Accelerated Algorithms for Monotone Inclusion and Constrained Nonconvex-Nonconcave Min-Max Optimization

Yang Cai YANG.CAI@YALE.EDU
Argyris Oikonomou ARGYRIS.OIKONOMOU@YALE.EDU
Weiqiang Zheng WEIQIANG.ZHENG@YALE.EDU
Yale University

Abstract

We study monotone inclusions and monotone variational inequalities, as well as their generalizations to non-monotone settings. We first show that the *Extra Anchored Gradient (EAG)* algorithm, originally proposed by Yoon and Ryu [37] for unconstrained convex-concave min-max optimization, can be applied to solve the more general problem of Lipschitz monotone inclusion. More specifically, we prove that the EAG solves Lipschitz monotone inclusion problems with an *accelerated convergence rate* of $O(\frac{1}{T})$, which is *optimal among all first-order methods* [9, 37]. Our second result is an accelerated forward-backward splitting algorithm (AS), which not only achieves the accelerated $O(\frac{1}{T})$ convergence rate for all monotone inclusion problems, but also exhibits the same accelerated rate for a family of general (non-monotone) inclusion problems that concern negative comonotone operators. As a special case of our second result, AS enjoys the $O(\frac{1}{T})$ convergence rate for solving a non-trivial class of nonconvex-nonconcave min-max optimization problems. Our analyses are based on simple potential function arguments, which might be useful for analysing other accelerated algorithms.

1. Introduction

We study the constrained single-valued *monotone inclusion* problem and the *monotone variational inequality*, as well as their generalizations in non-monotone settings. Given a closed convex set $\mathcal{Z} \subseteq \mathbb{R}^n$ and a single-valued and *monotone* operator $F : \mathcal{Z} \to \mathbb{R}^n$, i.e.,

$$\langle F(z) - F(z'), z - z' \rangle \geq 0, \quad \forall z, z' \in \mathcal{Z},$$

the constrained single-valued monotone inclusion problem (MI) consists in finding a $z^* \in \mathcal{Z}$ such that

$$\mathbf{0} \in F(z^*) + \partial \mathbb{I}_{\mathcal{Z}}(z^*),$$

where $\mathbb{I}_{\mathcal{Z}}(\cdot)$ is the indicator function for set $\mathcal{Z}$,[1] and $\partial \mathbb{I}_{\mathcal{Z}}(\cdot)$ is the subdifferential operator of $\mathbb{I}_{\mathcal{Z}}$. [2] The corresponding monotone variational inequality shares the same input, and asks for a $z^* \in \mathcal{Z}$ such that

$$\langle F(z^*), z^* - z \rangle \leq 0, \quad \forall z \in \mathcal{Z}.$$

1. $\mathbb{I}(z) = 0$ for all $z \in \mathcal{Z}$ and $+\infty$ otherwise.
2. In general, a monotone inclusion problem refers to finding a zero of a set-valued maximally monotone operator. When there is no confusion, we use monotone inclusion problems to refer to the constrained single-valued monotone inclusion problems.

The monotone inclusion problem (MI) and the related monotone variational inequality play a crucial role in mathematical programming, providing unified settings for the study of optimization and equilibrium problems. They also serve as computational frameworks for numerous important applications in fields such as economics, engineering, probability and statistics, and machine learning [2, 12, 31].

Although the set of exact solutions to the monotone inclusion problem (MI) coincides with the set of exact solutions to the corresponding variational inequality, the approximate solutions to these two problems differ due to different performance measures. An approximate solution to the monotone inclusion problem (MI) must have a small natural residual,[3] while an approximate solution to the variational inequality only satisfies a weaker condition, i.e., its gap function is small.[4] Indeed, it is well-known that an approximate solution to the monotone inclusion problem (MI) is also an approximate solution to the monotone variational equality, but the reverse is not true in general.

An important special case of the monotone inclusion problem (MI) is the convex-concave min-max optimization problem:

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

where $\mathcal{X}$ and $\mathcal{Y}$ are a closed convex sets in $\mathbb{R}^{n_x}$ and $\mathbb{R}^{n_y}$ respectively, and $f(\cdot,\cdot)$ is smooth, convex in x, and concave in y.[5] Besides its central importance in game theory, convex optimization, and online learning, the convex-concave min-max optimization problem has recently received a lot of attention from the machine learning community due to several novel applications such as the generative adversarial networks (GANs) (e.g., [1, 14]), adversarial examples (e.g., [24]), robust optimization (e.g., [4]), and reinforcement learning (e.g., [6, 11]).

Given the importance of the monotone inclusion problem (MI) , it is crucial to understand the following open question.

What is the optimal convergence rate achievable by a first-order method for monotone inclusions? (*)

We provide the first algorithm that achieves the optimal convergence rate and further extend it to inclusion problems with *negatively comonotone operators*, an important family of non-monotone operators. This generalization allows us to obtain the optimal convergence rate for a family of *structured nonconvex-nonconcave min-max* optimization problems. Prior to our work, even for the special case of convex-concave min-max optimization, the optimal convergence rate is only known for the relatively weak notion of duality gap [26, 27], which is also difficult to generalize to nonconvex-nonconcave settings, see [37] for more discussion.

3. The natural residual of a point z is simply the operator norm $\|F(z)\|$ in the unconstrained case, i.e., $\mathcal{Z} = \mathbb{R}^n$, and equals to the norm of its natural map $z - \Pi_{\mathcal{Z}}[z - F(z)]$ [12].
4. There are several variations of the gap function. Depending on the exact definition, a small gap function value could mean an approximate *weak* solution, i.e., approximately solve the Minty Variational Inequality (MVI), or an approximate *strong* solution, i.e., approximately solve the Stampacchia Variational Inequality (SVI). Formal definitions and discussions are in Section 2.1.
5. If we set $F(x,y) = \begin{pmatrix} \nabla_x f(x,y) \\ -\nabla_y f(x,y) \end{pmatrix}$ and $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, then (i) $F(x,y)$ is a Lipschitz and monotone operator, and (ii) the set of saddle points coincide with the solutions to the monotone inclusion (MI) problem for operator F and domain $\mathcal{Z}$.

1.1. Our Contributions

A point $z \in \mathcal{Z}$ is an ϵ-approximate solution to the constrained single-valued monotone inclusion problem (MI) if

$$\mathbf{0} \in F(z) + \partial \mathbb{I}_{\mathcal{Z}}(z) + \mathcal{B}(\mathbf{0}, \epsilon),$$

where $\mathcal{B}(\mathbf{0}, \epsilon)$ is the ball with radius ϵ centered at $\mathbf{0}$. As we argue in Section 2.3, this is equivalent to the tangent residual of z, a notion introduced in [5], being no more than ϵ. Our first contribution provides an answer to question (*).

Contribution 1: We extend the Extra Anchored Gradient algorithm (EAG), originally proposed by Yoon and Ryu [37] for unconstrained convex-concave min-max problems, to solve constrained single-valued monotone inclusion problems (MI) . Note that constrained convex-concave min-max optimization is a special case. We show in Theorem 1 that (EAG) finds an $O(\frac{L}{T})$-approximate solution in T iterations for monotone inclusions (MI) , where L is the Lipschitz constant of the operator F. The convergence rate we obtain for (EAG) matches the lower bound by [9, 37], and is therefore optimal for any first-order method.

For the second part of the paper, we go beyond the monotone case and study inclusion problems (CMI) with operators that are not necessarily monotone and only satisfy the weaker *ρ-comonotoncity* (Assumption 2) condition. Given a single-valued, L-Lipschitz, and possibly *non-monotone* operator F and a *set-valued maximally monotone* operator A, we denote $E = F + A$. The inclusion problem (CMI) consists in finding a point $z^* \in \mathbb{R}^n$ that satisfies

$$\mathbf{0} \in E(z^*) = F(z^*) + A(z^*),$$

under the assumption that E is a set-valued negatively comonotone operator.

The inclusion problem (CMI) captures (MI) (when $\rho = 0$ and $A = \partial \mathbb{I}_{\mathcal{Z}}$) and a class of *non-smooth nonconvex-nonconcave* min-max optimization problems (when $\rho < 0$ and A chosen appropriately, see Example 2). Our second contribution is a new algorithm that achieves accelerated rate for solving (CMI).

Contribution 2: We design an accelerated forward-backward splitting algorithm (AS) that finds a $O(\frac{L}{T})$-approximate solution to the inclusion problem (CMI) in T iterations as long as $E = F + A$ is a ρ-comonotone operator with $\rho > -\frac{1}{2 \cdot L}$. See Theorem 2 for a formal statement.

Our algorithm is inspired by the Fast Extra-Gradient (FEG) algorithm [20]. In particular, (AS) is identical to FEG in the unconstrained setting, i.e., $A(z) = \mathbf{0}$ for all z, and can be viewed as a generalization of FEG to accommodate arbitrary set-valued maximally monotone operator A. Our result is the first to obtain accelerated rate for constrained nonconvex-nonconcave min-max optimization and inclusion problems with negatively comonotone operators. Splitting methods are a family of algorithms for solving inclusion problems where the goal is to find a zero of an operator that can be represented as the sum of two or more operators. These methods only invoke each operator individually rather than their sum directly, and hence the name. There is a rich literature on splitting methods. See [32] and the references therein for more detailed discussion.

1.2. Related Work

There is a vast literature on inclusion problems and variational inequalities, e.g., see [2, 12, 31] and the references therein. We summarize most relevant results in Table 1 and provided more detailed discussion in Appendix A.

Table 1: Existing results for min-max optimization problem with monotone or non-monotone operators. The convergence rate is in terms of the operator norm (in the unconstrained setting) and the residual (in the constrained setting). $(*)$: the result only holds for MVI but not weak MVI.

Algorithm		Setting	Monotone	Non-Monotone	
				Comonotone	Weak MVI
Normal	EG [7]	general	$O(\frac{1}{\sqrt{T}})$		$O(\frac{1}{\sqrt{T}})^*$
	EG+ [10]	unconstrained	$O(\frac{1}{\sqrt{T}})$	$O(\frac{1}{\sqrt{T}})$	$O(\frac{1}{\sqrt{T}})$
	CEG+ [30]	general	$O(\frac{1}{\sqrt{T}})$	$O(\frac{1}{\sqrt{T}})$	$O(\frac{1}{\sqrt{T}})$
Accelerated	Halpern [9]	general	$O(\frac{\log T}{T})$		
	EAG [37]	unconstrained	$O(\frac{1}{T})$		
	FEG [20]	unconstrained	$O(\frac{1}{T})$	$O(\frac{1}{T})$	
	EAG [**This paper**]	general	$O(\frac{1}{T})$		
	AS [**This paper**]	general	$O(\frac{1}{T})$	$O(\frac{1}{T})$	

2. Preliminaries

We consider the Euclidean Space $(\mathbb{R}^n, \|\cdot\|)$, where $\|\cdot\|$ is the ℓ_2 norm and $\langle\cdot,\cdot\rangle$ denotes inner product on $\mathbb{R}^n$.

Basic Notions about Monotone Operators. A set-valued operator $A : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ maps $z \in \mathbb{R}^n$ to a subset $A(z) \subseteq \mathbb{R}^n$. We say A is single-valued if $|A(z)| \le 1$ for all $z \in \mathbb{R}^n$. Tha *graph* of an operator A is defined as $\mathrm{Gra}_A = \{(z,u) : z \in \mathbb{R}^n, u \in A(z)\}$. The inverse operator of A is denoted as A^{-1} whose graph is $\mathrm{Gra}_{A^{-1}} = \{(u,z) : (z,u) \in \mathrm{Gra}_A\}$. For two operators A and B, we denote $A+B$ as the operator with graph $\mathrm{Gra}_{A+B} = \{(z, u_A + u_B) : (z,u_A) \in \mathrm{Gra}_A, (z,u_B) \in \mathrm{Gra}_B\}$. We denote the identity operator as $I : \mathbb{R}^n \to \mathbb{R}^n$.

For $L \in (0,\infty)$, a single-valued operator $A : \mathbb{R}^n \to \mathbb{R}^n$ is *L-Lipschitz* if

$$\left\|A(z) - A(z')\right\| \le L \cdot \left\|z - z'\right\|, \quad \forall z, z' \in \mathbb{R}^n.$$

Moreover, A is *non-expansive* if it is 1-Lipschitz. A set-valued operator $A : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is *monotone* if

$$\left\langle u - u', z - z'\right\rangle \ge 0, \quad \forall (z,u), (z',u') \in \mathrm{Gra}_A .$$

For a closed convex set $\mathcal{Z} \subseteq \mathbb{R}^n$ and point $z \in \mathbb{R}^n$, we denote the *normal cone operator* as $N_{\mathcal{Z}}$:

$$N_{\mathcal{Z}}(z) = \begin{cases} \emptyset, & z \notin \mathcal{Z}, \\ \{v \in \mathbb{R}^n : \langle v, z' - z\rangle \le 0, \forall z' \in \mathcal{Z}\}, & z \in \mathcal{Z}. \end{cases}$$

Define the indicator function

$$\mathbb{I}_{\mathcal{Z}}(z) = \begin{cases} 0 & \text{if } z \in \mathcal{Z}, \\ +\infty & \text{otherwise.} \end{cases}$$

Then it is not hard to see that the subdifferential operator $\partial \mathbb{I}_{\mathcal{Z}} = N_{\mathcal{Z}}$. The projection operator $\Pi_{\mathcal{Z}} : \mathbb{R}^n \to \mathbb{R}^n$ is defined as $\Pi_{\mathcal{Z}}[z] := \text{argmin}_{z' \in \mathcal{Z}} \|z - z'\|^2$.

Maximally Monotone Operator. A is *maximally monotone* if A is monotone and there is no other monotone operator B such that $\text{Gra}_A \subset \text{Gra}_B$. When $f : \mathbb{R}^n \to \mathbb{R}$ is a convex closed proper function, then the subdifferential operator ∂f is maximally monotone. Therefore, $\partial \mathbb{I}_{\mathcal{Z}} = N_{\mathcal{Z}}$ is maximally monotone. We denote the **resolvent** of an operator A as $J_A := (I + A)^{-1}$.

ρ-comonotonicity. A generalized notion of monotonicity is the *ρ-comonotonicity* [3]: For $\rho \in \mathbb{R}$, an operator $A : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is ρ-comonotone if

$$\langle u - u', z - z' \rangle \geq \rho \|u - u'\|^2, \quad \forall (z, u), (z', u') \in \text{Gra}_A .$$

Note that when A is 0-comonotone, then A is monotone. If A is ρ-comonotone for $\rho > 0$, we also say A is *ρ-cocoercive* (a stronger assumption than monotonicity). When A satisfies negative comonotonicity, i.e., ρ-comonotonicity with $\rho < 0$, then A is possibly non-monotone. Negative comonotonicity is the focus of this paper in the non-monotone setting.

2.1. Monotone Inclusion and Variational Inequality

Constrained single-valued monotone Inclusion. Given a closed convex set $\mathcal{Z} \subseteq \mathbb{R}^n$ and a single-valued monotone operator F, the *constrained single-valued monotone inclusion* problem is to find a point $z^* \in \mathbb{R}^n$ that satisfies

$$\mathbf{0} \in F(z^*) + \partial \mathbb{I}_{\mathcal{Z}}(z^*). \tag{MI}$$

We focus on monotone inclusion problems of Lipschitz operators that have a solution.

Assumption 1 *In (MI) problem,*

1. *F is monotone and L-Lipschitz on $\mathcal{Z}$, i.e.,*

$$\langle F(z) - F(z'), z - z' \rangle \geq 0 \textit{ and } \|F(z) - F(z')\| \leq L \cdot \|z - z'\|, \quad \forall z, z' \in \mathcal{Z}.$$

2. *There exists a solution $z^* \in \mathcal{Z}$ such that $\mathbf{0} \in F(z^*) + \partial \mathbb{I}_{\mathcal{Z}}(z^*)$.*

We define the Minty Variational Inequality problem (MVI) and the Stampacchia Variational Inequality problem (SVI) and show how their solutions relate in Appendix B.

2.2. Inclusion Problems with Negatively Comonotone Operators.

We study inclusion problem (CMI) with (non)-monotone operators that satisfies *ρ-comonotoncity* (Assumption 2), which captures (MI) (when $\rho = 0$) and a class of nonconvex-nonconcave min-max optimization problems (when $\rho < 0$). Given a single-valued and possibly *non-monotone* operator F and a *set-valued maximally monotone* operator A, we denote $E = F + A$. The inclusion problem is to find a point $z^* \in \mathbb{R}^n$ that satisfies

$$\mathbf{0} \in E(z^*) = F(z^*) + A(z^*). \tag{CMI}$$

Similar to (MI) , we say z is an ϵ-approximate solution to (CMI) if

$$\mathbf{0} \in F(z) + A(z) + \mathcal{B}(\mathbf{0}, \epsilon).$$

We summarize the assumptions on (CMI) below.

Assumption 2 *In* (CMI) *,*

1. $F : \mathbb{R}^n \to \mathbb{R}^n$ *is* L*-Lipschitz.*
2. $A : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ *is maximally monotone.*
3. $E = F + A$ *is* ρ*-comonotone, i.e., there exists* $\rho \le 0$ *such that*

$$\langle u - u', z - z' \rangle \ge \rho \|u - u'\|^2, \quad \forall (z, u), (z', u') \in \mathrm{Gra}_E .$$

4. *There exists a solution* $z^* \in \mathbb{R}^n$ *such that* $\mathbf{0} \in E(z^*)$*.*

The formulation of (CMI) provides a unified treatment for a range of problems, such as min-max optimization and multi-player games. We present one detailed example in Example 2 in Appendix B and refer readers to [12] for more examples.

2.3. Convergence Criteria

An appropriate convergence criterion is the *tangent residual* defined in [5] $r^{tan}_{F,A}(z) := \min_{c \in A(z)} \|F(z) + c\|$. It is not hard to see that $r^{tan}_{F,A}(z) \le \epsilon$ implies that z is an ϵ-approximate solution to (CMI) . If $A = \partial \mathbb{I}_{\mathcal{Z}}$, and $\mathcal{Z}$ is bounded and has diameter no more than D, then z is an ϵ-approximate solution to (MI) .

Another commonly-used convergence criterion that captures the stationarity of a solution is the *natural residual* $r^{nat}_{F,A} := \|z - J_A[z - F(z)]\|$. Note that z^* is a solution to (CMI) iff $z^* = J_A[z^* - F(z^*)]$. The definition of the natural residual for (MI) is similar: $r^{nat}_{F,\partial \mathbb{I}_{\mathcal{Z}}} := \|z - \Pi_{\mathcal{Z}}[z - F(z)]\|$.

Fact 1 *In* (CMI) *,* $r^{nat}_{F,A}(z) \le r^{tan}_{F,A}(z)$*.*

We provide a proof of Fact 1 in Appendix B. In this paper, we state our convergence rates in terms of the tangent residual $r^{tan}_{F,A}(z)$, which implies (i) convergence rates in terms of the natural residual $r^{nat}_{F,A}(z)$, and (ii) z is an approximate solution to (CMI) or (MI) . [6]

3. Optimal Monotone Inclusion via EAG

In this section, we study the constrained single-valued monotone inclusion problem (MI) with closed convex feasible set $\mathcal{Z} \subseteq \mathbb{R}^n$ and monotone and L-Lipschitz operator F, as summarized in Assumption 1. We analyse the (projected) Extra Anchored Gradient Method (EAG) , which is proposed by [37] in the unprojected form for $\mathcal{Z} = \mathbb{R}^n$. Let $z_0 \in \mathcal{Z}$ be an arbitrary starting point and $\{z_k, z_{k+\frac{1}{2}}\}_{k \ge 0}$ be the iterates of (EAG) with step size $\eta > 0$, whose update rule is as follows:

$$\begin{aligned} z_{k+\frac{1}{2}} &= \Pi_{\mathcal{Z}}\left[z_k - \eta F(z_k) + \frac{1}{k+1}(z_0 - z_k)\right], \\ z_{k+1} &= \Pi_{\mathcal{Z}}\left[z_k - \eta F(z_{k+\frac{1}{2}}) + \frac{1}{k+1}(z_0 - z_k)\right]. \end{aligned} \tag{EAG}$$

6. We also provice convergence rates for approximate solutions of the Minty Variational Inequality problem (MVI) and the Stampacchia Variational Inequality problem (SVI). Please see Appendix B for the definition and solution concepts of the Minty Variational Inequality and the Stampacchia Variational Inequality problem and Theorem 12 in Appendix C.3 for the corresponding convergence rates.

Theorem 1 *Suppose Assumption 1 holds. Let $z_0 \in \mathcal{Z}$ be any starting point and $\{z_k, z_{k+\frac{1}{2}}\}_{k\geq 0}$ be the iterates of (EAG) with step size $\eta \in (0, \frac{1}{\sqrt{3}L})$. Denote $D := \|z_0 - z^*\|^2$. Then for any $T \geq 1$,*

$$r^{tan}_{F,\mathcal{Z}}(z_T)^2 \leq \frac{44}{\eta^2 L^2(1-3\eta^2L^2)} \cdot \frac{D^2L^2}{T^2}$$

We prove a more general version of Theorem 1 in Appendix C.3 where we also provide convergence rates for the Minty Variational Inequality and the Stampacchia Variational Inequality problem.

4. Accelerated Algorithm for Inclusion Problems with Negatively Comonotone Operators

In this section, we focus on the inclusion problem (CMI) with a single-valued L-Lipschitz (possibly non-monotone) operator F and a set-valued maximally monotone operator A, where $F + A$ satisfies ρ-comonotonicity with $\rho > -\frac{1}{2L}$ (Assumption 2). We propose an accelerated forward-backward splitting algorithm (AS) that is applicable to any maximally monotone operator A. Our algorithm (AS) generalizes previous algorithms such as EAG and FEG, which are limited to the unconstrained setting. We show that (AS) enjoys the optimal convergence rate of $O(\frac{1}{T})$ via a potential function argument.

In splitting methods, the update steps in each iteration of can be categorized as the *forward steps*, where they involve the forward evaluation of a single-valued operator, or *backward steps*, where they require computing the resolvent of an operator.

Accelerated Forward-Backward Splitting (AS). Given any initial point $z_0 \in \mathbb{R}^n$ and step size $\eta > 0$, (AS) sets $c_0 = 0$ and updates $\{z_{k+\frac{1}{2}}, z_{k+1}, c_{k+1}\}_{k\geq 0}$ as follows: for $k \geq 0$,

$$\begin{aligned} z_{k+\frac{1}{2}} &= z_k + \frac{1}{k+1}(z_0 - z_k) - \frac{k(\eta + 2\rho)}{k+1} \cdot (F(z_k) + c_k) \\ z_{k+1} &= J_{\eta A}\left[z_k + \frac{1}{k+1}(z_0 - z_k) - \eta F(z_{k+\frac{1}{2}}) - \frac{2k\rho}{k+1} \cdot (F(z_k) + c_k)\right] \\ c_{k+1} &= \frac{z_k + \frac{1}{k+1}(z_0 - z_k) - \eta F(z_{k+\frac{1}{2}}) - \frac{2k\rho}{k+1} \cdot (F(z_k) + c_k) - z_{k+1}}{\eta} \end{aligned} \tag{AS}$$

Note that by definition we have $c_k \in A(z_k)$ for all $k \geq 1$. Our algorithm is inspired by FEG [20]. In particular, when $A(z) = \mathbf{0}$ for all z, i.e., the unconstrained setting, c_k is always $\mathbf{0}$, and our algorithm is identical to FEG.

It is worth noting that (AS) only requires the computation of the resolvent for operator A and only computes it once in each iteration. When A is the subdifferential of the indicator function of a closed convex set $\mathcal{Z}$, the resolvent operator is exactly the projection to $\mathcal{Z}$. Compared to (EAG) , (AS) has an additional feature that it only performs one projection (rather than two) in each iteration when applied to (MI) . Kim [17] achieves the same accelerated rate in the monotone setting, i.e., $F + A$ is monotone, by assuming the ability to compute the resolvent of the operator $F + A$ (i.e., the accelerated proximal point method) or the resolvents of both F and A (i.e., the Accelerated Douglas-Rachford method). These resolvents could be substantially more difficult to compute in many applications than the resolvent of A, for example, when $A = \partial \mathbb{I}_{\mathcal{Z}}$. We prove Theorem 2 in Appendix D.1.

Theorem 2 *Suppose Assumption 2 holds for some $\rho \in (-\frac{1}{2L}, 0]$. Let $z_0 \in \mathbb{R}^n$ be any starting point and $\{z_k, z_{k+\frac{1}{2}}\}_{k\geq 1}$ be the iterates of (AS) with step size $\eta \in (max(0, -2\rho), \frac{1}{L})$. Then for any $T \geq 1$,*

$$\min_{c \in A(z_T)} \|F(z_T) + c\|^2 = r^{tan}_{F,A}(z_T)^2 \leq \frac{4}{(\eta + 2\rho)^2 L^2} \frac{H_0^2 L^2}{T^2},$$

where $H_0^2 = 4\|z_1 - z_0\|^2 + \|z_0 - z^*\|^2 \leq \frac{4 \cdot r^{tan}(z_0)^2}{L^2} + \|z_0 - z^*\|^2$.

Remark 3 *To interpret the convergence rate, one can think of a properly selected η such that $(\eta + 2\rho)L$ is an absolute constant, and the rate is $O(\frac{H_0^2 L^2}{T^2})$.*

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein Generative Adversarial Networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 214–223. PMLR, July 2017. URL https://proceedings.mlr.press/v70/arjovsky17a.html. ISSN: 2640-3498.

[2] Heinz H. Bauschke and Patrick L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, 2011. URL https://link.springer.com/book/10.1007/978-1-4419-9467-7.

[3] Heinz H Bauschke, Walaa M Moursi, and Xianfu Wang. Generalized monotone operators and their averaged resolvents. *Mathematical Programming*, 189(1):55–74, 2021.

[4] Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust Optimization*. Princeton University Press, December 2009. ISBN 978-1-4008-3105-0. doi: 10.1515/9781400831050. URL https://www.degruyter.com/document/doi/10.1515/9781400831050/html.

[5] Yang Cai, Argyris Oikonomou, and Weiqiang Zheng. Tight Last-Iterate Convergence of the Extragradient and the Optimistic Gradient Descent-Ascent Algorithm for Constrained Monotone Variational Inequalities. Technical Report arXiv:2204.09228, arXiv, May 2022. URL http://arxiv.org/abs/2204.09228. arXiv:2204.09228 [cs, math] type: article.

[6] Bo Dai, Albert Shaw, Lihong Li, Lin Xiao, Niao He, Zhen Liu, Jianshu Chen, and Le Song. Sbeed: Convergent reinforcement learning with nonlinear function approximation. In *International Conference on Machine Learning*, pages 1125–1134. PMLR, 2018.

[7] Cong D Dang and Guanghui Lan. On the convergence properties of non-euclidean extragradient methods for variational inequalities with generalized monotone operators. *Computational Optimization and applications*, 60(2):277–310, 2015.

[8] Constantinos Daskalakis, Stratis Skoulakis, and Manolis Zampetakis. The complexity of constrained min-max optimization. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 1466–1478, 2021.

[9] Jelena Diakonikolas. Halpern iteration for near-optimal and parameter-free monotone inclusion and strong solutions to variational inequalities. In *Conference on Learning Theory*, 2020.

[10] Jelena Diakonikolas, Constantinos Daskalakis, and Michael I. Jordan. Efficient Methods for Structured Nonconvex-Nonconcave Min-Max Optimization. *arXiv:2011.00364 [cs, math, stat]*, February 2021. URL http://arxiv.org/abs/2011.00364. arXiv: 2011.00364.

[11] Simon S. Du, Jianshu Chen, Lihong Li, Lin Xiao, and Dengyong Zhou. Stochastic variance reduction methods for policy evaluation. In *International Conference on Machine Learning*, pages 1049–1058. PMLR, 2017.

[12] Francisco Facchinei and Jong-Shi Pang. *Finite-Dimensional Variational Inequalities and Complementarity Problems*. Springer, 2003. URL https://link.springer.com/book/10.1007/b97544.

[13] Noah Golowich, Sarath Pattathil, and Constantinos Daskalakis. Tight last-iterate convergence rates for no-regret learning in multi-player games. *arXiv:2010.13724 [cs, math]*, October 2020. URL http://arxiv.org/abs/2010.13724. arXiv: 2010.13724.

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[15] Eduard Gorbunov, Nicolas Loizou, and Gauthier Gidel. Extragradient Method: $O(1/K)$ Last-Iterate Convergence for Monotone Variational Inequalities and Connections With Cocoercivity. *arXiv:2110.04261 [cs, math]*, October 2021. URL http://arxiv.org/abs/2110.04261. arXiv: 2110.04261.

[16] Benjamin Halpern. Fixed points of nonexpanding maps. *Bulletin of the American Mathematical Society*, 73(6):957–961, 1967.

[17] Donghwan Kim. Accelerated proximal point method for maximally monotone operators. *Mathematical Programming*, 190(1):57–87, November 2021. ISSN 1436-4646. doi: 10.1007/s10107-021-01643-0. URL https://doi.org/10.1007/s10107-021-01643-0.

[18] Ulrich Kohlenbach. On the proximal point algorithm and its halpern-type variant for generalized monotone operators in hilbert space. *Optimization Letters*, 16(2):611–621, 2022.

[19] G. M. Korpelevich. The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756, 1976. URL https://ci.nii.ac.jp/naid/10017556617/.

[20] Sucheol Lee and Donghwan Kim. Fast extra gradient methods for smooth structured nonconvex-nonconcave minimax problems. In *Annual Conference on Neural Information Processing Systems*, 2021.

[21] Felix Lieder. On the convergence rate of the Halpern-iteration. *Optimization Letters*, 15(2): 405–418, March 2021. ISSN 1862-4472, 1862-4480. doi: 10.1007/s11590-020-01617-9. URL https://link.springer.com/10.1007/s11590-020-01617-9.

[22] Mingrui Liu, Youssef Mroueh, Jerret Ross, Wei Zhang, Xiaodong Cui, Payel Das, and Tianbao Yang. Towards better understanding of adaptive gradient algorithms in generative adversarial nets. In *International Conference on Learning Representations*, 2019.

[23] Mingrui Liu, Hassan Rafique, Qihang Lin, and Tianbao Yang. First-order convergence theory for weakly-convex-weakly-concave min-max problems. *Journal of Machine Learning Research*, 22(169):1–34, 2021.

[24] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[25] Yura Malitsky. Golden ratio algorithms for variational inequalities. *Mathematical Programming*, 184(1):383–410, 2020.

[26] Arkadi 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*, 15(1):229–251, 2004. Publisher: SIAM.

[27] Yurii Nesterov. Dual extrapolation and its applications to solving variational inequalities and related problems. *Mathematical Programming*, 109(2):319–344, March 2007. ISSN 1436-4646. doi: 10.1007/s10107-006-0034-z. URL https://doi.org/10.1007/s10107-006-0034-z.

[28] Yuyuan Ouyang and Yangyang Xu. Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems. *Mathematical Programming*, 185(1):1–35, 2021.

[29] Jisun Park and Ernest K Ryu. Exact optimal accelerated complexity for fixed-point iterations. *arXiv preprint arXiv:2201.11413*, 2022.

[30] Thomas Pethick, Panagiotis Patrinos, Olivier Fercoq, Volkan Cevherå, et al. Escaping limit cycles: Global convergence for constrained nonconvex-nonconcave minimax problems. In *International Conference on Learning Representations*, 2022.

[31] Ernest K Ryu and Stephen Boyd. Primer on monotone operator methods. *Appl. Comput. Math*, 15(1):3–43, 2016.

[32] Ernest K Ryu and Stephen Boyd. A PRIMER ON MONOTONE OPERATOR METHODS. *APPL. COMPUT. MATH.*, page 41, 2016.

[33] Ernest K. Ryu and Wotao Yin. *Large-Scale Convex Optimization via Monotone Operators*. To be published with Cambridge University Press, 2022.

[34] Chaobing Song, Zhengyuan Zhou, Yichao Zhou, Yong Jiang, and Yi Ma. Optimistic dual extrapolation for coherent non-monotone variational inequalities. *Advances in Neural Information Processing Systems*, 33:14303–14314, 2020.

[35] Quoc Tran-Dinh. The connection between nesterov's accelerated methods and halpern fixed-point iterations. *arXiv preprint arXiv:2203.04869*, 2022.

[36] Quoc Tran-Dinh and Yang Luo. Halpern-type accelerated and splitting algorithms for monotone inclusions. *arXiv preprint arXiv:2110.08150*, 2021.

[37] TaeHo Yoon and Ernest K. Ryu. Accelerated Algorithms for Smooth Convex-Concave Minimax Problems with \$\mathcal{O}(1/k^2)\$ Rate on Squared Gradient Norm. *arXiv:2102.07922 [math]*, June 2021. URL http://arxiv.org/abs/2102.07922. arXiv: 2102.07922.

[38] Zhengyuan Zhou, Panayotis Mertikopoulos, Nicholas Bambos, Stephen Boyd, and Peter W Glynn. Stochastic mirror descent in variationally coherent optimization problems. *Advances in Neural Information Processing Systems*, 30, 2017.

Contents

Appendix A. Detailed Related Works

There is a vast literature on inclusion problems and variational inequalities, e.g., see [2, 12, 31] and the references therein. We only provide a brief discussion of the most relevant and recent results.

A.1. Convex-Concave and Monotone Settings

Convergence in Gap Function. Nemirovski and Nesterov [26, 27] show that the average iterate of extragradient-type methods has $O(\frac{1}{T})$ convergence rate in terms of gap function defined as $\max_{z'\in\mathcal{Z}} \langle F(z'), z - z'\rangle$, which means that their result only provides an approximate solution to the weak solution. The $O(\frac{1}{T})$ rate is optimal for first-order methods due to the lower bound by Ouyang and Xu [28].

Convergence of the Extragradient Method in Stronger Performance Measures. For stronger performance measures such as the norm of the operator (when $\mathcal{Z} = \mathbb{R}^n$) or the residual (in constrained setting), classical results [12, 19] show that the best-iterate of the extragradient method converges at a rate of $O(\frac{1}{\sqrt{T}})$. Recently, the same convergence rate is shown to hold even for the last-iterate of the extragradient method [5, 15]. Although $O(\frac{1}{\sqrt{T}})$ convergence on the residual is optimal for all p-SCIL algorithms [13], a subclass of first-order methods that includes the extragradient method and many of its variations, faster rate is possible for other first-order methods.

Faster Convergence Rate in Operator Norm or Residual. We provide a brief overview of results that achieve faster convergence rate in terms of the operator norm or residual. Note that these results also imply essentially the same convergence rate in terms of the gap function. The literature here is rich and fast-growing, we only discuss the ones that are close related to our paper. Recent results show accelerated rates through Halpern iteration [16] or a similar mechanism – anchoring. Implicit versions of Halpern iteration have $O(\frac{1}{T})$ convergence rate [17, 21, 29] for monotone operators and explicit variants of Halpern iteration achieve the same convergence rate when F is cocoercive [9, 17]. Diakonikolas [9] also provide a double-loop implementation of the algorithm for monotone operators at the expense of an additional logarithmic factor in the convergence rate. Yoon and Ryu [37] propose the extra anchored gradient (EAG) method, which is the first explicit method with accelerated $O(\frac{1}{T})$ rate in the unconstrained setting for monotone operators. They also established a matching $\Omega(\frac{1}{T})$ lower bound that holds for all first-order methods. Convergence analysis of past extragradient method with anchoring in the unconstrained setting is provided in [36]. Lee and Kim [20] proposed a generalization of EAG called fast extraradient (FEG), which applies to comonotone operators and improves the constants in the convergence rate, but their result only applies to the unconstrained setting. Very recently, Tran-Dinh [35] studies the connection between Halpern iteration and Nesterov accelerated method, and provides new algorithms for monotone operators and alternative analyses for EAG and FEG in the unconstrained setting. In Theorem 1, we show the projected version of EAG has $O(\frac{1}{T})$ convergence rate under arbitrary convex constraints, achieving the optimal convergence rate for all first-order methods in the constrained setting.

A.2. Nonconvex-Nonconcave Min-Max Optimization and Inclusions with Non-Monotone Operators

Many practical applications of min-max optimization in modern machine learning, such as GANs and multi-agent reinforcement learning, are nonconvex-nonconcave. Without any additional structure, the problem is intractable [8]. Hence, recent works study nonconvex-nonconcave min-max optimization problems under several structural assumptions. We only introduce the definitions in the unconstrained setting, as that is the setting considered by several recent results, and all convergence rates are with respect to the the operator norm. The *Minty variational inequality* (MVI) condition (also called coherence or variationally stable): there exits z^* such that

$$\langle F(z), z - z^* \rangle \geq 0, \quad \forall z \in \mathbb{R}^n$$

is studied in e.g., [7, 22, 23, 25, 34, 38]. Extragradient-type algorithms has $O(\frac{1}{\sqrt{T}})$ convergence rate for Lipschitz operators that satisfy the MVI condition [7]. Diakonikolas [10] proposes a weaker condition called *weak MVI*: there exits z^* and $\rho < 0$ such that

$$\langle F(z), z - z^* \rangle \geq \rho \cdot \|F(z)\|^2, \quad \forall z \in \mathbb{R}^n.$$

The weak MVI condition includes both MVI and negative comonotonicity [3] as special cases. Diakonikolas [10] proposes the EG+ algorithm, which has $O(\frac{1}{\sqrt{T}})$ convergence rate under the weak MVI condition in the unconstrained setting. Recently, Pethick et al [30] generalized EG+ to CEG+ algorithm which has $O(\frac{1}{\sqrt{T}})$ convergence rate under weak MVI condition in general (constrained) setting. The result for accelerated algorithms in the nonconvex-nonconcave setting is sparser. FEG achieves $O(\frac{1}{T})$ convergence rate for comonotone operators in the unconstrained setting [20]. In general (constrained) setting with comonotone operators, implicit methods such as the proximal point algorithm has been shown to converge [3, 18]. To the best of our knowledge, (AS) is the first explicit and efficient method that achieves the accelerated and optimal $O(\frac{1}{T})$ convergence rate in the constrained nonconvex-nonconcave setting (Theorem 2). We summarize previous results and our results in Table 1.

Appendix B. Additional Preliminaries

Maximally Monotone Operator. When A is maximally monotone, useful properties of J_A (See e.g., [31, 33]) include:

1. J_A is well-defined on $\mathbb{R}^n$;
2. J_A is non-expansive thus single-valued;
3. when $z = J_A(z')$, then $z' - z \in A(z)$;
4. when $A = \partial \mathbb{I}_{\mathcal{Z}}$ is the normal cone operator of a closed convex set $\mathcal{Z} \subseteq \mathbb{R}^n$, then $J_{\eta A} = \Pi_{\mathcal{Z}}$ is the projection operator for all $\eta > 0$.

Monotone Inclusion

Remark 4 *We remind readers that in general, monotone inclusion problem is to find $z^* \in \mathbb{R}^n$ such that $\mathbf{0} \in A(z^*)$, where A is a set-valued maximally monotone operator. To ease notation, we sometimes refer to the constrained single-valued monotone inclusion problem stated in* (MI) *as the* monotone inclusion problem.

Variational Inequality. A closely related problem to (MI) is the *monotone variational inequality* (VI) with operator F and feasible set $\mathcal{Z}$, which has two variants. The *Stampacchia Variational Inequality* (SVI) problem is to find $z^* \in \mathcal{Z}$ such that

$$\langle F(z^*), z^* - z \rangle \le 0, \quad \forall z \in \mathcal{Z}. \tag{SVI}$$

Such z^* is called a *strong* solution to VI. The *Minty Variational Inequality* (MVI) problem is to find $z^* \in \mathcal{Z}$ such that

$$\langle F(z), z^* - z \rangle \le 0, \quad \forall z \in \mathcal{Z}. \tag{MVI}$$

Such z^* is called a *weak* solution to VI. When F is continuous, then every solution to (MVI) is also a solution to (SVI). When F is monotone, every solution to (SVI) is also a solution to (MVI) and thus the two solution sets are equivalent. Moreover, the solution set to (MI) is the same as the solution set to (SVI).

Approximate Solutions. We say $z \in \mathcal{Z}$ is an ϵ-approximate solution to (MI) if

$$\mathbf{0} \in F(z) + \partial \mathbb{I}_{\mathcal{Z}}(z) + \mathcal{B}(\mathbf{0}, \epsilon),$$

where we use $\mathcal{B}(u, r)$ to denote a ball in $\mathbb{R}^n$ centered at u with radius r. We say $z \in \mathcal{Z}$ is an ϵ-approximate solution to (SVI) or (MVI) if

$$\begin{aligned} &\langle F(z), z - z' \rangle \leq \epsilon, \forall z' \in \mathcal{Z}, \text{ or} \\ &\langle F(z'), z - z' \rangle \leq \epsilon, \forall z' \in \mathcal{Z}, \text{ respectively.} \end{aligned}$$

When F is monotone, it is clear that every ϵ-approximate solution to (SVI) is also an ϵ-approximate solution to (MVI); but the reverse does not hold in general. When F is monotone and $\mathcal{Z}$ is bounded by D, then any $\frac{\epsilon}{D}$-approximate solution to (MI) is an ϵ-approximate solution to (SVI) [9, Fact 1]. Note that when $\mathcal{Z}$ is unbounded, neither (SVI) nor (MVI) can be approximated. If we restrict the domain to be a bounded subset of (possibly unbounded) $\mathcal{Z}$, then we can define the (restricted) gap functions as

$$\begin{aligned} \text{GAP}^{SVI}_{F,D}(z) &:= \max_{z' \in \mathcal{Z} \cap \mathcal{B}(z,D)} \langle F(z), z - z' \rangle, \\ \text{GAP}^{MVI}_{F,D}(z) &:= \max_{z' \in \mathcal{Z} \cap \mathcal{B}(z,D)} \langle F(z'), z - z' \rangle. \end{aligned}$$

The $O(\frac{1}{T})$ convergence rate for extragradient-type algorithm [26, 27] is provided in terms of $\text{GAP}^{MVI}_{F,D}(z)$, which means convergence to an approximate *weak* solution. Prior to our work, the $O(\frac{1}{T})$ convergence rate on $\text{GAP}^{SVI}_{F,D}(z)$ was only known in the unconstrained setting [37]. When F is monotone, then the tangent residual $r^{tan}_{F,D}(z) \leq \frac{\epsilon}{D}$ (definition in section 2.3) implies $\text{GAP}^{SVI}_{F,D}(z) \leq \epsilon$ [5, Lemma 2]. Therefore, our result also implies an $O(\frac{1}{T})$ convergence rate on $\text{GAP}^{SVI}_{F,D}(z)$ when $\mathcal{Z}$ is arbitrary convex set (Theorem 1).

Example 1 (Gap function is weaker than natural residual) *Consider an instance of the Monotone VI problem on the identity operator $F(x) = x$ in $\mathcal{Z} = [0, 1]$.*

- *Observe that the natural residual on $x \in \mathcal{Z}$ is $\|x - \Pi_{\mathcal{Z}}[x - F(x)]\| = x$.*
- *Moreover, since $\mathcal{Z} = [0, 1]$, observe that for any $x \in \mathcal{Z}$ and $D \geq 0$,*

$$\begin{aligned} \text{GAP}^{SVI}_{F,D}(x) \leq \text{GAP}^{SVI}_{F,1}(x) &= \max_{x' \in [0,1]} x \cdot (x - x') = x^2, \textit{ and} \\ \text{GAP}^{MVI}_{F,D}(x) \leq \text{GAP}^{MVI}_{F,1}(x) &= \max_{x' \in [0,1]} x' \cdot (x - x') = \frac{x^2}{4}. \end{aligned}$$

As a result, any algorithm with $O(\frac{1}{T})$ convergence rate with respect to the gap function only implies a $O(\frac{1}{\sqrt{T}})$ convergence rate for the corresponding (MI) or the natural residual.

Inclusion Problems with Negatively Comonotone Operators

Example 2 (Min-Max Optimization) *The following structured min-max optimization problem captures a wide range of applications in machine learning such as GANs, adversarial examples, robust optimization, and reinforcement learning:*

$$\min_{x \in \mathbb{R}^{n_x}} \max_{y \in \mathbb{R}^{n_y}} f(x,y) + g(x) - h(y), \tag{1}$$

where $f(\cdot,\cdot)$ is possibly non-convex in x and non-concave in y. Regularized and constrained min-max problems are covered by appropriate choices of lower semicontinuous and convex functions g and h. Examples include ℓ_1-norm, ℓ_2-norm, and indicator function of a convex feasible set. Let $z = (x,y)$, if we define $F(z) = (\partial_x f(x,y), -\partial_y f(x,y))$ and $A(z) = (\partial g(x), \partial h(y))$, where A is maximally monotone, then the first-order optimality condition of (1) has the form of (CMI) . See [20, Example 1] for examples of nonconvex-nonconcave conditions that are implied by negative comonotonicity.

Proof of Fact 1 **Proof** For any $c \in A(z)$, we have

$$\begin{aligned} r^{nat}_{F,A}(z) &= \|z - J_A[z - F(z)]\| \\ &= \|J_A[z+c] - J_A[z - F(z)]\| && (z = J_A[z+c]) \\ &\le \|F(z) + c\|. && \text{(non-expansiveness of } J_A) \end{aligned}$$

Thus $r^{nat}_{F,A}(z) \le \min_{c \in A(z)} \|F(z) + c\| = r^{tan}_{F,A}(z)$. ■

Appendix C. Missing Proofs in Section 3: Optimal Monotone Inclusion via EAG

Our analysis is based on the following potential function V_k: for $k \ge 1$,

$$\begin{aligned} V_k &:= \frac{k(k+1)}{2} \cdot \|\eta F(z_k) + \eta c_k\|^2 + k \cdot \langle \eta F(z_k) + \eta c_k, z_k - z_0 \rangle \\ \text{where} \quad c_k &:= \frac{z_{k-1} - \eta F(z_{k-\frac{1}{2}}) + \frac{1}{k}(z_0 - z_{k-1}) - z_k}{\eta}. \end{aligned} \tag{2}$$

From the update rule of (EAG) , we know $c_k \in N_{\mathcal{Z}}(z_k)$. Thus $\|F(z_k) + c_k\| \ge \min_{c \in N_{\mathcal{Z}}(z_k)} \|F(z_k) + c\| = r^{tan}(z_k)$. In Theorem 9, we show V_k is "approximately" non-increasing, i.e., $V_{k+1} \le V_k + O(1) \cdot \|\eta F(z_{k+1}) + \eta c_{k+1}\|^2$ for all $k \ge 1$. From this property, we get $O(\frac{1}{T})$ last-iterate convergence rate in terms of $\|F(z_T) + c_T\|$ and thus the same convergence rate in $r^{tan}(z_T)$ (Theorem 1).

Potential functions in a similar form as V_k have been used to analyse (MI) by Diakonikolas [9] for the Halpern iteration algorithm, and by Yoon and Ryu [37] for (EAG) in the unconstrained setting ($\mathcal{Z} = \mathbb{R}^n$). We emphasize that we use a different potential function with different analysis.

Diakonikolas [9] studied the Halpern iteration with operator $P := I - J_{F + \partial \mathbb{I}_{\mathcal{Z}}}$, which is $\frac{1}{2}$-cocoercive but can not be computed efficiently in general. She showed that the follwoing potential function is non-increasing.

$$P_k := \frac{k(k+1)}{2} \cdot \|P(z_k)\|^2 + k \cdot \langle P(z_k), z_k - z_0 \rangle,$$

which leads to $O(\frac{1}{T})$-approximate solution to (MI) after T iterations. However, since P can not be computed efficiently in general, the algorithm needs $O(\log(\frac{1}{\epsilon}))$ oracle queries for an $O(\epsilon)$-approximate value of P in each iteration, thus total oracle complexity $O(\frac{LD}{\epsilon} \cdot \log(\frac{1}{\epsilon}))$ for an ϵ-approximate solution to (MI) . In contrast, we use operator F in the potential function V_k, and we prove V_k is only "approximately" non-increasing (see Theorem 5 and 9). Moreover, (EAG) needs only 2 oracle queries in each iteration and achieves optimal $O(\frac{LD}{\epsilon})$ oracle complexity for an ϵ-approximate solution to (MI) (Theorem 1) matching the lower complexity bound $\Omega(\frac{LD}{\epsilon})$ [9].

Yoon and Ryu [37] studied convergence of (EAG) for (MI) in the unconstrained setting ($\mathcal{Z} = \mathbb{R}^n$). The specific algorithm they analysed uses anchoring term $\frac{1}{k+2}(z_0 - z_k)$ while we use $\frac{1}{k+1}(z_0 - z_k)$ (see Remark 10 for more discussion on the choice of the constant in the anchoring term). They use the following potential function

$$P_k := A_k \cdot \|F(z_k)\|^2 + B_k \cdot \langle F(z_k), z_k - z_0 \rangle,$$

where $B_k = k + 1$, and $A_k = O(k^2)$ is updated adaptively in a sophisticated way for each k. Their potential function P_k is more complicated compared to V_k as we choose $B_k = k$ and $A_k = \frac{k(k+1)}{2}$. For the analysis, their proof of the monotonicity of P_k is relatively involved. In contrast, we use a simple proof to show that V_k is "approximately" non-increasing (Theorem 5) which suffices to establish the $O(\frac{1}{T})$ convergence rate. Moreover, our analysis can be naturally extended to the constrained setting where $\mathcal{Z} \subseteq \mathbb{R}^n$ is an arbitrary closed convex set (Theorem 9).

C.1. Warm Up: Unconstrained Case

We begin with the unconstrained setting $\mathcal{Z} = \mathbb{R}^n$, which illustrates our main idea and proof techniques. Yoon and Ryu [37] also analyse the unconstrained setting but our proof is much simpler.

In the unconstrained setting, $c_k = 0$ by definition. Thus

$$V_k = \frac{k(k+1)}{2} \|\eta F(z_k)\|^2 + k \langle \eta F(z_k), z_k - z_0 \rangle, \forall k \geq 1.$$

It is not hard to see that $V_1 \leq (\eta^2 L^2 + 2\eta L)\|z_0 - z^*\|^2$: since the update rule for $z_{\frac{1}{2}}$ and z_1 of (EAG) coincides with the update rule of EG, by [5, Theorem 1], we have $\|\eta F(z_1)\|^2 \leq \|\eta F(z_0)\|^2 \leq \eta^2 L^2 \|z_0 - z^*\|^2$; by [19] and [12, Lemma 12.1.10], we have $\|z_1 - z^*\| \leq \|z_0 - z^*\|$

$$\begin{aligned} V_1 &= \|\eta F(z_1)\|^2 + \langle \eta F(z_1), z_1 - z_0 \rangle \\ &\leq \|\eta F(z_0)\|^2 + \|\eta F(z_1)\|(\|z_1 - z^*\| + \|z_0 - z^*\|) \\ &\leq (\eta^2 L^2 + 2\eta L)\|z_0 - z^*\|^2. \end{aligned}$$

Theorem 5 *Suppose Assumption 1 holds with $\mathcal{Z} = \mathbb{R}^n$. Then for any $k \geq 1$,* (EAG) *with any step size $\eta \in (0, \frac{1}{L})$ satisfies $V_{k+1} \leq V_k + \frac{\eta^2 L^2}{1-\eta^2 L^2} \|\eta F(z_{k+1})\|^2$.*

Proof Since F is monotone and L-Lipschitz, we have the following inequalities

$$\langle F(z_{k+1}) - F(z_k), z_k - z_{k+1} \rangle \leq 0$$

and

$$\left\| F(z_{k+\frac{1}{2}}) - F(z_{k+1}) \right\|^2 - L^2 \left\| z_{k+\frac{1}{2}} - z_{k+1} \right\|^2 \leq 0.$$

We simplify them using the update rule of (EAG) .

In particular, we replace $z_k - z_{k+1}$ with $\eta F(z_{k+\frac{1}{2}}) - \frac{1}{k+1}(z_0 - z_k)$ and $z_{k+\frac{1}{2}} - z_{k+1}$ with $\eta F(z_{k+\frac{1}{2}}) - \eta F(z_k)$.

$$\left\langle \eta F(z_{k+1}) - \eta F(z_k), \eta F(z_{k+\frac{1}{2}}) - \frac{1}{k+1}(z_0 - z_k) \right\rangle \leq 0, \tag{3}$$

$$\left\| \eta F(z_{k+\frac{1}{2}}) - \eta F(z_{k+1}) \right\|^2 - \eta^2 L^2 \left\| \eta F(z_{k+\frac{1}{2}}) - \eta F(z_k) \right\|^2 \leq 0. \tag{4}$$

It is not hard to verify that the following identity holds.

$$\begin{aligned}
&V_k - V_{k+1} + k(k+1) \cdot \text{LHS of Inequality(3)} + \frac{k(k+1)}{2\eta^2 L^2} \cdot \text{LHS of Inequality(4)} \\
&= \frac{k+1}{2\eta^2 L^2} \left\| \frac{(\eta^2 L^2 - 1)k + \eta^2 L^2}{\sqrt{(1-\eta^2 L^2)k}} \cdot \eta F(z_{k+1}) + \sqrt{(1-\eta^2 L^2)k} \cdot \eta F(z_{k+\frac{1}{2}}) \right\|^2 \\
&\quad - \frac{k+1}{2k} \cdot \frac{\eta^2 L^2}{1-\eta^2 L^2} \|\eta F(z_{k+1})\|^2.
\end{aligned}$$

Note that $\frac{k+1}{2k} \leq 1$ holds for all $k \geq 1$. Thus, $V_{k+1} \leq V_k + \frac{\eta^2 L^2}{1-\eta^2 L^2} \|\eta F(z_{k+1})\|^2$. ∎

Lemma 6 *For all $k \geq 2$,*

$$\begin{aligned}
&\left(\frac{k(k+1)}{4} - \frac{\eta^2 L^2}{1-\eta^2 L^2} \right) \|\eta F(z_k)\|^2 \\
&\leq (1+\eta L)^2 \|z_0 - z^*\|^2 + \frac{\eta^2 L^2}{1-\eta^2 L^2} \sum_{t=2}^{k-1} \|\eta F(z_t)\|^2.
\end{aligned}$$

Moreover, when $\eta \in (0, \frac{1}{\sqrt{3}L})$, we have

$$\frac{k^2}{4} \cdot \|\eta F(z_k)\|^2 \leq (1+\eta L)^2 \|z_0 - z^*\|^2 + \frac{\eta^2 L^2}{1-\eta^2 L^2} \sum_{t=2}^{k-1} \|\eta F(z_t)\|^2.$$

Proof Fix any $k \geq 2$. By definition, we have

$$\begin{aligned}
V_k &= \frac{k(k+1)}{2} \|\eta F(z_k)\|^2 + k\langle \eta F(z_k), z_k - z_0 \rangle \\
&\geq \frac{k(k+1)}{2} \|\eta F(z_k)\|^2 + k\langle \eta F(z_k), z^* - z_0 \rangle && (\langle F(z_k), z^* - z_k \rangle \leq 0) \\
&\geq \frac{k(k+1)}{2} \|\eta F(z_k)\|^2 - \frac{k(k+1)}{4} \|\eta F(z_k)\|^2 - \frac{k^2}{k(k+1)} \|z_0 - z^*\|^2 \\
& && (\langle a, b \rangle \geq -\tfrac{c}{4}\|a\|^2 - \tfrac{1}{c}\|b\|^2) \\
&\geq \frac{k(k+1)}{4} \|\eta F(z_k)\|^2 - \|z_0 - z^*\|^2. && (\tfrac{k}{k+1} \leq 1)
\end{aligned}$$

Using Theorem 5, we have

$$V_k \le V_1 + \frac{\eta^2 L^2}{1-\eta^2 L^2} \sum_{t=2}^{k} \|\eta F(z_t)\|^2.$$

Combining the two inequalities above and the fact that $V_1 \le (2\eta L + \eta^2 L^2)\|z_0 - z^*\|^2$ yields the first inequality in the statement. When $\eta L \in (0, \frac{1}{\sqrt{3}})$, we have $\frac{\eta^2 L^2}{1-\eta^2 L^2} \le \frac{1}{2} \le \frac{k}{4}$ for $k \ge 2$. Hence the second inequality in the statement holds. ■

Theorem 7 *Suppose Assumption 1 holds with $\mathcal{Z} = \mathbb{R}^n$. Let $z_0 \in \mathbb{R}^n$ be arbitrary starting point and $\{z_k, z_{k+\frac{1}{2}}\}_{k\ge 0}$ be the iterates of (EAG) with any step size $\eta \in (0, \frac{1}{\sqrt{3}L})$. Denote $D := \|z_0 - z^*\|$. Then for any $T \ge 1$,*

$$\|F(z_T)\|^2 \le \frac{4(1+\eta L)^2}{\eta^2 L^2 (1 - 3\eta^2 L^2)} \cdot \frac{D^2 L^2}{T^2},$$

$$\textsc{Gap}^{SVI}_{F,D}(z_T) \le \frac{2(1+\eta L)}{\eta L \sqrt{1 - 3\eta^2 L^2}} \cdot \frac{D^2 L}{T}.$$

If we set $\eta = \frac{1}{3L}$, then $\|F(z_T)\|^2 \le \frac{96 \cdot D^2 L^2}{T^2}$.

Proof Note that the second inequality is implied by the first inequality since $\textsc{Gap}^{SVI}_{F,D}(z) \le D \cdot \|F(z_T)\|$ [5, Lemma 2]. Denote $a_k := \frac{\|\eta F(z_k)^2\|}{\|z_0 - z^*\|^2}$. It suffices to prove for all $k \ge 1$,

$$a_k \le \frac{4(1+\eta L)^2}{(1 - 3\eta^2 L^2)k^2}. \tag{5}$$

Since the update rule for $z_{\frac{1}{2}}$ and z_1 of (EAG) coincides with the update rule of EG, by [5, Theorem 1], we have $\|\eta F(z_1)\|^2 \le \|\eta F(z_0)\|^2 \le \eta^2 L^2 \|z_0 - z^*\|^2$ and thus $a_1 \le \eta^2 L^2 < \frac{1}{3}$. Thus (5) holds for $k = 1$.

From Lemma 6, we know for $k \ge 2$,

$$\frac{k^2}{4} \cdot a_k \le (1+\eta L)^2 + \frac{\eta^2 L^2}{1-\eta^2 L^2} \sum_{t=2}^{k-1} a_t.$$

Applying Proposition 16 with $C_1 = (1+\eta L)^2$ and $p = \eta^2 L^2 < \frac{1}{3}$ completes the proof. ■

C.2. Convergence of EAG with Arbitrary Convex Constraints

In this section, we show how to extend the analysis in the unconstrained setting to the arbitrary convex constrained setting. Recall that the potential function V_k and c_k are defined in (2). We use the fact that $c_k \in N_{\mathcal{Z}}(z_k)$ extensively.

Proposition 8 $V_1 \le \frac{(1+\eta L + \eta^2 L^2)(2 + 2\eta L + \eta^2 L^2)}{1 - \eta^2 L^2} \|z_0 - z^*\|^2.$

Proof We first upper bound $\|\eta F(z_1)+\eta c_1\|$ and $\|z_1-z_0\|$. Note that $z_{\frac{1}{2}}, z_1$ are updated exactly as original EG. By definition, we have

$$\begin{aligned}
\|\eta F(z_1)+\eta c_1\| &= \left\|\eta F(z_1)+z_0-\eta F(z_{\frac{1}{2}})-z_1\right\| \\
&\le \left\|\eta F(z_1)-\eta F(z_{\frac{1}{2}})\right\|+\|z_0-z_1\| \\
&\le \eta L\left\|z_1-z_{\frac{1}{2}}\right\|+\|z_0-z_1\| && (L\text{-Lipschitzness of } F) \\
&\le (1+\eta L)\left\|z_1-z_{\frac{1}{2}}\right\|+\left\|z_{\frac{1}{2}}-z_0\right\| \\
&\le (1+\eta L+\eta^2L^2)\left\|z_{\frac{1}{2}}-z_0\right\| \\
&\le \frac{1+\eta L+\eta^2L^2}{\sqrt{1-\eta^2L^2}}\|z_0-z^*\|,
\end{aligned}$$

where in the last inequality we use a well-known result regarding EG [12, Lemma 12.1.10]: $\|z_{\frac{1}{2}}-z_0\|^2 \le \frac{\|z_0-z^*\|^2-\|z_1-z^*\|^2}{1-\eta^2L^2}$. Note that in the above sequence of inequalities, we also prove that $\|z_1-z_0\| \le \frac{1+\eta L}{\sqrt{1-\eta^2L^2}}\|z_0-z^*\|$.

By definition of V_1 and the above upper bound for $\|\eta F(z_1)+\eta c_1\|$ and $\|z_1-z_0\|$, we have

$$\begin{aligned}
V_1 &= \|\eta F(z_1)+\eta c_1\|^2+\langle \eta F(z_1)+\eta c_1, z_1-z_0\rangle \\
&\le \|\eta F(z_1)+\eta c_1\|^2+\|\eta F(z_1)+\eta c_1\|\cdot\|z_1-z_0\| \\
&\le \frac{(1+\eta L+\eta^2L^2)(2+2\eta L+\eta^2L^2)}{1-\eta^2L^2}\|z_0-z^*\|^2.
\end{aligned}$$

■

Theorem 9 *Suppose Assumption 1 holds. Let $z_0\in\mathcal{Z}$ be any starting point and $\{z_k, z_{k+\frac{1}{2}}\}_{k\ge 0}$ be the iterates of (EAG) with step size $\eta\in(0,\frac{1}{L})$. Then for any $k\ge 1$, $V_{k+1}\le V_k+\frac{\eta^2L^2}{1-\eta^2L^2}\|\eta F(z_{k+1})+\eta c_{k+1}\|^2$, where V_k and c_k are defined in (2).*

Proof We first present several inequalities. From the monotonicity and L-Lipschitzness of F, we have

$$(-\frac{k(k+1)}{2\eta^2L^2})\cdot\left(\eta^2L^2\cdot\left\|z_{k+\frac{1}{2}}-z_{k+1}\right\|^2-\left\|\eta F(z_{k+\frac{1}{2}})-\eta F(z_{k+1})\right\|^2\right)\le 0, \tag{6}$$

$$(-k(k+1))\cdot\langle \eta F(z_{k+1})-\eta F(z_k), z_{k+1}-z_k)\rangle \le 0. \tag{7}$$

Since $z_{k+\frac{1}{2}}=\Pi_{\mathcal{Z}}[z_k-\eta F(z_k)+\frac{1}{k+1}(z_0-z_k)]$, we can infer that $z_k-\eta F(z_k)+\frac{1}{k+1}(z_0-z_k)-z_{k+\frac{1}{2}}\in N_{\mathcal{Z}}(z_{k+\frac{1}{2}})$. Moreover, by definition of c_k and c_{k+1}, we know $c_k\in N_{\mathcal{Z}}(z_k)$ and $c_{k+1}\in N_{\mathcal{Z}}(z_{k+1})$. Therefore, we have

$$(-k(k+1))\cdot\left\langle z_k-\eta F(z_k)-z_{k+\frac{1}{2}}+\frac{1}{k+1}(z_0-z_k), z_{k+\frac{1}{2}}-z_{k+1}\right\rangle\le 0, \tag{8}$$

$$(-k(k+1))\cdot\langle \eta c_{k+1}, z_{k+1}-z_k\rangle\le 0, \tag{9}$$

$$(-k(k+1))\cdot\left\langle \eta c_k, z_k-z_{k+\frac{1}{2}}\right\rangle\le 0. \tag{10}$$

The following identity holds when we substitute ηc_{k+1} on both sides using $\eta c_{k+1} = z_k - \eta F(z_{k+\frac{1}{2}}) + \frac{1}{k+1}(z_0 - z_k) - z_{k+1}$, which follows from the definition. The correctness of the identity follows from Identity (19) in Proposition 15: we treat x_0 as z_0; x_t as $z_{k+\frac{t-1}{2}}$ for $t \in \{1,2,3\}$; y_t as $\eta F(z_{k+\frac{t-1}{2}})$ for $t \in \{1,2,3\}$; u_1 as ηc_k and u_3 as ηc_{k+1}; p as $\eta^2 L^2$ and q as k.

$$
\begin{aligned}
& V_k - V_{k+1} + \text{LHS of Inequality (6)} + \text{LHS of Inequality (7)} \\
& + \text{LHS of Inequality (8)} + \text{LHS of Inequality (9)} + \text{LHS of Inequality (10)} \\
= & \frac{k(k+1)}{2} \cdot \left\| z_{k+\frac{1}{2}} - z_k + \eta F(z_k) + \eta c_k + \frac{1}{k+1}(z_k - z_0) \right\|^2 \qquad (11) \\
& + \frac{(1-\eta^2 L^2)k(k+1)}{2\eta^2 L^2} \cdot \left\| \eta F(z_{k+\frac{1}{2}}) - \eta F(z_{k+1}) \right\|^2 \qquad (12) \\
& + (k+1) \cdot \left\langle \eta F(z_{k+\frac{1}{2}}) - \eta F(z_{k+1}), \eta F(z_{k+1}) + \eta c_{k+1} \right\rangle . \qquad (13)
\end{aligned}
$$

Since $\|a\|^2 + \langle a, b \rangle = \|a + \frac{b}{2}\|^2 - \frac{\|b\|^2}{4}$, we have

$$
\begin{aligned}
& \text{Expression(12)} + \text{Expression(13)} \\
= & \left\| A \cdot \left(\eta F(z_{k+\frac{1}{2}}) - \eta F(z_{k+1}) \right) + B \cdot (\eta F(z_{k+1}) + \eta c_{k+1}) \right\|^2 \\
& - \frac{k+1}{2k} \cdot \frac{\eta^2 L^2}{1 - \eta^2 L^2} \|\eta F(z_{k+1}) + \eta c_{k+1}\|^2,
\end{aligned}
$$

where $A = \sqrt{\frac{(1-\eta^2L^2)k(k+1)}{2\eta^2L^2}}$ and $B = \sqrt{\frac{\eta^2L^2(k+1)}{2(1-\eta^2L^2)k}}$. Since $k \geq 1$, we have $\frac{k+1}{2k} \leq 1$. Hence, we have $V_{k+1} \leq V_k + \frac{\eta^2L^2}{1-\eta^2L^2}\|\eta F(z_{k+1}) + \eta c_{k+1}\|^2$. ■

Remark 10 *The proof of Theorem 9 naturally extends to the following algorithm and potential function: Fix any $z_0 \in \mathcal{Z}$ and $\eta \in (0, \frac{1}{L}), \delta \geq 0$. Update $z_{\frac{1}{2}}$, z_1, c_1, V_1 as (EAG) and for $k \geq 1$:*

$$
\begin{aligned}
z_{k+\frac{1}{2}} &= \Pi_{\mathcal{Z}} \left[z_k - \eta F(z_k) + \frac{1}{k+\delta+1}(z_0 - z_k) \right], \\
z_{k+1} &= \Pi_{\mathcal{Z}} \left[z_k - \eta F(z_{k+\frac{1}{2}}) + \frac{1}{k+\delta+1}(z_0 - z_k) \right], \\
c_{k+1} &= \frac{z_k - \eta F(z_{k+\frac{1}{2}}) + \frac{1}{k+\delta+1}(z_0 - z_k) - z_{k+1}}{\eta}, \\
V_{k+1} &= \frac{(k+\delta+1)(k+\delta+2)}{2} \|\eta F(z_{k+1}) + \eta c_{k+1}\|^2 \\
&\quad + (k+\delta+1) \cdot \langle \eta F(z_{k+1}) + \eta c_{k+1}, z_{k+1} - z_0 \rangle .
\end{aligned}
$$

Since the identity in Proposition 15 holds for any $q \neq 0$, we only need to change every k to be $k+\delta$ in the proof of Theorem 9. It is possible that a choice of $\delta > 0$ leads to a better upper bound (better constant) than $\delta = 0$ which is chosen for (EAG) , but we do not optimize over δ here.

Lemma 11 *For $k \geq 2$,*

$$\left(\frac{k(k+1)}{4} - \frac{\eta^2 L^2}{1-\eta^2 L^2}\right) \cdot \|\eta F(z_k) + \eta c_k\|^2 \\ \leq V_1 + \|z_0 - z^*\|^2 + \frac{\eta^2 L^2}{1-\eta^2 L^2} \sum_{t=2}^{k-1} \|\eta F(z_t) + \eta c_t\|^2.$$

Moreover, when $\eta \in (0, \frac{1}{\sqrt{3}L})$, then

$$\frac{k^2}{4} \cdot \|\eta F(z_k) + \eta c_k\|^2 \leq V_1 + \|z_0 - z^*\|^2 + \frac{\eta^2 L^2}{1-\eta^2 L^2} \sum_{t=2}^{k-1} \|\eta F(z_t) + \eta c_t\|^2.$$

Proof Fix any $k \geq 2$. By definition, we have

$$\begin{aligned} V_k &= \frac{k(k+1)}{2}\|\eta F(z_k) + \eta c_k\|^2 + k\langle \eta F(z_k) + \eta c_k, z_k - z_0\rangle \\ &\geq \frac{k(k+1)}{2}\|\eta F(z_k) + \eta c_k\|^2 + k\langle \eta F(z_k) + \eta c_k, z^* - z_0\rangle \\ &\qquad (F + \partial \mathbb{I}_{\mathcal{Z}} \text{ is monotone and } \mathbf{0} \in F(z^*) + \partial \mathbb{I}_{\mathcal{Z}}(z^*)) \\ &\geq \frac{k(k+1)}{2}\|\eta F(z_k) + \eta c_k\|^2 - \frac{k(k+1)}{4}\|\eta F(z_k) + \eta c_k\|^2 - \frac{k}{k+1}\|z_0 - z^*\|^2 \\ &\qquad (\langle a, b\rangle \geq -\tfrac{c}{4}\|a\|^2 - \tfrac{1}{c}\|b\|^2) \\ &\geq \frac{k(k+1)}{4}\|\eta F(z_k) + \eta c_k\|^2 - \|z_0 - z^*\|^2. \end{aligned} \tag{14}$$

According to Theorem 9, $V_{t+1} - V_t \leq \frac{\eta^2 L^2}{1-\eta^2 L^2}\|\eta F(z_{t+1}) + \eta c_{t+1}\|^2$ for all $t \geq 1$. Through a telescoping sum, we obtain the following inequality:

$$V_k \leq V_1 + \frac{\eta^2 L^2}{1-\eta^2 L^2} \cdot \sum_{t=2}^{k} \|\eta F(z_t) + \eta c_t\|^2. \tag{15}$$

The first inequality in the statement follows from the combination of Inequality (14) and (15). The second inequality in the statement follows from the fact that $\frac{\eta^2 L^2}{1-\eta^2 L^2} \leq \frac{1}{2} \leq \frac{k}{4}$ when $\eta \in (0, \frac{1}{\sqrt{3}L})$. ■

C.3. Proof of Theorem 1

Theorem 12 *[More general version of Theorem 1.] Suppose Assumption 1 holds. Let $z_0 \in \mathcal{Z}$ be any starting point and $\{z_k, z_{k+\frac{1}{2}}\}_{k\geq 0}$ be the iterates of (EAG) with step size $\eta \in (0, \frac{1}{\sqrt{3}L})$. Denote $D := \|z_0 - z^*\|^2$. Then for any $T \geq 1$,*

$$r^{tan}_{F,\mathcal{Z}}(z_T)^2 \leq \frac{44}{\eta^2 L^2(1-3\eta^2 L^2)} \cdot \frac{D^2 L^2}{T^2}$$
$$\text{GAP}^{SVI}_{F,D}(z_T) \leq \frac{\sqrt{44}}{\eta L\sqrt{1-3\eta^2 L^2}} \cdot \frac{D^2 L}{T}.$$

Proof

The bound on $\text{GAP}^{SVI}_{F,D}(z_T)$ follows from the bound on $r^{tan}_{F,A}(z_T)$ since $\text{GAP}^{SVI}_{F,D}(z_T) \le D \cdot r^{tan}_{F,A}(z_T)$ [5, Lemma 2].

Denote $a_k := \frac{\|\eta F(z_k)+\eta c_k\|^2}{\|z_0-z^*\|^2}$. It suffices to prove that for all $k \ge 1$,

$$a_k \le \frac{44}{(1-3\eta^2L^2)k^2}. \tag{16}$$

Note that from the proof of Proposition 8, we have

$$\|\eta F(z_1) + \eta c_1\|^2 \le \frac{(1+\eta L+\eta^2L^2)^2}{1-\eta^2L^2}\|z_0 - z^*\|^2$$

which implies $a_1 \le \frac{(1+\eta L+\eta^2L^2)^2}{1-\eta^2L^2} \le 6$. Thus (16) holds for $k = 1$.

From (8), we also have

$$V_1 \le \frac{(1+\eta L+\eta^2L^2)(2+2\eta L+\eta^2L^2)}{1-\eta^2L^2}\|z_0 - z^*\|^2 \le 10 \cdot \|z_0 - z^*\|^2.$$

Thus by Lemma 11, we have

$$\frac{k^2}{4} \cdot \|\eta F(z_k) + \eta c_k\|^2 \le 11 \cdot \|z_0 - z^*\|^2 + \frac{\eta^2L^2}{1-\eta^2L^2}\sum_{t=2}^{k-1}\|\eta F(z_t) + \eta c_t\|^2$$

$$\Rightarrow \frac{k^2}{4} \cdot a_k \le 11 + \frac{\eta^2L^2}{1-\eta^2L^2}\sum_{t=2}^{k-1} a_t.$$

Applying Proposition 16 with $C_1 = 11$ and $p = \eta^2L^2 \in (0, \frac{1}{3})$, we have (16) holds for any $k \ge 2$. This completes the proof. ■

Appendix D. Missing Proofs in Section 4: Accelerated Algorithm for Inclusion Problems with Negatively Comonotone Operators

To analyze (AS), we adopt the following potential function: for $k \ge 1$,

$$U_k := \left(\frac{k^2}{2}\left(1+\frac{2\rho}{\eta}\right) - \frac{\rho}{\eta}k\right) \cdot \|\eta F(z_k) + \eta c_k\|^2 + k \cdot \langle \eta F(z_k) + \eta c_k, z_k - z_0\rangle.$$

The potential function is the same as the one used in the analysis for FEG [20] when c_k is always 0, and we properly adapted for non-zero c_k's. The convergence analysis builds on the following two properties of the potential function: Proposition 13 establishes an upper bound of U_1; Lemma 14 shows $U_{k+1} \le U_k$ for all $k \ge 1$.

Proposition 13 *$U_1 \le \frac{(1+\eta L)(3+\eta L)}{2} \cdot \|z_1 - z_0\|^2$ and $\|z_1 - z_0\|^2 \le \eta^2 \cdot r^{tan}_{F,A}(z_0)^2$.*

Proof Note that $z_{\frac{1}{2}} = z_0$ and $z_1 = J_{\eta A}[z_0 - \eta F(z_0)]$. Thus we have $\eta c_1 = z_0 - \eta F(z_0) - z_1$. We first bound $\|\eta F(z_1) + \eta c_1\|$ as follows:

$$
\begin{aligned}
\|\eta F(z_1) + \eta c_1\| &= \|z_0 - z_1 + \eta F(z_1) - \eta F(z_0)\| \\
&\le \|z_0 - z_1\| + \|\eta F(z_1) - \eta F(z_0)\| && \text{(Triangle inequality)} \\
&\le (1 + \eta L) \cdot \|z_0 - z_1\|. && \text{(F is L-Lipschitz)}
\end{aligned}
$$

Then we can bound U_1 as follows:

$$
\begin{aligned}
U_1 &= \frac{1}{2} \cdot \|\eta F(z_1) + \eta c_1\|^2 + \langle \eta F(z_1) + \eta c_1, z_1 - z_0 \rangle \\
&\le \frac{1}{2} \cdot \|\eta F(z_1) + \eta c_1\|^2 + \|\eta F(z_1) + \eta c_1\| \cdot \|z_1 - z_0\| && \text{(Cauchy-Schwarz Inequality)} \\
&\le \left(\frac{(1 + \eta L)^2}{2} + (1 + \eta L) \right) \cdot \|z_1 - z_0\|^2 \\
&\le \frac{(1 + \eta L)(3 + \eta L)}{2} \cdot \|z_1 - z_0\|^2.
\end{aligned}
$$

Moreover, for any $c \in A(z_0)$, we have

$$
\begin{aligned}
\|z_1 - z_0\| &= \|J_{\eta A}[z_0 - \eta F(z_0)] - J_{\eta A}[z_0 + \eta c]\| \\
&\le \eta \cdot \|F(z_0) + c\|. && \text{($J_{\eta A}$ is non-expansive)}
\end{aligned}
$$

Hence $\|z_1 - z_0\| \le \eta \cdot \min_{c \in A(z_0)} \|F(z_0) + c\| = \eta \cdot r^{tan}_{F,A}(z_0)$. ■

Lemma 14 *Assume Assumption 2 holds for some ρ. Let $z_0 \in \mathbb{R}^n$ be any initial point and $\{z_k, z_{k+\frac{1}{2}}\}_{k \ge 1}$ be the iterates of (AS) with step size $\eta > -2\rho$.*[7] *Then for all $k \ge 1$, we have $U_{k+1} \le U_k$.*

Proof Fix any $k \ge 1$. We first present several inequalities. Since F is L-Lipschitz, we have

$$
\left(-\frac{(k+1)^2}{2} \right) \cdot \left(\eta^2 L^2 \cdot \left\| z_{k+\frac{1}{2}} - z_{k+1} \right\|^2 - \left\| \eta F(z_{k+\frac{1}{2}}) - \eta F(z_{k+1}) \right\|^2 \right) \le 0. \tag{17}
$$

Additionally, since $F + A$ is ρ-comonotone, $c_k \in A(z_k)$, and $c_{k+1} \in A(z_{k+1})$, we have

$$
\begin{aligned}
(-k(k+1)) \cdot \Bigg(&\langle \eta F(z_{k+1}) + \eta c_{k+1} - \eta F(z_k) - \eta c_k, z_{k+1} - z_k \rangle \\
&- \frac{\rho}{\eta} \|\eta F(z_{k+1}) + \eta c_{k+1} - \eta F(z_k) - \eta c_k\|^2 \Bigg) \le 0.
\end{aligned} \tag{18}
$$

The following identity holds due to Identity (20) in Proposition 15: we treat x_0 as z_0; x_t as $z_{k+\frac{t-1}{2}}$ for $t \in \{1, 2, 3\}$; y_t as $\eta F(z_{k+\frac{t-1}{2}})$ for $t \in \{1, 2, 3\}$; u_1 as ηc_k, and u_3 as ηc_{k+1}; p as $\eta^2 L^2$, q

7. Lemma 14 holds for all step size η, but our potential function is no longer useful when $\eta \le -2\rho$.

as k, and r as $\frac{\rho}{\eta}$. Note that by the update rule of (AS) , we have $\eta c_k = \frac{z_k + \frac{1}{k+1}(z_0 - z_k) - \frac{k}{k+1}(1+2\frac{\rho}{\eta}) \cdot \eta F(z_k) - z_{k+\frac{1}{2}}}{\frac{k}{k+1}(1+2\frac{\rho}{\eta})}$, and by definition, we have $\eta c_{k+1} = z_k + \frac{1}{k+1}(z_0 - z_k) - \eta F(z_{k+\frac{1}{2}}) - \frac{2k\frac{\rho}{\eta}}{k+1} \cdot (\eta F(z_k) + \eta c_k) - z_{k+1}$.

$$
\begin{aligned}
&U_k - U_{k+1} + \text{LHS of Inequality (17)} + \text{LHS of Inequality (18)} \\
=&\frac{(1-\eta^2 L^2)(k+1)^2}{2} \cdot \left\|z_{k+1} - z_{k+\frac{1}{2}}\right\|^2.
\end{aligned}
$$

Hence, $U_{k+1} \le U_k$. ■

D.1. Proof of Theorem 2

Fix any $T \ge 1$. According to Lemma 14, we have $U_T \le U_1$. Then by definition of U_T, we have

$$
\begin{aligned}
U_T &= \left(\frac{T^2}{2}\left(1 + \frac{2\rho}{\eta}\right) - \frac{\rho}{\eta}T\right) \cdot \|\eta F(z_T) + \eta c_T\|^2 + T \cdot \langle \eta F(z_T) + \eta c_T, z_T - z^* \rangle \\
&\quad + T \cdot \langle \eta F(z_T) + \eta c_T, z^* - z_0 \rangle \\
&\ge \left(\frac{T^2}{2}\left(1 + \frac{2\rho}{\eta}\right) - \frac{\rho}{\eta}T\right) \cdot \|\eta F(z_T) + \eta c_T\|^2 + \frac{\rho}{\eta}T \cdot \|\eta F(z_T) + \eta c_T\|^2 \\
&\quad + T \cdot \langle \eta F(z_T) + \eta c_T, z^* - z_0 \rangle \\
&= \frac{\eta(\eta+2\rho)T^2}{2} \cdot \|F(z_T) + c_T\|^2 + T \cdot \langle \eta F(z_T) + \eta c_T, z^* - z_0 \rangle \\
&\ge \frac{\eta(\eta+2\rho)T^2}{2} \cdot \|F(z_T) + c_T\|^2 - \frac{\eta(\eta+2\rho)T^2}{4} \cdot \|F(z_T) + c_T\|^2 \\
&\quad - \frac{\eta}{\eta+2\rho}\|z_0 - z^*\|^2 \\
&= \frac{\eta(\eta+2\rho)T^2}{4} \cdot \|F(z_T) + c_T\|^2 - \frac{\eta}{\eta+2\rho}\|z_0 - z^*\|^2.
\end{aligned}
$$

In the first inequality, we use the fact that z^* is a solution of the (CMI) with the ρ-comonotone operator $F + A$. In the second inequality, we use $\langle a, b \rangle \ge -\frac{\delta}{4}\|a\|^2 - \frac{1}{\delta}\|b\|^2$ for $\delta > 0$.

By Proposition 13 and the assumption that $\eta L < 1$, $U_1 \le \frac{(1+\eta L)(3+\eta L)}{2} \cdot \|z_1 - z_0\|^2 \le 4\|z_1 - z_0\|^2 \le \frac{4}{L^2} \cdot r^{tan}(z_0)^2$. Therefore,

$$
\begin{aligned}
\|F(z_T) + c_T\|^2 &\le \frac{4}{\eta(\eta+2\rho)T^2}\left(U_1 + \frac{\eta}{\eta+2\rho}\|z_0 - z^*\|^2\right) \\
&= \frac{4}{(\eta+2\rho)^2 T^2}\left((1 + \frac{2\rho}{\eta})U_1 + \|z_0 - z^*\|^2\right) \\
&\le \frac{4}{(\eta+2\rho)^2 T^2}\left(4\|z_1 - z_0\|^2 + \|z_0 - z^*\|^2\right). \qquad (\rho \le 0)
\end{aligned}
$$

Appendix E. Auxiliary Propositions

Proposition 15 *Let $x_0, x_1, x_2, x_3, y_1, y_2, y_3, u_1, u_3$ be arbitrary vectors in $\mathbb{R}^n$. Let $q > 0$, $p > 0$, and $r \neq -\frac{1}{2}$ be real numbers.*

If $u_3 = x_1 - y_2 + \frac{1}{q+1}(x_0 - x_1) - x_3$*, then the following identity holds:*

$$\begin{aligned}
&\frac{q(q+1)}{2} \cdot \|y_1 + u_1\|^2 + q \cdot \langle y_1 + u_1, x_1 - x_0 \rangle \\
&- \left(\frac{(q+1)(q+2)}{2} \cdot \|y_3 + u_3\|^2 + (q+1) \cdot \langle y_3 + u_3, x_3 - x_0 \rangle \right) \\
&- \frac{q(q+1)}{2p} \cdot \left(p \cdot \|x_2 - x_3\|^2 - \|y_2 - y_3\|^2 \right) \\
&- q(q+1) \cdot \langle y_3 - y_1, x_3 - x_1 \rangle \\
&- q(q+1) \cdot \left\langle x_1 - y_1 - x_2 + \frac{1}{q+1}(x_0 - x_1), x_2 - x_3 \right\rangle \\
&- q(q+1) \cdot \langle u_3, x_3 - x_1 \rangle \\
&- q(q+1) \cdot \langle u_1, x_1 - x_2 \rangle \\
=&\frac{q(q+1)}{2} \cdot \left\| x_2 - x_1 + y_1 + u_1 + \frac{1}{q+1}(x_1 - x_0) \right\|^2 \\
&+ \frac{(1-p)q(q+1)}{2p} \cdot \|y_2 - y_3\|^2 \\
&+ (q+1) \cdot \langle y_2 - y_3, y_3 + u_3 \rangle
\end{aligned} \tag{19}$$

If $u_1 = \frac{x_1 + \frac{1}{q+1}(x_0 - x_1) - \frac{q}{q+1}(1+2r)y_1 - x_2}{\frac{q}{q+1}(1+2r)}$ *and* $u_3 = x_1 + \frac{1}{q+1}(x_0 - x_1) - y_2 - \frac{2rq}{q+1}(y_1 + u_1) - x_3$*, then the following identity holds:*

$$\begin{aligned}
&\left(\frac{q^2}{2}(1+2r) - rq \right) \cdot \|y_1 + u_1\|^2 + q \cdot \langle y_1 + u_1, x_1 - x_0 \rangle \\
&- \left(\frac{(q+1)^2}{2}(1+2r) - r(q+1) \right) \cdot \|y_3 + u_3\|^2 - (q+1) \cdot \langle y_3 + u_3, x_3 - x_0 \rangle \\
&- \frac{(q+1)^2}{2} \cdot \left(p \cdot \|x_2 - x_3\|^2 - \|y_2 - y_3\|^2 \right) \\
&- q(q+1) \cdot \left(\langle y_3 + u_3 - y_1 - u_1, x_3 - x_1 \rangle - r\|y_3 + u_3 - y_1 - u_1\|^2 \right) \\
=&\frac{(1-p)(q+1)^2}{2} \cdot \|x_2 - x_3\|^2.
\end{aligned} \tag{20}$$

Proof We verify the two identities using MATLAB. Readers can find the verification code here or go to the next url:

`https://github.com/weiqiangzheng1999/Accelerated-Non-Monotone-Inclusion`.

■

Proposition 16 *Let* $\{a_k \in \mathbb{R}^+\}_{k \geq 2}$ *be a sequence of real numbers. Let* $C_1 \geq 0$ *and* $p \in (0, \frac{1}{3})$ *be two real numbers. If the following condition holds for every* $k \geq 2$*,*

$$\frac{k^2}{4} \cdot a_k \leq C_1 + \frac{p}{1-p} \cdot \sum_{t=2}^{k-1} a_t, \tag{21}$$

then for each $k \geq 2$ we have

$$a_k \leq \frac{4 \cdot C_1}{1 - 3p} \cdot \frac{1}{k^2}. \tag{22}$$

Proof We prove the statement by induction.
Base Case: $k = 2$**.** From Inequality (21), we have

$$\frac{2^2}{4} \cdot a_2 \leq C_1 \quad \Rightarrow \quad a_2 \leq C_1 \leq \frac{4 \cdot C_1}{1 - 3p} \cdot \frac{1}{2^2}.$$

Thus, Inequality (22) holds for $k = 2$.
Inductive Step: for any $k \geq 3$**.** Fix some $k \geq 3$ and assume that Inequality (22) holds for all $2 \leq t \leq k - 1$. We slightly abuse notation and treat the summation in the form $\sum_{t=3}^{2}$ as 0. By Inequality (21), we have

$$\begin{aligned}
\frac{k^2}{4} \cdot a_k &\leq C_1 + \frac{p}{1-p} \cdot \sum_{t=2}^{k-1} a_t \\
&\leq \frac{C_1}{1-p} + \frac{p}{1-p} \cdot \sum_{t=3}^{k-1} a_t && (a_2 \leq C_1) \\
&\leq \frac{C_1}{1-p} + \frac{4p \cdot C_1}{(1-p)(1-3p)} \cdot \sum_{t=3}^{k-1} \frac{1}{t^2} && \text{(Induction assumption on Inequality (22))} \\
&\leq \frac{C_1}{1-p} + \frac{2p \cdot C_1}{(1-p)(1-3p)} && \left(\textstyle\sum_{t=3}^{\infty} \frac{1}{t^2} = \frac{\pi^2}{6} - \frac{5}{4} \leq \frac{1}{2}\right) \\
&= \frac{C_1}{1-3p}.
\end{aligned}$$

This complete the inductive step. Therefore, for all $k \geq 2$, we have $a_k \leq \frac{4 \cdot C_1}{1-3p} \cdot \frac{1}{k^2}$. ■