.

4 Applications

In this section we study applications of our PD extrapolation 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

$$\begin{array}{ll} \min & f(x) + P(x) \\ \text{s.t.} & -g(x) \in \mathcal{K}, \end{array} \tag{22}$$

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,

$$\vartheta g(x) + (1-\vartheta)g(y) - g(\vartheta x + (1-\vartheta)y) \in \mathcal{K} \qquad \forall x, y \in \mathbb{R}^n,\ \vartheta \in [0,1]. \tag{23}$$

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

Problem (22) 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 (22) reduces to an ordinary convex optimization problem

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

We make the following additional assumptions for problem (22).

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 (22) and its Lagrangian dual problem

$$\sup_{\lambda\in\mathcal{K}^*}\inf_{x}\{f(x)+P(x)+\langle\lambda,g(x)\rangle\} \tag{24}$$

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 (22) and (24) if and only if it satisfies the Karush-Kuhn-Tucker (KKT) condition

$$0\in\begin{pmatrix}\nabla f(x)+\nabla g(x)\lambda+\partial P(x)\\ -g(x)+\mathcal{N}_{\mathcal{K}^*}(\lambda)\end{pmatrix}. \tag{25}$$

In general, it is difficult to find an exact optimal solution of (22) and (24). Instead, for any given $\varepsilon>0$, we are interested in finding a pair of ε-KKT solutions (x,λ) of (22) and (24) that satisfies

$$\operatorname{dist}(0,\nabla f(x)+\nabla g(x)\lambda+\partial P(x))\le\varepsilon,\quad \operatorname{dist}(0,-g(x)+\mathcal{N}_{\mathcal{K}^*}(\lambda))\le\varepsilon. \tag{26}$$

Observe from (25) that problems (22) and (24) can be solved as the MI problem

$$0\in F(x,\lambda)+B(x,\lambda), \tag{27}$$

where

$$F(x,\lambda)=\begin{pmatrix}\nabla f(x)+\nabla g(x)\lambda\\ -g(x)\end{pmatrix},\quad B(x,\lambda)=\begin{pmatrix}\partial P(x)\\ \mathcal{N}_{\mathcal{K}^*}(\lambda)\end{pmatrix}. \tag{28}$$

Notice that $\lambda\in\mathcal{K}^*$ and g is $\mathcal{K}$-convex in the sense that (23) holds, which imply that $\langle\lambda,g(x)\rangle$ is convex in x. Based on this and the above assumptions, one can observe that $f(x)+\langle\lambda,g(x)\rangle$ is convex in x and concave in λ on $\operatorname{cl}(\operatorname{dom}B)$, which implies that F is monotone on $\operatorname{cl}(\operatorname{dom}B)$. One can also observe that F is locally Lipschitz continuous on $\operatorname{cl}(\operatorname{dom}B)$ and B is maximal monotone. As a result, Algorithm 2 can be suitably applied to the MI problem (27). It then follows from Theorem 3 that Algorithm 2, when applied to problem (27), finds an ε-residual solution (x,λ) of (27) within $\mathcal{O}\left(\varepsilon^{-1}\log\varepsilon^{-1}\right)$ evaluations of F and resolvent of B. Notice from (26) and (28) that such (x,λ) is also a pair of ε-KKT solutions of (22) and (24). In addition, the evaluation of F requires that of ∇f and ∇g, and also the resolvent of B can be computed as

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

The above discussion leads to the following result regarding Algorithm 2 for finding a pair of ε-KKT solutions of problems (22) and (24).

Theorem 4. *For any $\varepsilon>0$, Algorithm 2, when applied to the MI problem (27), outputs a pair of ε-KKT solutions of problems (22) and (24) within $\mathcal{O}\left(\varepsilon^{-1}\log\varepsilon^{-1}\right)$ evaluations of ∇f, ∇g, $\operatorname{prox}_{\gamma P}$ 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 (22) 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 [11] for finding a pair of ε-KKT solutions of a subclass of problems (22) and (24), which also requires $\mathcal{O}\left(\varepsilon^{-1}\log\varepsilon^{-1}\right)$ evaluations of ∇f, ∇g, $\operatorname{prox}_{\gamma P}$ and $\Pi_{\mathcal{K}^*}$. However, this method and its complexity analysis require that ∇f and ∇g be Lipschitz continuous on an open set containing $\operatorname{dom}P$ and also that $\operatorname{dom}P$ be bounded. As a result, it is generally not applicable to problem (22).

(iii) A variant of Tseng's MFBS method was proposed in [16, Section 6] for finding a pair of ε-KKT solutions of a special class of problems (22) and (24), where 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 (22). Additionally, this method has an operation complexity of $\mathcal{O}\left(\varepsilon^{-2}\right)$ (see [16, Theorem 6.3]). In contrast, our method achieves a significantly better operation complexity of of $\mathcal{O}\left(\varepsilon^{-1}\log\varepsilon^{-1}\right)$. Furthermore, an adaptive proximal algorithm was recently proposed in [8, Section 3.1] for solving a special case of problem* (22) *where g is an affine mapping. It has been shown in [8, Theorem3.4] that the iterates of this algorithm converges to a KKT solution of the problem.*

4.2 Conic constrained saddle point problems

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

$$\min_{-g(x)\in\mathcal{K}} \max_{-\tilde{g}(y)\in\tilde{\mathcal{K}}} \{\Psi(x,y) := f(x,y) + P(x) - \tilde{P}(y)\}, \tag{29}$$

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 (23), 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 (29).

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 mappings 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 (29) includes a rich class of saddle point problems as special cases. Several of them have been studied in the literature. For example, extragradient method [5], mirror-prox method [18], dual extrapolation method [19], and accelerated proximal point method [9] were developed for solving the special CCSP problem

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \tilde{f}(x,y), \tag{30}$$

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 [14] and extra anchored gradient method [29] were proposed for solving problem (30) with $\mathcal{X} = \mathbb{R}^n$ and $\mathcal{Y} = \mathbb{R}^m$. In addition, accelerated proximal gradient method [26], a variant of MFBS method [16], and also generalized extragradient method [16] were proposed for solving the special CCSP problem

$$\min_x \max_y \left\{\tilde{f}(x,y) + P(x) - \tilde{P}(y)\right\}, \tag{31}$$

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 (30) or (31) with a strongly-convex-(strongly)-concave $\tilde{f}$ (e.g., see [9, 28, 27]). Recently, extra-gradient method of multipliers [30] was proposed for solving a special case of problem (29) 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 [29] 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 [5, 19, 18, 26, 16, 14, 9, 28, 27, 29] are generally not suitable for solving the CCSP problem (29). We next apply our Algorithm 2 to find an ε-KKT solution of (29) 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 (29) if and only if it together with some $(\lambda, \tilde{\lambda})$ satisfies the KKT condition

$$0 \in \begin{pmatrix} \nabla_x f(x,y) + \nabla g(x)\lambda + \partial P(x) \\ -\nabla_y f(x,y) + \nabla \tilde{g}(y)\tilde{\lambda} + \partial \tilde{P}(y) \\ -g(x) + \mathcal{N}_{\mathcal{K}^*}(\lambda) \\ -\tilde{g}(y) + \mathcal{N}_{\tilde{\mathcal{K}}^*}(\tilde{\lambda}) \end{pmatrix}. \tag{32}$$

Generally, it is difficult to find a pair of exact optimal minimax solutions of (29). Instead, for any given $\varepsilon > 0$, we are interested in finding an ε-KKT solution $(x, y, \lambda, \tilde{\lambda})$ of (29) that satisfies

$$\mathrm{dist}(0, \nabla_x f(x,y) + \nabla g(x)\lambda + \partial P(x)) \le \varepsilon, \quad \mathrm{dist}(0, -\nabla_y f(x,y) + \nabla \tilde{g}(y)\tilde{\lambda} + \partial \tilde{P}(y)) \le \varepsilon, \tag{33}$$

$$\mathrm{dist}(0, -g(x) + \mathcal{N}_{\mathcal{K}^*}(\lambda)) \le \varepsilon, \quad \mathrm{dist}(0, -\tilde{g}(y) + \mathcal{N}_{\tilde{\mathcal{K}}^*}(\tilde{\lambda})) \le \varepsilon. \tag{34}$$

Observe from (32) that problem (29) can be solved as the MI problem

$$0 \in F(x, y, \lambda, \tilde{\lambda}) + B(x, y, \lambda, \tilde{\lambda}), \tag{35}$$

where

$$F(x, y, \lambda, \tilde{\lambda}) = \begin{pmatrix} \nabla_x f(x,y) + \nabla g(x)\lambda \\ -\nabla_y f(x,y) + \nabla \tilde{g}(y)\tilde{\lambda} \\ -g(x) \\ -\tilde{g}(y) \end{pmatrix}, \quad B(x, y, \lambda, \tilde{\lambda}) = \begin{pmatrix} \partial P(x) \\ \partial \tilde{P}(y) \\ \mathcal{N}_{\mathcal{K}^*}(\lambda) \\ \mathcal{N}_{\tilde{\mathcal{K}}^*}(\tilde{\lambda}) \end{pmatrix}. \tag{36}$$

Notice that $\lambda \in \mathcal{K}^*$, $\tilde{\lambda} \in \tilde{\mathcal{K}}^*$, and g and $\tilde{g}$ are respectively $\mathcal{K}$- and $\tilde{\mathcal{K}}$-convex in the sense of (23), which imply that $\langle \lambda, g(x) \rangle$ and $\langle \tilde{\lambda}, \tilde{g}(y) \rangle$ are convex in x and y, respectively. Based on this and the above assumptions, one can observe that $f(x,y) + \langle \lambda, g(x) \rangle - \langle \tilde{\lambda}, \tilde{g}(y) \rangle$ is convex in $(x, \tilde{\lambda})$ and concave in (y, λ) on $\mathrm{cl}(\mathrm{dom}\, B)$, which implies that F is monotone on $\mathrm{cl}(\mathrm{dom}\, B)$. One can also observe that F is 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 (35). It then follows from Theorem 3 that Algorithm 2, when applied to problem (35), finds an ε-residual solution $(x, y, \lambda, \tilde{\lambda})$ of (35) within $\mathcal{O}\left(\varepsilon^{-1} \log \varepsilon^{-1}\right)$ evaluations of F and resolvent of B. Notice from (33), (34) and (36) that such $(x, y, \lambda, \tilde{\lambda})$ is also an ε-KKT solution of problem (29). 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 (29).

Theorem 5. *For any $\varepsilon > 0$, Algorithm 2, when applied to the MI problem (35), outputs an ε-KKT solution of problem (29) 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 (29). 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:

$$\text{find } x \in \mathbb{R}^n \text{ such that } g(y) - g(x) + \langle y - x, F(x)\rangle \geq 0 \ \forall y \in \mathbb{R}^n, \tag{37}$$

where $g : \mathbb{R}^n \to (-\infty, \infty]$ is a proper closed convex function, and $F : \operatorname{dom} F \to \mathbb{R}^n$ is monotone and *locally Lipschitz continuous* on $\operatorname{cl}(\operatorname{dom} \partial g) \subseteq \operatorname{dom} F$. It shall be mentioned that $\operatorname{dom} g$ is possibly *unbounded*. Assume that problem (37) has at least one solution. For the details of VI and its applications, we refer the reader to [2] and the references therein.

Recently, an adaptive golden ratio algorithm was proposed in [12, Algorithm 1] for solving (37). In addition, some special cases of (37) have been well studied in the literature. For example, projection method [23], extragradient method [5], mirror-prox method [18], dual extrapolation method [19], operator extrapolation method [6], extra-point method [3, 4], and extra-momentum method [4] were developed for solving problem (37) 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 [16], and generalized extragradient method [16] were proposed for solving problem (37) with F being *Lipschitz continuous*. Iteration complexity of these methods except [6] 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 $\operatorname{cl}(\operatorname{dom} g)$ in our paper, these methods are generally not suitable for solving problem (37).

Generally, it is difficult to find an exact solution of problem (37). Instead, for any given $\varepsilon > 0$, we are interested in finding an ε-residual solution of (37), which is a point x satisfying $\operatorname{res}_{F+\partial g}(x) \leq \varepsilon$. To this end, we first observe that problem (37) is equivalent to the MI problem

$$0 \in (F + \partial g)(x). \tag{38}$$

Since F is monotone and locally Lipschitz continuous on $\operatorname{cl}(\operatorname{dom} \partial g)$ and ∂g is maximal monotone, Algorithm 2 can be suitably applied to the MI problem (38). It then follows from Theorem 3 that Algorithm 2, when applied to problem (38), finds an ε-residual solution x of (38), which is indeed also an ε-residual solution of (37), 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

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

The above discussion leads to the following result regarding Algorithm 2 for finding an ε-residual solution of problem (37).

Theorem 6. *For any $\varepsilon > 0$, Algorithm 2, when applied to the MI problem (38), outputs an ε-residual solution of problem (37) within $\mathcal{O}\left(\varepsilon^{-1} \log \varepsilon^{-1}\right)$ evaluations of F and $\operatorname{prox}_{\gamma g}$ for some $\gamma > 0$.*

Remark 7. *An adaptive golden ratio algorithm was recently proposed in [12, Algorithm 1]. While [12] did not specifically study the operation complexity of this algorithm for finding an ε-residual solution of (37), it can be shown that the algorithm achieves an operation complexity of $\mathcal{O}(\varepsilon^{-2})$ for such a solution by using [12, equation (34) and Lemma 2]. In contrast, the operation complexity of $\mathcal{O}\left(\varepsilon^{-1} \log \varepsilon^{-1}\right)$ achieved by our method is significantly better.*

5 Numerical results

In this section we conduct some preliminary experiments to test the performance of our proposed method (Algorithm 2), and compare it with FRBS method [13], MFBS method with an Armijo-Goldstein-type stepsize [25], and adaptive golden ratio (AGR) algorithm [12], respectively. All the methods are coded in Matlab and all the computations are performed on a desktop with a 3.60 GHz Intel i7-12700K 12-core processor and 32 GB of RAM.

We consider the problem

$$\min_{x\geq 0}\max_{\|y\|\leq 1}\Big\{\|Ax-b\|_4^4+\langle Bx,y\rangle-\|Cy-d\|_4^4\Big\},\tag{39}$$

where $A\in\mathbb{R}^{l\times n}$, $B\in\mathbb{R}^{m\times n}$, $C\in\mathbb{R}^{q\times m}$, $b\in\mathbb{R}^l$, $d\in\mathbb{R}^q$, and $\|z\|_4=(\sum_i z_i^4)^{1/4}$ for any vector z.

We randomly generate instances for problem (39). Specifically, we first randomly generate $U\in\mathbb{R}^{l\times(n/10)}$ and $V\in\mathbb{R}^{(n/10)\times n}$ with all the entries independently chosen from a normal distribution with mean 0 and standard deviation 0.1, and a diagonal matrix $D\in\mathbb{R}^{(n/10)\times(n/10)}$ with all the diagonal entries independently chosen from a uniform distribution between 0 and 1. Then we set $A=UDV$. In a similar vein, we randomly generate C. Besides, we randomly generate $P\in\mathbb{R}^{m\times l}$ with all the entries independently chosen from the standard normal distribution, and set $B=PA$. In addition, we randomly generate b and d with all the entries independently chosen from the standard normal distribution.

Notice that problem (39) is a special case of (29). As discussed in Subsection 4.2, (39) is equivalent to the monotone inclusion problem (1) with

$$F(x,y)=\begin{pmatrix}\nabla(\|Ax-b\|_4^4)+B^Ty\\ \nabla(\|Cy-d\|_4^4)-Bx\end{pmatrix},\qquad B(x,y)=\begin{pmatrix}\mathcal{N}_{\mathbb{R}^n_+}(x)\\ \mathcal{N}_{\mathcal{B}}(y)\end{pmatrix},$$

where $\mathcal{B}=\{z\in\mathbb{R}^m:\|z\|\leq 1\}$. Clearly, B is a maximal monotone operator and F is monotone and locally Lipschitz continuous, albeit *not globally* Lipschitz continuous. In addition, for any $\gamma>0$, the resolvent of γB can be calculated as

$$(I+\gamma B)^{-1}\begin{pmatrix}x\\ y\end{pmatrix}=\begin{pmatrix}\Pi_{\mathbb{R}^n_+}(x)\\ \Pi_{\mathcal{B}}(y)\end{pmatrix}.$$

As a result, (39) can be suitably solved by Algorithm 2, FRBS [13], MFBS [25], and AGR [12]. Our aim is to find a 10^{-4}-residual solution of the corresponding monotone inclusion problem of (39) for the above instances by using Algorithm 2, FRBS, MFBS and AGR, and compare their performance. Due to this, we terminate them once a 10^{-4}-residual solution is found. In addition, for all the methods, we choose 0 as the initial point and set the parameters as

- $(\varepsilon,\gamma_0,\delta,\nu,\rho_0,\tau_0,\zeta,\sigma,\eta)=(10^{-4},0.1,0.9,0.5,10,0.09,9,0.1,0.33)$ for Algorithm 2;
- $(\lambda_0,\delta,\sigma)=(0.1,0.5,0.9)$ for FRBS [13];
- $(\sigma,\theta,\beta)=(0.1,0.5,0.9)$ for MFBS [25].
- $(\lambda_0,\bar{\lambda},\phi)=(1,1,1.5)$ for AGR [12].

The computational results of Algorithm 2, FRBS, MFBS and AGR for the instances randomly generated above are presented in Table 1. In detail, the value of (n,m,l,q) is listed in the first four columns. For each instance, the number of gradient evaluations and the CPU time (in seconds) are given in the rest of the columns. One can observe that our method, namely Algorithm 2, substantially outperforms the other three methods in terms of number of gradient evaluations and CPU time. Notice that our method uses both primal and dual extrapolation schemes, while FRBS only uses a dual extrapolation scheme, and MFBS and AGR do not use any of them. The numerical results in Table 1 demonstrate that primal and dual extrapolation schemes have an acceleration effect.

				Gradient evaluations				CPU time (seconds)			
n	m	l	q	Algorithm 2	FRBS	MFBS	AGR	Algorithm 2	FRBS	MFBS	AGR
100	10	500	100	1.23×10^3	3.12×10^3	3.08×10^3	2.12×10^3	0.3	0.5	0.4	0.3
200	20	1000	200	2.81×10^3	8.15×10^3	1.20×10^4	7.41×10^3	1.4	6.2	6.3	3.7
300	30	1500	300	2.28×10^4	6.25×10^4	8.84×10^4	4.27×10^4	23.3	82.3	114.3	57.3
400	40	2000	400	7.62×10^4	1.78×10^5	3.84×10^5	1.24×10^5	109.4	314.4	598.4	231.9
500	50	2500	500	5.93×10^4	1.68×10^5	5.95×10^5	1.65×10^5	149.2	388.5	1741.8	464.4
600	60	3000	600	6.00×10^4	1.70×10^5	4.28×10^5	1.71×10^5	238.4	549.0	1588.8	561.7
700	70	3500	700	6.90×10^4	1.54×10^5	4.71×10^5	1.37×10^5	268.3	639.5	2005.9	596.7
800	80	4000	800	4.62×10^4	8.52×10^4	4.04×10^5	8.52×10^4	271.6	565.9	2363.5	577.5
900	90	4500	900	5.43×10^4	9.33×10^4	5.17×10^5	8.27×10^4	324.3	594.3	3459.8	562.8
1000	100	5000	1000	3.32×10^4	6.13×10^4	5.37×10^5	6.67×10^4	380.5	784.2	6206.5	854.8

Table 1: Numerical results for problem (39)

6 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.

6.1 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 $1 \leq t \in \mathbb{T}$, we define

$$\Delta^t = F(x^t) - F(x^{t-1}), \tag{40}$$

$$\tilde{\Delta}^t = \Delta^t - \eta\gamma_{t-1}^{-1}(x^t - x^{t-1}). \tag{41}$$

In addition, we define

$$v^t = \gamma_t^{-1}(x^t - x^{t+1} + \alpha_t(x^t - x^{t-1}) + \gamma_t\Delta^{t+1} - \gamma_t\beta_t\Delta^t) \qquad \forall 1 \leq t \in \mathbb{T} - 1, \tag{42}$$

$$\tilde{\gamma}_t = \frac{\gamma_t}{1-\eta}, \qquad \theta_t = \prod_{i=1}^{t-1}(1 + 2\mu\tilde{\gamma}_i) = \prod_{i=1}^{t-1}\left(1 + \frac{2\mu\gamma_i}{1-\eta}\right) \quad \forall 1 \leq t \in \mathbb{T} - 1.^{10} \tag{43}$$

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

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

$$v^t \in (F + B)(x^{t+1}), \tag{44}$$

$$v^t = \tilde{\gamma}_t^{-1}(x^t - x^{t+1} + \tilde{\gamma}_t\tilde{\Delta}^{t+1} - \tilde{\gamma}_t\beta_t\tilde{\Delta}^t). \tag{45}$$

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 (40) and (42) yields

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

and hence (44) holds. In addition, recall from (5) that $\alpha_t = \eta\gamma_t\beta_t/\gamma_{t-1}$. By this, (41) and (43), one has

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

Hence, (45) holds as desired. □

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

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

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*

$$\frac{1}{2}\theta_1\|x^0-x^*\|^2-\frac{1}{2}(1+2\mu\tilde{\gamma}_k)\theta_k\|x^{k+1}-x^*\|^2 \ge -\tilde{\gamma}_k\theta_k\langle\tilde{\Delta}^{k+1},x^{k+1}-x^*\rangle+R_k, \tag{46}$$

where

$$R_k=\sum_{t=1}^{k}\left(\tilde{\gamma}_t\beta_t\theta_t\langle\tilde{\Delta}^t,x^{t+1}-x^t\rangle+\frac{1}{2}\theta_t\|x^{t+1}-x^t\|^2\right). \tag{47}$$

Proof. By (2), (44) 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 (45) 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

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

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

$$\begin{aligned}&\sum_{t=1}^{k}\left(\frac{1}{2}\theta_t\|x^t-x^*\|^2-\frac{1}{2}(1+2\tilde{\gamma}_t\mu)\theta_t\|x^{t+1}-x^*\|^2\right)\\ &\ge\sum_{t=1}^{k}\theta_t\tilde{\gamma}_t\beta_t\langle\tilde{\Delta}^t,x^t-x^*\rangle-\sum_{t=1}^{k}\theta_t\tilde{\gamma}_t\langle\tilde{\Delta}^{t+1},x^{t+1}-x^*\rangle+R_k\\ &=\theta_1\tilde{\gamma}_1\beta_1\langle\tilde{\Delta}^1,x^1-x^*\rangle+\sum_{t=1}^{k-1}(\theta_{t+1}\tilde{\gamma}_{t+1}\beta_{t+1}-\theta_t\tilde{\gamma}_t)\langle\tilde{\Delta}^{t+1},x^{t+1}-x^*\rangle-\tilde{\gamma}_k\theta_k\langle\tilde{\Delta}^{k+1},x^{k+1}-x^*\rangle+R_k.\end{aligned} \tag{48}$$

In addition, it follows from $x^0=x^1$ and (41) that $\tilde{\Delta}^1=0$. Also, by the definition of $\tilde{\gamma}_t$, θ_t and β_t in (5) and (43), one has

$$\begin{aligned}\theta_{t+1}\tilde{\gamma}_{t+1}\beta_{t+1}-\theta_t\tilde{\gamma}_t=\theta_t\left(1+\frac{2\mu\gamma_t}{1-\eta}\right)\cdot\frac{\gamma_{t+1}}{1-\eta}\cdot\frac{\gamma_t}{\gamma_{t+1}}\left(1+\frac{2\mu\gamma_t}{1-\eta}\right)^{-1}-\theta_t\frac{\gamma_t}{1-\eta}=0,\\ \theta_{t+1}=(1+2\tilde{\gamma}_t\mu)\theta_t.\end{aligned} \tag{49}$$

Using these, $\tilde{\Delta}^1=0$ and (48), 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 (49) with $t=k$. □

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

$$\|x^{k+1}-x^*\|^2 \le \frac{1}{(1-2\nu^2)\theta_k}\|x^0-x^*\|^2 \qquad \forall 1 \le k \in \mathbb{T}-1. \tag{50}$$

Proof. By the definition of β_t and $\tilde{\gamma}_t$ in (5) and (43), one has $\tilde{\gamma}_{t-1}^{-1}\tilde{\gamma}_t\beta_t = (1+2\mu\gamma_{t-1}/(1-\eta))^{-1}$. Using this and the definition of θ_t in (43), we obtain

$$\begin{aligned}\theta_{t-1} - 4\nu^2\tilde{\gamma}_{t-1}^{-2}\tilde{\gamma}_t^2\beta_t^2\theta_t &= \theta_{t-1}\left(1-4\nu^2\left(1+\frac{2\mu\gamma_{t-1}}{1-\eta}\right)^{-2}\frac{\theta_t}{\theta_{t-1}}\right)\\ &\overset{(43)}{=} \theta_{t-1}\left(1-4\nu^2\left(1+\frac{2\mu\gamma_{t-1}}{1-\eta}\right)^{-1}\right) \ge 0,\end{aligned} \tag{51}$$

where the last inequality follows from the fact that $0<\nu\le 1/2$. In addition, it follows from (6), (40), (41), and the definition of $\tilde{\gamma}_t$ in (43) that

$$\|\tilde{\Delta}^t\| \le \nu\tilde{\gamma}_{t-1}^{-1}\|x^t-x^{t-1}\| \qquad \forall 2 \le t \in \mathbb{T}. \tag{52}$$

Recall that R_k is defined in (47). Letting $\theta_0=0$, and using (47), (51), (52) and $x^0=x^1$, we have

$$\begin{aligned}R_k &\overset{(47)}{\ge} \sum_{t=1}^{k}\left(-\tilde{\gamma}_t\beta_t\theta_t\|\tilde{\Delta}^t\|\|x^{t+1}-x^t\| + \frac{1}{2}\theta_t\|x^{t+1}-x^t\|^2\right)\\
&\overset{(52)}{\ge} \sum_{t=1}^{k}\left(-\nu\tilde{\gamma}_{t-1}^{-1}\tilde{\gamma}_t\beta_t\theta_t\|x^t-x^{t-1}\|\|x^{t+1}-x^t\| + \frac{1}{2}\theta_t\|x^{t+1}-x^t\|^2\right)\\
&= \sum_{t=1}^{k}\left(-\nu\tilde{\gamma}_{t-1}^{-1}\tilde{\gamma}_t\beta_t\theta_t\|x^t-x^{t-1}\|\|x^{t+1}-x^t\| + \frac{1}{4}\theta_t\|x^{t+1}-x^t\|^2 + \frac{1}{4}\theta_{t-1}\|x^t-x^{t-1}\|^2\right)\\
&\quad + \frac{1}{4}\theta_k\|x^{k+1}-x^k\|^2\\
&\ge \sum_{t=1}^{k}\left(\left(\sqrt{\theta_t\theta_{t-1}}/2 - \nu\tilde{\gamma}_{t-1}^{-1}\tilde{\gamma}_t\beta_t\theta_t\right)\|x^t-x^{t-1}\|\|x^{t+1}-x^t\|\right) + \frac{1}{4}\theta_k\|x^{k+1}-x^k\|^2\\
&\overset{(51)}{\ge} \frac{1}{4}\theta_k\|x^{k+1}-x^k\|^2.\end{aligned}$$

Using this, (46) and (52), we further obtain

$$\begin{aligned}\frac{1}{2}\theta_1\|x^0-x^*\|^2 - \frac{1}{2}(1+2\mu\tilde{\gamma}_k)\theta_k\|x^{k+1}-x^*\|^2 &\ge -\tilde{\gamma}_k\theta_k\langle\tilde{\Delta}^{k+1}, x^{k+1}-x^*\rangle + \frac{1}{4}\theta_k\|x^{k+1}-x^k\|^2\\
&\ge -\tilde{\gamma}_k\theta_k\|\tilde{\Delta}^{k+1}\|\|x^{k+1}-x^*\| + \frac{1}{4}\theta_k\|x^{k+1}-x^k\|^2\\
&\overset{(52)}{\ge} -\nu\theta_k\|x^{k+1}-x^k\|\|x^{k+1}-x^*\| + \frac{1}{4}\theta_k\|x^{k+1}-x^k\|^2\\
&\ge -\nu^2\theta_k\|x^{k+1}-x^*\|^2.\end{aligned}$$

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

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

□

In what follows, we will show that $\{n_t\}_{1\le t\in\mathbb{T}-1}$ 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

$$x^{t+1}(\gamma) = (I+\gamma B)^{-1}\left(x^t + \alpha_t(\gamma)(x^t-x^{t-1}) - \gamma\left(F(x^t)+\beta_t(\gamma)(F(x^t)-F(x^{t-1}))\right)\right) \quad \forall\gamma>0, \tag{53}$$

$$V_{t+1}(\gamma) = \|F(x^{t+1}(\gamma)) - F(x^t) - \eta\gamma^{-1}(x^{t+1}(\gamma)-x^t)\| \quad \forall\gamma>0, \tag{54}$$

where

$$\beta_t(\gamma) = \frac{\gamma_{t-1}}{\gamma}\left(1 + \frac{2\mu\gamma_{t-1}}{1-\eta}\right)^{-1}, \quad \alpha_t(\gamma) = \frac{\eta\gamma\beta_t(\gamma)}{\gamma_{t-1}}. \tag{55}$$

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* (10). *Assume that $x^t, x^{t-1} \in \mathcal{S}$ for some $1 \le t \in \mathbb{T} - 1$. Then $x^{t+1}(\gamma) \in \widehat{\mathcal{S}}$ for any $0 < \gamma \le \gamma_0$.*

Proof. Fix any $\gamma \in (0, \gamma_0]$. It follows from (53) that

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

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

$$\langle x^t - x^{t+1}(\gamma) + w, x^{t+1}(\gamma) - x^* \rangle \ge 0, \tag{56}$$

where

$$w = \alpha_t(\gamma)(x^t - x^{t-1}) - \gamma(F(x^t) - F(x^*)) - \gamma\beta_t(\gamma)(F(x^t) - F(x^{t-1})). \tag{57}$$

It follows from (56) that

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

which implies that

$$\|x^{t+1}(\gamma) - x^*\| \le \|x^t - x^* + w\| \le \|x^t - x^*\| + \|w\|. \tag{58}$$

Notice from Algorithm 1 that $0 < \gamma_{t-1} \le \gamma_0$ and $0 \le \eta < 1/3$. Using these and (55), we have

$$\gamma\beta_t(\gamma) = \gamma_{t-1}\left(1 + \frac{2\mu\gamma_{t-1}}{1-\eta}\right)^{-1} \le \gamma_0, \tag{59}$$

$$\alpha_t(\gamma) = \frac{\eta\gamma\beta_t(\gamma)}{\gamma_{t-1}} = \eta\left(1 + \frac{2\mu\gamma_{t-1}}{1-\eta}\right)^{-1} \le \frac{1}{3}. \tag{60}$$

Recall that $\mathcal{S}$, r_0 and w are given in (8) and (57), respectively. Using $x^t, x^{t-1}, x^* \in \mathcal{S}$, $0 < \gamma \le \gamma_0$, (8), (57), (59) and (60), we have

$$\begin{aligned}
\|w\| &\le \alpha_t(\gamma)\|x^t - x^{t-1}\| + \gamma L_{\mathcal{S}}\|x^t - x^*\| + \gamma\beta_t(\gamma)L_{\mathcal{S}}\|x^t - x^{t-1}\| \\
&\le \left(\frac{1}{3} + \gamma_0 L_{\mathcal{S}}\right)\|x^t - x^{t-1}\| + \gamma_0 L_{\mathcal{S}}\|x^t - x^*\| \\
&\le \left(\frac{1}{3} + \gamma_0 L_{\mathcal{S}}\right)(\|x^t - x^*\| + \|x^{t-1} - x^*\|) + \gamma_0 L_{\mathcal{S}}\|x^t - x^*\| \\
&\le \frac{2}{\sqrt{1-2\nu^2}}\left(\frac{1}{3} + \gamma_0 L_{\mathcal{S}}\right) r_0 + \frac{1}{\sqrt{1-2\nu^2}}\gamma_0 L_{\mathcal{S}} r_0 \le \frac{(2 + 9\gamma_0 L_{\mathcal{S}})r_0}{3\sqrt{1-2\nu^2}}.
\end{aligned}$$

This together with $x^t \in \mathcal{S}$, (8) and (58) yields

$$\|x^{t+1}(\gamma) - x^*\| \le \|x^t - x^*\| + \|w\| < \frac{(5 + 9\gamma_0 L_{\mathcal{S}})r_0}{3\sqrt{1-2\nu^2}}.$$

The conclusion then follows from this and the definition of $\widehat{\mathcal{S}}$ in (10). □

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 $t \ge 1$ 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 $n_t \le M + t - \sum_{i=1}^{t-1} n_i$, where M is given in* (11).

Proof. Recall that $x^{t+1}(\gamma)$ and $V_{t+1}(\gamma)$ are defined in (53) and (54), respectively. It follows from Lemma 6 that $x^{t+1}(\gamma) \in \widehat{\mathcal{S}}$ for any $0 < \gamma \le \gamma_0$. Also, notice that $x^t \in \mathcal{S} \subset \widehat{\mathcal{S}}$. By these and Lemma 1(ii), one has

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

Using this and (54), we obtain that for any $0 < \gamma \le \gamma_0$,

$$V_{t+1}(\gamma) \le \|F(x^{t+1}(\gamma)) - F(x^t)\| + \eta\gamma^{-1}\|x^{t+1}(\gamma) - x^t\| \le (L_{\widehat{\mathcal{S}}} + \eta\gamma^{-1})\|x^{t+1}(\gamma) - x^t\|. \tag{61}$$

In addition, notice from (5) that $\gamma_i \le \gamma_{i-1}\delta^{n_i-1}$ for $i = 1, \ldots, t-1$, which implies that $\gamma_{t-1} \le \gamma_0\delta^{\sum_{i=1}^{t-1}(n_i-1)}$. Let $\gamma = \min\{\gamma_0, \delta^{-1}\gamma_{t-1}\}\delta^{M+t-\sum_{i=1}^{t-1} n_i}$. In view of these, $\delta \in (0,1)$, and the definition of M in (11), one can verify that

$$0 < \gamma \le \delta^{-1}\gamma_{t-1}\delta^{M+t-\sum_{i=1}^{t-1} n_i} \le \gamma_0\delta^M \le \min\{\xi/L_{\widehat{\mathcal{S}}}, \gamma_0\}.$$

It then follows from this, (9), and (61) that

$$V_{t+1}(\gamma) \le (L_{\widehat{\mathcal{S}}}\gamma + \eta)\gamma^{-1}\|x^{t+1}(\gamma) - x^t\| \le \nu(1-\eta)\gamma^{-1}\|x^{t+1}(\gamma) - x^t\|,$$

which, together with (6), (54), the expression of γ, and the definition of n_t (see step 2 of Algorithm 1), implies that $n_t \le M + t - \sum_{i=1}^{t-1} n_i$ 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 $n_1 \le M + 1$. Hence, Algorithm 1 is well-defined at iteration 1. By this, (43), and (50) with $k = 1$, one has

$$\|x^2 - x^*\|^2 \overset{(50)}{\le} \frac{1}{(1-2\nu^2)\theta_1}\|x^0 - x^*\|^2 \overset{(43)}{=} \frac{1}{1-2\nu^2}\|x^0 - x^*\|^2,$$

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 $\sum_{i=1}^{t} n_i \le M + t$. Hence, Algorithm 1 is well-defined at iteration t. By this, (43), and (50) with $k = t$, one has

$$\|x^{t+1} - x^*\|^2 \overset{(50)}{\le} \frac{1}{(1-2\nu^2)\theta_t}\|x^0 - x^*\|^2 \overset{(43)}{\le} \frac{1}{1-2\nu^2}\|x^0 - x^*\|^2,$$

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 $0 \le \eta < 1/3$ and $0 < \gamma_t \le \gamma_0$ for all $0 \le t \in \mathbb{T} - 1$. Using these and (5), we have that for all $1 \le t \in \mathbb{T} - 1$,

$$\gamma_t\beta_t = \gamma_{t-1}\left(1 + \frac{2\mu\gamma_{t-1}}{1-\eta}\right)^{-1} \le \gamma_0, \quad \alpha_t = \frac{\eta\gamma_t\beta_t}{\gamma_{t-1}} = \eta\left(1 + \frac{2\mu\gamma_{t-1}}{1-\eta}\right)^{-1} \le \frac{1}{3}. \tag{62}$$

We next show by induction that

$$\gamma_t \ge \min\left\{L_{\widehat{\mathcal{S}}}^{-1}\delta\xi, \gamma_0\right\} \quad \forall 0 \le t \in \mathbb{T} - 1, \tag{63}$$

where ξ is defined in (9). Indeed, (63) clearly holds at $t = 0$. Suppose that (63) holds at some $0 \le t - 1 \in \mathbb{T} - 2$. We now show that (63) holds at t by considering the following two separate cases.

Case (a): $n_t = 0$. It follows from this and Algorithm 1 that $\gamma_t = \min\{\gamma_0, \delta^{-1}\gamma_{t-1}\}$, which together with $\delta \in (0,1)$ and (63) with t replaced by $t-1$ implies that (63) holds at t.

Case (b): $n_t > 0$. By this, (5), $\delta \in (0,1)$ and the definition of n_t, one can observe that $\gamma_t/\delta = \min\{\gamma_0, \delta^{-1}\gamma_{t-1}\}\delta^{n_t-1} \le \gamma_0$ and (6) will not hold if γ_t is replaced by γ_t/δ. Besides, from the proof of Lemma 7, one can see that (6) will hold if γ_t is replaced by $\tilde{\gamma}$ satisfying $0 < \tilde{\gamma} \le \min\{\xi/L_{\widehat{\mathcal{S}}}, \gamma_0\}$. Hence, it follows that $\gamma_t/\delta > \xi/L_{\widehat{\mathcal{S}}}$, which together with $\gamma_t \le \gamma_0$ implies that (63) holds at t.

By (8), (43), (50), and (63), one has that for all $1 \leq t \in \mathbb{T}-1$,

$$\begin{aligned}\|x^{t+1}-x^*\|^2 &\leq \frac{1}{(1-2\nu^2)\prod_{i=1}^{t-1}\left(1+\frac{2\mu\gamma_i}{1-\eta}\right)}\|x^0-x^*\|^2\\ &\leq \frac{r_0^2}{1-2\nu^2}\left(1+\frac{2\mu}{1-\eta}\min\left\{L_{\widehat{\mathcal{S}}}^{-1}\delta\xi,\gamma_0\right\}\right)^{1-t}.\end{aligned}$$

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

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

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 (42) 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, (40) and Lemma 1(i), one has

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

Also, notice from (12) that $T \geq 3$. By this, $\gamma_T \leq \gamma_0$, (12), (42), (62), (63), (64), and (65), one has

$$\begin{aligned}\|v^T\| &\overset{(42)}{\leq} \frac{1}{\gamma_T}\left(\|x^{T+1}-x^T\|+\alpha_T\|x^T-x^{T-1}\|+\gamma_T\|\Delta^{T+1}\|+\gamma_T\beta_T\|\Delta^T\|\right)\\ &\overset{(65)}{\leq} \frac{1}{\gamma_T}\left(\|x^{T+1}-x^T\|+\alpha_T\|x^T-x^{T-1}\|+\gamma_T L_{\mathcal{S}}\|x^{T+1}-x^T\|+\gamma_T\beta_T L_{\mathcal{S}}\|x^T-x^{T-1}\|\right)\\ &\overset{(64)}{\leq} \frac{2r_0}{\gamma_T\sqrt{1-2\nu^2}}\left(1+\alpha_T+\gamma_T L_{\mathcal{S}}+\gamma_T\beta_T L_{\mathcal{S}}\right)\left(1+\frac{2\mu}{1-\eta}\min\left\{L_{\widehat{\mathcal{S}}}^{-1}\delta\xi,\gamma_0\right\}\right)^{\frac{3-T}{2}}\\ &\overset{(62)}{\leq} \frac{2r_0}{\gamma_T\sqrt{1-2\nu^2}}\left(1+\frac{1}{3}+\gamma_0 L_{\mathcal{S}}+\gamma_0 L_{\mathcal{S}}\right)\left(1+\frac{2\mu}{1-\eta}\min\left\{L_{\widehat{\mathcal{S}}}^{-1}\delta\xi,\gamma_0\right\}\right)^{\frac{3-T}{2}}\\ &\overset{(63)}{\leq} \frac{r_0\left(8+12\gamma_0 L_{\mathcal{S}}\right)}{3\sqrt{1-2\nu^2}\min\left\{L_{\widehat{\mathcal{S}}}^{-1}\delta\xi,\gamma_0\right\}}\left(1+\frac{2\mu}{1-\eta}\min\left\{L_{\widehat{\mathcal{S}}}^{-1}\delta\xi,\gamma_0\right\}\right)^{\frac{3-T}{2}} \overset{(12)}{\leq} \epsilon,\end{aligned}$$

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 $t \leq T$. It then follows that (7) holds for such t. By this and (42), one can see that $\|v^t\| \leq \epsilon$, which together with (44) implies that $\mathrm{res}_{F+B}(x^{t+1}) \leq \epsilon$.

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

$$\sum_{t=1}^{|\mathbb{T}|-2}(n_t+1) \leq 2(|\mathbb{T}|-2)+M \leq 2T+M,$$

which together with (11) and (12) implies that the conclusion holds. □

6.2 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 $0 \leq k \in \mathbb{K}-1$ of Algorithm 2, Algorithm 1 is called to find an approximate solution of the following strongly MI problem

$$0 \in (F_k+B)(x) = (F+B)(x)+\rho_k^{-1}(x-z^k). \tag{66}$$

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

$$z_*^k = (I+\rho_k(F+B))^{-1}(z^k). \tag{67}$$

Moreover, z_*^k is the unique solution of problem (66) and thus

$$0 \in (F_k + B)(z_*^k) = (F + B)(z_*^k) + \rho_k^{-1}(z_*^k - z^k). \tag{68}$$

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*

$$\|z^{k+1} - z_*^k\| \le \rho_k \tau_k, \tag{69}$$

where z_*^k is defined in (67)*.*

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 (68) that

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

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

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

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*

$$\|z^s - x^*\| \le \|z^0 - x^*\| + \sum_{k=0}^{s-1} \rho_k\tau_k \qquad \forall 1 \le s \in \mathbb{K}, \tag{70}$$

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

Proof. By (68) 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

$$\|z_*^k - x^*\| \le \|z^k - x^*\|, \quad \|z^k - z_*^k\| \le \|z^k - x^*\|. \tag{72}$$

By the first relation in (72), 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 (69) implies that (70) holds. In addition, using (69) with $k = s$, (70) and (72), we have

$$\|z^{s+1} - z^s\| \le \|z^s - z_*^s\| + \|z^{s+1} - z_*^s\| \overset{(69),(72)}{\le} \|z^s - x^*\| + \rho_s\tau_s \overset{(70)}{\le} \|z^0 - x^*\| + \sum_{k=0}^{s} \rho_k\tau_k.$$

Hence, (71) holds as desired. □

Define

$$\mathcal{S}_k = \left\{x \in \operatorname{dom} B : \ \|x - z_*^k\| \le \frac{1}{\sqrt{1-2\nu^2}}\|z^k - z_*^k\|\right\} \qquad \forall 0 \le k \in \mathbb{K} - 1, \tag{73}$$

$$\widehat{\mathcal{S}}_k = \left\{x \in \operatorname{dom} B : \ \|x - z_*^k\| \le \frac{5 + 9\gamma_0 L_{\mathcal{Q}}}{3\sqrt{1-2\nu^2}}\|z^k - z_*^k\|\right\} \qquad \forall 0 \le k \in \mathbb{K} - 1, \tag{74}$$

where z_*^k is defined in (67), $L_{\mathcal{Q}}$ 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* (73) *and* (74)*. 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 (16) and (17). 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, (70), (72), (73), and $\sum_{i=0}^{\infty} \rho_i\tau_i = \rho_0\tau_0/(1-\sigma\zeta)$, we have

$$\begin{aligned}\|x - x^*\| \le \|x - z_*^k\| + \|z_*^k - x^*\| &\overset{(73)}{\le} \frac{1}{\sqrt{1-2\nu^2}}\|z^k - z_*^k\| + \|z_*^k - x^*\| \overset{(72)}{\le} \left(\frac{1}{\sqrt{1-2\nu^2}} + 1\right)\|z^k - x^*\| \\ &\overset{(70)}{\le} \left(\frac{1}{\sqrt{1-2\nu^2}} + 1\right)\left(\|z^0 - x^*\| + \sum_{i=0}^{k-1}\rho_i\tau_i\right) \le \left(\frac{1}{\sqrt{1-2\nu^2}} + 1\right)\left(\|z^0 - x^*\| + \frac{\rho_0\tau_0}{1-\sigma\zeta}\right),\end{aligned}$$

which together with (16) 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, η, $\{\rho_k\}$ and $\{\tau_k\}$ be given in Algorithm 2, and let ξ, $\bar{r}_0$ and Λ be defined in* (9), (16) *and* (18)*. 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 $\bar{M}_k$, where*

$$\bar{M}_k = 6 + 2\left\lceil \frac{2\log\frac{(\bar{r}_0+\Lambda)(8+12\gamma_0 L_{\mathcal{Q}})}{3\tau_k\sqrt{1-2\nu^2}\min\left\{L_{\widehat{\mathcal{Q}}}^{-1}\delta\xi, \gamma_0\right\}}}{\log\left(1 + \frac{2}{\rho_k(1-\eta)}\min\left\{L_{\widehat{\mathcal{Q}}}^{-1}\delta\xi, \gamma_0\right\}\right)}\right\rceil_+ + \left\lceil \frac{\log\left(\frac{\xi}{\gamma_0 L_{\widehat{\mathcal{Q}}}}\right)}{\log\delta}\right\rceil_+. \tag{75}$$

Proof. Recall that $F_k + B$ is $1/\rho_k$-strongly monotone and $(F_k + B)^{-1}(0) \ne \emptyset$. 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 (70), (72), 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 (67). The conclusion then follows from applying Theorem 2 to the subproblem (66) with ϵ, μ, $L_{\mathcal{S}}$, $L_{\widehat{\mathcal{S}}}$, and r_0 in (13) 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 (15) does not hold for $k = K$. On the other hand, by (71), $\rho_k = \rho_0\zeta^k$ 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 (20) 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 (15) 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

$$\rho_k^{-1}\|z^{k+1} - z^k\| + \tau_k \le \varepsilon. \tag{76}$$

In addition, by the definition of z^{k+1} (see step 2 of Algorithm 2), Theorem 2, and (14), there exists some v such that

$$v \in (F + B)(z^{k+1}) + \rho_k^{-1}(z^{k+1} - z^k), \quad \|v\| \le \tau_k. \tag{77}$$

Observe that $v - \rho_k^{-1}(z^{k+1} - z^k) \in (F + B)(z^{k+1})$. It follows from this, (76), (77), and the definition of res_{F+B} that

$$\mathrm{res}_{F+B}(z^{k+1}) \le \|v - \rho_k^{-1}(z^{k+1} - z^k)\| \le \|v\| + \rho_k^{-1}\|z^{k+1} - z^k\| \overset{(77)}{\le} \tau_k + \rho_k^{-1}\|z^{k+1} - z^k\| \overset{(76)}{\le} \varepsilon.$$

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 $\bar{M}_k$, where $\bar{M}_k$ is given in (75). In addition, using (18), (19) and (75), we have

$$\bar{M}_k = 6 + 2\left\lceil \frac{2C_1 - 2k\log\sigma}{\log\left(\min\left\{1 + \frac{2\delta\xi}{L_{\widehat{\mathcal{Q}}}\rho_0(1-\eta)\zeta^k}, 1 + \frac{2\gamma_0}{\rho_0(1-\eta)\zeta^k}\right\}\right)}\right\rceil_+ + C_2. \tag{78}$$

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

$$\begin{aligned}
&\log\left(\min\left\{1 + \frac{2\delta\xi}{\rho_0(1-\eta)\zeta^k L_{\widehat{\mathcal{Q}}}}, 1 + \frac{2\gamma_0}{\rho_0(1-\eta)\zeta^k}\right\}\right) \\
&= \min\left\{\log\left(1 + \zeta^{-k}\frac{2\delta\xi}{\rho_0(1-\eta)L_{\widehat{\mathcal{Q}}}}\right), \log\left(1 + \zeta^{-k}\frac{2\gamma_0}{\rho_0(1-\eta)}\right)\right\} \\
&\ge \min\left\{\zeta^{-k}\log\left(1 + \frac{2\delta\xi}{\rho_0(1-\eta)L_{\widehat{\mathcal{Q}}}}\right), \zeta^{-k}\log\left(1 + \frac{2\gamma_0}{\rho_0(1-\eta)}\right)\right\}.
\end{aligned} \tag{79}$$

By (78), (79), $\sigma \in (0, 1)$ and $\zeta > 1$, one has that for all $k \ge 0$,

$$\begin{aligned}
\bar{M}_k &\overset{(78)}{\le} 8 + \frac{(4C_1 - 4k\log\sigma)_+}{\log\left(\min\left\{1 + \frac{2\delta\xi}{\rho_0(1-\eta)\zeta^k L_{\widehat{\mathcal{Q}}}}, 1 + \frac{2\gamma_0}{\rho_0(1-\eta)\zeta^k}\right\}\right)} + C_2 \\
&\le 8 + \frac{4(C_1)_+ - 4k\log\sigma}{\log\left(\min\left\{1 + \frac{2\delta\xi}{\rho_0(1-\eta)\zeta^k L_{\widehat{\mathcal{Q}}}}, 1 + \frac{2\gamma_0}{\rho_0(1-\eta)\zeta^k}\right\}\right)} + C_2 \\
&\overset{(79)}{\le} 8 + \frac{4\zeta^k(C_1)_+ - 4k\zeta^k\log\sigma}{\log\left(\min\left\{1 + \frac{2\delta\xi}{\rho_0(1-\eta)L_{\widehat{\mathcal{Q}}}}, 1 + \frac{2\gamma_0}{\rho_0(1-\eta)}\right\}\right)} + C_2.
\end{aligned} \tag{80}$$

Observe that $|\mathbb{K}| \le K + 2$ and also the total number of inner iterations of Algorithm 2 is at most $\sum_{t=0}^{|\mathbb{K}|-2}\bar{M}_t$. It then follows from (19), (78) and (80) that the total number of evaluations of F and resolvent of B performed in Algorithm 2 is at most

$$\begin{aligned}
\sum_{k=0}^{|\mathbb{K}|-2}\bar{M}_k &\le \sum_{k=0}^{K}\left(8 + \frac{4\zeta^k(C_1)_+ - 4k\zeta^k\log\sigma}{\log\left(\min\left\{1 + \frac{2\delta\xi}{\rho_0(1-\eta)L_{\widehat{\mathcal{Q}}}}, 1 + \frac{2\gamma_0}{\rho_0(1-\eta)}\right\}\right)} + C_2\right) \\
&\le 8K + 8 + 4(C_1)_+ C_3\zeta^{K+1} + 4C_3(\log\sigma^{-1})K\zeta^{K+1} + (K+1)C_2,
\end{aligned} \tag{81}$$

where the second inequality is due to $\sum_{k=0}^{K}\zeta^k \le \zeta^{K+1}/(\zeta-1)$ and $\sum_{k=0}^{K}k\zeta^k \le K\zeta^{K+1}/(\zeta-1)$. By the definition of K in (20), one has

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

which together with $\zeta > 1$ implies that

$$\zeta^K \le \max\left\{\frac{2\zeta(\bar{r}_0+\Lambda)}{\varepsilon\rho_0}, \zeta\left(\frac{2\tau_0}{\varepsilon}\right)^{\frac{\log\zeta}{\log(1/\sigma)}}, 1\right\}. \tag{82}$$

Using (20), (21), (81) and (82), we can see that $\sum_{k=0}^{|\mathbb{K}|-2}\bar{M}_k \le \bar{M}$. □

7 Concluding remarks

We proposed primal-dual extrapolation methods enjoying an operation complexity of $\mathcal{O}\left(\log\varepsilon^{-1}\right)$ and $\mathcal{O}\left(\varepsilon^{-1}\log\varepsilon^{-1}\right)$, measured by the number of fundamental operations for finding an ϵ-residual solution of strongly and non-strongly monotone inclusion problems under local Lipschitz continuity, respectively. The latter complexity significantly improves upon the previously best operation complexity $\mathcal{O}(\varepsilon^{-2})$ achieved by the FRBS method [13].

One natural question is whether the aforementioned operation complexity of $\mathcal{O}\left(\varepsilon^{-1}\log\varepsilon^{-1}\right)$ can be improved to $\mathcal{O}\left(\varepsilon^{-1}\right)$, which would match the optimal complexity for solving non-strongly monotone inclusion problems under global Lipschitz continuity. Additionally, our proposed methods require the exact resolvent of B, which limits their applicability. Clearly, a method using the inexact resolvent of B for the monotone inclusion problem would be both practically and theoretically interesting. It is worthwhile to explore these as future research directions.

References

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

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

[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] K. Huang and S. Zhang. New first-order algorithms for stochastic variational inequalities. *SIAM Journal on Optimization*, 32(4):2745–2772, 2022.

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

[6] G. Kotsalis, G. Lan, and T. Li. Simple and optimal methods for stochastic variational inequalities, I: operator extrapolation. *SIAM Journal on Optimization*, 32(3):2041–2073, 2022.

[7] D. Kovalev and A. Gasnikov. The first optimal algorithm for smooth and strongly-convex-strongly-concave minimax optimization. *Advances in Neural Information Processing Systems*, 35:14691–14703, 2022.

[8] P. Latafat, A. Themelis, L. Stella, and P. Patrinos. Adaptive proximal algorithms for convex optimization under local lipschitz continuity of the gradient. *arXiv preprint arXiv:2301.04431*, page 4, 2023.

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

[10] 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.

[11] Z. Lu and Z. Zhou. Iteration-complexity of first-order augmented Lagrangian methods for convex conic programming. *SIAM Journal on Optimization*, 33(2):1159–1190, 2023.

[12] Y. Malitsky. Golden ratio algorithms for variational inequalities. *Mathematical Programming*, 184(1-2):383–410, 2020.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

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

[18] 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.

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

[20] 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.

[21] 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.

[22] R. T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14:877–898, 1976.

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

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

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

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

[27] 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.

[28] 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.

[29] 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.

[30] J. Zhang, M. Wang, M. Hong, and S. Zhang. Primal-dual first-order methods for affinely constrained multi-block saddle point problems. *SIAM Journal on Optimization*, 33(2):1035–1060, 2023.