Halpern Iteration for Near-Optimal and Parameter-Free Monotone Inclusion and Strong Solutions to Variational Inequalities

Jelena Diakonikolas JELENA@CS.WISC.EDU
Department of Computer Sciences, University of Wisconsin-Madison
1210 W. Dayton St, Madison, WI 53706

Editors: Jacob Abernethy and Shivani Agarwal

Abstract

We leverage the connections between nonexpansive maps, monotone Lipschitz operators, and proximal mappings to obtain near-optimal (i.e., optimal up to poly-log factors in terms of iteration complexity) and parameter-free methods for solving monotone inclusion problems. These results immediately translate into near-optimal guarantees for approximating strong solutions to variational inequality problems, approximating convex-concave min-max optimization problems, and minimizing the norm of the gradient in min-max optimization problems. Our analysis is based on a novel and simple potential-based proof of convergence of Halpern iteration, a classical iteration for finding fixed points of nonexpansive maps. Additionally, we provide a series of algorithmic reductions that highlight connections between different problem classes and lead to lower bounds that certify near-optimality of the studied methods.

Keywords: Halpern iteration, monotone inclusion, min-max optimization, variational inequalities.

1. Introduction

Given a closed convex set $\mathcal{U} \subseteq \mathbb{R}^d$ and a single-valued monotone operator $F : \mathbb{R}^d \to \mathbb{R}^d$, i.e., an operator that maps each vector to another vector and satisfies:

$$(\forall \mathbf{u}, \mathbf{v} \in \mathbb{R}^d): \quad \langle F(\mathbf{u}) - F(\mathbf{v}), \mathbf{u} - \mathbf{v} \rangle \geq 0, \tag{1}$$

the *monotone inclusion* problem consists in finding a point $\mathbf{u}^*$ that satisfies:

$$\mathbf{0} \in F(\mathbf{u}) + \partial I_{\mathcal{U}}(\mathbf{u}), \quad \text{where}$$
$$I_{\mathcal{U}}(\mathbf{u}) = \begin{cases} 0, & \text{if } \mathbf{u} \in \mathcal{U}, \\ \infty, & \text{otherwise} \end{cases} \tag{MI}$$

is the indicator function of the set $\mathcal{U} \subseteq \mathbb{R}^d$ and $\partial I_{\mathcal{U}}(\cdot)$ denotes the subdifferential operator (the set of all subgradients at the argument point) of $I_{\mathcal{U}}$.

Monotone inclusion is a fundamental problem in continuous optimization that is closely related to variational inequalities (VIs) with monotone operators, which model a plethora of problems in mathematical programming, game theory, engineering, and finance (Facchinei and Pang, 2003, Section 1.4). Within machine learning, VIs with monotone operators and associated monotone inclusion problems arise, for example, as an abstraction of convex-concave min-max optimization problems, which naturally model adversarial training (Madry et al., 2018; Arjovsky et al., 2017; Arjovsky and Bottou, 2017; Goodfellow et al., 2014).

When it comes to convex-concave min-max optimization, approximating the associated VI leads to guarantees in terms of the optimality gap. Such guarantees are generally possible only when the feasible set $\mathcal{U}$ is bounded; a simple example that demonstrates this fact is $\Phi(\mathbf{x}, \mathbf{y}) = \langle \mathbf{x}, \mathbf{y} \rangle$ with the feasible set $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$. The only (min-max or saddle-point) solution in this case is obtained when both $\mathbf{x}$ and $\mathbf{y}$ are the all-zeros vectors. However, if either $\mathbf{x} \neq \mathbf{0}$ or $\mathbf{y} \neq \mathbf{0}$, then the optimality gap $\max_{\mathbf{y}' \in \mathbb{R}^d} \Phi(\mathbf{x}, \mathbf{y}') - \min_{\mathbf{x}' \in \mathbb{R}^d} \Phi(\mathbf{x}', \mathbf{y})$ is infinite.

On the other hand, approximate monotone inclusion is well-defined even for unbounded feasible sets. In the context of min-max optimization, it corresponds to guarantees in terms of stationarity. Specifically, in the unconstrained setting, solving monotone inclusion corresponds to minimizing the norm of the gradient of Φ. Note that even in the special setting of convex optimization, convergence in norm of the gradient is much less understood than convergence in optimality gap (Nesterov, 2012; Kim and Fessler, 2018). Further, unlike standard results for VIs that provide convergence guarantees for approximating *weak* solutions (Nemirovski, 2004; Nesterov, 2007; Bach and Levy, 2019), approximations to monotone inclusion lead to approximations to *strong* solutions (see Section 1.2 for definitions of weak and strong solutions and their relationship to monotone inclusion).

We leverage the connections between nonexpansive maps, structured monotone operators, and proximal maps to obtain near-optimal algorithms for solving monotone inclusion over different classes of problems with Lipschitz-continuous operators. In particular, we make use of the classical Halpern iteration, which is defined by (Halpern, 1967):

$$\mathbf{u}_{k+1} = \lambda_{k+1}\mathbf{u}_0 + (1 - \lambda_{k+1})T(\mathbf{u}_k), \tag{Hal}$$

where $T : \mathbb{R}^d \to \mathbb{R}^d$ is a nonexpansive map, i.e., $\forall \mathbf{u}, \mathbf{v} \in \mathbb{R}^d : \|T(\mathbf{u}) - T(\mathbf{v})\| \leq \|\mathbf{u} - \mathbf{v}\|$.

In addition to its simplicity, Halpern iteration is particularly relevant to machine learning applications, as it is an *implicitly regularized* method with the following property: if the set of fixed points of T is non-empty, then Halpern iteration (Hal) started at a point $\mathbf{u}_0$ and applied with any choice of step sizes $\{\lambda_k\}_{k \geq 1}$ that satisfy all of the following conditions:

$$(i) \lim_{k \to \infty} \lambda_k = 0, \quad (ii) \sum_{k=1}^{\infty} \lambda_k = \infty, \quad (iii) \sum_{k=1}^{\infty} |\lambda_{k+1} - \lambda_k| < \infty \tag{2}$$

converges to the fixed point of T with the *minimum* ℓ_2 *distance to* $\mathbf{u}_0$. This result was proved by Wittmann (1992), who extended a similar though less general result previously obtained by Browder (1967). The result of Wittmann (1992) has since been extended to various other settings (Bauschke, 1996; Xu, 2002; Kohlenbach, 2011; Körnlein, 2015; Lieder, 2017, and references therein).

1.1. Contributions and Related Work

A special case of what is now known as the Halpern iteration (Hal) was introduced and its asymptotic convergence properties were analyzed by Halpern (1967) in the setting of $\mathbf{u}_0 = \mathbf{0}$ and $T : \mathcal{B}_2 \to \mathcal{B}_2$, where $\mathcal{B}_2$ is the unit Euclidean ball. Using the proof-theoretic techniques of Kohlenbach (2008), Leustean (2007) extracted from the asymptotic convergence result of Wittmann (1992) the rate at which Halpern iteration converges to a fixed point. The results obtained by Leustean (2007) are rather loose and provide guarantees of the form $\|T(\mathbf{u}_k) - \mathbf{u}_k\| = O(\frac{M}{\log(k)})$ in the best case (obtained for $\lambda_k = \Theta(\frac{1}{k})$), where $M \geq \|\mathbf{u}_0\| + \|T(\mathbf{u}_0)\| + \|\mathbf{u}_k\|, \forall k$. A tighter result that shows that $\|T(\mathbf{u}_k) - \mathbf{u}_k\|$ decreases at rate that is at least as good as $1/\sqrt{k}$ was obtained by Kohlenbach

(2011). The results of Leustean (2007) and Kohlenbach (2011) apply to general normed spaces. The work of Kohlenbach (2011) also provided an explicit rate of metastability that characterizes the convergence of the sequence of iterates $\{\mathbf{u}_k\}$ in Hilbert spaces.

More recently, Lieder (2017) proved that under the standard assumption that T has a fixed point $\mathbf{u}^*$ and for the step size $\lambda_k = \frac{1}{k+1}$, Halpern iteration converges to a fixed point as $\|T(\mathbf{u}_k) - \mathbf{u}_k\| = \frac{2\|\mathbf{u}_0 - \mathbf{u}^*\|}{k+1}$. A similar result but for an alternative algorithm was recently obtained by Kim (2019). These two results (as well as all the results from this paper) only apply to Euclidean spaces. Unlike Halpern iteration, the algorithm introduced by Kim (2019) is not known to possess the implicit regularization property discussed earlier in this paper. The results of Lieder (2017) and Kim (2019) can be used to obtain the same $1/k$ convergence rate for monotone inclusion with a cocoercive operator *but only if the cocoercivity parameter is known*, which is rarely the case in practice. Similarly, those results can also be extended to more general monotone Lipschitz operators *but only if the proximal map (or resolvent) of F can be computed exactly*, an assumption that can rarely be met (see Section 1.2 for definitions of cocoercive operators and proximal maps). We also note that the results of Lieder (2017) and Kim (2019) were obtained using the performance estimation (PEP) framework of Drori and Teboulle (2014). The convergence proofs resulting from the use of PEP are computer-assisted: they are generated as solutions to large semidefinite programs, which typically makes them hard to interpret and generalize.

Our approach is arguably simpler, as it relies on the use of a potential function, which allows us to remove the assumptions about the knowledge of the problem parameters and availability of exact proximal maps. Our main contributions are summarized as follows:

Results for cocoercive operators. We introduce a new, potential-based, proof of convergence of Halpern iteration that applies to more general step sizes λ_k than handled by the analysis of Lieder (2017) (Section 2). The proof is simple and only requires elementary algebra. Further, the proof is derived for cocoercive operators and leads to a *parameter-free* algorithm for monotone inclusion. We also extend this parameter-free method to the constrained setting using the concept of gradient mapping generalized to monotone operators (Section 2.1). To the best of our knowledge, this is the first work to obtain the $1/k$ convergence rate with a parameter-free method.

Results for monotone Lipschitz operators. Up to a logarithmic factor, we obtain the same $1/k$ convergence rate for the parameter-free setting of the more general monotone Lipschitz operators (Section 2.2). The best known convergence rate established by previous work for the same setting was of the order $1/\sqrt{k}$ (Dang and Lan, 2015; Ryu et al., 2019), and it was for the "best" iterate (all of our results are for the *last* iterate). We obtain the improved convergence rate through the use of the Halpern iteration with *inexact* proximal maps that can be implemented efficiently. The idea of coupling inexact proximal maps with another method is similar in spirit to the Catalyst framework (Lin et al., 2017) and other instantiations of the inexact proximal-point method, such as, e.g., in the work of Davis and Drusvyatskiy (2019); Asi and Duchi (2019); Lin et al. (2018); Thekumparampil et al. (2019). However, we note that, unlike in the previous work, the coupling used here is with a method (Halpern iteration) whose convergence properties were not well-understood and for which no simple potential-based convergence proof existed prior to our work.

Results for strongly monotone Lipschitz operators. We show that a simple restarting-based approach applied to our method for operators that are only monotone and Lipschitz (described above) leads to a *parameter-free* method for strongly monotone and Lipschitz operators (Section 2.3). Un-

der mild assumptions about the problem parameters and up to a poly-logarithmic factor, the resulting algorithm is *iteration-complexity-optimal*. To the best of our knowledge, this is *the first near-optimal parameter-free method* for the setting of strongly monotone Lipschitz operators and any of the associated problems – monotone inclusion, VIs, or convex-concave min-max optimization.

Lower bounds. To certify near-optimality of the analyzed methods, we provide lower bounds that rely on algorithmic reductions between different problem classes and highlight connections between them (Section 3). The lower bounds are derived by leveraging the recent lower bound of Ouyang and Xu (2019) for approximating the optimality gap in convex-concave min-max optimization.

1.2. Notation and Preliminaries

While it is possible to extend all of our results to (infinite-dimensional) Hilbert spaces in a relatively straightforward manner, we will focus here on real d-dimensional normed vector spaces $(E, \|\cdot\|)$, where $\|\cdot\|$ is induced by an inner product $\langle\cdot,\cdot\rangle$ associated with the space, i.e., $\|\cdot\| = \sqrt{\langle\cdot,\cdot\rangle}$. The most important example is $(\mathbb{R}^d, \|\cdot\|_2)$, which can be taken as a canonical example to have in mind when reading this paper. Our focus on finite-dimensional spaces is due to the applications that motivate this work and were mentioned at the beginning of the paper.

Variational Inequalities and Monotone Operators. Let $\mathcal{U} \subseteq E$ be closed and convex, and let $F : E \to E$ be an L-Lipschitz-continuous operator defined on $\mathcal{U}$. Namely, we assume that:

$$(\forall \mathbf{u}, \mathbf{v} \in \mathcal{U}): \quad \|F(\mathbf{u}) - F(\mathbf{v})\| \leq L\|\mathbf{u} - \mathbf{v}\|. \tag{3}$$

The definition of monotonicity was already provided in Eq. (1), and easily specializes to monotonicity on the set $\mathcal{U}$ by restricting $\mathbf{u}, \mathbf{v}$ to be from $\mathcal{U}$. Further, F is said to be:

1. *strongly monotone* (or *coercive*) on $\mathcal{U}$ with parameter m, if:

$$(\forall \mathbf{u}, \mathbf{v} \in \mathcal{U}): \quad \langle F(\mathbf{u}) - F(\mathbf{v}), \mathbf{u} - \mathbf{v}\rangle \geq m\|\mathbf{u} - \mathbf{v}\|^2; \tag{4}$$

2. *cocoercive* on $\mathcal{U}$ with parameter γ, if:

$$(\forall \mathbf{u}, \mathbf{v} \in \mathcal{U}): \quad \langle F(\mathbf{u}) - F(\mathbf{v}), \mathbf{u} - \mathbf{v}\rangle \geq \gamma\|F(\mathbf{u}) - F(\mathbf{v})\|^2. \tag{5}$$

It is immediate from the definition of cocoercivity that every γ-cocoercive operator is monotone and $1/\gamma$-Lipschitz. The latter follows by applying the Cauchy-Schwarz inequality to the left-hand side of Eq. (5) and then dividing both sides by $\gamma\|F(\mathbf{u}) - F(\mathbf{v})\|$.

Examples of monotone operators include the gradient of a convex function and appropriately modified gradient of a convex-concave function. Namely, if a function $\Phi(\mathbf{x}, \mathbf{y})$ is convex in $\mathbf{x}$ and concave in $\mathbf{y}$, then $F(\left[\begin{smallmatrix}\mathbf{x}\\ \mathbf{y}\end{smallmatrix}\right]) = \left[\begin{smallmatrix}\nabla_{\mathbf{x}}\Phi(\mathbf{x},\mathbf{y})\\ -\nabla_{\mathbf{y}}\Phi(\mathbf{x},\mathbf{y})\end{smallmatrix}\right]$ is monotone.

The Stampacchia Variational Inequality (SVI) problem consists in finding $\mathbf{u}^* \in \mathcal{U}$ such that:

$$(\forall \mathbf{u} \in \mathcal{U}): \quad \langle F(\mathbf{u}^*), \mathbf{u} - \mathbf{u}^*\rangle \geq 0. \tag{SVI}$$

In this case, $\mathbf{u}^*$ is also referred to as a *strong* solution to the variational inequality (VI) corresponding to F and $\mathcal{U}$. The Minty Variational Inequality (MVI) problem consists in finding $\mathbf{u}^*$ such that:

$$(\forall \mathbf{u} \in \mathcal{U}): \quad \langle F(\mathbf{u}), \mathbf{u}^* - \mathbf{u}\rangle \leq 0, \tag{MVI}$$

in which case $\mathbf{u}^*$ is referred to as a *weak* solution to the variational inequality corresponding to F and $\mathcal{U}$. In general, if F is continuous, then the solutions to (MVI) are a subset of the solutions to (SVI). If we assume that F is monotone, then Eq. (1) implies that every solution to (SVI) is also a solution to (MVI), and thus the two solution sets are equivalent. The solution set to monotone inclusion is the same as the solution set to (SVI).

Approximate versions of variational inequality problems (SVI) and (MVI) are defined as follows: Given $\epsilon > 0$, find an ϵ-approximate solution $\mathbf{u}_\epsilon^* \in \mathcal{U}$, which is a solution that satisfies:

$$(\forall \mathbf{u} \in \mathcal{U}): \quad \langle F(\mathbf{u}_\epsilon^*), \mathbf{u}_\epsilon^* - \mathbf{u} \rangle \leq \epsilon, \quad \text{or}$$
$$(\forall \mathbf{u} \in \mathcal{U}): \quad \langle F(\mathbf{u}), \mathbf{u}_\epsilon^* - \mathbf{u} \rangle \leq \epsilon, \quad \text{respectively.}$$

Clearly, when F is monotone, an ϵ-approximate solution to (SVI) is also an ϵ-approximate solution to (MVI); the reverse does *not* hold in general.

Similarly, ϵ-approximate monotone inclusion can be defined as fidning $\mathbf{u}_\epsilon^*$ that satisfies:

$$\mathbf{0} \in F(\mathbf{u}_\epsilon^*) + \partial I_{\mathcal{U}}(\mathbf{u}_\epsilon^*) + \mathcal{B}(\epsilon), \tag{6}$$

where $\mathcal{B}(\epsilon)$ is the ball w.r.t. $\|\cdot\|$, centered at $\mathbf{0}$ and of radius ϵ. We will sometimes write Eq. (6) in the equivalent form $-F(\mathbf{u}_\epsilon^*) \in \partial I_{\mathcal{U}}(\mathbf{u}_\epsilon^*) + \mathcal{B}(\epsilon)$. The following fact is immediate from Eq. (6).

Fact 1 *Given F and $\mathcal{U}$, let $\mathbf{u}_\epsilon^*$ satisfy Eq. (6). Then:*

$$(\forall \mathbf{u} \in \{\mathcal{U} \cap \mathcal{B}_{\mathbf{u}_\epsilon^*}\}): \quad \langle F(\mathbf{u}_\epsilon^*), \mathbf{u}_\epsilon^* - \mathbf{u} \rangle \leq \epsilon,$$

where $\mathcal{B}_{\mathbf{u}_\epsilon^*}$ denotes the unit ball w.r.t. $\|\cdot\|$, centered at $\mathcal{B}_{\mathbf{u}_\epsilon^*}$.

Further, if the diameter of $\mathcal{U}$, $D = \sup_{\mathbf{u},\mathbf{v} \in \mathcal{U}} \|\mathbf{u} - \mathbf{v}\|$, is bounded, then:

$$(\forall \mathbf{u} \in \mathcal{U}): \quad \langle F(\mathbf{u}_\epsilon^*), \mathbf{u}_\epsilon^* - \mathbf{u} \rangle \leq \epsilon D.$$

Thus, when the diameter D is bounded, any $\frac{\epsilon}{D}$-approximate solution to monotone inclusion is an ϵ-approximate solution to (SVI) (and thus also to (MVI)); the converse does *not* hold in general. Recall that when D is unbounded, neither (SVI) nor (MVI) can be approximated.

We assume throughout the paper that a solution to monotone inclusion (MI) exists. This assumption implies that solutions to both (SVI) and (MVI) exist as well. Existence of solutions follows from standard results and is guaranteed whenever e.g., $\mathcal{U}$ is compact, or, if there exists a compact set $\mathcal{U}'$ such that $\mathrm{Id} - \frac{1}{L}F$ maps $\mathcal{U}'$ to itself (Facchinei and Pang, 2003).

Nonexpansive Maps. Let $T : E \to E$. We say that T is *nonexpansive* on $\mathcal{U} \subseteq E$, if $\forall \mathbf{u}, \mathbf{v} \in \mathcal{U}$:

$$\|T(\mathbf{u}) - T(\mathbf{v})\| \leq \|\mathbf{u} - \mathbf{v}\|.$$

Nonexpansive maps are closely related to cocoercive operators, and here we summarize some of the basic properties that are used in our analysis. More information can be found in, e.g., the book by Bauschke and Combettes (2011).

Fact 2 *T is nonexpansive if and only if $\mathrm{Id} - T$ is $\frac{1}{2}$-cocoercive, where Id is the identity map.*

T is said to be *firmly nonexpansive* or *averaged*, if $\forall \mathbf{u}, \mathbf{v} \in \mathcal{U}$:

$$\|T(\mathbf{u}) - T(\mathbf{v})\|^2 + \|(\mathrm{Id} - T)\mathbf{u} - (\mathrm{Id} - T)\mathbf{v}\|^2 \leq \|\mathbf{u} - \mathbf{v}\|^2.$$

Useful properties of firmly nonexpansive maps are summarized in the following fact.

Fact 3 *For any firmly nonexpansive operator T, $\mathrm{Id} - T$ is also firmly non-expansive, and, moreover, both T and $\mathrm{Id} - T$ are 1-cocoercive.*

2. Halpern Iteration for Monotone Inclusion and Variational Inequalities

Halpern iteration is typically stated for nonexpansive maps T as in (Hal). Because our interest is in cocoercive operators F with the unknown parameter $1/L$, we instead work with the following version of the Halpern iteration:

$$\mathbf{u}_{k+1} = \lambda_{k+1}\mathbf{u}_0 + (1-\lambda_{k+1})\Big(\mathbf{u}_k - \frac{2}{L_{k+1}}F(\mathbf{u}_k)\Big), \tag{H}$$

where $L_k \in (0,\infty)$, $\forall k$. If L were known, we could simply set $L_{k+1} = L$, in which case (H) would be equivalent to the standard Halpern iteration, due to Fact 2. We assume throughout that $\lambda_1 = \frac{1}{2}$.

We start with the assumption that the setting is unconstrained: $\mathcal{U} \equiv E$. We will see in Section 2.1 how the result can be extended to the constrained case. Section 2.2 will consider the case of operators that are monotone and Lipschitz, while Section 2.3 will deal with the strongly monotone and Lipschitz case. Some of the proofs are omitted and are instead provided in Appendix A.

To analyze the convergence of (H) for the appropriate choices of sequences $\{\lambda_i\}_{i\geq 1}$ and $\{L_i\}_{i\geq 1}$, we make use of the following potential function:

$$\mathcal{C}_k = \frac{1}{L_k}\|F(\mathbf{u}_k)\|^2 - \frac{\lambda_k}{1-\lambda_k}\langle F(\mathbf{u}_k), \mathbf{u}_0 - \mathbf{u}_k\rangle\,. \tag{7}$$

Let us first show that if $A_k\mathcal{C}_k$ is non-increasing with k for an appropriately chosen sequence of positive numbers $\{A_k\}_{k\geq 1}$, then we can deduce a property that, under suitable conditions on $\{\lambda_i\}_{i\geq 1}$ and $\{L_i\}_{i\geq 1}$, implies a convergence rate for (H).

Lemma 4 *Let $\mathcal{C}_k$ be defined as in Eq. (7) and let $\mathbf{u}^*$ be the solution to (MI) that minimizes $\|\mathbf{u}_0 - \mathbf{u}^*\|$. Assume further that $\langle F(\mathbf{u}_1) - F(\mathbf{u}_0), \mathbf{u}_1 - \mathbf{u}_0\rangle \geq \frac{1}{L_1}\|F(\mathbf{u}_1) - F(\mathbf{u}_0)\|^2$. If $A_{k+1}\mathcal{C}_{k+1} \leq A_k\mathcal{C}_k$, $\forall k \geq 1$, where $\{A_i\}_{i\geq 1}$ is a sequence of positive numbers that satisfies $A_1 = 1$, then:*

$$(\forall k \geq 1): \quad \|F(\mathbf{u}_k)\| \leq L_k\frac{\lambda_k}{1-\lambda_k}\|\mathbf{u}_0 - \mathbf{u}^*\|.$$

Using Lemma 4, our goal is now to show that we can choose $L_k = O(L)$ and $\lambda_k = O(\frac{1}{k})$, which in turn would imply the desired $1/k$ convergence rate: $\|F(\mathbf{u}_k)\| = O(\frac{L\|\mathbf{u}_0-\mathbf{u}^*\|}{k})$. The following lemma provides sufficient conditions for $\{A_i\}_{i\geq 1}$, $\{\lambda_i\}_{i\geq 1}$, and $\{L_i\}_{i\geq 1}$ to ensure that $A_{k+1}\mathcal{C}_{k+1} \leq A_k\mathcal{C}_k$, $\forall k \geq 1$, so that Lemma 4 applies.

Lemma 5 *Let $\mathcal{C}_k$ be defined as in Eq. (7). Let $\{A_i\}_{i\geq 1}$ be defined recursively as $A_1 = 1$ and $A_{k+1} = A_k\frac{\lambda_k}{(1-\lambda_k)\lambda_{k+1}}$ for $k \geq 1$. Assume that $\{\lambda_i\}_{i\geq 1}$ is chosen so that $\lambda_1 = \frac{1}{2}$ and for $k \geq 1$: $\frac{\lambda_{k+1}}{1-2\lambda_{k+1}} \geq \frac{\lambda_k L_k}{(1-\lambda_k)L_{k+1}}$. Finally, assume that $L_k \in (0,\infty)$ and $\langle F(\mathbf{u}_k) - F(\mathbf{u}_{k-1}), \mathbf{u}_k - \mathbf{u}_{k-1}\rangle \geq \frac{1}{L_k}\|F(\mathbf{u}_k) - F(\mathbf{u}_{k-1})\|^2$, $\forall k$. Then,*

$$(\forall k \geq 1): \quad A_{k+1}\mathcal{C}_{k+1} \leq A_k\mathcal{C}_k.$$

Observe first the following. If we knew L and set $L_k = L$, $\lambda_k = \frac{1}{k+1}$, and $A_k = k(k+1)/2$, then all of the conditions from Lemma 5 would be satisfied, and Lemma 4 would then imply $\|F(\mathbf{u}_k)\| \leq \frac{L\|\mathbf{u}_0-\mathbf{u}^*\|}{k}$, which recovers the result of Lieder (2017). The choice $\lambda_k = \frac{1}{k+1}$ is also the tightest possible that satisfies the conditions from Lemma 5 – the inequality relating λ_{k+1} and λ_k is

satisfied with equality. This result is in line with the numerical observations made by Lieder (2017), who observed that the convergence of Halpern iteration is fastest for $\lambda_k = \frac{1}{k+1}$.

To construct a parameter-free method, we use that F is L-cocoercive; namely, that there exists a constant $L < \infty$ such that F satisfies Eq. (5) with $\gamma = 1/L$. The idea is to start to with a "guess" of L (e.g., $L_0 = 1$) and double the guess L_k as long as $\langle F(\mathbf{u}_k) - F(\mathbf{u}_{k-1}), \mathbf{u}_k - \mathbf{u}_{k-1}\rangle < \frac{1}{L_k}\|F(\mathbf{u}_k) - F(\mathbf{u}_{k-1})\|^2$. The total number of times that the guess can be doubled is bounded above by $\max\{0, \log_2(2L/L_0)\}$. Parameter λ_k is simply chosen to satisfy the condition from Lemma 5. The algorithm pseudocode is stated in Algorithm 1 for a given accuracy specified at the input.

Algorithm 1: Parameter-Free Halpern – Cocoercive Case

Input: $L_0 > 0$, $\epsilon > 0$, $\mathbf{u}_0$. If not provided at the input, set $L_0 = 1$.

$\lambda_1 = \frac{1}{2}$, $k = 0$

while $\|F(\mathbf{u}_k)\| > \epsilon$ **do**

  $k = k + 1$

  $L_k = L_{k-1}$

  $p_k = \frac{L_{k-1}}{L_k}\frac{\lambda_{k-1}}{1-\lambda_{k-1}}$, $\lambda_k = \frac{p_k}{1+2p_k}$

  $\mathbf{u}_k = \lambda_k \mathbf{u}_0 + (1-\lambda_k)(\mathbf{u}_{k-1} - 2F(\mathbf{u}_{k-1})/L_k)$

  while $\langle F(\mathbf{u}_k) - F(\mathbf{u}_{k-1}), \mathbf{u}_k - \mathbf{u}_{k-1}\rangle < \frac{1}{L_k}\|F(\mathbf{u}_k) - F(\mathbf{u}_{k-1})\|^2$ **do**

    $L_k = 2 \cdot L_k$

    $p_k = \frac{L_{k-1}}{L_k}\frac{\lambda_{k-1}}{1-\lambda_{k-1}}$, $\lambda_k = \frac{p_k}{1+2p_k}$

    $\mathbf{u}_k = \lambda_k \mathbf{u}_0 + (1-\lambda_k)(\mathbf{u}_{k-1} - 2F(\mathbf{u}_{k-1})/L_k)$

  end

end

return $\mathbf{u}_k$

We now prove the first of our main results. Note that the total number of arithmetic operations in Algorithm 1 is of the order of the number of oracle queries to F multiplied by the complexity of evaluating F at a point. The same will be true for all the algorithms stated in this paper, except that the complexity of evaluating F may be replaced by the complexity of projections onto $\mathcal{U}$.

Theorem 6 *Given $\mathbf{u}_0 \in \mathcal{U}$ and an operator F that is $\frac{1}{L}$-cocoercive on E, Algorithm 1 returns a point $\mathbf{u}_k$ such that $\|F(\mathbf{u}_k)\| \le \epsilon$ after at most $\frac{\max\{2L, L_0\}\|\mathbf{u}_0 - \mathbf{u}^*\|}{\epsilon} + \max\{0, \log_2(2L/L_0)\}$ oracle queries to F.*

Proof As F is $\frac{1}{L}$-cocoercive, $L_k \le \max\{2L, L_0\}$ and the total number of times that the algorithm enters the inner while loop is at most $\max\{0, \log_2(2L/L_0)\}$. The parameters satisfy the assumptions of Lemmas 4 and 5, and, thus, $\|F(\mathbf{u}_k)\| \le L_k \frac{\lambda_k}{1-\lambda_k}\|\mathbf{u}_0 - \mathbf{u}^*\|$. Hence, we only need to show that λ_k decreases sufficiently fast with k. As L_k can only be increased in any iteration, we have that

$$\lambda_{k+1} \le \frac{\frac{\lambda_k}{1-\lambda_k}}{1 + 2\frac{\lambda_k}{1-\lambda_k}} = \frac{\lambda_k}{1+\lambda_k} \le \frac{\lambda_{k-1}}{1+2\lambda_{k-1}} \le \cdots \le \frac{\lambda_1}{1+k\lambda_1} = \frac{1}{k+2}.$$

Hence, the total number of outer iterations is at most $\frac{\max\{2L, L_0\}\|\mathbf{u}_0 - \mathbf{u}^*\|}{\epsilon}$. Combining with the maximum total number of inner iterations from the beginning of the proof, the result follows. ■

2.1. Constrained Setups with Cocoercive Operators

Assume now that $\mathcal{U} \subseteq E$. We will make use of a counterpart to *gradient mapping* (Nesterov, 2018, Chapter 2) that we refer to as the *operator mapping*, defined as:

$$G_\eta(\mathbf{u}) = \eta\Big(\mathbf{u} - \Pi_{\mathcal{U}}\Big(\mathbf{u} - \frac{1}{\eta}F(\mathbf{u})\Big)\Big), \tag{8}$$

where $\Pi_{\mathcal{U}}\big(\mathbf{u} - \frac{1}{\eta}F(\mathbf{u})\big)$ is the projection operator, namely:

$$\Pi_{\mathcal{U}}\Big(\mathbf{u} - \frac{1}{\eta}F(\mathbf{u})\Big) = \operatorname*{argmin}_{\mathbf{v}\in\mathcal{U}}\Big\{\frac{1}{2}\|\mathbf{v} - \mathbf{u} + F(\mathbf{u})/\eta\|^2\Big\} = \operatorname*{argmin}_{\mathbf{v}\in\mathcal{U}}\Big\{\langle F(\mathbf{u}), \mathbf{v}\rangle + \frac{\eta}{2}\|\mathbf{v} - \mathbf{u}\|^2\Big\}.$$

Operator mapping generalizes a cocoercive operator to the constrained case: when $\mathcal{U} \equiv E$, $G_\eta \equiv F$.

It is a well-known fact that the projection operator is firmly-nonexpansive (Bauschke and Combettes, 2011, Proposition 4.16). Thus, Fact 3 can be used to show that, if F is $\frac{1}{L}$-cocoercive and $\eta \geq L$, then G_η is $\frac{3}{4\eta}$-cocoercive. This is shown in the following (simple) proposition, which can be found in a similar form in (Beck, 2017, Lemma 10.11) and is provided here for completeness.

Proposition 7 *Let F be a $\frac{1}{L}$-cocoercive operator and let G_η be defined as in Eq. (1), where $\eta \geq L$. Then G_η is $\frac{3}{4\eta}$-cocoercive.*

As G_η is $\frac{3}{4\eta}$-cocoercive, applying results from the beginning of the section to G_η, it is now immediate that Algorithm 2 (provided for completeness) produces $\mathbf{u}_k$ with $\|G_{L_k}(\mathbf{u}_k)\| \leq \epsilon$ after at most $\frac{\max\{8L/3, L_0\}\|\mathbf{u}_0 - \mathbf{u}^*\|}{\epsilon} + \max\{0, \log_2(8L/(3L_0))\}$ oracle queries to F (as each computation of G_η requires one oracle query to F). To complete this subsection, it remains to show that G_η

Algorithm 2: Parameter-Free Halpern – Cocoercive and Constrained Case

Input: $L_0 > 0$, $\epsilon > 0$, $\mathbf{u}_0 \in \mathcal{U}$. If not provided at the input, set $L_0 = 1$.

$\lambda_1 = \frac{1}{2}$, $k = 0$

$\bar{\mathbf{u}}_0 = \Pi_{\mathcal{U}}(\mathbf{u}_0 - F(\mathbf{u}_0)/L_0)$, $\bar{L}_0 = \frac{\|F(\bar{\mathbf{u}}_0) - F(\mathbf{u}_0)\|}{\|\bar{\mathbf{u}}_0 - \mathbf{u}_0\|}$

while $\|G_{L_k}(\mathbf{u}_k)\| > \epsilon/(1 + \bar{L}_k/L_k)$ **do**

1. $k = k + 1$
2. $L_k = L_{k-1}$
3. $p_k = \frac{L_{k-1}}{L_k}\frac{\lambda_{k-1}}{1-\lambda_{k-1}}$, $\lambda_k = \frac{p_k}{1+2p_k}$
4. $\mathbf{u}_k = \lambda_k\mathbf{u}_0 + (1 - \lambda_k)\bar{\mathbf{u}}_{k-1}$

   while $\langle G_{L_k}(\mathbf{u}_k) - G_{L_k}(\mathbf{u}_{k-1}), \mathbf{u}_k - \mathbf{u}_{k-1}\rangle < \frac{3}{4L_k}\|G_{L_k}(\mathbf{u}_k) - G_{L_k}(\mathbf{u}_{k-1})\|^2$ **do**

5. $L_k = 2 \cdot L_k$
6. $p_k = \frac{L_{k-1}}{L_k}\frac{\lambda_{k-1}}{1-\lambda_{k-1}}$, $\lambda_k = \frac{p_k}{1+2p_k}$
7. $\mathbf{u}_k = \lambda_k\mathbf{u}_0 + (1 - \lambda_k)(\mathbf{u}_{k-1} - G_{L_k}(\mathbf{u}_{k-1})/L_k)$

   end
8. $\bar{\mathbf{u}}_k = \Pi_{\mathcal{U}}(\mathbf{u}_k - F(\mathbf{u}_k)/L_k)$, $\bar{L}_k = \frac{\|F(\bar{\mathbf{u}}_k) - F(\mathbf{u}_k)\|}{\|\bar{\mathbf{u}}_k - \mathbf{u}_k\|}$, $L_k = \max\{L_k, \bar{L}_k\}$

end

return $\bar{\mathbf{u}}_k$, $\mathbf{u}_k$

is a good surrogate for approximating (MI) (and (SVI)). This is indeed the case and it follows as a suitable generalization of Lemma 3 from Ghadimi and Lan (2016), which is provided here for completeness.

Lemma 8 *Let G_η be defined as in Eq. (8). Denote $\bar{\mathbf{u}} = \Pi_{\mathcal{U}}(\mathbf{u} - F(\mathbf{u})/\eta)$, so that $G_\eta(\mathbf{u}) = \eta(\mathbf{u} - \bar{\mathbf{u}})$. If, for some $\mathbf{u} \in \mathcal{U}$, $\|G_\eta(\mathbf{u})\| \leq \epsilon$, then*

$$F(\bar{\mathbf{u}}) \in -\partial I_{\mathcal{U}}(\bar{\mathbf{u}}) + \mathcal{B}((1 + L_{\text{loc}}/\eta)\epsilon),$$

where $L_{\text{loc}} = \frac{\|F(\bar{\mathbf{u}}) - F(\mathbf{u})\|}{\|\bar{\mathbf{u}} - \mathbf{u}\|} \leq L$.

Proof As, by definition, $\bar{\mathbf{u}} = \operatorname{argmin}_{\mathbf{v} \in \mathcal{U}} \left\{ \langle F(\mathbf{u}), \mathbf{v} \rangle + \frac{\eta}{2}\|\mathbf{v} - \mathbf{u}\|^2 \right\}$, by first-order optimality of $\bar{\mathbf{u}}$, we have: $\mathbf{0} \in F(\mathbf{u}) + \eta(\bar{\mathbf{u}} - \mathbf{u}) + \partial I_{\mathcal{U}}(\bar{\mathbf{u}})$. Equivalently: $-F(\bar{\mathbf{u}}) \in F(\mathbf{u}) - F(\bar{\mathbf{u}}) - G_\eta(\mathbf{u}) + \partial I_{\mathcal{U}}(\bar{\mathbf{u}})$. The rest of the proof follows simply by using $\|G_\eta\| \leq \epsilon$ and $\|F(\mathbf{u}) - F(\bar{\mathbf{u}})\| = L_{\text{loc}}\|\mathbf{u} - \bar{\mathbf{u}}\| = \frac{L_{\text{loc}}}{\eta}\|G_\eta(\mathbf{u})\| \leq \frac{L_{\text{loc}}}{\eta}\epsilon$. ■

Lemma 8 implies that when the operator mapping is small in norm $\|\cdot\|$, then $\bar{\mathbf{u}} = \Pi_{\mathcal{U}}(\mathbf{u} - F(\mathbf{u})/\eta)$ is an approximate solution to (MI) corresponding to F on $\mathcal{U}$. We can now formally bound the number of oracle queries to F needed to approximate (MI) and (SVI).

Theorem 9 *Given $\mathbf{u}_0 \in \mathcal{U}$ and a $\frac{1}{L}$-cocoercive operator F, Algorithm 2 returns $\bar{\mathbf{u}}_k \in \mathcal{U}$ such that*

1. *$\|G_{L_k}(\bar{\mathbf{u}}_k)\| \leq \epsilon$, $\max_{\mathbf{v} \in \{\mathcal{U} \cap \mathcal{B}_{\bar{\mathbf{u}}_k}\}} \langle F(\bar{\mathbf{u}}_k), \bar{\mathbf{u}}_k - \mathbf{v} \rangle \leq \epsilon$ after at most $\frac{4\max\{8L/3, L_0\}\|\mathbf{u}_0 - \mathbf{u}^*\|}{\epsilon} + 2\max\{0, \log_2(\frac{8L}{3L_0})\}$ oracle queries to F;*
2. *$\max_{\mathbf{v} \in \mathcal{U}} \langle F(\bar{\mathbf{u}}), \bar{\mathbf{u}} - \mathbf{v} \rangle \leq \epsilon$ after at most $\frac{4\max\{8L/3, L_0\}\|\mathbf{u}_0 - \mathbf{u}^*\|D}{\epsilon} + 2\max\{0, \log_2(\frac{8L}{3L_0})\}$ oracle queries to F.*

Further, every point $\mathbf{u}_k$ that Algorithm 2 constructs is from the feasible set: $\mathbf{u}_k \in \mathcal{U}, \forall k \geq 0$, and a simple modification to the algorithm takes at most $\frac{\max\{8L/3, L_0\}\|\mathbf{u}_0 - \mathbf{u}^*\|}{\epsilon} + \max\{0, \log_2(8L/(3L_0))\}$ oracle queries to F to construct a point such that $\|G_{L_k}(\mathbf{u}_k)\| \leq \epsilon$.

Proof By the definition of G_η, if $\mathbf{u}_0 \in \mathcal{U}$, then $\mathbf{u}_k \in \mathcal{U}$, for all k. This follows simply as:

$$\begin{aligned}\mathbf{u}_{k+1} &= \lambda_{k+1}\mathbf{u}_0 + (1 - \lambda_{k+1})\Big(\mathbf{u}_k - \frac{1}{L_{k+1}}G_{L_{k+1}}(\mathbf{u}_k)\Big) \\ &= \lambda_{k+1}\mathbf{u}_0 + (1 - \lambda_{k+1})\Pi_{\mathcal{U}}(\mathbf{u}_k - F(\mathbf{u}_k)/L_{k+1}).\end{aligned}$$

Observe that, due to Line 8 of Algorithm 2, $L_k \geq \bar{L}_k$. The rest of the proof follows using Lemma 8, Fact 1, and the same reasoning as in the proof of Theorem 6. Observe that if the goal is to only output a point $\mathbf{u}_k$ such that $\|G_{L_k}(\mathbf{u}_k)\| \leq \epsilon$, then computing $\bar{\mathbf{u}}_k$ and $F(\bar{\mathbf{u}}_k)$ is not needed, and the algorithm can instead use $\|G_{L_k}(\mathbf{u}_k)\| > \epsilon$ as the exit condition in the outer while loop. ■

2.2. Setups with non-Cocoercive Lipschitz Operators

We now consider the case in which F is not cocoercive, but only monotone and L-Lipschitz. To obtain the desired convergence result, we make use of the resolvent operator, defined as $J_{F+\partial I_{\mathcal{U}}} = (\text{Id} + F + \partial I_{\mathcal{U}})^{-1}$. A useful property of the resolvent is that it is firmly-nonexpansive (Ryu and Boyd, 2016, and references therein), which, due to Fact 3, implies that $P = \text{Id} - J_{F+\partial I_{\mathcal{U}}}$ is $\frac{1}{2}$-cocoercive.

Finding a point $\mathbf{u} \in \mathcal{U}$ such that $\|P(\mathbf{u})\| \leq \epsilon$ is sufficient for approximating monotone inclusion (and (SVI)). This is shown in the following simple proposition, provided here for completeness.

Proposition 10 *Let $P = \mathrm{Id} - J_{F+\partial I_{\mathcal{U}}}$. If $\|P(\mathbf{u})\| \leq \epsilon$, then $\bar{\mathbf{u}} = \mathbf{u} - P(\mathbf{u}) = J_{F+\partial I_{\mathcal{U}}}(\mathbf{u})$ satisfies*

$$F(\bar{\mathbf{u}}) \in -\partial I_{\mathcal{U}}(\bar{\mathbf{u}}) + \mathcal{B}(\epsilon).$$

Proof By the definition of P and $J_{F+\partial I_{\mathcal{U}}}$, $\mathbf{u} - P(\mathbf{u}) = (\mathrm{Id} + F + \partial I_{\mathcal{U}})^{-1}(\mathbf{u})$. Equivalently:

$$\mathbf{u} - P(\mathbf{u}) + F(\mathbf{u} - P(\mathbf{u})) + \partial I_{\mathcal{U}}(\mathbf{u} - P(\mathbf{u})) \ni \mathbf{u}.$$

As $\|P(\mathbf{u})\| \leq \epsilon$, the result follows. ■

If we could compute the resolvent exactly, it would suffice to directly apply the result of Lieder (2017). However, excluding very special cases, computing the exact resolvent efficiently is generally not possible. However, since F is Lipschitz, the resolvent $J_{F+\partial I_{\mathcal{U}}}$ can be *approximated* efficiently. This is because it corresponds to solving a VI defined on a closed convex set $\mathcal{U}$ with the operator $F + \mathrm{Id}$ that is 1-strongly monotone and $(L+1)$-Lipschitz. Thus, it can be computed by solving a strongly monotone and Lipschitz VI, for which one can use the results of e.g., Nesterov and Scrimali (2011); Mokhtari et al. (2019); Gidel et al. (2019) if L is known, or Stonyakin et al. (2018), if L is not known. For completeness, we provide a simple modification to the Extragradient algorithm of Korpelevich (1977) in Algorithm 4 (Appendix A), for which we prove that it attains the optimal convergence rate without the knowledge of L. The convergence result is summarized in the following lemma, whose proof is provided in Appendix A.

Lemma 11 *Let $\bar{\mathbf{u}}_k^* = J_{F+I_{\mathcal{U}}}(\mathbf{u}_k)$, where $\mathbf{u}_k \in \mathcal{U}$ and F is L-Lipschitz. Then, there exists a parameter-free algorithm that queries F at most $O((L+1)\log(\frac{L\|\mathbf{u}_k - \bar{\mathbf{u}}_k^*\|}{\epsilon}))$ times and outputs a point $\bar{\mathbf{u}}_k$ such that $\|\bar{\mathbf{u}}_k - \bar{\mathbf{u}}_k^*\| \leq \epsilon$.*

To obtain the desired result, we need to prove the convergence of a Halpern iteration with inexact evaluations of the cocoercive operator P. Note that here we do know the cocoercivity parameter of P – it is equal to $1/2$. The resulting inexact version of Halpern's iteration for P is:

$$\begin{aligned} \mathbf{u}_{k+1} &= \lambda_{k+1}\mathbf{u}_0 + (1-\lambda_{k+1})(\mathbf{u}_k - \tilde{P}(\mathbf{u}_k)) \\ &= \lambda_{k+1}\mathbf{u}_0 + (1-\lambda_{k+1})\tilde{J}_{F+\partial I_{\mathcal{U}}}(\mathbf{u}_k), \end{aligned} \tag{9}$$

where $\tilde{P}(\mathbf{u}_k) - P(\mathbf{u}_k) = J_{F+\partial I_{\mathcal{U}}}(\mathbf{u}_k) - \tilde{J}_{F+\partial I_{\mathcal{U}}}(\mathbf{u}_k) = \mathbf{e}_k$ is the error.

To analyze the convergence of (9), we again use the potential function $\mathcal{C}_k$ from Eq. (7), with P as the operator. For simplicity of exposition, we take the best choice of $\lambda_i = \frac{1}{i+1}$ that can be obtained from Lemma 4 for $L_i = L = 2$, $\forall i$. The key result for this setting is provided in the following lemma, whose proof is deferred to the appendix.

Lemma 12 *Let $\mathcal{C}_k$ be defined as in Eq. (7) with P as the $\frac{1}{2}$-cocoercive operator, and let $L_k = 2$, $\lambda_k = \frac{1}{k+1}$, and $A_k = \frac{k(k+1)}{2}$, $\forall k \geq 1$. If the iterates $\mathbf{u}_k$ evolve according to (9) for an arbitrary initial point $\mathbf{u}_0 \in \mathcal{U}$, then:*

$$(\forall k \geq 1): \quad A_{k+1}\mathcal{C}_{k+1} \leq A_k\mathcal{C}_k + A_{k+1}\left\langle \mathbf{e}_k, (1-\lambda_{k+1})P(\mathbf{u}_k) - P(\mathbf{u}_{k+1})\right\rangle.$$

Further, if, $\forall k \geq 1$, $\|\mathbf{e}_{k-1}\| \leq \frac{\epsilon}{4k(k+1)}$, then $\|P(\mathbf{u}_K)\| \leq \epsilon$ after at most $K = \frac{4\|\mathbf{u}_0 - \mathbf{u}^*\|}{\epsilon}$ iterations.

We are now ready to state the algorithm and prove the main theorem for this subsection.

Algorithm 3: Parameter-Free Halpern – Monotone and Lipschitz Case
Input: $\epsilon > 0$, $\mathbf{u}_0 \in \mathcal{U}$
$k = 0$, $\epsilon_0 = \frac{\epsilon}{8}$
$\bar{\mathbf{u}}_0 = \tilde{J}_{F+\partial I_{\mathcal{U}}}(\mathbf{u}_0)$, where $\|\tilde{J}_{F+\partial I_{\mathcal{U}}}(\mathbf{u}_0) - J_{F+\partial I_{\mathcal{U}}}(\mathbf{u}_0)\| \le \epsilon_0$
$\tilde{P}(\mathbf{u}_0) = \mathbf{u}_0 - \bar{\mathbf{u}}_0$
while $\|\tilde{P}(\mathbf{u}_k)\| > \frac{3\epsilon}{4}$ **do**
  $k = k + 1$, $\lambda_k = \frac{1}{k+1}$, $\epsilon_k = \frac{\epsilon}{8(k+1)(k+2)}$
  $\mathbf{u}_k = \lambda_k \mathbf{u}_0 + (1-\lambda_k)\bar{\mathbf{u}}_{k-1}$
  $\bar{\mathbf{u}}_k = \tilde{J}_{F+\partial I_{\mathcal{U}}}(\mathbf{u}_k)$, where $\|\tilde{J}_{F+\partial I_{\mathcal{U}}}(\mathbf{u}_k) - J_{F+\partial I_{\mathcal{U}}}(\mathbf{u}_k)\| \le \epsilon_k$
  $\tilde{P}(\mathbf{u}_k) = \mathbf{u}_k - \bar{\mathbf{u}}_k$
end
return $\bar{\mathbf{u}}_k$, $\mathbf{u}_k$

Theorem 13 *Let F be a monotone and L-Lipschitz operator and let $\mathbf{u}_0 \in \mathcal{U}$ be an arbitrary initial point. For any $\epsilon > 0$, Algorithm 3 outputs a point with $\|P(\mathbf{u}_k)\| \le \epsilon$ after at most $\frac{8\|\mathbf{u}^* - \mathbf{u}_0\|}{\epsilon}$ iterations, where each iteration can be implemented with $O((L+1)\log(\frac{(L+1)\|\mathbf{u}_0 - \mathbf{u}^*\|}{\epsilon}))$ oracle queries to F. Hence, the total number of oracle queries to F is: $O\big(\frac{(L+1)\|\mathbf{u}_0 - \mathbf{u}^*\|}{\epsilon}\log\big(\frac{(L+1)\|\mathbf{u}_0 - \mathbf{u}^*\|}{\epsilon}\big)\big)$.*

Proof Recall that $\tilde{P}(\mathbf{u}_k) - P(\mathbf{u}_k) = \mathbf{e}_k$ and $\|\mathbf{e}_k\| = \epsilon_k = \frac{\epsilon}{8(k+1)(k+2)} < \frac{\epsilon}{4}$. Hence, as Algorithm 3 outputs a point $\mathbf{u}_k$ with $\|\tilde{P}(\mathbf{u}_k)\| \le \frac{3\epsilon}{4}$, by the triangle inequality, $\|P(\mathbf{u}_k)\| \le \epsilon$.

To bound the number of iterations until $\|\tilde{P}(\mathbf{u}_k)\| \le \frac{3\epsilon}{4}$, note that, again by the triangle inequality, if $\|P(\mathbf{u}_k)\| \le \epsilon/2$, then $\|\tilde{P}(\mathbf{u}_k)\| \le \frac{3\epsilon}{4}$. Applying Lemma 12, $\|P(\mathbf{u}_k)\| \le \epsilon/2$ after at most $k = \frac{8\|\mathbf{u}_0 - \mathbf{u}^*\|}{\epsilon}$ iterations, completing the proof of the first part of the theorem.

For the remaining part, using Lemma 11, $\tilde{J}_{F+\partial I_{\mathcal{U}}}(\mathbf{u}_k)$ can be computed (with target error ϵ_k) in $O((L+1)\log(\frac{(L+1)\|\mathbf{u}_k - J_{F+\partial I_{\mathcal{U}}}(\mathbf{u}_k)\|}{\epsilon_k})) = O((L+1)\log(\frac{(L+1)\|P(\mathbf{u}_k)\|}{\epsilon}))$ iterations, as $O(\log(\frac{1}{\epsilon_k})) = O(\log(\frac{1}{\epsilon}))$ and $P(\mathbf{u}_k) = \mathbf{u}_k - J_{F+\partial I_{\mathcal{U}}}(\mathbf{u}_k)$, by definition. It remains to use that $\|P(\mathbf{u}_k)\| = O(\|\mathbf{u}_0 - \mathbf{u}^*\|)$, which can be deduced from, e.g., Eq. (15) in the proof of Lemma 12. ■

Similarly as before, $\|P(\mathbf{u}_k)\| \le \epsilon$ implies an ϵ-approximate solution to (MI), by Proposition 10. When the diameter D is bounded, $\|P(\mathbf{u}_k)\| \le \frac{\epsilon}{D}$ implies an ϵ-approximate solution to (SVI).

Remark 14 *In degenerate cases where $L << 1$, instead of using the resolvent of $F + \partial I_{\mathcal{U}}$, one could use the resolvent of $F/\eta + \partial I_{\mathcal{U}}$ for $\eta = O(L)$, assuming the order of magnitude of L is known (this is typically a mild assumption). Then, each approximate computation of the resolvent would take $O((L/\eta + 1)\log(\frac{(L/\eta+1)\|\mathbf{u}_0 - \mathbf{u}^*\|}{\epsilon})$ oracle queries to F, and we would need to require that $\|\tilde{P}(\mathbf{u}_k)\| \le \frac{3\epsilon}{4\eta}$. Thus, the total number of queries to F would be $O((L+\eta)\log(\frac{(L+\eta)\|\mathbf{u}_0 - \mathbf{u}^*\|}{\epsilon}))$.*

2.3. Setups with Strongly Monotone and Lipschitz Operators

We now show that by restarting Algorithm 3, we can obtain a parameter-free method with near-optimal oracle complexity. To simplify the exposition, we assume w.l.o.g. that $L = \Omega(1)$. The proof of the following theorem is provided in Appendix A.

Theorem 15 *Given F that is L-Lipschitz and m-strongly monotone, consider running the following algorithm $\mathcal{A}$, starting with $\mathbf{u}_0 \in \mathcal{U}$:*

$(\mathcal{A})$: *At iteration k, call Algorithm 3 with error $\epsilon_k = \frac{7}{16}\|\tilde{P}(\mathbf{u}_{k-1})\|$ and initial point $\mathbf{u}_{k-1}$.*

Then, $\mathcal{A}$ outputs $\mathbf{u}_k \in \mathcal{U}$ with $\|P(\mathbf{u}_k)\| \leq \epsilon$ after at most $1 + \log_2(\frac{\|\mathbf{u}_0 - \mathbf{u}^*\|}{\epsilon})$ iterations, for any $\epsilon \in (0, \frac{1}{2}]$. The total number of queries to F until $\|P(\mathbf{u}_k)\| \leq \epsilon$ is $O\big((L+\frac{L}{m})\log(\frac{\|\mathbf{u}_0 - \mathbf{u}^*\|}{\epsilon})\log(L+\frac{L}{m})\big)$.

3. Lower Bound Reductions

In this section, we only state the lower bounds, while more details about the oracle model and the proof are deferred to Appendix A.

Lemma 16 *For any deterministic algorithm working in the operator oracle model and any $L, D > 0$, there exists an L-Lipschitz-continuous operator F and a closed convex feasible set $\mathcal{U}$ with diameter D such that:*

(a) For all $\epsilon > 0$ such that $k = \frac{LD^2}{\epsilon} = O(d)$, $\max_{\mathbf{u}\in\mathcal{U}} \langle F(\mathbf{u}_k), \mathbf{u}_k - \mathbf{u}\rangle = \Omega(\epsilon)$;

(b) For all $\epsilon > 0$ such that $k = \frac{LD}{\epsilon} = O(d)$, $\max_{\mathbf{u}\in\{\mathcal{U}\cap\mathcal{B}_{\mathbf{u}_k}\}} \langle F(\mathbf{u}_k), \mathbf{u}_k - \mathbf{u}\rangle = \Omega(\epsilon)$;

(c) If F is $\frac{1}{L}$-cocoercive, then for all $\epsilon > 0$ such that $k = \frac{LD}{\epsilon\log(D/\epsilon)} = O(d)$, it holds that

$$\max_{\mathbf{u}\in\{\mathcal{U}\cap\mathcal{B}_{\mathbf{u}_k}\}} \langle F(\mathbf{u}_k), \mathbf{u}_k - \mathbf{u}\rangle = \Omega(\epsilon);$$

(d) If F is m-strongly monotone, then for all $\epsilon > 0$ such that $k = \frac{L}{m} = O(d)$, it holds that

$$\max_{\mathbf{u}\in\{\mathcal{U}\cap\mathcal{B}_{\mathbf{u}_k}\}} \langle F(\mathbf{u}_k), \mathbf{u}_k - \mathbf{u}\rangle = \Omega(\epsilon).$$

Parts (a) and (b) of Lemma 16 certify that Algorithm 3 is optimal up to a logarithmic factor, due to Theorem 13. This is true because we can run Algorithm 3 with accuracy $\frac{\epsilon}{D}$ to obtain $\max_{\mathbf{u}\in\mathcal{U}} \langle F(\mathbf{u}_k), \mathbf{u}_k - \mathbf{u}\rangle = O(\epsilon)$ in $k = O(\frac{LD^2}{\epsilon}\log(\frac{LD}{\epsilon}))$ iterations, or with accuracy ϵ to obtain $\max_{\mathbf{u}\in\{\mathcal{U}\cap\mathcal{B}_{\mathbf{u}_k}\}} \langle F(\mathbf{u}_k), \mathbf{u}_k - \mathbf{u}\rangle = O(\epsilon)$ in $k = O(\frac{LD}{\epsilon}\log(\frac{LD}{\epsilon}))$ iterations (see Proposition 10).

Part (c) of Lemma 16 certifies that Algorithm 2 is optimal up to a $\log(D/\epsilon)$ factor, due to Theorem 9. Part (d) certifies that the restarting algorithm from Theorem 15 is optimal up to a factor $\log(D/\epsilon)\log(L/m)$ whenever $L = \Omega(L/m)$. Note that $L = \Omega(L/m)$ can be ensured by a proper scaling of the problem instance, as any such scaling would leave the condition number L/m unaffected and would only impact the target error ϵ, which only appears under a logarithm.

4. Conclusion

We showed that variants of Halpern iteration can be used to obtain near-optimal methods for solving different classes of monotone inclusion problems with Lipschitz operators. The results highlight connections between monotone inclusion, variational inequalities, fixed points of nonexpansive maps, and proximal-point-type algorithms. Some interesting questions that merit further investigation remain. In particular, one open question that arises is to close the gap between the upper and lower bounds provided here. We conjecture that the optimal complexity of monotone inclusion is: (i) $\Theta(\frac{LD}{\epsilon})$ when the operator is either L-Lipschitz or $\frac{1}{L}$-cocoercive, and (ii) $\Theta(\frac{L}{m}\log(\frac{LD}{\epsilon}))$ when the operator is L-Lipschitz and m-strongly monotone.

Acknowledgements

We thank Prof. Ulrich Kohlenbach for useful comments and pointers to the literature. We also thank Howard Heaton for pointing out a typo in the proof of Lemma 2.1 in a previous version of this paper.

References

Martin Arjovsky and Leon Bottou. Towards principled methods for training generative adversarial networks. In *Proc. ICLR'17*, 2017.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN. *arXiv preprint arXiv:1701.07875*, 2017.

Hilal Asi and John C Duchi. Stochastic (approximate) proximal point methods: Convergence, optimality, and adaptivity. *SIAM Journal on Optimization*, 29(3):2257–2290, 2019.

Francis Bach and Kfir Y Levy. A universal algorithm for variational inequalities adaptive to smoothness and noise. In *Proc. COLT'19*, 2019.

Heinz H Bauschke. The approximation of fixed points of compositions of nonexpansive mappings in Hilbert space. *Journal of Mathematical Analysis and Applications*, 202(1):150–159, 1996.

Heinz H Bauschke and Patrick L Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*, volume 408. Springer, 2011.

Amir Beck. *First-order methods in optimization*, volume 25. SIAM, 2017.

Felix E Browder. Convergence of approximants to fixed points of nonexpansive nonlinear mappings in Banach spaces. *Archive for Rational Mechanics and Analysis*, 24(1):82–90, 1967.

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.

Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

Yoel Drori and Marc Teboulle. Performance of first-order methods for smooth convex minimization: a novel approach. *Mathematical Programming*, 145(1-2):451–482, 2014.

Francisco Facchinei and Jong-Shi Pang. *Finite-dimensional variational inequalities and complementarity problems*. Springer Science & Business Media, 2003.

Saeed Ghadimi and Guanghui Lan. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Mathematical Programming*, 156(1-2):59–99, 2016.

Gauthier Gidel, Hugo Berard, Gaëtan Vignoud, Pascal Vincent, and Simon Lacoste-Julien. A variational inequality perspective on generative adversarial networks. In *Proc. ICLR'19*, 2019.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proc. NIPS'14*, 2014.

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

Donghwan Kim. Accelerated proximal point method and forward method for monotone inclusions. *arXiv preprint arXiv:1905.05149*, 2019.

Donghwan Kim and Jeffrey A Fessler. Optimizing the efficiency of first-order methods for decreasing the gradient of smooth convex functions. *arXiv preprint arXiv:1803.06600*, 2018.

Ulrich Kohlenbach. *Applied proof theory: proof interpretations and their use in mathematics*. Springer Science & Business Media, 2008.

Ulrich Kohlenbach. On quantitative versions of theorems due to fe browder and r. wittmann. *Advances in Mathematics*, 226(3):2764–2795, 2011.

Daniel Körnlein. Quantitative results for halpern iterations of nonexpansive mappings. *Journal of Mathematical Analysis and Applications*, 428(2):1161–1172, 2015.

GM Korpelevich. Extragradient method for finding saddle points and other problems. *Matekon*, 13 (4):35–49, 1977.

Laurentiu Leustean. Rates of asymptotic regularity for halpern iterations of nonexpansive mappings. *Journal of Universal Computer Science*, 13(11):1680–1691, 2007.

Felix Lieder. On the convergence rate of the Halpern-iteration, 2017. http://www.optimization-online.org/DB_FILE/2017/11/6336.pdf.

Hongzhou Lin, Julien Mairal, and Zaid Harchaoui. Catalyst acceleration for first-order convex optimization: From theory to practice. *The Journal of Machine Learning Research*, 18(1):7854–7907, 2017.

Qihang Lin, Mingrui Liu, Hassan Rafique, and Tianbao Yang. Solving weakly-convex-weakly-concave saddle-point problems as weakly-monotone variational inequality. *arXiv preprint arXiv:1810.10207*, 2018.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *Proc. ICLR'18*, 2018.

Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. *arXiv preprint arXiv:1901.08511*, 2019.

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.

Yurii Nesterov. Dual extrapolation and its applications to solving variational inequalities and related problems. *Mathematical Programming*, 109(2-3):319–344, 2007.

Yurii Nesterov. How to make the gradients small. *Optima. Mathematical Optimization Society Newsletter*, (88):10–11, 2012.

Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

Yurii Nesterov and Laura Scrimali. Solving strongly monotone variational and quasi-variational inequalities. *Discrete & Continuous Dynamical Systems-A*, 31(4):1383–1396, 2011.

Yuyuan Ouyang and Yangyang Xu. Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems. *Mathematical Programming*, Aug 2019.

Ernest K Ryu and Stephen Boyd. Primer on monotone operator methods. *Applied and Computational Mathematics*, 15(1):3–43, 2016.

Ernest K Ryu, Kun Yuan, and Wotao Yin. ODE analysis of stochastic gradient methods with optimism and anchoring for minimax problems and GANs. *arXiv preprint arXiv:1905.10899*, 2019.

Fedor Stonyakin, Alexander Gasnikov, Pavel Dvurechensky, Mohammad Alkousa, and Alexander Titov. Generalized mirror prox for monotone variational inequalities: Universality and inexact oracle. *arXiv preprint arXiv:1806.05140*, 2018.

Kiran K Thekumparampil, Prateek Jain, Praneeth Netrapalli, and Sewoong Oh. Efficient algorithms for smooth minimax optimization. In *Proc. NeurIPS'19*, 2019.

Rainer Wittmann. Approximation of fixed points of nonexpansive mappings. *Archiv der Mathematik*, 58(5):486–491, 1992.

Hong-Kun Xu. Iterative algorithms for nonlinear operators. *Journal of the London Mathematical Society*, 66(1):240–256, 2002.

Appendix A. Omitted Proofs

A.1. Unconstrained Setting with a Cocoercive Operator

Lemma 4 *Let $\mathcal{C}_k$ be defined as in Eq. (7) and let $\mathbf{u}^*$ be the solution to (MI) that minimizes $\|\mathbf{u}_0 - \mathbf{u}^*\|$. Assume further that $\langle F(\mathbf{u}_1) - F(\mathbf{u}_0), \mathbf{u}_1 - \mathbf{u}_0\rangle \geq \frac{1}{L_1}\|F(\mathbf{u}_1) - F(\mathbf{u}_0)\|^2$. If $A_{k+1}\mathcal{C}_{k+1} \leq A_k\mathcal{C}_k$, $\forall k \geq 1$, where $\{A_i\}_{i\geq 1}$ is a sequence of positive numbers that satisfies $A_1 = 1$, then:*

$$(\forall k \geq 1): \quad \|F(\mathbf{u}_k)\| \leq L_k \frac{\lambda_k}{1-\lambda_k}\|\mathbf{u}_0 - \mathbf{u}^*\|.$$

Proof The statement holds trivially if $\|F(\mathbf{u}_k)\| = 0$, so assume that $\|F(\mathbf{u}_k)\| > 0$. Under the assumption of the lemma, we have that $A_k\mathcal{C}_k \leq \mathcal{C}_1$, $\forall k \geq 1$. From (H) and $\lambda_1 = \frac{1}{2}$, $\mathbf{u}_1 = \mathbf{u}_0 - \frac{1}{L_1}F(\mathbf{u}_0)$, and thus: $\mathcal{C}_1 = \frac{1}{L_1}\|F(\mathbf{u}_1)\|^2 - \frac{1}{L_1}\langle F(\mathbf{u}_1), F(\mathbf{u}_0)\rangle$.

Let $\mathbf{u}^*$ be an arbitrary solution to (MI) (and thus also to (MVI)). As $\langle F(\mathbf{u}_1) - F(\mathbf{u}_0), \mathbf{u}_1 - \mathbf{u}_0\rangle \geq \frac{1}{L_1}\|F(\mathbf{u}_1) - F(\mathbf{u}_0)\|^2$ and $\mathbf{u}_1 = \mathbf{u}_0 - \frac{1}{L_1}F(\mathbf{u}_0)$, it follows that $\|F(\mathbf{u}_1)\|^2 \leq \langle F(\mathbf{u}_0), F(\mathbf{u}_1)\rangle$, and, thus $\mathcal{C}_1 \leq 0$. Further, as $A_k > 0$, we also have $\mathcal{C}_k \leq 0$, and, hence:

$$\begin{aligned}
\|F(\mathbf{u}_k)\|^2 &\leq L_k \frac{\lambda_k}{1-\lambda_k}\langle F(\mathbf{u}_k), \mathbf{u}_0 - \mathbf{u}_k\rangle \\
&= L_k \frac{\lambda_k}{1-\lambda_k}\langle F(\mathbf{u}_k), \mathbf{u}_0 - \mathbf{u}^*\rangle + L_k \frac{\lambda_k}{1-\lambda_k}\langle F(\mathbf{u}_k), \mathbf{u}^* - \mathbf{u}_k\rangle \\
&\leq L_k \frac{\lambda_k}{1-\lambda_k}\langle F(\mathbf{u}_k), \mathbf{u}_0 - \mathbf{u}^*\rangle \leq L_k \frac{\lambda_k}{1-\lambda_k}\|F(\mathbf{u}_k)\| \cdot \|\mathbf{u}_0 - \mathbf{u}^*\|,
\end{aligned}$$

where the last line is by $\mathbf{u}^*$ being a solution to (MVI) and by the Cauchy-Schwarz inequality. The conclusion of the lemma now follows by dividing both sides of $\|F(\mathbf{u}_k)\|^2 \leq L_k \frac{\lambda_k}{1-\lambda_k}\|F(\mathbf{u}_k)\| \cdot \|\mathbf{u}_0 - \mathbf{u}^*\|$ by $\|F(\mathbf{u}_k)\|$ and observing that the statement holds for an arbitrary solution $\mathbf{u}^*$ to (MI), and thus, it also holds for the one that minimizes the distance to $\mathbf{u}_0$. ■

Lemma 5 *Let $\mathcal{C}_k$ be defined as in Eq. (7). Let $\{A_i\}_{i\geq 1}$ be defined recursively as $A_1 = 1$ and $A_{k+1} = A_k \frac{\lambda_k}{(1-\lambda_k)\lambda_{k+1}}$ for $k \geq 1$. Assume that $\{\lambda_i\}_{i\geq 1}$ is chosen so that $\lambda_1 = \frac{1}{2}$ and for $k \geq 1$: $\frac{\lambda_{k+1}}{1-2\lambda_{k+1}} \geq \frac{\lambda_k L_k}{(1-\lambda_k)L_{k+1}}$. Finally, assume that $L_k \in (0,\infty)$ and $\langle F(\mathbf{u}_k) - F(\mathbf{u}_{k-1}), \mathbf{u}_k - \mathbf{u}_{k-1}\rangle \geq \frac{1}{L_k}\|F(\mathbf{u}_k) - F(\mathbf{u}_{k-1})\|^2$, $\forall k$. Then,*

$$(\forall k \geq 1): \quad A_{k+1}\mathcal{C}_{k+1} \leq A_k \mathcal{C}_k.$$

Proof By the assumption of the lemma,

$$\frac{1}{L_{k+1}}\|F(\mathbf{u}_{k+1}) - F(\mathbf{u}_k)\|^2 \leq \langle F(\mathbf{u}_{k+1}) - F(\mathbf{u}_k), \mathbf{u}_{k+1} - \mathbf{u}_k\rangle,$$

which, after expanding the left-hand side, can be equivalently written as:

$$\frac{1}{L_{k+1}}\|F(\mathbf{u}_{k+1})\|^2 \leq \big\langle F(\mathbf{u}_{k+1}), \mathbf{u}_{k+1} - \mathbf{u}_k + \frac{2}{L_{k+1}}F(\mathbf{u}_k)\big\rangle - \big\langle F(\mathbf{u}_k), \mathbf{u}_{k+1} - \mathbf{u}_k + \frac{1}{L_{k+1}}F(\mathbf{u}_k)\big\rangle.$$

From (H), we have that $\mathbf{u}_{k+1} - \mathbf{u}_k = \frac{\lambda_{k+1}}{1-\lambda_{k+1}}(\mathbf{u}_0 - \mathbf{u}_{k+1}) - \frac{2}{L_{k+1}}F(\mathbf{u}_k)$ and $\mathbf{u}_{k+1} - \mathbf{u}_k = \lambda_{k+1}(\mathbf{u}_0 - \mathbf{u}_k) - \frac{2(1-\lambda_{k+1})}{L_{k+1}}F(\mathbf{u}_k)$. Hence:

$$\begin{aligned}\frac{1}{L_{k+1}}\|F(\mathbf{u}_{k+1})\|^2 \leq{}& \frac{\lambda_{k+1}}{1-\lambda_{k+1}}\langle F(\mathbf{u}_{k+1}), \mathbf{u}_0 - \mathbf{u}_{k+1}\rangle - \lambda_{k+1}\langle F(\mathbf{u}_k), \mathbf{u}_0 - \mathbf{u}_k\rangle \\ &+ \frac{1-2\lambda_{k+1}}{L_{k+1}}\|F(\mathbf{u}_k)\|^2.\end{aligned}$$

Rearranging the last inequality and multiplying both sides by A_{k+1}, we have:

$$\begin{aligned}&A_{k+1}\Big(\frac{1}{L_{k+1}}\|F(\mathbf{u}_{k+1})\|^2 - \frac{\lambda_{k+1}}{1-\lambda_{k+1}}\langle F(\mathbf{u}_{k+1}), \mathbf{u}_0 - \mathbf{u}_{k+1}\rangle\Big) \\ &\qquad \leq \frac{A_{k+1}(1-2\lambda_{k+1})}{L_{k+1}}\|F(\mathbf{u}_k)\|^2 - A_{k+1}\lambda_{k+1}\langle F(\mathbf{u}_k), \mathbf{u}_0 - \mathbf{u}_k\rangle.\end{aligned}$$

The left-hand side of the last inequality if precisely $A_{k+1}\mathcal{C}_{k+1}$. The right-hand side is at most $A_k\mathcal{C}_k$, by the choice of sequences $\{A_i\}_{i\geq 1}$, $\{\lambda_i\}_{i\geq 1}$. ■

A.2. Operator Mapping

Proposition 7 *Let F be a $\frac{1}{L}$-cocoercive operator and let G_η be defined as in Eq. (1), where $\eta \geq L$. Then G_η is $\frac{3}{4\eta}$-cocoercive.*

Proof As $\Pi_{\mathcal{U}}$ is firmly nonexpansive, it is also 1-cocoercive (by Fact 3). Thus, using that, by definition of G_η, $\mathbf{u} - \Pi_{\mathcal{U}}(\mathbf{u}) = \frac{1}{\eta} G_\eta(\mathbf{u})$, $\forall \mathbf{u} \in \mathcal{U}$, we have, $\forall \mathbf{u}, \mathbf{v} \in E$:

$$\begin{aligned}
&\left\langle \Pi_{\mathcal{U}}\Big(\mathbf{u} - \frac{1}{\eta}F(\mathbf{u})\Big) - \Pi_{\mathcal{U}}\Big(\mathbf{v} - \frac{1}{\eta}F(\mathbf{v})\Big), \mathbf{u} - \frac{1}{\eta}F(\mathbf{u}) - \Big(\mathbf{v} - \frac{1}{\eta}F(\mathbf{v})\Big) \right\rangle \\
&\qquad = \left\langle \frac{1}{\eta}(G_\eta(\mathbf{v}) - G_\eta(\mathbf{u})) + \mathbf{u} - \mathbf{v}, \mathbf{u} - \mathbf{v} - \frac{1}{\eta}(F(\mathbf{u}) - F(\mathbf{v})) \right\rangle \\
&\qquad \geq \left\| \frac{1}{\eta}(G_\eta(\mathbf{v}) - G_\eta(\mathbf{u})) + \mathbf{u} - \mathbf{v} \right\|^2.
\end{aligned}$$

Expanding the square on the right-hand side and rearranging, we get:

$$\begin{aligned}
\frac{1}{\eta^2}\|G_\eta(\mathbf{v}) - G_\eta(\mathbf{u})\|^2 \leq \frac{1}{\eta}\left\langle G_\eta(\mathbf{u}) - G_\eta(\mathbf{v}), \mathbf{u} - \mathbf{v}\right\rangle \\
+ \frac{1}{\eta^2}\left\langle G_\eta(\mathbf{u}) - G_\eta(\mathbf{v}), F(\mathbf{u}) - F(\mathbf{v})\right\rangle - \frac{1}{\eta}\left\langle F(\mathbf{u}) - F(\mathbf{v}), \mathbf{u} - \mathbf{v}\right\rangle.
\end{aligned}$$

As $\eta \geq L$ and F is $\frac{1}{L}$-cocoercive, $\frac{1}{\eta}\left\langle F(\mathbf{u}) - F(\mathbf{v}), \mathbf{u} - \mathbf{v}\right\rangle \geq \frac{1}{\eta^2}\|F(\mathbf{u}) - F(\mathbf{v})\|^2$. It remains to apply Young's inequality, which implies $\left\langle G_\eta(\mathbf{u}) - G_\eta(\mathbf{v}), F(\mathbf{u}) - F(\mathbf{v})\right\rangle \leq \frac{\varepsilon}{2}\|G_\eta(\mathbf{u}) - G_\eta(\mathbf{v})\|^2 + \frac{1}{2\varepsilon}\|F(\mathbf{u}) - F(\mathbf{v})\|^2$, $\forall \varepsilon > 0$, and choose $\varepsilon = \frac{1}{2}$. ■

A.3. Approximating the Resolvent

Let us start by proving the convergence of a variant of the Extragradient method of Korpelevich (1977) that does not require the knowledge of the Lipschitz constant L (but does require knowledge of the strong monotonicity parameter m; when computing the resolvent we have $m = 1$). The algorithm is summarized in Algorithm 4. Observe that the update step for $\mathbf{u}_k$ from Lines 6 and 10 can be written in the form of a projection onto $\mathcal{U}$; we chose to write it in the current form as it is more convenient for the analysis.

We now bound the convergence of Algorithm 4.

Lemma 17 *Let $a_0 > 0$ and let F be m-strongly monotone and L-Lipschitz. Then, Algorithm 4 outputs a point $\mathbf{u}_k$ with $\|\mathbf{u}_k - \mathbf{u}^*\| \leq \epsilon$ after at most $k = O\big(\frac{L}{m}\log(\frac{L\|\mathbf{u}_0 - \mathbf{u}^*\|}{m\epsilon})\big)$ oracle queries to F, where $\mathbf{u}^*$ solves (SVI).*

Proof Define $A_k = \sum_{i=0}^k a_i$. To prove the lemma, we will use the following gap (or merit) functions:

$$f_k = \frac{1}{A_k}\sum_{i=0}^k a_i\Big(\left\langle F(\bar{\mathbf{u}}_i), \bar{\mathbf{u}}_i - \mathbf{u}^*\right\rangle - \frac{m}{2}\|\bar{\mathbf{u}}_i - \mathbf{u}^*\|^2\Big).$$

As F is strongly monotone and $\mathbf{u}^*$ solves (SVI), $f_k \geq 0$, $\forall k$. By convention, we take $f_{-1} = 0$ and $A_{-1} = 0$, so that $A_k f_k - A_{k-1}f_{k-1} = a_k\Big(\left\langle F(\bar{\mathbf{u}}_k), \bar{\mathbf{u}}_k - \mathbf{u}^*\right\rangle - \frac{m}{2}\|\bar{\mathbf{u}}_k - \mathbf{u}^*\|^2\Big)$. Let us now bound $A_k f_k - A_{k-1}f_{k-1}$, and observe that $A_k f_k - A_{k-1}f_{k-1} \geq 0$. First, write

$$\begin{aligned}
A_k f_k - A_{k-1}f_{k-1} &= a_k\Big(\left\langle F(\bar{\mathbf{u}}_k), \bar{\mathbf{u}}_k - \mathbf{u}^*\right\rangle - \frac{m}{2}\|\bar{\mathbf{u}}_k - \mathbf{u}^*\|^2\Big) \\
&= a_k\left\langle F(\bar{\mathbf{u}}_k), \mathbf{u}_{k+1} - \mathbf{u}^*\right\rangle + a_k\left\langle F(\mathbf{u}_k), \bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\right\rangle \\
&\quad + a_k\left\langle F(\bar{\mathbf{u}}_k) - F(\mathbf{u}_k), \bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\right\rangle - \frac{a_k m}{2}\|\bar{\mathbf{u}}_k - \mathbf{u}^*\|^2.
\end{aligned} \tag{10}$$

Algorithm 4: EG Without the Knowledge of L

Input: $a_0,\ \mathbf{u}_0 \in \mathcal{U},\ m, \epsilon$. If not provided at the input or $> 1/m$, set $a_0 = 1/m$.

$\bar{\mathbf{u}}_0 = \Pi_{\mathcal{U}}(\mathbf{u}_0 - a_k F(\mathbf{u}_0))$

$k = 0$, $\delta_0 = \frac{a_0 m \epsilon}{5\sqrt{2}}$

while $\|\bar{\mathbf{u}}_k - \mathbf{u}_k\| > \delta_k$ **do**

  $k = k + 1$, $a_k = a_{k-1}$

  $\bar{\mathbf{u}}_k = \Pi_{\mathcal{U}}(\mathbf{u}_k - a_k F(\mathbf{u}_k))$

  $\mathbf{u}_{k+1} = \operatorname{argmin}_{\mathbf{u}\in\mathcal{U}} \left\{ a_k \langle F(\bar{\mathbf{u}}_k), \mathbf{u}\rangle + \frac{a_k m}{2}\|\mathbf{u} - \bar{\mathbf{u}}_k\|^2 + \frac{1}{2}\|\mathbf{u} - \mathbf{u}_k\|^2 \right\}$

  while $a_k \langle F(\bar{\mathbf{u}}_k) - F(\mathbf{u}_k), \bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\rangle > \frac{1}{4}\|\mathbf{u}_{k+1} - \bar{\mathbf{u}}_k\|^2 + \frac{1}{4}\|\bar{\mathbf{u}}_k - \mathbf{u}_k\|^2$ **do**

    $a_k = \min\left\{ \frac{a_k}{2},\ \frac{\|\bar{\mathbf{u}}_k - \mathbf{u}_k\|}{\|F(\bar{\mathbf{u}}_k) - F(\mathbf{u}_k)\|} \right\}$

    $\bar{\mathbf{u}}_k = \Pi_{\mathcal{U}}(\mathbf{u}_k - a_k F(\mathbf{u}_k))$

    $\mathbf{u}_{k+1} = \operatorname{argmin}_{\mathbf{u}\in\mathcal{U}} \left\{ a_k \langle F(\bar{\mathbf{u}}_k), \mathbf{u}\rangle + \frac{a_k m}{2}\|\mathbf{u} - \bar{\mathbf{u}}_k\|^2 + \frac{1}{2}\|\mathbf{u} - \mathbf{u}_k\|^2 \right\}$

  end

  $\delta_k = \frac{a_k m \epsilon}{5\sqrt{2}}$

end

return $\mathbf{u}_k$

By the first-order optimality of $\mathbf{u}_{k+1}$ in its definition, we have, $\forall \mathbf{u}$:

$$\langle a_k F(\bar{\mathbf{u}}_k) + a_k m(\mathbf{u}_{k+1} - \bar{\mathbf{u}}_k) + \mathbf{u}_{k+1} - \mathbf{u}_k, \mathbf{u} - \mathbf{u}_{k+1}\rangle \geq 0,$$

and, thus:

$$a_k \langle F(\bar{\mathbf{u}}_k), \mathbf{u}_{k+1} - \mathbf{u}\rangle \leq a_k m \langle \mathbf{u}_{k+1} - \bar{\mathbf{u}}_k, \mathbf{u} - \mathbf{u}_{k+1}\rangle + \langle \mathbf{u}_{k+1} - \mathbf{u}_k, \mathbf{u} - \mathbf{u}_{k+1}\rangle .$$

By the standard three-point identity (which can also be verified directly):

$$\langle \mathbf{u}_{k+1} - \mathbf{u}_k, \mathbf{u} - \mathbf{u}_{k+1}\rangle = \frac{1}{2}\|\mathbf{u} - \mathbf{u}_k\|^2 - \frac{1}{2}\|\mathbf{u} - \mathbf{u}_{k+1}\|^2 - \frac{1}{2}\|\mathbf{u}_k - \mathbf{u}_{k+1}\|^2.$$

Thus, setting $\mathbf{u} = \mathbf{u}^*$:

$$\begin{aligned} a_k \langle F(\bar{\mathbf{u}}_k), \mathbf{u}_{k+1} - \mathbf{u}^*\rangle = {} & a_k m \langle \mathbf{u}_{k+1} - \bar{\mathbf{u}}_k, \mathbf{u}^* - \mathbf{u}_{k+1}\rangle \\ & + \frac{1}{2}\|\mathbf{u}^* - \mathbf{u}_k\|^2 - \frac{1}{2}\|\mathbf{u}^* - \mathbf{u}_{k+1}\|^2 - \frac{1}{2}\|\mathbf{u}_k - \mathbf{u}_{k+1}\|^2. \end{aligned}$$

Observe also that:

$$\langle \mathbf{u}_{k+1} - \bar{\mathbf{u}}_k, \mathbf{u}^* - \mathbf{u}_{k+1}\rangle = \frac{1}{2}\|\mathbf{u}^* - \bar{\mathbf{u}}_k\|^2 - \frac{1}{2}\|\mathbf{u}_{k+1} - \bar{\mathbf{u}}_k\|^2 - \|\mathbf{u}^* - \mathbf{u}_{k+1}\|^2.$$

Thus, we have:

$$\begin{aligned} a_k \langle F(\bar{\mathbf{u}}_k), \mathbf{u}_{k+1} - \mathbf{u}^*\rangle = {} & \frac{1}{2}\|\mathbf{u}^* - \mathbf{u}_k\|^2 - \frac{1 + a_k m}{2}\|\mathbf{u}^* - \mathbf{u}_{k+1}\|^2 \\ & - \frac{1}{2}\|\mathbf{u}_k - \mathbf{u}_{k+1}\|^2 + \frac{a_k m}{2}\|\mathbf{u}^* - \bar{\mathbf{u}}_k\|^2 - \frac{a_k m}{2}\|\mathbf{u}_{k+1} - \bar{\mathbf{u}}_k\|^2. \end{aligned} \tag{11}$$

By similar arguments:

$$a_k \langle F(\mathbf{u}_k), \bar{\mathbf{u}}_k - \mathbf{u}_{k+1} \rangle = \frac{1}{2}\|\mathbf{u}_{k+1} - \mathbf{u}_k\|^2 - \frac{1}{2}\|\mathbf{u}_{k+1} - \bar{\mathbf{u}}_k\|^2 - \frac{1}{2}\|\bar{\mathbf{u}}_k - \mathbf{u}_k\|^2. \tag{12}$$

Combining Eq. (10)-(12):

$$\begin{aligned} A_k f_k - A_{k-1} f_{k-1} = \frac{1}{2}\|\mathbf{u}^* - \mathbf{u}_k\|^2 - \frac{1 + a_k m}{2}\|\mathbf{u}^* - \mathbf{u}_{k+1}\|^2 + a_k \langle F(\bar{\mathbf{u}}_k) - F(\mathbf{u}_k), \bar{\mathbf{u}}_k - \mathbf{u}_{k+1} \rangle \\ - \frac{1 + a_k m}{2}\|\mathbf{u}_{k+1} - \bar{\mathbf{u}}_k\|^2 - \frac{1}{2}\|\bar{\mathbf{u}}_k - \mathbf{u}_k\|^2. \end{aligned}$$

By the condition of the while loop in Line 7 of Algorithm 4, and because $A_k f_k - A_{k-1} f_{k-1} \geq 0$,

$$\frac{1 + a_k m}{2}\|\mathbf{u}^* - \mathbf{u}_{k+1}\|^2 + \frac{1 + 2a_k m}{4}\|\mathbf{u}_{k+1} - \bar{\mathbf{u}}_k\|^2 + \frac{1}{4}\|\bar{\mathbf{u}}_k - \mathbf{u}_k\|^2 \leq \frac{1}{2}\|\mathbf{u}^* - \mathbf{u}_k\|^2. \tag{13}$$

The condition of the while loop in Line 7 of Algorithm 4 is satisfied for any $a_k \leq \frac{1}{2L}$, as

$$\begin{aligned} a_k \langle F(\bar{\mathbf{u}}_k) - F(\mathbf{u}_k), \bar{\mathbf{u}}_k - \mathbf{u}_{k+1} \rangle &\leq a_k L \|\bar{\mathbf{u}}_k - \mathbf{u}_k\| \cdot \|\bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\| \\ &\leq \frac{a_k L}{2}\big(\|\bar{\mathbf{u}}_k - \mathbf{u}_k\|^2 + \|\bar{\mathbf{u}}_k - \mathbf{u}_{k+1}\|^2\big), \end{aligned}$$

where we have used the Cauchy-Schwarz inequality, the fact that F is L-Lipschitz, and the Young inequality. Thus, in any iteration, $a_k > \frac{1}{4L}$, and the total number of times the while loop from Line 7 is entered is at most $\log_2(4L/a_0)$.

From Eq. (13), $\|\mathbf{u}^* - \mathbf{u}_{k+1}\|^2 \leq \frac{1}{1+m/(4L)}\|\mathbf{u}^* - \mathbf{u}_k\|^2 \leq (1 - \frac{m}{8L})\|\mathbf{u}^* - \mathbf{u}_k\|^2$. Thus, for any $\delta > 0$, $\|\mathbf{u}^* - \mathbf{u}_k\| \leq \delta$ for $k \geq \frac{16L}{m}\log(\frac{\|\mathbf{u}^* - \mathbf{u}_0\|}{\delta})$. Consequently, from Eq. (13), $\|\bar{\mathbf{u}}_k - \mathbf{u}_k\| \leq \sqrt{2}\delta$ whenever $\|\mathbf{u}^* - \mathbf{u}_k\| \leq \delta$. In particular, for $\delta = \frac{a_k m \epsilon}{5\sqrt{2}} \geq \frac{m\epsilon}{20\sqrt{2}L}$, $\|\bar{\mathbf{u}}_k - \mathbf{u}_k\| \leq \sqrt{2}\delta = \frac{a_k m}{5}\epsilon$ after at most $k = \frac{16L}{m}\log(\frac{20\sqrt{2}L\|\mathbf{u}^* - \mathbf{u}_0\|}{m\epsilon}$ (outer loop) iterations.

It remains to show that when $\|\bar{\mathbf{u}}_k - \mathbf{u}_k\| \leq \delta$, $\|\mathbf{u}_k - \mathbf{u}^*\| \leq \epsilon$, and so Algorithm 4 terminates. Observe that $\mathbf{u}_k - \bar{\mathbf{u}}_k = a_k G_{1/a_k}(\mathbf{u}_k)$, where G_{1/a_k} is the operator mapping defined in Eq. (8). Thus, using Lemma 8 and noting that $a_k \leq 1/L_{\text{loc}} = \frac{\|\bar{\mathbf{u}}_k - \mathbf{u}_k\|}{\|F(\bar{\mathbf{u}}_k) - F(\mathbf{u}_k)\|}$, if $\|\bar{\mathbf{u}}_k - \mathbf{u}_k\| \leq \frac{a_k m}{5}\epsilon$, we have

$$\langle F(\bar{\mathbf{u}}_k), \bar{\mathbf{u}}_k - \mathbf{u}^* \rangle \leq \frac{2m}{5}\epsilon\|\bar{\mathbf{u}}_k - \mathbf{u}^*\|.$$

On the other hand, as F is m-strongly monotone, we also have $\langle F(\bar{\mathbf{u}}_k), \bar{\mathbf{u}}_k - \mathbf{u}^* \rangle \geq \frac{m}{2}\|\bar{\mathbf{u}}_k - \mathbf{u}^*\|^2$. Hence, $\|\bar{\mathbf{u}}_k - \mathbf{u}^*\| \leq \frac{4\epsilon}{5}$. Finally, applying the triangle inequality and as $a + k \leq 1/m$:

$$\|\mathbf{u}_k - \mathbf{u}^*\| \leq \|\mathbf{u}_k - \bar{\mathbf{u}}_k\| + \|\bar{\mathbf{u}}_k - \mathbf{u}^*\| \leq \frac{\epsilon}{5} + \frac{4\epsilon}{5} = \epsilon.$$

Note that we have already bounded the total number of inner and outer loop iterations. Observing that each inner iteration makes 2 oracle queries to F and each outer iteration makes 2 oracle queries to F outside of the inner iteration, the bound on the total number of oracle queries to F follows. ■

Lemma 11 *Let $\bar{\mathbf{u}}_k^* = J_{F+I_{\mathcal{U}}}(\mathbf{u}_k)$, where $\mathbf{u}_k \in \mathcal{U}$ and F is L-Lipschitz. Then, there exists a parameter-free algorithm that queries F at most $O((L+1)\log(\frac{(L+1)\|\mathbf{u}_k - \bar{\mathbf{u}}_k^*\|}{\epsilon}))$ times and outputs a point $\bar{\mathbf{u}}_k$ such that $\|\bar{\mathbf{u}}_k - \bar{\mathbf{u}}_k^*\| \leq \epsilon$.*

Proof Observe first that $\bar{\mathbf{u}}_k^*$ solves (SVI) for operator $\bar{F}(\mathbf{u}) = F(\mathbf{u}) + \mathbf{u} - \mathbf{u}_k$ over the set $\mathcal{U}$. This follows from the definition of the resolvent, which implies:

$$\bar{\mathbf{u}}_k^* + F(\bar{\mathbf{u}}_k^*) + \partial I_{\mathcal{U}}(\bar{\mathbf{u}}_k^*) \ni \mathbf{u}_k.$$

Equivalently: $\mathbf{0} \in \bar{F}(\bar{\mathbf{u}}_k^*) + \partial I_{\mathcal{U}}(\bar{\mathbf{u}}_k^*)$.

The rest of the proof follows by applying Lemma 17 to $\bar{F}$, which is $(L+1)$-Lipschitz and 1-strongly monotone. ■

A.4. Inexact Halpern Iteration

We start by first proving the following auxiliary result.

Proposition 18 *Given an initial point $\mathbf{u}_0 \in \mathcal{U}$, let $\mathbf{u}_k$ evolve according to Eq. (9), where $\lambda_k = \frac{1}{k+1}$. Then,*

$$(\forall k \geq 1): \quad \|\mathbf{u}_k - \mathbf{u}^*\| \leq \|\mathbf{u}_0 - \mathbf{u}^*\| + \frac{1}{k+1}\sum_{i=1}^{k} i\|\mathbf{e}_{i-1}\|,$$

where $\mathbf{u}^*$ is such that $\|P(\mathbf{u}^*)\| = 0$.

Proof Let $T = \mathrm{Id} - P$. Then $T(\mathbf{u}^*) = \mathbf{u}^*$. By Fact 2, T is nonexpansive. Observe that we can equivalently write Eq. (9) as $\mathbf{u}_k = \lambda_k \mathbf{u}_0 + (1-\lambda_k)T(\mathbf{u}_{k-1}) + (1-\lambda_k)\mathbf{e}_{k-1}$. Thus, using that $\mathbf{u}^* = T(\mathbf{u}^*)$:

$$\begin{aligned}\|\mathbf{u}_k - \mathbf{u}^*\| &= \|\lambda_k(\mathbf{u}_0 - \mathbf{u}^*) + (1-\lambda_k)(T(\mathbf{u}_{k-1}) - T(\mathbf{u}^*)) + (1-\lambda_k)\mathbf{e}_{k-1}\| \\ &\leq \lambda_k\|\mathbf{u}_0 - \mathbf{u}^*\| + (1-\lambda_k)\|\mathbf{u}_{k-1} - \mathbf{u}^*\| + (1-\lambda_k)\|\mathbf{e}_{k-1}\|,\end{aligned}$$

where we have used the triangle inequality and nonexpansivity of T. The result follows by recursively applying the last inequality and observing that $\prod_{j=i}^{k}(1-\lambda_j) = \frac{i}{k+1}$. ■

Using this proposition, we can now prove the following lemma.

Lemma 12 *Let $\mathcal{C}_k$ be defined as in Eq. (7) with P as the $\frac{1}{2}$-cocoercive operator, and let $L_k = 2$, $\lambda_k = \frac{1}{k+1}$, and $A_k = \frac{k(k+1)}{2}$, $\forall k \geq 1$. If the iterates $\mathbf{u}_k$ evolve according to (9) for an arbitrary initial point $\mathbf{u}_0 \in \mathcal{U}$, then:*

$$(\forall k \geq 1): \quad A_{k+1}\mathcal{C}_{k+1} \leq A_k\mathcal{C}_k + A_{k+1}\left\langle \mathbf{e}_k, (1-\lambda_{k+1})P(\mathbf{u}_k) - P(\mathbf{u}_{k+1})\right\rangle.$$

Further, if, $\forall k \geq 1$, $\|\mathbf{e}_{k-1}\| \leq \frac{\epsilon}{4k(k+1)}$, then $\|P(\mathbf{u}_K)\| \leq \epsilon$ after at most $K = \frac{4\|\mathbf{u}_0 - \mathbf{u}^*\|}{\epsilon}$ iterations.

Proof By the same arguments as in the proof of Lemma 4 with P in place of F:

$$\frac{1}{2}\|P(\mathbf{u}_{k+1})\|^2 \leq \langle P(\mathbf{u}_{k+1}), \mathbf{u}_{k+1} - \mathbf{u}_k + P(\mathbf{u}_k)\rangle - \left\langle P(\mathbf{u}_k), \mathbf{u}_{k+1} - \mathbf{u}_k + \frac{1}{2}P(\mathbf{u}_k)\right\rangle.$$

From (9) and the definition of $\tilde{P}$, we have that

$$\begin{aligned}\mathbf{u}_{k+1} - \mathbf{u}_k &= \frac{\lambda_{k+1}}{1-\lambda_{k+1}}(\mathbf{u}_0 - \mathbf{u}_{k+1}) - P(\mathbf{u}_k) - \mathbf{e}_k, \text{and} \\ \mathbf{u}_{k+1} - \mathbf{u}_k &= \lambda_{k+1}(\mathbf{u}_0 - \mathbf{u}_k) - (1-\lambda_{k+1})P(\mathbf{u}_k) - (1-\lambda_{k+1})\mathbf{e}_k.\end{aligned}$$

Hence:

$$\frac{1}{2}\|P(\mathbf{u}_{k+1})\|^2 \leq \frac{\lambda_{k+1}}{1-\lambda_{k+1}}\langle P(\mathbf{u}_{k+1}), \mathbf{u}_0 - \mathbf{u}_{k+1}\rangle - \lambda_{k+1}\langle P(\mathbf{u}_k), \mathbf{u}_0 - \mathbf{u}_k\rangle + \frac{1-2\lambda_{k+1}}{2}\|P(\mathbf{u}_k)\|^2 + \langle \mathbf{e}_k, (1-\lambda_{k+1})P(\mathbf{u}_k) - P(\mathbf{u}_{k+1})\rangle .$$

Plugging $\lambda_{k+1} = \frac{1}{k+2}$ in the last inequality and using the definition of $\mathcal{C}_k$ and the choice of A_k from the statement of the lemma completes the proof of the first part.

Using the same arguments as in the proof of Lemma 5, we can conclude from $A_{k+1}\mathcal{C}_{k+1} \leq A_k\mathcal{C}_k + A_{k+1}\langle \mathbf{e}_k, (1-\lambda_{k+1})P(\mathbf{u}_k) - P(\mathbf{u}_{k+1})\rangle$, $\forall k \geq 1$ that:

$$\begin{aligned}\frac{\|P(\mathbf{u}_k)\|^2}{2} &\leq \frac{1}{k}\|P(\mathbf{u}_k)\|\|\mathbf{u}_0 - \mathbf{u}^*\| + \frac{1}{A_k}\sum_{i=1}^{k} A_i\langle \mathbf{e}_{i-1}, (1-\lambda_i)P(\mathbf{u}_{i-1}) - P(\mathbf{u}_i)\rangle \\ &= \frac{1}{k}\|P(\mathbf{u}_k)\|\|\mathbf{u}_0 - \mathbf{u}^*\| + \frac{1}{k(k+1)}\sum_{i=1}^{k} i(i+1)\left\langle \mathbf{e}_{i-1}, \frac{i}{i+1}P(\mathbf{u}_{i-1}) - P(\mathbf{u}_i)\right\rangle . \end{aligned} \tag{14}$$

Let us now bound each $\left\langle \mathbf{e}_{i-1}, \frac{i}{i+1}P(\mathbf{u}_{i-1}) - P(\mathbf{u}_i)\right\rangle$ term. Recall that $P(\mathbf{u}^*) = \mathbf{0}$ and P is 2-Lipschitz (as discussed in Section 1.2, this follows from P being $\frac{1}{2}$-cocoercive). Thus, we have:

$$\begin{aligned}\left\langle \mathbf{e}_{i-1}, \frac{i}{i+1}P(\mathbf{u}_{i-1}) - P(\mathbf{u}_i)\right\rangle &= \left\langle \mathbf{e}_{i-1}, \frac{i}{i+1}(P(\mathbf{u}_{i-1}) - P(\mathbf{u}^*)) - (P(\mathbf{u}_i) - P(\mathbf{u}^*))\right\rangle \\ &\leq 2\|\mathbf{e}_{i-1}\|\Big(\frac{i}{i+1}\|\mathbf{u}_{i-1} - \mathbf{u}^*\| + \|\mathbf{u}_i - \mathbf{u}^*\|\Big) \\ &\leq 2\|\mathbf{e}_{i-1}\|\Big(\frac{i+2}{i+1}\|\mathbf{u}_0 - \mathbf{u}^*\| + \frac{i}{i+1}\|\mathbf{e}_{i-1}\| + \frac{2}{i+1}\sum_{j=1}^{i-1} j\|\mathbf{e}_{j-1}\|\Big),\end{aligned}$$

where we have used Proposition 18 in the last inequality. In particular, if $\|\mathbf{e}_{i-1}\| \leq \frac{\epsilon}{4i(i+1)}$, then, $\forall i \geq 1$:

$$\left\langle \mathbf{e}_{i-1}, \frac{i}{i+1}P(\mathbf{u}_{i-1}) - P(\mathbf{u}_i)\right\rangle \leq \frac{\epsilon}{2i(i+1)}\Big(\frac{i+2}{i+1}\|\mathbf{u}_0 - \mathbf{u}^*\| + \epsilon/2\Big).$$

Combining with Eq. (14):

$$\frac{\|P(\mathbf{u}_k)\|^2}{2} \leq \frac{1}{k}\|P(\mathbf{u}_k)\|\|\mathbf{u}_0 - \mathbf{u}^*\| + \frac{\epsilon}{2k}(\|\mathbf{u}_0 - \mathbf{u}^*\| + \epsilon/2). \tag{15}$$

Observe that if $\|\mathbf{u}_0 - \mathbf{u}^*\| \leq \epsilon/2$, as P is 2-Lipschitz and $P(\mathbf{u}^*) = \mathbf{0}$, we would have $\|P(\mathbf{u}_0)\| \leq \epsilon$, and the statement of the second part of the lemma would hold trivially. Assume from now on that $\|\mathbf{u}_0 - \mathbf{u}^*\| > \epsilon/2$. Suppose that $\|P(\mathbf{u}_k)\| > \epsilon$ and $k \geq \frac{4\|\mathbf{u}_0 - \mathbf{u}^*\|}{\epsilon}$. Then, dividing both sides of Eq. (15) by $\|P(\mathbf{u}_k)\|/2$ and using that $\|P(\mathbf{u}_k)\| > \epsilon$ and $\|\mathbf{u}_0 - \mathbf{u}^*\| > \epsilon/2$, we get:

$$\|P(\mathbf{u}_k)\| < \frac{2\|\mathbf{u}_0 - \mathbf{u}^*\|(1+1/2)}{k} + \frac{2\cdot\epsilon/4}{k} < \frac{3\epsilon}{4} + \frac{\epsilon}{4} \leq \epsilon,$$

contradicting the assumption that $\|P(\mathbf{u}_k)\| > \epsilon$ and completing the proof. ■

A.5. Strongly Monotone Lipschitz Operators

Theorem 15 *Given F that is L-Lipschitz and m-strongly monotone, consider running the following algorithm $\mathcal{A}$, starting with $\mathbf{u}_0 \in \mathcal{U}$:*

$(\mathcal{A})$: *At iteration k, call Algorithm 3 with error $\epsilon_k = \frac{7}{16}\|\tilde{P}(\mathbf{u}_{k-1})\|$ and initial point $\mathbf{u}_{k-1}$.*

Then, $\mathcal{A}$ outputs $\mathbf{u}_k \in \mathcal{U}$ with $\|P(\mathbf{u}_k)\| \le \epsilon$ after at most $1 + \log_2(\frac{\|\mathbf{u}_0 - \mathbf{u}^*\|}{\epsilon})$ iterations, for any $\epsilon \in (0, \frac{1}{2}]$. The total number of queries to F until $\|P(\mathbf{u}_k)\| \le \epsilon$ is $O\big((L+\frac{L}{m})\log(\frac{\|\mathbf{u}_0 - \mathbf{u}^*\|}{\epsilon})\log(L+\frac{L}{m})\big)$.

Proof The first part of the theorem is immediate, as each call to Algorithm 3 ensures, due to Theorem 13, that

$$\|P(\mathbf{u}_k)\| \le \frac{7\|\tilde{P}(\mathbf{u}_{k-1})\|}{16} \le \frac{7\|P(\mathbf{u}_{k-1})\|}{16} + \frac{\epsilon_k}{8} \le \frac{\|P(\mathbf{u}_{k-1})\|}{2},$$

and $\|P(\mathbf{u}_0)\| \le 2\|\mathbf{u}_0 - \mathbf{u}^*\|$ as P is 2-Lipschitz (because it is $\frac{1}{2}$-cocoercive) and $P(\mathbf{u}^*) = \mathbf{0}$.

It remains to bound the number of calls to F for each call to Algorithm 3. Using Theorem 13 and $\|\tilde{P}(\mathbf{u}_k)\| = \Theta(\|P(\mathbf{u}_k)\|)$, each call to Algorithm 3 takes $O(\frac{L\|\mathbf{u}_{k-1} - \mathbf{u}^*\|}{\|P(\mathbf{u}_{k-1})\|}\log(\frac{L\|\mathbf{u}_{k-1} - \mathbf{u}^*\|}{\|P(\mathbf{u}_{k-1})\|}))$ calls to F. Denote $\bar{\mathbf{u}}^*_{k-1} = J_{F+\partial I_{\mathcal{U}}}(\mathbf{u}_{k-1}) = \mathbf{u}_{k-1} - P(\mathbf{u}_{k-1})$. Using Proposition 10:

$$\left\langle F(\bar{\mathbf{u}}^*_{k-1}), \bar{\mathbf{u}}^*_{k-1} - \mathbf{u}^* \right\rangle \le \|P(\mathbf{u}_{k-1})\|\|\bar{\mathbf{u}}^*_{k-1} - \mathbf{u}^*\|.$$

On the other hand, as F is m-strongly monotone and $\mathbf{u}^*$ is an (MVI) solution,

$$m\|\bar{\mathbf{u}}^*_{k-1} - \mathbf{u}^*\|^2 \le \left\langle F(\bar{\mathbf{u}}^*_{k-1}), \bar{\mathbf{u}}^*_{k-1} - \mathbf{u}^* \right\rangle.$$

Hence: $\|\bar{\mathbf{u}}^*_{k-1} - \mathbf{u}^*\| \le \frac{1}{m}\|P(\mathbf{u}_{k-1})\|$. It remains to use the triangle inequality and $P(\mathbf{u}_{k-1}) = \mathbf{u}_{k-1} - \bar{\mathbf{u}}^*_{k-1}$ to obtain:

$$\|\mathbf{u}_{k-1} - \mathbf{u}^*\| \le \left(1 + \frac{1}{m}\right)\|P(\mathbf{u}_{k-1})\|, \tag{16}$$

which completes the proof. ■

A.6. Lower Bounds

We make use of the lower bound from Ouyang and Xu (2019) and the algorithmic reductions between the problems considered in previous sections to derive (near-tight) lower bounds for all of the problems considered in this paper.

The lower bounds are for deterministic algorithms working in a (first-order) oracle model. For convex-concave saddle-point problems with the objective $\Phi(\mathbf{x}, \mathbf{y})$ and closed convex feasible set $\mathcal{X} \times \mathcal{Y}$, any such algorithm $\mathcal{A}$ can be described as follows: in each iteration k, $\mathcal{A}$ queries a pair of points $(\bar{\mathbf{x}}_k, \bar{\mathbf{y}}_k) \in \mathcal{X} \times \mathcal{Y}$ to obtain $(\nabla_{\mathbf{x}}\Phi(\bar{\mathbf{x}}_k, \bar{\mathbf{y}}_k),\ \nabla_{\mathbf{y}}\Phi(\bar{\mathbf{x}}_k, \bar{\mathbf{y}}_k))$, and outputs a candidate solution pair $(\mathbf{x}_k, \mathbf{y}_k) \in \mathcal{X} \times \mathcal{Y}$. Both the query points pair $(\bar{\mathbf{x}}_k, \bar{\mathbf{y}}_k)$ and the candidate solution pair $(\mathbf{x}_k, \mathbf{y}_k)$ can only depend on (i) global problem parameters (such as the Lipschitz constant of Φ's gradients or the feasible sets $\mathcal{X}, \mathcal{Y}$) and (ii) oracle queries and answers up to iteration k :

$$\{\bar{\mathbf{x}}_i,\ \bar{\mathbf{y}}_i,\ \nabla_{\mathbf{x}}\Phi(\bar{\mathbf{x}}_i, \bar{\mathbf{y}}_i),\ \nabla_{\mathbf{y}}\Phi(\bar{\mathbf{x}}_i, \bar{\mathbf{y}}_i)\}_{i=0}^{k-1}.$$

We start by summarizing the result from (Ouyang and Xu, 2019, Theorem 9).

Theorem 19 *For any deterministic algorithm working in the first-order oracle model described above and any $L, R_{\mathcal{X}}, R_{\mathcal{Y}} > 0$, there exists a problem instance with a convex-concave function $\Phi(\mathbf{x}, \mathbf{y}) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ whose gradients are L-Lipschitz, such that $\forall k = O(d)$:*

$$\max_{\mathbf{y} \in \mathbf{y}} \Phi(\mathbf{x}_k, \mathbf{y}) - \min_{\mathbf{x} \in \mathcal{X}} \Phi(\mathbf{x}, \mathbf{y}_k) = \Omega\Big(\frac{L({R_{\mathcal{X}}}^2 + R_{\mathcal{X}} R_{\mathcal{Y}})}{k}\Big),$$

where $(\mathbf{x}_k, \mathbf{y}_k) \in \mathcal{X} \times \mathcal{Y}$ is the algorithm output after k iterations and $R_{\mathcal{X}}$, $R_{\mathcal{Y}}$ denote the diameters of the feasible sets $\mathcal{X}$, $\mathcal{Y}$, respectively, and where both $\mathcal{X}$, $\mathcal{Y}$, are closed and convex.

The assumption of the theorem that $k = O(d)$ means that the lower bound applies in the high-dimensional regime $d = \Omega(\frac{L({R_{\mathcal{X}}}^2 + R_{\mathcal{X}} R_{\mathcal{Y}})}{\epsilon})$, which is standard and generally unavoidable.

In the setting of VIs, we consider a related model in which an algorithm has oracle access to F and refer to it as the operator oracle model. Similarly as for the saddle-point problems, we consider deterministic algorithms that on a given problem instance described by $(F, \mathcal{U})$ operate as follows: in each iteration k the algorithm queries a point $\bar{\mathbf{u}}_k \in \mathcal{U}$, receives $F(\bar{\mathbf{u}}_k)$, and outputs a solution candidate $\mathbf{u}_k \in \mathcal{U}$. Both $\mathbf{u}_k$ and $\bar{\mathbf{u}}_k$ can only depend on (i) global problem parameters (such as the feasible set $\mathcal{U}$ and the Lipschitz parameter of F), and (ii) oracle queries and answers up to iteration $k : \{\bar{\mathbf{u}}_i, F(\bar{\mathbf{u}}_i)\}_{i=0}^{k-1}$. Note that all methods described in this paper and most of the commonly used methods for solving VIs, such as, e.g., the mirror-prox method of Nemirovski (2004) and dual extrapolation method of Nesterov (2007), work in this oracle model.

Lemma 16 *For any deterministic algorithm working in the operator oracle model described above and any $L, D > 0$, there exists a VI described by an L-Lipschitz-continuous operator F and a closed convex feasible set $\mathcal{U}$ with diameter D such that:*

(a) For all $\epsilon > 0$ such that $k = \frac{LD^2}{\epsilon} = O(d)$, $\max_{\mathbf{u} \in \mathcal{U}} \langle F(\mathbf{u}_k), \mathbf{u}_k - \mathbf{u} \rangle = \Omega(\epsilon)$;

(b) For all $\epsilon > 0$ such that $k = \frac{LD}{\epsilon} = O(d)$, $\max_{\mathbf{u} \in \{\mathcal{U} \cap \mathcal{B}_{\mathbf{u}_k}\}} \langle F(\mathbf{u}_k), \mathbf{u}_k - \mathbf{u} \rangle = \Omega(\epsilon)$;

(c) If F is $\frac{1}{L}$-cocoercive, then for all $\epsilon > 0$ such that $k = \frac{LD}{\epsilon \log(D/\epsilon)} = O(d)$, it holds that

$$\max_{\mathbf{u} \in \{\mathcal{U} \cap \mathcal{B}_{\mathbf{u}_k}\}} \langle F(\mathbf{u}_k), \mathbf{u}_k - \mathbf{u} \rangle = \Omega(\epsilon)$$

(d) If F is m-strongly monotone, then for all $\epsilon > 0$ such that $k = \frac{L}{m} = O(d)$, it holds that

$$\max_{\mathbf{u} \in \{\mathcal{U} \cap \mathcal{B}_{\mathbf{u}_k}\}} \langle F(\mathbf{u}_k), \mathbf{u}_k - \mathbf{u} \rangle = \Omega(\epsilon).$$

Proof

Proof of (a): Suppose that this claim was not true. Then we would be able to solve any instance with L-Lipschitz F and $\mathcal{U}$ with diameter bounded by D and obtain $\mathbf{u}_k$ with $\max_{\mathbf{u} \in \mathcal{U}} \langle F(\mathbf{u}_k), \mathbf{u}_k - \mathbf{u} \rangle \leq \epsilon$ in $o(\frac{LD^2}{\epsilon})$ iterations, assuming the appropriate high-dimensional regime. In particular, given any fixed convex-concave $\Phi(\mathbf{x}, \mathbf{y})$ with L-Lipschitz gradients and feasible sets $\mathcal{X}, \mathcal{Y}$ whose diameter is $D/2$, let $\mathbf{u} = \left[\begin{smallmatrix}\mathbf{x}\\ \mathbf{y}\end{smallmatrix}\right]$, $F(\mathbf{u}) = \left[\begin{smallmatrix}\nabla_{\mathbf{x}} \Phi(\mathbf{x}, \mathbf{y})\\ -\nabla_{\mathbf{y}} \Phi(\mathbf{x}, \mathbf{y})\end{smallmatrix}\right]$, $\mathcal{U} = \mathcal{X} \times \mathcal{Y}$. Then, it is not hard to verify that F is monotone and L-Lipschitz (see, e.g., Nemirovski (2004); Facchinei and Pang (2003)) and the diameter of $\mathcal{U}$ is D. Thus, by assumption, we would be able to construct a point $\mathbf{u}_k = \left[\begin{smallmatrix}\mathbf{x}_k\\ \mathbf{y}_k\end{smallmatrix}\right]$ for which

$\max_{\mathbf{u}\in\mathcal{U}} \langle F(\mathbf{u}_k), \mathbf{u}_k - \mathbf{u}\rangle \leq \epsilon$ in $o(\frac{LD^2}{\epsilon})$ iterations. But then, because Φ is convex-concave, we would also have, for any $\mathbf{x}\in\mathcal{X}, \mathbf{y}\in\mathcal{Y}$:

$$\begin{aligned}\Phi(\mathbf{x}_k, \mathbf{y}) - \Phi(\mathbf{x}, \mathbf{y}_k) &\leq \max_{\mathbf{y}\in\mathcal{Y}} \Phi(\mathbf{x}_k, \mathbf{y}) - \Phi(\mathbf{x}_k, \mathbf{y}_k) + \Phi(\mathbf{x}_k, \mathbf{y}_k) - \min_{\mathbf{x}\in\mathcal{X}} \Phi(\mathbf{x}, \mathbf{y}_k)\\ &\leq \langle \nabla_{\mathbf{y}}\Phi(\mathbf{x}_k, \mathbf{y}_k), \mathbf{y} - \mathbf{y}_k\rangle + \langle \nabla_{\mathbf{x}}\Phi(\mathbf{x}_k, \mathbf{y}_k), \mathbf{x}_k - \mathbf{x}\rangle = \langle F(\mathbf{u}_k), \mathbf{u}_k - \mathbf{u}\rangle.\end{aligned}$$

In particular, we would get:

$$\max_{\mathbf{y}\in\mathcal{Y}} \Phi(\mathbf{x}_k, \mathbf{y}) - \min_{\mathbf{x}\in\mathcal{X}} \Phi(\mathbf{x}, \mathbf{y}_k) \leq \max_{\mathbf{u}\in\mathcal{U}} \langle F(\mathbf{u}_k), \mathbf{u}_k - \mathbf{u}\rangle \leq \epsilon.$$

Because we obtained this bound for an arbitrary L-Lipschitz convex-concave Φ and arbitrary feasible sets $\mathcal{X}, \mathcal{Y}$ with diameters $D/2$, Theorem 19 leads to a contradiction.

Proof of (b): If (b) was not true, then we would be able to obtain a point $\mathbf{u}_k$ with

$$\max_{\mathbf{u}\in\{\mathcal{U}\cap\mathcal{B}_{\mathbf{u}_k}\}} \langle F(\mathbf{u}_k), \mathbf{u}_k - \mathbf{u}\rangle = o(\epsilon/D)$$

in $k = \frac{LD^2}{\epsilon}$ iterations. But the same point would satisfy $\max_{\mathbf{u}\in\mathcal{U}} \langle F(\mathbf{u}_k), \mathbf{u}_k - \mathbf{u}\rangle = o(\epsilon)$, which is a contradiction, due to (a).

Proof of (c): We prove the claim for $L = 2$. This is w.l.o.g., due to the standard rescaling argument: if F is $\frac{1}{L}$-cocoercive, then $\bar{F} = F/(2L)$ is $\frac{1}{2}$-cocoercive. Further, if, for some $\mathbf{u}_k \in \mathcal{U}$,

$$\max_{\mathbf{u}\in\{\mathcal{U}\cap\mathcal{B}_{\mathbf{u}_k}\}} \left\langle \bar{F}(\mathbf{u}_k), \mathbf{u}_k - \mathbf{u}\right\rangle = \Omega(\epsilon),$$

then $\max_{\mathbf{u}\in\{\mathcal{U}\cap\mathcal{B}_{\mathbf{u}_k}\}} \langle F(\mathbf{u}_k), \mathbf{u}_k - \mathbf{u}\rangle = \Omega(L\epsilon)$.

Suppose that the claim was not true for a $\frac{1}{2}$-cocoercive operator F. Then for any M-Lipschitz monotone operator G, we would be able to use the strategy from Section 2.2 to obtain a point $\mathbf{u}_k$ with

$$\max_{\mathbf{u}\in\{\mathcal{U}\cap\mathcal{B}_{\mathbf{u}_k}\}} \langle G(\mathbf{u}_k), \mathbf{u}_k - \mathbf{u}\rangle = o(\epsilon)$$

in $k = \frac{MD}{\epsilon}$ iterations. This is a contradiction, due to (b).

Proof of (d): Suppose that the claim was not true, i.e., that there existed an algorithm that, for any $m, L > 0$, could output $\mathbf{u}_k$ with $\max_{\mathbf{u}\in\{\mathcal{U}\cap\mathcal{B}_{\mathbf{u}_k}\}} \left\langle \bar{F}(\mathbf{u}_k), \mathbf{u}_k - \mathbf{u}\right\rangle = \epsilon/2$ in $k = o(L/m)$ iterations, for any m-strongly monotone and L-Lipschitz operator. Then for any L-Lipschitz monotone operator F, we could apply that algorithm to $\bar{F}(\cdot) = F(\cdot) + \frac{\epsilon}{2D}(\cdot - \mathbf{u}_0)$ to obtain a point $\mathbf{u}_k$ with $\max_{\mathbf{u}\in\{\mathcal{U}\cap\mathcal{B}_{\mathbf{u}_k}\}} \left\langle \bar{F}(\mathbf{u}_k), \mathbf{u}_k - \mathbf{u}\right\rangle = \epsilon/2$ in $k = o(LD/\epsilon)$ iterations. But then we would also have:

$$\max_{\mathbf{u}\in\{\mathcal{U}\cap\mathcal{B}_{\mathbf{u}_k}\}} \langle F(\mathbf{u}_k), \mathbf{u}_k - \mathbf{u}\rangle = \max_{\mathbf{u}\in\{\mathcal{U}\cap\mathcal{B}_{\mathbf{u}_k}\}} \left\langle \bar{F}(\mathbf{u}_k) - \frac{\epsilon}{2D}(\mathbf{u}_k - \mathbf{u}_0), \mathbf{u}_k - \mathbf{u}\right\rangle \leq \epsilon,$$

which is a contradiction, due to (b). ■