Regularized Operator Extrapolation Method For Stochastic Hierarchical Variational Inequality Problems

Mohammad Khalafi **Digvijay Boob**
Department of Operations Research and Engineering Management,
Southern Methodist University, Dallas, TX, USA
{mkhalafi, dboob}@smu.edu

Abstract

The hierarchical variational inequality (HVI) problem is a broad framework covering optimal equilibrium selection and equilibrium problems with equilibrium constraints (EPECs). We propose Regularized Operator Extrapolation (R-OpEx), a single-loop first-order algorithm for smooth and nonsmooth HVIs with stochastic monotone operators. R-OpEx combines Tikhonov regularization with operator extrapolation, requires only one operator evaluation per iteration, and tracks a single sequence of iterates. We show that R-OpEx obtains an ϵ-solution in $\mathcal{O}(\epsilon^{-4})$ iterations for nonsmooth stochastic HVIs. If the inner operator is smooth and stochastic, we show an improved complexity of $\mathcal{O}(\epsilon^{-2})$ for the outer level operator while maintaining $\mathcal{O}(\epsilon^{-4})$ complexity for the inner level operator. For a smooth deterministic inner level operator, the complexity reduces to $\mathcal{O}(\epsilon^{-2})$. Finally, we improve the complexities substantially when the outer level is strongly monotone. To our knowledge, this is the first work to establish such guarantees for nonsmooth stochastic HVIs. We validate our results through numerical studies.

1 INTRODUCTION

The variational inequality (VI) problem associated with a monotone operator $\tilde{F}$ over a convex set $\tilde{X}$, denoted as VI$(\tilde{F}, \tilde{X})$, needs to find $x^* \in \tilde{X}$ such that

$$\langle \tilde{F}(x^*), x - x^* \rangle \geq 0, \quad \forall x \in \tilde{X}. \tag{1}$$

In this paper, we study the following monotone hierarchical VI (HVI) problem:

$$\text{Find } x^* \in X_F^* : \langle H(x^*), x - x^* \rangle \geq 0, \quad \forall x \in X_F^*,$$
$$\text{where } X_F^* := \{x_F^* \in X : \langle F(x_F^*), y - x_F^* \rangle \geq 0, \forall y \in X\}. \tag{2}$$

where $X \subseteq \mathbb{R}^n$ is a closed convex set, $H : X \to \mathbb{R}^n$ and $F : X \to \mathbb{R}^n$ are possibly discontinuous monotone operators. We denote *outer* and *inner* VI problems as VI(H, X_F^*) and VI(F, X), respectively. It is easy to see that X_F^* is the solution set of VI(F, X).

1.1 Motivation

Generalized Nash Equilibrium: Many real-world applications in game theory and optimization can be modeled as HVIs. A prominent example is the *generalized Nash equilibrium problem* (GNEP) that captures many applications such as Multi-agent reinforcement learning (MARL) and fairness-constrained optimization. For player $i \in [d]$, we form the GNEP

$$\min_{x^i} g_i(x^i, x^{-i}) \quad \text{s.t. } x^i \in X_i(x^{-i}),$$

where $X_i(x^{-i})$ denotes the feasible strategy set of player i given the opponents' strategies. The joint feasible set is $X(x) = \prod_{i=1}^d X_i(x^{-i})$, and $\bar{x} = (\bar{x}^1, \ldots, \bar{x}^d)$ is a Nash equilibrium if

$$g_i(\bar{x}^i, \bar{x}^{-i}) \leq g_i(x^i, \bar{x}^{-i}), \quad \forall i \in [d], \forall x^i \in X_i(\bar{x}^{-i}).$$

Consider a GNEP with shared nonlinear constraints, where the feasible set is

$$S := \{x \in \mathbb{R}^n \mid F(x) = 0\},$$

for a continuous monotone mapping F. Then, S coincides with the solution set of VI$(F, \mathbb{R}^n)$ which is convex by (Facchinei and Pang, 2003, Thm. 2.3.5).

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

The convexity of S implies that each feasible set

$$X_i(x^{-i}) := \{x^i \in \mathbb{R}^{n_i} \mid (x^i, x^{-i}) \in S\}$$

are convex for fixed x^{-i}. Hence, the GNEP is jointly convex. Under appropriate convexity properties of g_i, we can show that the gradient mapping

$$H(x) := \left[\nabla_{x^1} g_1(x)^\top, \ldots, \nabla_{x^d} g_d(x)^\top\right]^\top,$$

is monotone and if $x^* \in S$ solves $\mathrm{VI}(H, S)$, then x^* is a Nash equilibrium of the GNEP. Since S is the solution set of $\mathrm{VI}(F, \mathbb{R}^n)$, this is a special case of the HVI problem.

Moreover, when the shared constraints are modeled with nonlinear complementarity conditions as

$$S := \{x \in \mathbb{R}^n \mid 0 \leq x \perp F(x) \geq 0\}.$$

It is easy to see that x^* solves $\mathrm{VI}(F, \mathbb{R}^n_+)$ iff $x^* \in S$.

Optimal solution selection: In optimization, when the primary problem admits multiple optimal solutions, one often seeks a "best" solution according to an additional performance criterion. This *optimal solution selection problem* naturally leads to a bilevel structure, where the lower level ensures feasibility and optimality with respect to the primary objective. At the same time, the upper-level discriminates among solutions by enforcing the secondary criterion. A parallel situation occurs in noncooperative games, where multiple Nash equilibria may exist, each differing in terms of efficiency or fairness. Selecting equilibria that minimize or maximize a given metric—such as the Price of Anarchy or Price of Stability—can likewise be cast in a bilevel form, with equilibrium conditions at the lower level and metric optimization at the upper level. Thus, both optimization and game-theoretic contexts illustrate how HVIs provide a unified framework for systematically addressing problems of solution and equilibrium selection. For more details, please refer to Samadi and Yousefian (2025).

Research Gap: There are some key challenges in solving HVIs. First, it is hard to project onto the feasible set X_F^* of the outer VI problem in many cases. Therefore, projection-based methods, including projected gradient, extragradient (EG) Korpelevich (1976); Nemirovski (2004), and optimistic GDA Mokhtari et al. (2020), do not apply directly to the problem $\mathrm{VI}(H, X_F^*)$. Second, consider $H(x) = \nabla f(x)$ for some convex function f; thus (2) reduces to a VI-constrained convex minimization problem. However, unlike typical bilevel optimization problems, which have a finite number of functional constraints at the inner level, the corresponding constraint set is characterized by an infinite number of inequalities. Hence, we cannot apply methods based on Lagrangian relaxation and Karush–Kuhn–Tucker (KKT) conditions, as in functional-constrained problems. In the following, we present a literature review on VIs and HVIs.

1.2 Related work

Variational inequalities. The VI problem was first introduced by Minty (1962) and Stampacchia (1964). Since then, VIs have played a vital role in different engineering problems Facchinei and Pang (2003); Pang et al. (2010); Shanbhag (2013). A popular solution method for the classical VIs is EG method by Korpelevich (1976) (see also Nemirovski (2004); Censor et al. (2011); Iusem and Nasri (2011); Juditsky et al. (2011); Iusem et al. (2017); Chen et al. (2017)). In a seminal work, Nemirovski (2004) introduced a generalized EG-type algorithm called mirror-prox and showed $\mathcal{O}(\epsilon^{-1})$ complexity for smooth and deterministic VIs which improved the existing $\mathcal{O}(\epsilon^{-2})$ complexity of projected gradient method. Here, ϵ denotes the error for the VI problem in terms of a gap function (c.f. Definition 1 and 2). This result matches the lower bounds provided by Ouyang and Xu (2021) for the bilinear saddle point problem, and hence is an optimal complexity. Further, Nesterov et al. (2006) obtained linear convergence for the strongly monotone VI problem. In addition, simple yet efficient algorithms such as projected reflected gradient methods Malitsky (2015) and operator extrapolation Kotsalis et al. (2022) were proposed for solving monotone VIs. Recently, an operator constraint extrapolation method Boob et al. (2024) was introduced for solving VI problems with complex but finitely many function constraints. Unlike EG methods, these algorithms need only one projection onto the "simple" set X in each iteration and track a single sequence of iterates. The VI problem becomes extremely challenging in the stochastic setting, where we cannot access the operator's exact evaluation. Jiang and Xu (2008) established stochastic approximation (SA) to obtain the asymptotic convergence for the stochastic VIs with monotone and smooth operators. Later, Juditsky et al. (2011); Koshal et al. (2012); Yousefian et al. (2017) showed stochastic approximation methods for merely monotone and non-Lipschitz stochastic VIs. In particular, the stochastic mirror-prox method exhibited complexity of $\mathcal{O}(\epsilon^{-2})$ Juditsky et al. (2011). Yousefian et al. (2014) achieved $\mathcal{O}(\epsilon^{-1})$ complexity for stochastic VIs under weak sharpness assumption. Boob et al. (2024) proved the optimal complexity of $\mathcal{O}(\epsilon^{-2})$ for VI problems with stochastic operator and stochastic function constraints.

Hierarchical variational inequalities. The ex-

isting literature on HVIs and VI-constrained optimization problems is scarce Xu (2004); Yamada et al. (2011); Facchinei et al. (2014); Yousefian et al. (2017); Kaushik and Yousefian (2021); Kaushik et al. (2023); Samadi and Yousefian (2025). Yamada et al. (2011) proposed a *hybrid steepest descent method* (HSDM) to solve a nonsmooth bilevel optimization problem. Xu (2004) presented a sequential averaging method on the HVI problem. However, neither of those works has provided any convergence rate in inner or outer levels. Few works on HVIs and VI-constrained optimization provide explicit complexity results for optimality (outer-level) and feasibility (inner-level) gaps. Kaushik and Yousefian (2021) showed $\mathcal{O}(\epsilon^{-4})$ convergence rate for optimality and feasibility gaps for a convex VI-constrained optimization problem. In the follow-up work, Samadi and Yousefian (2025) improved the rate to $\mathcal{O}(\epsilon^{-2})$ for a deterministic smooth HVI problem. Building on Samadi and Yousefian (2025), Alves et al. (2025) developed an inertial variant of extragradient methods for smooth and deterministic HVIs. To the best of our knowledge, Alves et al. (2025) is the most recent work establishing optimality and feasibility gap measures for hierarchical variational inequality problems.

Regularization and gradient-based methods. As discussed before, HVIs are difficult to solve given that (i) it is difficult to project onto the feasible set X_F^*, and (ii) X_F^* consists of infinitely many functional constraints. Indeed, it implies the Lagrangian dual variable must have infinite dimension, which is impractical. The well-known Tikhonov's regularization method Tikhonov (1963) introduces a simple idea to handle this challenge. It uses iterative methods to solve the inner and outer problems in (2) simultaneously. In the context of HVIs, Tikhonov's regularization uses an alternative operator $O(\cdot, \eta_k) := F(\cdot) + \eta_k H(\cdot)$ with regularization parameter $\eta_k > 0$, satisfying $\eta_k \to 0$. Using $\eta_k \to 0$ and $\sum_{k=1}^{\infty} \eta_k = \infty$, Cabot (2005) and Solodov (2007a) proposed a projected gradient method and showed asymptotic convergence. Motivated by regularization-based approaches, some works have combined standard first-order methods, such as block coordinate or extragradient, with an iteratively regularized scheme to solve deterministic HVI Kaushik and Yousefian (2021); Samadi and Yousefian (2025). A few papers address more general bilevel optimization problems. Solodov (2007b) used a bundle method for nonsmooth bilevel convex optimization problems to obtain asymptotic convergence. Recently, Doron and Shtern (2023) developed the Iterative Approximation and the Level-Set Expansion (ITALEX) approach, which provides explicit convergence rates for both inner and outer levels. Other important works addressing nonsmooth bilevel problems include Sabach and Shtern (2017) and Merchav and Sabach (2023).

1.3 Contributions

In this paper, we introduce a new method based on operator extrapolation Kotsalis et al. (2022); Boob et al. (2024) and Tikhonov's regularization for a general class of hierarchical variational inequality problems where both operators H and F are nonsmooth and stochastic. We summarize our contributions below.

First, we provide an explicit and best-known convergence rate for optimality and feasibility gaps for stochastic nonsmooth HVIs. Among the handful of works that measure suboptimality and infeasibility of the solutions Kaushik and Yousefian (2021); Kaushik et al. (2023); Samadi and Yousefian (2025), none of them study a fully-stochastic HVI problem with nonsmooth operators. Specifically, we show $\mathcal{O}(\epsilon^{-4})$ oracle complexity for a general nonsmooth HVI.

Second, in case we have a smooth stochastic operator F, we show $\mathcal{O}(\epsilon^{-2})$ complexity in operator H by doing mini-batching in operator F and maintaining $\mathcal{O}(\epsilon^{-4})$ complexity for F. We further improve the complexity to $\mathcal{O}(\epsilon^{-2})$ when the operator F is smooth and deterministic.

Third, if H is strongly monotone, we improve the complexity to $\mathcal{O}(\epsilon^{-4/5})$ in the general nonsmooth stochastic case. This complexity is, in fact, better than the best-known result in the more restrictive smooth deterministic case. We further improve this complexity to $\mathcal{O}(\epsilon^{-2/3})$ if F is smooth and deterministic.

Finally, Unlike the recent EG-based approach Samadi and Yousefian (2025), our method does not require an additional operator evaluation and only needs to store the current evaluation for the next iteration. We verify the advantage of our method over the EG-type method, along with our theoretical convergence results for the newly proposed stochastic setting, using numerical experiments. In Table 1, we compare the most relevant methods in HVIs, namely Kaushik and Yousefian (2021); Kaushik et al. (2023); Samadi and Yousefian (2025), with our algorithm in terms of the assumptions, settings used, and convergence rates.

Outline: Section 2 presents notations, preliminaries, and assumptions. Section 3 describes R-OpEx in detail to solve problem (2). Section 4 provides a

Table 1: Comparison of algorithms under various settings and convergence rates.

Algorithm	Setting				Operator Complexity		
	Stochastic Operator		Nonsmooth Operator		Monotone		Strongly Monotone
	F	H	F	H	F	H	H
aRB-IRG*Kaushik and Yousefian (2021)	×	×	×	×	$\mathcal{O}(1/\epsilon^4)$	$\mathcal{O}(1/\epsilon^4)$	–
pair-IG†Kaushik et al. (2023)	×	×	×	×	$\mathcal{O}(1/\epsilon^4)$	$\mathcal{O}(1/\epsilon^4)$	–
IR-EG‡Samadi and Yousefian (2025)	×	×	×	×	$\mathcal{O}(1/\epsilon^2)$	$\mathcal{O}(1/\epsilon^2)$	$\mathcal{O}(\ln(1/\epsilon)/\epsilon)$
R-OpEx	✓	✓	✓	✓	$\mathcal{O}(1/\epsilon^4)$	$\mathcal{O}(1/\epsilon^4)$	$\mathcal{O}(1/\epsilon^{4/5})$
R-OpEx (Mini-batching)	✓	✓	×	✓	$\mathcal{O}(1/\epsilon^4)$	$\mathcal{O}(1/\epsilon^2)$	–
R-OpEx	×	✓	×	✓	$\mathcal{O}(1/\epsilon^2)$	$\mathcal{O}(1/\epsilon^2)$	$\mathcal{O}(1/\epsilon^{2/3})$

* Averaging randomized block iteratively regularized gradient.
† Projected averaging regularized incremental subgradient.
‡ Iteratively regularized extragradient.

unified convergence analysis of the R-OpEx method for various cases, including general HVI and HVI with a smooth inner level. We illustrate numerical experiments in Section 5 and conclude in Section 6.

2 NOTATION AND PRELIMINARIES

We use the following notation throughout the paper. We denote $[m] := \{1, \cdots, m\}$. A vector $x \in \mathbb{R}^n$ is a column vector, and its transpose is shown as $x^\top$. Furthermore, $\mathbb{R}^n_+$ represents the non-negative orthant of an n-dimensional Euclidean space. Euclidean norm is denoted as $\|\cdot\|$ and the standard inner product is defined as $\langle \cdot, \cdot \rangle$. We say the operator $F : X \to \mathbb{R}^n$ is monotone on the convex set $X \subseteq \mathbb{R}^n$ if $\langle F(x) - F(y), x - y \rangle \geq 0$ for any $x, y \in X$. Moreover, the operator F is said to be μ_F-strongly monotone if $\langle F(x) - F(y), x - y \rangle \geq \mu \|x - y\|^2$ for any $x, y \in X$. We denote the Euclidean projection of a vector x onto a closed convex set X as $\mathbf{proj}_X(x)$, where $\mathbf{proj}_X(x) := \text{argmin}_{y \in X} \|x - y\|$. Consequently, we define the distance function $\text{dist}(x, X)$ for a given vector x and closed convex set X as $\text{dist}(x, X) := \min_{y \in X} \|x - y\| = \|x - \mathbf{proj}_X(x)\|$. The operator F is said to be L_F-*Lipschitz continuous* if $\|F(x) - F(y)\| \leq L_F \|x - y\|$, for all $x, y \in X$. We will explore the concepts of smoothness and nonsmoothness later. We define the diameter of a compact set X by $D_X := \max_{x,y \in X} \frac{\|x-y\|}{2}$. We now discuss the definitions and assumptions used throughout this paper. First, we denote the solution set of (2) as X^*_H and assume it is nonempty. Next, we consider the general case where F and H are possibly composed of Lipschitz continuous and bounded discontinuous monotone operators, satisfying

$$\|F(x_1) - F(x_2)\| \leq L_F \|x_1 - x_2\| + M_F, \quad \forall x_1, x_2 \in X, \tag{3a}$$

$$\|H(x_1) - H(x_2)\| \leq L_H \|x_1 - x_2\| + M_H, \quad \forall x_1, x_2 \in X, \tag{3b}$$

where the L_F, L_H terms stem from the Lipschitz continuous components and the M_F, M_H correspond to the discontinuous components of operators F, H, respectively. If $M_F = 0$ or $M_H = 0$, the corresponding operator is Lipschitz continuous, which is known as a "smooth operator" in the VI literature. Consequently, one can classify the hierarchical VI problem in (2) into smooth and nonsmooth cases. The problem (2) is a *smooth HVI* if both operators F and H are smooth, i.e., $M_F = M_H = 0$. If at least one of M_F or M_H is positive, then (2) is a *nonsmooth HVI* problem - the main focus of this paper. Apart from smoothness and nonsmoothness assumptions, we say the operators are bounded in sets X and X^*_F as follows:

$$\|F(x)\| \leq C_F, \forall x \in X, \quad \|F(x)\| \leq B_F, \forall x \in X^*_F, \tag{4a}$$

$$\|H(x)\| \leq C_H, \forall x \in X, \quad \|H(x)\| \leq B_H, \forall x \in X^*_F. \tag{4b}$$

Convergence criteria. In this paper, we define three criteria to evaluate an obtained solution. The first two measures correspond to the gap functions for each inner and outer VI problem. First, let us define the inner gap function or *the feasibility gap*.

Definition 1. Let $X \subseteq \mathbb{R}^n$ be a nonempty, closed, and convex set and $F : X \to \mathbb{R}^n$ be a monotone operator. Then, we say that $\tilde{x}$ is an ϵ-feasible solution associated with the inner $\text{VI}(F, X)$ if

$$\text{Gap}(\tilde{x}, F, X) := \max_{x \in X} \langle F(x), \tilde{x} - x \rangle \leq \epsilon, \tag{5}$$

Similarly, $\tilde{x}$ is a stochastic ϵ-feasible solution if

$$\mathbb{E}[\text{Gap}(\tilde{x}, F, X)] := \mathbb{E}[\max_{x \in X} \langle F(x), \tilde{x} - x \rangle] \leq \epsilon.$$

Furthermore, one can define the outer gap function, also known as *the optimality gap*, as follows.

Definition 2. Let $X \subseteq \mathbb{R}^n$ be a nonempty, closed, and convex set, and $H : X \to \mathbb{R}^n$ be a monotone operator. Then, we say that $\tilde{x}$ is an ϵ-optimal solution

associated with the outer VI(H, X_F^*) if

$$\text{Gap}(\tilde{x}, H, X_F^*) := \max_{x \in X_F^*} \langle H(x), \tilde{x} - x \rangle \leq \epsilon. \quad (6)$$

Similarly, $\tilde{x}$ is a stochastic ϵ-optimal solution if $\mathbb{E}[\text{Gap}(\tilde{x}, H, X_F^*)] := \mathbb{E}[\max_{x \in X_F^*} \langle H(x), \tilde{x} - x \rangle] \leq \epsilon$,

Henceforth, we say that an ϵ-feasible and ϵ-optimal solution as simply ϵ-solution. Note that $\text{Gap}(\tilde{x}, F, X) = 0$ (respectively, $\text{Gap}(\tilde{x}, H, X_F^*) = 0$) if and only if $\tilde{x}$ is a solution to VI(F, X) (respectively, VI(H, X_F^*)). Moreover, it is easy to see $\text{Gap}(\tilde{x}, F, X) \geq 0$ for all $\tilde{x} \in X$ and $\text{Gap}(\tilde{x}, H, X_F^*) \geq 0$ for all $\tilde{x} \in X_F^*$. However, if $\tilde{x} \notin X$, then one might obtain $\text{Gap}(\tilde{x}, F, X) \leq 0$. Similarly, if $\tilde{x} \notin X_F^*$ then we might have $\text{Gap}(\tilde{x}, H, X_F^*) \leq 0$. Due to the structure of our proposed method, the latter case with negative values in optimality may occur. Therefore, we also need to obtain lower bounds for the optimality gap.
Lastly, we consider an additional measure of feasibility. In particular, we measure the feasibility of a given point $\tilde{x} \in X$ as its distance from the set X_F^*, i.e., $\text{dist}(\tilde{x}, X_F^*)$.

Stochastic approximation of operators. We assume operators F and H are in expectation form. We refer to this as a fully-stochastic hierarchical VI. We use stochastic oracles (SOs) to generate random vectors that estimate F and H. In particular, given a random variable ξ which is independent of the search point x, SOs generate random samples of $\mathfrak{F}(x, \xi)$ and $\mathfrak{H}(x, \xi)$ such that

$$\forall x \in X, \quad F(x) = \mathbb{E}[\mathfrak{F}(x, \xi)], \quad \mathbb{E}[\|F(x) - \mathfrak{F}(x, \xi)\|^2] \leq \sigma_F^2, \quad (7a)$$

$$\forall x \in X, \quad H(x) = \mathbb{E}[\mathfrak{H}(x, \xi)], \quad \mathbb{E}[\|H(x) - \mathfrak{H}(x, \xi)\|^2] \leq \sigma_H^2. \quad (7b)$$

Weak sharpness of operators. Throughout this paper, we employ the weak sharpness assumption to derive lower bounds only. Our upper bounds hold without the weak sharpness assumption. Its primary role is to control potential negative optimality gaps that may arise during intermediate iterations when the iterates lie outside the feasible set of the outer problem. When the feasible set coincides with the entire space, i.e., $X_F^* = \mathbb{R}^n$, this assumption becomes unnecessary and can be relaxed (see Kaushik and Yousefian (2021); Kaushik et al. (2023)).

Definition 3. We say that VI(F, X) is α-weakly sharp if there exists a scalar $\alpha > 0$ such that for all $x \in X$ and $x^* \in X_F^*$, the following holds

$$\langle F(x^*), x - x^* \rangle \geq \alpha \text{dist}(x, X_F^*).$$

3 R-OPEX METHOD

We present the R-OpEx method in Algorithm 1 below. This algorithm, in its k-th iteration with the search point x_k, generates random vectors $\mathfrak{F}_k := \mathfrak{F}(x_k, \xi_k)$ and $\mathfrak{H}_k := \mathfrak{H}(x_k, \xi_k)$ using the SOs defined in (7a) and (7b). The most crucial idea in R-OpEx is that it incorporates regularization techniques with operator extrapolation methods. Specifically, we write the regularized operator as $\mathfrak{O}_k^{\eta_k} := \mathfrak{F}_k + \eta_k \mathfrak{H}_k$ with regularization weight η_k. Then, we perform a descent step with an extrapolated regularized operator as $\mathfrak{O}_k^{\eta_k} + \theta_k[\mathfrak{O}_k^{\eta_{k-1}} - \mathfrak{O}_{k-1}^{\eta_{k-1}}]$ where θ_k is the extrapolation parameter. See line 2 of Algorithm 1 for an expanded form. Notice here that γ_k is the step-size for the projection operator and τ_k is the averaging parameter to obtain $\bar{x}_K$.
Unlike the usual extragradient methods, R-OpEx is a simple algorithm that does not require maintaining two sequences or evaluating operators F of H (or their stochastic versions $\mathfrak{F}$ or $\mathfrak{H}$) twice. Indeed, in every iteration, we only require a single evaluation F and H, i.e., $\mathfrak{F}_k$ and $\mathfrak{H}_k$ respectively. Moreover, we maintain a single sequence $\{x_k\}$ in Algorithm 1. Note that the regularization parameter η_k plays a weighting role that controls the balance between the inner problem (feasibility, operator $\mathfrak{F}$) and the outer problem (optimality, operator $\mathfrak{H}$). For example, a very small η_k ($\eta_k \approx 0$) prioritizes the inner problem, making the signal from $\mathfrak{H}$ negligible, while a very large η_k ($\eta_k \to \infty$) renders the inner problem insignificant. Hence, to get convergence in both levels, we need to balance η_k appropriately.

4 CONVERGENCE ANALYSIS

This section provides an integrated convergence analysis for the R-OpEx method. We first provide convergence results for the general HVI with nonsmooth and stochastic operators. Then, we assume the inner problem VI(F, X) is smooth and mention the corresponding rates. In the following lemma, we provide an essential relation that holds for each iteration of Algorithm 1.

Lemma 1 (One-iteration bound for R-OpEx)**.** *Consider the problem* (2) *with* $\mu_H \geq 0$. *Let* $\{x_k\}_{k \geq 1}$ *be a sequence generated by Algorithm 1. Suppose* (3a),(3b) *holds. Then we have the following relation*

Algorithm 1 **R**egularized **Op**erator **Ex**trapolation (R-OpEx) method

1: **Input:** $x_0 = x_1 \in X$, $\{\tau_1, \eta_0, \eta_1, \gamma_1, \theta_1\}$, K
2: **for** k $= 1, \dots, K-1$ **do**

$$x_{k+1} \leftarrow \underset{x\in X}{\operatorname{argmin}} \langle \mathfrak{F}_k + \eta_k \mathfrak{H}_k + \theta_k \big[\mathfrak{F}_k + \eta_{k-1}\mathfrak{H}_k - [\mathfrak{F}_{k-1} + \eta_{k-1}\mathfrak{H}_{k-1}]\big], x\rangle + \tfrac{1}{2\gamma_k}\|x - x_k\|^2 \tag{8}$$

3: **end for**
4: **return** $\bar{x}_K = \frac{\sum_{k=0}^{K-1} \tau_k x_{k+1}}{\sum_{k=0}^{K-1} \tau_k}$.

for any $x \in X$

$$\begin{aligned}
&\langle F(x) + \eta_k H(x), x_{k+1} - x\rangle \le \tfrac{1}{2\gamma_k}[\|x - x_k\|^2 - \|x_{k+1} - x_k\|^2] \\
&- (\tfrac{1}{2\gamma_k} + \eta_k \mu_H)\|x - x_{k+1}\|^2 + \langle \Delta\mathfrak{O}_{k+1}, x_{k+1} - x\rangle \\
&- \theta_k \langle \Delta\mathfrak{O}_k, x_k - x\rangle \\
&+ \theta_k (L_F + \eta_{k-1} L_H)\|x_k - x_{k-1}\|\|x_{k+1} - x_k\| \\
&- \langle \delta^F_{k+1} + \eta_k \delta^H_{k+1}, x_{k+1} - x\rangle \\
&- \theta_k \langle \delta^F_k - \delta^F_{k-1} + \eta_{k-1}[\delta^H_k - \delta^H_{k-1}], x_{k+1} - x_k\rangle \\
&+ \theta_k (M_F + \eta_{k-1} M_H)\|x_{k+1} - x_k\|,
\end{aligned} \tag{9}$$

where $\Delta\mathfrak{O}_k = \mathfrak{F}_k - \mathfrak{F}_{k-1} + \eta_{k-1}[\mathfrak{H}_k - \mathfrak{H}_{k-1}]$ *and* $\delta^F_k = \mathfrak{F}_k - F(x_k)$, $\delta^H_k = \mathfrak{H}_k - H(x_k)$.

Now, we are ready to state the results for the general nonsmooth and stochastic problem.

4.1 General nonsmooth and stochastic HVI

In this section, we consider the most general case where we assume both operators F and H are in nonsmooth and stochastic forms. Notably, we assume M_F, σ_F and M_H, σ_H have positive values. We discuss the convergence results for the monotone case in Theorems 1-2 and discuss the similar results for the strongly monotone case in Section A.2.2.

Theorem 1. *Let us assume problem* (2) *is monotone* ($\mu_H = 0$). *Also let us assume* (3a),(3b),(4a),(4b) *and* (7a),(7b) *hold. Suppose we have the following step-size policy*

$$\tau_k = \tau = 1, \quad \theta_k = \theta = 1, \quad \eta_k = \eta = K^{-\frac{1}{4}},$$
$$\gamma_k = \gamma = \frac{D_X}{8D_X(L_F + \eta L_H) + \sqrt{K(M_F^2 + 2\sigma_F^2 + \eta^2(M_H^2 + 2\sigma_H^2))}}. \tag{10}$$

Then, Algorithm 1 gives the following optimality and feasibility upper bounds for $K \ge 1$

$$\begin{aligned}
&- B_H \mathbb{E}[dist(\bar{x}_K, X_F^*)] \le \mathbb{E}[Gap(\bar{x}_K, H, X_F^*)] \le \\
&D_X \Bigg[16 D_X (\tfrac{L_F}{K^{3/4}} + \tfrac{L_H}{K}) + \tfrac{2\sqrt{M_F^2 + 2\sigma_F^2 + \eta^2(M_H^2 + 2\sigma_H^2)}}{K^{1/4}} + \\
&\tfrac{5(M_F^2 + 2\sigma_F^2 + \eta^2(M_H^2 + 2\sigma_H^2))}{8D_X K^{3/4}(L_F + \eta L_H) + K^{5/4}\sqrt{M_F^2 + 2\sigma_F^2 + \eta^2(M_H^2 + \sigma_H^2)}} \\
&+ \tfrac{5(M_F^2 + 2\sigma_F^2 + \eta^2(M_H^2 + 2\sigma_H^2)) + \sigma_F^2 + \eta^2\sigma_H^2}{8D_X K^{-1/4}(L_F + \eta L_H) + K^{1/4}\sqrt{M_F^2 + 2\sigma_F^2 + \eta^2(M_H^2 + 2\sigma_H^2)}}\Bigg].
\end{aligned} \tag{11}$$

and

$$\begin{aligned}
&\mathbb{E}[Gap(\bar{x}_K, F, X)] \\
&\le D_X \Bigg[16 D_X (\tfrac{L_F}{K} + \tfrac{L_H}{K^{5/4}}) + \tfrac{2\sqrt{M_F^2 + 2\sigma_F^2 + \eta^2(M_H^2 + 2\sigma_H^2)}}{K^{1/2}} \\
&+ \tfrac{5(M_F^2 + 2\sigma_F^2 + \eta^2(M_H^2 + 2\sigma_H^2)) + \sigma_F^2 + \eta^2\sigma_H^2}{8D_X(L_F + \eta L_H) + K^{1/2}\sqrt{M_F^2 + 2\sigma_F^2 + \eta^2(M_H^2 + 2\sigma_H^2)}} + \\
&\tfrac{5(M_F^2 + 2\sigma_F^2 + \eta^2(M_H^2 + \sigma_H^2))}{8D_X K(L_F + \eta L_H) + K^{3/2}\sqrt{M_F^2 + 2\sigma_F^2 + \eta^2(M_H^2 + 2\sigma_H^2)}} + \tfrac{2C_H}{K^{1/4}}\Bigg].
\end{aligned} \tag{12}$$

Remark 1. *Note that using step-size policy in* (10)*, Algorithm 1 gives convergence rate of* $\mathcal{O}\big(D_X(\frac{D_X L_F}{K^{3/4}} + \frac{D_X L_H}{K} + \frac{M_F + \sigma_F}{K^{1/4}} + \frac{M_H + \sigma_H}{K^{1/2}})\big)$. *This rate is the best-known rate for a stochastic and nonsmooth HVI problem.*

Next, observe that constructing a lower bound on the optimality gap is crucial, given that Algorithm 1 uses a single projection onto set X and not set X_F^*. Therefore, it is also possible that we obtain a negative optimality gap. In Theorem 2, we construct an explicit lower bound for the optimality gap.

Theorem 2. *Suppose VI*(F, X) *is* α*-weakly sharp. Then using the same assumptions in Theorem 1, we have*

$$\begin{aligned}
&\mathbb{E}[Gap(\bar{x}_K, H, X_F^*)] \ge \\
&- \tfrac{B_H D_X}{\alpha}\Bigg[16 D_X (\tfrac{L_F}{K} + \tfrac{L_H}{K^{5/4}}) + \tfrac{2\sqrt{M_F^2 + 2\sigma_F^2 + \eta^2(M_H^2 + 2\sigma_H^2)}}{K^{1/2}} \\
&+ \tfrac{5(M_F^2 + 2\sigma_F^2 + \eta^2(M_H^2 + 2\sigma_H^2)) + \sigma_F^2 + \eta^2\sigma_H^2}{8D_X(L_F + \eta L_H) + K^{1/2}\sqrt{M_F^2 + 2\sigma_F^2 + \eta^2(M_H^2 + 2\sigma_H^2)}} + \\
&\tfrac{5(M_F^2 + 2\sigma_F^2 + \eta^2(M_H^2 + \sigma_H^2))}{8D_X K(L_F + \eta L_H) + K^{3/2}\sqrt{M_F^2 + 2\sigma_F^2 + \eta^2(M_H^2 + 2\sigma_H^2)}} + \tfrac{2C_H}{K^{1/4}}\Bigg].
\end{aligned} \tag{13}$$

Additionally, we provide explicit oracle complexity in case the threshold $\frac{\alpha}{\|H(x^*)\|}$ is available. Remark 2 states the convergence rate in the presence of $\frac{\alpha}{\|H(x^*)\|}$. For more details, the reader is referred to Theorem 6 in Section A.2.1.

Remark 2. *Consider we change* $\eta = K^{-1/4}$ *to a value satisfying* $\eta \le \frac{\alpha}{\|H(x^*)\|}$ *and keep the rest of parameters as* (10)*. Then Algorithm 1 gives rate of* $\mathcal{O}\big(D_X(\frac{D_X L_F}{K} + \frac{D_X L_H}{K^{5/4}} + \frac{M_F + M_H + \sigma_F + \sigma_H}{K^{1/2}})\big)$. *This*

improves the convergence rate from $\mathcal{O}(1/K^{1/4})$ *to* $\mathcal{O}(1/K^{1/2})$.

Next, we provide the convergence rates for the strongly monotone case, where $\mu_H > 0$. Section A.2.2 presents a complete analysis of this case.

Theorem 3. *Suppose that problem* (2) *is strongly monotone* ($\mu_H > 0$)*, and that the same assumptions as in Theorem 1 hold. Consider* $\tau_k = k+1, \theta_k = \frac{k}{k+1}$ *and keep* η_k, γ_k *as* (10)*. Then, we obtain* $\mathcal{O}\big(D_X(\frac{D_X L_F}{K^{7/4}} + \frac{D_X L_H}{K^2} + \frac{M_F+\sigma_F}{K^{5/4}} + \frac{M_H+\sigma_H}{K^{3/2}})\big)$*. Further, suppose* $\frac{\alpha}{\|H(x^*)\|}$ *is available then the convergence rate is improved to* $\mathcal{O}\big(D_X(\frac{D_X L_F}{K^2} + \frac{D_X L_H}{K^{9/4}} + \frac{M_F+M_H+\sigma_F+\sigma_H}{K^{3/2}})\big)$.

As one can observe from (10), we should have an estimate of the number of iterations K a priori. In Section 4.2, we propose a new step-size policy that eliminates the need to know the number of iterations in advance.

4.2 Step-size policy for η_k independent of K

This section specifies a new step-size policy independent of K. The full analysis is in Section A.3. While we provide analysis of such a policy for only the nonsmooth fully stochastic case, similar claims can be made for the remaining cases. We drop their discussion for the sake of brevity.

Theorem 4. *Consider the following updating policy*

$$\tau_k = \tau = 1, \quad \eta_k = (k+1)^{-1/4}, \quad \theta_k = (\tfrac{k}{k+1})^{1/4},$$
$$\gamma_k = \frac{D_X}{8D_X(L_F+\eta_k L_H)+\sqrt{k(M_F^2+2\sigma_F^2+\eta_k^2(M_H^2+2\sigma_H^2))}}.$$

Then by using the same assumption in Theorem 1, Algorithm 1 gives $\mathcal{O}\big(D_X(\frac{D_X L_F}{K^{3/4}} + \frac{D_X L_H}{K} + \frac{M_F+\sigma_F}{K^{1/4}} + \frac{M_H+\sigma_H}{K^{1/2}})\big)$ *convergence rate.*

It is worth mentioning that the choice of η_k is optimal for both fixed (i.e., K-dependent) and time-varying policies. In the fixed scheme, the convergence requirements (cf. (24a)) fix η_k across iterations. From (9), dividing by η_k and using (25) shows that choosing $\eta_k \ll K^{-1/4}$ enlarges the coefficient of the optimality term and slows the rate. Conversely, (35) adds the feasibility penalty $2C_H D_X \eta_k$ on the left-hand side, so taking $\eta_k \gg K^{-1/4}$ weakens the inner-level signal and deteriorates feasibility. Thus the best trade-off is $\eta_k \propto K^{-1/4}$. In the time-varying scheme, we relax the fixed-step requirement (see (54)) and allow $\eta_k \approx k^{-a}$; the same trade-off argument yields the optimal exponent $a = 1/4$, i.e., $\eta_k \propto k^{-1/4}$.

4.3 Smooth inner VI

In this section, we analyze the convergence of Algorithm 1 under the assumption that the operator F is smooth, i.e., $M_F = 0$. We can break down this setting into two subcategories depending on whether we have stochasticity in F ($\sigma_F > 0$) or not ($\sigma_F = 0$). In case of stochastic F, we use mini-batching to reduce the effect of σ_F. In particular, for size B mini-batching to the operator $\mathfrak{F}$, we have $\bar{\mathfrak{F}}_B(x) = \frac{1}{B}\sum_{i=1}^{B}\mathfrak{F}(x,\xi_i)$ to approximate operator operator F in line two of Algorithm 1. Theorem 5 states the convergence rate and operator complexity for an inner-smooth HVI with stochastic or deterministic F. The comprehensive analysis for the smooth inner VI case is presented in Section A.4.

Theorem 5. Stochasitc F**:** *Assume that problem* (2) *is monotone* ($\mu_H = 0$)*. Additionally, let* $M_F = 0$*, and suppose we have the following step-size policy with mini-batching of size* $B = K$

$$\tau_k = \tau = 1, \quad \theta_k = \theta = 1, \quad \eta_k = \eta = K^{-\frac{1}{2}},$$
$$\gamma_k = \gamma = \frac{D_X}{8D_X(L_F+\eta L_H)+\sqrt{M_H^2+2(\sigma_H^2+\sigma_F^2)}}.$$

Then, R-OpEx gives the convergence rate of $\mathcal{O}\big(D_X(\frac{D_X L_F}{K^{1/2}} + \frac{D_X L_H}{K} + \frac{M_H+\sigma_F+\sigma_H}{K^{1/2}})\big)$.

Deterministic F**:** *If we have a smooth and deterministic inner VI* ($M_F = \sigma_F = 0$)*, then we obtain the same convergence rate as above. Finally, if* ($M_F = \sigma_F = 0$) *and suppose HVI is strongly monotone* ($\mu_H > 0$)*, then, the convergence rate improves to* $\mathcal{O}\big(D_X(\frac{D_X L_F}{K^{3/2}} + \frac{D_X L_H}{K^2} + \frac{M_H+\sigma_F+\sigma_H}{K^{3/2}})\big)$.

We improve the convergence rate from $\mathcal{O}(K^{-1/4})$ in the general nonsmooth HVI to $\mathcal{O}(K^{-1/2})$. While the oracle complexity with respect to the stochastic operator F remains at $\mathcal{O}(\epsilon^{-4})$ under mini-batching, the complexity in H improves to $\mathcal{O}(\epsilon^{-2})$. In the strongly monotone setting for stochastic F, however, no mini-batching scheme yields an improvement beyond the current $\mathcal{O}(\epsilon^{-4/5})$ complexity in both F and H. For deterministic F, our results further show that R-OpEx achieves $\mathcal{O}(\epsilon^{-2})$ complexity in H, which can be strengthened to $\mathcal{O}(\epsilon^{-2/3})$ when $\mu_H > 0$.

5 NUMERICAL EXPERIMENTS

In this section, we validate the results from Section 4, particularly Theorem 1 and Theorem 3. Therefore, in our main experiments, we use step-sizes derived directly from our theoretical results. We present the first experimental setting for stochastic convex optimization with a feasible region defined by a stochastic two-player zero-sum

Nash game in Appendix A.5. Here, we focus on the traffic equilibrium problem and study two distinct cases. First, we consider a stochastic setting in which both operators incur randomness, and we report the performance of R-OpEx under this scenario. Second, to establish a performance baseline against state-of-the-art methods, we compare R-OpEx with IR-EG Samadi and Yousefian (2025) on a traffic equilibrium problem. All experiments are conducted on a 64-bit Windows 11 machine with Intel i7-1260P @ 2.10GHz and 16GB RAM.

Stochastic traffic equilibrium problem. As shown in Figure 1, we consider a traffic flow network. The traffic network we are interested in has

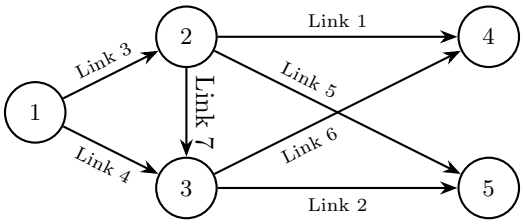


Figure 1: Traffic Network five nodes seven links

five nodes, seven links, two origin-destination (O-D) pairs ($1\rightarrow$ 4, $1\rightarrow$ 5), and six paths $p_1 = \{3,7,6\}, p_2 = \{3,1\}, p_3 = \{4,6\}, p_4 = \{3,7,2\}, p_5 = \{3,5\}, p_6 = \{4,2\}$. Travel equilibrium problems are modeled according to travel demand between each (O-D) pair and capacity on each link. Let $d = [d_1 + \xi_1, d_2 + \xi_2]$ be the stochastic travel demand between (O-D) pairs where $\xi_i, i = 1,2$ is the noise following $\mathcal{N}(0,1)$. Also, denote $cap = [cap_1, \ldots, cap_7]$ as the capacity vector associated with network links. The corresponding traffic flows for paths and links are denoted by $h = [h_1, \ldots, h_6]$ and $f = [f_1, \ldots, f_7]$. Furthermore, Δ represents link-path incident matrix, where $\delta_{a,p} = 1$ if path p contains link a, and $\delta_{a,p} = 0$ otherwise. Similarly, Ω denotes (O-D)-path matrix. In these experiments, we follow a generalized bureau of public roads (GBPR) to compute the link travel time function Yin et al. (2009) as follows.

$$c_a(f_a) = t_a^0(1 + 0.15(\tfrac{f_a}{cap_a})^{n_a}),$$

where $n_a \geq 1$, and t_a^0 are given parameters. One can see $f = \Delta h$ implying the following

$$C(h) = \Delta^\top c(\Delta h) \tag{14}$$

Let $u = [u_1, u_2]$ be the minimum travel costs between each (O-D) pair. Consequently, for a realization of the random vector $\xi = [\xi_1, \xi_2]$, one can use Wardrop's user equilibrium for the decision variable $x = [h, u]$ and construct the complementarity problem $x \geq 0, F(x) \geq 0$, and $x^\top F(x) = 0$ where we have the following operator $F \in \mathbb{R}^8$

$$F(x;\xi) = \begin{bmatrix} C(h) - \Omega^\top u \\ \Omega h - d \end{bmatrix}$$

We set $n_a = 1 \forall a = 1, \ldots, 7$, yielding a linear complementarity problem (LCP). Moreover, we use total travel cost of the network $\psi(x) = \mathbb{E}[\zeta^\top C(h)]$ as the function we want to minimize in traffic equilibrium, where ζ is a vector with each element $\zeta_i \sim \mathcal{U}_{0,2}, i \in [6]$. Given that ψ is convex and F is a monotone operator for $n_a = 1, a \in [7]$ (see Section 6.2 of Samadi and Yousefian (2025)), we can formulate the problem as a stochastic VI constrained minimization where the outer problem is associated with minimization of $\psi(x)$ and the inner problem is $\mathbf{VI}(F, \mathbb{R}^8_+)$. For the experiment setting, we set the capacity of each link a to $cap_a = 400$. Let the stochastic travel demand between (O-D) be $d = [200 + \xi_1, 220 + \xi_2]$. Moreover, we choose $t_a^0 = 1, \forall a \in [7]$. We use $\|\bar{x}_{k+1} - \bar{x}_k\|$ to measure suboptimality where $\bar{x}_k$ is the solution generated by R-OpEx. Since the inner problem is an LCP, we use the function $\phi(x) = \|\min\{x, 0\}\| + \|\min\{F(x), 0\}\| + \|x^\top F(x)\|$ to measure the infeasibility. Figure 2 demonstrates the performance of Algorithm 1 for $R = 10$ i.i.d instances and $K = 5 \times 10^6$ iterations. Similar to the previous section, we plot the gaps in terms of the number of iterations and time in seconds.

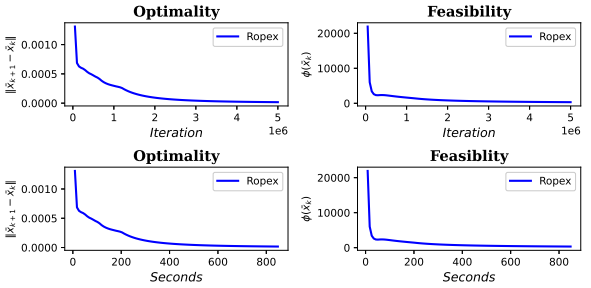


Figure 2: Optimality and feasibility gaps for 10 replications of Algorithm 1 with 5×10^5 iterations.

Comparison with IR-EG method. In the second part, we compare the performance of our algorithm with IR-EG Samadi and Yousefian (2025), the most recent and relevant baseline. Since IR-EG does not account for stochasticity or nonsmoothness, we restrict the traffic equilibrium problem to a smooth and deterministic setting for a fair comparison. To ensure a fair comparison, we set the step sizes and other parameters of each algorithm according to the policies prescribed in their respective convergence analysis. In other words, each method is run with

the parameters that are theoretically justified for achieving its stated convergence rate. To better illustrate the ability of both methods to handle nonlinearity, we evaluate them under different values of n_a. The results in Figure 3 show the optimality and feasibility gaps as functions of the number of operator evaluations. Consistent with our theoretical analysis, R-OpEx achieves uniformly lower gaps than IR-EG.

6 CONCLUSION AND DISCUSSION

This paper studies hierarchical variational inequality problems with nonsmooth and stochastic operators. We propose a novel and easy-to-implement algorithm based on operator extrapolation and Tikhonov regularization, achieving the best-known explicit convergence rate in both levels of HVI. These convergence rate guarantees are obtained using fixed or variable step-size policies, showcasing our method's flexibility. In addition, we improve the convergence rate and operator complexity if we have a smooth inner problem. Our proposed scheme matches the best existing convergence rates in the literature, whereas no work assumes stochasticity of the operators. Promising directions include stochastic VI-constrained nonconvex optimization and extending recent advances in last-iterate methods (see Boob and Khalafi (2024)) to HVIs.

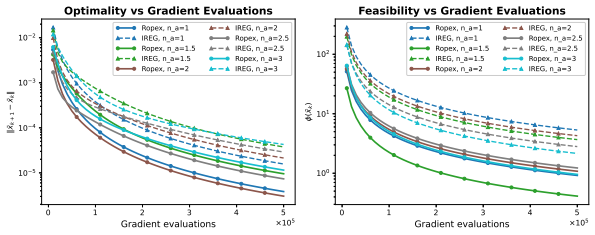


Figure 3: Comparison of R-OpEx and IR-EG algorithms on the traffic equilibrium problem under different congestion exponents $n_a \in \{1.0, 1.5, 2.0, 2.5, 3.0\}$. The plots illustrate the evolution of the optimality gap (left) and feasibility gap (right) with respect to the number of gradient evaluations (up to 5×10^5).

Acknowledgments.
D. Boob and M. Khalafi were partially supported by the National Science Foundation [Grant 2340858] and Office of Naval Research [Grant N000142412749].

References

Alves, M. M., Chen, K., and Fukuda, E. H. (2025). An inertial iteratively regularized extragradient method for bilevel variational inequality problems.

Boob, D., Deng, Q., and Khalafi, M. (2024). First-order methods for stochastic variational inequality problems with function constraints.

Boob, D., Deng, Q., and Lan, G. (2023). Stochastic first-order methods for convex and nonconvex functional constrained optimization. *Mathematical Programming*, 197(1):215–279.

Boob, D. and Khalafi, M. (2024). Optimal primal-dual algorithm with last iterate convergence guarantees for stochastic convex optimization problems.

Cabot, A. (2005). Proximal point algorithm controlled by a slowly vanishing term: applications to hierarchical minimization. *SIAM Journal on Optimization*, 15(2):555–572.

Censor, Y., Gibali, A., and Reich, S. (2011). The subgradient extragradient method for solving variational inequalities in hilbert space. *Journal of Optimization Theory and Applications*, 148(2):318–335.

Chen, Y., Lan, G., and Ouyang, Y. (2017). Accelerated schemes for a class of variational inequalities. *Mathematical Programming*, 165:113–149.

Doron, L. and Shtern, S. (2023). Methodology and first-order algorithms for solving nonsmooth and non-strongly convex bilevel optimization problems. *Mathematical Programming*, 201(1):521–558.

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

Facchinei, F., Pang, J.-S., Scutari, G., and Lampariello, L. (2014). Vi-constrained hemivariational inequalities: distributed algorithms and power control in ad-hoc networks. *Mathematical Programming*, 145(1):59–96.

Iusem, A. N., Jofré, A., Oliveira, R. I., and Thompson, P. (2017). Extragradient method with variance reduction for stochastic variational inequalities. *SIAM Journal on Optimization*, 27(2):686–724.

Iusem, A. N. and Nasri, M. (2011). Korpelevich's method for variational inequality problems in banach spaces. *Journal of Global Optimization*, 50:59–76.

Jiang, H. and Xu, H. (2008). Stochastic approximation approaches to the stochastic variational inequality problem. *IEEE Transactions on Automatic Control*, 53(6):1462–1475.

Juditsky, A., Nemirovski, A., and Tauvel, C. (2011). Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58.

Kaushik, H. D., Samadi, S., and Yousefian, F. (2023). An incremental gradient method for optimization problems with variational inequality constraints. *IEEE Transactions on Automatic Control*, 68(12):7879–7886.

Kaushik, H. D. and Yousefian, F. (2021). A method with convergence rates for optimization problems with variational inequality constraints. *SIAM Journal on Optimization*, 31(3):2171–2198.

Korpelevich, G. M. (1976). The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756.

Koshal, J., Nedic, A., and Shanbhag, U. V. (2012). Regularized iterative stochastic approximation methods for stochastic variational inequality problems. *IEEE Transactions on Automatic Control*, 58(3):594–609.

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

Malitsky, Y. (2015). Projected reflected gradient methods for monotone variational inequalities. *SIAM Journal on Optimization*, 25(1):502–520.

Merchav, R. and Sabach, S. (2023). Convex bi-level optimization problems with nonsmooth outer objective function. *SIAM Journal on Optimization*, 33(4):3114–3142.

Minty, G. J. (1962). Monotone (nonlinear) operators in hilbert space.

Mokhtari, A., Ozdaglar, A. E., and Pattathil, S. (2020). Convergence rate of o(1/k) for optimistic gradient and extragradient methods in smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 30(4):3230–3251.

Nemirovski, A. (2004). Prox-method with rate of convergence o (1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251.

Nesterov, Y., Scrimali, L., et al. (2006). Solving strongly monotone variational and quasi-variational inequalities. Technical report, CORE.

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

Pang, J.-S., Scutari, G., Palomar, D. P., and Facchinei, F. (2010). Design of cognitive radio systems under temperature-interference constraints: A variational inequality approach. *IEEE Transactions on Signal Processing*, 58(6):3251–3271.

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

Samadi, S. and Yousefian, F. (2025). Improved guarantees for optimal nash equilibrium seeking and bilevel variational inequalities. *SIAM Journal on Optimization*, 35(1):369–399.

Shanbhag, U. V. (2013). Stochastic variational inequality problems: Applications, analysis, and algorithms. In *Theory driven by influential applications*, pages 71–107. INFORMS.

Solodov, M. (2007a). An explicit descent method for bilevel convex optimization. *Journal of Convex Analysis*, 14(2):227.

Solodov, M. V. (2007b). A bundle method for a class of bilevel nonsmooth convex minimization problems. *SIAM Journal on Optimization*, 18(1):242–259.

Stampacchia, G. (1964). Formes bilineaires coercitives sur les ensembles convexes. *Comptes Rendus Hebdomadaires Des Seances De L Academie Des Sciences*, 258(18):4413.

Tikhonov, A. N. (1963). Solution of incorrectly formulated problems and the regularization method. *Sov Dok*, 4:1035–1038.

Xu, H.-K. (2004). Viscosity approximation methods for nonexpansive mappings. *Journal of Mathematical Analysis and Applications*, 298(1):279–291.

Yamada, I., Yukawa, M., and Yamagishi, M. (2011). Minimizing the moreau envelope of nonsmooth convex functions over the fixed point set of certain quasi-nonexpansive mappings. *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, pages 345–390.

Yin, Y., Madanat, S. M., and Lu, X.-Y. (2009). Robust improvement schemes for road networks under demand uncertainty. *European Journal of Operational Research*, 198(2):470–479.

Yousefian, F., Nedić, A., and Shanbhag, U. V. (2014). Optimal robust smoothing extragradi-

ent algorithms for stochastic variational inequality problems. In *53rd IEEE conference on decision and control*, pages 5831–5836. IEEE.

Yousefian, F., Nedić, A., and Shanbhag, U. V. (2017). On smoothing, regularization, and averaging in stochastic approximation methods for stochastic variational inequality problems. *Mathematical Programming*, 165:391–431.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A Appendix

A.1 ANALYSIS OF LEMMA 1

First, we must state a key lemma known as the "three-point" lemma, which is crucial in the proof of Lemma 1.

Lemma 2. *(See the proof in Boob et al. (2023)) Let x^* be a solution of problem $\min_{x\in X}\{h(x)+\frac{\lambda}{2}\|x-\hat{x}\|^2\}$ where $h(x)$ is a convex function. Then,*

$$h(x^*)-h(x)\leq \tfrac{\lambda}{2}[\|x-\hat{x}\|^2-\|x^*-x\|^2-\|x^*-\hat{x}\|^2] \quad \forall x\in X. \tag{15}$$

A.1.1 Proof of Lemma 1

Proof. From optimality of x_{k+1} and Lemma 2, we have the following

$$\langle \mathfrak{F}_k+\eta_k\mathfrak{H}_k, x_{k+1}-x\rangle \leq \tfrac{1}{2\gamma_k}[\|x-x_k\|^2-\|x-x_{k+1}\|^2-\|x_{k+1}-x_k\|^2]-\theta_k\langle \Delta\mathfrak{O}_k, x_{k+1}-x\rangle.$$

Then from the definition of $\Delta\mathfrak{O}_k$ we have

$$\begin{aligned}\langle \mathfrak{F}_{k+1}+\eta_k\mathfrak{H}_{k+1}, x_{k+1}-x\rangle \leq& \tfrac{1}{2\gamma_k}[\|x-x_k\|^2-\|x-x_{k+1}\|^2-\|x_{k+1}-x_k\|^2]\\ &+\langle \Delta\mathfrak{O}_{k+1}, x_{k+1}-x\rangle-\theta_k\langle \Delta\mathfrak{O}_k, x_k-x\rangle-\theta_k\langle \Delta\mathfrak{O}_k, x_{k+1}-x_k\rangle.\end{aligned} \tag{16}$$

Note that from definition of δ_k^F and δ_k^H, and from strong monotonicity of H with modules μ_H, we have the following relations

$$\begin{aligned}\langle \mathfrak{F}_{k+1}+\eta_k\mathfrak{H}_{k+1}, x_{k+1}-x\rangle &= \langle F(x_{k+1})+\eta_k H_{k+1}, x_{k+1}-x\rangle+\langle \delta_{k+1}^F+\eta_k\delta_{k+1}^H, x_{k+1}-x\rangle\\ &\geq \langle F(x)+\eta_k H(x), x_{k+1}-x\rangle+\langle \delta_{k+1}^F+\eta_k\delta_{k+1}^H, x_{k+1}-x\rangle+\eta_k\mu_H\|x-x_{k+1}\|^2,\end{aligned} \tag{17}$$

$$\theta_k\langle \Delta\mathfrak{O}_k, x_{k+1}-x_k\rangle = \theta_k\langle \Delta O_k, x_{k+1}-x_k\rangle+\theta_k\langle \delta_k^F-\delta_{k-1}^F+\eta_{k-1}[\delta_k^H-\delta_{k-1}^H], x_{k+1}-x_k\rangle, \tag{18}$$

where $\Delta O_k = F(x_k)-F(x_{k-1})+\eta_{k-1}[H(x_k)-H(x_{k-1})]$. Therefore, from (16), (17) and (18), we have

$$\begin{aligned}&\langle F(x)+\eta_k H(x), x_{k+1}-x\rangle \leq \tfrac{1}{2\gamma_k}[\|x-x_k\|^2-\|x_{k+1}-x_k\|^2]\\ &-(\tfrac{1}{2\gamma_k}+\eta_k\mu_H)\|x-x_{k+1}\|^2\\ &+\langle \Delta\mathfrak{O}_{k+1}, x_{k+1}-x\rangle-\theta_k\langle \Delta\mathfrak{O}_k, x_k-x\rangle-\langle \delta_{k+1}^F+\eta_k\delta_{k+1}^H, x_{k+1}-x\rangle\\ &-\theta_k\langle \Delta O_k, x_{k+1}-x_k\rangle-\theta_k\langle \delta_k^F-\delta_{k-1}^F+\eta_{k-1}[\delta_k^H-\delta_{k-1}^H], x_{k+1}-x_k\rangle.\end{aligned} \tag{19}$$

Consequently, one can derive the upper bound for -$\theta_k\langle \Delta O_k, x_{k+1}-x_k\rangle$ from (3a) and (3b) and use (19), to obtain (9). □

A.2 CONVERGENCE ANALYSIS OF THE GENERAL HVI

This section is dedicated to the detailed illustration of the proofs and materials we used to obtain the results of R-OpEx for the general problem. We divide this section into two cases depending on whether we assume strong monotonicity.

A.2.1 Convergence analysis for the monotone problem

This section provides the necessary analysis for the proof of Theorems 1, 2. Moreover, we provide a detailed illustration analysis regarding Remark 2. The following lemma states the necessary conditions to obtain the convergence results. First, let us mention the following practical propositions.

Proposition 1. *Let $\rho_1,\ldots,\rho_j$ be a sequence of elements in $\mathbb{R}^n$ and let S be a convex set in $\mathbb{R}^n$. Define the sequence $v_t, t=1,\ldots,$ as follows: $v_1\in S$ and $v_{t+1}=\operatorname{argmin}_{x\in S}\langle \rho_t, x\rangle+\frac{1}{2}\|x-v_t\|^2$. Then, for any $x\in S$ and $t\geq 0$, the following inequality holds*

$$\langle \rho_t, v_t-x\rangle \leq \tfrac{1}{2}\|x-v_t\|^2-\tfrac{1}{2}\|x-v_{t+1}\|^2+\tfrac{1}{2}\|\rho_t\|^2. \tag{20}$$

Proof. Using Lemma 2 with $g(x)=\langle \rho_t, x\rangle$, we have, due to the optimality of v_{t+1},

$$\langle \rho_t, v_{t+1}-x\rangle+\tfrac{1}{2}\|v_{t+1}-v_t\|_2^2+\tfrac{1}{2}\|x-v_{t+1}\|^2\leq \tfrac{1}{2}\|x-v_t\|^2,$$

is satisfied for all $x\in S$. The above relation and the fact $\langle \rho_t, v_t-v_{t+1}\rangle-\frac{1}{2}\|v_{t+1}-v_t\|^2\leq\frac{1}{2}\|\rho_t\|^2$, imply (20). Hence, we conclude the proof. □

Lemma 3. *Consider the sequences $\{x_k^a\}_{k\geq 1}$, as follows*

$$x_2^a = x_1^a = x_1, \qquad x_{k+1}^a := \operatorname*{argmin}_{x\in X_F^*} -\langle \gamma_{k-1}\delta_k, x\rangle + \tfrac{1}{2}\|x - x_k^a\|^2, \qquad \forall k \geq 2. \tag{21}$$

For $\delta_k = \delta_k^F + \eta_k \delta_{k+1}^H$. Suppose $\frac{\tau_k}{\gamma_k\eta_k} \leq \frac{\tau_{k-1}}{\gamma_{k-1}\eta_{k-1}}$, for $k \geq 2$. Then, we have

$$\forall x \in X_F^*, \quad \textstyle\sum_{k=1}^{K-1} \frac{\tau_k}{\eta_k}\langle \delta_{k+1}, x - x_{k+1}^a\rangle \quad \leq \frac{\tau_1}{2\gamma_1\eta_1}\|x - x_1\|^2 + \sum_{k=1}^{K-1}\frac{\gamma_k\tau_k}{2\eta_k}\|\delta_{k+1}\|^2. \tag{22}$$

Proof. Noting the definition of $\{x_k^a\}_{k\geq 2}$ in (21) and applying Proposition 1 we have

$$-\langle \gamma_k\delta_{k+1}, x_{k+1}^a - x\rangle \leq \tfrac{1}{2}\|x - x_{k+1}^a\|^2 - \tfrac{1}{2}\|x - x_{k+2}^a\|^2 + \tfrac{\gamma_k^2}{2}\|\delta_{k+1}\|^2. \tag{23}$$

Multiplying the above relation by $\frac{\tau_k}{\eta_k}$, summing it over $k = 1$ to $K-1$ given that $\frac{\tau_k}{\gamma_k\eta_k} \leq \frac{\tau_{k-1}}{\gamma_{k-1}\eta_{k-1}}$, we get (22). □

Analysis of Theorem 1

Lemma 4. *Let us consider the step-size policy that satisfies the following conditions*

$$\tfrac{\tau_k}{\gamma_k\eta_k} \leq \tfrac{\tau_{k-1}}{\gamma_{k-1}\eta_{k-1}}, \quad \tfrac{\tau_k\theta_k}{\eta_k} = \tfrac{\tau_{k-1}}{\eta_{k-1}}, \quad \theta_k(L_F^2 + \eta_k^2 L_H^2) \leq \tfrac{1}{50\gamma_k\gamma_{k-1}} \tag{24a}$$

$$\tfrac{\tau_k}{\gamma_k} \leq \tfrac{\tau_{k-1}}{\gamma_{k-1}}, \quad \tau_k\theta_k = \tau_{k-1}. \tag{24b}$$

Then, we have the optimality and feasibility bounds mentioned in Theorem 1.

Proof. First, note that (10) satisfies (24a) and (24b). Regarding Lemma 1 and taking $x = x_F^* \in X_F^*$ as an arbitrary solution to the inner problem VI(F, X), we know that $\langle F(x_F^*), x_{k+1} - x_F^*\rangle \geq 0$. Thus, noting $\eta_k \geq 0$, one can rewrite (9) as follows

$$\begin{aligned}
&\langle H(x_F^*), x_{k+1} - x_F^*\rangle \leq \tfrac{1}{2\gamma_k\eta_k}[\|x_F^* - x_k\|^2 - \|x_F^* - x_{k+1}\|^2 - \|x_{k+1} - x_k\|^2]\\
&+ \tfrac{1}{\eta_k}\langle \Delta\mathfrak{O}_{k+1}, x_{k+1} - x_F^*\rangle - \tfrac{\theta_k}{\eta_k}\langle \Delta\mathfrak{O}_k, x_k - x_F^*\rangle + \tfrac{\theta_k}{\eta_k}(L_F + \eta_{k-1}L_H)\|x_k - x_{k-1}\|\|x_{k+1} - x_k\|\\
&+ \tfrac{\theta_k}{\eta_k}(M_F + \eta_{k-1}M_H)\|x_{k+1} - x_k\|.\\
&- \tfrac{1}{\eta_k}\langle \delta_{k+1}^F + \eta_k\delta_{k+1}^H, x_{k+1} - x_F^*\rangle - \tfrac{\theta_k}{\eta_k}\langle \delta_k^F - \delta_{k-1}^F + \eta_{k-1}[\delta_k^H - \delta_{k-1}^H], x_{k+1} - x_k\rangle.
\end{aligned} \tag{25}$$

Multiplying both sides by $\tau_k \geq 0$ and summing up (25) from $k = 1$ to $K-1$, we have the following

$$\begin{aligned}
&\textstyle\sum_{k=1}^{K-1}\langle H(x_F^*), \tau_k(x_{k+1} - x_F^*)\rangle \leq \sum_{k=1}^{K-1}\frac{\tau_k}{2\gamma_k\eta_k}[\|x_F^* - x_k\|^2 - \|x_F^* - x_{k+1}\|^2]\\
&+ \textstyle\sum_{k=1}^{K-1}\frac{\tau_k}{\eta_k}\langle \Delta\mathfrak{O}_{k+1}, x_{k+1} - x_F^*\rangle - \frac{\tau_k\theta_k}{\eta_k}\langle \Delta\mathfrak{O}_k, x_k - x_F^*\rangle\\
&+ \textstyle\sum_{k=1}^{K-1}\big[\frac{\tau_k\theta_k}{\eta_k}(L_F + \eta_{k-1}L_H)\|x_k - x_{k-1}\|\|x_{k+1} - x_k\| - \frac{\tau_k}{10\gamma_k\eta_k}\|x_{k+1} - x_k\|^2\\
&- \tfrac{\tau_{k-1}}{10\gamma_{k-1}\eta_{k-1}}\|x_k - x_{k-1}\|^2\big] + \textstyle\sum_{k=1}^{K-1}\frac{\tau_k\theta_k}{\eta_k}(M_F + \eta_{k-1}M_H)\|x_{k+1} - x_k\| - \frac{\tau_k}{10\gamma_k\eta_k}\|x_{k+1} - x_k\|^2\\
&- \textstyle\sum_{k=1}^{K-1}\frac{\tau_k}{\eta_k}\langle \delta_{k+1}^F + \eta_k\delta_{k+1}^H, x_{k+1} - x_F^*\rangle\\
&- \textstyle\sum_{k=1}^{K-1}\frac{\tau_k\theta_k}{\eta_k}\langle \delta_k^F - \delta_{k-1}^F, x_{k+1} - x_k\rangle - \frac{\tau_k}{10\gamma_k\eta_k}\|x_{k+1} - x_k\|^2\\
&- \textstyle\sum_{k=1}^{K-1}\frac{\tau_k\theta_k\eta_{k-1}}{\eta_k}\langle \delta_k^H - \delta_{k-1}^H, x_{k+1} - x_k\rangle - \frac{\tau_k}{10\gamma_k\eta_k}\|x_{k+1} - x_k\|^2.
\end{aligned} \tag{26}$$

Moreover from Young's inequality ($ab \leq \frac{\epsilon}{2}a^2 + \frac{1}{2\epsilon}b^2$), we have the followings

$$\tfrac{\tau_k\theta_k}{\eta_k}(M_F + \eta_{k-1}M_H)\|x_{k+1} - x_k\| - \tfrac{\tau_k}{10\gamma_k\eta_k}\|x_{k+1} - x_k\|^2 \leq \tfrac{5\theta^2\gamma_k\tau_k}{\eta_k}(M_F^2 + \eta_{k-1}^2M_H^2), \tag{27a}$$

$$\tfrac{\tau_k\theta_k}{\eta_k}\langle \delta_k^F - \delta_{k-1}^F, x_{k+1} - x_k\rangle - \tfrac{\tau_k}{10\gamma_k\eta_k}\|x_{k+1} - x_k\|^2 \leq \tfrac{5\theta^2\gamma_k\tau_k\|\delta_k^F - \delta_{k-1}^F\|^2}{\eta_k}, \tag{27b}$$

$$\tfrac{\tau_k\theta_k\eta_{k-1}}{\eta_k}\langle \delta_k^H - \delta_{k-1}^H, x_{k+1} - x_k\rangle - \tfrac{\tau_k}{10\gamma_k\eta_k}\|x_{k+1} - x_k\|^2 \leq \tfrac{5\theta^2\gamma_k\tau_k\eta_{k-1}^2\|\delta_k^H - \delta_{k-1}^H\|^2}{\eta_k}. \tag{27c}$$

Now, from the monotonicity of H and the conditions (24a)-(24b), (26), and (27), we have

$$\begin{aligned}
&\langle H(x_F^*), \bar{x}_K - x_F^*\rangle \leq \tfrac{1}{\sum_{k=1}^{K-1}\tau_k}\Big[\big[\tfrac{\tau_1}{2\gamma_1\eta_1}\|x_F^* - x_1\|^2 - \tfrac{\tau_{K-1}}{2\gamma_{K-1}\eta_{K-1}}\|x_F^* - x_K\|^2\big] + \tfrac{\tau_{K-1}}{\eta_{K-1}}\langle \Delta\mathfrak{O}_K, x_K - x_F^*\rangle\\
&+ \textstyle\sum_{k=1}^{K-1}\frac{5\theta^2\gamma_k\tau_k}{\eta_k}(M_F^2 + \|\delta_k^F - \delta_{k-1}^F\|^2 + \eta_{k-1}^2(M_H^2 + \|\delta_k^F - \delta_{k-1}^F\|^2))\\
&- \textstyle\sum_{k=1}^{K-1}\frac{\tau_k}{\eta_k}\langle \delta_{k+1}^F + \eta_k\delta_{k+1}^H, x_{k+1} - x_F^*\rangle\Big].
\end{aligned} \tag{28}$$

Note $-\sum_{k=1}^{K-1}\frac{\tau_k}{\eta_k}\langle\delta_{k+1}^F+\eta_k\delta_{k+1}^H, x_{k+1}-x_F^*\rangle$ is written as below

$$\begin{aligned}-\sum_{k=1}^{K-1}\tfrac{\tau_k}{\eta_k}\langle\delta_{k+1}^F+\eta_k\delta_{k+1}^H, x_{k+1}-x_F^*\rangle =&\sum_{k=1}^{K-1}\tfrac{\tau_k}{\eta_k}\langle\delta_{k+1}^F+\eta_k\delta_{k+1}^H, x_F^*-x_{k+1}^a\rangle\\ &+\sum_{k=1}^{K-1}\tfrac{\tau_k}{\eta_k}\langle\delta_{k+1}^F+\eta_k\delta_{k+1}^H, x_{k+1}^a-x_{k+1}\rangle,\end{aligned}\tag{29}$$

where x_{k+1}^a is a sequence defined in Lemma 3. Moreover, let us consider $\xi_{[k]}:=(\xi_1,\ldots,\xi_k)$. Thus, from (8) and (21), one can see that x_{k+1} and x_{k+1}^a depend on $\xi_{[k]}$. Therefore, from (7a), (7b), and from the tower property of expectations, we have the following relation

$$\mathbb{E}[\langle\delta_{k+1}^F+\eta_k\delta_{k+1}^H, x_{k+1}^a-x_{k+1}\rangle]=\mathbb{E}[\mathbb{E}[\langle\delta_{k+1}^F+\eta_k\delta_{k+1}^H, x_{k+1}^a-x_{k+1}\rangle|\xi_{[k]}]]=0.\tag{30}$$

Furthermore, from (22), we can have the following upper bound for the first summation in the right-hand side of (29).

$$\begin{aligned}\mathbb{E}[\sum_{k=1}^{K-1}\tfrac{\tau_k}{\eta_k}\langle\delta_{k+1}^F+\eta_k\delta_{k+1}^H, x_F^*-x_{k+1}^a\rangle]&\le\mathbb{E}[\tfrac{\tau_1}{2\gamma_1\eta_1}\|x_F^*-x_1\|^2]+\mathbb{E}[\sum_{k=1}^{K-1}\tfrac{\gamma_k\tau_k}{2\eta_k}\|\delta_{k+1}^F+\eta_k\delta_{k+1}^H\|^2]\\ &\le\tfrac{\tau_1}{\gamma_1\eta_1}D_X^2+\sum_{k=1}^{K-1}\tfrac{\gamma_k\tau_k(\sigma_F^2+\eta_k^2\sigma_H^2)}{\eta_k},\end{aligned}\tag{31}$$

where the last inequality comes from $\|a+b\|^2\le 2\|a\|^2+2\|b\|^2$, (7a) and (7b). Therefore by taking the expectations from both sides of (28) and in view of (29)-(31), $\|a+b\|^2\le 2\|a\|^2+2\|b\|^2$, Young's inequality for $\langle\Delta\mathfrak{O}_K, x_K-x_F^*\rangle$, $\tau_k=1$ and the third condition of (24a) in (28), we obtain the following

$$\begin{aligned}\mathbb{E}[\langle H(x_F^*),\bar{x}_K-x_F^*\rangle]\le&\tfrac{2}{K\gamma_1\eta_1}D_X^2+\tfrac{5\gamma_{K-1}}{K\eta_{K-1}}(M_F^2+2\sigma_F^2+\eta_{K-1}^2(M_H^2+2\sigma_H^2))\\ &+\sum_{k=1}^{K-1}\tfrac{5\theta^2\gamma_k}{K\eta_k}(M_F^2+2\sigma_F^2+\eta_{k-1}^2(M_H^2+2\sigma_H^2))+\sum_{k=1}^{K-1}\tfrac{\gamma_k(\sigma_F^2+\eta_k^2\sigma_H^2)}{K\eta_k}.\end{aligned}\tag{32}$$

Using the step-size policy in (10), one can derive the following bound for (32).

$$\begin{aligned}\mathbb{E}[\langle H(x_F^*),\bar{x}_K-x_F^*\rangle]\le&D_X\Bigg[16D_X(\tfrac{L_F}{K^{3/4}}+\tfrac{L_H}{K})+\tfrac{2\sqrt{M_F^2+2\sigma_F^2+\eta^2(M_H^2+2\sigma_H^2)}}{K^{1/4}}\\ &+\tfrac{5(M_F^2+2\sigma_F^2+\eta^2(M_H^2+2\sigma_H^2))+\sigma_F^2+\eta^2\sigma_H^2}{8K^{-1/4}(L_F+\eta L_H)+K^{1/4}\sqrt{M_F^2+2\sigma_F^2+\eta^2(M_H^2+2\sigma_H^2)}}\\ &+\tfrac{5(M_F^2+2\sigma_F^2+\eta^2(M_H^2+2\sigma_H^2))}{8K^{3/4}(L_F+\eta L_H)+K^{5/4}\sqrt{M_F^2+2\sigma_F^2+\eta^2(M_H^2+\sigma_H^2)}}\Bigg].\end{aligned}\tag{33}$$

Therefore, by taking the maximum on both sides of (33) concerning the set X_F^* and using the definition of the optimality gap function in (6), one obtains the following

$$\begin{aligned}\mathbb{E}[\mathrm{Gap}(\bar{x}_K,H,X_F^*)]\le&D_X\Bigg[16D_X(\tfrac{L_F}{K^{3/4}}+\tfrac{L_H}{K})+\tfrac{2\sqrt{M_F^2+2\sigma_F^2+\eta^2(M_H^2+2\sigma_H^2)}}{K^{1/4}}\\ &+\tfrac{5(M_F^2+2\sigma_F^2+\eta^2(M_H^2+2\sigma_H^2))+\sigma_F^2+\eta^2\sigma_H^2}{8K^{-1/4}(L_F+\eta L_H)+K^{1/4}\sqrt{M_F^2+2\sigma_F^2+\eta^2(M_H^2+2\sigma_H^2)}}\\ &+\tfrac{5(M_F^2+2\sigma_F^2+\eta^2(M_H^2+2\sigma_H^2))}{8K^{3/4}(L_F+\eta L_H)+K^{5/4}\sqrt{M_F^2+2\sigma_F^2+\eta^2(M_H^2+\sigma_H^2)}}\Bigg].\end{aligned}\tag{34}$$

Now, let us move to the feasibility gap function at the inner level. Considering Lemma 1 and from Cauchy-Schwarz inequality, we obtain the following

$$\begin{aligned}\langle F(x),x_{k+1}-x\rangle\le&\tfrac{1}{2\gamma_k}[\|x-x_k\|^2-\|x-x_{k+1}\|^2-\|x_{k+1}-x_k\|^2]+2C_HD_X\eta_k\\ &+\langle\Delta\mathfrak{O}_{k+1},x_{k+1}-x\rangle-\theta_k\langle\Delta\mathfrak{O}_k,x_k-x\rangle\\ &+\theta_k(L_F+\eta_{k-1}L_H)\|x_k-x_{k-1}\|\|x_{k+1}-x_k\|\\ &+\theta_k(M_F+\eta_{k-1}M_H)\|x_{k+1}-x_k\|\\ &-\langle\delta_{k+1}^F+\eta_k\delta_{k+1}^H,x_{k+1}-x\rangle-\theta_k\langle\delta_k^F-\delta_{k-1}^F+\eta_{k-1}[\delta_k^H-\delta_{k-1}^H],x_{k+1}-x_k\rangle.\end{aligned}\tag{35}$$

Using a similar approach to obtain an optimality gap and using the conditions in (24a)-(24b), we obtain the following

$$\begin{aligned}\mathbb{E}[\langle F(x),\bar{x}_K-x\rangle]\le&\tfrac{2}{K\gamma_1}D_X^2+\tfrac{5\gamma_{K-1}}{K}(M_F^2+2\sigma_F^2+\eta_{K-1}^2(M_H^2+\sigma_H^2))+\tfrac{2C_HD_X\sum_{k=1}^{K-1}\eta_k}{K}\\ &+\sum_{k=1}^{K-1}\tfrac{5\theta^2\gamma_k}{K}(M_F^2+2\sigma_F^2+\eta_k^2(M_H+2\sigma_H^2))+\sum_{k=1}^{K-1}\tfrac{\gamma_k(\sigma_F^2+\eta_k^2\sigma_H^2)}{K}.\end{aligned}\tag{36}$$

From the step-size policy in (10), one can drive the following

$$\begin{aligned}\mathbb{E}[\langle F(x), \bar{x}_K - x\rangle] \leq & D_X\Bigg[16D_X\left(\frac{L_F}{K} + \frac{L_H}{K^{5/4}}\right) + \frac{2\sqrt{M_F^2+2\sigma_F^2+\eta^2(M_H^2+2\sigma_H^2)}}{K^{1/2}} \\ & + \frac{5(M_F^2+2\sigma_F^2+\eta^2(M_H^2+2\sigma_H^2))+\sigma_F^2+\eta^2\sigma_H^2}{8(L_F+\eta L_H)+K^{1/2}\sqrt{M_F^2+2\sigma_F^2+\eta^2(M_H^2+2\sigma_H^2)}} \\ & + \frac{5(M_F^2+2\sigma_F^2+\eta^2(M_H^2+\sigma_H^2))}{8K(L_F+\eta L_H)+K^{3/2}\sqrt{M_F^2+2\sigma_F^2+\eta^2(M_H^2+2\sigma_H^2)}} + \frac{2C_H}{K^{1/4}}\Bigg].\end{aligned} \tag{37}$$

Taking the maximum from both sides of (37) concerning set X, we obtain the following feasibility bound

$$\begin{aligned}\mathbb{E}[\mathrm{Gap}(\bar{x}_K, F, X)] \leq & D_X\Bigg[16D_X\left(\frac{L_F}{K} + \frac{L_H}{K^{5/4}}\right) + \frac{2\sqrt{M_F^2+2\sigma_F^2+\eta^2(M_H^2+2\sigma_H^2)}}{K^{1/2}} \\ & + \frac{5(M_F^2+2\sigma_F^2+\eta^2(M_H^2+2\sigma_H^2))+\sigma_F^2+\eta^2\sigma_H^2}{8(L_F+\eta L_H)+K^{1/2}\sqrt{M_F^2+2\sigma_F^2+\eta^2(M_H^2+2\sigma_H^2)}} \\ & + \frac{5(M_F^2+2\sigma_F^2+\eta^2(M_H^2+\sigma_H^2))}{8K(L_F+\eta L_H)+K^{3/2}\sqrt{M_F^2+2\sigma_F^2+\eta^2(M_H^2+2\sigma_H^2)}} + \frac{2C_H}{K^{1/4}}\Bigg].\end{aligned} \tag{38}$$

For the lower bound in (11), note that $\bar{x}_K \in X$. From Definition 2, and for any $\tilde{x} \in X_F^*$, one can write

$$\langle H(\tilde{x}), \bar{x}_K - \tilde{x}\rangle \leq \max_{x \in X_F^*} \langle H(x), \bar{x}_K - x\rangle = \mathrm{Gap}(\bar{x}_K, H, X_F^*)$$

From Cauchy-Schwarz inequality and Assumption (4b), we have

$$-B_H\|\bar{x}_K - \tilde{x}\| \leq \mathrm{Gap}(\bar{x}_K, H, X_F^*).$$

Letting $\tilde{x} = \mathbf{proj}_{X_F^*}(\bar{x}_K)$ and taking the expectation from both sides, we obtain the results. □

Analysis of Theorem 2

Proof. Let us take $x_F^* \in X_F^*$ as an arbitrary vector. Then from Definition 3, one can obtain

$$\mathbb{E}[\langle F(x_F^*), \bar{x}_K - x_F^*\rangle] \geq \alpha\mathbb{E}[\mathrm{dist}(\bar{x}_K, X_F^*)].$$

Therefore, from the definition of the feasibility gap function in (5), we have the following

$$\mathrm{Gap}(\bar{x}_K, F, X) = \max_{x \in X} \mathbb{E}[\langle F(x), \bar{x}_K - x\rangle] \geq \mathbb{E}[\langle F(x_F^*), \bar{x}_K - x_F^*\rangle] \geq \alpha\mathbb{E}[\mathrm{dist}(\bar{x}_K, X_F^*)]. \tag{39}$$

From (11), (12) and (39), we obtain (13). □

Analysis of Remark 2 The following Theorem formally states the rates concerning Remark 2.

Theorem 6. *Suppose VI(F, X) is α-weakly sharp, and we have the following step-size policy*

$$\tau_k = \tau = 1, \quad \theta_k = \theta = 1, \quad \eta_k = \eta \leq \frac{\alpha}{2\|H(x^*)\|}, \quad \gamma_k = \gamma = \frac{D_X}{8D_X(L_F+\eta L_H)+\sqrt{K(M_F^2+2\sigma_F^2+\eta^2(M_H^2+2\sigma_H^2))}}, \tag{40}$$

then, we have the following

$$\begin{aligned}\mathbb{E}[Gap(\bar{x}_K, H, X_F^*)] \leq & \frac{\|H(x^*)\|D_X}{\alpha}\Bigg[16D_X\left(\frac{L_F}{K} + \frac{L_H}{K^{5/4}}\right) + \frac{2\sqrt{M_F^2+2\sigma_F^2+\eta^2(M_H^2+2\sigma_H^2)}}{K^{1/2}} \\ & + \frac{2[5(M_F^2+2\sigma_F^2+\eta^2(M_H^2+2\sigma_H^2))+\sigma_F^2+\eta^2\sigma_H^2]}{8D_X(L_F+\eta L_H)+K^{1/2}\sqrt{M_F^2+2\sigma_F^2+\eta^2(M_H^2+2\sigma_H^2)}} \\ & + \frac{10(M_F^2+2\sigma_F^2+\eta^2(M_H^2+2\sigma_H^2))}{8KD_X(L_F+\eta L_H)+K^{3/2}\sqrt{M_F^2+2\sigma_F^2+\eta^2(M_H^2+2\sigma_H^2)}}\Bigg].\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}[dist(\bar{x}_K, X_F^*)] \leq & \frac{D_X}{\alpha}\Bigg[32D_X\left(\frac{L_F}{K} + \frac{L_H}{K^{5/4}}\right) + \frac{4\sqrt{M_F^2+2\sigma_F^2+\eta^2(M_H^2+2\sigma_H^2)}}{K^{1/2}} \\ & + \frac{10(M_F^2+2\sigma_F^2+\eta^2(M_H^2+2\sigma_H^2))+2(\sigma_F^2+\eta^2\sigma_H^2)}{8D_X(L_F+\eta L_H)+K^{1/2}\sqrt{M_F^2+2\sigma_F^2+\eta^2(M_H^2+2\sigma_H^2)}} \\ & + \frac{10(M_F^2+2\sigma_F^2+\eta^2(M_H^2+2\sigma_H^2))}{8KD_X(L_F+\eta L_H)+K^{3/2}\sqrt{M_F^2+2\sigma_F^2+\eta^2(M_H^2+2\sigma_H^2)}}\Bigg].\end{aligned}$$

Proof. Take an arbitrary solution $x^* \in X_H^*$ where X_H^* denotes the solution set of problem (2). Using Lemma

1 and the similar relations we used in Lemma 4, we have the following

$$\begin{aligned}\mathbb{E}[\langle F(x^*)+\eta H(x^*),\bar{x}_K-x^*\rangle] \leq & \tfrac{2}{K\gamma_1}D_X^2+\tfrac{5\gamma_{K-1}\theta_{K-1}^2}{K}(M_F^2+2\sigma_F^2+\eta_{K-1}^2(M_H^2+2\sigma_H^2))\\ &+\textstyle\sum_{k=1}^{K-1}\tfrac{5\theta^2\gamma_k}{K}(M_F^2+2\sigma_F^2+\eta_{k-1}^2(M_H^2+\sigma_H^2))+\sum_{k=1}^{K-1}\tfrac{\gamma_k(\sigma_F^2+\eta_k^2\sigma_H^2)}{K}.\end{aligned} \tag{41}$$

Given that $x^* \in X_F^*$ and from Definition 3, we have

$$\mathbb{E}[\langle F(x^*),\bar{x}_K-x^*\rangle] \geq \alpha\mathbb{E}[\text{dist}(\bar{x}_K, X*_F)].$$

Then

$$\begin{aligned}\mathbb{E}[\langle H(x^*),\bar{x}_K-x^*\rangle] =&\mathbb{E}[\langle H(x^*),\bar{x}_K-\mathbf{proj}_{X_F^*}(\bar{x}_K)+\mathbf{proj}_{X_F^*}(\bar{x}_K)-x^*\rangle]\\ \geq& -\|H(x^*)\|\mathbb{E}[\|\bar{x}_K-\mathbf{proj}_{X_F^*}(\bar{x}_K)\|].\end{aligned} \tag{42}$$

Note that given $\mathbf{proj}_{X_F^*}(\bar{x}_K) \in X_F^*$ and x^* solves $\text{VI}(H, X_F^*)$, we know $\mathbb{E}[\langle H(x^*), \mathbf{proj}_{X_F^*}(\bar{x}_K)-x^*\rangle] \geq 0$. Therefore from (41), we have

$$\begin{aligned}\mathbb{E}[\alpha\text{dist}(\bar{x}_K,X_F^*)-\eta\|H(x^*)\|\text{dist}(\bar{x}_K,X_F^*)] \leq & \tfrac{2D_X^2}{K\gamma_1}+\tfrac{5\gamma_{K-1}\theta_{K-1}^2}{K}(M_F^2+2\sigma_F^2+\eta_{K-1}^2(M_H^2+2\sigma_H^2))\\ &+\textstyle\sum_{k=1}^{K-1}\tfrac{5\theta^2\gamma_k}{K}(M_F^2+2\sigma_F^2+\eta_k^2(M_H^2+2\sigma_H^2))+\sum_{k=1}^{K-1}\tfrac{\gamma_k(\sigma_F^2+\eta_k^2\sigma_H^2)}{K}.\end{aligned} \tag{43}$$

Taking step-size policy in (40), we have

$$\begin{aligned}\mathbb{E}[\text{dist}(\bar{x}_K,X_F^*)] \leq & \tfrac{D_X}{\alpha}\Bigg[32D_X(\tfrac{L_F}{K}+\tfrac{L_H}{K^{5/4}})+\tfrac{4\sqrt{M_F^2+2\sigma_F^2+\eta^2(M_H^2+2\sigma_H^2)}}{K^{1/2}}\\ &+\tfrac{10(M_F^2+2\sigma_F^2+\eta^2(M_H^2+2\sigma_H^2))+2(\sigma_F^2+\eta^2\sigma_H^2)}{8D_X(L_F+\eta L_H)+K^{1/2}\sqrt{M_F^2+2\sigma_F^2+\eta^2(M_H^2+2\sigma_H^2)}}\\ &+\tfrac{10(M_F^2+2\sigma_F^2+\eta^2(M_H^2+2\sigma_H^2))}{8KD_X(L_F+\eta L_H)+K^{3/2}\sqrt{M_F^2+2\sigma_F^2+\eta^2(M_H^2+2\sigma_H^2)}}\Bigg]\end{aligned}$$

Moreover, from (32) and applying the step-size policy in (40), one can get the following

$$\begin{aligned}\mathbb{E}[\langle H(x_F^*),\bar{x}_K-x_F^*\rangle] \leq & \tfrac{\|H(x^*)\|D_X}{\alpha}\Bigg[16D_X(\tfrac{L_F}{K}+\tfrac{L_H}{K^{5/4}})+\tfrac{2\sqrt{M_F^2+2\sigma_F^2+\eta^2(M_H^2+2\sigma_H^2)}}{K^{1/2}}\\ &+\tfrac{2[5(M_F^2+2\sigma_F^2+\eta^2(M_H^2+2\sigma_H^2))+\sigma_F^2+\eta^2\sigma_H^2]}{8D_X(L_F+\eta L_H)+K^{1/2}\sqrt{M_F^2+2\sigma_F^2+\eta^2(M_H^2+2\sigma_H^2)}}\\ &+\tfrac{10(M_F^2+2\sigma_F^2+\eta^2(M_H^2+2\sigma_H^2))}{8KD_X(L_F+\eta L_H)+K^{3/2}\sqrt{M_F^2+2\sigma_F^2+\eta^2(M_H^2+2\sigma_H^2)}}\Bigg].\end{aligned} \tag{44}$$

Taking the maximum from both sides, we obtain

$$\begin{aligned}\mathbb{E}[\text{Gap}(\bar{x}_K,H,X_F^*)] \leq & \tfrac{\|H(x^*)\|D_X}{\alpha}\Bigg[16D_X(\tfrac{L_F}{K}+\tfrac{L_H}{K^{5/4}})+\tfrac{2\sqrt{M_F^2+2\sigma_F^2+\eta^2(M_H^2+2\sigma_H^2)}}{K^{1/2}}\\ &+\tfrac{2[5(M_F^2+2\sigma_F^2+\eta^2(M_H^2+2\sigma_H^2))+\sigma_F^2+\eta^2\sigma_H^2]}{8D_X(L_F+\eta L_H)+K^{1/2}\sqrt{M_F^2+2\sigma_F^2+\eta^2(M_H^2+2\sigma_H^2)}}\\ &+\tfrac{10(M_F^2+2\sigma_F^2+\eta^2(M_H^2+2\sigma_H^2))}{8KD_X(L_F+\eta L_H)+K^{3/2}\sqrt{M_F^2+2\sigma_F^2+\eta^2(M_H^2+2\sigma_H^2)}}\Bigg].\end{aligned}$$

□

A.2.2 Convergence results for strongly monotone problem

Lemma 5. *Let us assume problem* (2) *is strongly monotone* ($\mu_H > 0$). *Suppose we have the following step-size policy*

$$\tau_k = k+1, \quad \theta_k = \tfrac{k}{k+1}, \quad \eta_k = \eta = K^{-\frac{1}{4}}, \quad \gamma_k = \gamma = \tfrac{D_X}{8D_X(L_F+\eta L_H)+\sqrt{K(M_F^2+2\sigma_F^2+\eta^2(M_H^2+2\sigma_H^2))}}, \tag{45}$$

Then, for $K \geq \frac{1}{2\gamma_k\eta_{k-1}\mu_H}$, *Algorithm 1 gives the following optimality and feasibility gaps*

$$\begin{aligned}-B_H\mathbb{E}[dist(\bar{x}_K,X_F^*)] \leq \mathbb{E}[Gap(\bar{x}_K,H,X_F^*)] \leq & D_X\Bigg[\big(16D_X(\tfrac{L_F}{K^{7/4}}+\tfrac{L_H}{K^2})+\tfrac{2\sqrt{M_F^2+2\sigma_F^2+\eta^2(M_H^2+2\sigma_H^2)}}{K^{5/4}}\big)\\ &+\tfrac{5(M_F^2+2\sigma_F^2+\eta^2(M_H^2+2\sigma_H^2))+\sigma_F^2+\eta^2\sigma_H^2}{8K^{3/4}D_X(L_F+\eta L_H)+K^{5/4}\sqrt{M_F^2+2\sigma_F^2+\eta^2(M_H^2+2\sigma_H^2)}}\\ &+\tfrac{5(M_F^2+2\sigma_F^2+\eta^2(M_H^2+2\sigma_H^2))}{8K^{7/4}D_X(L_F+\eta L_H)+K^{9/4}\sqrt{M_F^2+2\sigma_F^2+\eta^2(M_H^2+\sigma_H^2)}}\Bigg],\end{aligned} \tag{46}$$

and

$$\mathbb{E}[Gap(\bar{x}_K, F, X)] \leq D_X \Bigg[\big(16 D_X (\tfrac{L_F}{K^2} + \tfrac{L_H}{K^{9/4}}) + \tfrac{2\sqrt{M_F^2+2\sigma_F^2+\eta^2(M_H^2+2\sigma_H^2)}}{K^{3/2}}\big) + \tfrac{5(M_F^2+2\sigma_F^2+\eta^2(M_H^2+2\sigma_H^2))+\sigma_F^2+\eta^2\sigma_H^2}{8K(L_F+\eta L_H)+K^{3/2}\sqrt{M_F^2+2\sigma_F^2+\eta^2(M_H^2+2\sigma_H^2)}} + \tfrac{5(M_F^2+2\sigma_F^2+\eta^2(M_H^2+\sigma_H^2))}{8K^2(L_F+\eta L_H)+K^{5/2}\sqrt{M_F^2+2\sigma_F^2+\eta^2(M_H^2+2\sigma_H^2)}} + \tfrac{2C_H}{K^{5/4}} \Bigg]. \tag{47}$$

Further, the explicit lower bound for the strongly monotone case has the following form

$$\mathbb{E}[Gap(\bar{x}_K, H, X_F^*)] \geq -\tfrac{B_H}{\alpha} \Bigg[D_X \big[\big(16 D_X (\tfrac{L_F}{K^2} + \tfrac{L_H}{K^{9/4}}) + \tfrac{2\sqrt{M_F^2+2\sigma_F^2+\eta^2(M_H^2+2\sigma_H^2)}}{K^{3/2}}\big) + \tfrac{5(M_F^2+2\sigma_F^2+\eta^2(M_H^2+2\sigma_H^2))+\sigma_F^2+\eta^2\sigma_H^2}{8KD_X(L_F+\eta L_H)+K^{3/2}\sqrt{M_F^2+2\sigma_F^2+\eta^2(M_H^2+2\sigma_H^2)}} + \tfrac{5(M_F^2+2\sigma_F^2+\eta^2(M_H^2+\sigma_H^2))}{8K^2D_X(L_F+\eta L_H)+K^{5/2}\sqrt{M_F^2+2\sigma_F^2+\eta^2(M_H^2+2\sigma_H^2)}} \big] + \tfrac{2C_H}{K^{5/4}} \Bigg]. \tag{48}$$

Moreover, consider the following step-size policy

$$\tau_k = k+1, \quad \theta_k = \tfrac{k}{k+1}, \quad \eta_k = \eta = \tfrac{\alpha}{2\|H(x^*)\|}, \quad \gamma_k = \gamma = \tfrac{D_X}{8D_X(L_F+\eta L_H)+\sqrt{K(M_F^2+2\sigma_F^2+\eta^2(M_H^2+2\sigma_H^2))}}, \tag{49}$$

Then, similar to results in Theorem 6 and under the weak sharpness assumption for F, we obtain the following bounds for $K \geq \frac{1}{2\gamma_k\eta_{k-1}\mu_H}$

$$\mathbb{E}[Gap(\bar{x}_K, H, X_F^*)] \leq \tfrac{\|H(x^*)\| D_X^2}{\alpha} \Bigg[32(\tfrac{L_F}{K^2} + \tfrac{L_H}{K^{9/4}}) + \tfrac{4\sqrt{M_F^2+2\sigma_F^2+\eta^2(M_H^2+2\sigma_H^2)}}{K^{3/2}} + \tfrac{2[5(M_F^2+2\sigma_F^2+\eta^2(M_H^2+2\sigma_H^2))+\sigma_F^2+\eta^2\sigma_H^2]}{8K(L_F+\eta L_H)+K^{3/2}\sqrt{M_F^2+2\sigma_F^2+\eta^2(M_H^2+2\sigma_H^2)}} + \tfrac{10(M_F^2+2\sigma_F^2+\eta^2(M_H^2+2\sigma_H^2))}{8K^2(L_F+\eta L_H)+K^{5/2}\sqrt{M_F^2+2\sigma_F^2+\eta^2(M_H^2+2\sigma_H^2)}} \Bigg], \tag{50}$$

$$\mathbb{E}[dist(\bar{x}_K, X_F^*)] \leq \tfrac{D_X}{\alpha} \Bigg[32 D_X (\tfrac{L_F}{K^2} + \tfrac{L_H}{K^{9/4}}) + \tfrac{4\sqrt{M_F^2+2\sigma_F^2+\eta^2(M_H^2+2\sigma_H^2)}}{K^{3/2}} + \tfrac{10(M_F^2+2\sigma_F^2+\eta^2(M_H^2+2\sigma_H^2))+2(\sigma_F^2+\eta^2\sigma_H^2)}{8KD_X(L_F+\eta L_H)+K^{3/2}\sqrt{M_F^2+2\sigma_F^2+\eta^2(M_H^2+2\sigma_H^2)}} + \tfrac{10(M_F^2+2\sigma_F^2+\eta^2(M_H^2+2\sigma_H^2))}{8K^2D_X(L_F+\eta L_H)+K^{5/2}\sqrt{M_F^2+2\sigma_F^2+\eta^2(M_H^2+2\sigma_H^2)}} \Bigg], \tag{51}$$

Convergence analysis for the strongly monotone problem

Proof. Since the proof of Lemma 5 is very similar to the proof we had for Theorems 1-6, we only mention the differences. To prove (46) and (47), we have to satisfy the changes we have in conditions (24a) and (24b). In particular, the new conditions to satisfy are the following

$$\tfrac{\tau_k}{2\gamma_k\eta_k} \leq \tfrac{\tau_{k-1}}{\eta_{k-1}}(\tfrac{1}{2\gamma_{k-1}} + \eta_{k-1}\mu_H), \quad \tfrac{\tau_k\theta_k}{\eta_k} = \tfrac{\tau_{k-1}}{\eta_{k-1}}, \quad \theta_k(L_F^2 + \eta_k^2 L_H^2) \leq \tfrac{1}{50\gamma_k\gamma_{k-1}} \tag{52a}$$

$$\tfrac{\tau_k}{2\gamma_k} \leq \tau_{k-1}(\tfrac{1}{2\gamma_k}\eta_{k-1}\mu_H), \quad \tau_k\theta_k = \tau_{k-1}, \tag{52b}$$

Considering the new conditions in (52a) and (52b), and following the same procedure used in Lemma 4 and the analysis of Theorem 1—specifically with respect to the step-size policy in (45)—we derive (46) and (47). The lower bound in (48) follows from the same argument presented in Theorem 2. Finally, the upper bounds in (51) and (50) are established using the approach taken in Theorem 6. □

A.3 Convergence analysis of variable step-size policy

In this section, we analyze the convergence rates of the update policy introduced in Section 4.2 where η_k does not depend on K.

Lemma 6. *Consider the following conditions*

$$\tfrac{\tau_k\theta_k}{\eta_k} = \tfrac{\tau_{k-1}}{\eta_{k-1}}, \quad \theta_k(L_F^2 + \eta_k^2 L_H^2) \leq \tfrac{1}{50\gamma_k\gamma_{k-1}}. \tag{53}$$

Then the following step-size policy satisfies (53)

$$\tau_k = \tau = 1, \quad \eta_k = (k+1)^{-1/4}, \quad \theta_k = (\tfrac{k}{k+1})^{1/4}, \quad \gamma_k = \frac{D_X}{8D_X(L_F+\eta_k L_H)+\sqrt{k(M_F^2+2\sigma_F^2+\eta_k^2(M_H^2+2\sigma_H^2))}}, \tag{54}$$

and gives the following upper bounds for optimality and feasibility gaps

$$\begin{aligned} -B_H\mathbb{E}[dist(\bar{x}_K, X_F^*)] \le \mathbb{E}[Gap(\bar{x}_K, H, X_F^*)] &\le \frac{D_X^2+D_{X_F^*}^2}{D_X}\left[\frac{16D_X(L_F+\eta_{K-1}L_H)}{K^{3/4}} + \frac{2\sqrt{M_F^2+2\sigma_F^2+\eta_{K-1}^2(M_H^2+2\sigma_H^2)}}{K^{1/4}}\right] \\ &+ \frac{5D_X(M_F^2+2\sigma_F^2+\eta_{K-1}^2(M_H^2+2\sigma_H^2))}{8K^{3/4}D_X(L_F+\eta_{K-1}L_H)+K^{5/4}\sqrt{M_F^2+2\sigma_F^2+\eta_{K-1}(M_H^2+2\sigma_H^2)}} + \frac{D_X(20(M_F^2+2\sigma_F^2)+\sigma_F^2)}{3K^{1/4}\sqrt{M_F^2+2\sigma_F^2}} + \frac{D_X(10(M_H^2+2\sigma_H^2)+\sigma_H^2)}{K^{1/2}\sqrt{M_H^2+2\sigma_H^2}} \end{aligned} \tag{55}$$

$$\begin{aligned} \mathbb{E}[Gap(\bar{x}_K, F, X)] &\le D_X\left[\frac{32D_X(L_F+\eta_{K-1}L_H)}{K} + \frac{2\sqrt{M_F^2+2\sigma_F^2+\eta_{K-1}^2(M_H^2+2\sigma_H^2)}}{K^{1/2}}\right] \\ &+ 2\tfrac{D_X}{K}[2D_X(2L_F+(\eta_{K-1}+\eta_1)L_H)+2M_F+4\sigma_F+(\eta_{K-1}+\eta_1)(M_H+2\sigma_H)] \\ &+ 4\tfrac{D_X(K+2)^{3/4}}{K}[2D_X(L_F+\eta_1 L_H)+M_F+2\sigma_F+\eta_1(M_H+2\sigma_H)] \\ &+ \tfrac{D_X(20(M_F^2+2\sigma_F^2)+\sigma_F^2)}{3K^{1/2}\sqrt{M_F^2+2\sigma_F^2}} + \tfrac{D_X(10(M_H^2+2\sigma_H^2)+\sigma_H^2)}{K^{3/4}\sqrt{M_H^2+2\sigma_H^2}} + \tfrac{2C_H D_X}{K^{1/4}}. \end{aligned} \tag{56}$$

Additionally, similar to theorem 2, we obtain the following lower bound for optimality gap

$$\begin{aligned} \mathbb{E}[Gap(\bar{x}_K, H, X_F^*)] &\ge -\tfrac{B_H}{\alpha}\Bigg[D_X\big[\tfrac{32D_X(L_F+\eta_{K-1}L_H)}{K} + \tfrac{2\sqrt{M_F^2+2\sigma_F^2+\eta_{K-1}^2(M_H^2+2\sigma_H^2)}}{K^{1/2}}\big] \\ &+ 2\tfrac{D_X}{K}[2D_X(2L_F+(\eta_{K-1}+\eta_1)L_H)+2M_F+4\sigma_F+(\eta_{K-1}+\eta_1)(M_H+2\sigma_H)] \\ &+ 4\tfrac{D_X(K+2)^{3/4}}{K}[2D_X(L_F+\eta_1 L_H)+M_F+2\sigma_F+\eta_1(M_H+2\sigma_H)] \\ &+ \tfrac{D_X(20(M_F^2+2\sigma_F^2)+\sigma_F^2)}{3K^{1/2}\sqrt{M_F^2+2\sigma_F^2}} + \tfrac{D_X(10(M_H^2+2\sigma_H^2)+\sigma_H^2)}{K^{3/4}\sqrt{M_H^2+2\sigma_H^2}} + \tfrac{2C_H D_X}{K^{1/4}}\Bigg]. \end{aligned} \tag{57}$$

Proof. First, it is easy to see that step-size policy in (54) satisfies (53). Therefore, similar to (25) and multiplying both sides of (25) we obtain the following

$$\begin{aligned} &\langle H(x_F^*), \tau_k(x_{k+1}-x_F^*)\rangle \le \tfrac{\tau_k}{2\gamma_k\eta_k}[\|x_F^*-x_k\|^2 - \|x_F^*-x_{k+1}\|^2 - \|x_{k+1}-x_k\|^2] \\ &+ \tfrac{\tau_k}{\eta_k}\langle\Delta\mathfrak{O}_{k+1}, x_{k+1}-x_F^*\rangle - \tfrac{\tau_k\theta_k}{\eta_k}\langle\Delta\mathfrak{O}_k, x_k-x_F^*\rangle + \tfrac{\tau_k\theta_k}{\eta_k}(L_F+\eta_{k-1}L_H)\|x_k-x_{k-1}\|\|x_{k+1}-x_k\| \\ &+ \tfrac{\tau_k\theta_k}{\eta_k}(M_F+\eta_{k-1}M_H)\|x_{k+1}-x_k\| \\ &- \tfrac{\tau_k}{\eta_k}\langle\delta_{k+1}^F+\eta_k\delta_{k+1}^H, x_{k+1}-x_F^*\rangle - \tfrac{\tau_k\theta_k}{\eta_k}\langle\delta_k^F-\delta_{k-1}^F+\eta_{k-1}[\delta_k^H-\delta_{k-1}^H], x_{k+1}-x_k\rangle. \end{aligned} \tag{58}$$

Next, considering $-\frac{\tau_k}{\eta_k}\langle\delta_{k+1}^F+\eta_k\delta_{k+1}^H, x_{k+1}-x_F^*\rangle$, one can rewrite it as below

$$\begin{aligned} -\tfrac{\tau_k}{\eta_k}\langle\delta_{k+1}^F+\eta_k\delta_{k+1}^H, x_{k+1}-x_F^*\rangle &= -\tfrac{\tau_k}{\eta_k}\langle\delta_{k+1}^F+\eta_k\delta_{k+1}^H, x_{k+1}^a-x_F^*\rangle \\ &\quad - \tfrac{\tau_k}{\eta_k}\langle\delta_{k+1}^F+\eta_k\delta_{k+1}^H, x_{k+1}-x_{k+1}^a\rangle. \end{aligned} \tag{59}$$

Moreover, we add and subtract $\frac{\tau_{k-1}}{2\gamma_{k-1}\eta_{k-1}}\|x_F^*-x_k\|$ to (58). Thus, we have the following

$$\begin{aligned} &\langle H(x_F^*), \tau_k(x_{k+1}-x_F^*)\rangle \le \tfrac{\tau_{k-1}}{2\gamma_{k-1}\eta_{k-1}}\|x_F^*-x_k\|^2 - \tfrac{\tau_k}{2\gamma_k\eta_k}\|x_F^*-x_{k+1}\|^2 - \tfrac{\tau_k}{2\gamma_k\eta_k}\|x_{k+1}-x_k\|^2 \\ &+ \tfrac{\tau_k}{\eta_k}\langle\Delta\mathfrak{O}_{k+1}, x_{k+1}-x_F^*\rangle - \tfrac{\tau_k\theta_k}{\eta_k}\langle\Delta\mathfrak{O}_k, x_k-x_F^*\rangle + \tfrac{\tau_k\theta_k}{\eta_k}(L_F+\eta_{k-1}L_H)\|x_k-x_{k-1}\|\|x_{k+1}-x_k\| \\ &+ \tfrac{\tau_k\theta_k}{\eta_k}(M_F+\eta_{k-1}M_H)\|x_{k+1}-x_k\| - \tfrac{\tau_k\theta_k}{\eta_k}\langle\delta_k^F-\delta_{k-1}^F+\eta_{k-1}[\delta_k^H-\delta_{k-1}^H], x_{k+1}-x_k\rangle \\ &- \tfrac{\tau_k}{\eta_k}\langle\delta_{k+1}^F+\eta_k\delta_{k+1}^H, x_{k+1}^a-x_F^*\rangle - \tfrac{\tau_k}{\eta_k}\langle\delta_{k+1}^F+\eta_k\delta_{k+1}^H, x_{k+1}-x_{k+1}^a\rangle \\ &+ [\tfrac{\tau_k}{2\gamma_k\eta_k} - \tfrac{\tau_{k-1}}{2\gamma_{k-1}\eta_{k-1}}]\|x_F^*-x_k\|^2. \end{aligned} \tag{60}$$

Moreover, regarding (23) in Lemma 3 and by adding and subtracting $\frac{\tau_{k-1}}{2\gamma_{k-1}\eta_{k-1}}\|x_F^*-x_{k+1}^a\|$ to right-hand side of (23), one can rewrite (60) as follows

$$\begin{aligned} &\langle H(x_F^*), \tau_k(x_{k+1}-x_F^*)\rangle \le \tfrac{\tau_{k-1}}{2\gamma_{k-1}\eta_{k-1}}\|x_F^*-x_k\|^2 - \tfrac{\tau_k}{2\gamma_k\eta_k}\|x_F^*-x_{k+1}\|^2 - \tfrac{\tau_k}{2\gamma_k\eta_k}\|x_{k+1}-x_k\|^2 \\ &+ \tfrac{\tau_k}{\eta_k}\langle\Delta\mathfrak{O}_{k+1}, x_{k+1}-x_F^*\rangle - \tfrac{\tau_k\theta_k}{\eta_k}\langle\Delta\mathfrak{O}_k, x_k-x_F^*\rangle + \tfrac{\tau_k\theta_k}{\eta_k}(L_F+\eta_{k-1}L_H)\|x_k-x_{k-1}\|\|x_{k+1}-x_k\| \\ &+ \tfrac{\tau_k\theta_k}{\eta_k}(M_F+\eta_{k-1}M_H)\|x_{k+1}-x_k\| - \tfrac{\tau_k\theta_k}{\eta_k}\langle\delta_k^F-\delta_{k-1}^F+\eta_{k-1}[\delta_k^H-\delta_{k-1}^H], x_{k+1}-x_k\rangle \\ &- \tfrac{\tau_k}{\eta_k}\langle\delta_{k+1}^F+\eta_k\delta_{k+1}^H, x_{k+1}-x_{k+1}^a\rangle + [\tfrac{\tau_k}{2\gamma_k\eta_k} - \tfrac{\tau_{k-1}}{2\gamma_{k-1}\eta_{k-1}}]\|x_F^*-x_k\|^2 + [\tfrac{\tau_k}{2\gamma_k\eta_k} - \tfrac{\tau_{k-1}}{2\gamma_{k-1}\eta_{k-1}}]\|x_F^*-x_{k+1}^a\|^2 \\ &+ \tfrac{\tau_{k-1}}{2\gamma_{k-1}\eta_{k-1}}\|x_F^*-x_{k+1}^a\|^2 - \tfrac{\tau_k}{2\gamma_k\eta_k}\|x_F^*-x_{k+2}^a\|^2 + \tfrac{\tau_k\gamma_k}{2\eta_k}\|\delta_{k+1}^F+\eta_k\delta_{k+1}^H\|^2. \end{aligned} \tag{61}$$

Now, let us sum (61) from $k=1$ to $K-1$. We have

$$\begin{aligned}
&\textstyle\sum_{k=1}^{K-1}\langle H(x_F^*),\tau_k(x_{k+1}-x_F^*)\rangle \le \frac{\tau_0}{2\gamma_0\eta_0}\|x_F^*-x_1\|^2-\frac{\tau_{K-1}}{2\gamma_{K-1}\eta_{K-1}}\|x_F^*-x_K\|^2+\frac{\tau_0}{2\gamma_0\eta_0}\|x_F^*-x_2^a\|^2\\
&-\tfrac{\tau_{K-1}}{2\gamma_{K-1}\eta_{K-1}}\|x_F^*-x_{K+1}^a\|^2+\textstyle\sum_{k=1}^{K-1}\frac{\tau_k}{\eta_k}\langle\Delta\mathfrak{O}_{k+1},x_{k+1}-x_F^*\rangle-\frac{\tau_k\theta_k}{\eta_k}\langle\Delta\mathfrak{O}_k,x_k-x_F^*\rangle\\
&+\textstyle\sum_{k=1}^{K-1}[\frac{\tau_k\theta_k}{\eta_k}(L_F+\eta_{k-1}L_H)\|x_k-x_{k-1}\|\|x_{k+1}-x_k\|-\frac{\tau_k}{10\gamma_k\eta_k}\|x_{k+1}-x_k\|^2\\
&-\tfrac{\tau_{k-1}}{10\gamma_{k-1}\eta_{k-1}}\|x_k-x_{k-1}\|^2]+\textstyle\sum_{k=1}^{K-1}\frac{\tau_k\theta_k}{\eta_k}(M_F+\eta_{k-1}M_H)\|x_{k+1}-x_k\|-\frac{\tau_k}{10\gamma_k\eta_k}\|x_{k+1}-x_k\|^2\\
&-\textstyle\sum_{k=1}^{K-1}\frac{\tau_k\theta_k}{\eta_k}\langle\delta_k^F-\delta_{k-1}^F,x_{k+1}-x_k\rangle-\frac{\tau_k}{10\gamma_k\eta_k}\|x_{k+1}-x_k\|^2\\
&-\textstyle\sum_{k=1}^{K-1}\frac{\tau_k\theta_k\eta_{k-1}}{\eta_k}\langle\delta_k^H-\delta_{k-1}^H,x_{k+1}-x_k\rangle-\frac{\tau_k}{10\gamma_k\eta_k}\|x_{k+1}-x_k\|^2\\
&-\textstyle\sum_{k=1}^{K-1}\frac{\tau_k}{\eta_k}\langle\delta_{k+1}^F+\eta_k\delta_{k+1}^H,x_{k+1}-x_{k+1}^a\rangle\\
&+[\tfrac{2\tau_{K-1}}{\gamma_{K-1}\eta_{K-1}}-\tfrac{2\tau_0}{\gamma_0\eta_0}](D_X^2+D_{X_F^*}^2)\\
&+\textstyle\sum_{k=1}^{K-1}\frac{\tau_k\gamma_k}{2\eta_k}\|\delta_{k+1}^F+\eta_k\delta_{k+1}^H\|^2.
\end{aligned}$$

Note that

$$\tfrac{\tau_0}{2\gamma_0\eta_0}\|x_F^*-x_1\|^2\le\tfrac{2\tau_0}{\gamma_0\eta_0}D_X^2,\quad \tfrac{\tau_0}{2\gamma_0\eta_0}\|x_F^*-x_2^a\|^2\le\tfrac{2\tau_0}{\gamma_0\eta_0}D_{X_F^*}^2. \tag{62}$$

Then, considering conditions (53) and relations in (27a), one obtains the following

$$\begin{aligned}
&\textstyle\sum_{k=1}^{K-1}\langle H(x_F^*),\tau_k(x_{k+1}-x_F^*)\rangle\le[\frac{2\tau_{K-1}}{\gamma_{K-1}\eta_{K-1}}](D_X^2+D_{X_F^*}^2)+\frac{\tau_{K-1}}{\eta_{K-1}}\langle\Delta\mathfrak{O}_K,x_K-x_F^*\rangle\\
&+\textstyle\sum_{k=1}^{K-1}\frac{5\theta^2\gamma_k\tau_k}{\eta_k}(M_F^2+\|\delta_k^F-\delta_{k-1}^F\|^2+\eta_{k-1}^2(M_H^2+\|\delta_k^F-\delta_{k-1}^F\|^2))\\
&+\textstyle\sum_{k=1}^{K-1}\frac{\tau_k\gamma_k}{2\eta_k}\|\delta_{k+1}^F+\eta_k\delta_{k+1}^H\|^2-\sum_{k=1}^{K-1}\frac{\tau_k}{\eta_k}\langle\delta_{k+1}^F+\eta_k\delta_{k+1}^H,x_{k+1}-x_{k+1}^a\rangle.
\end{aligned}\tag{63}$$

Further, considering (30) and using the similar arguments we used to obtain (32), we have the following bound for the optimality

$$\begin{aligned}
&\mathbb{E}[\langle H(x_F^*),\bar{x}_K-x_F^*\rangle]\le[\tfrac{2\tau_{K-1}}{K\gamma_{K-1}\eta_{K-1}}](D_X^2+D_{X_F^*}^2)+\tfrac{5\gamma_{K-1}}{K\eta_{K-1}}(M_F^2+2\sigma_F^2+\eta_{K-1}^2(M_H^2+2\sigma_H^2))\\
&+\textstyle\sum_{k=1}^{K-1}\frac{5\theta^2\gamma_k\tau_k}{K\eta_k}(M_F^2+2\sigma_F^2+\eta_{k-1}^2(M_H^2+2\sigma_H^2))+\sum_{k=1}^{K-1}\frac{\tau_k\gamma_k}{2K\eta_k}(\sigma_F^2+\eta_k\sigma_H^2).
\end{aligned}\tag{64}$$

Using the updating rule in (54) we obtain the following

$$\begin{aligned}
&\mathbb{E}[\langle H(x_F^*),\bar{x}_K-x_F^*\rangle]\le\tfrac{D_X^2+D_{X_F^*}^2}{D_X}\left[\tfrac{16D_X(L_F+\eta_{K-1}L_H)}{K^{3/4}}+\tfrac{2\sqrt{M_F^2+2\sigma_F^2+\eta_{K-1}^2(M_H^2+2\sigma_H^2)}}{K^{1/4}}\right]\\
&+\tfrac{5D_X(M_F^2+2\sigma_F^2+\eta_{K-1}^2(M_H^2+2\sigma_H^2))}{8K^{3/4}D_X(L_F+\eta_{K-1}L_H)+K^{5/4}\sqrt{M_F^2+2\sigma_F^2+\eta_{K-1}(M_H^2+2\sigma_H^2)}}+\tfrac{D_X(20(M_F^2+2\sigma_F^2)+\sigma_F^2)}{3K^{1/4}\sqrt{M_F^2+2\sigma_F^2}}+\tfrac{D_X(10(M_H^2+2\sigma_H^2)+\sigma_H^2)}{K^{1/2}\sqrt{M_H^2+2\sigma_H^2}}.
\end{aligned}\tag{65}$$

Finally, taking the maximum concerning the set X_F^* and the definition of the optimality gap function, we obtain (55). The left-hand side of optimality uses the same argument we mentioned in Theorem 1. Now, we move to the feasibility gap. Considering Lemma 1 and using the same procedure we used to get (61), we have

$$\begin{aligned}
&\langle F(x),\tau_k(x_{k+1}-x)\rangle\le\tfrac{\tau_{k-1}}{2\gamma_{k-1}}\|x-x_k\|^2-\tfrac{\tau_k}{2\gamma_k}\|x-x_{k+1}\|^2-\tfrac{\tau_k}{2\gamma_k}\|x_{k+1}-x_k\|^2\\
&+\tau_k\langle\Delta\mathfrak{O}_{k+1},x_{k+1}-x\rangle-\tau_k\theta_k\langle\Delta\mathfrak{O}_k,x_k-x\rangle+\tau_k\theta_k(L_F+\eta_{k-1}L_H)\|x_k-x_{k-1}\|\|x_{k+1}-x_k\|\\
&+\tau_k\theta_k(M_F+\eta_{k-1}M_H)\|x_{k+1}-x_k\|-\tau_k\theta_k\langle\delta_k^F-\delta_{k-1}^F+\eta_{k-1}[\delta_k^H-\delta_{k-1}^H],x_{k+1}-x_k\rangle\\
&-\tau_k\langle\delta_{k+1}^F+\eta_k\delta_{k+1}^H,x_{k+1}-x_{k+1}^a\rangle+[\tfrac{\tau_k}{2\gamma_k}-\tfrac{\tau_{k-1}}{2\gamma_{k-1}}]\|x-x_k\|^2+[\tfrac{\tau_k}{2\gamma_k}-\tfrac{\tau_{k-1}}{2\gamma_{k-1}}]\|x-x_{k+1}^a\|^2\\
&+\tfrac{\tau_{k-1}}{2\gamma_{k-1}}\|x-x_{k+1}^a\|^2-\tfrac{\tau_k}{2\gamma_k}\|x-x_{k+2}^a\|^2+\tfrac{\tau_k\gamma_k}{2}\|\delta_{k+1}^F+\eta_k\delta_{k+1}^H\|^2+2C_HD_X\eta_k.
\end{aligned}\tag{66}$$

Summing (66) from $k=1$ to $K-1$, in view of (62) and using the same argument we used in (63) we have

$$\begin{aligned}
&\textstyle\sum_{k=1}^{K-1}\langle F(x),\tau_k(x_{k+1}-x)\rangle\le\frac{4\tau_{K-1}}{\gamma_{K-1}}D_X^2+\sum_{k=1}^{K-1}\tau_k\langle\Delta\mathfrak{O}_{k+1},x_{k+1}-x\rangle-\tau_k\theta_k\langle\Delta\mathfrak{O}_k,x_k-x\rangle\\
&+\textstyle\sum_{k=1}^{K-1}5\theta^2\gamma_k\tau_k(M_F^2+\|\delta_k^F-\delta_{k-1}^F\|^2+\eta_{k-1}^2(M_H^2+\|\delta_k^F-\delta_{k-1}^F\|^2))\\
&+\textstyle\sum_{k=1}^{K-1}\frac{\tau_k\gamma_k}{2}\|\delta_{k+1}^F+\eta_k\delta_{k+1}^H\|^2+\sum_{k=1}^{K-1}2C_HD_X\eta_k-\sum_{k=1}^{K-1}\frac{\tau_k}{\eta_k}\langle\delta_{k+1}^F+\eta_k\delta_{k+1}^H,x_{k+1}-x_{k+1}^a\rangle.
\end{aligned}\tag{67}$$

Moreover, we need to simplify $\sum_{k=1}^{K-1}\tau_k\langle\Delta\mathfrak{O}_{k+1},x_{k+1}-x\rangle-\tau_k\theta_k\langle\Delta\mathfrak{O}_k,x_k-x\rangle$ as below

$$\begin{aligned}
\textstyle\sum_{k=1}^{K-1}\tau_k\langle\Delta\mathfrak{O}_{k+1},x_{k+1}-x\rangle-\tau_k\theta_k\langle\Delta\mathfrak{O}_k,x_k-x\rangle&=\tau_{K-1}\langle\Delta\mathfrak{O}_K,x_K-x\rangle\\
&+\textstyle\sum_{k=1}^{K-1}(\tau_k-\tau_{k+1}\theta_{k+1})\langle\Delta\mathfrak{O}_{k+1},x_{k+1}-x\rangle-\tau_1\langle\Delta\mathfrak{O}_1,x_1-x\rangle
\end{aligned}\tag{68}$$

Further, from (68) and (54), one can bound $\mathbb{E}[\sum_{k=1}^{K-1}\tau_k\langle\Delta\mathfrak{O}_{k+1}, x_{k+1}-x\rangle - \tau_k\theta_k\langle\Delta\mathfrak{O}_k, x_k-x\rangle]$ as follows

$$\begin{aligned}\mathbb{E}[\textstyle\sum_{k=1}^{K-1}\tau_k\langle\Delta\mathfrak{O}_{k+1}, x_{k+1}-x\rangle - \tau_k\theta_k\langle\Delta\mathfrak{O}_k, x_k-x\rangle] &\le 2D_X[2D_X(2L_F+(\eta_{K-1}+\eta_1)L_H)\\ &+2M_F+4\sigma_F+(\eta_{K-1}+\eta_1)(M_H+2\sigma_H)]\\ &+2D_X[2D_X(L_F+\eta_1L_H)+M_F+2\sigma_F+\eta_1(M_H+2\sigma_H)]\textstyle\sum_{k=1}^{K-1}(1-(\frac{k+1}{k+2})^{1/4}).\end{aligned} \tag{69}$$

Note that $\sum_{k=1}^{K-1}(1-(\frac{k+1}{k+2})^{1/4}) = \sum_{k=1}^{K-1}(1-(1-\frac{1}{k+2})^{1/4}) \le 2(K+2)^{3/4}$. Therefore we can rewrite (69) below

$$\begin{aligned}\mathbb{E}[\textstyle\sum_{k=1}^{K-1}\tau_k\langle\Delta\mathfrak{O}_{k+1}, x_{k+1}-x\rangle - \tau_k\theta_k\langle\Delta\mathfrak{O}_k, x_k-x\rangle] &\le 2D_X[2D_X(2L_F+(\eta_{K-1}+\eta_1)L_H)\\ &+2M_F+4\sigma_F+(\eta_{K-1}+\eta_1)(M_H+2\sigma_H)]+2D_X[2D_X(L_F+\eta_1L_H)\\ &+M_F+2\sigma_F+\eta_1(M_H+2\sigma_H)]2(K+2)^{3/4}.\end{aligned} \tag{70}$$

By an analogous reasoning we used to obtain (64), we have

$$\begin{aligned}\mathbb{E}[\langle F(x), \bar{x}_K-x\rangle] &\le \tfrac{4\tau_{K-1}}{K\gamma_{K-1}}D_X^2+2\tfrac{D_X}{K}[2D_X(2L_F+(\eta_{K-1}+\eta_1)L_H)\\ &+2M_F+4\sigma_F+(\eta_{K-1}+\eta_1)(M_H+2\sigma_H)]\\ &+4\tfrac{D_X(K+2)^{3/4}}{K}[2D_X(L_F+\eta_1L_H)+M_F+2\sigma_F+\eta_1(M_H+2\sigma_H)]\\ &+\textstyle\sum_{k=1}^{K-1}\frac{5\theta^2\gamma_k\tau_k}{K}(M_F^2+2\sigma_F^2+\eta_{k-1}^2(M_H^2+2\sigma_H^2))\\ &+\textstyle\sum_{k=1}^{K-1}\frac{\tau_k\gamma_k}{2K}(\sigma_F^2+\eta_k\sigma_H^2)+\sum_{k=1}^{K-1}\frac{2C_HD_X\eta_k}{K}.\end{aligned} \tag{71}$$

Using the step-size policy in (54) and taking the maximum from both sides regarding set X of (71), we obtain the following upper bound

$$\begin{aligned}\mathbb{E}[\mathrm{Gap}(\bar{x}_K, F, X)] &\le D_X\left[\tfrac{32D_X(L_F+\eta_{K-1}L_H)}{K}+\tfrac{2\sqrt{M_F^2+2\sigma_F^2+\eta_{K-1}^2(M_H^2+2\sigma_H^2)}}{K^{1/2}}\right]\\ &+2\tfrac{D_X}{K}[2D_X(2L_F+(\eta_{K-1}+\eta_1)L_H)+2M_F+4\sigma_F+(\eta_{K-1}+\eta_1)(M_H+2\sigma_H)]\\ &+4\tfrac{D_X(K+2)^{3/4}}{K}[2D_X(L_F+\eta_1L_H)+M_F+2\sigma_F+\eta_1(M_H+2\sigma_H)]\\ &+\tfrac{D_X(20(M_F^2+2\sigma_F^2)+\sigma_F^2)}{3K^{1/2}\sqrt{M_F^2+2\sigma_F^2}}+\tfrac{D_X(10(M_H^2+2\sigma_H^2)+\sigma_H^2)}{K^{3/4}\sqrt{M_H^2+2\sigma_H^2}}+\tfrac{2C_HD_X}{K^{1/4}}.\end{aligned}$$

The lower bound for the optimality gap is obtained through the same line of argument we used in Theorem 2. □

Note that the convergence results for the strongly monotone case are the same as Theorem 3 and we do not mention it again for brevity.

A.4 Convergence analysis for the smooth VI

Lemma 7. *Let us assume that problem* (2) *is monotone* ($\mu_H = 0$). *Additionally, let* $M_F = 0$, *and suppose we have the following step-size policy with mini-batching of size* $B = K$

$$\tau_k = \tau = 1, \quad \theta_k = \theta = 1, \quad \eta_k = \eta = K^{-\frac{1}{2}}, \quad \gamma_k = \gamma = \tfrac{D_X}{8D_X(L_F+\eta L_H)+\sqrt{M_H^2+2(\sigma_H^2+\sigma_F^2)}}. \tag{72}$$

Then, Algorithm 1 gives the following optimality and feasibility gaps for $K \ge 1$

$$\begin{aligned}-B_H\mathbb{E}[dist(\bar{x}_K, X_F^*)] \le \mathbb{E}[Gap(\bar{x}_K, H, X_F^*)] \le & D_X\Big[\big(16D_X(\tfrac{L_F}{K^{1/2}}+\tfrac{L_H}{K})+\tfrac{2\sqrt{M_H^2+2(\sigma_H^2+\sigma_F^2)}}{K^{1/2}}\big)\\ &+\tfrac{5(M_H^2+2(\sigma_H^2+\sigma_F^2))+\sigma_H^2+\sigma_F^2}{8D_XK^{1/2}(L_F+\eta L_H)+K^{1/2}\sqrt{M_H^2+2(\sigma_H^2+\sigma_F^2)}}\\ &+\tfrac{5(M_H^2+2(\sigma_H^2+\sigma_F^2))}{8D_XK^{3/2}(L_F+\eta L_H)+K^{3/2}\sqrt{M_H^2+2(\sigma_H^2+\sigma_F^2)}}\Big].\end{aligned} \tag{73}$$

and

$$\begin{aligned}\mathbb{E}[Gap(\bar{x}_K, F, X)] \le & D_X\Big[\big(16D_X(\tfrac{L_F}{K}+\tfrac{L_H}{K^{3/2}})+\tfrac{\sqrt{M_H^2+2(\sigma_H^2+\sigma_F^2)}}{K}\big)+\tfrac{5(M_H^2+2(\sigma_H^2+\sigma_F^2))+\sigma_H^2+\sigma_F^2}{8D_XK(L_F+\eta L_H)+K\sqrt{M_H^2+2(\sigma_H^2+\sigma_F^2)}}\\ &+\tfrac{5(M_H^2+2(\sigma_H^2+\sigma_F^2))}{8D_XK^2(L_F+\eta L_H)+K^2\sqrt{M_H^2+2(\sigma_H^2+\sigma_F^2)}}+\tfrac{2C_H}{K^{1/2}}].\end{aligned} \tag{74}$$

Proof. The proof stems from the same reasoning we used in Theorem 1. □

Theorem 7. *Let us assume problem* (2) *is monotone* ($\mu_H = 0$) *and* $H_F = \sigma_F = 0$. *Suppose we have the following step-size policy*

$$\tau_k = \tau = 1, \quad \theta_k = \theta = 1, \quad \eta_k = \eta = K^{-\frac{1}{2}}, \quad \gamma_k = \gamma = \frac{D_X}{8D_X(L_F+\eta L_H)+\sqrt{M_H^2+2\sigma_H^2}}, \tag{75}$$

Then, Algorithm 1 gives the following optimality and feasibility gaps for $K \geq 1$

$$\begin{aligned} -B_H\mathbb{E}[dist(\bar{x}_K, X_F^*)] \leq \mathbb{E}[Gap(\bar{x}_K, H, X_F^*)] \leq & D_X\Bigg[\Big(16D_X\big(\tfrac{L_F}{K^{1/2}} + \tfrac{L_H}{K}\big) + \tfrac{2\sqrt{M_H^2+2\sigma_H^2}}{K^{1/2}}\Big) \\ & + \tfrac{5(M_H^2+2\sigma_H^2)+\sigma_H^2}{8D_XK^{1/2}(L_F+\eta L_H)+K^{1/2}\sqrt{M_H^2+2\sigma_H^2}} \\ & + \tfrac{5(M_H^2+2\sigma_H^2)}{8D_XK^{3/2}(L_F+\eta L_H)+K^{3/2}\sqrt{M_H^2+2\sigma_H^2}}\Bigg]. \end{aligned} \tag{76}$$

and

$$\begin{aligned} \mathbb{E}[Gap(\bar{x}_K, F, X)] \leq & D_X\Bigg[\Big(16D_X\big(\tfrac{L_F}{K} + \tfrac{L_H}{K^{3/2}}\big) + \tfrac{\sqrt{M_H^2+2\sigma_H^2}}{K}\Big) + \tfrac{5(M_H^2+2\sigma_H^2)+\sigma_H^2}{8D_XK(L_F+\eta L_H)+K\sqrt{M_H^2+2\sigma_H^2}} \\ & + \tfrac{5(M_H^2+2\sigma_H^2)}{8D_XK^2(L_F+\eta L_H)+K^2\sqrt{M_H^2+2\sigma_H^2}} + \tfrac{2C_H}{K^{1/2}}\Bigg]. \end{aligned} \tag{77}$$

Moreover, considering the following policy in the strongly monotone case ($\mu_H > 0$),

$$\tau_k = k+1, \quad \theta_k = \tfrac{k}{k+1}, \quad \eta_k = \eta = K^{-\frac{1}{2}}, \quad \gamma_k = \gamma = \frac{D_X}{8D_X(L_F+\eta L_H)+\sqrt{M_H^2+2\sigma_H^2}}. \tag{78}$$

Then, for $K \geq \frac{1}{2\gamma_k\eta_{k-1}\mu_H}$, *Algorithm 1 gives the following optimality and feasibility gaps*

$$\begin{aligned} -B_H\mathbb{E}[dist(\bar{x}_K, X_F^*)] \leq \mathbb{E}[Gap(\bar{x}_K, H, X_F^*)] \leq & D_X\Bigg[\Big(16D_X\big(\tfrac{L_F}{K^{3/2}} + \tfrac{L_H}{K^2}\big) + \tfrac{2\sqrt{M_H^2+2\sigma_H^2}}{K^{3/2}}\Big) \\ & + \tfrac{5(M_H^2+2\sigma_H^2)+\sigma_H^2}{8D_XK^{3/2}(L_F+\eta L_H)+K^{3/2}\sqrt{M_H^2+2\sigma_H^2}} \\ & + \tfrac{5(M_H^2+2\sigma_H^2}{8D_XK^{3/2}(L_F+\eta L_H)+K^{5/2}\sqrt{M_H^2+2\sigma_H^2}}\Bigg]. \end{aligned} \tag{79}$$

$$\begin{aligned} \mathbb{E}[Gap(\bar{x}_K, F, X)] \leq & D_X\Bigg[\Big(16D_X\big(\tfrac{L_F}{K^2} + \tfrac{L_H}{K^{5/2}}\big) + \tfrac{\sqrt{M_H^2+2\sigma_H^2}}{K^2}\Big) + \tfrac{5(M_H^2+2\sigma_H^2)+\sigma_H^2}{8D_XK^2(L_F+\eta L_H)+K^2\sqrt{M_H^2+2\sigma_H^2}} \\ & + \tfrac{5(M_H^2+2\sigma_H^2)}{8D_XK^3(L_F+\eta L_H)+K^3\sqrt{M_H^2+2\sigma_H^2}} + \tfrac{2C_H}{K^{3/2}}\Bigg]. \end{aligned} \tag{80}$$

Proof. The proof is similar to Theorem 1 for the monotone case and Lemma 5 for the strongly monotone case. □

Theorem 7 shows the improvement in convergence rate in terms of M_H and σ_H from $\mathcal{O}(K^{-1/4})$ to $\mathcal{O}(K^{-1/2})$. Similar to the nonsmooth stochastic case, we provide a lower bound for the monotone problem with a smooth deterministic operator F.

Theorem 8. *Suppose VI*(F, X) *is* α*-weakly sharp. Then we have*

$$\begin{aligned} \mathbb{E}[Gap(\bar{x}_K, H, X_F^*)] \geq & -\tfrac{B_H}{\alpha}\Bigg[2D_X^2\Big(8\big(\tfrac{L_F}{K^{1/2}} + \tfrac{L_H}{K}\big) + \tfrac{\sqrt{M_H^2+2\sigma_H^2}}{K^{1/2}}\Big) \\ & + \tfrac{5(M_H^2+2\sigma_H^2)+\sigma_H^2}{8K^{1/2}(L_F+\eta L_H)+K^{1/2}\sqrt{M_H^2+2\sigma_H^2}} \\ & + \tfrac{5(M_H^2+2\sigma_H^2)}{8K^{3/2}(L_F+\eta L_H)+K^{3/2}\sqrt{M_H^2+\sigma_H^2}} + \tfrac{2C_HD_X}{K^{1/2}}\Bigg]. \end{aligned} \tag{81}$$

Proof. The proof uses the same arguments used in Theorem 2 □

A.5 VI-constrained optimization problem.

Consider the following optimization problem

$$\min_{x \in X_F^*} \psi(x) := \mathbb{E}[\tfrac{1}{2}\|x+\zeta\|^2] \tag{82}$$

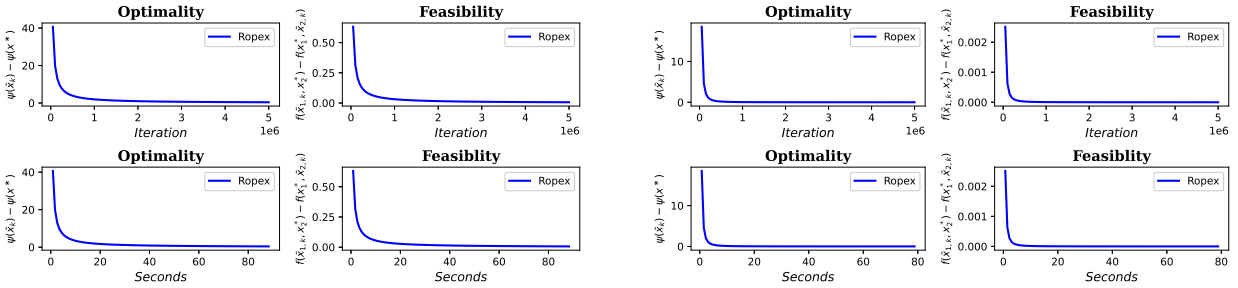
where $\zeta \in \mathbb{R}^2$ is a random vector whose elements $\zeta_i, i = 1, 2$ are independent and identically distributed following $\mathcal{N}(0, 1)$. Moreover, considering the set $X = [20, 50] \times [5, 15]$, we define the feasible region $X_F^* \subseteq X$ as the solution set of the following Nash equilibrium (NE) problem

$$\min_{x_1 \in [20,50]} \max_{x_2 \in [5,15]} f(x_1, x_2) := \mathbb{E}[25 - 2x_1x_2 + \xi x_1], \tag{83}$$

where the random variable ξ is following $\mathcal{N}(10, 1)$. One can notice that the explicit form of X_F^* can be stated as $X_F^* := \{(x_1, x_2) | x_1 \in [20, 50], x_2 = 5\}$. Furthermore, considering the explicit form of X_F^*, the minimum of (82) is obtained at $x^* = (20, 5)$. Note that the corresponding operators $\mathfrak{F}$ and $\mathfrak{H}$ have the following expressions

$$\mathfrak{F}(x_1, x_2; \xi) = \begin{bmatrix} -2x_2 + \xi \\ 2x_1 \end{bmatrix} \quad \mathfrak{H}(x_1, x_2; \zeta) = \begin{bmatrix} x_1 + \zeta_1 \\ x_2 + \zeta_2 \end{bmatrix}$$

We implement Algorithm 1 on problem (82) for monotone and strongly monotone cases. Note that it is a smooth problem with $L_F = \|[0, -2; 2, 0]\|$ and $L_H = 1$. We set the number of iterations $K = 5 \times 10^6$ for $R = 10$ i.i.d. replications. Figure 4a depicts the optimality and feasibility gaps of the solutions generated by R-OpEx regarding the number of iterations and time in the monotone setting. For the optimality and feasibility gaps, we use $\psi(\bar{x}_k) - \psi(x^*)$ and $f(\bar{x}_{1,k}, x_2^*) - f(x_1^*, \bar{x}_{2,k})$, respectively. Each gap is plotted against the number of iterations and the time in seconds. Figure 4b shows similar results for the strongly monotone case of implementing Algorithm 1 on 82. As expected, we achieve better runtime and accuracy with R-OpEx in the strongly monotone setting.



(a) Monotone case

(b) Strongly monotone case

Figure 4: Optimality and feasibility gaps for 10 replications of Algorithm 1 with 5×10^6 iterations.