Extending the Reach of First-Order Algorithms for Nonconvex Min-Max Problems with Cohypomonotonicity

Ahmet Alacaoglu* Donghwan Kim† Stephen J. Wright‡

Abstract

We focus on constrained, L-smooth, nonconvex-nonconcave min-max problems either satisfying ρ-cohypomonotonicity or admitting a solution to the ρ-weakly Minty Variational Inequality (MVI), where larger values of the parameter $\rho > 0$ correspond to a greater degree of nonconvexity. These problem classes include examples in two player reinforcement learning, interaction dominant min-max problems, and certain synthetic test problems on which classical min-max algorithms fail. It has been conjectured that first-order methods can tolerate value of ρ no larger than $\frac{1}{L}$, but existing results in the literature have stagnated at the tighter requirement $\rho < \frac{1}{2L}$. With a simple argument, we obtain optimal or best-known complexity guarantees with cohypomonotonicity or weak MVI conditions for $\rho < \frac{1}{L}$. The algorithms we analyze are inexact variants of Halpern and Krasnosel'skiĭ-Mann (KM) iterations. We also provide algorithms and complexity guarantees in the stochastic case with the same range on ρ. Our main insight for the improvements in the convergence analyses is to harness the recently proposed *conic nonexpansiveness* property of operators. As byproducts, we provide a refined analysis for inexact Halpern iteration and propose a stochastic KM iteration with a multilevel Monte Carlo estimator.

1 Introduction

We consider the problem

$$\min_{u\in U}\max_{v\in V} f(u,v), \tag{1.1}$$

where $U \subseteq \mathbb{R}^m, V \subseteq \mathbb{R}^n$ are closed convex sets admitting efficient projection operators and $f\colon \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$ is a function such that $\nabla_u f(u,v)$ and $\nabla_v f(u,v)$ are Lipschitz continuous. The general setting where $f(u,v)$ is allowed to be nonconvex-nonconcave is *extremely* relevant in machine learning (ML), with applications in generative adversarial networks (GANs) [Goodfellow et al., 2014] and adversarial ML [Madry et al., 2018]. Yet, at the same time, such problems are *extremely* challenging to solve, with documented hardness results, see e.g., Daskalakis et al. [2021]. As a result, an extensive literature has arisen about special cases of the nonconvex-nonconcave problem (1.1) for which algorithms with good convergence and complexity properties can be derived [Diakonikolas et al., 2021, Bauschke et al., 2021, Lee and Kim, 2021, Pethick et al., 2022, 2023a,b, Gorbunov et al., 2023, Böhm, 2022, Cai et al., 2022, Cai and Zheng, 2022, Hajizadeh et al., 2023, Kohlenbach, 2022, Anonymous, 2024a,b, Grimmer et al., 2023, Tran-Dinh and Luo, 2023].

To describe these special cases of (1.1), we state the following *nonmonotone* inclusion problem, which generalizes (1.1):

$$\text{Find } x^\star \in \mathbb{R}^d \text{ such that } 0 \in F(x^\star) + G(x^\star), \tag{1.2}$$

where $F\colon \mathbb{R}^d \to \mathbb{R}^d$ is L-Lipschitz and $G\colon \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is maximally monotone. Mapping this problem to finding stationary points of (1.1) is standard by setting $x = \binom{u}{v}$, $F(x) = \binom{\nabla_u f(u,v)}{-\nabla_v f(u,v)}$ and $G(x) = \binom{\partial \iota_U}{\partial \iota_V}$, where ι_U is the indicator function for set U. The nonmonotonicity in problem (1.2) is due to nonconvex-nonconcavity of problem (1.1).

*Wisconsin Institute for Discovery, University of Wisconsin–Madison. alacaoglu@wisc.edu
†Department of Mathematical Sciences, KAIST. donghwankim@kaist.ac.kr
‡Department of Computer Sciences, University of Wisconsin–Madison. swright@cs.wisc.edu

The main additional assumption we make is that $F+G$ is *ρ-cohypomonotone*. Recalling the standard definition $\operatorname{gra}(F+G) = \{(x,u) \in \mathbb{R}^d \times \mathbb{R}^d \colon\ u \in (F+G)(x)\}$, ρ-cohypomonotonicity is defined as

$$\begin{aligned} &\langle u-v, x-y\rangle \geq -\rho\|u-v\|^2 \\ &\forall (x,u) \in \operatorname{gra}(F+G) \text{ and } \forall (y,v) \in \operatorname{gra}(F+G), \end{aligned} \tag{1.3}$$

for $\rho > 0$, see [Bauschke et al., 2021, Def. 2.4]. When (1.3) holds only for $y = x^\star$, it is also called the *weak MVI condition* or *ρ-star-cohypomonotonicity*, due to [Diakonikolas et al., 2021]. For $\rho > 0$, the weak MVI condition requires the existence of a solution $x^\star$ to the ρ-weakly MVI:

$$\langle u, x - x^\star\rangle \geq -\rho\|u\|^2 \quad \forall (x,u) \in \operatorname{gra}(F+G). \tag{1.4}$$

For standard *monotone operators* (or convex-concave instances of (1.1)), the inner product in (1.3) is lower bounded by 0. The assumption (1.3) allows the right-hand side to be negative, allowing *nonmonotonicity of $F+G$* or *nonconvex-nonconcavity of $f(u,v)$*, while the limit of nonmonotonicity is determined by $\rho > 0$. These two assumptions, cohypomonotonicity or weak MVI, are required in the extensive literature cited at the end of the first paragraph.

In this paper, we extend the range of ρ, doubling the upper limit of $\frac{1}{2L}$ considered in the previous works, thus allowing a wider range of nonconvex problems of the form (1.1) to be solved by first-order algorithms, while ensuring optimal or best-known complexity guarantees.

Motivation. Cohypomonotonicity and weak MVI conditions, defined in (1.3) and (1.4), allowed progress to be made in understanding the behavior of first-order algorithms for structured nonconvex-nonconcave problems, in a wide variety of works cited at the end of first paragraph. On the one hand, these assumptions are not as general as one might desire: They have not been shown to hold for problems arising in generative or adversarial ML. On the other hand, they have been proven to hold for other relevant problems in ML.

Examples where cohypomonotonicity holds include the *interaction dominant min-max problems* (Example 1.2) and some stylized worst-case nonconvex-nonconcave instances [Hsieh et al., 2021, Pethick et al., 2023b] (see also [Bauschke et al., 2021, Sections 5, 6]). The relaxed assumption of having a weak MVI solution is implied by star (and quasi-strong) monotonicity [Loizou et al., 2021] or existence of a solution to MVI [Dang and Lan, 2015], the latter being relevant in the context of policy gradient algorithms for reinforcement learning (RL) [Lan, 2023]. Weak MVI condition is satisfied for the following RL problem.

Example 1.1. *von Neumann's ratio game:* This is a simple two player stochastic game [Neumann, 1945, Daskalakis et al., 2020, Diakonikolas et al., 2021]. Using the standard definition of the simplex $\Delta^d = \{x \in \mathbb{R}^d \colon , x \geq 0, \sum_{i=1}^d x_i = 1\}$, the problem is

$$\min_{x\in\Delta^m} \max_{y\in\Delta^n} \frac{\langle x, Ry\rangle}{\langle x, Sy\rangle},$$

where $R \in \mathbb{R}^{m\times n}$, $S \in \mathbb{R}_+^{m\times n}$ and $\langle x, Sy\rangle > 0\ \forall (x,y) \in \Delta^m \times \Delta^n$. As described in Diakonikolas et al. [2021], it is easy to construct instances of this problem where it satisfies ρ-weakly MVI condition, but not cohypomonotonicity. ⧫

Example 1.2. *Interaction dominant min-max problems [Grimmer et al., 2023]:* We say that f in (1.1) is $\alpha(r)$-interaction dominant if it satisfies for all $z = \binom{u}{v} \in \mathbb{R}^{n+m}$ that

$$\begin{aligned} \nabla^2_{xx} f(z) + \nabla^2_{xy} f(z)(r^{-1}\mathrm{Id} - \nabla^2_{yy} f(z))^{-1}\nabla^2_{yx} f(z) &\succeq \alpha(r)\mathrm{Id}, \\ -\nabla^2_{yy} f(z) + \nabla^2_{yx} f(z)(r^{-1}\mathrm{Id} + \nabla^2_{xx} f(z))^{-1}\nabla^2_{xy} f(z) &\succeq \alpha(r)\mathrm{Id}. \end{aligned}$$

Interaction is captured by the second terms on the left-hand side of each condition. The problem is called (nonnegative) *interaction dominant* if these terms *dominate* the smallest eigenvalue of $\nabla^2_{xx} f$ and largest eigenvalue of $\nabla^2_{yy} f$, i.e., $\alpha(r) \geq 0$. This is equivalent to the r-cohypomonotonicity of F [Hajizadeh et al., 2023, Prop. 1]. ⧫

The limit for the parameter ρ in (1.3) and (1.4) for which convergence can be proved in most algorithms seems to have stagnated at $\rho < \frac{1}{2L}$. Two exceptions exist for a special case of our setting when $G \equiv 0$, which

Assumption	Reference	Upper bound of ρ	Constraints	Oracle complexity
cohypomonotone	Cai and Zheng [2022]	$\frac{1}{60L}$	✓	$O(\varepsilon^{-1})$
	Cai et al. [2022], Pethick et al. [2023b] Lee and Kim [2021], Tran-Dinh [2023] Gorbunov et al. [2023]	$\frac{1}{2L}$	✓	$O(\varepsilon^{-1})$
	Cai et al. [2024]	$\frac{0.7}{L}$	×	$\widetilde{O}(\varepsilon^{-1})$
	Theorem 2.1	$\frac{1}{L}$	✓	$\widetilde{O}(\varepsilon^{-1})$
weak MVI	Diakonikolas et al. [2021]‡	$\frac{1}{8L}$	×	$O(\varepsilon^{-2})$
	Böhm [2022]‡	$\frac{1}{2L}$	×	$O(\varepsilon^{-2})$
	Cai and Zheng [2022]	$\frac{1}{12\sqrt{3}L}$	✓	$O(\varepsilon^{-2})$
	Anonymous [2024a]‡	$\frac{1}{3L}$	✓	$\widetilde{O}(\varepsilon^{-2})$
	Pethick et al. [2022]	$\frac{1}{2L}$	✓	$O(\varepsilon^{-2})$
	Anonymous [2024b]	$\frac{0.63}{L}$	×	$O(\varepsilon^{-2})$
	Theorem 3.1	$\frac{1}{L}$	✓	$\widetilde{O}(\varepsilon^{-2})$

Table 1: Comparison of first-order algorithms for deterministic problems. Complexity refers to the number of oracle calls to get $\operatorname{dist}(0,(F+G)(x)) \le \varepsilon$. See also Remark 2.3. ‡These works defined weak MVI as $\langle F(x), x - x^\star\rangle \ge -\frac{\gamma}{2}\|F(x)\|^2$, i.e., $\gamma = 2\rho$.

corresponds in view of (1.1) to an unconstrained problem. First is the recent work [Anonymous, 2024b] that claimed to improve the limit of ρ for weak MVI to $\approx \frac{0.63}{L}$ with a rather complicated analysis. The rate obtained is also suboptimal under cohypomonotonicity. This work conjectured (but did not prove) $\frac{1}{L}$ as the maximum limit for ρ and also did not provide any algorithm achieving this. For an unconstrained cohypomonotone problem, [Cai et al., 2024, Corollary 4.5] also showed possibility of obtaining guarantees with $\rho < \frac{1}{\sqrt{2}L} \approx \frac{0.7}{L}$. Relevant citations and discussions appear in Table 1 and Appendix D.

First-order oracles. As standard in the operator splitting literature (see e.g., Bauschke and Combettes [2017]), a *first-order oracle call* for (1.2) consists of one evaluation of F and one *resolvent* of G (see (1.5)). In the context of the min-max problem (1.1), this requires computation of gradients $\nabla_u f(u,v), \nabla_v f(u,v)$ together with projections on sets U, V. (All works in Table 1 have the same oracle access.)

Contributions. We show how to increase the range of the cohypomonotonicity parameter to $\rho < \frac{1}{L}$ while maintaining first-order oracle complexity $\widetilde{O}(\varepsilon^{-1})$ for finding a point x such that $\operatorname{dist}(0,(F+G)(x)) \le \varepsilon$. Such a complexity is optimal (up to a log factor) even for monotone problems [Yoon and Ryu, 2021, Section 3]. With weak MVI and the improved range of $\rho < \frac{1}{L}$, we show complexity $\widetilde{O}(\varepsilon^{-2})$ for $\operatorname{dist}(0,(F+G)(x)) \le \varepsilon$ which is the best-known (up to a log factor) under this assumption. Table 1 summarizes known results on complexity and the upper bound of ρ.

Thanks to the modularity of our approach, we show some corollaries. First, we provide complexity guarantees for stochastic versions of our problems where the operator F is accessed via unbiased oracles $F(\cdot,\xi)$ (that is, $\mathbb{E}_\xi[F(x,\xi)] = F(x)$). We also discuss how to improve the best-known *ρ-independent* and *ρ-agnostic* complexity bounds. On a technical side, we tighten the analysis of Halpern iteration with inexact resolvent computations, an ingredient that is critical for the stochastic extension. Similarly, to obtain the best-known complexity for stochastic problems under weak MVI, we incorporate the multilevel Monte Carlo estimator to KM iteration to control the bias in subproblem solutions.

1.1 Preliminaries

Notation. We denote the ℓ_2 norm as $\|\cdot\|$. Given $G\colon \mathbb{R}^d \rightrightarrows \mathbb{R}^d$, we use standard definitions $\operatorname{gra} G = \{(x,u) \in \mathbb{R}^d \times \mathbb{R}^d \colon u \in G(x)\}$ and $\operatorname{dist}(0, G(x)) = \min_{u\in G(x)} \|u\|$. Domain of an operator is defined as $\operatorname{dom} G = \{x \in \mathbb{R}^d \colon G(x) \neq \emptyset\}$. The operator G is *maximally* monotone (resp. cohypomonotone or

hypomonotone) if its graph is not strictly contained in the graph of any other monotone (resp. cohypomonotone or hypomonotone) operator.

An operator $F\colon \mathbb{R}^d \rightrightarrows \mathbb{R}^d$, given $(x,u) \in \operatorname{gra} F$ and $(y,v) \in \operatorname{gra} F$, is **(i)** γ-*strongly monotone* if $\langle u-v, x-y\rangle \geq \gamma\|x-y\|^2$ with $\gamma > 0$ and *monotone* if the inequality holds with $\gamma = 0$; **(ii)** ρ-*hypomonotone* if $\langle u-v, x-y\rangle \geq -\rho\|x-y\|^2$ with $\rho > 0$. An operator $F\colon \mathbb{R}^d \to \mathbb{R}^d$ is **(iii)** L-*Lipschitz* if $\|F(x)-F(y)\| \leq L\|x-y\|$; **(iv)** *nonexpansive* if F is 1-Lipschitz; **(v)** γ-*cocoercive* if $\langle F(x)-F(y), x-y\rangle \geq \gamma\|F(x)-F(y)\|^2$ with $\gamma > 0$. We refer to *star* variants of these properties (e.g., *star-cocoercive*) when they are required only at $(y,v) = (x^\star, 0)$ where $0 \in F(x^\star)$.

The *resolvent* of an operator $F\colon \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is defined as

$$J_F = (\mathrm{Id} + F)^{-1}. \tag{1.5}$$

The resolvent generalizes the well-known *proximal operator* that has been ubiquitous in optimization and ML, where F is typically the subdifferential of a regularizer function, e.g., ℓ_1 norm. Favorable properties of the resolvent are well-known when F is monotone [Bauschke and Combettes, 2017]. Meanwhile, in our nonmonotone case, immense care must be taken in utilizing this object, as it might even be undefined. A comprehensive reference for the properties of resolvent of a nonmonotone operator is [Bauschke et al., 2021]. We review and explain the results relevant to our work in the sequel.

The algorithms we analyze are based on the classical Halpern [Halpern, 1967] and Krasnosel'skiĭ-Mann (KM) [Krasnosel'skii, 1955, Mann, 1953] iterations. Given an operator $T\colon \mathbb{R}^d \to \mathbb{R}^d$, Halpern iteration is defined as

$$x_{k+1} = \beta_k x_0 + (1-\beta_k)T(x_k), \tag{1.6}$$

for a decreasing sequence $\{\beta_k\} \in (0,1)$ and initial point x_0. The KM iteration, with a fixed $\beta \in (0,1)$, is defined as

$$x_{k+1} = \beta x_k + (1-\beta)T(x_k). \tag{1.7}$$

Conic nonexpansiveness. The key to relaxing the range of ρ parameter for both assumptions is to harness the algorithmic consequences of *conic nonexpansiveness*, the notion introduced by the influential work of Bauschke et al. [2021] that also inspired our developments. We say that $T\colon \mathbb{R}^d \to \mathbb{R}^d$ is α-conically nonexpansive with $\alpha > 0$ when there exists a nonexpansive operator $N\colon \mathbb{R}^d \to \mathbb{R}^d$ such that $T = (1-\alpha)\mathrm{Id} + \alpha N$, see [Bauschke et al., 2021, Def. 3.1]. This equivalently means that a particular combination of Id and T is nonexpansive: $\|((1-\alpha^{-1})\mathrm{Id} + \alpha^{-1}T)(x-y)\| \leq \|x-y\|^2$. An important characterization of this property given in [Bauschke et al., 2021, Cor. 3.5(iii)] is that T is α-conically nonexpansive if and only if $\mathrm{Id} - T$ is $\frac{1}{2\alpha}$-cocoercive. We also consider the *star* variants (in the sense defined in the Notation paragraph) of these properties and characterizations, which are detailed in Appendix B.1.1.

Assumption 1. *The operator $F\colon \mathbb{R}^d \to \mathbb{R}^d$ is L-Lipschitz and $G\colon \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is maximally monotone. The solution set for the problem (1.2) is nonempty.*

Assumption 1 is standard, see Facchinei and Pang [2003], and is required throughout the text. Monotonicity is not assumed for F. Lipschitzness of F corresponds to smoothness of f in context of (1.1) and maximal monotonicity of G is satisfied when we have constraint sets given in (1.1) but also when we have convex regularizers that can be added on (1.1) (e.g., $\|\cdot\|_1$).

Assumption 2. *The operator $F+G$ is maximally ρ-cohypomonotone (see (1.3) for the definition).*

Assumption 2 is abundant in the recent literature for nonconvex-nonconcave optimization [Lee and Kim, 2021, Bauschke et al., 2021, Cai et al., 2022, Cai and Zheng, 2022, Gorbunov et al., 2023, Pethick et al., 2023b]. An instance is provided in Example 1.2 with further pointers to related problems are given in Section 1. Assumption 2 is required only for the results in Sections 2 and 4.1.

Assumption 3. *There exists a ρ-weakly MVI solution to the problem (1.2) (see (1.4) for the definition).*

Assumption 3 is weaker than Assumption 2 as it is only required on the ray to a solution, see also Example 1.1. Assumption 3, used in Sections 3 and 4.2, is also widespread in the recent literature for nonconvex-nonconcave optimization [Diakonikolas et al., 2021, Pethick et al., 2022, 2023a, Cai et al., 2022, Anonymous, 2024a,b, Böhm, 2022].

2 Algorithm and Analysis under Cohypomonotonicity

Algorithm 1 Inexact Halpern iteration for problems with cohypomonotonicity

Input: Parameters $\beta_k = \frac{1}{k+2}, \eta, L, \rho$, $\alpha = 1 - \frac{\rho}{\eta}$, $K \geq 1$, initial iterate $x_0 \in \mathbb{R}^d$, subroutine FBF given in Algorithm 2

for $k = 0, 1, 2, \ldots, K-1$ **do**

$\widetilde{J}_{\eta(F+G)}(x_k) = \texttt{FBF}\left(x_k, T, G, \mathrm{Id} + \eta F - x_k, 1 + \eta L\right)$ where $T = \left\lceil \frac{4(1+\eta L)}{1-\eta L} \log(98\sqrt{k+2}\log(k+2)) \right\rceil$

$x_{k+1} = \beta_k x_0 + (1-\beta_k)((1-\alpha)x_k + \alpha \widetilde{J}_{\eta(F+G)}(x_k))$

end for

Algorithm 2 $\texttt{FBF}(z_0, T, A, B, L_B)$ from [Tseng, 2000]

Input: Parameter $\tau = \frac{1}{2L_B}$, initial iterate $z_0 \in \mathbb{R}^d$

for $t = 0, 1, 2, \ldots, T-1$ **do**

$z_{t+1/2} = J_{\tau A}(z_t - \tau B(z_t))$

$z_{t+1} = z_{t+1/2} + \tau B(z_t) - \tau B(z_{t+1/2})$

end for

2.1 Algorithm Construction and Analysis Ideas

Recall the definitions of resolvent (1.5) and cohypomonotonicity (1.3). We sketch the algorithmic construction and analysis ideas which will be expanded on in Section 2.2.

(I) We know that Halpern iteration in (1.6) with $\beta_k = \frac{1}{k+2}$ has optimal rate when $\mathrm{Id} - T$ is cocoercive, see [Sabach and Shtern, 2017, Lieder, 2021, Kim, 2021]. That is, one gets $\eta^{-1}\|x_k - J_{\eta(F+G)}(x_k)\| \leq \varepsilon$ with $O(\varepsilon^{-1})$ evaluations of $J_{\eta(F+G)}$.

(II) When $F+G$ is maximally ρ-cohypomonotone (per Assumption 2), we know from Bauschke et al. [2021] (with precise pointers in App. A.1) that $J_{\eta(F+G)}$ is $\frac{1}{2\alpha}$-conically nonexpansive where $\alpha = 1 - \frac{\rho}{\eta}$, its domain is $\mathbb{R}^d$ and it is single-valued when $\frac{\rho}{\eta} < 1$. Consequently, $\mathrm{Id} - J_{\eta(F+G)}$ is α-cocoercive. Then, one can use the result in (I).

We next see a high level discussion on the approximate computation of $J_{\eta(F+G)}$.

(III) Since F is L-Lipschitz, we have that F is L-hypomonotone by Cauchy-Schwarz inequality, i.e.,

$$\langle F(x) - F(y), x - y\rangle \geq -L\|x-y\|^2.$$

Hence, $\mathrm{Id} + \eta F$ is $(1-\eta L)$-strongly monotone.

By definition, we have $x_k^\star = J_{\eta(F+G)}(x_k) = (\mathrm{Id} + \eta(F+G))^{-1}(x_k)$. Existence and uniqueness of $x_k^\star$ is guaranteed by (II) when $\rho < \eta$. By definition, $x_k^\star$ is the solution of the problem

$$0 \in (\mathrm{Id} + \eta(F+G))(x_k^\star) - x_k.$$

Hence, computation of the resolvent is a strongly monotone inclusion problem where $\mathrm{Id} + \eta F$ is $(1-\eta L)$-strongly monotone and $(\eta L + 1)$-Lipschitz, and G is maximally monotone. In view of (1.1) this also corresponds to a strongly convex-strongly concave problem, also known as the proximal operator of f with a center point x_k over constraint sets U, V.

(IV) Any optimal variational inequality (or monotone inclusion) algorithm, such as the forward-backward-forward (FBF) [Tseng, 2000], gives $\hat{x}_k$ with $\|\hat{x}_k - J_{\eta(F+G)}(x_k)\|^2 \leq \varepsilon_k^2$ with complexity $\widetilde{O}\left(\frac{1+\eta L}{1-\eta L}\right)$.

In summary, our requirements are $\frac{\rho}{\eta} < 1$ for ensuring well-definedness of the resolvent, as per (II), and $1 - \eta L > 0$ for ensuring strong monotonicity for efficient approximation of the resolvent, as per (III). Hence, we need $\rho < \eta < \frac{1}{L}$, leading to the claimed improved range on ρ.

Item (II) refers to the resolvent of $\eta(F+G)$, which cannot be evaluated exactly in general with standard first-order oracles. We approximate $J_{\eta(F+G)}$, which leads to the inexact Halpern iteration, similar to Diakonikolas et al. [2021], Cai et al. [2024]. Note that in the context of problem (1.1), approximating the resolvent corresponds to computing approximation of *proximal operator* for function f which is a strongly convex-strongly concave min-max problem.

In the next section, by extending the arguments in [Diakonikolas, 2020, Lemma 12] and [Cai et al., 2024, Lemma C.3] to accommodate conic nonexpansiveness, we show that $\eta^{-1}\|x_k - J_{\eta(F+G)}(x_k)\| \le \varepsilon$, where the number of (outer) Halpern iterations is $O\left(\frac{\|x_0 - x^\star\|}{(\eta-\rho)\varepsilon}\right)$, when we approximate the resolvent to an accuracy of poly $\left(\frac{1}{k}\right)$. To achieve this, we can run a subsolver as per (IV), with $\widetilde{O}\left(\frac{1+\eta L}{1-\eta L}\right)$ calls to evaluations of F and resolvents of G. By combining the complexities at outer and inner levels, we obtain the optimal first-order complexity under $\rho < \frac{1}{L}$.

Discussion. From the construction (I)-(IV), we see that the ingredients of our approach are based on known results. This raises the question: *what insight makes it possible to go beyond the $\rho < \frac{1}{2L}$ barrier?* The key is *conic nonexpansiveness*, the influential notion introduced by Bauschke et al. [2021]. In particular, previous results on first-order complexity for nonmonotone problems (including Pethick et al. [2023b] who utilized a similar algorithmic construction based on KM as ours in Section 3) used *nonexpansiveness* of the resolvent, which asks for the stringent requirement $\rho < \frac{1}{2L}$. This yields $\frac{1}{2}$-cocoercivity of $\mathrm{Id} - J_{\eta(F+G)}$, which allows Halpern or KM iteration to be analyzed in a standard way.

Our main starting insight is that, from the viewpoint of the analysis of Halpern iteration (or KM in Section 3), we need only cocoercivity of $\mathrm{Id} - J_{\eta(F+G)}$, not necessarily with the constant $\frac{1}{2}$. In particular, with conic nonexpansiveness, which relaxes nonexpansiveness, we still obtain that $\mathrm{Id} - J_{\eta(F+G)}$ is cocoercive, just with a constant (other than $\frac{1}{2}$) that now depends on ρ, that is, $1 - \frac{\rho}{\eta}$. Hence, as long as we stay in the range $\rho < \eta$, Halpern iteration can be analyzed for $\rho < \eta < \frac{1}{L}$, at essentially no cost.

2.2 Analysis

We now analyze the construction described in the previous section, given as Algorithm 1. We start with the main result, see Appendix A.4 for its proof.

Theorem 2.1. *Let Assumptions 1 and 2 hold. Let $\eta < \frac{1}{L}$ in Algorithm 1 and suppose $\rho < \eta$. For any $k = 1, \dots, K$, we have that (x_k) from Algorithm 1 satisfies*

$$\frac{1}{\eta^2}\|x_k - J_{\eta(F+G)}(x_k)\|^2 \le \frac{16\|x_0 - x^\star\|^2}{(\eta-\rho)^2(k+1)^2}.$$

The number of first-order oracles used at iteration k is upper bounded by $2T$ where T is defined in Algorithm 1.

Corollary 2.2. *Under the setting of Theorem 2.1, for any $\varepsilon > 0$, we have $\eta^{-1}\|(\mathrm{Id} - J_{\eta(F+G)})(x_K)\| \le \varepsilon$, for $K \le \left\lceil \frac{4\|x_0 - x^*\|}{(\eta-\rho)\varepsilon} \right\rceil$ and first-order oracle complexity*

$$\widetilde{O}\left(\frac{(1+\eta L)\|x_0 - x^\star\|}{\varepsilon\,(\eta-\rho)\,(1-\eta L)}\right).$$

Remark 2.3. The definition of $x^\star$ gives that $(\mathrm{Id} - J_{\eta(F+G)})(x^\star) = 0$ and $(\mathrm{Id} - J_{\eta(F+G)})(x_k)$ is indeed the fixed point residual, which is a standard way to measure optimality for fixed point iterations, see e.g., [Ryu and Yin, 2022, Section 2.4.2]. Based on Cor. 2.2, it is straightforward to produce x^{out} with $\mathrm{dist}(0, (F+G)(x^{\mathrm{out}})) \le \varepsilon$ as claimed in Table 1, with no change in the worst-case complexity. This is clear when $G \equiv 0$. In the general case, see [Cai et al., 2024, Lem. C.4].

Remark 2.4. The constant in our complexity deteriorates as ρ gets close to η which is the same case as most of the works included in Table 1. It is straightforward to make our bound *ρ-independent* in view of

Pethick et al. [2023b] by simply expressing ρ as a fraction of η, e.g. assume $\rho < \frac{9\eta}{10}$. Then, at the expense of a constant multiple of 10, we have the complexity $\widetilde{O}\left(\frac{(1+\eta L)\|x_0 - x^\star\|}{\varepsilon\eta(1-\eta L)}\right)$, valid for the range $\rho < \frac{9}{10L}$. In comparison, the ρ-independent complexity result in Pethick et al. [2023b] had $\widetilde{O}(\varepsilon^{-2})$ for $\rho < \frac{1}{2L}$. A similar reasoning by slightly restricting the range of ρ can also make the algorithms agnostic to the knowledge of ρ.

Outline of the analysis. We follow the steps sketched in Section 2.1. First, we compute the required number of outer iterations by using the tools mentioned in (I), (II). Second, we analyze the inner loop (Algorithm 2) as mentioned in (IV). Finally we piece together these ingredients.

2.2.1 Outer-Loop Complexity

We now analyze Halpern iteration with inexactness in the resolvent computation. See Appendix A.2 for the proof.

Lemma 2.5. *Let Assumptions 1 and 2 hold. Suppose that the iterates (x_k) of Algorithm 1 satisfy $\|J_{\eta(F+G)}(x_i) - \widetilde{J}_{\eta(F+G)}(x_i)\| \le \varepsilon_i$ for some $\varepsilon_i > 0$ and $\rho < \eta$. Then, we have for any $K \ge 1$ that*

$$\frac{K(K+1)}{4}\|x_K - J_{\eta(F+G)}(x_K)\|^2 - \frac{K+1}{K\alpha^2}\|x^\star - x_0\|^2 \le \sum_{k=0}^{K-1}\left(\frac{(k+1)(k+2)\varepsilon_k^2}{2} + (k+1)\|R(x_k)\|\varepsilon_k\right),$$

where $\|x_k - x^\star\| \le \|x_0 - x^\star\| + \frac{\alpha}{k+1}\sum_{i=0}^{k-1}(i+1)\varepsilon_i$ and $R = \mathrm{Id} - J_{\eta(F+G)}$.

In (2.1) below, we define appropriate values for ε_i, and show that the number of inner iterations T selected for `FBF` in Algorithm 1 suffices to achieve the inexactness level ε_i.

This analysis extends Diakonikolas [2020], who studied monotone inclusions, in two aspects. First, we analyze the rate under conic nonexpansiveness which is the relevant property when the parameter ρ lies in the range $[\frac{1}{2L}, \frac{1}{L})$. Second, and more importantly, we conduct a tighter error analysis that allows the inexactness on the error in resolvent computation (ε_k) to be $\widetilde{O}(k^{-3/2})$ instead of the tolerance $\widetilde{O}(k^{-3})$ used in [Diakonikolas, 2020, Yoon and Ryu, 2022, Cai et al., 2024]. Even though it is not immediately obvious, this is because the bottleneck term on the bound in Lemma 2.5 is $\sum_{k=0}^{K-1}(k+1)(k+2)\varepsilon_k^2$ which sums to a log with $\varepsilon_k = \widetilde{O}(k^{-3/2})$. This tightening becomes important in the stochastic case in Section 4, where the inner loop does not have an exponential convergence rate. The improvement derives from using a slightly smaller step size, which helps avoid the main source of *looseness* in the previous analysis. We discuss this further following (A.9). See Remark A.3 for a discussion from the viewpoint of nonexpansive operators.

2.2.2 Inner-Loop Complexity

The seminal FBF algorithm of Tseng [2000] is optimal for solving the resolvent subproblem, which is a strongly monotone inclusion. We provide the derivation of the precise constants appearing in the statement in Appendix A.3.

Theorem 2.6. *(See [Tseng, 2000, Theorem 3.4]) Let B be μ-strongly monotone with $\mu > 0$ and L_B-Lipschitz; A be maximally monotone, and $z^\star = (A+B)^{-1}(0) \neq \emptyset$. For any $\zeta > 0$, after running Algorithm 2 with initial point z_0 for $T = \left\lceil \frac{4L_B}{\mu}\log\frac{\|z_0 - z^\star\|}{\zeta}\right\rceil$ iterations and $\tau = \frac{1}{2L_B}$, we get*

$$\|z_T - z^\star\| \le \zeta,$$

where the number of calls to evaluations of B and resolvents of A is upper bounded by $2T$.

2.2.3 Total complexity

Section 2.1 already shows the key steps in our analysis, but we combine the preliminary results above into a proof sketch here, to highlight the simplicity of our approach. Full proof is given in Appendix A.4.

Proof sketch of Theorem 2.1. Denote $R = \mathrm{Id} - J_{\eta(F+G)}$ for brevity. Suppose that ε_k in Lemma 2.5 satisfies

$$\varepsilon_k = \frac{\gamma\|R(x_k)\|}{\sqrt{k+2}\log(k+2)}, \text{ with } \gamma = \frac{1}{98}. \tag{2.1}$$

We justify this supposition further below. Then we have by Lemma 2.5 (after multiplying both sides by α) that

$$\frac{\alpha K(K+1)}{4}\|R(x_K)\|^2 - \frac{K+1}{K\alpha}\|x_0 - x^\star\|^2 \le \sum_{k=0}^{K-1}\|R(x_k)\|^2\left(\frac{\alpha\gamma^2(k+1)}{2\log^2(k+2)} + \frac{\alpha\gamma\sqrt{k+2}}{\log(k+2)}\right).$$

We can show by induction from this bound that

$$\|R(x_k)\| \le \frac{4\|x_0 - x^\star\|}{\alpha(k+1)} \quad \forall k \ge 0.$$

We see that for $K \le \lceil \frac{4\|x_0 - x^\star\|}{\eta\alpha\varepsilon} \rceil$, we are guaranteed to have $\eta^{-1}\|R(x_K)\| \le \varepsilon$.

We now calculate the number of inner iterations to reach the accuracy ε_k (see (2.1)). At iteration k, as per the setup in Theorem 2.6, we set

$$A \equiv \eta G, \quad B(\cdot) \equiv (\mathrm{Id} + \eta F)(\cdot) - x_k, \quad z_0 \equiv x_k, \quad z_N \equiv \widetilde{J}_{\eta(F+G)}(x_k), \quad z^\star \equiv J_{\eta(F+G)}(x_k), \quad \zeta \equiv \varepsilon_k.$$

hence $z_0 - z^\star = (\mathrm{Id} - J_{\eta(F+g)})(x_k) = R(x_k)$. B is $L_B \equiv (1+\eta L)$-Lipschitz and $(1-\eta L)$-strongly monotone due to Fact A.1(iv). Existence of $z^\star$ is guaranteed by Fact A.1(i).

By matching these definitions with Algorithm 1, we see by invoking Theorem 2.6 that the number of inner iterations used at step k to obtain $\|J_{\eta(F+G)}(x_k) - \widetilde{J}_{\eta(F+G)}(x_k)\| \le \varepsilon_k$ is

$$T \equiv \left\lceil \frac{4(1+\eta L)}{1-\eta L}\log\frac{\|R(x_k)\|}{\varepsilon_k} \right\rceil,$$

by the settings of z_0, $z^\star$, $R(x_k)$, and ζ above, along with ε_k defined in (2.1). This value is precisely T used in Algorithm 1, which justifies our application of Lemma 2.5. By combining the bounds on inner and outer iterations, we conclude. □

3 Algorithm and Analysis under weak MVI

3.1 Algorithm Construction and Analysis Ideas

Algorithm 3 Inexact KM iteration for problems with weak MVI

Input: Parameters η, L, ρ, $\alpha = 1 - \frac{\rho}{\eta}$, $K > 0$, initial iterate $x_0 \in \mathbb{R}^d$, subroutine `FBF` given in Algorithm 2

for $k = 0, 1, 2, \ldots, K-1$ **do**

$\widetilde{J}_{\eta(F+G)}(x_k) = \texttt{FBF}\,(x_k, T, G, \mathrm{Id} + \eta F, 1 + \eta L)$, where $T = \left\lceil \frac{4(1+\eta L)}{1-\eta L}\log(8(k+1)\log^2(k+2)) \right\rceil$

$x_{k+1} = (1-\alpha)x_k + \alpha\widetilde{J}_{\eta(F+G)}(x_k)$

end for

We turn to the *weak MVI condition* of Assumption 3, which (as mentioned in Section 1.1) is weaker than cohypomonotonicity. The best-known complexity under this assumption is $O(\varepsilon^{-2})$: the lower part of Table 1 outlines existing results. Our aim is to obtain $\widetilde{O}(\varepsilon^{-2})$ complexity for the extended range $\rho < \frac{1}{L}$. The steps of our construction are as follows.

(i) We know that KM iteration (1.7), when $\mathrm{Id} - T$ is star-cocoercive, gets $\eta^{-1}\|x_k - J_{\eta(F+G)}(x_k)\| \le \varepsilon$ with $O(\varepsilon^{-2})$ evaluations of $J_{\eta(F+G)}$ [Groetsch, 1972].

(ii) We learn from Bauschke et al. [2021] that $J_{\eta(F+G)}$ has domain $\mathbb{R}^d$ and is single-valued when F is L-Lipschitz and $\eta < \frac{1}{L}$. Lemma B.2 gives that $J_{\eta(F+G)}$ is $\frac{1}{2\alpha}$-conically star-nonexpansive, with $\alpha = 1 - \frac{\rho}{\eta}$, leading to $\mathrm{Id} - J_{\eta(F+G)}$ being α-star-cocoercive.

Thus, we require $\rho < \eta$. Then, as per (ii), KM applied to $\mathrm{Id} - J_{\eta(F+G)}$ requires $O(\varepsilon^{-2})$ evaluations of $J_{\eta(F+G)}$ to find x such that $\eta^{-1}\|x - J_{\eta(F+G)}(x)\| \leq \varepsilon$.

(iii) Since F is Lipschitz and G is maximally monotone, we can estimate $J_{\eta(F+G)}$ as before (via (III) and (IV) of Section 2), with a linear rate of convergence when $\eta < \frac{1}{L}$. The existence of a solution to the subproblem is guaranteed by item (ii). The inner iterations introduce a logarithmic factor into the total complexity. As a result, the range for ρ is again $\rho < \eta < \frac{1}{L}$.

Even with inexactness, Alg. 3 is classical; see [Facchinei and Pang, 2003, Theorem 12.3.7], Combettes [2001] and Combettes and Pennanen [2002]. We analyze this scheme for problems with weak MVI solutions and characterize the first-order oracle complexity. Pethick et al. [2023b] recently analyzed a similar scheme under cohypomonotonicity, by using star-nonexpansiveness of the resulting operator.[1] Our main difference regarding the results in this section is that we harness the milder property of star-conic nonexpansiveness to improve the range of ρ (see also Bartz et al. [2022] for a similar idea by using exact resolvent). We also approximate the resolvent differently by viewing it as a strongly monotone problem and applying an optimal algorithm for this problem. FBF can be replaced with other optimal algorithms like [Malitsky and Tam, 2020], showing the modularity of our approach.

The key insight for extending the upper bound of ρ to $\frac{1}{L}$ is similar to that of Section 2. The difference is that the analysis of Halpern iteration requires conic nonexpansiveness between any pair of points in the space, making it unsuitable with weak MVI. By contrast, the KM iteration can be analyzed with conic nonexpansiveness restricted on a ray to the solution, a property that is a consequence of weak MVI. Star-conic nonexpansiveness, while not defined explicitly in Bauschke et al. [2021], directly follows by adapting the corresponding results therein by using ρ-weak MVI condition instead of cohypomonotonicity; see App. B.1.1.

3.2 Analysis

Similar to Section 2, we start with the main complexity result, under weak MVI. Its proof appears in Appendix B.4.

Theorem 3.1. *Let Assumptions 1 and 3 hold. Let $\eta < \frac{1}{L}$ in Algorithm 3 and suppose $\rho < \eta$. For any $K \geq 1$, we have*

$$\frac{1}{K}\sum_{k=0}^{K-1}\frac{1}{\eta^2}\|x_k - J_{\eta(F+G)}(x_k)\|^2 \leq \frac{11\|x_0 - x^\star\|^2}{(\eta-\rho)^2 K}.$$

The number of first-order oracles used at iteration k is upper bounded by $2T$ where T is defined in Algorithm 3.

Corollary 3.2. *Under the setting of Theorem 3.1, for any $\varepsilon > 0$, we have for some $x^{\mathrm{out}} \in \{x_0, \ldots, x_{K-1}\}$ that $\eta^{-1}\|(\mathrm{Id} - J_{\eta(F+G)})(x^{\mathrm{out}})\| \leq \varepsilon$ for $K \leq \left\lceil \frac{11\|x_0 - x^\star\|^2}{(\eta-\rho)^2\varepsilon^2} \right\rceil$ with first-order oracle complexity*

$$\widetilde{O}\left(\frac{(1+\eta L)\|x_0 - x^\star\|^2}{\varepsilon^2\,(\eta-\rho)^2\,(1-\eta L)}\right).$$

See Remark 2.3 for details to convert this result to produce a point with $\mathrm{dist}(0, (F+G)(x^{\mathrm{out}})) \leq \varepsilon$ as in Table 1.

Remark 3.3. This result is for the *best iterate*, that is, $x^{\mathrm{out}} = \arg\min_{x \in \{x_0, \ldots, x_{k-1}\}} \|(\mathrm{Id} - J_{\eta(F+G)})(x)\|$, consistent with existing results for weak MVI, see Diakonikolas et al. [2021], Pethick et al. [2022], Cai and Zheng [2022].

[1] This work claimed that some of their results extend to accommodate weak MVI condition as well.

Remark 3.4. Note that x^{out} as defined in Remark 3.3 is not computable since we do not have access to $J_{\eta(F+G)}(x_k)$. For the unconstrained case, i.e., $G \equiv 0$, we can show the result with $x^{\text{out}} = \arg\min_{x\in\{x_0,\dots,x_{K-1}\}} \|Fx\|^2$, which is computable. For the constrained problem (1.1), we can handle this issue by slightly changing how $\widetilde{J}_{\eta(F+G)}$ is calculated and requiring the knowledge of the target accuracy ε, with no change in the order of complexity bounds. We present Alg. 3 in its current form so that it is *anytime*, not requiring the target accuracy as an input. The details for making x^{out} computable are in App. B.5. We can also present this result as an *expected* bound for a *randomly selected* x^{out}, like [Diakonikolas et al., 2021, Thm. 3.2(ii)].

Outer-loop complexity. We now analyze the iteration complexity of the outer loop; see App. B.2 for a proof which is a modification of Combettes [2001] and Bartz et al. [2022] to accommodate star-conic nonexpansiveness and inexact resolvent computations.

Lemma 3.5. *Let Assumptions 1 and 3 hold. Suppose that the iterates (x_k) of Algorithm 3 satisfy $\|J_{\eta(F+G)}(x_k) - \widetilde{J}_{\eta(F+G)}(x_k)\| \le \varepsilon_k$ for some $\varepsilon_k > 0$ and $\rho < \eta$. Then, we have for $K \ge 1$ that*

$$\sum_{k=0}^{K-1} \|(\mathrm{Id} - J_{\eta(F+G)})(x_k)\|^2 - \frac{2\eta^2}{(\eta-\rho)^2}\|x_0 - x^\star\|^2 \le 6\sum_{k=0}^{K-1}\varepsilon_k^2 + \frac{4\eta}{\eta-\rho}\sum_{k=0}^{K-1}\|x_k - x^\star\|\varepsilon_k,$$

where $\|x_k - x^\star\| \le \|x_{k-1} - x^\star\| + (1-\rho/\eta)\varepsilon_{k-1}$.

Total Complexity. The sketch of the proof of Theorem 3.1 follows Section 2.2.3 closely. We use Lemma 3.5 instead of Lemma 2.5. The definition of ε_k is slightly different, as can be noticed by the number of inner iterations T in Algorithm 3. However, with the same argument in Section 2.2.3, we can show that T as in Algorithm 3 is sufficient to obtain ε_k.

4 Algorithms and Analyses with Stochasticity

In this case, F in (1.2) is accessed via unbiased oracles. Let Ξ denote the underlying distribution that we can sample from.

Assumption 4. *The* stochastic first-order oracle (SFO) $F_\xi \colon \mathbb{R}^d \to \mathbb{R}^d$ *satisfies*

$$F(x) = \mathbb{E}_{\xi\sim\Xi}[F_\xi(x)] \quad \text{and} \quad \mathbb{E}_{\xi\sim\Xi}\|F_\xi(x) - F(x)\|^2 \le \sigma^2.$$

In view of (1.1), this corresponds to using *stochastic gradients* $F(x) = \left(\begin{smallmatrix}\nabla_u f_\xi(u,v)\\ -\nabla_v f_\xi(u,v)\end{smallmatrix}\right)$. Table 2, with comparisons for stochastic problems, is in Appendix C.

4.1 Cohypomonotone Case

In this case, Algorithm 1 will call `FBF` with stochastic oracles $\widetilde{F}(x_t) := F_{\xi_t}(x_t)$ for $\xi_t \sim \Xi$ to approximate $\widetilde{J}_{\eta(F+G)}$:

$$\widetilde{J}_{\eta(F+G)}(x_k) = \texttt{FBF}\left(x_k, T, G, \mathrm{Id} + \eta\widetilde{F}, 1+\eta L\right), \tag{4.1}$$

where $T = \lceil 1734(k+2)^3\log^2(k+2)(1-\eta L)^{-2}\rceil$.

Corollary 4.1. *Let Assumptions 1, 2 and 4 hold. Let $\eta < \frac{1}{L}$ in Alg. 1, $\rho < \eta$ and use (4.1) for computing $\widetilde{J}_{\eta(F+G)}$ (see Alg. 4). For any $\varepsilon > 0$, we have $\eta^{-1}\mathbb{E}\|(\mathrm{Id} - J_{\eta(F+G)})(x_K)\| \le \varepsilon$ for the last iterate, with SFO complexity $\widetilde{O}(\varepsilon^{-4})$.*

The proof, provided in App. C.2.1 is the stochastic adaptation of Section 2. Our tighter analysis for the level of inexactness (which is highlighted after Lemma 2.5) is the main reason we could get the $\widetilde{O}(\varepsilon^{-4})$ complexity. The inexactness level required in the existing analyses in Diakonikolas [2020], Cai et al. [2024] would instead result in $\widetilde{O}(\varepsilon^{-7})$ complexity.

Remark 4.2. The previous *last iterate* result for constrained, cohypomonotone, stochastic problems by [Pethick et al., 2023b, Corollary E.3(ii)] was $\widetilde{O}(\varepsilon^{-16})$. This result also required increasing batch sizes in the inner loop and $\rho < \frac{1}{2L}$. For unconstrained problems, Chen and Luo [2022] showed an improved $\widetilde{O}(\varepsilon^{-2})$ expected complexity for $\rho < \frac{1}{2L}$ with some drawbacks described in App. D. It is an open question to get a similar complexity improvement in our setup.

Remark 4.3. Pethick et al. [2023a] has complexity $\widetilde{O}(\varepsilon^{-4})$ for a constrained problem with weak MVI. However, this work additionally assumed mean-square (MS)-Lipschitzness: $\mathbb{E}_{\xi\sim\Xi}\|F_\xi(x) - F_\xi(y)\|^2 \le L^2\|x-y\|^2$ and additional oracle access to query the operator for the same seed for two different points: $F_\xi(x_k), F_\xi(x_{k-1})$. These two assumptions define a fundamentally different template. For nonconvex minimization, for example, lower bounds improve with these assumptions compared to our standard *stochastic approximation* setting in Assumption 4, see Arjevani et al. [2023]. Moreover, the additional assumption might not hold even for trivial problems: $F_1(x) = x^2$, $F_2(x) = -x^2$ where $F = F_1 + F_2$ is clearly Lipschitz but not MS-Lipschitz.

4.2 Weak MVI Case

We next modify Algorithm 3 for the stochastic case. The main observation from the analysis (see Lemma C.8) is that bounding the *bias* $\|\mathbb{E}[\widetilde{J}_{\eta(F+G)}(x_k)] - J_{\eta(F+G)}(x_k)\|$ with square root of *variance* $\mathbb{E}\|\widetilde{J}_{\eta(F+G)}(x_k) - J_{\eta(F+G)}(x_k)\|^2$ by Jensen's inequality is too loose and would give complexity $\widetilde{O}(\varepsilon^{-6})$, like [Pethick et al., 2023b, Cor. E.3(i)].

A natural candidate for a careful bias analysis is the multilevel Monte Carlo (MLMC) technique which helps control the bias-variance tradeoff [Giles, 2008, Blanchet and Glynn, 2015, Asi et al., 2021, Hu et al., 2021]. The high level idea is that stochastic KM iteration, in our setting would give $O(\varepsilon^{-4})$ complexity if we had unbiased samples of $J_{\eta(F+G)}$ (see, e.g., Bravo and Cominetti [2024]). Obtaining such unbiased samples is highly non-trivial since $J_{\eta(F+G)}$ is an optimization problem. Fortunately, MLMC is a way to get an estimator with bias $O(\varepsilon)$ and variance $\widetilde{O}(1)$ by making, in expectation, $\widetilde{O}(1)$ calls to the oracle defined in Assumption 4. MLMC is used in Asi et al. [2021] for the related proximal point algorithm.

Estimator. Given $T \ge 1$, $M \ge 1$, set for $m = 1, \dots, M$,

$$\begin{aligned} &\widetilde{J}^{(m)}_{\eta(F+G)}(x_k) = \begin{cases} y^0 + 2^I(y^I - y^{I-1}) \text{ if } I \le T, \\ y^0, \quad \text{otherwise,} \end{cases} \\ &\text{where } I \sim \text{Geom}(1/2) \\ &\text{and } y^i = \texttt{FBF}(x_k, 2^i, G, \text{Id} + \eta\widetilde{F}, 1 + \eta L) \ \forall i \ge 0. \end{aligned} \tag{4.2}$$

Given M independent draws of this estimator, we define $\widetilde{J}_{\eta(F+G)}(x_k) = \frac{1}{M}\sum_{m=1}^M \widetilde{J}^{(m)}_{\eta(F+G)}(x_k)$. To show that the scheme is *implementable* we give the (non-optimized) values of M, T. This is to illustrate that they are agnostic to unknown quantities $\{\|x_0 - x^\star\|^2, \sigma^2\}$, unlike some MLMC methods [Chen and Luo, 2022]. Proof is in App. C.3.1.

Corollary 4.4. *Let Assumptions 1, 3 and 4 hold. In Algorithm 3, set* $\eta < \frac{1}{L}$, $\alpha \leftarrow \frac{\alpha}{\sqrt{k+2}\log(k+3)}$, *suppose that* $\rho < \eta$ *and use* (4.2) *for computing* $\widetilde{J}_{\eta(F+G)}$ *(see Algorithm 6) with* $T \equiv \lceil \frac{96(1-\eta L)^{-2}}{\min\{\frac{\alpha_k}{120\alpha(k+1)}, \frac{1}{120}\}} \rceil$ *and* $M \equiv \lceil \frac{672\times 120(\log_2 T)}{(1-\eta L)^2} \rceil$. *For any* $\varepsilon > 0$, *we have that* $\eta^{-1}\mathbb{E}\|(\text{Id} - J_{\eta(F+G)})(x^{\text{out}})\| \le \varepsilon$, *with* expected *SFO complexity* $\widetilde{O}(\varepsilon^{-4})$ *where* x^{out} *is selected uniformly at random from* $\{x_0, \dots, x_{K-1}\}$.

This result is an alternative to Pethick et al. [2023a] which required additional assumptions as explained in Remark 4.3. In our setting under Assumption 4, the only $O(\varepsilon^{-4})$ complexity was known in the special case of unconstrained problems ($G \equiv 0$), due to Diakonikolas et al. [2021] (also obtained in Choudhury et al. [2023] for a different algorithm). Because of the use of MLMC, our complexity result is *expected* number of stochastic oracle calls and hence the four results mentioned in this paragraph complement each other. See also Table 2.

MLMC is used in conditional/compositional stochastic minimization [Hu et al., 2021], distributionally robust optimization [Levy et al., 2020], and stochastic minimization with non-i.i.d. data [Dorfman and Levy, 2022]. Our development of the KM iteration with MLMC can provide the potential to extend some of these results to stochastic min-max setting.

A Proofs for Section 2

A.1 Preliminary Results

We start with the properties of the resolvent of a cohypomonotone operator and the properties of the subproblem for approximating this resolvent. These important points are also sketched in Section 2.1. We present this preliminary result here for the ease of reference throughout the proofs. Most of the conclusions follow from the results of Bauschke et al. [2021]. Note that ρ-cohypomonotone in our notation is $-\rho$-comonotone in the notation of Bauschke et al. [2021]. See also [Bauschke et al., 2021, Remark 2.5] for these two conventions.

Fact A.1. *Let Assumptions 1 and 2 hold and let $\eta > 0$. Then, we have*

(i) The operator $J_{\eta(F+G)}$ is single-valued and $\operatorname{dom} J_{\eta(F+G)} = \mathbb{R}^d$ when $\rho < \eta$.

(ii) The operator $J_{\eta(F+G)}$ is $\frac{1}{2\left(1-\frac{\rho}{\eta}\right)}$-conically nonexpansive and $\mathrm{Id} - J_{\eta(F_G)}$ is $\left(1-\frac{\rho}{\eta}\right)$-cocoercive when $\rho < \eta$.

(iii) For any $\bar{x} \in \mathbb{R}^d$, computing $J_{\eta(F+G)}(\bar{x})$ is equivalent to solving the problem:

$$\text{Find } x \in \mathbb{R}^d \text{ such that } 0 \in (\mathrm{Id} + \eta(F+G))(x) - \bar{x}. \tag{A.1}$$

The problem (A.1) has a unique solution when $\rho < \eta$.

(iv) The operator $\mathrm{Id} + \eta F$ is $(1+\eta L)$-Lipschitz and $(1-\eta L)$-strongly monotone when $\eta < \frac{1}{L}$.

Proof. (i) By Assumption 2 and the definition of cohypomonotonicity in (1.3), we have that $\eta(F+G)$ is maximally $\frac{\rho}{\eta}$-cohypomonotone. Then for $\frac{\rho}{\eta} < 1$, [Bauschke et al., 2021, Corollary 2.14] gives the result.

(ii) Since $\eta(F+G)$ is maximally $\frac{\rho}{\eta}$-cohypomonotone, [Bauschke et al., 2021, Prop. 3.11(ii)] gives $\frac{1}{2\left(1-\frac{\rho}{\eta}\right)}$-conic nonexpansiveness. Cocoercivity of $\mathrm{Id} - J_{\eta(F+G)}$ then follows from [Bauschke et al., 2021, Corollary 3.5(iii)].

(iii) Let us denote $\bar{x}^\star = J_{\eta(F+G)}(\bar{x})$ and use the definition of a resolvent to obtain

$$\bar{x}^\star = J_{\eta(F+G)}(\bar{x}) = (\mathrm{Id} + \eta(F+G))^{-1}(\bar{x}) \iff \bar{x}^\star + \eta(F+G)(\bar{x}^\star) \ni \bar{x},$$

where the existence of $\bar{x}^\star$ is guaranteed by (i). Rearranging the inclusion gives (A.1). Uniqueness of the solution is due to (i).

(iv) By Lipschitzness of F and Cauchy-Schwarz inequality, we have

$$\langle \eta F(x) - \eta F(y), x - y\rangle \geq -\eta \|F(x) - F(y)\| \|x - y\| \geq -\eta L \|x - y\|^2.$$

As a result, we have that $\mathrm{Id} + \eta F$ is $(1-\eta L)$-strongly monotone. We also have by triangle inequality that

$$\|(\mathrm{Id} + \eta F)(x) - (\mathrm{Id} + \eta F)y\| \leq \|x - y\| + \eta \|F(x) - F(y)\| \leq (1 + \eta L)\|x - y\|,$$

completing the proof. □

A.2 Complexity of the Outer loop

Bounding the norm of the iterates.

Lemma A.2. *Let Assumptions 1 and 2 hold. Suppose that the iterates (x_k) of Algorithm 1 satisfy $\|J_{\eta(F+G)}(x_k) - \widetilde{J}_{\eta(F+G)}(x_k)\| \leq \varepsilon_k$ for some $\varepsilon_k > 0$ and $\rho < \eta$. Then, we have for $k \geq 0$ that*

$$\|x_{k+1} - x^\star\| \leq \|x_0 - x^\star\| + \left(1 - \frac{\rho}{\eta}\right) \frac{1}{k+2} \sum_{i=0}^{k} (i+1)\varepsilon_i.$$

Proof. Recall the following notation from Algorithm 1:

$$\alpha = 1 - \frac{\rho}{\eta} = \frac{\eta - \rho}{\eta}.$$

Then, by Fact A.1(ii), we know that $J_{\eta(F+G)}$ is $\frac{1}{2\alpha}$-conically nonexpansive. This means that we can write $J_{\eta(F+G)} = (1 - \frac{1}{2\alpha})\mathrm{Id} + \frac{1}{2\alpha}N$ for a nonexpansive operator N.

Adding and subtracting $\alpha(1-\beta_k)J_{\eta(F+G)}(x_k)$ in the definition of x_{k+1} in Algorithm 1 and rearranging gives

$$\begin{aligned}
x_{k+1} &= \beta_k x_0 + (1-\beta_k)\left((1-\alpha)x_k + \alpha \widetilde{J}_{\eta(F+G)}(x_k)\right) \\
&= \beta_k x_0 + (1-\beta_k)\left((1-\alpha)x_k + \alpha J_{\eta(F+G)}(x_k)\right) + \alpha(1-\beta_k)\left(\widetilde{J}_{\eta(F+G)}(x_k) - J_{\eta(F+G)}(x_k)\right) \\
&= \beta_k x_0 + \frac{1-\beta_k}{2}x_k + \frac{1-\beta_k}{2}N(x_k) + \alpha(1-\beta_k)\left(\widetilde{J}_{\eta(F+G)}(x_k) - J_{\eta(F+G)}(x_k)\right),
\end{aligned}$$

where the last step is because $J_{\eta(F+G)} = \frac{2\alpha-1}{2\alpha}\mathrm{Id} + \frac{1}{2\alpha}N$ for a nonexpansive operator N.

We now use triangle inequality, nonexpansiveness of N, the definition of ε_k, and the last equality to obtain

$$\begin{aligned}
\|x_{k+1} - x^\star\| &\le \beta_k\|x_0 - x^\star\| + \frac{1-\beta_k}{2}\|x_k - x^\star\| + \frac{1-\beta_k}{2}\|N(x_k) - x^\star\| \\
&\quad + \alpha(1-\beta_k)\|\widetilde{J}_{\eta(F+G)}(x_k) - J_{\eta(F+G)}(x_k)\| \\
&\le \beta_k\|x_0 - x^\star\| + (1-\beta_k)\|x_k - x^\star\| + \alpha(1-\beta_k)\varepsilon_k, \qquad \text{(A.2)}
\end{aligned}$$

where the inequality used that $Nx^\star = x^\star$ since $N = 2\alpha J_{\eta(F+G)} + (1-2\alpha)\mathrm{Id}$ and that $J_{\eta(F+G)}(x^\star) = x^\star$ by the definition of $x^\star$ and Fact A.1(i).

The result of the lemma now follows by induction after using the definition $\beta_k = \frac{1}{k+2}$. In particular, the assertion is true for $k = 0$ by inspection. Assume the assertion holds for $k = K-1$, then (A.2) gives

$$\begin{aligned}
\|x_{K+1} - x^\star\| &\le \frac{1}{K+2}\|x_0 - x^\star\| + \frac{K+1}{K+2}\|x_K - x^\star\| + \frac{\alpha(K+1)}{K+2}\varepsilon_K \\
&\le \frac{1}{K+2}\|x_0 - x^\star\| + \frac{K+1}{K+2}\left(\|x_0 - x^\star\| + \frac{\alpha}{K+1}\sum_{i=0}^{K-1}(i+1)\varepsilon_i\right) + \frac{\alpha(K+1)}{K+2}\varepsilon_K \\
&= \|x_0 - x^\star\| + \frac{\alpha}{K+2}\sum_{i=0}^{K}(i+1)\varepsilon_i,
\end{aligned}$$

which completes the induction. The result follows after using $\alpha = 1 - \frac{\rho}{\eta}$. □

Iteration complexity

Lemma 2.5. *Let Assumptions 1 and 2 hold. Suppose that the iterates (x_k) of Algorithm 1 satisfy $\|J_{\eta(F+G)}(x_k) - \widetilde{J}_{\eta(F+G)}(x_k)\| \le \varepsilon_k$ for some $\varepsilon_k > 0$ and $\rho < \eta$. Then, we have for any $K \ge 1$ that*

$$\frac{\alpha K(K+1)}{4}\|(\mathrm{Id} - J_{\eta(F+G)})(x_K)\|^2 - \frac{K+1}{K\alpha}\|x^\star - x_0\|^2 \le \sum_{k=0}^{K-1}\left(\frac{\alpha}{2}(k+1)(k+2)\varepsilon_k^2 + \alpha(k+1)\|R(x_k)\|\varepsilon_k\right),$$

where $\alpha = 1 - \frac{\rho}{\eta}$, as defined in Algorithm 1.

Remark A.3. Halpern iteration with a nonexpansive operator N is

$$x_{k+1} = \beta_k x_0 + (1-\beta_k)N(x_k).$$

Without using smaller step size for the operator, as we propose (corresponding to strict inexactness requirements, such as Diakonikolas [2020], Cai et al. [2024]), we have $N = (1-2\alpha)\mathrm{Id} + 2\alpha J$ that is nonexpansive

(since $(\mathrm{Id} - J)$ is α-cocoercive):

$$\begin{aligned}
&\|x - 2\alpha(\mathrm{Id} - J)x - (y - 2\alpha(\mathrm{Id} - J)y)\|^2 \\
&= \|x - y\|^2 - 4\alpha\langle(\mathrm{Id} - J)x - (\mathrm{Id} - J)y, x - y\rangle + 4\alpha^2\|(\mathrm{Id} - J)(x - y)\|^2 \\
&\leq \|x - y\|^2,
\end{aligned}$$

where the inequality is by α-cocoercivity of $\mathrm{Id} - J$.

This can be viewed as a Cayley (or reflection) operator of a firmly nonexpansive operator $N' = (1-\alpha)\mathrm{Id} + \alpha J$, since $N = 2N' - \mathrm{Id}$. Our choice to make the algorithm more robust (to work with a relaxed inexactness requirement), is setting a smaller step size (named the halved-step Halpern iteration) which corresponds to

$$x_{k+1} = \beta_k x_0 + (1 - \beta_k)N'(x_k).$$

With our choice, the operator N' is firmly-nonexpansive, which is defined as

$$\|N'x - N'y\| \leq \|x - y\| - \|(\mathrm{Id} - N')x - (\mathrm{Id} - N')y\|,$$

instead of N which is just nonexpansive. We believe this helps us make the algorithm more robust to inexactness in the resolvent computation.

Proof of Lemma 2.5. By Fact A.1(ii), we have that $\mathrm{Id} - J_{\eta(F+G)}$ is $\left(1 - \frac{\rho}{\eta}\right)$ cocoercive. Recall the definition of α from Algorithm 1 and introduce a new notation for $\mathrm{Id} - J_{\eta(F+G)}$ as:

$$\alpha = 1 - \frac{\rho}{\eta} \quad \text{and} \quad R = \mathrm{Id} - J_{\eta(F+G)}.$$

With these notations, we use α-cocoercivity of R:

$$\langle R(x_{k+1}) - R(x_k), x_{k+1} - x_k\rangle \geq \alpha\|R(x_{k+1}) - R(x_k)\|^2. \tag{A.3}$$

By rearranging the update rule of x_{k+1} in Algorithm 1, we have for $k \geq 0$ that

$$\begin{aligned}
x_{k+1} &= \beta_k x_0 + (1 - \beta_k)x_k - \alpha(1 - \beta_k)(\mathrm{Id} - \widetilde{J}_{\eta(F+G)})(x_k) \\
&= \beta_k x_0 + (1 - \beta_k)x_k - \alpha(1 - \beta_k)R(x_k) + \alpha(1 - \beta_k)(\widetilde{J}_{\eta(F+G)} - J_{\eta(F+G)})(x_k),
\end{aligned} \tag{A.4}$$

where we added and subtracted $\alpha(1 - \beta_k)J_{\eta(F+G)}(x_k)$ and used the definition $R = \mathrm{Id} - J_{\eta(F+G)}$.

We now use a step that is common in the rate analysis of Halpern-type methods, which can be seen for example in Diakonikolas [2020] or Yoon and Ryu [2021]. In particular, from (A.4), we obtain two identical representations for $x_{k+1} - x_k$:

$$x_{k+1} - x_k = \beta_k(x_0 - x_k) - \alpha(1 - \beta_k)R(x_k) + \alpha(1 - \beta_k)(\widetilde{J}_{\eta(F+G)} - J_{\eta(F+G)})(x_k), \tag{A.5a}$$

$$x_{k+1} - x_k = \frac{\beta_k}{1 - \beta_k}(x_0 - x_{k+1}) - \alpha R(x_k) + \alpha(\widetilde{J}_{\eta(F+G)} - J_{\eta(F+G)})(x_k), \tag{A.5b}$$

where the second representation follows from subtracting $\beta_k x_{k+1}$ from both sides of (A.4) and rearranging. With these at hand, we develop the left-hand side of (A.3). First, by using (A.5b), we have that

$$\begin{aligned}
\langle R(x_{k+1}), x_{k+1} - x_k\rangle &= \frac{\beta_k}{1 - \beta_k}\langle R(x_{k+1}), x_0 - x_{k+1}\rangle - \alpha\langle R(x_{k+1}), R(x_k)\rangle \\
&\quad + \alpha\langle R(x_{k+1}), (\widetilde{J}_{\eta(F+G)} - J_{\eta(F+G)})(x_k)\rangle \\
&= \frac{\beta_k}{1 - \beta_k}\langle R(x_{k+1}), x_0 - x_{k+1}\rangle - \frac{\alpha}{2}\left(\|R(x_{k+1})\|^2 + \|R(x_k)\|^2 - \|R(x_{k+1}) - R(x_k)\|^2\right) \\
&\quad + \alpha\langle R(x_{k+1}), (\widetilde{J}_{\eta(F+G)} - J_{\eta(F+G)})(x_k)\rangle,
\end{aligned} \tag{A.6}$$

where the last step used the expansion $\|a - b\|^2 = \|a\|^2 - 2\langle a, b\rangle + \|b\|^2$.

Second, by using (A.5a), we have that

$$-\langle R(x_k), x_{k+1} - x_k\rangle = -\beta_k\langle R(x_k), x_0 - x_k\rangle + \alpha(1-\beta_k)\|R(x_k)\|^2 \\ - \alpha(1-\beta_k)\langle R(x_k), (\widetilde{J}_{\eta(F+G)} - J_{\eta(F+G)})(x_k)\rangle. \tag{A.7}$$

After using (A.6) and (A.7) on (A.3) and rearranging, we obtain

$$\begin{aligned}&\frac{\alpha}{2}\|R(x_{k+1})\|^2 + \frac{\beta_k}{1-\beta_k}\langle R(x_{k+1}), x_{k+1} - x_0\rangle \\ &\leq \frac{\alpha}{2}\left(1-2\beta_k\right)\|R(x_k)\|^2 + \beta_k\langle R(x_k), x_k - x_0\rangle \\ &\quad + \alpha\langle R(x_{k+1}) - (1-\beta_k)R(x_k), (\widetilde{J}_{\eta(F+G)} - J_{\eta(F+G)})(x_k)\rangle - \frac{\alpha}{2}\|R(x_{k+1}) - R(x_k)\|^2.\end{aligned} \tag{A.8}$$

For the third term on the right-hand side of (A.8), we apply Cauchy-Schwarz, triangle and Young's inequalities along with the definition of ε_k to obtain

$$\begin{aligned}&\alpha\langle R(x_{k+1}) - (1-\beta_k)R(x_k), (\widetilde{J}_{\eta(F+g)} - J_{\eta(F+G)})(x_k)\rangle \\ &\leq \alpha\|R(x_{k+1}) - (1-\beta_k)R(x_k)\|\varepsilon_k \\ &\leq \alpha\left(\|R(x_{k+1}) - R(x_k)\| + \beta_k\|R(x_k)\|\right)\varepsilon_k \\ &= \alpha\|R(x_{k+1}) - R(x_k)\|\varepsilon_k + \alpha\beta_k\|R(x_k)\|\varepsilon_k \\ &\leq \frac{\alpha}{2}\|R(x_{k+1}) - R(x_k)\|^2 + \frac{\alpha}{2}\varepsilon_k^2 + \alpha\beta_k\|R(x_k)\|\varepsilon_k.\end{aligned} \tag{A.9}$$

This is the main point of departure from the standard analyses where this inequality is bounded by $O\left(\|x_k - x^\star\|\varepsilon_k\right)$, cf. [Diakonikolas, 2020, display equation after (14)]. We instead use the last term in (A.8) (which we obtained by using a smaller step size) to cancel the corresponding error term in (A.9). We use this last estimate in (A.8) and get

$$\begin{aligned}\frac{\alpha}{2}\|R(x_{k+1})\|^2 + \frac{\beta_k}{1-\beta_k}\langle R(x_{k+1}), x_{k+1} - x_0\rangle &\leq \frac{\alpha}{2}\left(1-2\beta_k\right)\|R(x_k)\|^2 + \beta_k\langle R(x_k), x_k - x_0\rangle \\ &\quad + \frac{\alpha}{2}\varepsilon_k^2 + \alpha\beta_k\|R(x_k)\|\varepsilon_k.\end{aligned} \tag{A.10}$$

Noting the identities

$$\beta_k = \frac{1}{k+2} \implies 1-\beta_k = \frac{k+1}{k+2}, \quad \frac{\beta_k}{1-\beta_k} = \frac{1}{k+1}, \quad 1-2\beta_k = \frac{k}{k+2},$$

on(A.10) we obtain

$$\begin{aligned}\frac{\alpha}{2}\|R(x_{k+1})\|^2 + \frac{1}{k+1}\langle R(x_{k+1}), x_{k+1} - x_0\rangle &\leq \frac{\alpha}{2}\frac{k}{k+2}\|R(x_k)\|^2 + \frac{1}{k+2}\langle R(x_k), x_k - x_0\rangle \\ &\quad + \frac{\alpha}{2}\varepsilon_k^2 + \frac{\alpha}{k+2}\|R(x_k)\|\varepsilon_k,\end{aligned}$$

which holds for $k \geq 0$. Multiplying both sides by $(k+1)(k+2)$ gives

$$\begin{aligned}&\frac{\alpha(k+1)(k+2)}{2}\|R(x_{k+1})\|^2 + (k+2)\langle R(x_{k+1}), x_{k+1} - x_0\rangle \\ &\leq \frac{\alpha k(k+1)}{2}\|R(x_k)\|^2 + (k+1)\langle R(x_k), x_k - x_0\rangle \\ &\quad + \frac{\alpha}{2}(k+1)(k+2)\varepsilon_k^2 + \alpha(k+1)\|R(x_k)\|\varepsilon_k.\end{aligned}$$

We sum the inequality for $k = 0, 1, \ldots, K-1$ to get

$$\begin{aligned}&\frac{\alpha K(K+1)}{2}\|R(x_K)\|^2 + (K+1)\langle R(x_K), x_K - x_0\rangle \\ &\leq \sum_{k=0}^{K-1}\left(\frac{\alpha}{2}(k+1)(k+2)\varepsilon_k^2 + \alpha(k+1)\|R(x_k)\|\varepsilon_k\right).\end{aligned} \tag{A.11}$$

By the standard estimation for the inner product on this left-hand side (using *(i)* monotonicity of R, which is implied by α-cocoercivity of R with $\alpha > 0$; *(ii)* definition of $x^\star$ as $R(x^\star) = (\mathrm{Id} - J_{\eta(F+G)})(x^\star) = 0$ which uses Fact A.1(i); *(iii)* Young's inequality), we derive

$$\begin{aligned}(K+1)\langle R(x_K), x_K - x_0\rangle &= (K+1)\langle R(x_K), x^\star - x_0\rangle + (K+1)\langle R(x_K), x_K - x^\star\rangle \\ &\geq (K+1)\langle R(x_K), x^\star - x_0\rangle + (K+1)\langle R(x^\star), x_K - x^\star\rangle \\ &= (K+1)\langle R(x_K), x^\star - x_0\rangle \\ &\geq -\frac{\alpha K(K+1)}{4}\|R(x_K)\|^2 - \frac{K+1}{K\alpha}\|x^\star - x_0\|^2.\end{aligned}$$

We use this lower bound on (A.11) to conclude. □

A.3 Complexity of the Inner Loop

Theorem 2.6. *(See e.g., [Tseng, 2000, Theorem 3.4]) Let B be μ-strongly monotone with $\mu > 0$ and L_B-Lipschitz; A be maximally monotone, and $z^\star = (A+B)^{-1}(0) \neq \emptyset$. For any $\zeta > 0$, after running Algorithm 2 with initial point z_0 for $T = \left\lceil \frac{4L_B}{\mu} \log \frac{\|z_0 - z^\star\|}{\zeta} \right\rceil$ iterations and $\tau = \frac{1}{2L_B}$, we get*

$$\|z_T - z^\star\| \leq \zeta,$$

with the number of calls to evaluations of B and resolvents of A is upper bounded by $2\left\lceil \frac{4L_B}{\mu} \log \frac{\|z_0 - z^\star\|}{\zeta} \right\rceil$.

Proof. We only derive the number of iterations for ease of reference which follows trivially from [Tseng, 2000, Theorem 3.4]. In particular, in the notation of [Tseng, 2000, Theorem 3.4(c)], we select $\theta = \frac{1}{2}$, $\alpha = \frac{1}{2L_B}$ and assume without loss of generality that $\frac{\mu}{L_B} \leq \frac{1}{2}$ to obtain

$$\|z_{t+1} - z^\star\|^2 \leq \left(1 - \frac{\mu}{2L_B}\right)\|z_t - z^\star\|^2,$$

which after unrolling gives that

$$\|z_T - z^\star\|^2 \leq \left(1 - \frac{\mu}{2L_B}\right)^T \|z_0 - z^\star\|^2.$$

Standard manipulations give that after $T = \left\lceil \frac{4L_B}{\mu} \log \frac{\|z_0 - z^\star\|}{\zeta} \right\rceil$ iterations, we have $\|z_T - z^\star\|^2 \leq \zeta^2$. □

A.4 Total complexity

Theorem 2.1. *Let Assumptions 1 and 2 hold. Let $\eta < \frac{1}{L}$ in Algorithm 1 and $\rho < \eta$. For any $k = 1, \ldots, K$, we have that (x_k) from Algorithm 1 satisfies*

$$\frac{1}{\eta^2}\|x_k - J_{\eta(F+G)}(x_k)\|^2 \leq \frac{16\|x_0 - x^\star\|^2}{(\eta - \rho)^2(k+1)^2}.$$

The number of first-order oracles used at iteration k of Algorithm 1 is upper-bounded by

$$\left\lceil \frac{4(1+\eta L)}{1 - \eta L} \log(98\sqrt{k+2}\log(k+2)) \right\rceil.$$

Proof of Theorem 2.1. We recall the notations

$$\alpha = 1 - \frac{\rho}{\eta} \quad \text{and} \quad R = \mathrm{Id} - J_{\eta(F+G)}$$

and start from the result of Lemma 2.5 which states for $K \geq 1$ that

$$\frac{\alpha K(K+1)}{4}\|R(x_K)\|^2 \leq \frac{K+1}{K\alpha}\|x^\star - x_0\|^2 + \sum_{k=0}^{K-1}\left(\frac{\alpha}{2}(k+1)(k+2)\varepsilon_k^2 + \alpha(k+1)\|R(x_k)\|\varepsilon_k\right).$$

Let us set

$$\varepsilon_k = \frac{\gamma \|R(x_k)\|}{\sqrt{k+2}\log(k+2)} \tag{A.12}$$

and note that we will not evaluate ε_k but we will show that for a computable number of inner iterations, this error criterion will be proven to be satisfied.

We substitute the definition of ε_k to the previous inequality and get

$$\frac{\alpha K(K+1)}{4}\|R(x_K)\|^2 \leq \frac{K+1}{K\alpha}\|x_0 - x^\star\|^2 + \sum_{k=0}^{K-1}\left(\frac{\alpha\gamma^2(k+1)\|R(x_k)\|^2}{2\log^2(k+2)} + \frac{\alpha\gamma\sqrt{k+2}\|R(x_k)\|^2}{\log(k+2)}\right). \tag{A.13}$$

We now show by induction that

$$\|R(x_k)\| \leq \frac{4\|x_0 - x^\star\|}{\alpha(k+1)} \quad \forall k \geq 1. \tag{A.14}$$

Note that α^{-1}-Lipschitzness of R and $R(x^\star) = 0$ gives $\|R(x_0)\| \leq \frac{1}{\alpha}\|x_0 - x^\star\|$. For $k = 1$, we have by α^{-1}-Lipschitzness of R, $R(x^\star) = 0$ and Lemma A.2 that

$$\|R(x_1)\| \leq \frac{1}{\alpha}\|x_1 - x^\star\| \leq \frac{1}{\alpha}\left(\|x_0 - x^\star\| + \frac{\gamma\|x_0 - x^\star\|}{2\sqrt{2}\log 2}\right) < \frac{2\|x_0 - x^\star\|}{\alpha}, \tag{A.15}$$

for $\gamma = \frac{1}{98}$, which establishes the base case of induction. Now we assume (A.14) holds for all $k \leq K-1$. Then, we use (A.13) for $K \geq 2$ (where we also use $\frac{K+1}{K} \leq 2$):

$$\begin{aligned}\frac{\alpha K(K+1)}{4}\|R(x_K)\|^2 &\leq \frac{2}{\alpha}\|x_0 - x^\star\|^2 + \sum_{k=0}^{K-1}\left(\frac{\alpha\gamma^2(k+1)\|R(x_k)\|^2}{2\log^2(k+2)} + \frac{\alpha\gamma\sqrt{k+2}\|R(x_k)\|^2}{\log(k+2)}\right)\\ &\leq \frac{2}{\alpha}\|x_0 - x^\star\|^2 + \sum_{k=0}^{K-1}\left(\frac{16\gamma^2\|x_0 - x^\star\|^2}{\alpha(k+1)\log^2(k+2)} + \frac{16\gamma\sqrt{k+2}\|x_0 - x^\star\|^2}{\alpha(k+1)^2\log(k+2)}\right).\end{aligned}$$

Since we have that

$$\sum_{k=0}^{K-1}\frac{16}{(k+1)\log^2(k+2)} < 55 \text{ and } \sum_{k=0}^{K-1}\frac{16\sqrt{k+2}}{(k+1)^2\log(k+2)} < 49,$$

the value $\gamma = \frac{1}{98}$ results in

$$\frac{\alpha K(K+1)}{4}\|R(x_K)\|^2 \leq \frac{2.6}{\alpha}\|x_0 - x^\star\|^2.$$

A direct implication of this inequality is that

$$\begin{aligned}\|R(x_K)\|^2 &\leq \frac{10.4}{\alpha^2 K(K+1)}\|x_0 - x^\star\|^2\\ &\leq \frac{15.6}{\alpha^2(K+1)^2}\|x_0 - x^\star\|^2,\end{aligned}$$

where we used $\frac{1}{K(K+1)} \leq \frac{1.5}{(K+1)^2}$ which holds when $K \geq 2$. This completes the induction.

We next see that with T set as in Algorithm 1, we get the inexactness level specified by ε_k and the oracle complexity of each iteration is as claimed in the statement.

At iteration k, to apply the result in Theorem 2.6, we identify the following settings stemming from Algorithm 1

$$\begin{aligned}&A \equiv \eta G, \quad B(\cdot) \equiv (\mathrm{Id} + \eta F)(\cdot) - x_k, \quad z_0 \equiv x_k, \quad z^\star \equiv J_{\eta(F+G)}(x_k), \quad \zeta \equiv \varepsilon_k\\ \implies\ &z_0 - z^\star = (\mathrm{Id} - J_{\eta(F+g)})(x_k) = R(x_k)\end{aligned}$$

hence B is $(1+\eta L)$-Lipschitz and $(1-\eta L)$-strongly monotone due to Fact A.1(iv). Existence of $z^\star$ is guaranteed by Fact A.1(iii).

We now see that by the setting of

$$T = \left\lceil \frac{4(1+\eta L)}{1-\eta L} \log(98\sqrt{k+2}\log(k+2)) \right\rceil = \left\lceil \frac{4(1+\eta L)}{1-\eta L} \log \frac{\|R(x_k)\|}{\varepsilon_k} \right\rceil,$$

Theorem 2.6 gives us that

$$\|\widetilde{J}_{\eta(F+G)}(x_k) - J_{\eta(F+G)}(x_k)\| \leq \varepsilon_k$$

as claimed.

Since each iteration of Algorithm 2 uses 2 evaluations of F and 1 resolvent for G, the first-order oracle complexity is $2T$ and the result follows. □

We now continue with the proof of Corollary 2.2 which follows trivially from Theorem 2.1.

Proof of Corollary 2.2. By Theorem 2.1, we have that after at most $\left\lceil \frac{4\|x_0 - x^\star\|}{(\eta-\rho)\varepsilon} \right\rceil$ iterations, i.e., for a K such that

$$K \leq \left\lceil \frac{4\|x_0 - x^\star\|}{(\eta-\rho)\varepsilon} \right\rceil, \tag{A.16}$$

we are guaranteed to have

$$\eta^{-1}\|(\mathrm{Id} - J_{\eta(F+G)})(x_K)\| \leq \varepsilon.$$

Total number of first-oracle calls during the run of the algorithm then be calculated as

$$\sum_{k=1}^{K} \left\lceil \frac{4(1+\eta L)}{1-\eta L} \log(98\sqrt{k+2}\log(k+2)) \right\rceil \leq K \cdot \left(\frac{4(1+\eta L)}{1-\eta L} \log(98\sqrt{K+2}\log(K+2)) + 1 \right).$$

We conclude after using (A.16). □

B Proofs for Section 3

B.1 Preliminary results

We now derive similar properties to Fact A.1 but with Assumption 3. These proofs are slightly more involved than Fact A.1 to accommodate the weaker assumption.

For example, for the well-definedness of the resolvent in Fact A.1, we could directly use the corresponding result from Bauschke et al. [2021] since these results are shown with cohypomonotonicity. However, showing this with only *star*-cohypomonotonicity (or equivalently weak MVI condition) requires a couple more additional tools from the operator splitting literature to show, for example, that $\eta(F+G)$ is *maximally* ηL-hypomonotone which is required to utilize existence results from Bauschke et al. [2021] in our setting.

We start with the definition of *star*-conic nonexpansiveness that will be used in Fact B.1. Recall that an operator N is star-nonexpansive when $\|Nx - x^\star\| \leq \|x - x^\star\|$ where $x^\star$ is a fixed point of N.

Definition 1. *$T\colon \mathbb{R}^d \to \mathbb{R}^d$ is α-conically star-nonexpansive if there exists a star-nonexpansive operator $N\colon \mathbb{R}^d \to \mathbb{R}^d$ such that $T = (1-\alpha)\mathrm{Id} + \alpha N$.*

This is a direct relaxation of conic nonexpansiveness in [Bauschke et al., 2021, Definition 3.1]. In Appendix B.1.1, we show the star-conic nonexpansiveness (and related properties) of the resolvent of a star-cohypomonotone operator in view of Assumption 3, by invoking the corresponding arguments of Bauschke et al. [2021] restricted to a point in the domain and a fixed point of the resolvent. Then we show *star*-cocoercivity of $\mathrm{Id} - J_{\eta(F+G)}$ which facilitates the analysis of KM iteration.

Fact B.1. *Let Assumptions 1 and 3 hold. Then, we have*

(i) The operator $J_{\eta(F+G)}$ is single-valued and $\operatorname{dom} J_{\eta(F+G)} = \mathbb{R}^d$ when $\eta < \frac{1}{L}$.

(ii) The operator $J_{\eta(F+G)}$ is $\frac{1}{2\left(1-\frac{\rho}{\eta}\right)}$-star-conically nonexpansive and $\mathrm{Id} - J_{\eta(F_G)}$ is $\left(1-\frac{\rho}{\eta}\right)$-star-cocoercive when $\rho < \eta$.

(iii) For any $\bar{x} \in \mathbb{R}^d$, computing $J_{\eta(F+G)}(\bar{x})$ is equivalent to solving the problem

$$\text{Find } x \in \mathbb{R}^d \text{ such that } 0 \in (\mathrm{Id} + \eta(F+G))(x) - \bar{x}. \tag{B.1}$$

The problem (B.1) has a unique solution when $\eta < \frac{1}{L}$.

(iv) The operator $\mathrm{Id} + \eta F$ is $(1+\eta L)$-Lipschitz and $(1-\eta L)$-strongly monotone.

Proof.

(i) Since $F+G$ has a ρ-weak MVI solution under Assumption 3, we have that $\eta(F+G)$ has ρ/η-weak MVI solution, i.e., by simple change of variables, we have for some $\eta > 0$

$$\begin{aligned} &\langle \eta u, x - x^* \rangle \geq \eta\rho\|u\|^2 \quad \text{where} \quad u \in (F+G)(x) \\ \iff &\langle v, x - x^* \rangle \geq \frac{\rho}{\eta}\|v\|^2 \quad \text{where} \quad v \in \eta(F+G)(x). \end{aligned}$$

Additionally, when F is L-Lipschitz, it is maximally L-hypomonotone (see e.g., [Giselsson and Moursi, 2021, Lemma 2.12]) and ηF is maximally ηL-hypomonotone since

$$\begin{aligned} &\langle F(x) - F(y), x - y \rangle \geq -\|F(x) - F(y)\|\|x - y\| \geq -L\|x - y\|^2 \\ \implies &\eta\langle F(x) - F(y), x - y \rangle \geq -\eta L\|x - y\|^2. \end{aligned}$$

By [Dao and Phan, 2019, Lemma 3.2(ii)], we know that $\eta F + \mathrm{Id}$ is maximally $(1-\eta L)$-(strongly) monotone. Then, using this and maximal monotonicity of G, we have by [Bauschke and Combettes, 2017, Corollary 25.5] that $\mathrm{Id} + \eta(F+G)$ is maximally $(1-\eta L)$-(strongly) monotone. Invoking [Dao and Phan, 2019, Lemma 3.2(ii)] again gives us that $\eta(F+G)$ is maximally ηL-hypomonotone.

We can then use [Bauschke et al., 2021, Lemma 2.8] to obtain that $(\eta(F+G))^{-1}$ is maximally ηL-cohypomonotone. This can be combined with [Bauschke et al., 2021, Corollary 2.14] to get the result when $\eta L < 1$.

(ii) As shown in the proof of (i), we have that $\eta(F+G)$ is $\frac{\rho}{\eta}$-star-cohypomonotone (i.e., a solution exists to the $\frac{\rho}{\eta}$-weakly MVI). Lemma B.2 then gives us that $J_{\eta(F+G)}$ is $\frac{1}{2\left(1-\frac{\rho}{\eta}\right)}$-conically star-nonexpansive and as a result $\mathrm{Id} - J_{\eta(F+G)}$ is $\left(1 - \frac{\rho}{\eta}\right)$ star-cocoercive by Corollary B.3.

(iii) The proof is the same as Fact A.1(iii) where the only difference is that now we ensure the existence of $J_{\eta(F+G)}$ with (i). Uniqueness of the solution is apparent from combining (i) and (iii).

(iv) The proof is the same as Fact A.1(iv). □

B.1.1 Properties of Conic Star-Nonexpansiveness

This section particularizes the notion and properties of the α-conic nonexpansiveness in Bauschke et al. [2021] to their *star* variants. The aim is to show that the properties extend to their *star*-variants when we use weak MVI condition instead of cohypomonotonicity. This sections implicitly assumes that J_A for operator $A\colon \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is well-defined, which will be detailed later. We say that an operator N is star-nonexpansive when $\|Nx - x^\star\| \leq \|x - x^\star\|$ where $x^\star$ is a fixed point of N.

Lemma B.2. *(See [Bauschke et al., 2021, Lemma 3.4]) Consider $T\colon \mathbb{R}^d \to \mathbb{R}^d$ and let $T = (1-\alpha)\mathrm{Id} + \alpha N$. Then, N is star-nonexpansive if and only if we have, for all $x \in \mathbb{R}^d$,*

$$2\alpha \langle Tx - x^\star, (\mathrm{Id} - T)x \rangle \geq (1 - 2\alpha)\|(\mathrm{Id} - T)x\|^2,$$

or equivalently

$$\left\| \left(1 - \frac{1}{\alpha}\right) x + \frac{1}{\alpha} Tx - x^\star \right\| \leq \|x - x^\star\|. \tag{B.2}$$

Proof. Using $\alpha^2\|a\|^2 - \|(\alpha-1)a+b\|^2 = 2\alpha\langle b, a-b\rangle - (1-2\alpha)\|a-b\|^2$ (see [Bauschke et al., 2021, Lemma 3.3]) with $a = x - x^\star$ and $b = Tx - x^\star$, we have

$$\begin{aligned}
0 &\le 2\alpha\langle Tx - x^\star, (\mathrm{Id}-T)x\rangle - (1-2\alpha)\|(\mathrm{Id}-T)x\|^2 \\
&= \alpha^2\|x-x^\star\|^2 - \|(\alpha-1)(x-x^\star) + Tx - x_*\|^2 \\
&= \alpha^2\|x-x^\star\|^2 - \|(\alpha-1)(x-x^\star) + (1-\alpha)(x-x^\star) + \alpha(Nx-x^\star)\|^2 \\
&= \alpha^2(\|x-x^\star\|^2 - \|Nx-x^\star\|^2),
\end{aligned}$$

which gives the assertion. Last claim follows by substituting $N = \frac{1}{\alpha}T + \left(1-\frac{1}{\alpha}\right)\mathrm{Id}$ in the definition of star-nonexpansiveness for N. □

Corollary B.3. *(See [Bauschke et al., 2021, Corollary 3.5(iii)]) $T\colon \mathbb{R}^d \to \mathbb{R}^d$ is α-conically star-nonexpansive if and only if $\mathrm{Id} - T$ is $\frac{1}{2\alpha}$-star-cocoercive.*

Proof. We use Lemma B.2:

$$\langle Tx - x^\star, (\mathrm{Id}-T)x\rangle \ge \left(\frac{1}{2\alpha} - 1\right)\|(\mathrm{Id}-T)x\|^2 \quad \Leftrightarrow \quad \langle x - x^\star, (\mathrm{Id}-T)x\rangle \ge \frac{1}{2\alpha}\|(\mathrm{Id}-T)x\|^2,$$

which is simply adding to both sides $\|(\mathrm{Id}-T)(x)\|^2$. □

Proposition B.4. *(See [Bauschke et al., 2021, Proposition 3.6(i)]) Let $A = T^{-1} - \mathrm{Id}$ and set $N = \frac{1}{\alpha}T - \frac{1-\alpha}{\alpha}\mathrm{Id}$, i.e., $T = J_A = (\mathrm{Id}+A)^{-1} = (1-\alpha)\mathrm{Id} + \alpha N$. Then, T is α-conically star-nonexpansive if and only if A is $\left(1-\frac{1}{2\alpha}\right)$-star-cohypomonotone, i.e.,*

$$\langle x - x^\star, Ax\rangle \ge -\left(1 - \frac{1}{2\alpha}\right)\|Ax\|^2.$$

Proof. We see the two directions:

"⇒" Let $(x,u) \in \operatorname{gra} A$. Then by definition of $A = T^{-1} - \mathrm{Id}$ and manipulations, it follows that $(x,u) = (T(x+u), (\mathrm{Id}-T)(x+u))$. By Lemma B.2 invoked with $x \leftarrow x+u$, we have

$$\begin{aligned}
& 2\alpha\langle T(x+u) - x^\star, (\mathrm{Id}-T)(x+u)\rangle \ge (1-2\alpha)\|(\mathrm{Id}-T)(x+u)\|^2 \\
\Leftrightarrow\ & 2\alpha\langle x - x^\star, u\rangle \ge (1-2\alpha)\|u\|^2,
\end{aligned}$$

where the last step substituted $(x,u) = (T(x+u), (\mathrm{Id}-T)(x+u))$.

"⇐" Since $(Tx, (\mathrm{Id}-T)x) \in \operatorname{gra} A$, we have by star-cohypomonotonicity that $\langle Tx - x^\star, (\mathrm{Id}-T)x\rangle \ge \left(\frac{1}{2\alpha} - 1\right)\|(\mathrm{Id}-T)x\|^2$. In view of Lemma B.2, we deduce conic star-nonexpansiveness.

□

B.2 Complexity of the Outer Loop

Bounding the norm of iterates. Just like Appendix A, we start with the bound of the norms of the iterates.

Lemma B.5. *Let Assumptions 1 and 3 hold. Suppose that the iterates (x_k) of Algorithm 3 satisfy $\|J_{\eta(F+G)}(x_k) - \widetilde{J}_{\eta(F+G)}(x_k)\| \le \varepsilon_k$ for some $\varepsilon_k > 0$ and $\rho < \eta$. Then, we have for $k \ge 0$ that*

$$\|x_{k+1} - x^\star\| \le \|x_k - x^\star\| + \left(1 - \frac{\rho}{\eta}\right)\varepsilon_k.$$

Proof. From Fact B.1(ii), we know that $J_{\eta(F+G)}$ is $\frac{1}{2\left(1-\frac{\rho}{\eta}\right)}$-conically star-nonexpansive. Then, by property (B.2) derived in Lemma B.2, since $J_{\eta(F+G)}$ is also $\frac{1}{1-\frac{\rho}{\eta}}$-conically star-nonexpansive due to $2\left(1-\frac{\rho}{\eta}\right) \ge 1 - \frac{\rho}{\eta}$ (see also Corollary B.3), we have

$$\left\|\frac{\rho}{\eta}x_k + \left(1-\frac{\rho}{\eta}\right)J_{\eta(F+G)}(x_k) - x^\star\right\| \le \|x_k - x^\star\|. \tag{B.3}$$

By the definition of x_{k+1} in Algorithm 3, the definition of ε_k and triangle inequality, we have for $k \geq 0$ that

$$\begin{aligned}\|x_{k+1} - x^\star\| &\leq \left\| \frac{\rho}{\eta} x_k + \left(1 - \frac{\rho}{\eta}\right) J_{\eta(F+G)}(x_k) - x^\star \right\| + \left(1 - \frac{\rho}{\eta}\right) \|J_{\eta(F+G)}(x_k) - \widetilde{J}_{\eta(F+G)}(x_k)\| \\ &\leq \left\| \frac{\rho}{\eta} x_k + \left(1 - \frac{\rho}{\eta}\right) J_{\eta(F+G)}(x_k) - x^\star \right\| + \left(1 - \frac{\rho}{\eta}\right) \varepsilon_k.\end{aligned}$$

Combining with (B.3) gives the result. □

Iteration complexity. Equipped with this result, we proceed to deriving the iteration complexity of the outer loop.

Lemma 3.5. *Let Assumptions 1 and 3 hold. Suppose that the iterates (x_k) of Algorithm 3 satisfy $\|J_{\eta(F+G)}(x_k) - \widetilde{J}_{\eta(F+G)}(x_k)\| \leq \varepsilon_k$ for some $\varepsilon_k > 0$ and $\rho < \eta$. Then, we have for $K \geq 1$ that*

$$\sum_{k=0}^{K-1} \|(\mathrm{Id} - J_{\eta(F+G)})(x_i)\|^2 \leq \frac{2}{\left(1 - \frac{\rho}{\eta}\right)^2} \|x_0 - x^\star\|^2 + 6 \sum_{k=0}^{K-1} \varepsilon_k^2 + \frac{4}{1 - \frac{\rho}{\eta}} \sum_{k=0}^{K-1} \|x_k - x^\star\| \varepsilon_k, \tag{B.4}$$

where

$$\|x_k - x^\star\| \leq \|x_{k-1} - x^\star\| + \left(1 - \frac{\rho}{\eta}\right) \varepsilon_{k-1}.$$

Proof. From Fact B.1(ii), we have that $\mathrm{Id} - J_{\eta(F+G)}$ is $\left(1 - \frac{\rho}{\eta}\right)$-star cocoercive. Let us recall our running notations:

$$\alpha = 1 - \frac{\rho}{\eta}, \quad R = \mathrm{Id} - J_{\eta(F+G)}, \quad \widetilde{R} = \mathrm{Id} - \widetilde{J}_{\eta(F+G)}.$$

As a result, we have the following equivalent representation of x_{k+1} (see the definition in Algorithm 3):

$$\begin{aligned}x_{k+1} &= x_k - \left(1 - \frac{\rho}{\eta}\right) \left(\mathrm{Id} - \widetilde{J}_{\eta(F+G)}\right)(x_k) \\ &= x_k - \alpha \widetilde{R}(x_k).\end{aligned} \tag{B.5}$$

Then, by α-star-cocoercivity of R, we have

$$\langle R(x_k), x_k - x^\star \rangle \geq \alpha \|R(x_k)\|^2. \tag{B.6}$$

A simple decomposition gives

$$\langle R(x_k), x_k - x^\star \rangle = \langle \widetilde{R}(x_k), x_k - x^\star \rangle + \langle R(x_k) - \widetilde{R}(x_k), x_k - x^\star \rangle. \tag{B.7}$$

We estimate the first term on the right-hand side of (B.7) as

$$\begin{aligned}\langle \widetilde{R}(x_k), x_k - x^\star \rangle &= \frac{1}{\alpha} \langle x_k - x_{k+1}, x_k - x^\star \rangle \\ &= \frac{1}{2\alpha} \left(\|x_k - x_{k+1}\|^2 + \|x_k - x^\star\|^2 - \|x_{k+1} - x^\star\|^2 \right) \\ &\leq \frac{1}{2\alpha} \left(\|x_k - x^\star\|^2 - \|x_{k+1} - x^\star\|^2 \right) + \frac{3\alpha}{4} \|R(x_k)\|^2 + \frac{3\alpha}{2} \|\widetilde{R}(x_k) - R(x_k)\|^2 \\ &\leq \frac{1}{2\alpha} \left(\|x_k - x^\star\|^2 - \|x_{k+1} - x^\star\|^2 \right) + \frac{3\alpha}{4} \|R(x_k)\|^2 + \frac{3\alpha \varepsilon_k^2}{2},\end{aligned} \tag{B.8}$$

where we used the definition of x_{k+1} from (B.5) in the first step, standard expansion $\|a - b\|^2 = \|a\|^2 - 2\langle a, b \rangle + \|b\|^2$ for the second step, the definition of x_{k+1} from (B.5) and Young's inequality in the third step, and the definitions of $R_k, \widetilde{R}_k, \varepsilon_k$ in the last step.

For the second term on the right-hand side of (B.7), we have by Cauchy-Schwarz inequality and the definition of $\widetilde{R}$ and ε_k that

$$\begin{aligned}\langle R(x_k) - \widetilde{R}(x_k), x_k - x^\star\rangle &\leq \|R(x_k) - \widetilde{R}(x_k)\|\|x_k - x^\star\| \\ &\leq \|x_k - x^\star\|\varepsilon_k. \end{aligned} \tag{B.9}$$

We combine (B.8) and (B.9) in (B.7), plug in the result to (B.6) and rearrange to obtain

$$\frac{\alpha}{4}\|R(x_k)\|^2 \leq \frac{1}{2\alpha}\left(\|x_k - x^\star\|^2 - \|x_{k+1} - x^\star\|^2\right) + \frac{3\alpha\varepsilon_k^2}{2} + \|x_k - x^\star\|\varepsilon_k.$$

The result follows by multiplying both sides by $4/\alpha$, summing for $k = 0, 1, \ldots, K-1$, and using the definition of α. The bound on $\|x_k - x^\star\|^2$ follows by Lemma B.5. □

B.3 Complexity of the Inner Loop

In a modular fashion, we will use precisely the same algorithm for the inner loop, i.e., the Forward-Backward-Forward (FBF) algorithm of Tseng [2000] like the Section A.3. Hence the complexity of the inner loop is the same as Theorem 2.6. As we see in the next section, the accuracy required by $\widetilde{J}_{\eta(F+G)}$ is slightly different leading to the number of inner loop iterations T in Algorithm 3 to be slightly different than Algorithm 1.

B.4 Total Complexity

Theorem 3.1. *Let Assumptions 1 and 3 hold. Let $\eta < \frac{1}{L}$ in Algorithm 3 and suppose that $\rho < \eta$. For any $K > 1$, we have that*

$$\frac{1}{K}\sum_{k=0}^{K-1}\frac{1}{\eta^2}\|x_k - J_{\eta(F+G)}(x_k)\|^2 \leq \frac{11\|x_0 - x^\star\|^2}{(\eta-\rho)^2 K}.$$

The number of first-order oracles used at iteration k of Algorithm 3 is upper bounded by

$$\left\lceil \frac{4(1+\eta L)}{1-\eta L}\log(8(k+2)\log^2(k+2)) \right\rceil.$$

Remark B.6. It is straightforward to convert this to a last-iterate result if we additionally assume cohypomonotonicity as in Pethick et al. [2023b], but we refrain from doing so since the main point of this section is to *relax* cohypomonotonicity.

Proof of Theorem 3.1. Recall the notations $\alpha = 1 - \frac{\rho}{\eta}$ and $R = \mathrm{Id} - J_{\eta(F+G)}$. Let us set

$$\varepsilon_k = \frac{1}{8(k+1)\log^2(k+2)}\|x_k - J_{\eta(F+G)}(x_k)\| \tag{B.10}$$

and note, just as in the proof of Theorem 2.1, that we will not evaluate the value of ε_k but we will show that for the number of iterations that FBF runs at each KM iteration in Algorithm 3, the error criterion dictated by ε_k is satisfied.

By using the definition of ε_k in Lemma B.5 gives

$$\|x_{k+1} - x^\star\| \leq \|x_k - x^\star\| + \frac{\alpha}{8(k+1)\log^2(k+2)}\|x_k - J_{\eta(F+G)}(x_k)\|. \tag{B.11}$$

We note that $\alpha = 1 - \frac{\rho}{\eta} \leq 1$ and since $R = \mathrm{Id} - J_{\eta(F+G)}$ is α-star cocoercive (see, e.g., Corollary B.3), we have that R is α^{-1}-star Lipschitz and hence by $(\mathrm{Id} - J_{\eta(F+G)})(x^\star) = 0$ we have

$$\|(\mathrm{Id} - J_{\eta(F+G)})(x_k)\| = \|(\mathrm{Id} - J_{\eta(F+G)})(x_k) - (\mathrm{Id} - J_{\eta(F+G)})(x^\star)\| \leq \alpha^{-1}\|x_k - x^\star\|. \tag{B.12}$$

Consequently, (B.11) becomes, after summing for $k = 0, 1, \ldots, K-1$ that

$$\|x_K - x^\star\| \leq \|x_0 - x^\star\| + \sum_{i=0}^{K-1}\frac{1}{8(i+1)\log^2(i+2)}\|x_i - x^\star\|.$$

We can show, by induction, that

$$\|x_k - x^\star\| \le 2\|x_0 - x^\star\| \quad \forall k \ge 0, \tag{B.13}$$

because $\sum_{i=0}^{\infty} \frac{1}{(i+1)\log^2(i+2)} < 4$.

We use (B.13) in the result of Lemma 3.5 to obtain (also noting the definitions of α and R)

$$\sum_{k=0}^{K-1} \|R(x_k)\|^2 \le \frac{2}{\alpha^2}\|x_0 - x^\star\|^2 + \sum_{k=0}^{K-1} 6\varepsilon_k^2 + \frac{4}{\alpha}\sum_{k=0}^{K-1} 2\|x_0 - x^\star\|\varepsilon_k. \tag{B.14}$$

By using (B.13) and (B.12) in (B.10) we also know the following upper bound on ε_k:

$$\varepsilon_k \le \frac{\|x_0 - x^\star\|}{4\alpha(k+1)\log^2(k+2)}.$$

With this, (B.14) becomes

$$\begin{aligned}\sum_{k=0}^{K-1} \|R(x_k)\|^2 &\le \frac{2}{\alpha^2}\|x_0 - x^\star\|^2 + \sum_{k=0}^{K-1} \frac{3\|x_0 - x^\star\|^2}{8\alpha^2(k+1)^2\log^4(k+2)} + \sum_{k=0}^{K-1} \frac{2\|x_0 - x^\star\|^2}{\alpha^2(k+1)\log^2(k+2)} \\ &< \frac{11}{\alpha^2}\|x_0 - x^\star\|^2,\end{aligned} \tag{B.15}$$

since $\sum_{k=0}^{K-1} \frac{3}{8(k+1)^2\log^4(k+2)} < 2$ and $\sum_{k=0}^{K-1} \frac{2}{(k+1)\log^2(k+1)} < 7$. This establishes the first part of the assertion.

We next see that, with T set as in Algorithm 1, we get the inexactness level specified by ε_k and we verify that the oracle complexity of each iteration is as claimed in the statement.

For the second part of the result, we proceed similar to the proof of Theorem 2.1. Namely, at iteration k, we apply the result in Theorem 2.6. For this, let us identify the following from the definitions in Algorithm 3

$$\begin{aligned}&A \equiv \eta G, \quad B(\cdot) \equiv (\mathrm{Id} + \eta F)(\cdot) - x_k, \quad z_0 \equiv x_k, \quad z^\star \equiv J_{\eta(F+G)}(x_k), \quad \zeta \equiv \varepsilon_k \\ \implies\ & z_0 - z^\star = (\mathrm{Id} - J_{\eta(F+g)})(x_k) = R(x_k).\end{aligned}$$

As before, we have that B is $(1+\eta L)$-Lipschitz and $(1-\eta L)$-strongly monotone due to Fact B.1(iv). Existence of $z^\star$ is guaranteed by Fact B.1(iii) since $\eta < \frac{1}{L}$.

We now see that by the setting of T from Algorithm 3 and definition of ε_k, we have

$$T = \left\lceil \frac{4(1+\eta L)}{1-\eta L}\log(8(k+1)\log^2(k+2)) \right\rceil = \left\lceil \frac{4(1+\eta L)}{1-\eta L}\log\frac{\|R(x_k)\|}{\varepsilon_k} \right\rceil.$$

With this value, Theorem 2.6 gives us

$$\|\widetilde{J}_{\eta(F+G)}(x_k) - J_{\eta(F+G)}(x_k)\| \le \varepsilon_k$$

as claimed.

Since each iteration of Algorithm 2 uses 2 evaluations of F and 1 resolvent for G, the first-order oracle complexity is $2T$ and the result follows. □

We continue with the proof of Corollary 3.2 which follows trivially from Theorem 3.1.

Proof of Corollary 3.2. Based on Theorem 3.1, we have that after K iterations where

$$K \le \left\lceil \frac{11\|x_0 - x^\star\|^2}{\eta^2\alpha^2\varepsilon^2} \right\rceil \tag{B.16}$$

we are guaranteed to obtain

$$\min_{0\le k\le K-1} \eta^{-1}\|R(x_k)\| \le \frac{1}{K}\sum_{k=0}^{K-1} \eta^{-1}\|R(x_k)\| \le \varepsilon$$

Total number of first-oracle calls during the run of the algorithm then be calculated as

$$\sum_{k=1}^{K}\left\lceil\frac{4(1+\eta L)}{1-\eta L}\log(8(k+2)\log^2(k+2))\right\rceil \leq K\cdot\left(\frac{4(1+\eta L)}{1-\eta L}\log(8(K+2)\log^2(K+2))+1\right).$$

We conclude after using (B.16). □

B.5 Additional Results

Let us re-emphasize the strategy in the previous proof: we set a value for ε_k and then we show that when we run the inner algorithm FBF for a certain *computable* number of iterations, the criterion enforced on $\widetilde{J}_{\eta(F+G)}$ by ε_k is satisfied. However, this number of inner iterations is *worst-case*. Another alternative, which could be more useful in practice is to set ε_k to a computable value and monitor the progress of the inner algorithm and break when ε_k is attained. One sidenote is that this is attainable in the deterministic case considered in this section, however it cannot be done in the stochastic case since the convergence guarantees are generally given in expectation.

This described strategy can be made rigorous with the only change being in the constants in our deterministic case. We now see this in the next proposition.

Corollary B.7. *Let Assumptions 1 and 3 hold and let $G=\partial\iota_C$ for a convex closed set C. Let $\eta<\frac{1}{L}$ and $\rho<\eta$ in Algorithm 1 with $\|\widetilde{J}_{\eta(F+G)}(x_k)-J_{\eta(F+G)}(x_k)\|\leq\frac{c}{k\log^2(k+2)}$ for any $c>0$ and use [Diakonikolas, 2020, Algorithm 4] to obtain such $\widetilde{J}_{\eta(F+G)}(x_k)$ at iteration k. Then, we have that*

$$\frac{1}{K}\sum_{k=0}^{K-1}\eta^{-1}\|(\mathrm{Id}-J_{\eta(F+G)})(x_k)\|\leq\varepsilon,$$

with the number of calls to evaluation of F and resolvent of G is bounded by $\tilde{O}\left(\frac{(1+\eta L)((1+c)\|x_0-x^\star\|^2+c^2)}{\varepsilon^2(\eta-\rho)^2(1-\eta L)}\right)$.

Remark B.8. Note that [Diakonikolas, 2020, Algorithm 4] has a built-in stopping criterion to terminate the algorithm when the required accuracy is achieved. The value for ε_k defined in this corollary is computable since it only depends on k and a user-defined constant c. This is an alternative to `FBF` we used in the main text where we use a computable number of iterations to run the inner algorithm rather than using a stopping criterion as [Diakonikolas, 2020, Algorithm 4]. On the one hand, in practice, a stopping criterion can be more desirable since the worst-case number of iterations can be pessimistic. On the other hand, the strategy of using a stopping criterion is inherently more complicated in the stochastic case whereas using a worst-case computable number of inner iteration is still easily implementable. This is why we considered the latter setting throughout the paper. However, this corollary is still included for the former strategy.

Proof of Corollary B.7. We obtain the result for modifying the proof of Theorem 3.1. We set

$$\varepsilon_k=\frac{c}{(k+1)\log^2(k+2)},$$

for any $c>0$.

By using this on Lemma B.5 and summing the result for $k=0,1,\ldots,K-1$ we obtain

$$\begin{aligned}\|x_k-x^\star\|&\leq\|x_0-x^\star\|+\alpha\sum_{k=0}^{K-1}\frac{c}{(k+1)\log^2(k+2)}\\&\leq\|x_0-x^\star\|+4\alpha c,\end{aligned}$$

since $\sum_{k=0}^{K-1}\frac{1}{(k+1)\log^2(k+2)}<4$. We use this bound on the result of Lemma 3.5 to obtain

$$\begin{aligned}\sum_{k=0}^{K-1}\|R(x_k)\|^2&\leq\frac{2}{\alpha^2}\|x_0-x^\star\|^2+\sum_{k=0}^{K-1}\frac{6c^2}{(k+1)^2\log^4(k+2)}+\frac{4}{\alpha}\sum_{k=0}^{K-1}\frac{c(\|x_0-x^\star\|+4\alpha c)}{(k+1)\log^2(k+2)}\\&\leq\left(\frac{2}{\alpha^2}+\frac{16c}{\alpha}\right)\|x_0-x^\star\|^2+30c^2+64c^2,\end{aligned}$$

which gives the result after dividing by η^2 and noting that [Diakonikolas, 2020, Lemma 17] gives complexity $\widetilde{O}\left(\frac{1+\eta L}{1-\eta L}\right)$ for obtaining such a $\widetilde{J}_{\eta(F+G)}(x_k)$. □

We continue with the result mentioned in Remark 3.4.

Corollary B.9. *Let Assumptions 1 and 3 hold. Let $\eta < \frac{1}{L}$ and $\rho < \eta$.*

(i) Let $G \equiv 0$ and consider Algorithm 3. Then we have that $\min_{0\le k\le K-1} \|F(x_k)\| \le 2\varepsilon$ with the first-order oracle calls bounded by

$$\widetilde{O}\left(\frac{(1+\eta L)\|x_0 - x^\star\|^2}{\varepsilon^2 (\eta-\rho)^2 (1-\eta L)}\right).$$

(ii) Let $G \equiv \partial \iota_C$ for a convex closed set $C \subseteq \mathbb{R}^d$. Given $\varepsilon > 0$, consider Algorithm 3 with the update $\widetilde{J}_{\eta(F+G)}(x_k)$ replaced with [Diakonikolas, 2020, Algorithm 4] with error criterion $\|\widetilde{J}_{\eta(F+G)}(x_k) - J_{\eta(F+G)}(x_k)\| \le \frac{\eta\varepsilon^2}{(k+1)\log^2(k+3)}$. Then, for $x^{out} = \arg\min_{x\in\{x_0,\dots,x_{k-1}\}} \|x - \widetilde{J}_{\eta(F+G)}(x)\|$, we have that $\eta^{-1}\|(\mathrm{Id} - J_{\eta(F+G)})(x^{out})\| \le 2\varepsilon + 3\varepsilon^2$ with the first-order oracle calls bounded by

$$\widetilde{O}\left(\frac{(1+\eta L)\|x_0 - x^\star\|^2}{\varepsilon^2 (\eta-\rho)^2 (1-\eta L)}\right). \tag{B.17}$$

See also Remark 2.3 for details on how we can use this result to further obtain a guarantee like $\mathrm{dist}(0, (F+G)(x^{\mathrm{out}})) \le \varepsilon$.

Proof of Corollary B.9. (i) In this case, we start from the final steps of the proof of Theorem 3.1 (see (B.15)) which, after using $R = \mathrm{Id} - J_{\eta(F+G)}$, gives us that

$$\frac{1}{K}\sum_{k=0}^{K-1} \eta^{-2}\|x_k - J_{\eta F}(x_k)\|^2 \le \varepsilon^2, \tag{B.18}$$

with the prescribed complexity bound given in Theorem 3.1. Let us define $\bar{x}_k = J_{\eta F}(x_k)$.

On the one hand, we use the definition of resolvent to obtain

$$\bar{x}_k = J_{\eta F}(x_k) \iff \bar{x}_k + \eta F(\bar{x}_k) = x_k \iff x_k - \bar{x}_k = \eta F(\bar{x}_k),$$

which, in view of (B.18), means that we have

$$\frac{1}{K}\sum_{k=0}^{K-1} \|F(\bar{x}_k)\|^2 \le \varepsilon^2. \tag{B.19}$$

On the other hand, we know by Young's inequality and Lipschitzness of F that

$$\begin{aligned}
\frac{1}{K}\sum_{k=0}^{K-1} \|F(x_k)\|^2 &\le \frac{1}{K}\sum_{k=0}^{K-1} 2\|F(\bar{x}_k)\|^2 + \frac{1}{K}\sum_{k=0}^{K-1} 2\|F(x_k) - F(\bar{x}_k)\|^2 \\
&\le \frac{1}{K}\sum_{k=0}^{K-1} 2\|F(\bar{x}_k)\|^2 + \frac{1}{K}\sum_{k=0}^{K-1} 2L^2\|x_k - \bar{x}_k\|^2 \\
&\le (2 + 2\eta^2 L^2)\varepsilon^2 \\
&< 4\varepsilon^2,
\end{aligned}$$

where we used (B.18) and (B.19).

(ii) A slight modification of the proof of Corollary B.7 by using $\varepsilon_k = \frac{\eta\varepsilon^2}{(k+1)\log^2(k+3)}$ gives us that

$$\frac{1}{K}\sum_{k=0}^{K-1}\eta^{-2}\|(\mathrm{Id}-J_{\eta(F+G)})(x_k)\|^2 \le \varepsilon^2 \tag{B.20}$$

with the complexity bound (B.17). This is because [Diakonikolas, 2020, Lemma 17] showed that [Diakonikolas, 2020, Algorithm 4] outputs a $\widetilde{J}_{\eta(F+g)}(x_k)$ satisfying the requirement set by $\varepsilon_k = \frac{\eta\varepsilon^2}{(k+1)\log^2(k+3)}$, with the same worst-case complexity as Theorem 2.6. The difference is that [Diakonikolas, 2020, Algorithm 4] has a computable stopping criterion (instead of the maximum number of iterations Algorithm 2 takes) where we can check if $\varepsilon_k = \frac{\eta\varepsilon^2}{(k+1)\log^2(k+3)}$ accuracy is achieved and break the loop.

Since we have the pointwise bound $\|J_{\eta(F+G)}(x_k) - \widetilde{J}_{\eta(F+G)}(x_k)\|^2 \le \eta^2\varepsilon^4$, we derive from (B.20) that

$$\frac{1}{K}\sum_{k=0}^{K-1}\eta^{-2}\|(\mathrm{Id}-\widetilde{J}_{\eta(F+G)})(x_k)\|^2 \le 2(\varepsilon^2+\varepsilon^4).$$

Hence, for x^{out} defined in the statement, we get

$$\eta^{-2}\|(\mathrm{Id}-\widetilde{J}_{\eta(F+G)})(x^{\mathrm{out}})\|^2 \le 2(\varepsilon^2+\varepsilon^4). \tag{B.21}$$

Then, by using the pointwise bound again, we know that

$$\begin{aligned}\eta^{-1}\|(\mathrm{Id}-J_{\eta(F+G)})(x^{\mathrm{out}})\| &\le \eta^{-1}\|(\mathrm{Id}-\widetilde{J}_{\eta(F+G)})(x^{\mathrm{out}})\| + \eta^{-1}\|(J_{\eta(F+G)}-\widetilde{J}_{\eta(F+G)})(x^{\mathrm{out}})\| \\ &\le \varepsilon^2 + \sqrt{2(\varepsilon^4+\varepsilon^2)} < 2\varepsilon + 3\varepsilon^2,\end{aligned}$$

which uses (B.21) and the implication of the error criterion $\|J_{\eta(F+G)}(x_k) - \widetilde{J}_{\eta(F+G)}(x_k)\| \le \eta\varepsilon^2$, completing the proof. □

C Proofs for Section 4

Notation. We use the following definitions for conditional expectations: For expectation conditioned on the filtration generated by the randomness of $x_k, \dots, x_1$, we use $\mathbb{E}_k[\cdot]$ while analyzing Algorithm 4 and Algorithm 6. In the notation of Algorithm 5, we similarly use $\mathbb{E}_{t+1/2}[\cdot]$ for the expectation conditioned on the filtration generated by the randomness of $z_{t+1/2}, z_t, \dots, z_1, z_{1/2}$. Unif denotes the uniform distribution and Geom denotes the geometric distribution.

Table 2 summarizes the existing works for stochastic min-max problems satisfying cohypomonotonicity or weak MVI conditions.

C.1 Analysis of the inner loop for stochastic problems

The main change for algorithms in the stochastic case is computing the resolvent approximation $\widetilde{J}_{\eta(F+G)}(x_k)$. We now need to invoke `FBF` with unbiased oracles for F, see for example (4.1). For ease of reference, we specify the algorithm below. Note that Algorithm 4 is precisely Algorithm 1 when (4.1) is used for estimating the resolvent and Algorithm 5 is precisely Algorithm 2 when unbiased oracle $\widetilde{B}$ is inputted rather than full operator B. Algorithm 5 is a stochastic version of FBF, which is analyzed in the monotone case by Böhm et al. [2022].

More particularly, we solve the following stochastic strongly monotone inclusion problem:

$$\text{Find } x^\star \in \mathbb{R}^d \text{ such that } 0 \in (A+B)(x^\star), \text{ where } B = \mathbb{E}_{\xi\sim\Xi}[B_\xi].$$

Similar results to next theorem appeared in Hsieh et al. [2019], Kotsalis et al. [2022]. We provide a proof for being complete and precise since we could not find a particular reference for stochastic FBF with strong monotonicity and explicit constants. It is also worth noting that we do not focus on optimizing the non-dominant terms. A tight bound for all the terms can be found in Kotsalis et al. [2022].

Algorithm 4 Stochastic Inexact Halpern iteration for problems with cohypomonotonicity

Input: Parameters $\beta_k = \frac{1}{k+2}, \eta, L, \rho$, $\alpha = 1 - \frac{\rho}{\eta}$, $K > 0$, initial iterate $x_0 \in \mathbb{R}^d$, subroutine `FBF` given in Algorithm 5

for $k = 0, 1, 2, \ldots, K-1$ **do**
  $\widetilde{J}_{\eta(F+G)}(x_k) = \texttt{FBF}\left(x_k, T, G, \mathrm{Id} + \eta\widetilde{F}, 1 + \eta L\right)$, where $T = \left\lceil \frac{1734(k+2)^3 \log^2(k+2)}{(1-\eta L)^2} \right\rceil$
  $x_{k+1} = \beta_k x_0 + (1-\beta_k)((1-\alpha)x_k + \alpha \widetilde{J}_{\eta(F+G)}(x_k))$
end for

Algorithm 5 $\texttt{FBF}(z_0, T, A, \widetilde{B}, L_B)$ from [Tseng, 2000] – Stochastic

Input: Parameter $\tau_t = \frac{2}{(t+1)\mu + 6L_B}$, initial iterate $z_0 \in \mathbb{R}^d$

for $t = 0, 1, 2, \ldots, T-1$ **do**
  $z_{t+1/2} = J_{\tau_t A}(z_t - \tau_t B_{\xi_t}(z_t))$
  $z_{t+1} = z_{t+1/2} + \tau_t B_{\xi_t}(z_t) - \tau_t B_{\xi_{t+1/2}}(z_{t+1/2})$
end for

Theorem C.1. *Let $z^\star = (A+B)^{-1}(0) \neq \emptyset$, the operator B be L_B-Lipschitz and μ-strongly monotone with $\mu > 0$, A be maximally monotone. Let $\widetilde{B}(\cdot, \xi)$ satisfy $\mathbb{E}_{\xi\sim\Xi}[\widetilde{B}(\cdot,\xi)] = B(\cdot)$ and $\mathbb{E}_{\xi\sim\Xi}\|\widetilde{B}(x,\xi) - B(x)\|^2 \leq \sigma^2$. Then, we have that the last iterate of Algorithm 5 after running for T iterations, when initialized with z_0, and step size $\tau_t = \frac{2}{(t+1)\mu + 6L_B}$ satisfies the bound*

$$\mathbb{E}\|z_T - z^\star\|^2 \leq \frac{6L_B/\mu \|z_0 - z^\star\|^2 + 48\sigma^2/\mu^2}{T + 6L_B}.$$

Each iteration of the algorithm uses two evaluations of $\widetilde{B}$ and one resolvent of A

Proof. Note that the definition of $z_{t+1/2}$ implies $\tau_t A(z_{t+1/2}) \ni z_t - z_{t+1/2} - \tau_t B_{\xi_t}(z_t)$. The definition of $z^\star$ implies $\tau_t A(z^\star) \ni -\tau_t B(z^\star)$ By using this with monotonicity of A, we get

$$\langle z_{t+1/2} - z_t + \tau_t B_{\xi_t}(z_t) - \tau_t B(z^\star), z^\star - z_{t+1/2}\rangle \geq 0.$$

By the definition of z_{t+1}, we then have

$$\langle z_{t+1} - z_t + \tau_t B_{\xi_{t+1/2}}(z_{t+1/2}) - \tau_t B(z^\star), z^\star - z_{t+1/2}\rangle \geq 0. \tag{C.1}$$

By taking expectation conditioned on $z_{t+1/2}$ and also using strong monotonicity of B, we also have

$$\begin{aligned}
\mathbb{E}_{t+1/2}\langle \tau_t B_{\xi_{t+1/2}}(z_{t+1/2}) - \tau_t B(z^\star), z_{t+1/2} - z^\star\rangle &= \langle \tau_t B(z_{t+1/2}) - \tau_t B(z^\star), z_{t+1/2} - z^\star\rangle \\
&\geq \mu\tau_t \|z^\star - z_{t+1/2}\|^2 \\
&\geq \frac{\mu\tau_t}{2}\|z^\star - z_{t+1}\|^2 - \mu\tau_t\|z_{t+1} - z_{t+1/2}\|^2 \\
&\geq \frac{\mu\tau_t}{2}\|z^\star - z_{t+1}\|^2 - \frac{1}{3}\|z_{t+1} - z_{t+1/2}\|^2,
\end{aligned} \tag{C.2}$$

where the third step is by Young's inequality and last step is by the definition of τ_t, i.e., $\tau_t\mu = \frac{2\mu}{(t+1)\mu + 6L} \leq \frac{2\mu}{6L} \leq \frac{1}{3}$ since $\mu \leq L$.

We have, by the elementary identities $\langle a, b\rangle = \frac{1}{2}\left(\|a\|^2 + \|b\|^2 - \|a-b\|^2\right) = \frac{1}{2}\left(-\|a\|^2 - \|b\|^2 + \|a+b\|^2\right)$, that

$$\begin{aligned}
\langle z_{t+1} - z_t, z^\star - z_{t+1/2}\rangle &= \langle z_{t+1} - z_t, z^\star - z_{t+1}\rangle + \langle z_{t+1} - z_t, z_{t+1} - z_{t+1/2}\rangle \\
&= \frac{1}{2}\left(\|z_t - z^\star\|^2 - \|z_{t+1} - z^\star\|^2 - \|z_t - z_{t+1/2}\|^2 + \|z_{t+1} - z_{t+1/2}\|^2\right).
\end{aligned} \tag{C.3}$$

Assumption	Reference	Limit of ρ	Constraints	Batch size	Oracle†	Complexity
weak MVI	Diakonikolas et al. [2021]	$\frac{1}{4\sqrt{2}L}$	×	$O(\varepsilon^{-2})$	Single	$O(\varepsilon^{-4})$
	Choudhury et al. [2023]	$\frac{1}{2L}$	×	$O(\varepsilon^{-2})$	Single	$O(\varepsilon^{-4})$
	Böhm [2022]	$\frac{1}{2L}$	×	$O(\varepsilon^{-2})$	Single	$O(\varepsilon^{-6})$
	Pethick et al. [2023a]	$\frac{1}{2L}$	✓	1	Multiple	$O(\varepsilon^{-4})$
	Theorem C.11	$\frac{1}{L}$	✓	$\widetilde{O}(1)$	Single	$\widetilde{O}(\varepsilon^{-4})$
cohypomonotone	Pethick et al. [2023b]	$\frac{1}{2L}$	✓	k^2	Single	$\widetilde{O}(\varepsilon^{-6})$ (best)‡
	Pethick et al. [2023b]	$\frac{1}{2L}$	✓	k^3	Single	$\widetilde{O}(\varepsilon^{-16})$ (last)
	Chen and Luo [2022]*	$\frac{1}{2L}$	×	$\widetilde{O}(1)$	Single	$\widetilde{O}(\varepsilon^{-2})$
	Corollary C.5*	$\frac{1}{L}$	✓	1	Single	$\widetilde{O}(\varepsilon^{-4})$

Table 2: Comparison of first order algorithms for stochastic problems. Complexity refers to the number of oracle calls to get the fixed point residual $\mathbb{E}\|(\mathrm{Id} - J_{\eta(F+G)})(x^{\mathrm{out}})\| \leq \varepsilon$. See also Remark 2.3. †Oracle access refers to the number of operator evaluations algorithm makes with one random seed. For example, "Single" refers to algorithms that only access one sample per seed, i.e., only $F_{\xi_t}(x_t)$, "Multiple" is for algorithms that access multiple samples per seed, i.e., $F_{\xi_t}(x_t)$ and $F_{\xi_t}(x_{t-1})$. Algorithms with "Multiple" access also make the additional assumption that $\mathbb{E}_{\xi\sim\Xi}\|F_\xi(x) - F_\xi(y)\|^2 \leq L^2\|x-y\|^2$ which is stronger than mere Lipschitzness of F. ‡(best) refers to *best iterate* in view of Remark 3.3; (last) refers to a last iterate convergence rate. *These works have complexity as *expected* number of oracle calls due to the use of MLMC estimator. See also Appendix D.1 for derivations of the complexities when they are not written explicitly in the existing works.

Using (C.2) and (C.3) on (C.1) after taking total expectation, using tower rule and dividing both sides by τ_t gives

$$\left(\frac{1}{2\tau_t} + \frac{\mu}{2}\right)\mathbb{E}\|z^\star - z_{t+1}\|^2 \leq \frac{1}{2\tau_t}\mathbb{E}\|z^\star - z_t\|^2 + \frac{5}{6\tau_t}\mathbb{E}\|z_{t+1} - z_{t+1/2}\|^2 - \frac{1}{2\tau_t}\mathbb{E}\|z_t - z_{t+1/2}\|^2. \tag{C.4}$$

Definition of z_{t+1} in Algorithm 2 gives

$$\begin{aligned}\frac{5}{6}\|z_{t+1} - z_{t+1/2}\| &= \frac{5\tau_t^2}{6}\|B_{\xi_t}(z_t) - B_{\xi_{t+1/2}}(z_{t+1/2})\|^2 \\ &\leq \frac{5\tau_t^2}{2}\left(\|B_{\xi_t}(z_t) - B(z_t)\|^2 + \|B(z_t) - B(z_{t+1/2})\|^2 + \|B(z_{t+1/2}) - B_{\xi_{t+1/2}}(z_{t+1/2})\|^2\right) \\ &\leq 5\tau_t^2\sigma^2 + \frac{5\tau_t^2 L_B^2}{2}\|z_t - z_{t+1/2}\|^2.\end{aligned}$$

With this, we get in place of (C.4) that

$$\left(\frac{1}{2\tau_t} + \frac{\mu}{2}\right)\mathbb{E}\|z^\star - z_{t+1}\|^2 \leq \frac{1}{2\tau_t}\mathbb{E}\|z^\star - z_t\|^2 + \frac{1}{\tau_t}\left(\frac{5\tau_t^2 L_B^2}{2} - \frac{1}{2}\right)\mathbb{E}\|z_t - z_{t+1/2}\|^2 + 5\tau_t\sigma^2. \tag{C.5}$$

The definition of $\tau_t = \frac{2}{(t+1)\mu + 6L_B}$ has two consequences:

$$\frac{1}{2\tau_t} + \frac{\mu}{2} = \frac{6L_B + (t+3)\mu}{4}$$

and

$$\tau_t = \frac{2}{(t+1)\mu + 6L_B} \leq \frac{1}{3L_B} \Longrightarrow \tau_t^2 \leq \frac{1}{5L_B^2} \iff 5\tau_t^2 L_B^2 \leq 1.$$

As a result, we obtain, after multiplying both sides of (C.5) by $\left(\frac{1}{2\tau_t} + \frac{\mu}{2}\right)^{-1} = \frac{4}{6L_B + (t+3)\mu}$ that

$$\mathbb{E}\|z^\star - z_{t+1}\|^2 \leq \left(\frac{(t+1)\mu + 6L}{(t+3)\mu + 6L}\right)\|z^\star - z_t\|^2 + \frac{40\sigma^2}{(6L + (t+1)\mu)(6L + (t+3)\mu)}. \tag{C.6}$$

We next show by induction that

$$\mathbb{E}\|z^\star - z_t\|^2 \le \frac{6L/\mu\|z_0 - z^\star\|^2 + 48\sigma^2/\mu^2}{t + 6L/\mu} \quad \forall t \ge 0.$$

For brevity, let us denote $\kappa = 6L/\mu$.

The base case $t = 0$ holds by inspection. Next we assume the assertion holds for $t = T$ and consider (C.6) to deduce

$$\begin{aligned}
&\mathbb{E}\|z^\star - z_{T+1}\|^2 \\
&\le \frac{T+1+\kappa}{T+3+\kappa}\frac{6L/\mu\|z_0 - z^\star\|^2 + 48\sigma^2/\mu^2}{T+\kappa} + \frac{40\sigma^2/\mu^2}{(T+1+\kappa)(T+3+\kappa)} \\
&\le \left(\frac{(T+1+\kappa)}{(T+3+\kappa)(T+\kappa)} + \frac{1}{1.2(T+1+\kappa)(T+3+\kappa)}\right)\left(6L/\mu\|z_0 - z^\star\|^2 + 48\sigma^2/\mu^2\right).
\end{aligned}$$

As a result, the inductive step will be implied by

$$\left(\frac{(T+1+\kappa)^2}{(T+1+\kappa)(T+3+\kappa)(T+\kappa)} + \frac{(T+\kappa)}{1.2(T+1+\kappa)(T+3+\kappa)(T+\kappa)}\right) \le \frac{1}{T+1+\kappa},$$

which, after letting $\alpha = T + \kappa$, is equivalent to

$$\left(\frac{1.2(\alpha+1)^2}{(\alpha+3)\alpha} + \frac{\alpha}{(\alpha+3)\alpha}\right) \le 1.2 \iff 1.2(\alpha+1)^2 \le 1.2\alpha^2 + 2.6\alpha \iff 1.2 \le 0.2\alpha \iff 6 \le \alpha.$$

This holds because $\alpha = T + \kappa = T + 6L/\mu \ge 6$ since $L/\mu > 1$. This completes the induction. □

C.2 Stochastic Problem with Cohypomonotonicity

We have a stochastic version of Lemma A.2 proof of which is almost equivalent.

Lemma C.2. *Let Assumptions 1 and 2 hold. For the sequence (x_k) generated by Algorithm 4 with $\mathbb{E}_k\|J_{\eta(F+G)}(x_k) - \widetilde{J}_{\eta(F+G)}(x_k)\|^2 \le \varepsilon_k^2$, we have for $k \ge 0$ that*

$$\mathbb{E}\|x_{k+1} - x^\star\| \le \|x_0 - x^\star\| + \frac{\alpha}{k+2}\sum_{i=0}^{k}(i+1)\mathbb{E}[\varepsilon_i].$$

Proof. The proof follows the same steps as Lemma A.2 after taking expectation on (A.2) and using Jensen's inequality since

$$\mathbb{E}_k\left[\|\widetilde{J}_{\eta(F+G)}(x_k) - J_{\eta(F+G)}(x_k)\|\right] \le \sqrt{\mathbb{E}_k\left[\|\widetilde{J}_{\eta(F+G)}(x_k) - J_{\eta(F+G)}(x_k)\|^2\right]} \le \varepsilon_k.$$

Hence the result follows by tower rule and the same induction as the proof of Lemma A.2. □

Lemma C.3. *Let Assumptions 1 and 2 hold. Consider Algorithm 4 with $\mathbb{E}_k\|\widetilde{J}_{\eta(F+G)}(x_k) - J_{\eta(F+G)}(x_k)\|^2 \le \varepsilon_k^2$. Then, we have for any $\gamma > 0$ and $K \ge 1$ that*

$$\frac{\alpha K(K+1)}{4}\mathbb{E}\|R(x_K)\|^2 \le \frac{K+1}{K\alpha}\|x^\star - x_0\|^2 + \sum_{k=0}^{K-1}\left(\frac{\alpha(\gamma+1)}{2\gamma}(k+1)(k+2)\mathbb{E}[\varepsilon_k^2] + \frac{\gamma\alpha\mathbb{E}\|R(x_k)\|^2}{2}\right).$$

Proof. We follow the proof of Lemma 2.5 until (A.8) and then we take expectation to obtain

$$\begin{aligned}
&\frac{\alpha}{2}\mathbb{E}\|R(x_{k+1})\|^2 + \frac{\beta_k}{1-\beta_k}\mathbb{E}\langle R(x_{k+1}), x_{k+1} - x_0\rangle \\
&\le \frac{\alpha}{2}(1-2\beta_k)\,\mathbb{E}\|R(x_k)\|^2 + \beta_k\mathbb{E}\langle R(x_k), x_k - x_0\rangle \\
&\quad + \alpha\mathbb{E}\langle R(x_{k+1}) - (1-\beta_k)R(x_k), (\widetilde{J}_{\eta(F+G)} - J_{\eta(F+G)})(x_k)\rangle - \frac{\alpha}{2}\mathbb{E}\|R(x_{k+1}) - R(x_k)\|^2. \qquad \text{(C.7)}
\end{aligned}$$

We then consider (A.9) after taking expectation and using Cauchy-Schwarz, triangle and Young's inequalities to obtain

$$\begin{aligned}
&\alpha\mathbb{E}\langle R(x_{k+1}) - (1-\beta_k)R(x_k), (\widetilde{J}_{\eta(F+g)} - J_{\eta(F+G)})(x_k)\rangle \\
&\leq \alpha\mathbb{E}\left[(\|R(x_{k+1}) - R(x_k)\| + \beta_k\|R(x_k)\|)\,\|\widetilde{J}_{\eta(F+G)}(x_k) + J_{\eta(F+G)}(x_k)\|\right] \\
&\leq \frac{\alpha}{2}\mathbb{E}\|R(x_{k+1}) - R(x_k)\|^2 + \frac{\gamma\alpha\beta_k^2}{2}\mathbb{E}\|R(x_k)\|^2 + \frac{\alpha}{2}\left(1+\frac{1}{\gamma}\right)\mathbb{E}\|\widetilde{J}_{\eta(F+G)}(x_k) - J_{\eta(F+G)}(x_k)\|^2 \\
&\leq \frac{\alpha}{2}\mathbb{E}\|R(x_{k+1}) - R(x_k)\|^2 + \frac{\gamma\alpha\beta_k^2}{2}\mathbb{E}\|R(x_k)\|^2 + \frac{\alpha}{2}\left(1+\frac{1}{\gamma}\right)\mathbb{E}[\varepsilon_k^2], \qquad \text{(C.8)}
\end{aligned}$$

where the last step also used the tower rule along with the definition of ε_k.

We then use the same arguments as those after (A.10) to get the result. □

C.2.1 Proof for Corollary 4.1

Corollary 4.1 is essentially the summary of the results proven below.

Theorem C.4. *Let Assumptions 1, 2, and 4 hold. Let $\eta < \frac{1}{L}$ in Algorithm 4 and $\rho < \eta$. For any $K \geq 1$, we have that*

$$\frac{1}{\eta^2}\mathbb{E}\|x_K - J_{\eta(F+G)}(x_K)\|^2 \leq \frac{36(\|x_0 - x^\star\|^2 + \sigma^2)}{(\eta-\rho)^2K^2}.$$

The number of first-order oracles used at iteration k of Algorithm 1 is upper bounded by

$$2\left\lceil \frac{1734(k+2)^3\log^2(k+2)}{(1-\eta L)^2}\right\rceil. \qquad \text{(C.9)}$$

Corollary C.5. *Let Assumptions 1 and 2 hold. Let $\eta < \frac{1}{L}$ in Algorithm 4 and $\rho < \eta$. For any $\varepsilon > 0$, we have that $\mathbb{E}\left[\eta^{-1}\|x_k - J_{\eta(F+G)}(x_k)\|\right] \leq \varepsilon$ with stochastic first-order oracle complexity*

$$\widetilde{O}\left(\frac{\|x_0 - x^\star\|^4 + \sigma^4}{(\eta-\rho)^4(1-\eta L)^2\varepsilon^4}\right)$$

Proof. This corollary immediately follows from Theorem C.4 by combining the number of outer iterations and the number of stochastic first-order oracle calls for each outer iteration. □

Remark C.6. The complexity in the previous corollary has the same dependence on $\|x_0 - x^\star\|, \sigma$ as Pethick et al. [2023a], Bravo and Cominetti [2024]. As we see in the next remark, the dependence on $(\eta - \rho)$ can be improved by using the knowledge of the target accuracy ε and the variance upper bound σ^2 as done in Diakonikolas et al. [2021], Kim [2021], Chen and Luo [2022].

Remark C.7. By using parameters depending on target accuracy ε and noise variance σ^2, we can improve the complexity to

$$\widetilde{O}\left(\frac{\|x_0 - x^\star\|^2\sigma^2}{(\eta-\rho)^2(1-\eta L)^2\varepsilon^4}\right)$$

Proof of Theorem C.4. Let us set

$$\varepsilon_k^2 = \frac{\gamma^2(\alpha^2\|R(x_k)\|^2 + 8\sigma^2)}{\alpha^2(k+2)^3\log^2(k+2)} \qquad \text{(C.10)}$$

and plug this in to the result of Lemma C.3 to obtain

$$\begin{aligned}
&\frac{\alpha K(K+1)}{4}\mathbb{E}\|R(x_K)\|^2 \\
&\leq \frac{K+1}{K\alpha}\|x_0 - x^\star\|^2 + \sum_{k=0}^{K-1}\mathbb{E}\left(\frac{(\gamma^2+\gamma)(\alpha^2\|R(x_k)\|^2 + 8\sigma^2)}{2\alpha(k+2)\log^2(k+2)} + \frac{\gamma\alpha\|R(x_k)\|^2}{2}\right) \\
&= \frac{K+1}{K\alpha}\|x_0 - x^\star\|^2 + \sum_{k=0}^{K-1}\left(\frac{4(\gamma^2+\gamma)\sigma^2}{\alpha(k+2)\log^2(k+2)} + \frac{\alpha(\gamma^2+\gamma)\mathbb{E}\|R(x_k)\|^2}{2(k+2)\log^2(k+2)} + \frac{\gamma\alpha\mathbb{E}\|R(x_k)\|^2}{2}\right) \\
&< \frac{K+1}{K\alpha}\|x_0 - x^\star\|^2 + \frac{12(\gamma^2+\gamma)\sigma^2}{\alpha} + \sum_{k=0}^{K-1}\left(\frac{\alpha(\gamma^2+\gamma)\mathbb{E}\|R(x_k)\|^2}{2(k+2)\log^2(k+2)} + \frac{\gamma\alpha\mathbb{E}\|R(x_k)\|^2}{2}\right),
\end{aligned} \tag{C.11}$$

since $\sum_{k=0}^{K-1}\frac{1}{(k+2)\log^2(k+2)} < 3$.

We now show by induction that

$$\mathbb{E}\|R(x_k)\|^2 \leq \frac{36(\|x_0 - x^\star\|^2 + \sigma^2)}{\alpha^2(k+1)^2}$$

The base case for the induction with $K = 0, 1$ hold the same way as the proof of Theorem 2.1 where the only change is we use Lemma C.2 and the definition of ε_k in (C.10), see also (A.15).

Let us consider (C.11) for $K \geq 2$ and assume the assertion holds for $k \leq K-1$. We then have, by also noting $\frac{K+1}{K} \leq 2$ that

$$\begin{aligned}
&\frac{\alpha K(K+1)}{4}\mathbb{E}\|R(x_K)\|^2 \\
&\leq \frac{2}{\alpha}\|x_0 - x^\star\|^2 + \frac{12(\gamma^2+\gamma)\sigma^2}{\alpha} + \sum_{k=0}^{K-1}\left(\frac{18(\gamma^2+\gamma)(\|x_0 - x^\star\|^2 + \sigma^2)}{\alpha(k+2)(k+1)^2\log^2(k+2)} + \frac{18\gamma(\|x_0 - x^\star\|^2 + \sigma^2)}{\alpha(k+1)^2}\right).
\end{aligned}$$

By using $\sum_{k=0}^{\infty}\frac{18}{(k+1)^2} < 30$ and $\sum_{k=0}^{\infty}\frac{18}{(k+2)(k+1)^2\log^2(k+2)} < 21$.

With $\gamma = \frac{1}{17}$, we have that

$$\frac{\alpha K(K+1)}{4}\mathbb{E}\|R(x_K)\|^2 \leq \frac{6(\|x_0 - x^\star\|^2 + \sigma^2)}{\alpha}.$$

We use $\frac{1}{K(K+1)} \leq \frac{1.5}{(K+1)^2}$ which holds for $K \geq 2$ to complete the induction.

To see the number of first-order oracles, we use the result for stochastic FBF in Theorem C.1. For our subproblem at iteration k, this result gives

$$\begin{aligned}
\mathbb{E}_k\left[\|\widetilde{J}_{\eta(F+G)}(x_k) - J_{\eta(F+G)}(x_k)\|^2\right] &\leq \frac{6\left(\frac{1+\eta L}{1-\eta L}\|x_k - J_{\eta(F+G)}(x_k)\|^2 + \frac{8\sigma^2}{(1-\eta L)^2}\right)}{T} \\
&\leq \frac{\frac{6}{(1-\eta L)^2}\left(\|x_k - J_{\eta(F+G)}(x_k)\|^2 + 8\sigma^2\right)}{T}.
\end{aligned}$$

Recall that (C.10), with $\gamma = \frac{1}{17}$ and $R = \mathrm{Id} - J_{\eta(F+G)}$, requires

$$\mathbb{E}_k\left[\|\widetilde{J}_{\eta(F+G)}(x_k) - J_{\eta(F+G)}(x_k)\|^2\right] \leq \frac{(\alpha^2\|(\mathrm{Id} - J_{\eta(F+G)})(x_k)\|^2 + 8\sigma^2)}{289\alpha^2(k+2)^3\log^2(k+2)}$$

Noting that $\frac{1}{\alpha^2} > 1$, a sufficient condition to attain this requirement is

$$\frac{\frac{6}{(1-\eta L)^2}\left(\|x_k - J_{\eta(F+G)}(x_k)\|^2 + 8\sigma^2\right)}{T} \leq \frac{\|x_k - J_{\eta(F+G)}(x_k)\|^2 + 8\sigma^2}{289(k+2)^3\log^2(k+2)},$$

verifying the required number of iterations T as given in Algorithm 4 to be sufficient for the inexactness criterion. Since each iteration of FBF takes 2 stochastic operator evaluations $\widetilde{F}$ and one resolvent of G, we have the result. □

C.3 Stochastic Problem with weak MVI condition

As motivated in Section 4.2, we use the multilevel Monte Carlo (MLMC) estimator [Giles, 2008, Blanchet and Glynn, 2015, Asi et al., 2021, Hu et al., 2021]. In Section 4.2, we only sketched the main changes in Algorithm 3 because of space limitations. We start by explicitly writing down the algorithm with MLMC estimator.

Algorithm 6 Inexact KM iteration for problems with weak MVI

Input: Parameters η, L, ρ, $\alpha = 1 - \frac{\rho}{\eta}$, $\alpha_k = \frac{\alpha}{\sqrt{k+2}\log(k+3)}$ $K > 0$, initial iterate $x_0 \in \mathbb{R}^d$, subroutine `MLMC-FBF` given in Algorithm 7

for $k = 0, 1, 2, \ldots, K-1$ **do**

$T \leftarrow \lceil \frac{96(1-\eta L)^{-2}}{\min\{\frac{\alpha_k}{120\alpha(k+1)}, \frac{1}{120}\}} \rceil$ and $M \leftarrow \lceil \frac{672 \times 120(\log_2 T)}{(1-\eta L)^2} \rceil$

$\widetilde{J}^{(m)}_{\eta(F+G)}(x_k) = \texttt{MLMC-FBF}\left(x_k, T, \mathrm{Id} + \eta\widetilde{F}, G, 1+\eta L\right)$ independently for each $m = 1, \ldots, M$

$\widetilde{J}_{\eta(F+G)}(x_k) = \frac{1}{M}\sum_{i=1}^{M} \widetilde{J}^{(i)}_{\eta(F+G)}(x_k)$

$x_{k+1} = (1-\alpha_k)x_k + \alpha_k \widetilde{J}_{\eta(F+G)}(x_k)$

end for

Algorithm 7 `MLMC-FBF`(z_0, T, A, B, L_B)

Input: Initial iterate $z_0 \in \mathbb{R}^d$, subsolver `FBF` from Algorithm 5

Define $y^i = \texttt{FBF}(z_0, 2^i, \widetilde{B}, A, L_B)$ for any $i \geq 0$. Draw $I \sim \mathrm{Geom}(1/2)$

Output: $y^{\mathrm{out}} = y^0 + 2^I(y^I - y^{I-1})$ if $2^I \leq T$, otherwise $y^{\mathrm{out}} = y^0$.

We start by modifying the proof of Lemma 3.5 for the stochastic problem, which is the most important for getting the final complexity.

Lemma C.8. *Let Assumptions 1 and 3 hold. Suppose that the iterates generated by Algorithm 6 satisfy* $\mathbb{E}_k\|\widetilde{J}_{\eta(F+G)}(x_k) - J_{\eta(F+G)}(x_k)\|^2 \leq \varepsilon^2_{k,v}$ *and* $\|\mathbb{E}_k[\widetilde{J}_{\eta(F+G)}(x_k)] - J_{\eta(F+G)}(x_k)\| \leq \varepsilon_{k,b}$. *Then, we have that*

$$\frac{\alpha}{4}\sum_{k=0}^{K-1}\alpha_k\mathbb{E}\|(\mathrm{Id} - J_{\eta(F+G)})(x_k)\|^2 \leq \frac{1}{2}\|x_0 - x^\star\|^2 + \frac{3}{2}\sum_{k=0}^{K-1}\alpha_k^2\mathbb{E}[\varepsilon^2_{k,v}] + \sum_{k=0}^{K-1}\alpha_k\mathbb{E}[\|x_k - x^\star\|\varepsilon_{k,b}].$$

Proof. We proceed mostly as the proof of Lemma 3.5 apart from minor changes due to the stochastic setting such as iteration-dependent step sizes.

From Fact B.1(ii), we have that $\mathrm{Id} - J_{\eta(F+G)}$ is $\left(1 - \frac{\rho}{\eta}\right)$-star cocoercive. Recall our running notations:

$$\alpha = 1 - \frac{\rho}{\eta}, \quad R = \mathrm{Id} - J_{\eta(F+G)}, \quad \widetilde{R} = \mathrm{Id} - \widetilde{J}_{\eta(F+G)}.$$

As a result, we have the following equivalent representation of x_{k+1} (see the definition in Algorithm 6):

$$x_{k+1} = x_k - \alpha_k\widetilde{R}(x_k). \tag{C.12}$$

By α-star-cocoercivity of $R = \mathrm{Id} - J_{\eta(F+G)}$, we have

$$\langle R(x_k), x_k - x^\star\rangle \geq \alpha\|R(x_k)\|^2. \tag{C.13}$$

By a simple decomposition, we write

$$\langle R(x_k), x_k - x^\star\rangle = \langle\widetilde{R}(x_k), x_k - x^\star\rangle + \langle R(x_k) - \widetilde{R}(x_k), x_k - x^\star\rangle. \tag{C.14}$$

For the expectation of the first term on the right-hand side of (C.14), we derive that (*cf.* (B.8))

$$\begin{aligned}
\mathbb{E}\langle \widetilde{R}(x_k), x_k - x^\star\rangle &= \frac{1}{\alpha_k}\mathbb{E}\langle x_k - x_{k+1}, x_k - x^\star\rangle \\
&= \frac{1}{2\alpha_k}\mathbb{E}\left(\|x_k - x_{k+1}\|^2 + \|x_k - x^\star\|^2 - \|x_{k+1} - x^\star\|^2\right) \\
&\le \frac{1}{2\alpha_k}\mathbb{E}\left(\|x_k - x^\star\|^2 - \|x_{k+1} - x^\star\|^2\right) + \frac{3\alpha_k}{4}\mathbb{E}\|R(x_k)\|^2 + \frac{3\alpha_k}{2}\mathbb{E}\|\widetilde{R}(x_k) - R(x_k)\|^2 \\
&\le \frac{1}{2\alpha_k}\mathbb{E}\left(\|x_k - x^\star\|^2 - \|x_{k+1} - x^\star\|^2\right) + \frac{3\alpha}{4}\mathbb{E}\|R(x_k)\|^2 + \frac{3\alpha_k\mathbb{E}[\varepsilon_{k,v}^2]}{2},
\end{aligned} \tag{C.15}$$

where we used the definition of x_{k+1} from (C.12) in the first step, standard expansion $\|a-b\|^2 = \|a\|^2 - 2\langle a, b\rangle + \|b\|^2$ for the second step, the definition of x_{k+1} from (C.12) and Young's inequality in the third step, the definition of $\varepsilon_{k,v}$ with tower rule and $\alpha_k \le \alpha$ in the last step.

For the second term on the right-hand side of (C.14), we have, by Cauchy-Schwarz inequality and the definition of $\widetilde{R}$ and $\varepsilon_{k,b}$, that

$$\begin{aligned}
\mathbb{E}\langle R(x_k) - \widetilde{R}(x_k), x_k - x^\star\rangle &= \mathbb{E}[\mathbb{E}_k\langle R(x_k) - \widetilde{R}(x_k), x_k - x^\star\rangle] \\
&= \mathbb{E}\langle \mathbb{E}_k[R(x_k) - \widetilde{R}(x_k)], x_k - x^\star\rangle \\
&\le \mathbb{E}\left[\|R(x_k) - \mathbb{E}_k[\widetilde{R}(x_k)]\|\|x_k - x^\star\|\right] \\
&\le \mathbb{E}\left[\|x_k - x^\star\|\varepsilon_{k,b}\right],
\end{aligned} \tag{C.16}$$

where the first step is by tower rule and the second step is by $x_k - x^\star$ being measurable under the conditioning of $\mathbb{E}_k$.

We combine (C.15) and (C.16) in (C.14), plug in the result to (C.13) and rearrange to obtain

$$\frac{\alpha}{4}\mathbb{E}\|R(x_k)\|^2 \le \frac{1}{2\alpha_k}\mathbb{E}\left(\|x_k - x^\star\|^2 - \|x_{k+1} - x^\star\|^2\right) + \frac{3\alpha_k\mathbb{E}[\varepsilon_{k,v}^2]}{2} + \mathbb{E}[\|x_k - x^\star\|\varepsilon_{k,b}].$$

We conclude after multiplying both sides by α_k and summing for $k = 0, 1, \ldots, K-1$. □

The next lemma considers the bias and variance of the MLMC estimator and follows the same arguments as [Asi et al., 2021, Proposition 1]. The only change is that we use the algorithm FBF (see Algorithm 5 for a stochastic version) as the subsolver and we consider a strongly monotone inclusion problem rather than minimization. These do not alter the estimations significantly as can be seen in the proof.

Lemma C.9. *Under the same setting as Theorem C.1 and $T \ge 2$, for the output of Algorithm 7, we have that*

$$\begin{aligned}
\|\mathbb{E}[y^{\text{out}}] - z^\star\|^2 &\le \frac{12L/\mu\|z_0 - z^\star\|^2 + 96\sigma^2/\mu^2}{T} \\
\mathbb{E}\|y^{\text{out}} - z^\star\|^2 &\le 14(6L/\mu\|z_0 - z^\star\|^2 + 48\sigma^2/\mu^2)\log_2(T)
\end{aligned}$$

where the expected number of calls to $\widetilde{F}$ is $O(\log_2 T)$.

Proof. We argue as [Asi et al., 2021, Property 1]. The only difference is that we call Theorem C.1 which is our main solver for the strongly monotone problem.

Let us denote $i_T = \max\{i \ge 0 \colon 2^i \le T\}$. For a given event E, consider also the following notation for the characteristic function: $\mathbf{1}_E = 1$ if E is true and $\mathbf{1}_E = 0$ if E is false.

Then, we have by the definition of y^{out} in Algorithm 7 that

$$\begin{aligned}
\mathbb{E}[y^{\text{out}}] &= \mathbb{E}[y^0] + \mathbb{E}[\mathbf{1}_{\{2^I \le T\}} \cdot 2^I(y^I - y^{I-1})] \\
&= \mathbb{E}[y^0] + \sum_{i=1}^{i_T} \Pr(I = i) 2^i \mathbb{E}[y^i - y^{i-1}] \\
&= \mathbb{E}[y^0] + \mathbb{E}[y^{i_T} - y^0] \\
&= \mathbb{E}[y^{i_T}].
\end{aligned} \tag{C.17}$$

By the definition of i_T, we have that $2^{i_T} \geq \frac{T}{2}$ and hence, by Jensen's inequality and Theorem C.1, we have

$$\begin{aligned}\|\mathbb{E}[y^{i_T}] - z^\star\|^2 &\leq \mathbb{E}\|y^{i_T} - z^\star\|^2 \\ &\leq \frac{12L\|z_0 - z^\star\|^2 + 96\sigma^2/\mu}{T\mu},\end{aligned}$$

which is the claimed bound on the bias due to (C.17).

We continue with estimating the variance of y^{out}. First, Young's inequality gives that

$$\mathbb{E}\|y^{\text{out}} - z^\star\|^2 \leq 2\mathbb{E}\|y^{\text{out}} - y^0\|^2 + 2\mathbb{E}\|y^0 - z^\star\|^2. \tag{C.18}$$

We estimate the first term on the right-hand side:

$$\begin{aligned}\mathbb{E}\|y^{\text{out}} - y^0\|^2 &= \sum_{i=1}^{i_T} \Pr(I = i)\mathbb{E}\|2^i(y^i - y^{i-1})\|^2 \\ &= \sum_{i=1}^{i_T} 2^i\mathbb{E}\|y^i - y^{i-1}\|^2 \\ &\leq \sum_{i=1}^{i_T} 2^{i+1}\left(\mathbb{E}\|y^i - z^\star\|^2 + \mathbb{E}\|y^{i-1} - z^\star\|^2\right),\end{aligned} \tag{C.19}$$

where the last step is by Young's inequality.

By the definitions of y^i, y^{i-1} and Theorem C.1, we have that

$$\begin{aligned}\mathbb{E}\|y^i - z^\star\|^2 &\leq \frac{6L\|z_0 - z^\star\|^2 + 48\sigma^2/\mu}{2^i\mu}, \\ \mathbb{E}\|y^{i-1} - z^\star\|^2 &\leq \frac{6L\|z_0 - z^\star\|^2 + 48\sigma^2/\mu}{2^{i-1}\mu}.\end{aligned}$$

This gives, in view of (C.19), that

$$\mathbb{E}\|y^{\text{out}} - y^0\|^2 \leq \frac{6(6L\|z_0 - z^\star\|^2 + 48\sigma^2/\mu)}{\mu} i_T.$$

The second term on the right-hand side of (C.18) is estimated the same way by using Theorem C.1:

$$\mathbb{E}\|y^0 - z^\star\|^2 \leq \frac{6L\|z_0 - z^\star\|^2 + 48\sigma^2/\mu}{\mu}.$$

Combining the last two estimates in (C.18) gives the claimed bound on the variance after using $i_T \leq \log_2 T$.

The expected number of calls to $\widetilde{B}$ is calculated as

$$2 + 2\sum_{i=1}^{i_T} P(I = i)(2^i + 2^{i-1}) = O(1 + i_T) = O(1 + \log_2 T),$$

since each iteration of stochastic FBF uses 2 unbiased samples of F. This completes the proof. □

Let us consider $M \geq 1$ independent draws of `MLMC-FBF` and denote for i-th draw:

$$\widetilde{J}^{(i)}_{\eta(F+G)}(x_k) = \texttt{MLMC-FBF}\left(x_k, T, G, \mathrm{Id} + \eta\widetilde{F}, 1 + \eta L\right).$$

Then, we define

$$\widetilde{J}_{\eta(F+G)}(x_k) = \frac{1}{M}\sum_{i=1}^{M} \widetilde{J}^{(i)}_{\eta(F+G)}(x_k). \tag{C.20}$$

This will help us get a better control on the variance as [Asi et al., 2021, Theorem 1].

Corollary C.10. *For $\widetilde{J}_{\eta(F+G)}(x_k)$ as defined in (C.20) and under the setting of Theorem C.1, we have the bias and variance bounds given as*

$$\|\mathbb{E}_k[\widetilde{J}_{\eta(F+G)}(x_k)] - J_{\eta(F+G)}(x_k)\|^2 \le b_k^2(\|(\mathrm{Id} + J_{\eta(F+G)})(x_k)\|^2 + \sigma^2),$$
$$\mathbb{E}_k\|\widetilde{J}_{\eta(F+G)}(x_k) - J_{\eta(F+G)}(x_k)\|^2 \le v^2(\|(\mathrm{Id} + J_{\eta(F+G)})(x_k)\|^2 + \sigma^2),$$

where

$$T = \left\lceil \frac{\max\{12L/\mu, 96/\mu^2\}}{\min\{b_k^2, \frac{v^2}{2}\}} \right\rceil, \quad \text{and} \quad M = \left\lceil \frac{2\log_2 T \max\{84L/\mu, 672/\mu^2\}}{v^2} \right\rceil.$$

Each iteration makes in expectation $O(\log T \cdot M)$ calls to stochastic first-order oracle.

Proof. This proof follows the arguments in [Asi et al., 2021, Theorem 1]. The difference is that we set the values of T, M independent of $\|R(x_k)\|^2$ and σ^2, to make T, M computable, which results in these terms appearing in the bias and variance upper bounds.

We first note that $\mathbb{E}_k[\widetilde{J}_{\eta(F+G)}(x_k)] = \mathbb{E}_k[\widetilde{J}^{(1)}_{\eta(F+G)}(x_k)]$. We next have by direct expansion that

$$\begin{aligned}\mathbb{E}_k\|\widetilde{J}_{\eta(F+G)}(x_k) - J_{\eta(F+G)}(x_k)\|^2 = \frac{1}{M}\mathbb{E}_k\|\widetilde{J}^{(1)}_{\eta(F+G)}(x_k) - J_{\eta(F+G)}(x_k)\|^2 \\ + \left(1 - \frac{1}{M}\right)\|\mathbb{E}_k[\widetilde{J}^{(1)}_{\eta(F+G)}(x_k)] - J_{\eta(F+G)}(x_k)\|^2,\end{aligned}$$

since $\widetilde{J}^{(i)}_{\eta(F+G)}$ are independent draws of the same estimator.

By applying the identity $\mathbb{E}\|X - \mathbb{E}X\|^2 = \mathbb{E}\|X\|^2 - \|\mathbb{E}X\|^2$ with $X = \widetilde{J}^{(1)}_{\eta(F+G)}(x_k) - J_{\eta(F+G)}(x_k)$, we obtain

$$\begin{aligned}\mathbb{E}_k\|\widetilde{J}_{\eta(F+G)}(x_k) - J_{\eta(F+G)}(x_k)\|^2 = \frac{1}{M}\mathbb{E}_k\|\widetilde{J}^{(1)}_{\eta(F+G)}(x_k) - \mathbb{E}_k[\widetilde{J}^{(1)}_{\eta(F+G)}(x_k)]\|^2 \\ + \|\mathbb{E}_k[\widetilde{J}^{(1)}_{\eta(F+G)}(x_k)] - J_{\eta(F+G)}(x_k)\|^2. \qquad \text{(C.21)}\end{aligned}$$

On the one hand, the fact $\mathbb{E}\|X - \mathbb{E}X\|^2 \le \mathbb{E}\|X\|^2$ gives

$$\mathbb{E}_k\|\widetilde{J}^{(1)}_{\eta(F+G)}(x_k) - \mathbb{E}_k[\widetilde{J}^{(1)}_{\eta(F+G)}(x_k)]\|^2 \le \mathbb{E}_k\|\widetilde{J}^{(1)}_{\eta(F+G)}(x_k) - J_{\eta(F+G)}(x_k)\|^2. \qquad \text{(C.22)}$$

On the other hand, the bounds in Lemma C.9 gives, after substituting $z_0 = x_k$ and $z^\star = J_{\eta(F+G)}(x_k)$ that

$$\|\mathbb{E}_k[\widetilde{J}^{(1)}_{\eta(F+G)}(x_k)] - J_{\eta(F+G)}(x_k)\|^2 \le \frac{12L/\mu\|(\mathrm{Id} + J_{\eta(F+G)})(x_k)\|^2 + 96\sigma^2/\mu^2}{T}, \qquad \text{(C.23a)}$$
$$\mathbb{E}_k\|\widetilde{J}^{(1)}_{\eta(F+G)}(x_k) - J_{\eta(F+G)}(x_k)\|^2 \le \left(84L/\mu\|(\mathrm{Id} + J_{\eta(F+G)})(x_k)\|^2 + 672\sigma^2/\mu^2\right)\log_2 T. \qquad \text{(C.23b)}$$

Using $\mathbb{E}_k[\widetilde{J}_{\eta(F+G)}(x_k)] = \mathbb{E}_k[\widetilde{J}^{(1)}_{\eta(F+G)}(x_k)]$ gives the bias bound after using the definition of T and (C.23a)

Plugging in (C.23b) and (C.22) in (C.21) gives the variance bound after substituting the values of T and M. □

C.3.1 Proof for Corollary 4.4

Corollary 4.4 is essentially the summary of the results proven below.

Let us remark the recent work [Bravo and Cominetti, 2024, Corollary 5.4] that studied stochastic KM iteration for nonexpansive operators assumes access to an unbiased oracle and get the complexity $\widetilde{O}(\varepsilon^{-4})$. As mentioned in Section 4.2, this corresponds to requiring unbiased samples of $J_{\eta(F+G)}$ in our setting, which is difficult due to the definition of the resolvent. We get the same complexity up to logarithmic factors without access to unbiased samples of $J_{\eta(F+G)}$, which we go around by using the MLMC technique. We also do not require nonexpansiveness from $J_{\eta(F+G)}$ and work with star-conic nonexpansiveness.

Theorem C.11. *Let Assumptions 1, 3, and 4 hold. Consider Algorithm 6 with $\eta < \frac{1}{L}$ and $\rho < \eta$. Then, we have for $K \geq 1$ that*

$$\mathbb{E}_{x^{\text{out}} \sim \text{Unif}\{x_0,\dots,x_{K-1}\}}[\mathbb{E}\|(\text{Id} - J_{\eta(F+G)})(x^{\text{out}})\|^2] \leq \frac{64(\|x_0 - x^\star\|^2 + \alpha^2\sigma^2)\log(K+3)}{\alpha^2\sqrt{K}},$$

where $\alpha = 1 - \frac{\rho}{\eta}$. Each iteration makes, in expectation, $O(\log^2(k+2))$ calls to stochastic oracle $\widetilde{B}$ and resolvent of A. Hence to obtain $\mathbb{E}\|(\text{Id} - J_{\eta(F+G)})(x^{\text{out}})\| \leq \varepsilon$, we have the expected stochastic first-order complexity $\widetilde{O}(\varepsilon^{-4})$.

The main reason for the length of the following proof is the lack of boundedness of (x_k). In particular, proving this theorem is rather straightforward when we assume a bounded domain. We have to handle the complications without this assumption. There are also additional difficulties that arise because we are making sure that the inputs to `MLMC-FBF` will not involve unknown quantities such as $\|x_0 - x^\star\|$ or σ to run the algorithm. These are, for example, used in Chen and Luo [2022] for setting the parameters. Because of this reason, the bounds for $\varepsilon_{k,v}$ and $\varepsilon_{k,b}$ involve $\|(\text{Id} + J_{\eta(F+G)})(x_k)\|$ and σ^2.

The main reason for the difficulty here is $\|(\text{Id} + J_{\eta(F+G)})(x_k)\|^2$, since we do not have a uniform bound on this quantity, unlike σ^2 and this term appears in many terms. We will carry these terms coming from the MLMC bounds to get a recursion involving the sum of $\|(\text{Id} + J_{\eta(F+G)})(x_k)\|^2$ for different ranges on both sides. We then go around the issue of lacking of a bound on (x_k) by using an inductive argument on $\sum_{k=0}^{K} \|(\text{Id} + J_{\eta(F+G)})(x_k)\|^2$.

Proof of Theorem C.11. Recall our running notations:

$$\alpha = 1 - \frac{\rho}{\eta}, \quad R = \text{Id} - J_{\eta(F+G)}, \quad \widetilde{R} = \text{Id} - \widetilde{J}_{\eta(F+G)}.$$

We start by following the proof of Lemma B.5. By α-star-cocoercivity of $\text{Id} - J_{\eta(F+G)}$ and $\alpha \geq \alpha_k$ (which gives that $J_{\eta(F+G)}$ is $\frac{1}{\alpha_k}$-star-conic nonexpansive), we can use property (B.2) derived in Lemma B.2 to obtain

$$\left\|(1-\alpha_k)x_k + \alpha_k J_{\eta(F+G)}(x_k) - x^\star\right\| \leq \|x_k - x^\star\|$$

and by the definition of x_{k+1}, we get

$$\begin{aligned}\|x_{k+1} - x^\star\| &\leq \|(1-\alpha_k)x_k + \alpha_k J_{\eta(F+G)}(x_k) - x^\star\| + \alpha_k\|\widetilde{J}_{\eta(F+G)}(x_k) - J_{\eta(F+G)}(x_k)\| \\ &\leq \|x_k - x^\star\| + \alpha_k\|\widetilde{J}_{\eta(F+G)}(x_k) - J_{\eta(F+G)}(x_k)\|. \end{aligned} \tag{C.24}$$

Summing the inequality for $0, \dots, k-1$ gives

$$\begin{aligned}&\|x_k - x^\star\| \leq \|x_0 - x^\star\| + \sum_{i=0}^{k-1} \alpha_i\|\widetilde{J}_{\eta(F+G)}(x_i) - J_{\eta(F+G)}(x_i)\| \\ \implies &\mathbb{E}\|x_k - x^\star\|^2 \leq 2\mathbb{E}\|x_0 - x^\star\|^2 + 2k\sum_{i=0}^{k-1} \alpha_i^2\mathbb{E}\|\widetilde{J}_{\eta(F+G)}(x_i) - J_{\eta(F+G)}(x_i)\|^2, \end{aligned} \tag{C.25}$$

where we first squared both sides, used Young's inequality and then took expectation.

We continue by restating the result of Lemma C.8 after applying Young's inequality on the last term to obtain

$$\begin{aligned}\frac{\alpha}{4}\sum_{k=0}^{K-1} \alpha_k\mathbb{E}\|(\text{Id} - J_{\eta(F+G)})(x_k)\|^2 \leq \frac{1}{2}\|x_0 - x^\star\|^2 + \frac{3}{2}\sum_{k=0}^{K-1} \alpha_k^2\mathbb{E}[\varepsilon_{k,v}^2] \\ + \sum_{k=0}^{K-1}\left(\frac{\alpha_k^2}{2\alpha^2(k+1)}\mathbb{E}\|x_k - x^\star\|^2 + \frac{(k+1)\alpha^2}{2}\mathbb{E}[\varepsilon_{k,b}^2]\right). \end{aligned} \tag{C.26}$$

We now estimate the second and third terms on the right-hand side. By using Corollary C.10 and the definition of $R(x_k)$, $\alpha_k = \frac{\alpha}{\sqrt{k+2}\log(k+2)} < \frac{\alpha}{\sqrt{2}\log 3}$ and using $v^2 = \frac{1}{60} \leq \frac{\sqrt{2}\log 3}{24}$, we obtain

$$\begin{aligned}\frac{3}{2}\sum_{k=0}^{K-1}\alpha_k^2\mathbb{E}[\varepsilon_{k,v}^2] &\leq \frac{3}{2}\sum_{k=0}^{K-1}\alpha_k^2 v^2\left(\mathbb{E}\|R(x_k)\|^2+\sigma^2\right)\\ &\leq \frac{\alpha\alpha_{K-1}}{16}(\mathbb{E}\|R(x_{K-1})\|^2+\sigma^2)+\frac{3}{2}\sum_{k=0}^{K-2}\alpha_k^2 v^2\left(\mathbb{E}\|R(x_k)\|^2+\sigma^2\right). \end{aligned} \tag{C.27}$$

We continue with the first part of the third term on the right-hand side of (C.26) and bound it using (C.25):

$$\begin{aligned}\sum_{k=0}^{K-1}\frac{\alpha_k^2}{2\alpha^2(k+1)}\mathbb{E}\|x_k-x^\star\|^2 &\leq \frac{1}{2}\|x_0-x^\star\|^2+\sum_{k=1}^{K-1}\frac{\alpha_k^2}{2\alpha^2(k+1)}\mathbb{E}\|x_k-x^\star\|^2\\ &\leq \frac{1}{2}\|x_0-x^\star\|^2+\sum_{k=1}^{K-1}\frac{\alpha_k^2}{2\alpha^2(k+1)}\left(2\|x_0-x^\star\|^2+2k\sum_{i=0}^{k-1}\alpha_i^2\mathbb{E}[\varepsilon_{i,v}^2]\right). \end{aligned} \tag{C.28}$$

We focus on the last term here to get

$$\begin{aligned}\sum_{k=1}^{K-1}\frac{\alpha_k^2}{2\alpha^2(k+1)}\cdot 2k\sum_{i=0}^{k-1}\alpha_i^2\mathbb{E}[\varepsilon_{i,v}^2] &= \frac{1}{\alpha^2}\sum_{i=0}^{K-2}\sum_{k=i+1}^{K-1}\frac{k}{k+1}\alpha_k^2\alpha_i^2\mathbb{E}[\varepsilon_{i,v}^2]\\ &\leq \frac{1}{\alpha^2}\left(\sum_{k=0}^{K-1}\alpha_k^2\right)\sum_{i=0}^{K-2}\alpha_i^2\mathbb{E}[\varepsilon_{i,v}^2]\\ &\leq \frac{1}{\alpha^2}\left(\sum_{k=0}^{K-1}\alpha_k^2\right)\sum_{i=0}^{K-2}\alpha_i^2 v^2\left(\mathbb{E}\|R(x_i)\|^2+\sigma^2\right), \end{aligned}$$

where the last step used Corollary C.10.

Plugging in back to (C.28) gives

$$\begin{aligned}\sum_{k=0}^{K-1}\frac{\alpha_k^2}{2\alpha^2(k+1)}\mathbb{E}\|x_k-x^\star\|^2 &\leq \left(\frac{1}{2}+\sum_{k=1}^{K-1}\frac{\alpha_k^2}{\alpha^2(k+1)}\right)\|x_0-x^\star\|^2\\ &\quad+\left(\sum_{k=0}^{K-1}\frac{\alpha_k^2}{\alpha^2}\right)\sum_{i=0}^{K-2}\alpha_i^2 v^2\left(\mathbb{E}\|R(x_i)\|^2+\sigma^2\right). \end{aligned} \tag{C.29}$$

We finally estimate the second part of the third term on the right-hand side of (C.26) by using Corollary C.10:

$$\alpha^2\sum_{k=0}^{K-1}\frac{k+1}{2}\mathbb{E}[\varepsilon_{k,b}^2] \leq \alpha^2\sum_{k=0}^{K-1}(k+1)b_k^2\mathbb{E}[\|R(x_k)\|^2+\sigma^2].$$

We use the setting $b_k^2 = \frac{\alpha_k}{120\alpha(k+1)}$ and $b_{K-1}^2 < \frac{\alpha_{K-1}}{16\alpha K}$ to obtain

$$\alpha^2\sum_{k=0}^{K-1}\frac{k+1}{2}\mathbb{E}[\varepsilon_{k,b}^2] \leq \frac{\alpha\alpha_{K-1}}{16}(\mathbb{E}\|R(x_{K-1})\|^2+\sigma^2)+\alpha^2\sum_{k=0}^{K-2}(k+1)b_k^2\mathbb{E}[\|R(x_k)\|^2+\sigma^2]. \tag{C.30}$$

We collect (C.27), (C.29), and (C.30) in (C.26) to get

$$\begin{aligned}\frac{\alpha}{8}\sum_{k=0}^{K-1}\alpha_k\mathbb{E}\|R(x_k)\|^2 &\leq \left(1+\sum_{k=1}^{K-1}\frac{\alpha_k^2}{\alpha^2(k+1)}\right)\|x_0-x^\star\|^2+\frac{\alpha\alpha_{K-1}}{8}\sigma^2\\ &\quad+\left(\frac{3}{2}+\sum_{k=0}^{K-1}\frac{\alpha_k^2}{\alpha^2}\right)\sum_{k=0}^{K-2}\alpha_k^2 v^2\left(\mathbb{E}\|R(x_k)\|^2+\sigma^2\right)\\ &\quad+\alpha^2\sum_{k=0}^{K-2}(k+1)b_k^2\mathbb{E}[\|R(x_k)\|^2+\sigma^2]. \end{aligned} \tag{C.31}$$

We now show by induction that

$$\alpha \sum_{k=0}^{K-1} \alpha_k \mathbb{E}\|R(x_k)\|^2 \leq C\left(\|x_0 - x^\star\|^2 + \alpha^2\sigma^2\right) \quad \forall K \geq 1, \tag{C.32}$$

for some C to be determined. Let us set

$$\alpha_k = \frac{\alpha}{\sqrt{k+2}\log(k+3)},$$

which gives

$$\sum_{k=0}^{K-1} \frac{\alpha_k^2}{\alpha^2} < 3, \quad \sum_{k=0}^{K-1} \frac{\alpha_k^2}{\alpha^2(k+1)} < 0.25, \quad \alpha_k \leq \alpha \ \forall k \geq 0.$$

With these estimations and $\alpha < 1$, (C.31) becomes

$$\begin{aligned}\frac{\alpha}{8} \sum_{k=0}^{K-1} \alpha_k \mathbb{E}\|R(x_k)\|^2 \leq\ & 1.25\|x_0 - x^\star\|^2 + \alpha^2\sigma^2 \\ & + 4.5 \sum_{k=0}^{K-2} v^2(\alpha\alpha_k \mathbb{E}\|R(x_k)\|^2 + \alpha^2\sigma^2) \\ & + \alpha \sum_{k=0}^{K-2} \frac{(k+1)b_k^2}{\alpha_k} \mathbb{E}[\alpha\alpha_k\|R(x_k)\|^2 + \alpha^2\sigma^2].\end{aligned} \tag{C.33}$$

Let us set

$$C = 32, \quad b_k^2 = \frac{\alpha_k}{120\alpha(k+1)}, \quad v^2 = \frac{1}{60}$$

and use the inductive assumption $\alpha \sum_{k=0}^{K-2} \alpha_k \mathbb{E}\|R(x_k)\|^2 \leq 32(\|x_0 - x^\star\|^2 + \alpha^2\sigma^2)$ in (C.33) to obtain

$$\frac{\alpha}{8} \sum_{k=0}^{K-1} \alpha_k \mathbb{E}\|R(x_k)\|^2 \leq 4(\|x_0 - x^\star\|^2 + \alpha^2\sigma^2),$$

which verifies $\alpha \sum_{k=0}^{K-1} \alpha_k \mathbb{E}\|R(x_k)\|^2 \leq 32(\|x_0 - x^\star\|^2 + \alpha^2\sigma^2)$.

For the base case, we use $\alpha_0 = \frac{\alpha}{\sqrt{2}\log 3} < 1$ and α^{-1}-star-Lipschitzness of $R = \mathrm{Id} - J_{\eta(F+G)}$ to get $\alpha\alpha_0\|R(x_0)\|^2 \leq \|x_0 - x^\star\|^2$. This establishes the base case and completes the induction.

Finally, in view of Corollary C.10, and definitions of b_k, v, each iteration makes expected number of calls $O(\log^2(k+1))$. By using $\alpha_k \geq \alpha_K = \frac{\alpha}{\sqrt{K+2}\log(K+3)}$ in (C.32) with $C = 32$ and multiplying both sides by $\frac{1}{K\alpha_K}$ and using $\frac{\sqrt{K+2}}{K} \leq \frac{2}{\sqrt{K}}$ which is true for $K \geq 1$, we get the claimed rate result. By using the expected cost of each iteration, we also get the final expected stochastic first-order complexity result. □

D Additional Remarks on Related Work

There exist a line of works that attempted to construct local estimation of Lipschitz constants to offer an improved range for ρ depending on the curvature [Pethick et al., 2022, Alacaoglu et al., 2023]. However, these results cannot bring global improvements in the worst-case range of ρ where the limit for ρ is still $\frac{1}{2L}$. This is because it is easy to construct examples where the local Lipschitz constants are the same as the global Lipschitz constant.

The work Hajizadeh et al. [2023] gets linear rate of convergence for interaction dominant problems which is shown to be closely related to cohypomonoconity, see Example 1.2. One important difference is that cohypomonotonicity is equivalent to α interaction dominance with $\alpha \geq 0$ whereas Hajizadeh et al. [2023] requires $\alpha > 0$ for linear convergence. This is an important difference because we know that cohypomonotonicity relaxes monotonicity and that that even monotonicity is not sufficient for linear convergence. Even

for monotone problems $O(\varepsilon^{-1})$ is the optimal first-order oracle complexity (see, e.g., Yoon and Ryu [2021]) and hence it is also optimal with cohypomonotonicity.

In the literature for fixed point iterations, several works considered inexact Halpern or KM iterations without characterizing explicit first-order complexity results, see for example Leuştean and Pinto [2021], Bartz et al. [2022], Kohlenbach [2022], Combettes and Pennanen [2004]. In particular, Bartz et al. [2022] used conic nonexpansiveness to analyze KM iteration. The dependence of the range of ρ on L arises when we start characterizing the first-order complexity. This is the reason these works have not been included in comparisons in Table 1.

For the stochastic cohypomonotone problems, the best complexity result to our knowledge is due to Chen and Luo [2022]. This paper can obtain the optimal complexity $\widetilde{O}(\varepsilon^{-2})$ with cohypomonotone stochastic problems with a 6-loop algorithm using many carefully designed regularization techniques, extending the work of Allen-Zhu [2018] that focused on minimization. Some disadvantages of this approach compared to ours: *(i)* the bound for cohypomonotonicity is $\rho \leq \frac{1}{2L}$; *(ii)* the algorithm needs estimates of variance upper bound σ^2 and, more importantly, $\|x^0 - x^\star\|^2$; *(iii)* the result is only given for unconstrained problems, which also makes it difficult to assume a bounded domain since there is no guarantee a priori for the iterates to stay bounded for an unconstrained problem. Given that the 6-loop algorithm and analysis of Chen and Luo [2022] is rather complicated, it is not clear to us if their arguments generalize to constraints or if the other drawbacks can be alleviated.

The work Tran-Dinh and Luo [2023] focused on problems with ρ-weakly MVI solutions for $\rho < \frac{1}{8L}$ and derived $O(\varepsilon^{-2})$ for a randomized coordinate algorithm. Due to randomization, the complexity result in this work holds for the expectation of the optimality measure. Because of the coordinatewise updates, the problem focused in this work is deterministic, similar to the setup in Section 3.

D.1 Clarifications about Table 2

Since the complexity results have not been written explicitly in some of the references, we provide details on how we computed the complexities that we report for the existing works.

Choudhury et al. [2023]: We use Theorem 4.5 in this corresponding paper to see that squared operator norm is upper bounded by $O(K^{-1})$. To make the operator norm smaller than ε, the order of K is ε^{-2}. The batch-size has order K and hence the total number of oracle calls is $O(K^2) = O(\varepsilon^{-4})$.

Böhm et al. [2022]: We use Theorem 3.3 in this corresponding paper. The paper stated that to make the squared operator norm smaller than ε, number of iterations is $O(\varepsilon^{-2})$ and the batch size is $O(\varepsilon^{-3})$. This gives complexity $O(\varepsilon^{-3})$ for making the *squared* operator norm smaller than ε. Hence, to make the operator norm smaller than ε, the complexity is $O(\varepsilon^{-6})$.

[Pethick et al., 2023b]: *(i)* For "best rate" result, we use Corollary E.3(i) in this corresponding paper. The dominant term in the bound for the squared residual is $O(K^{-1})$. Hence to make the norm of the residual smaller than ε (equivalently, the squared norm smaller than ε^{-2}), one needs K to be of the order ε^{-2}. Then, the squared variance is assumed to decrease at the order of k^2 which requires the batch size at iteration k to be k^2. Then the complexity is upper bounded by $\sum_{k=1}^{K} \tau k^2 = \widetilde{O}(K^3) = \widetilde{O}(\varepsilon^{-6})$, *(ii)* for the "last iterate", we use the Corollary E.3(ii) given in the paper to see that the dominant term in the bound of the squared residual is $O\left(\frac{1}{\sqrt{K}}\right)$. To make the squared residual smaller than ε^2, this means K is of the order ε^{-4}. The squared variance is assumed to decrease at the rate k^3 which requires a batch size of k^3 at iteration k. Then, with the same calculation as before, the complexity of stochastic first-order oracles to make the residual less than ε is $\sum_{k=1}^{K} \tau k^3 = \widetilde{O}(K^4) = \widetilde{O}(\varepsilon^{-16})$.

Acknowledgments

This work was supported in part by the NSF grant 2023239, the NSF grant 2224213, the AFOSR award FA9550-21-1-0084, National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2019R1A5A1028324, 2022R1C1C1003940), and the Samsung Science & Technology Foundation grant (No. SSTF-BA2101-02).

References

A. Alacaoglu, A. Böhm, and Y. Malitsky. Beyond the golden ratio for variational inequality algorithms. *Journal of Machine Learning Research*, 24(172):1–33, 2023.

Z. Allen-Zhu. How to make the gradients small stochastically: Even faster convex and nonconvex sgd. *Advances in Neural Information Processing Systems*, 31, 2018.

Anonymous. Semi-anchored gradient methods for nonconvex-nonconcave minimax problems, 2024a. URL https://openreview.net/forum?id=rmLTwKGiSP.

Anonymous. Weaker MVI condition: Extragradient methods with multi-step exploration. In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=RNGUbTYSjk.

Y. Arjevani, Y. Carmon, J. C. Duchi, D. J. Foster, N. Srebro, and B. Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199(1-2):165–214, 2023.

H. Asi, Y. Carmon, A. Jambulapati, Y. Jin, and A. Sidford. Stochastic bias-reduced gradient methods. *Advances in Neural Information Processing Systems*, 34:10810–10822, 2021.

S. Bartz, M. N. Dao, and H. M. Phan. Conical averagedness and convergence analysis of fixed point algorithms. *Journal of Global Optimization*, 82(2):351–373, 2022.

H. H. Bauschke and P. L. Combettes. Convex analysis and monotone operator theory in hilbert spaces. *CMS Books in Mathematics*, 2017.

H. H. Bauschke, W. M. Moursi, and X. Wang. Generalized monotone operators and their averaged resolvents. *Mathematical Programming*, 189:55–74, 2021.

J. H. Blanchet and P. W. Glynn. Unbiased monte carlo for optimization and functions of expectations via multi-level randomization. In *2015 Winter Simulation Conference (WSC)*, pages 3656–3667. IEEE, 2015.

A. Böhm. Solving nonconvex-nonconcave min-max problems exhibiting weak minty solutions. *Transactions on Machine Learning Research*, 2022.

M. Bravo and R. Cominetti. Stochastic fixed-point iterations for nonexpansive maps: Convergence and error bounds. *SIAM Journal on Control and Optimization*, 62(1):191–219, 2024.

A. Böhm, M. Sedlmayer, E. R. Csetnek, and R. I. Bot. Two steps at a time—taking gan training in stride with tseng's method. *SIAM Journal on Mathematics of Data Science*, 4(2):750–771, 2022.

X. Cai, A. Alacaoglu, and J. Diakonikolas. Variance reduced halpern iteration for finite-sum monotone inclusions. In *The Twelfth International Conference on Learning Representations*, 2024.

Y. Cai and W. Zheng. Accelerated single-call methods for constrained min-max optimization. In *The Eleventh International Conference on Learning Representations*, 2022.

Y. Cai, A. Oikonomou, and W. Zheng. Accelerated algorithms for monotone inclusions and constrained nonconvex-nonconcave min-max optimization. *arXiv preprint arXiv:2206.05248*, 2022.

L. Chen and L. Luo. Near-optimal algorithms for making the gradient small in stochastic minimax optimization. *arXiv preprint arXiv:2208.05925*, 2022.

S. Choudhury, E. Gorbunov, and N. Loizou. Single-call stochastic extragradient methods for structured non-monotone variational inequalities: Improved analysis under weaker conditions. In *Advances in Neural Information Processing Systems*, 2023.

P. L. Combettes. Quasi-fejérian analysis of some optimization algorithms. In *Studies in Computational Mathematics*, volume 8, pages 115–152. Elsevier, 2001.

P. L. Combettes and T. Pennanen. Generalized mann iterates for constructing fixed points in hilbert spaces. *Journal of Mathematical Analysis and Applications*, 275(2):521–536, 2002.

P. L. Combettes and T. Pennanen. Proximal methods for cohypomonotone operators. *SIAM journal on control and optimization*, 43(2):731–742, 2004.

C. D. Dang and G. Lan. On the convergence properties of non-euclidean extragradient methods for variational inequalities with generalized monotone operators. *Computational Optimization and applications*, 60:277–310, 2015.

M. N. Dao and H. M. Phan. Adaptive douglas–rachford splitting algorithm for the sum of two operators. *SIAM Journal on Optimization*, 29(4):2697–2724, 2019.

C. Daskalakis, D. J. Foster, and N. Golowich. Independent policy gradient methods for competitive reinforcement learning. *Advances in neural information processing systems*, 33:5527–5540, 2020.

C. Daskalakis, S. Skoulakis, and M. 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.

J. Diakonikolas. Halpern iteration for near-optimal and parameter-free monotone inclusion and strong solutions to variational inequalities. In *Conference on Learning Theory*, pages 1428–1451. PMLR, 2020.

J. Diakonikolas, C. Daskalakis, and M. I. Jordan. Efficient methods for structured nonconvex-nonconcave min-max optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 2746–2754. PMLR, 2021.

R. Dorfman and K. Y. Levy. Adapting to mixing time in stochastic optimization with markovian data. In *International Conference on Machine Learning*, pages 5429–5446. PMLR, 2022.

F. Facchinei and J.-S. Pang. *Finite-dimensional variational inequalities and complementarity problems*. Springer, 2003.

M. B. Giles. Multilevel monte carlo path simulation. *Operations research*, 56(3):607–617, 2008.

P. Giselsson and W. M. Moursi. On compositions of special cases of lipschitz continuous operators. *Fixed Point Theory and Algorithms for Sciences and Engineering*, 2021(1):1–38, 2021.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

E. Gorbunov, A. Taylor, S. Horváth, and G. Gidel. Convergence of proximal point and extragradient-based methods beyond monotonicity: the case of negative comonotonicity. In *International Conference on Machine Learning*, pages 11614–11641. PMLR, 2023.

B. Grimmer, H. Lu, P. Worah, and V. Mirrokni. The landscape of the proximal point method for nonconvex–nonconcave minimax optimization. *Mathematical Programming*, 201(1-2):373–407, 2023.

C. Groetsch. A note on segmenting mann iterates. *Journal of Mathematical Analysis and Applications*, 40(2): 369–372, 1972.

S. Hajizadeh, H. Lu, and B. Grimmer. On the linear convergence of extragradient methods for nonconvex–nonconcave minimax problems. *INFORMS Journal on Optimization*, 2023.

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

Y.-G. Hsieh, F. Iutzeler, J. Malick, and P. Mertikopoulos. On the convergence of single-call stochastic extra-gradient methods. *Advances in Neural Information Processing Systems*, 32, 2019.

Y.-P. Hsieh, P. Mertikopoulos, and V. Cevher. The limits of min-max optimization algorithms: Convergence to spurious non-critical sets. In *International Conference on Machine Learning*, pages 4337–4348. PMLR, 2021.

Y. Hu, X. Chen, and N. He. On the bias-variance-cost tradeoff of stochastic optimization. *Advances in Neural Information Processing Systems*, 34:22119–22131, 2021.

D. Kim. Accelerated proximal point method for maximally monotone operators. *Mathematical Programming*, 190(1-2):57–87, 2021.

U. 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.

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

M. A. Krasnosel'skii. Two remarks on the method of successive approximations. *Uspekhi matematicheskikh nauk*, 10(1):123–127, 1955.

G. Lan. Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes. *Mathematical programming*, 198(1):1059–1106, 2023.

S. Lee and D. Kim. Fast extra gradient methods for smooth structured nonconvex-nonconcave minimax problems. *Advances in Neural Information Processing Systems*, 34:22588–22600, 2021.

L. Leuştean and P. Pinto. Quantitative results on a halpern-type proximal point algorithm. *Computational Optimization and Applications*, 79(1):101–125, 2021.

D. Levy, Y. Carmon, J. C. Duchi, and A. Sidford. Large-scale methods for distributionally robust optimization. *Advances in Neural Information Processing Systems*, 33:8847–8860, 2020.

F. Lieder. On the convergence rate of the halpern-iteration. *Optimization letters*, 15(2):405–418, 2021.

N. Loizou, H. Berard, G. Gidel, I. Mitliagkas, and S. Lacoste-Julien. Stochastic gradient descent-ascent and consensus optimization for smooth games: Convergence analysis under expected co-coercivity. *Advances in Neural Information Processing Systems*, 34:19095–19108, 2021.

A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

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

W. R. Mann. Mean value methods in iteration. *Proceedings of the American Mathematical Society*, 4(3): 506–510, 1953.

J. v. Neumann. A model of general economic equilibrium. *The Review of Economic Studies*, 13(1):1–9, 1945.

T. Pethick, P. Patrinos, O. Fercoq, V. Cevher, and P. Latafat. Escaping limit cycles: Global convergence for constrained nonconvex-nonconcave minimax problems. In *International Conference on Learning Representations*, 2022.

T. Pethick, O. Fercoq, P. Latafat, P. Patrinos, and V. Cevher. Solving stochastic weak minty variational inequalities without increasing batch size. In *International Conference on Learning Representations*, 2023a.

T. Pethick, W. Xie, and V. Cevher. Stable nonconvex-nonconcave training via linear interpolation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023b.

E. K. Ryu and W. Yin. *Large-scale convex optimization: algorithms & analyses via monotone operators*. Cambridge University Press, 2022.

S. Sabach and S. Shtern. A first order method for solving convex bilevel optimization problems. *SIAM Journal on Optimization*, 27(2):640–660, 2017.

Q. Tran-Dinh. Sublinear convergence rates of extragradient-type methods: A survey on classical and recent developments. *arXiv preprint arXiv:2303.17192*, 2023.

Q. Tran-Dinh and Y. Luo. Randomized block-coordinate optimistic gradient algorithms for root-finding problems. *arXiv preprint arXiv:2301.03113*, 2023.

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

T. Yoon and E. K. Ryu. Accelerated algorithms for smooth convex-concave minimax problems with o (1/kˆ 2) rate on squared gradient norm. In *International Conference on Machine Learning*, pages 12098–12109. PMLR, 2021.

T. Yoon and E. K. Ryu. Accelerated minimax algorithms flock together. *arXiv preprint arXiv:2205.11093*, 2022.