Generalized Smooth Stochastic Variational Inequalities: Almost Sure Convergence and Convergence Rates

Anonymous authors
Paper under double-blind review

Abstract

This paper focuses on solving a stochastic variational inequality (SVI) problem under relaxed smoothness assumption for a class of structured non-monotone operators. The SVI problem has attracted significant interest in the machine learning community due to its immediate application to adversarial training and multi-agent reinforcement learning. In many such applications, the resulting operators do not satisfy the smoothness assumption. To address this issue, we focus on a weaker generalized smoothness assumption called α-symmetric. Under p-quasi sharpness and α-symmetric assumptions on the operator, we study clipped projection (gradient descent-ascent) and clipped Korpelevich (extragradient) methods. For these clipped methods, we provide the first almost-sure convergence results without making any assumptions on the boundedness of either the stochastic operator or the stochastic samples. We also provide the first in-expectation unbiased convergence rate results for these methods under a relaxed smoothness assumption for $\alpha \leq \frac{1}{2}$.

1 Introduction

This paper focuses on the stochastic variational inequality (SVI) problem, which consists of finding a point $u^* \in U$, such that

$$\langle F(u^*), u - u^* \rangle \geq 0 \qquad \text{for all } u \in U,$$

where the operator $F(\cdot)$ is specified as the expected value of a stochastic operator $\Phi(\cdot, \cdot) : U \times \Xi \to \mathbb{R}^m$, i.e.,

$$F(u) = \mathbb{E}[\Phi(u, \xi)] \qquad \text{for all } u \in U,$$

where $\xi \in \Xi$ is a random vector. Variational Inequality (VI) problems encompass many practical applications, such as optimization, min-max problems, and multi-agent games. In particular, they play a vital role in modeling equilibrium problems where it's important to capture an interaction between many agents. In machine learning literature, the increasing focus on VIs is due to their relevance to generative adversarial networks (GANs) Gemp & Mahadevan (2018); Gidel et al. (2019), actor-critic methods Pfau & Vinyals (2016), adversarial training, and multi-agent reinforcement learning Sokota et al. (2022); Kotsalis et al. (2022). In many such applications, the corresponding operator is defined as an expected value of stochastic or finite sum of operators, which motivates us to study SVIs. One of the pivotal works Nemirovski (2004); Juditsky et al. (2011) on SVIs proposed and studied the celebrated Mirror-Prox method under assumptions on monotonicity and Lipschitz continuity of an operator.[1] These assumptions become classical for the analysis of first-order methods for solving SVIs Beznosikov et al. (2022); Hsieh et al. (2019; 2020); Loizou et al. (2021).

In adversarial and multi-agent training, where the corresponding operator is a gradient of a highly non-linear neural network model, these classical assumptions might not be satisfied. It is well-known that one possible remedy for such non-convergent behavior is in clipping, normalization, or adaptive stepsizes, such as ADAM (Kingma & Ba, 2015). This effect might be explained by the experiment conducted in Zhang et al.

[1] The well-studied Mirror-Prox method Nemirovski (2004) has been proven to be optimal for solving VIs under strong monotonicity and Lipschitz continuity assumptions. In fact, this method is the stochastic version of the classical extragradient method.

(2020). In this work, authors observed that when training deep neural network, the norm of the hessian of the loss function correlates with a norm of a gradient along the optimization trajectory.

This observation motivated Zhang et al. (2020) to introduce a new and more realistic assumption on the linear growth of the hessian. This led to a great number of works in optimization investigating new assumptions on generalized smoothness and convergence behavior of classical gradient (Li et al., 2023), normalized (Chen et al., 2023), clipped (Koloskova et al., 2023), and adaptive methods (Wang et al., 2023; Zhang et al., 2024). Despite this progress in optimization, there are only a few works on generalized smooth min-max (Xian et al., 2024) and VI (Vankov et al., 2024) problems. This motivates us to delve into investigation of first-order methods for generalized smooth SVIs.

1.1 Related Work

Weaker Assumption on SVIs. More work has focused on stochastic methods for SVI under more relaxed assumptions to develop and analyze the methods applicable to broader problem classes. In particular, some studies have explored SVIs under pseudo-monotonicity Kannan & Shanbhag (2019), quasi-monotonicity Loizou et al. (2021), co-coercivity Beznosikov et al. (2023) and quasi-sharpness Vankov et al. (2023). Diakonikolas et al. (2021) showed that such conditions may not be satisfied even in two played Markov games and introduced the weakest known structured non-monotone assumption. Later, weak Minty SVIs were studied Pethick et al. (2023); Choudhury et al. (2023); Alacaoglu et al. (2024) under Lipschitz's assumption on the operator. In our work, we consider the generalized smooth assumption that goes beyond the existing settings.

Normalized and Clipped Methods for SVIs. Jelassi et al. (2022) studied the performance of normalized stochastic gradient descent-ascent and ADAM and suggested the crucial role of normalization for training GANs. It is worth noting that, with the right clipping parameters, clipped and normalized step sizes are equivalent up to a constant. Another line of works Gorbunov et al. (2022) focuses on smooth SVIs under heavy-tail noise. Using Lipschitz's continuity of operator and the right choice of clipping parameters, the authors showed a high probability convergence rate for the clipped stochastic Korpelevich method. Recent work Xian et al. (2024) considered generalized smooth stochastic nonconvex strongly-concave min-max problems and provided $\mathcal{O}(\frac{1}{\sqrt{K}})$ convergence rates for variants of stochastic gradient sescent ascent (SGDA) with normalized stepsizes. Due to the specific structure of the minmax problem and the fact that the gradient of the corresponding function is nonmonotone in one variable and strongly monotone in another, it is difficult to compare this work with ours. Moreover, in this work, the crucial part of the analysis is in the fact that the norm of a gradient can be upper bounded by a function residual. One can not use such bounds in SVIs due to the absence of function values. In our analysis, we develop a new technique to bound the operator norm in almost sure (*a.s.*) sense and in expectation.

Stochastic Analysis of Clipped Methods for Generalized Smooth Optimization. In Zhang et al. (2020), authors analyzed clipped gradient method under *a.s.* bounded error assumption. Koloskova et al. (2023) showed that the gradient method with standard clipping may not converge to a solution even with small stepsizes. The authors analyzed clipped gradient descent as a biased method and provided a convergence rate for non-convex functions. Later, Li et al. (2024) developed a new technique allowing to bound stochastic gradients by the function value residual along the optimization trajectory, which helps to find the convergence rate for the gradient with the right choice of stepsizes. In our work, we do not make an assumption on *a.s.* bounded noise and bounded stochastic gradients. Furthermore, we provide not only in-expectations but also *a.s.* convergence of the considered clipped methods.

Contributions. In light of the existing literature, we consider stochastic VIs with p-quasi sharp *generalized smooth* α-symmetric operators. We assume a bounded variance of the noise and do not use a restrictive assumption of bounded stochastic operators or bounded samples. Our key contributions are summarized below (see also Table 1):

- We provide the first known analysis of the clipped stochastic projection method (clipped SGDA) for solving stochastic generalized smooth VIs with p-quasi sharp and α-symmetric operators. The key feature of our analysis is the use of cleverly chosen clipped *stochastic stepsizes* γ_k. We use two

different samples of stochastic the operator, one for clipping stepsizes γ_k and another for the direction of the method update. This choice allows us to separate the clipping part from the stochastic error and analyze the method in an unbiased manner. To show *a.s.* convergence, we prove that the series of clipped *stochastic* stepsizes is not summable *a.s.*, i.e. $\mathbb{P}(\sum_{k=0}^{\infty} \mathbb{E}[\gamma_k \mid \mathcal{F}_{k-1}] = \infty) = 1$.

- We also provide convergence rate for stochastic clipped projection method for $\alpha \leq 1/2$. For $p = 2$ we achieve $\mathcal{O}(k^{-1})$ last iterate convergence. For $p > 2$, we show the best iterate convergence rate of $\mathcal{O}(k^{-2(1-q)/p})$, where $1 > q > 1/2$ is a parameter of the stepsize choice.
- We provide the first known analysis of the stochastic clipped Korpelevich method for solving stochastic generalized smooth VIs with p-quasi sharp and α-symmetric operators. By reusing clipping stepsizes γ_k for both iterates updates h_k and u_k, we separate stochastic stepsize from the stochastic error, similar to the projection method analysis. To show *a.s.* convergence, we prove that the series of clipped *stochastic* stepsizes is not summable *a.s.*, i.e. $\mathbb{P}(\sum_{k=0}^{\infty} \mathbb{E}[\gamma_k \mid \mathcal{F}_{k-1}] = \infty) = 1$.
- Moreover, we prove in-expectation convergence rates for the stochastic clipped Korpelevich methods for $\alpha \leq 1/2$. For $p = 2$, we show the last iterate sublinear convergence rate $\mathcal{O}(k^{-1})$. For $p > 2$, we show the best iterate convergence rate of $\mathcal{O}(k^{-2(1-q)/p})$, where $1 > q > 1/2$ is a parameter of the stepsize choice.
- Finally, we present numerical experiments where we compare the performance of the methods with proposed stochastic clipping for different stepsize parameter $q > 1/2$ and quasi-sharpness parameter p.

	Stochastic Projection	Stochastic Korpelevich
$p > 0, \alpha \in (0, 1]$	Asym (Thm 3.2)	Asym (Thm 4.2)
$p = 2, \alpha \in (0, \frac{1}{2}]$	$\mathcal{O}\left(\frac{D_0}{k^2} + \frac{\sigma^2(C_F+\sigma)^2}{\mu^2 k}\right)$	$\mathcal{O}\left(\frac{(C_F+\sigma)^2 K^2 D_0}{\mu^2 k^2} + \frac{(\sigma^2 + K_1\sigma^{2\alpha})(C_F+\sigma)^2}{\mu^2 k}\right)$
$p > 2, \alpha \in (0, \frac{1}{2}]$	$\mathcal{O}\left(\frac{(D_0+\sigma^2)^{2/p}(C_F+\sigma)^{2/p}}{\mu^{2/p} k^{2(1-q)/p}}\right)$	$\mathcal{O}\left(\frac{\sigma^{4/p}(D_0+\sigma^2+K_1\sigma^{2\alpha})^{2/p}(C_F+\sigma)^{2/p}}{\mu^{2/p} k^{2(1-q)/p}}\right)$

Table 1: Summary of convergence rate results showing the decrease of certain performance measures with the number k of iterations. We use "Asym" as an abbreviation for asymptotic almost sure convergence results. For p-quasi sharp operators, with $p = 2$, and for stochastic projection and Korpelevich methods, the performance measures are $D_k = \mathbb{E}[\mathrm{dist}^2(u_k, U^*)]$ and $D_k = \mathbb{E}[\mathrm{dist}^2(h_k, U^*)]$, respectively. For $p > 2$, the performance measure for both methods is $D_k^{\mathrm{best}} = \min_{t=0,\ldots,k} D_t$. The constant C_F denotes the upper bound on $\mathbb{E}[\|F(u_k)\|]$ and $\mathbb{E}[\|F(h_k)\|]$ for stochastic projection and Korpelevich methods, respectively.

The rest of the paper is organized as follows. In Section 2, we define the assumption on the operator class we consider and define the first-order methods we focus on. In Section 3 we show the almost sure convergence result of clipped stochastic projection method. In Section 4, we provide *a.s.* convergence results and in-expectation convergence rates for the clipped stochastic Korpelevich method. In Section 5, we conduct experiments on solving generalized smooth SVIs and compare the performance of the stochastic clipped projection and Korpelevich method for different problem and stepsize parameters. Section 6 concludes our work and presents some further research directions.

2 Preliminaries

In this section, we provide the necessary concepts and assumptions for the considered SVI problem. We start with a standard definition; the operator F is said to be Lipschitz continuous on a set U if there exists $L > 0$ such that

$$\|F(u) - F(v)\| \leq L\|u - v\| \quad \text{for all } u, v \in U. \tag{1}$$

So far, Lipschitz continuity of operator was the most common assumption to study SVIs Nemirovski (2004); Yousefian et al. (2014; 2017); Hsieh et al. (2019); Loizou et al. (2021); Alacaoglu et al. (2024). However, this

assumption does not hold in modern deep-learning applications. Based on the experiments on neural network training provided in Zhang et al. (2020), the norm of Jacobian of the operator correlates with the norm of the operator. Motivated by this observation, Zhang et al. (2020) proposed a new, more realistic, and weaker assumption named (L_0, L_1)-smooth: a differentiable operator F is (L_0, L_1)-smooth operator on a set U when

$$\|\nabla F(u)\| \leq (L_0 + L_1\|F(u))\|) \quad \text{for all } u \in U, \tag{2}$$

where $w_\theta = \theta u + (1-\theta)v$ and $\alpha \in (0, 1]$. When the operator $F(\cdot)$ is L-Lipshitz continuous, it satisfies (2) with $L_0 = L$ and $L_1 = 0$. Recent work Chen et al. (2023) generalized a class of (L_0, L_1)-smooth operators and introduced a new class termed α-symmetric. The class of α-symmetric operators includes the class of (L_0, L_1)-smooth operators and coincides with it when an operator is differentiable for $\alpha = 1$. Given that the class of α-symmetric operators includes the class of (L_0, L_1)-smooth and Lipschitz continuous operators we focus on this class in our work.

Assumption 2.1. Given a convex set $U \subseteq \mathbb{R}^m$, the operator $F(\cdot) : U \to \mathbb{R}^m$ is α-symmetric over U, i.e., for some $\alpha \in (0, 1]$ and $L_0, L_1 \geq 0$, we have for all $u, v \in U$,

$$\|F(u) - F(v)\| \leq \left(L_0 + L_1 \max_{\theta\in(0,1)} \|F(w_\theta)\|^\alpha\right) \|u - v\|, \tag{3}$$

where $w_\theta = \theta u + (1-\theta)v$.

An alternative characterization of α-symmetric operators has been proved in Chen et al. (2023), as given in the following proposition.

Proposition 2.2 (Chen et al. (2023), Proposition 1)**.** *Let $U \subseteq \mathbb{R}^m$ be a nonempty convex set and let $F(\cdot) : U \to \mathbb{R}^m$ be an operator. Then, the following statements hold:*

(a) $F(\cdot)$ is α-symmetric with $\alpha \in (0, 1)$ and constants $L_0, L_1 \geq 0$ if and only if the following relation holds for all $y, y' \in U$,

$$\|F(y) - F(y')\| \leq \|y - y'\|(K_0 + K_1\|F(y')\|^\alpha + K_2\|y - y'\|^{\alpha/(1-\alpha)}), \tag{4}$$

where $K_0 = L_0(2^{\alpha^2/(1-\alpha)} + 1)$, $K_1 = L_1 2^{\alpha^2/(1-\alpha)} 3^\alpha$, and $K_2 = L_1^{1/(1-\alpha)} 2^{\alpha^2/(1-\alpha)} 3^\alpha (1-\alpha)^{\alpha/(1-\alpha)}$.

(b) $F(\cdot)$ is α-symmetric with $\alpha = 1$ and constants $L_0, L_1 \geq 0$ if and only if the following relation holds for all $y, y' \in U$,

$$\|F(y) - F(y')\| \leq \|y - y'\|(L_0 + L_1\|F(y')\|) \exp(L_1\|y - y'\|). \tag{5}$$

Proposition 2.2 is useful for our analysis, since it describes an α-symmetric operator by using two points $y, y' \in U$, and bypasses the evaluation of $\max_{\theta\in(0,1)} \|F(w_\theta)\|^\alpha$. The solution set for the variational inequality problem defined by the set U and operator F, denoted U^*, is given by

$$U^* = \{u^* \in U \mid \langle F(u^*), u - u^*\rangle \geq 0 \ \text{ for all } u \in U\}.$$

Throughout this paper, we make the following assumption on the set U and the solution set.

Assumption 2.3. The set $U \subseteq \mathbb{R}^m$ is a nonempty closed convex set, and the solution set U^* is nonempty and closed.

In our analysis we assume that operator F is p-quasi sharp Vankov et al. (2023).

Assumption 2.4. The operator $F(\cdot) : U \to \mathbb{R}^m$ has a p-quasi sharpness property over U relative to the solution set U^*, i.e., for some $p > 0$, $\mu > 0$, and for all $u \in U$ and $u^* \in U^*$,

$$\langle F(u), u - u^*\rangle \geq \mu \operatorname{dist}^p(u, U^*). \tag{6}$$

In the optimization community, such assumptions are called error bounds and are widely studied Zhou & So (2017). This class of operators encompasses strongly monotone, p-monotone Facchinei & Pang (2003); Lin & Jordan (2025), strongly quasi-monotone Loizou et al. (2021) and strongly coherent Song et al. (2020) operators and aligns with the class of operators that satisfy the Saddle-Point Metric Subregularity Wei et al. (2021) for $p > 2$. The are many applications where operator satisfy (6), for example robust learning Wang et al. (2023); Zarifis et al. (2024), network congestion games Xiao & Shanbhag (2025). Assumption 2.4 is one of the most general assumptions with positive inner product between an operator value $F(u)$ and $u - u^*$ is positive, which is crucial in our analysis.

For solving the SVI problem, we consider stochastic variants of projection and Korpelevich (1976) methods, where stochastic approximations $\Phi(u_k, \xi_k)$ and $\Phi(h_k, \xi_k^1)$ are used, respectively, instead of the directions $F(u_k)$ and $F(h_k)$. The iterates of each of the stochastic methods are defined as follows:

Stochastic projection method:

$$u_{k+1} = P_U(u_k - \gamma_k \Phi(u_k, \xi_k)), \tag{7}$$

Stochastic Korpelevich method:

$$\begin{aligned} u_k &= P_U(h_k - \gamma_k \Phi(h_k, \xi_k^1)), \\ h_{k+1} &= P_U(h_k - \gamma_k \Phi(u_k, \xi_k^2)), \end{aligned} \tag{8}$$

where $\{\gamma_k > 0\}$ is a sequence of stochastic stepsizes, and $u_0, h_0 \in U$ are arbitrary deterministic initial points[2]. At each operator evaluation of these stochastic methods, a random sample ξ_k is drawn according to the distribution of the random variable ξ. We assume that the stochastic approximation error $\Phi(u, \xi) - F(u)$ is unbiased and has finite variance, leading to the following formal assumption.

Assumption 2.5. The random sample ξ is such that for all $u \in U$,

$$\mathbb{E}[\Phi(u, \xi) - F(u) \mid u] = 0, \qquad \mathbb{E}[\|\Phi(u, \xi) - F(u)\|^2 \mid u] \leq \sigma^2.$$

Our proof techniques in the following sections can be applied to analyze the $(a.s.)$ convergence and convergence rate of the stochastic Popov (Popov, 1980) method with an appropriate selection of stochastic clipping. However, due to space constraints, we leave this exploration for future research.

3 Projection method

Common approaches to developing convergent methods for generalized smooth optimization and VI problems are normalized or clipping stepsizes. We focus on the latter one and present stepsizes for the stochastic projection method for α-symmetric operators:

$$\gamma_k = \beta_k \min\left\{1, \frac{1}{\|\Phi(u_k, \xi_k^2)\|}\right\}, \tag{9}$$

where $\beta_k > 0$ for all $k \geq 0$ and ξ_k^2 is a random variable, such that ξ_k^2 and ξ_k are independent conditionally on u_k. In other words, at every iteration of the projection method, having u_k, two independent samples of the stochastic operator are drawn: (1) $\Phi(u_k, \xi_k)$ for the direction of update and (2) $\Phi(u_k, \xi_k^2)$ for clipping stepsize γ_k. We define the sigma-algebra $\mathcal{F}_k$ for the method:

$$\mathcal{F}_k = \{\xi_0, \xi_0^2, \ldots, \xi_k, \xi_k^2\} \qquad \text{for all } k \geq 0, \tag{10}$$

with $\mathcal{F}_{-1} = \emptyset$. In the sequel, we provide important results on the behavior of the iterates of the clipped stochastic projection method.

[2]The results easily extend to the case when the initial points are random as long as $\mathbb{E}[\|u_0\|^2]$ and $\mathbb{E}[\|h_0\|^2]$ are finite.

3.1 Almost sure convergence

The following lemma establishes a key relation for the iterate sequence $\{u_k\}$ generated by the stochastic projection method with stochastic clipping stepsizes. Its proof is in Appendix B.1

Lemma 3.1. *Let Assumptions 2.1, 2.3, 2.4, 2.5 hold, and $\{u_k\}$ be the iterate sequence generated by stochastic projection method (7) with stepsizes γ_k defined in (9). Let parameter β_k be such that $\sum_{k=0}^{\infty}\beta_k = \infty$ and $\sum_{k=0}^{\infty}\beta_k^2 < \infty$. Then, the following relation holds almost surely for all $k \geq 0$,*

$$\mathbb{E}[\|u_{k+1} - u^*\|^2 \mid \mathcal{F}_{k-1}] \leq \|u_k - u^*\|^2 - 2\mu\mathbb{E}[\gamma_k \mid \mathcal{F}_{k-1}]\mathrm{dist}^p(u_k, U^*) + 3\beta_k^2(2\sigma^2 + 1). \tag{11}$$

Furthermore, almost surely, we have

$$\sum_{k=0}^{\infty}\mathbb{E}[\gamma_k \mid \mathcal{F}_{k-1}]\,\mathrm{dist}^p(u_k, U^*) < \infty, \tag{12}$$

and, the sequence $\{\|u_k - u^*\|^2\}$ is bounded almost surely for all $u^* \in U^*$.

In the conventional analysis of methods for SVIs with Lipschitz continuous operators, the sequence $\{\gamma_k\}$ of stepsizes is deterministic and such that $\sum_k \gamma_k = \infty$. In our case, γ_k is a random variable, and to show *a.s.* convergence we have to show that the series $\sum_{k=0}^{\infty}\{\mathbb{E}[\gamma_k \mid \mathcal{F}_{k-1}]\}$ is not summable. We do so, providing a sequence of lower bounds for the series and by showing that random variable $\|F(u_k)\|$ is *a.s.* upper bounded for all $k \geq 0$. In the next theorem, we present the first results on *a.s.* convergence of the stochastic projection method.

Theorem 3.2. *Let Assumptions 2.1, 2.3, 2.4, and 2.5 hold, and $\{u_k\}$ be the iterate sequence generated by stochastic projection method (7) with stepsizes γ_k defined in (9). Let parameter β_k be such that $\sum_{k=0}^{\infty}\beta_k = \infty$ and $\sum_{k=0}^{\infty}\beta_k^2 < \infty$. Then, the iterates u_k converge almost surely to a point $\bar{u}$ such that $\bar{u} \in U^*$ almost surely.*

The full proof of Theorem 3.2 can be found in Appendix B.2. Notice that in an unconstrained setting ($U = \mathbb{R}^m$) according to Theorems 3.1 and 3.2 in Koloskova et al. (2023) for any clipping parameters $\beta > 0, c > 0$ there exist a stochastic gradient operator $\nabla_\xi f(\cdot)$ which satisfies Assumptions 2.1, 2.4 (with $p = 2$), 2.5 for which there exists a fixed point $\hat{v}$ of a standard clipping with one-sample which there exists a solution

$$\mathbb{E}_\xi[\beta_k \min\{1, \frac{c}{\|\nabla_\xi f(\hat{v})\|}\}\hat{v}] = 0 \quad \text{and} \quad \|\mathbb{E}_\xi[\nabla_\xi f(\hat{v})]\| \geq \sigma^2/12,$$

where $c > 0$ is a constant independent from a step sizes parameter β_k. This observation leads to an unavoidable bias in one-sample clipped SGD (Koloskova et al., 2023). In contrast, by using two sample in clipped projection method we overcome this problem and provide *a.s.* convergence to a solution.

3.2 Convergence rate

The difficulty of the convergence rate analysis is in the randomness of stepsizes γ_k. To show in-expectation convergence, we can take a total expectation on both sides of equation (11) of Lemma 3.1. However, since γ_k is a random variable, we have to provide a lower bound on $\mathbb{E}[\gamma_k \mathrm{dist}^p(u_k, U^*)]$. With this goal in mind, in the next lemma we show that the sequence $\{\mathbb{E}[\|F(u_k)\|]\}$ of expected norms is bounded. The proof of the lemma is in Appendix B.3.

Lemma 3.3. *Let Assumption 2.1 hold, with $\alpha \in (0, 1/2]$, Assumptions 2.3, 2.4, 2.5 hold, and $\{u_k\}$ be iterate sequence generated by stochastic projection method (7) with stepsizes γ_k defined in (9). Let parameter β_k be such that $\sum_{k=0}^{\infty}\beta_k = \infty$, and $\sum_{k=0}^{\infty}\beta_k^2 < \infty$. Then, the sequence $\{\mathbb{E}[\|F(u_k)\|]\}$ is bounded by some constant $C_F > 0$.*

To prove the preceding lemma, we show that the expected norms of the operator are bounded by some constant C_F on the trajectory of the method. Unfortunately, even though the sequence $\{\|F(u_k)\|\}$ is bounded almost surely, it does not imply that $\{\mathbb{E}[\|F(h_k)\|]\}$ is bounded. . To show this, we rely on the properties of

the method and the generalized smoothness of the operator in Proposition 2.2 to obtain that for all $k \geq 0$, and arbitrary solution v^*,

$$\begin{aligned}\|F(u_k)\| &\leq \|F(u_k) - F(v^*)\| + \|F(v^*)\| \\ &\leq \|u_k - v^*\|(K_0 + K_1\|F(v^*)\|^\alpha + K_2\|u_k - v^*\|^{\alpha/(1-\alpha)}) + \|F(v^*)\|. \end{aligned} \tag{13}$$

Notice that by taking an expectation in (13), the RHS is undefined for $\alpha > 1/2$. For $\alpha \in (0, 1/2]$, using (13) and boundedness of $\mathbb{E}[\|u_k - v^*\|]$ we achieve the desired bound on $\mathbb{E}[\|F(u_k)\|]$. Using this result, in the next theorem, we provide a convergence rate for the projection method with clipping.

Theorem 3.4. *Let Assumption 2.1, with $\alpha \in (0, 1/2]$, and Assumptions 2.3, 2.4, 2.5 hold. Let $\{u_k\}$ be the sequence generated by stochastic projection method (7) with stepsizes γ_k defined in (9). Let $D_k = \mathbb{E}[\text{dist}^2(u_k, U^*)]$ and C_F be an upperbound on $\mathbb{E}[\|F(u_k)\|]$. Then, we have:*

Case *$p = 2$. Let $\beta_k = \frac{2}{a(2+k)}$ with $a = \mu \min\left\{1, \frac{1}{2(C_F+\sigma)}\right\}$. Then, the following inequality holds*

$$D_{k+1} \leq \frac{8D_0}{k^2} + \frac{6(2\sigma^2+1)}{a^2 k} \qquad \text{for all } k \geq 1. \tag{14}$$

Case *$p \geq 2$. Let $\beta_k = \frac{b}{(k+1)^q}$, where $1/2 < q < 1$ and $b > 0$. Then, the following inequality holds*

$$\bar{D}_k \leq \frac{(1-q)^{2/p}\left(D_0 + 3b^2(2\sigma^2+1)/(2q-1)\right)^{2/p}}{(ab)^{2/p}\left((k+1)^{1-q} - 2^{1-q}\right)^{2/p}} \qquad \text{for all } k \geq 1, \tag{15}$$

where $\bar{D}_k = \mathbb{E}[\text{dist}^2(\bar{u}_k, U^*)]$, $\bar{u}_k = (\sum_{t=0}^k \beta_t)^{-1}\sum_{t=0}^k \beta_t u_t$, and $a = \mu \min\left\{1, \frac{1}{2(C_F+\sigma)}\right\}$.

To derive the convergence rate in terms of $\mathbb{E}[\text{dist}^2(u_k, U^*)]$ we need to relate the progress at each iteration, measured by $\mathbb{E}[\gamma_k \text{dist}^p(u_k, U^*)]$ to $\mathbb{E}[\text{dist}^2(u_k, U^*)]$. Using the boundedness of $\mathbb{E}[\|F(u_k)\|]$, we first bound $\mathbb{E}[\gamma_k \text{dist}^p(u_k, U^*)]$ in terms of $\mathbb{E}[\text{dist}^p(u_k, U^*)]$. Finally, we estimate $\mathbb{E}[\text{dist}^p(h_k, U^*)]$ through $\mathbb{E}[\text{dist}^2(u_k, U^*)]$ by applying Jensens inequality, which holds for $p \geq 2$. The proof of Theorem 3.4 is in Appendix B.4.

For the simplicity of convergence rate comparison, assume $2(C_F + \sigma) \geq 1$. Then, from Theorem 3.4 we obtain $\mathcal{O}(\frac{D_0}{k^2} + \frac{\sigma^2(C_F+\sigma)^2}{\mu^2 k})$ last iterate convergence rate for $p = 2$, and $\mathcal{O}(\frac{(D_0+\sigma^2)^{2/p}(C_F+\sigma)^{2/p}}{\mu^{2/p}k^{2(1-q)/p}})$ average (or best) iterate convergence rate for $p > 2$ with $q \in (1/2, 1)$. It is worth mentioning that obtained rates are unbiased, unlike the analysis in Koloskova et al. (2023). However, it comes with the price of two oracle calls per iteration. For $p = 2$, the rate from Theorem 3.4 matches the rate $\mathcal{O}(\frac{1}{k})$ obtained in Theorem 4.3 (Loizou et al., 2021) for SGDA under stronger assumption on quasi-strong monotonicity and Lipschitz continuity of the operator. Interestingly, for $p = 2$, the parameter μ appears in the rate as $\frac{1}{\mu^2}$ in both (Loizou et al., 2021) and in our Theorem 3.4, which is known to be the optimal dependence on μ (Beznosikov et al., 2022). The rate for $p > 2$ is new in the stochastic case and generalizes the convergence results in deterministic setting (Vankov et al., 2024). From a theoretical perspective, the projection clipped method and its different-sample variant differ in several key aspects. In terms of asymptotic convergence, the standard projection clipped method suffers from an unavoidable bias, whereas our projection clipped method with different samples enjoys almost sure convergence. For convergence rates, our theorem shows a sublinear rate for the different-sample variant, while there are no known results for the standard projection clipped method in the setting of stochastic VIs. This makes the different-sample projection clipped method more favorable from a theoretical point of view. We also compare the performance of the two methods in the numerical experiments section.

4 Korpelevich method

The stepsizes for the stochastic Korpelevich method for α-symmetric operators are as given below

$$\gamma_k = \beta_k \min\left\{1, \frac{1}{\|\Phi(h_k, \xi_k^1)\|}\right\}, \tag{16}$$

where $\beta_k > 0$ for all $k \geq 0$ and ξ_k^1 is a random variable associated with the stochastic approximation $\Phi(h_k, \xi_k^1)$ of $F(h_k)$. We define the sigma-algebra $\mathcal{F}_k$ for the method, as follows:

$$\mathcal{F}_k = \{\xi_0^1, \xi_0^2, \ldots, \xi_k^1, \xi_k^2\} \qquad \text{for all } k \geq 0, \tag{17}$$

with $\mathcal{F}_{-1} = \emptyset$. Notice that to obtain h_{k+1} from a point u_k, the stepsize γ_k clips $\Phi(h_k, \xi_k^1)$, not the stochastic approximation $\Phi(u_k, \xi_k^2)$ of the operator at point u_k, i.e., we have

This is correct.

$$h_{k+1} = P_U\left(h_k - \beta_k \min\left\{1, \frac{1}{\|\Phi(h_k, \xi_k^1)\|}\right\}\Phi(u_k, \xi_k^2)\right). \tag{18}$$

Thus, sample ξ_k^2 is drawn after ξ_k^1, and $\Phi(h_k, \xi_k^1)$ is measurable with respect to $\mathcal{F}_{k-1} \cup \xi_k^1$. This property of the stochastic Korpelevich method with clipping stepsizes γ_k is crucial for further convergence analysis of the method. In the sequel, we provide important results on the behavior of the iterates of the clipped stochastic Korpelevich method.

4.1 Almost sure convergence

In the forthcoming lemma, we provide some basic relations that hold almost surely for the iterates of the stochastic Korpelevich method with clipped stochastic stepsize.

Lemma 4.1. *Let Assumptions 2.1, 2.3, 2.4, and 2.5 hold. Also, let $\{h_k\}$ and $\{u_k\}$ be iterates generated by stochastic Korpelevich method (8) with stepsizes γ_k defined in (16) and with parameter β_k such that $\sum_{k=0}^{\infty} \beta_k = \infty$ and $\sum_{k=0}^{\infty} \beta_k^2 < \infty$. Then, the following relation holds almost surely*

This is correct. Let $v_k = \|h_k - u^*\|^2 + \frac{1}{2}\|h_{k-1} - u_{k-1}\|^2 + 2\gamma_k \mu \mathrm{dist}^p(u_k, U^*)$, then

$$\begin{aligned}\mathbb{E}[v_{k+1} \mid \mathcal{F}_{k-1}] \leq v_k - \frac{1}{2}\|h_{k-1} - u_{k-1}\|^2 - 2\mu\gamma_{k-1}\mathrm{dist}^p(u_{k-1}, U^*) \\ + 6\beta_k^2(\sigma^2 + C_e(\beta_k, \alpha)\sigma^{2\alpha}) \qquad \textit{for all } k \geq 0,\end{aligned} \tag{19}$$

where $C_e(\beta_k, \alpha) = K_1$, when $\alpha \in (0, 1)$, and $C_e(\beta_k, \alpha) = \exp(L_1\beta_k)$, when $\alpha = 1$. Moreover, the following relations hold almost surely,

$$\sum_{k=0}^{\infty} \gamma_k \mathrm{dist}^p(u_k, U^*) < \infty, \quad \sum_{k=0}^{\infty} \|h_k - u_k\|^2 < \infty. \tag{20}$$

Furthermore, the sequence $\{\|h_k - u^*\|\}$ is bounded almost surely for all $u^* \in U^*$.

The proof of Lemma 4.1 is in Appendix C.1.

In a standard analysis of the Korpelevich method for SVI with Lipschitz operators Kannan & Shanbhag (2019); Vankov et al. (2023), *a.s.* convergence results were achieved for a deterministic sequence $\{\gamma_k\}$. In our case, similarly to projection method analysis, $\{\gamma_k\}$ is a sequence of random variables, which makes the analysis of the methods more difficult and involved. By the choice of stepsizes γ_k as given in (16), the following relation holds true

$$\mathbb{E}[\gamma_k \langle \Phi(u_k, \xi_k^2) - F(u_k), u_k - u^* \rangle \mid \mathcal{F}_{k-1}] = 0 \qquad \text{for all } k \geq 0. \tag{21}$$

To prove (21) for the stochastic Korpelevich method, we note that the clipping stepsize is using $\|\Phi(h_k, \xi_k^1)\|$, which decouples from $\Phi(u_k, \xi_k^2)$ by properly using conditional expectation. Specifically, we first take the expectation conditioned on $\mathcal{F}_{k-1} \cup \xi_k^1$. Since γ_k is measurable with respect to $\mathcal{F}_{k-1} \cup \xi_k^1$, we use Assumption 2.5 and the law of the total expectation. Interestingly, we do not have to take another sample for the clipping in the stochastic Korpelevich method, as we have done in the stochastic projection method. Thus, to perform one iteration, we use two oracle calls in both methods.

Using Lemma 4.1, we next present the almost sure convergence of the clipped Korpelevich method.

Theorem 4.2. *Let Assumptions 2.1, 2.3, 2.4, and 2.5 hold and $\{h_k\}$, $\{u_k\}$ be iterates generated by stochastic Korpelevich method (8) with stepsizes γ_k defined in (16). Let parameter β_k be such that $\sum_{k=0}^{\infty}\beta_k = \infty$, and $\sum_{k=0}^{\infty}\beta_k^2 < \infty$. Then, the iterates h_k and u_k converge almost surely to a point $\bar{u}$ such that $\bar{u} \in U^*$ almost surely.*

To prove *a.s.* convergence, we firstly show that $\sum_{k=0}^{\infty}\gamma_k = \infty$ *a.s.*, by providing a sequence of lower bounds on γ_k, using *a.s.* boundedness of $\|h_k - u^*\|$ of Lemma 4.1, and proving that $\|F(h_k)\|$ is *a.s.* bounded. The full proof can be found in Appendix C.2.

4.2 Convergence rate

We start our analysis by taking the total expectation on both sides of equation (19) from Lemma 4.1. For further analysis, similar to the clipped stochastic projection methods, the challenge lies in the randomness of the stepsizes γ_k. To handle this, firstly, we establish a lower bound for $\mathbb{E}[\gamma_k \mathrm{dist}^p(u_k, U^*)]$ by showing that the sequence $\{\mathbb{E}[|F(u_k)|]\}$ of expected norms remains bounded, as shown in the next lemma. The proof of the lemma can be found in Appendix B.3.

Lemma 4.3. *Let Assumption 2.1, with $\alpha \in (0, 1/2]$, and Assumptions 2.3, 2.4, 2.5 hold. Let $\{u_k\}$, $\{h_k\}$ be iterates generated by stochastic Korpelevich method (8) with stepsizes γ_k defined in (16) and the parameter β_k such that $\sum_{k=0}^{\infty}\beta_k = \infty$ and $\sum_{k=0}^{\infty}\beta_k^2 < \infty$. Then, $\mathbb{E}[\|F(h_k)\|]$ is bounded by some constant $C_F > 0$ for all $k \geq 0$.*

Similarly to the analysis presented in Section 3, we bound $F(h_k)$ by using a triangle inequality and the property of α-symmetric operators, and by taking the total expectation, we obtain

$$\mathbb{E}[\|F(u_k)\|] \leq K_0\mathbb{E}[\|u_k - v^*\|] + K_2\mathbb{E}[\|u_k - v^*\|^{\alpha/(1-\alpha)})] + \|F(v^*)\| + K_1\|F(v^*)\|^{\alpha}. \tag{22}$$

We can show that the preceding bound has a finite expectation only for $0 < \alpha \leq 1/2$, which motivates the restriction on α in Lemma 4.3. Equipped with the boundedness of the sequence $\{\mathbb{E}[\|F(h_k)\|\}$ of expected norms of the operator along the iterates $\{h_k\}$, we present the next convergence rate theorem.

Theorem 4.4. *Let Assumption 2.1, with $\alpha \in (0, 1/2]$, and Assumptions 2.3, 2.4, 2.5 hold. Let $\{u_k\}$, $\{h_k\}$ be iterates generated by stochastic Korpelevich method (8) with stepsizes γ_k defined in (16). Let $D_k = \mathbb{E}[\mathrm{dist}^2(h_k, U^*)]$ and C_F be an upperbound on $\mathbb{E}[\|F(h_k)\|]$ then the following results holds:*

Case *$p = 2$. Let $\beta_k = \dfrac{2}{a(\frac{2d}{a} + k)}$, with $a = \mu \min\left\{1, \frac{1}{2(C_F+\sigma)}\right\}$, $d = \max\{4\mu, 2\sqrt{3}(K_0 + K_1 + K_2)\}$ where K_0, K_1, and K_2 are from Proposition 2.2(a). Then, the following relation holds*

$$D_{k+1} \leq \frac{8d^2 D_0}{a^2 k^2} + \frac{12(\sigma^2 + K_1\sigma^{2\alpha})}{a^2 k} \qquad \text{for all } k \geq 1. \tag{23}$$

Case *$p \geq 2$. Let $\beta_k = \frac{b}{(k+1)^q}$, where $1/2 < q < 1$ and $0 < b \leq \min\left\{\frac{1}{4\mu}, \frac{1}{2\sqrt{3}(K_0+K_1+K_2)}\right\}$. Then, the following inequality holds for all $k \geq 1$,*

$$\bar{D}_k \leq \frac{2^{2(p-2)/p}(1-q)^{2/p}\left(D_0 + 6b^2(\sigma^2 + K_1\sigma^{2\alpha})(2\sigma^2+1)/(2q-1)\right)^{2/p}}{(ab)^{2/p}\left((k+1)^{1-q} - 2^{1-q}\right)^{2/p}}, \tag{24}$$

where $\bar{D}_k = \mathbb{E}[\mathrm{dist}^2(\bar{u}_k, U^*)]$, $\bar{u}_k = (\sum_{t=0}^{k}\beta_t)^{-1}\sum_{t=0}^{k}\beta_t u_t$, and $a = \mu \min\left\{1, \frac{1}{2(C_F+\sigma)}\right\}$.

Similarly to the proof of Theorem 3.4, to derive the convergence rate in terms of $\mathbb{E}[\mathrm{dist}^2(h_k, U^*)]$ we need to relate the progress at each iteration, measured by $\mathbb{E}[\gamma_k \mathrm{dist}^p(h_k, U^*)]$ to $\mathbb{E}[\mathrm{dist}^2(h_k, U^*)]$. Using the boundedness of $\mathbb{E}[\|F(h_k)\|]$ and applying Jensens inequality, which holds for $p \geq 2$ we obtain the final rates. The proof of Theorem 4.4 is provided in Appendix C.4.

For the simplicity of convergence rate comparison, assume $2(C_F + \sigma) \geq 1$ and $K_0 + K_1 + K_2 \geq \frac{2\mu}{\sqrt{3}}$. Then by denoting $K = K_0 + K_2 + K_3$, from Theorem 4.4, we obtain $\mathcal{O}(\frac{(C_F+\sigma)^2 K^2 D_0}{\mu^2 k^2} + \frac{(\sigma^2+K_1\sigma^{2\alpha})(C_F+\sigma)^2}{\mu^2 k})$ last

iterate convergence for $p = 2$, and $\mathcal{O}\left(\frac{\sigma^{4/p}(D_0+\sigma^2+K_1\sigma^{2\alpha})^{2/p}(C_F+\sigma)^{2/p}}{\mu^{2/p}k^{2(1-q)/p}}\right)$ average (or best) iterate convergence rate for $p > 2$ with $q \in (1/2, 1)$. In both cases $p = 2$ and $p > 2$ the convergence rate of clipped stochastic projection method in Theorem 3.4 and the rate of clipped stochastic Korpelevich method in Theorem 4.4 have the same dependency in k. For $p = 2$ the rate from Theorem 4.4 matches the rate $\mathcal{O}(\frac{1}{k})$ obtained in Proposition 5 (Kannan & Shanbhag, 2019) for stochastic Korpelevich method under stronger assumption on strong pseudo monotonicity and Lipschitz continuity of the operator. Interestingly, for $p = 2$, the parameter μ appears in the rate as $\frac{1}{\mu^2}$ in both Kannan & Shanbhag (2019) and in our Theorem 4.4, which is known to be the optimal dependence on μ (Beznosikov et al., 2022). For $p > 2$, the obtained rate is new in stochastic case and generalize the results in deterministic setting for Lipschitz continuous operators (Wei et al., 2021) and α-symmetric operators (Vankov et al., 2024).

5 Numerical Experiments

We study the performance of the clipped stochastic projection and Korpelevich methods, for different values of parameters $\alpha > 0$ and $p > 0$. Despite the absence of analysis, we also implement the clipped stochastic Popov method with $\gamma_k = \beta_k \min\{1, \frac{1}{\|F(h_k)\|}, \frac{1}{(\|u_k - h_{k-1}\|+1)^{\alpha/(1-\alpha)}}\}$:

$$u_{k+1} = P_U(u_k - \gamma_k \Phi(h_k, \xi_k)), \quad h_{k+1} = P_U(u_{k+1} - \gamma_{k+1}\Phi(h_k, \xi_k)),$$

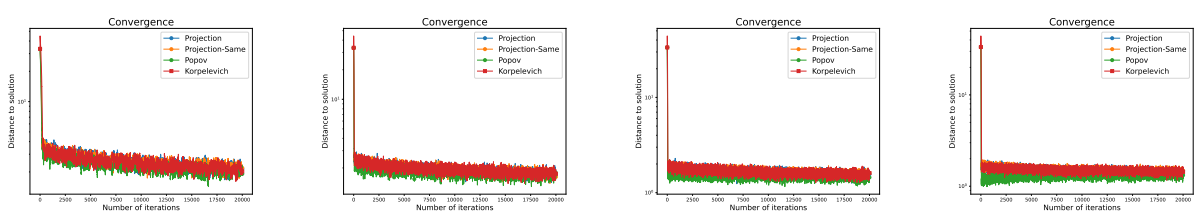
where $u_0, h_0 \in U$ are arbitrary deterministic initial. We consider an unconstrained minmax game:

$$\min_{u_1} \max_{u_2} \frac{1}{p}\|u_1\|^p + \langle u_1, u_2\rangle - \frac{1}{p}\|u_2\|^p,$$

$p > 1$, and $u_1 \in \mathbb{R}^d, u_2 \in \mathbb{R}^d$, and then corresponding operator $F : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ is defined by

$$F(u) = \begin{bmatrix} \|u_1\|^{p-2}u_1 + u_2 \\ \|u_2\|^{p-2}u_2 - u_1 \end{bmatrix}. \tag{25}$$

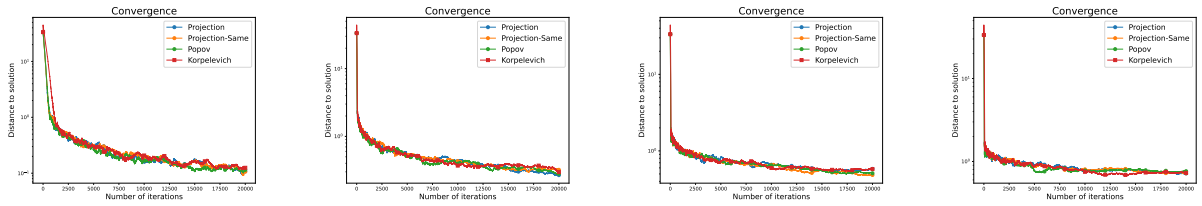
We assume that we have an access only to a noise evaluation of the corresponding operator and aim to solve unconstrained SVI($\mathbb{R}^{2d}, F$) with the following stochastic operator $\Phi(u, \xi) = F(u) + \xi$, where ξ is a random vector with independent zero-mean Gaussian entries and with variance $\sigma^2 = 1$. Then, $F(u) = \mathbb{E}[\Phi(u, \xi)]$ is an α-symmetric and p-quasi sharp operator due to Vankov et al. (2024). We set these parameters to be $\{(\alpha \approx 0.33, p = 2.5), (\alpha \approx 0.5, p = 3.0), (\alpha \approx 0.8, p = 6.0)\}$. We also compare our results with the projection method that uses the same sample clipping, meaning stepsizes γ_k clip $\|\Phi(u_k, \xi_k)\|$ instead of a different sample $\|\Phi(u_k, \xi_k^2)\|$.



(a) $(\alpha \approx 0.33, p = 1.5)$ (b) $(\alpha \approx 0.33, p = 2.5)$ (c) $(\alpha \approx 0.8, p = 4.0)$ (d) $(\alpha \approx 0.8, p = 6.0)$

Figure 1: Comparison of the clipped stochastic projection, same-sample projection, Korpelevich, and Popov methods with $\beta_k = 100/(100 + k^{1/2+\epsilon})$.

In Figure 1, we plot an average distance to solution from the current iterate over twenty runs to the solution set as a function of the number of iterations. In particular, the stepsizes for clipped stochastic projection and Korpelevich methods are chosen according to Theorems 3.4 and 4.4, respectively, with $\beta_k = \frac{100}{100+k^q}$ for $q = 1/2 + \epsilon$ with $\epsilon > 0$. Note that, according to Theorems 3.4 and 4.4, the parameter q should be greater than 1/2; meanwhile, the rates in these theorems are better for smaller choices of q. We also set $\beta_k = \frac{100}{100+k^q}$



(a) $(\alpha \approx 0.33, p = 1.5)$ (b) $(\alpha \approx 0.33, p = 2.5)$ (c) $(\alpha \approx 0.8, p = 4.0)$ (d) $(\alpha \approx 0.8, p = 6.0)$

Figure 2: Comparison of the clipped stochastic projection, same-sample projection, Korpelevich, and Popov methods with $\beta_k = 100/(100 + k^{1-\epsilon})$.

for clipped stochastic Popov method and the clipped stochastic projection method using the same sample $\Phi(u_k, \xi_k)$ for clipping.

Based on this experiment, we made three important observations. Firstly, the clipped stochastic projection method and same-sample clipped stochastic projection method show similar results despite the fact that the same-sample stochastic projection method has a biased error. . Secondly, for $\alpha \leq 1/2$ as predicted in theory (Theorems 3.4, 4.4), both projection and Korpelevich methods show in-expectation convergence, while for $\alpha > 1/2$ Korpelevich method has less stable performance. Finally, despite the fact that for stochastic Lipschitz SVI, Korpelevich method outperforms the projection method, we don't see this in generalized smooth SVIs.

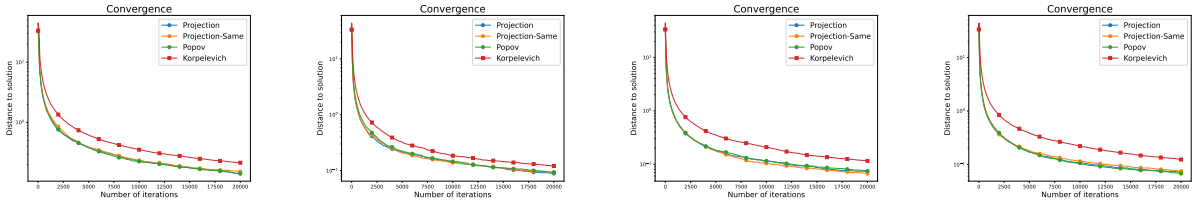
Based on this experiment, we made three important observations. Firstly, the clipped stochastic projection method and the same-sample clipped stochastic projection method show similar results, despite the fact that the latter has a biased error. We speculate that this is because the biased error is relatively small, and in practice dominated by the term $\gamma_k \text{dist}^p(u_k, U^*)$. Secondly, although in the stochastic Lipschitz SVI setting the Korpelevich method outperforms the projection method, we do not observe this advantage in the generalized smooth SVI setting. This aligns with our theoretical results, since both methods have the same complexity and require two oracle calls per iteration. Moreover, even in the standard Lipschitz continuous strongly monotone case, both methods achieve the same order $O(1/k)$ rate in the leading stochastic term, with the Korpelevich method enjoying a smaller non-leading term because it can use a larger stepsize of $\frac{1}{L}$ Beznosikov et al. (2022) compared to $\frac{\mu}{L^2}$ for the projection method Loizou et al. (2021). However, in the generalized smooth case, where stepsizes are clipped, the Korpelevich method no longer benefits from larger stepsizes.

Next, we investigate the performance of the methods for larger values of q. In Figure 2, we set $q = 1 - \epsilon$, and corresponding $\beta_k = \frac{100}{100+k^{1-\epsilon}}$ and run all four methods for the same problem parameter setting. We observe that for all considered α, despite the theory, a larger choice of q improved the performance of all methods in the σ-neighborhood.

For the same setting, in Figures 3 and 4, now we plot the distance to the solution from the average iterate $\bar{u_k} = (\sum_{i=0}^{k} \beta_i)^{-1} \sum_{i=0}^{k} \beta_i u_i$. In terms of average iterates, we observed that smaller values of q are preferable. Interestingly, the clipped projection methods outperform the clipped Korpelevich method, even though both enjoy the same convergence guarantee of order $\mathcal{O}(1/k^{2(1-q)/p})$.

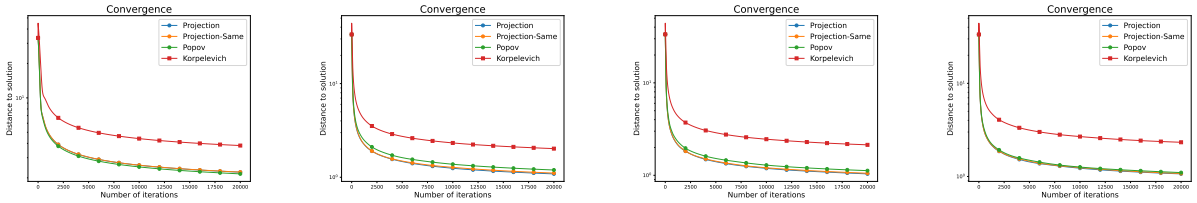
6 Conclusion

This paper studied the SVI problem under generalized smooth and structured non-monotone assumptions. Specifically, we consider α-symmetric and p-quasi-sharp operators, a class of generalized smooth and structured non-monotone operators for SVIs. For this wide class of operators, we proved the first-known almost sure convergence of clipped stochastic projection and Korpelevich methods for all parameters p. We also provided $\mathcal{O}(1/k)$ convergence rate for both considered methods when the operator is p-quasi sharp with $p = 2$. For $p > 2$ we provided $\mathcal{O}(k^{-2(1-q)/p})$ average (or best) iterate convergence rate for both methods, where q is a



(a) $(\alpha \approx 0.33, p = 1.5)$ (b) $(\alpha \approx 0.33, p = 2.5)$ (c) $(\alpha \approx 0.8, p = 4.0)$ (d) $(\alpha \approx 0.8, p = 6.0)$

Figure 3: Comparison of the clipped stochastic projection, same-sample projection, Korpelevich, and Popov methods with $\beta_k = 100/(100 + k^{1/2+\epsilon})$.



(a) $(\alpha \approx 0.33, p = 1.5)$ (b) $(\alpha \approx 0.33, p = 2.5)$ (c) $(\alpha \approx 0.8, p = 4.0)$ (d) $(\alpha \approx 0.8, p = 6.0)$

Figure 4: Comparison of the clipped stochastic projection, same-sample projection, Korpelevich, and Popov methods with $\beta_k = 100/(100 + k^{1-\epsilon})$.

stepsizes parameter $1/2 < q < 1$. Despite the generality of our results, there are still open questions remain. In particular, it would be interesting to know if it is possible to show in-expectation convergence rates for α-smooth SVI $\alpha > 1/2$. Another attractive direction of further research in generalized smooth SVIs is in relaxation of p-quasi sharpness assumption to Minty ($\mu = 0$) or weak Minty conditions ($\mu < 0$). We also believe that our technique for proving almost sure convergence and in-expectation rates can be used for the analysis of other methods whose stepsizes are random variables, for example, clipped stochastic Popov method or first-order methods with adaptive stepsizes.

References

Ahmet Alacaoglu, Donghwan Kim, and Stephen Wright. Revisiting inexact fixed-point iterations for min-max problems: Stochasticity and structured nonconvexity. In *Forty-first International Conference on Machine Learning*, 2024.

Aleksandr Beznosikov, Boris Polyak, Eduard Gorbunov, Dmitry Kovalev, and Alexander Gasnikov. Smooth monotone stochastic variational inequalities and saddle point problems–survey. *arXiv preprint arXiv:2208.13592*, 2022.

Aleksandr Beznosikov, Eduard Gorbunov, Hugo Berard, and Nicolas Loizou. Stochastic gradient descent-ascent: Unified theory and new efficient methods. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 172–235. PMLR, 2023.

Ziyi Chen, Yi Zhou, Yingbin Liang, and Zhaosong Lu. Generalized-smooth nonconvex optimization is as efficient as smooth nonconvex optimization. In *International Conference on Machine Learning*, 2023.

Sayantan Choudhury, Eduard Gorbunov, and Nicolas Loizou. Single-call stochastic extragradient methods for structured non-monotone variational inequalities: Improved analysis under weaker conditions. *preprint arXiv:2302.14043*, 2023.

Jelena Diakonikolas, Constantinos Daskalakis, and Michael I Jordan. Efficient methods for structured nonconvex-nonconcave min-max optimization. In *International Conference on Artificial Intelligence and Statistics*, pp. 2746–2754. PMLR, 2021.

Rick Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 2019.

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

Ian Gemp and Sridhar Mahadevan. Global convergence to the equilibrium of gans using variational inequalities. *arXiv preprint arXiv:1808.01531*, 2018.

Gauthier Gidel, Hugo Berard, Gaëtan Vignoud, Pascal Vincent, and Simon Lacoste-Julien. A variational inequality perspective on generative adversarial networks. In *International Conference on Learning Representations*, 2019.

Eduard Gorbunov, Marina Danilova, David Dobre, Pavel Dvurechenskii, Alexander Gasnikov, and Gauthier Gidel. Clipped stochastic methods for variational inequalities with heavy-tailed noise. *Advances in Neural Information Processing Systems*, 35:31319–31332, 2022.

Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. On the convergence of single-call stochastic extra-gradient methods. *Advances in Neural Information Processing Systems*, 32, 2019.

Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. Explore aggressively, update conservatively: Stochastic extragradient methods with variable stepsize scaling. *Advances in Neural Information Processing Systems*, 33:16223–16234, 2020.

Samy Jelassi, David Dobre, Arthur Mensch, Yuanzhi Li, and Gauthier Gidel. Dissecting adaptive methods in gans. *arXiv preprint arXiv:2210.04319*, 2022.

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

Aswin Kannan and Uday V Shanbhag. Optimal stochastic extragradient schemes for pseudomonotone stochastic variational inequality problems and their variants. *Computational Optimization and Applications*, 74(3):779–820, 2019.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Anastasia Koloskova, Hadrien Hendrikx, and Sebastian U Stich. Revisiting gradient clipping: Stochastic bias and tight convergence guarantees. In *International Conference on Machine Learning*, pp. 17343–17363. PMLR, 2023.

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

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

Haochuan Li, Ali Jadbabaie, and Alexander Rakhlin. Convergence of adam under relaxed assumptions. *arXiv preprint arXiv:2304.13972*, 2023.

Haochuan Li, Jian Qian, Yi Tian, Alexander Rakhlin, and Ali Jadbabaie. Convex and non-convex optimization under generalized smoothness. *Advances in Neural Information Processing Systems*, 36, 2024.

Tianyi Lin and Michael I Jordan. Perseus: A simple and optimal high-order method for variational inequalities. *Mathematical Programming*, 209(1):609–650, 2025.

Nicolas Loizou, Hugo Berard, Gauthier Gidel, Ioannis Mitliagkas, and Simon Lacoste-Julien. Stochastic gradient descent-ascent and consensus optimization for smooth games: Convergence analysis under expected co-coercivity. *Advances in Neural Information Processing Systems*, 34:19095–19108, 2021.

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

Thomas Pethick, Olivier Fercoq, Puya Latafat, Panagiotis Patrinos, and Volkan Cevher. Solving stochastic weak minty variational inequalities without increasing batch size. In *The Eleventh International Conference on Learning Representations, ICLR*, 2023.

David Pfau and Oriol Vinyals. Connecting generative adversarial networks and actor-critic methods. *arXiv preprint arXiv:1610.01945*, 2016.

Boris T Polyak. Introduction to optimization. optimization software. *Inc., Publications Division, New York*, 1:32, 1987.

Leonid Denisovich Popov. A modification of the arrow-hurwicz method for search of saddle points. *Mathematical notes of the Academy of Sciences of the USSR*, 28(5):845–848, 1980.

Samuel Sokota, Ryan D'Orazio, J Zico Kolter, Nicolas Loizou, Marc Lanctot, Ioannis Mitliagkas, Noam Brown, and Christian Kroer. A unified approach to reinforcement learning, quantal response equilibria, and two-player zero-sum games. *arXiv preprint arXiv:2206.05825*, 2022.

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

Sebastian U Stich. Unified optimal analysis of the (stochastic) gradient method. *arXiv preprint arXiv:1907.04232*, 2019.

Daniil Vankov, Angelia Nedić, and Lalitha Sankar. Last iterate convergence of popov method for non-monotone stochastic variational inequalities. *arXiv preprint arXiv:2310.16910*, 2023.

Daniil Vankov, Angelia Nedić, and Lalitha Sankar. Generalized smooth variational inequalities: Methods with adaptive stepsizes. In *Proceedings of the 41st International Conference on Machine Learning, PMLR*, volume 235, pp. 49137–49170, 2024.

Bohan Wang, Huishuai Zhang, Zhiming Ma, and Wei Chen. Convergence of adagrad for non-convex objectives: Simple proofs and relaxed assumptions. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 161–190. PMLR, 2023.

Chen-Yu Wei, Chung-Wei Lee, Mengxiao Zhang, and Haipeng Luo. Linear last-iterate convergence in constrained saddle-point optimization. In *International Conference on Learning Representations*, 2021.

Wenhan Xian, Ziyi Chen, and Heng Huang. Delving into the convergence of generalized smooth minimax optimization. In *Forty-first International Conference on Machine Learning*, 2024.

Zhuoyu Xiao and Uday V Shanbhag. Computing equilibria in stochastic nonconvex and non-monotone games via gradient-response schemes. *arXiv preprint arXiv:2504.14056*, 2025.

Farzad Yousefian, Angelia Nedić, and Uday V. Shanbhag. Optimal robust smoothing extragradient algorithms for stochastic variational inequality problems. In *53rd IEEE Conference on Decision and Control (CDC), Los Angeles, CA, USA, December 15-17, 2014*, pp. 5831–5836. IEEE, 2014.

Farzad Yousefian, Angelia Nedić, and Uday V. Shanbhag. On smoothing, regularization, and averaging in stochastic approximation methods for stochastic variational inequality problems. *Mathematical Programming*, 165:391–431, 2017.

Nikos Zarifis, Puqian Wang, Ilias Diakonikolas, and Jelena Diakonikolas. Robustly learning single-index models via alignment sharpness. *arXiv preprint arXiv:2402.17756*, 2024.

Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. In *International Conference on Learning Representations*, 2020.

Qi Zhang, Yi Zhou, and Shaofeng Zou. Convergence guarantees for rmsprop and adam in generalized-smooth non-convex optimization with affine noise variance. *arXiv preprint arXiv:2404.01436*, 2024.

Zirui Zhou and Anthony Man-Cho So. A unified approach to error bounds for structured convex optimization problems. *Mathematical Programming*, 165:689–728, 2017.

A Technical Lemmas

In our analysis, we use the properties of the projection operator $P_U(\cdot)$ given in the following lemma.

Lemma A.1. *(Theorem 1.5.5 and Lemma 12.1.13 in Facchinei & Pang (2003)) Given a nonempty convex closed set $U \subset \mathbb{R}^m$, the projection operator $P_U(\cdot)$ has the following properties:*

$$\langle v - P_U(v), u - P_U(v)\rangle \leq 0 \quad \text{for all } u \in U, v \in \mathbb{R}^m, \tag{26}$$

$$\|u - P_U(v)\|^2 \leq \|u - v\|^2 - \|v - P_U(v)\|^2 \quad \text{for all } u \in U, v \in \mathbb{R}^m, \tag{27}$$

$$\|P_U(u) - P_U(v)\| \leq \|u - v\| \quad \text{for all } u, v \in \mathbb{R}^m. \tag{28}$$

In the forthcoming analysis, we use Lemma 11 Polyak (1987), which is stated below.

Lemma A.2. *[Lemma 11 Polyak (1987)] Let $\{v_k\}, \{z_k\}, \{a_k\}$, and $\{b_k\}$ be nonnegative random scalar sequences such that almost surely for all $k \geq 0$,*

$$\mathbb{E}[v_{k+1} \mid \mathcal{F}_k] \leq (1 + a_k)v_k - z_k + b_k, \tag{29}$$

where $\mathcal{F}_k = \{v_0, \ldots, v_k, z_0, \ldots, z_k, a_0, \ldots, a_k, b_0, \ldots, b_k\}$, and a.s. $\sum_{k=0}^{\infty} a_k < \infty$, $\sum_{k=0}^{\infty} b_k < \infty$. *Then, almost surely,* $\lim_{k\to\infty} v_k = v$ *for some nonnegative random variable v and* $\sum_{k=0}^{\infty} z_k < \infty$.

As a direct consequence of Lemma A.2, when the sequences $\{v_k\}, \{z_k\}, \{a_k\}, \{b_k\}$ are deterministic, we obtain the following result.

Lemma A.3. *Let $\{\bar{v}_k\}, \{\bar{z}_k\}, \{\bar{a}_k\}, \{\bar{b}_k\}$ be nonnegative scalar sequences such that for all $k \geq 0$,*

$$\bar{v}_{k+1} \leq (1 + \bar{a}_k)\bar{v}_k - \bar{z}_k + \bar{b}_k, \tag{30}$$

where $\sum_{k=0}^{\infty} \bar{a}_k < \infty$ *and* $\sum_{k=0}^{\infty} \bar{b}_k < \infty$. *Then,* $\lim_{k\to\infty} \bar{v}_k = \bar{v}$ *for some scalar $\bar{v} \geq 0$ and* $\sum_{k=0}^{\infty} \bar{z}_k < \infty$.

Lemma A.4. *Let X be a non-negative random variable such that $\mathbb{E}[X^\rho]$ is defined for some $\rho \geq 1$, and $\mathbb{E}[X^\rho] \neq 0$, then for every $a > 0$ it holds*

$$\mathbb{P}(X > a(\mathbb{E}[X^\rho])^{1/\rho}) \leq \frac{1}{a^\rho}. \tag{31}$$

Proof. Let $Y = X^\rho$. By the conditions of the lemma, the expectation $\mathbb{E}[Y] = \mathbb{E}[X^\rho]$ is well defined. Then, by Markov's inequality:

$$\begin{aligned}\mathbb{P}(X > a(\mathbb{E}[X^\rho])^{1/\rho}) &= \mathbb{P}(Y > a^\rho \mathbb{E}[X^\rho]) \\ &\leq \frac{\mathbb{E}[X^\rho]}{a^\rho \mathbb{E}[X^\rho]}.\end{aligned}$$

■

Lemma A.5. *Let a_1, a_2 be nonnegative scalar and $p > 0$. Then the following inequality holds:*

$$(a_1 + a_2)^p \leq 2^{p-1}(a_1^p + a_2^p).$$

Proof. Let $a = (a_1, a_2), b = (1, 1)$, then by Hölder inequality:

$$\begin{aligned} a_1 + a_2 &= \|ab\| \\ &\leq \|a\|_p \|b\|_{p/(p-1)} \\ &\leq (a_1^p + a_2^p)^{1/p}(1+1)^{(p-1)/p}. \end{aligned}$$

Raising the inequality in the power p we get the desired relation. ■

A.1 Auxiliary Results

In our analysis we make use of Lemma 3 and Lemma 7 from Stich (2019), as well as the sequences provided in the proofs in Stich (2019).

Lemma A.6. *Let $\{r_k\}$ and $\{s_k\}$ be nonnegative scalar sequences that satisfy the following relation*

$$r_{k+1} \leq (1 - a\alpha_k)r_k - b\alpha_k s_k + c\gamma_k^2 \qquad \text{for all } k \geq 0,$$

where $a > 0$, $b > 0$, $c \geq 0$, and

$$\gamma_k = \frac{2}{a\left(\frac{2d}{a} + k\right)} \qquad \text{for all } k \geq 0,$$

where $d \geq a$. Then, for any given $K \geq 0$, the following relation holds:

$$\frac{b}{W_K}\sum_{k=0}^{K} w_k s_k + a r_{K+1} \leq \frac{8d^2}{aK^2} r_0 + \frac{2c}{aK},$$

where $w_k = 2d/a + k$, $0 \leq k \leq K$, and $W_K = \sum_{k=0}^{K} w_k$.

Lemma A.7. *For $1 > q \geq 1/2$ and $K \geq 1$, we have*

$$\sum_{t=0}^{K} \frac{1}{(t+1)^q} \geq \frac{1}{1-q}((K+1)^{1-q} - 2^{1-q}). \tag{32}$$

For $q = 1/2$ and $K \geq 1$,

$$\sum_{t=0}^{K} \frac{1}{(t+1)^{2q}} \leq \log(K+1). \tag{33}$$

For $q > 1/2$ and $K \geq 1$,

$$\sum_{t=0}^{K} \frac{1}{(t+1)^{2q}} \leq \frac{1}{2q-1}. \tag{34}$$

Proof. Let $1 > q \geq 1/2$ and $K \geq 1$. Then, it holds

$$\sum_{t=0}^{K} \frac{1}{(t+1)^q} \geq \int_{s=1}^{K} \frac{ds}{(s+1)^q} = \frac{1}{1-q}((K+1)^{1-q} - 2^{1-q}). \tag{35}$$

When $q = 1/2$ and $K \geq 1$, then

$$\sum_{t=0}^{K} \frac{1}{t+1} \leq \int_{s=0}^{K} \frac{ds}{s+1} = \log(K+1). \tag{36}$$

When $q > 1/2$ and $K \geq 1$, we have that

$$\sum_{t=0}^{K} \frac{1}{(t+1)^{2q}} \leq \int_{s=0}^{K} \frac{ds}{(s+1)^{2q}} = \frac{1}{2q-1} - \frac{1}{(2q-1)(K+1)^{2q-1}} < \frac{1}{2q-1}. \tag{37}$$

■

B Projection Method Analysis

B.1 Proof of Lemma 3.1

Proof. Let $k \geq 0$ be arbitrary but fixed. From the definition of u_{k+1} in (7), we have $\|u_{k+1} - y\|^2 = \|P_U(u_k - \gamma_k \Phi(u_k, \xi_k)) - y\|^2$ for all $y \in U$. Using the non-expansiveness property of projection operator (28) we obtain for all $y \in U$ and $k \geq 0$,

$$\begin{aligned}
\|u_{k+1} - y\|^2 &\leq \|u_k - \gamma_k \Phi(u_k, \xi_k) - y\|^2 \\
&= \|u_k - y\|^2 - 2\gamma_k \langle \Phi(u_k, \xi_k), u_k - y\rangle + \gamma_k^2 \|\Phi(u_k, \xi_k)\|^2 \\
&= \|u_k - y\|^2 + \gamma_k^2 \|\Phi(u_k, \xi_k)\|^2 \\
&\quad - 2\gamma_k \langle F(u_k), u_k - y\rangle + 2\gamma_k \langle e_k, u_k - y\rangle,
\end{aligned} \tag{38}$$

where $e_k = F(u_k) - \Phi(u_k, \xi_k)$. By the definition of the stepsizes (9), $\gamma_k = \beta_k \min\{1, \frac{1}{\|\Phi(u_k, \xi_k^2)\|}\}$, then the term $\gamma_k^2 \|\Phi(u_k, \xi_k)\|^2$ can be upper bounded as follows

$$\begin{aligned}
\gamma_k^2 \|\Phi(u_k, \xi_k)\|^2 &= \gamma_k^2 \|\Phi(u_k, \xi_k) - F(u_k) + F(u_k) - \Phi(u_k, \xi_k^2) + \Phi(u_k, \xi_k^2)\|^2 \\
&\leq \beta_k^2 \min\left\{1, \frac{1}{\|\Phi(u_k, \xi_k^2)\|^2}\right\} 3(\|e_k\|^2 + \|e_k^2\|^2 + \|\Phi(u_k, \xi_k^2)\|^2) \\
&\leq 3\beta_k^2 \|e_k\|^2 + 3\beta_k^2 \|e_k^2\|^2 + 3\beta_k^2,
\end{aligned} \tag{39}$$

where $e_k = \Phi(u_k, \xi_k) - F(u_k), e_k^2 = \Phi(u_k, \xi_k^2) - F(u_k)$. Thus,

$$\begin{aligned}
\|u_{k+1} - y\|^2 \leq \|u_k - y\|^2 &- 2\gamma_k \langle F(u_k), u_k - y\rangle \\
&+ 2\gamma_k \langle e_k, u_k - y\rangle + 3\beta_k^2 (\|e_k\|^2 + \|e_k^2\|^2 + 1).
\end{aligned} \tag{40}$$

Plugging in $y = u^* \in U^*$, where u^* is an arbitrary solution, and using p-quasi sharpness we get:

$$\begin{aligned}
\|u_{k+1} - u^*\|^2 \leq \|u_k - u^*\|^2 &- 2\mu\gamma_k \mathrm{dist}^p(u_k, U^*) \\
&+ 2\gamma_k \langle e_k, u_k - u^*\rangle + 3\beta_k^2 (\|e_k\|^2 + \|e_k^2\|^2 + 1).
\end{aligned} \tag{41}$$

Using stochastic properties of ξ_k and ξ_k^2 imposed by Assumption 2.5, and the conditional independence of ξ_k and ξ_k^2, we have:

$$\mathbb{E}[\gamma_k \langle e_k, u_k - u^*\rangle | \mathcal{F}_{k-1}] = \mathbb{E}[\gamma_k \mid \mathcal{F}_{k-1}] \langle \mathbb{E}[e_k \mid \mathcal{F}_{k-1}], u_k - u^*\rangle = 0.$$

$$\mathbb{E}[\|e_k\|^2 \mid \mathcal{F}_{k-1}] \leq \sigma^2, \quad \mathbb{E}[\|e_k^2\|^2 \mid \mathcal{F}_{k-1}] \leq \sigma^2.$$

Thus, by taking the conditional expectation on $\mathcal{F}_{k-1} = \{\xi_0, \xi_0^2, \ldots, \xi_{k-1}, \xi_{k-1}^2\}$ in relation (41) we obtain for all $u^* \in U^*$ and for all $k \geq 0$:

$$\begin{aligned}
\mathbb{E}[\|u_{k+1} - u^*\|^2 | \mathcal{F}_{k-1}] \leq \|u_k - u^*\|^2 &+ 3\beta_k^2 (2\sigma^2 + 1) \\
&- 2\mu\beta_k \mathbb{E}\left[\min\left\{1, \frac{1}{\|\Phi(u_k, \xi_k^2)\|}\right\} \mid \mathcal{F}_{k-1}\right] \mathrm{dist}^p(u_k, U^*).
\end{aligned} \tag{42}$$

The equation (42) satisfies the condition of Lemma A.2 with

$$v_k = \|u_k - u^*\|^2, \quad a_k = 0, \quad z_k = 2\mu\mathbb{E}[\gamma_k \mid \mathcal{F}_{k-1}]\,\mathrm{dist}^p(u_k, U^*), \quad b_k = 3\beta_k^2(2\sigma^2 + 1). \tag{43}$$

By Lemma A.2, it follows that the sequence $\{v_k\}$ converges *a.s.* to a non-negative scalar for any $u^* \in U^*$, and almost surely we have

$$\sum_{k=0}^{\infty} \mathbb{E}[\gamma_k \mid \mathcal{F}_{k-1}]\,\mathrm{dist}^p(u_k, U^*) < \infty. \tag{44}$$

Since the sequence $\{\|u_k - u^*\|^2\}$ converges *a.s.* for all $u^* \in U^*$, it follows that the sequence $\{\|u_k - u^*\|\}$ is bounded *a.s.* for all $u^* \in U^*$. ■

B.2 Proof of Theorem 3.2

Lemma B.1. *Let γ_k are given by (9) then the series of $\{\mathbb{E}[\gamma_k \mid \mathcal{F}_{k-1}]\}$ is nonsummable almost surely,*

$$\sum_{k=0}^{\infty} \mathbb{E}[\gamma_k \mid \mathcal{F}_{k-1}] = \infty \quad \text{a.s.} \tag{45}$$

Proof. We will show that $\sum_{k=0}^{\infty} \beta_k \min\left\{1, \frac{1}{\|\Phi(u_k,\xi_k^2)\|}\right\} = \infty$ almost surely by the sequences of lower bound on this series. Consider the following event:

$$A_k = \{\|e_k^2\| \le 2\sigma\},$$

where $e_k^2 = \Phi(u_k, \xi_k^2) - F(u_k)$ is a stochastic error from the sample for the clipping stepsize γ_k. Define $x_k = \min\left\{1, \frac{1}{\|\Phi(u_k,\xi_k^2)\|}\right\}$, then,

$$x_k = x_k\mathbb{I}(A_k) + x_k\mathbb{I}(\overline{A}_k) \ge x_k\mathbb{I}(A_k), \tag{46}$$

where the random variable $\mathbb{I}(A_k) =$ is an indicator function $\mathbb{I}(A_k) = \begin{cases} 1, & \omega \in A_k, \\ 0, & \omega \notin A_k \end{cases}$, where $\omega = (\omega_0, \omega_2, \ldots)$ is a full path in the path space $\Omega = \mathbb{R}^{\mathbb{N}_0}$.

By the definition of x_k, the triangle inequality and definition of $\mathbb{I}(A_k)$, we have

$$\begin{aligned}
\mathbb{E}[x_k\mathbb{I}(A_k) \mid \mathcal{F}_{k-1}] &= \mathbb{E}[\min\left\{1, \frac{1}{\|\Phi(u_k,\xi_k^2)\|}\right\}\mathbb{I}(A_k) \mid \mathcal{F}_{k-1}] \\
&\ge \mathbb{E}[\min\left\{1, \frac{1}{\|F(u_k)\| + \|e_k^2\|}\right\}\mathbb{I}(A_k) \mid \mathcal{F}_{k-1}] \\
&\ge \mathbb{E}[\min\left\{1, \frac{1}{\|F(u_k)\| + 2\sigma}\right\}\mathbb{I}(A_k) \mid \mathcal{F}_{k-1}].
\end{aligned} \tag{47}$$

Taking conditional expectation on (46) and combining it with (47) we have the lowerbound

$$\sum_{k=0}^{\infty} \gamma_k \ge \sum_{k=0}^{\infty} \beta_k\mathbb{E}[\min\left\{1, \frac{1}{\|F(u_k)\| + 2\sigma}\right\}\mathbb{I}(A_k)|\mathcal{F}_{k-1}]. \tag{48}$$

To bound $p_k := \mathbb{E}[\mathbb{I}(A_k)|\mathcal{F}_{k-1}] = \mathbb{P}(A_k \mid \mathcal{F}_{k-1})$ we provide an upperbound on $\mathbb{P}(\overline{A}_k \mid \mathcal{F}_{k-1})$ using Markov's inequality and Assumption 2.5:

$$\mathbb{P}(\overline{A}_k \mid \mathcal{F}_{k-1}) = \mathbb{P}(\|e_k^2\| > 2\mathbb{E}[\|e_k^2\| \mid \mathcal{F}_{k-1}]\}) \le \frac{\mathbb{E}[\|e_k^2\| \mid \mathcal{F}_{k-1}]}{2\mathbb{E}[\|e_k^2\| \mid \mathcal{F}_{k-1}])} = \frac{1}{2}. \tag{49}$$

This implies $\mathbb{E}[\mathbb{I}(A_k)|\mathcal{F}_{k-1}] \geq \frac{1}{2}$ a.s.

To further lower bound $x_k\mathbb{I}(A_k)$ we show *a.s.* boundedness of $\|F(u_k)\|$ for all $k \geq 0$, using property of α-symmetric operators. To estimate $\|F(u_k)\|$, we add and subtract $F(v^*)$, where $v^* \in U^*$ is an arbitrary but fixed solution, and get

$$\|F(u_k)\| = \|F(u_k) - F(v^*) + F(v^*)\| \leq \|F(u_k) - F(v^*)\| + \|F(v^*)\|.$$

Define the following event:

$$A = \{\omega \in \Omega : \exists\, C(\omega) \in \mathbb{R} \text{ s.t.} \|u_k(\omega) - v^*\| < C(\omega)\ \forall\, k \geq 0\}.$$

Based on Lemma 3.1, the sequence $\{\|u_k - v^*\|\}$ is bounded *a.s.*, and thus $\mathbb{P}(A) = 1$. Let $\omega \in A$, now we can estimate $\|F(u_k(\omega))\|$ using the α-symmetric assumption on the operator.
Case $\alpha \in (0, 1)$.

$$\|F(u_k(\omega)) - F(v^*)\| \leq \|u_k(\omega) - v^*\|(K_0 + K_1\|F(v^*)\|^\alpha + K_2\|u_k(\omega) - v^*\|^{\alpha/(1-\alpha)}). \tag{50}$$

Since $\omega \in A$, it follows that $\|u_k(\omega) - v^*\| \leq C(\omega)$ for all $k \geq 0$. Using this fact and (50) we obtain that for all $k \geq 0$,

$$\|F(u_k(\omega))\| \leq C(\omega)(K_0 + K_1\|F(v^*)\|^\alpha + K_2 C(\omega)^{\alpha/(1-\alpha)}) + \|F(v^*)\|. \tag{51}$$

Therefore, the sequence $\{\|F(u_k(\omega))\|\}$ is upper bounded by $C_1(\omega) = C(\omega)(K_0 + K_1\|F(v^*)\|^\alpha + K_2 C(\omega)^{\alpha/(1-\alpha)}) + \|F(v^*)\|$.
Case $\alpha = 1$.
For $\alpha = 1$ by Proposition 2.2 we have

$$\|F(u_k(\omega)) - F(v^*)\| \leq \|u_k(\omega) - v^*\|(L_0 + L_1\|F(v^*)\|)\exp(L_1\|u_k(\omega) - v^*\|). \tag{52}$$

Therefore, for all $k \geq 0$,

$$\begin{aligned}\|F(u_k(\omega))\| &\leq \|F(u_k(\omega)) - F(v^*)\| + \|F(v^*)\| \\ &\leq \|u_k(\omega) - v^*\|(L_0 + L_1\|F(v^*)\|)\exp(L_1\|u_k(\omega) - v^*\|) + \|F(v^*)\|.\end{aligned} \tag{53}$$

Since $\omega \in A$, we have $\|u_k(\omega) - v^*\| \leq C(\omega)$ for all $k \geq 0$, which when used in (53), implies that for all $k \geq 0$,

$$\begin{aligned}\|F(u_k(\omega))\| &\leq \|u_k(\omega) - v^*\|(L_0 + L_1\|F(v^*)\|)\exp(L_1\|u_k(\omega) - v^*\|) + \|F(v^*)\| \\ &\leq C(\omega)(L_0 + L_1\|F(v^*)\|)\exp(L_1 C(\omega)) + \|F(v^*)\|.\end{aligned} \tag{54}$$

Hence, the sequence $\{\|F(u_k(\omega))\|\}$ is upper bounded by $\overline{C}_1(\omega)$, where $\overline{C}_1(\omega) = C(\omega)(L_0 + L_1\|F(v^*)\|)\exp(L_1 C(\omega)) + \|F(v^*)\|$. Now, for both cases $\alpha \in (0, 1)$ and $\alpha = 1$ in (51) and (54), respectively, we have that $\|F(u_k(\omega))\|$ is upper bounded by $\max\{C_1(\omega), \overline{C}_1(\omega)\}$. Thus

$$\mathbb{P}(F(u_k) \text{ is bounded}) = 1.$$

Then almost surely we have (i) $F(u_k)$ is bounded (ii) $\mathbb{E}[\mathbb{I}(A_k)|\mathcal{F}_{k-1}] \geq \frac{1}{2}$. Consider $\omega \in \Omega$ such that (i), (ii) hold, then in a view of (48) we have

$$\begin{aligned}\sum_{k=0}^{\infty} \beta_k x_k(\omega) &\geq \frac{1}{2}\min\left\{1, \frac{1}{\overline{C}_1(\omega) + 2\sigma}\right\}\sum_{k=0}^{\infty}\beta_k \\ &= \infty\end{aligned} \tag{55}$$

where the last bound comes from $\sum_{k=0}^{\infty}\beta_k = \infty$, which concludes the proof. ■

Proof of Theorem 3.2.

Proof. By Lemma 3.1, we have

$$\sum_{k=0}^{\infty} \mathbb{E}[\gamma_k \mid \mathcal{F}_{k-1}] \operatorname{dist}^p(u_k, U^*) < \infty \quad \text{a.s.} \tag{56}$$

Due to Lemma B.1, the series $\sum_{k=0}^{\infty} \mathbb{E}[\gamma_k \mid \mathcal{F}_{k-1}] = \infty$ almost surely, then it follows that

$$\liminf_{k\to\infty} \operatorname{dist}^p(u_k, U^*) = 0 \qquad a.s. \tag{57}$$

Since $\|u_k - u^*\|$ converges *a.s.* for any given $u^* \in U^*$, the sequence $\{u_k\}$ is bounded *a.s.* and has accumulation points *a.s.* Let $\{k_i\}$ be an index sequence, such that

$$\lim_{i\to\infty} \operatorname{dist}^p(u_{k_i}, U^*) = \liminf_{k\to\infty} \operatorname{dist}^p(u_k, U^*) = 0 \quad a.s. \tag{58}$$

We assume that the sequence $\{u_{k_i}\}$ is convergent with a limit point $\bar{u}$; otherwise, we choose a convergent subsequence. Therefore,

$$\lim_{i\to\infty} \|u_{k_i} - \bar{u}\| = 0 \quad a.s. \tag{59}$$

Then, by (57), $\operatorname{dist}(\bar{u}, U^*) = 0$, thus $\bar{u} \in U^*$ *a.s.* since U^* is closed. Since the sequence $\{\|u_k - u^*\|\}$ converges *a. s.* for all $u^* \in U^*$, by (59) we have

$$\lim_{k\to\infty} \|u_k - \bar{u}\| = 0 \quad a.s. \tag{60}$$

■

B.3 Proof of Lemma 3.3

Proof. By taking the total expectation in (42) in Lemma 3.1 and using the definition of the stepsize γ_k, we obtain for any solution $u^* \in U^*$ and all $k \geq 0$,

$$\mathbb{E}[\|u_{k+1} - u^*\|^2] \leq \mathbb{E}[\|u_k - u^*\|^2] - 2\mu\mathbb{E}[\gamma_k \operatorname{dist}^p(u_k, U^*)] + 3\beta_k^2(2\sigma^2 + 1). \tag{61}$$

The equation (61) satisfies the conditions of Lemma A.3 with

$$\bar{v}_k = \mathbb{E}[\|u_k - u^*\|^2], \quad \bar{a}_k = 0, \quad \bar{z}_k = 2\mu\mathbb{E}[\gamma_k \operatorname{dist}^p(u_k, U^*)], \quad \bar{b}_k = 3\beta_k^2(2\sigma^2 + 1). \tag{62}$$

Thus, by Lemma A.3, it follows that the sequence $\{\mathbb{E}[\|u_k - u^*\|^2]\}$ converges to a non-negative scalar for any $u^* \in U^*$. Therefore, the sequence $\{\mathbb{E}[\|u_k - u^*\|^2]\}$ is bounded for all $u^* \in U^*$. Next, using the property of α-symmetric operators, we show that $\{\mathbb{E}[\|F(u_k)\|]\}$ is bounded. Let $v^* \in U^*$ be an arbitrary, but fixed solution. Then, by the α-symmetric property of F, we have that

$$\begin{aligned} \|F(u_k)\| &\leq \|F(u_k) - F(v^*)\| + \|F(v^*)\| \\ &\leq \|u_k - v^*\|(K_0 + K_1\|F(v^*)\|^\alpha + K_2\|u_k - v^*\|^{\alpha/(1-\alpha)}) + \|F(v^*)\|. \end{aligned} \tag{63}$$

Taking expectation, we obtain

$$\mathbb{E}[\|F(u_k)\|] \leq (K_0 + K_1\|F(v^*)\|^\alpha)\mathbb{E}[\|u_k - v^*\|] + K_2\mathbb{E}[\|u_k - v^*\|^{1+\alpha/(1-\alpha)}] + \|F(v^*)\|. \tag{64}$$

Notice, that $\mathbb{E}[\|u_k - v^*\|^{1+\alpha/(1-\alpha)})] = \mathbb{E}[(\|u_k - v^*\|^2)^{1/2(1-\alpha)})]$, and for $\alpha \leq 1/2$, the quantity $1/2(1-\alpha) \leq 1$. Thus, we can apply Jensen inequality for concave function

$$\mathbb{E}[(\|u_k - v^*\|^2)^{1/2(1-\alpha)})] \leq \mathbb{E}[\|u_k - v^*\|^2]^{1/2(1-\alpha)}.$$

Therefore, using these results and Jensen inequality for the first term in equation (64), we obtain

$$\mathbb{E}[\|F(u_k)\|] \leq (K_0 + K_1\|F(v^*)\|^\alpha)\mathbb{E}[\|u_k - v^*\|^2]^{1/2} + K_2\mathbb{E}[\|u_k - v^*\|^2]^{1/2(1-\alpha)} + \|F(v^*)\|. \tag{65}$$

Since $\mathbb{E}[\|u_k - v^*\|^2]$ is bounded, $\mathbb{E}[\|F(u_k)\|]$ is bounded by some constant $C_F > 0$ for all $k \geq 0$. ■

B.4 Proof of Theorem 3.4

Proof. Letting $y = P_{U^*}(u_k)$ in equation (40) in Lemma 3.1 and using p-quasi sharpness we obtain

$$\begin{aligned}\|u_{k+1} - P_{U^*}(u_k)\|^2 \leq \|u_k - P_{U^*}(u_k)\|^2 - 2\mu\gamma_k \mathrm{dist}^p(u_k, U^*) \\ + 2\gamma_k\langle e_k, u_k - P_{U^*}(u_k)\rangle + 3\beta_k^2(\|e_k\|^2 + \|e_k^2\|^2 + 1).\end{aligned} \tag{66}$$

By the definition of the distance function, we have

$$\mathrm{dist}^2(u_{k+1}, U^*) \leq \|u_{k+1} - P_{U^*}(u_k)\|^2.$$

Thus,

$$\begin{aligned}\mathrm{dist}^2(u_{k+1}, U^*) \leq \mathrm{dist}^2(u_k, U^*) - 2\mu\gamma_k \mathrm{dist}^p(u_k, U^*) \\ + 2\gamma_k\langle e_k, u_k - P_{U^*}(u_k)\rangle + 3\beta_k^2(\|e_k\|^2 + \|e_k^2\|^2 + 1).\end{aligned} \tag{67}$$

By Assumption 2.5 and the law of total expectation, and independence of samples ξ_k and ξ_k^2, it follows that

$$\begin{aligned}\mathbb{E}[\gamma_k\langle e_k, u_k - P_{U^*}(u_k)\rangle] &= \mathbb{E}[\mathbb{E}[\gamma_k\langle e_k, u_k - P_{U^*}(u_k)\rangle \mid \mathcal{F}_{k-1}]] \\ &= \mathbb{E}[\mathbb{E}[\gamma_k \mid \mathcal{F}_{k-1}]\langle \mathbb{E}[e_k \mid |\mathcal{F}_{k-1}], u_k - P_{U^*}(u_k)\rangle] \\ &= 0.\end{aligned} \tag{68}$$

Also, we have $\mathbb{E}[\mathbb{E}[\|e_k^1\| \mid \mathcal{F}_{k-1}] \leq \sigma^2$ and $\mathbb{E}[\mathbb{E}[\|e_k^2\| \mid \mathcal{F}_{k-1}] \leq \sigma^2$. Thus, by taking the total expectation in (67), we obtain

$$\mathbb{E}[\mathrm{dist}^2(u_{k+1}, U^*)] \leq \mathbb{E}[\mathrm{dist}^2(u_k, U^*)] - 2\mu\mathbb{E}[\gamma_k \mathrm{dist}^p(u_k, U^*)] + 3\beta_k^2(2\sigma^2 + 1). \tag{69}$$

We aim to upper bound $2\mu\mathbb{E}[\gamma_k \mathrm{dist}^p(u_k, U^*)]$. To do so consider an event A_k, defined as follows:

$$A_k = \{\|F(u_k)\| + \|e_k\| \leq 2(\mathbb{E}[\|F(u_k)\|] + \mathbb{E}[\|e_k\|])\}.$$

Then, by the law of total expectation, we obtain

$$\mathbb{E}[\gamma_k \mathrm{dist}^p(u_k, U^*)] = \mathbb{E}[\gamma_k \mathrm{dist}^p(u_k, U^*)|A_k]\mathbb{P}(A_k) + \mathbb{E}[\gamma_k \mathrm{dist}^p(u_k, U^*)|\overline{A}_k]\mathbb{P}(\overline{A}_k), \tag{70}$$

where $\overline{A}$ denotes the complement of an event A. We want to provide a lower bound on $\mathbb{P}(A_k)$. To do so, we upperbound $\mathbb{P}(\overline{A}_k)$ using Markov's inequality, as follows:

$$\begin{aligned}\mathbb{P}(\overline{A}_k) &= \mathbb{P}\left(\{\|F(u_k)\| + \|e_k\| > 2(\mathbb{E}[\|F(u_k)\|] + \mathbb{E}[\|e_k\|])\}\right) \\ &\leq \frac{\mathbb{E}[\|F(u_k)\|] + \mathbb{E}[\|e_k\|]}{2(\mathbb{E}[\|F(u_k)\|] + \mathbb{E}[\|e_k\|])} \\ &= \frac{1}{2}.\end{aligned} \tag{71}$$

Thus,

$$\begin{aligned}\mathbb{E}[\gamma_k \mathrm{dist}^p(u_k, U^*)] &= \mathbb{E}[\gamma_k \mathrm{dist}^p(u_k, U^*)|A_k](1 - \mathbb{P}(\overline{A}_k)) + \mathbb{E}[\gamma_k \mathrm{dist}^p(u_k, U^*)|\overline{A}_k]\mathbb{P}(\overline{A}_k) \\ &\geq \frac{1}{2}\mathbb{E}[\gamma_k \mathrm{dist}^p(u_k, U^*)|A_k] + \mathbb{E}[\gamma_k \mathrm{dist}^p(u_k, U^*)|\overline{A}_k]\mathbb{P}(\overline{A}_k) \\ &\geq \frac{1}{2}\mathbb{E}[\gamma_k \mathrm{dist}^p(u_k, U^*)|A_k].\end{aligned} \tag{72}$$

By the definition of the event A_k, we have

$$\begin{aligned}\mathbb{E}[\gamma_k \mathrm{dist}^p(u_k, U^*)|A_k] &= \beta_k\mathbb{E}\left[\min\left\{1, \frac{1}{\|\Phi(u_k, \xi_k)\|}\right\}\mathrm{dist}^p(u_k, U^*)|A_k\right] \\ &\geq \beta_k\mathbb{E}\left[\min\left\{1, \frac{1}{\|F(u_k)\| + \|e_k\|}\right\}\mathrm{dist}^p(u_k, U^*)|A_k\right] \\ &\geq \beta_k \min\left\{1, \frac{1}{2(\mathbb{E}[\|F(u_k)\|] + \mathbb{E}[\|e_k\|])}\right\}\mathbb{E}[\mathrm{dist}^p(u_k, U^*)|A_k].\end{aligned} \tag{73}$$

By Lemma 3.3, $\mathbb{E}[\|F(u_k)\|] \leq C_F$ for all $k \geq 0$, and by Assumption 2.5 and Jensen inequality, we have $\mathbb{E}[\|e_k\|] \leq \mathbb{E}[\|e_k\|^2]^{1/2} \leq \sigma$. Thus, it follows that

$$\mathbb{E}[\gamma_k \mathrm{dist}^p(u_k, U^*)] \geq \frac{1}{2}\beta_k \min\left\{1, \frac{1}{2(C_F+\sigma)}\right\} \mathbb{E}[\mathrm{dist}^p(u_k, U^*)]. \tag{74}$$

Combining equations (69) and (74), and using $a = \mu \min\left\{1, \frac{1}{2(C_F+\sigma)}\right\}$, we obtain

$$\mathbb{E}[\mathrm{dist}^2(u_{k+1}, U^*)] \leq \mathbb{E}[\mathrm{dist}^2(u_k, U^*)] - a\beta_k \mathbb{E}[\mathrm{dist}^p(u_k, U^*)] + 3\beta_k^2(2\sigma^2+1). \tag{75}$$

Now let $D_k = \mathbb{E}[\mathrm{dist}^2(u_k, U^*)]$, and consider the following two cases:

Case $p = 2$. When $p = 2$, equation (75) satisfies the assumptions of Lemma A.6 with

$$r_k = D_k, \quad \alpha_k = \beta_k, \quad s_k = 0, \quad d = a, \quad c = 3(2\sigma^2+1). \tag{76}$$

Then, by Lemma A.6, we get the following convergence rate for all $k \geq 1$,

$$D_{k+1} \leq \frac{8D_0}{k^2} + \frac{6(2\sigma^2+1)}{a^2 k}. \tag{77}$$

Case $p > 2$. When $p \geq 2$, by applying telescoping sum to inequality (75) and rearranging the terms we obtain

$$\mathbb{E}[a\sum_{t=0}^{k} \beta_k \mathrm{dist}^p(u_k, U^*)] \leq D_0 - D_{k+1} + 3(2\sigma^2+1)\sum_{t=0}^{k} \beta_k^2. \tag{78}$$

Since $p \geq 2$, the function $\mathrm{dist}^p(\cdot, U^*)$ is convex, thus by defining $\bar{u}_k = (\sum_{t=0}^{k} \beta_k)^{-1} \sum_{t=0}^{k} \beta_k u_t$ and applying Jensen inequality be obtain

$$(\sum_{t=0}^{k} \beta_k)\mathbb{E}[\mathrm{dist}^p(\bar{u}_k, U^*)] \leq \mathbb{E}[\sum_{t=0}^{k} \beta_k \mathrm{dist}^p(u_k, U^*)].$$

Since $p \geq 2$, by applying Jensen inequality one more time, we obtain

$$(\bar{D}_k)^{p/2} = \left(\mathbb{E}[\mathrm{dist}^2(\bar{u}_k, U^*)]\right)^{p/2} \leq \mathbb{E}\left[\left(\mathrm{dist}^2(\bar{u}_k, U^*)\right)^{p/2}\right] = \mathbb{E}[\mathrm{dist}^p(\bar{u}_k, U^*)].$$

Applying these estimates, we get

$$(\bar{D}_k)^{p/2} \sum_{t=0}^{k} \beta_t \leq \sum_{t=0}^{k} \beta_t D_t^{p/2} \leq \frac{1}{a}\left(D_0 - D_{k+1} + 3(2\sigma^2+1)\sum_{t=0}^{k} \beta_t^2\right). \tag{79}$$

Since $\beta_k = \frac{b}{(k+1)^q}$, with $b > 0, 1 > q > 1/2$, then $\{\beta_k\}$ satisfies the conditions of Lemma 3.3. Also, by Lemma A.7 the following inequalities hold: for all $k \geq 1$,

$$\sum_{t=0}^{k} \beta_t \geq \frac{b}{1-q}((k+1)^{1-q} - 2^{1-q}), \qquad \sum_{t=0}^{k} \beta_t^2 \leq \frac{b^2}{2q-1}. \tag{80}$$

Combining equations (79) and (80), and omitting D_{k+1}, we obtain

$$(\bar{D}_k)^{p/2} \leq \frac{(1-q)\left(D_0 + 3b^2(2\sigma^2+1)/(2q-1)\right)}{ab\left((k+1)^{1-q} - 2^{1-q}\right)}. \tag{81}$$

Raising both sides of the preceding inequality in power $2/p$, we obtain

$$\bar{D}_k \leq \frac{(1-q)^{2/p}\left(D_0 + 3b^2(2\sigma^2+1)/(2q-1)\right)^{2/p}}{(ab)^{2/p}\left((k+1)^{1-q} - 2^{1-q}\right)^{2/p}}. \tag{82}$$

■

C Korpelevich Method analysis

Lemma C.1. *Let U be a closed convex set. Then, for the iterate sequences $\{u_k\}$ and $\{h_k\}$ generated by the stochastic Korpelevich method (8) and $y \in U$ and $k \geq 0$,*

$$\begin{aligned}\|h_{k+1} - y\|^2 \leq \|h_k - y\|^2 - \|h_k - u_k\|^2 - 2\gamma_k\langle F(u_k), u_k - y\rangle - 2\gamma_k\langle e_k^2, u_k - y\rangle \\ + 3\gamma_k^2\|F(h_k) - F(u_k)\|^2 + 3\gamma_k^2(\|e_k^2\|^2 + \|e_k^1\|^2),\end{aligned}$$

where $e_k^1 = \Phi(h_k, \xi_k^1) - F(h_k)$, $e_k^2 = \Phi(u_k, \xi_k^2) - F(u_k)$ for all $k \geq 0$.

Proof. Let $k \geq 0$ be arbitrary but fixed. By the definition of h_{k+1} in (8), we have $\|h_{k+1} - y\| = \|P_U(h_k - \gamma_k\Phi(u_k, \xi_k^2)) - y\|$ for any $y \in U$. Using the projection inequality, we obtain for any $y \in U$,

$$\begin{aligned}\|h_{k+1} - y\|^2 &\leq \|h_k - \gamma_k\Phi(u_k, \xi_k^2) - y\|^2 - \|h_{k+1} - h_k + \gamma_k\Phi(u_k, \xi_k^2)\|^2 \\ &= \|h_k - y\|^2 - \|h_{k+1} - h_k\|^2 + 2\gamma_k\langle\Phi(u_k, \xi_k^2), y - h_{k+1}\rangle.\end{aligned} \tag{83}$$

Next, we consider the term $\|h_{k+1} - h_k\|^2$, where we add and subtract u_k, thus

$$\|h_{k+1} - h_k\|^2 = \|h_{k+1} - u_k\|^2 + \|h_k - u_k\|^2 - 2\langle h_{k+1} - u_k, h_k - u_k\rangle. \tag{84}$$

Adding and subtracting $2\gamma_k\langle\Phi(h_k, \xi_k^1), u_k - h_{k+1}\rangle$, and combining (83) and (84) we obtain

$$\begin{aligned}\|h_{k+1} - y\|^2 &\leq \|h_k - y\|^2 - \|h_{k+1} - u_k\|^2 - \|h_k - u_k\|^2 + 2\langle h_{k+1} - u_k, h_k - u_k\rangle \\ &\quad + 2\gamma_k\langle\Phi(u_k, \xi_k^2), y - u_k + u_k - h_{k+1}\rangle + 2\gamma_k\langle\Phi(h_k, \xi_k^1) - \Phi(h_k, \xi_k^1), u_k - h_{k+1}\rangle \\ &\leq \|h_k - y\|^2 - \|h_{k+1} - u_k\|^2 - \|h_k - u_k\|^2 + 2\langle h_{k+1} - u_k, h_k - \gamma_k\Phi(h_k, \xi_k^1) - u_k\rangle \\ &\quad + 2\gamma_k\langle\Phi(u_k, \xi_k^2), y - u_k\rangle + 2\gamma_k\langle\Phi(h_k, \xi_k^1) - \Phi(u_k, \xi_k^2), h_{k+1} - u_k\rangle.\end{aligned} \tag{85}$$

Since $u_k = P_U(h_k - \gamma_k\Phi(h_k, \xi_k^1))$ and $h_{k+1} \in U$, by the projection inequality in (26), it follows that

$$2\langle h_{k+1} - u_k, h_k - \gamma_k\Phi(h_k, \xi_k^1) - u_k\rangle \leq 0.$$

Using Cauchy-Schwarz inequality and relation $2ab \leq a^2 + b^2$ for $a, b \in \mathbb{R}$, we obtain

$$\begin{aligned}2\gamma_k\langle\Phi(h_k, \xi_k^1) - \Phi(u_k, \xi_k^2), h_{k+1} - u_k\rangle &\leq 2\gamma_k\|\Phi(h_k, \xi_k^1) - \Phi(u_k, \xi_k^2)\|\|h_{k+1} - u_k\| \\ &\leq \gamma_k^2\|\Phi(h_k, \xi_k^1) - \Phi(u_k, \xi_k^2)\|^2 + \|h_{k+1} - u_k\|^2.\end{aligned}$$

Using triangle inequality and relation $(\sum_{i=1}^m a_i)^2 \leq m\sum_{i=1}^m a_i^2$ we get

$$\begin{aligned}\|\Phi(h_k, \xi_k^1) - \Phi(u_k, \xi_k^2)\|^2 &= \|\Phi(h_k, \xi_k^1) - F(h_k) + F(h_k) - F(u_k) + F(u_k) - \Phi(u_k, \xi_k^2)\|^2 \\ &\leq 3(\|e_k^1\|^2 + \|F(h_k) - F(u_k)\|^2 + \|e_k^2\|^2).\end{aligned}$$

Combining the preceding three estimates with (85), we get the stated relation

$$\begin{aligned}\|h_{k+1} - y\|^2 \leq \|h_k - y\|^2 - \|h_k - u_k\|^2 - 2\gamma_k\langle F(u_k), u_k - y\rangle - 2\gamma_k\langle e_k^2, u_k - y\rangle \\ + 3\gamma_k^2\|F(h_k) - F(u_k)\|^2 + 3\gamma_k^2(\|e_k^2\|^2 + \|e_k^1\|^2).\end{aligned}$$

■

C.1 Proof of Lemma 4.1

Proof. By Lemma C.1 we have for all $k \geq 0$ and for all $y \in U$,

$$\begin{aligned}\|h_{k+1} - y\|^2 \leq \|h_k - y\|^2 - \|h_k - u_k\|^2 - 2\gamma_k\langle F(u_k), u_k - y\rangle - 2\gamma_k\langle e_k^2, u_k - y\rangle \\ + 3\gamma_k^2\|F(h_k) - F(u_k)\|^2 + 3\gamma_k^2(\|e_k^2\|^2 + \|e_k^1\|^2),\end{aligned} \tag{86}$$

with $e_k^1 = \Phi(h_k, \xi_k^1) - F(h_k)$ and $e_k^2 = \Phi(u_k, \xi_k^2) - F(u_k)$ for all $k \geq 0$. We want to estimate the term $\|F(h_k) - F(u_k)\|^2$ on the RHS of the inequality using the fact that $F(\cdot)$ is an α-symmetric operator for two cases **(a)** $\alpha \in (0,1)$ and **(b)** $\alpha = 1$.

Case $\alpha \in (0,1)$. Using the alternative characterization of α-symmetric operators from Proposition 2.2(a) (as given in (4)), when $\alpha \in (0,1)$, the next inequality holds for any $k \geq 0$,

$$\|F(h_k) - F(u_k)\| \leq \|h_k - u_k\|(K_0 + K_1\|F(h_k)\|^\alpha + K_2\|h_k - u_k\|^{\alpha/(1-\alpha)}). \tag{87}$$

We want to separate $\|F(h_k)\|$ into two parts: stochastic approximation of operator $\Phi(h_k, \xi_k^1)$ and error e_k^1. Recall that $e_k^1 = F(h_k) - \Phi(h_k, \xi_k^1)$, then based on triangle inequality $\|F(h_k)\| \leq \|\Phi(h_k, \xi_k^1)\| + \|e_k^1\|$, and since $\alpha \leq 1$ we obtain

$$\|F(h_k)\|^\alpha \leq \|\Phi(h_k, \xi_k^1)\|^\alpha + \|e_k^1\|^\alpha. \tag{88}$$

Thus, combining this fact with (87) we get the following estimation

$$\|F(h_k) - F(u_k)\| \leq \|h_k - u_k\|(K_0 + K_1\|\Phi(h_k, \xi_k^1)\|^\alpha + K_1\|e_k^1\|^\alpha + K_2\|h_k - u_k\|^{\alpha/(1-\alpha)}). \tag{89}$$

By the projection property (27) and the stepsize choice (16), we have

$$\|h_k - u_k\| \leq \gamma_k\|\Phi(h_k, \xi_k^1)\| = \beta_k \min\{1, \frac{1}{\|\Phi(h_k, \xi_k^1)\|}\}\|\Phi(h_k, \xi_k^1)\| \leq \beta_k \leq 1. \tag{90}$$

Then, $K_2\|h_k - u_k\|^{\alpha/(1-\alpha)} \leq K_2$, and

$$\begin{aligned}
\gamma_k\|F(h_k) - F(u_k)\| &\leq \gamma_k(K_0 + K_1\|\Phi(h_k, \xi_k^1)\|^\alpha + K_1\|e_k^1\|^\alpha + K_2)\|h_k - u_k\| \\
&\leq \beta_k(K_0 \min\{1, \frac{1}{\|\Phi(h_k, \xi_k^1)\|}\} + K_1 \min\{1, \frac{1}{\|\Phi(h_k, \xi_k^1)\|}\}\|\Phi(h_k, \xi_k^1)\|^\alpha)\|h_k - u_k\| \\
&\quad + \beta_k(K_1\|e_k^1\|^\alpha + K_2 \min\{1, \frac{1}{\|\Phi(h_k, \xi_k^1)\|}\})\|h_k - u_k\| \\
&\leq \beta_k(K_0 + K_1 + K_2)\|h_k - u_k\| + \beta_k K_1\|e_k^1\|^\alpha\|h_k - u_k\|.
\end{aligned} \tag{91}$$

By inequality (90), we have $\|h_k - u_k\| \leq 1$, and using this estimate in equation (91) we obtain

$$\gamma_k\|F(h_k) - F(u_k)\| \leq \beta_k(K_0 + K_1 + K_2)\|h_k - u_k\| + \beta_k K_1\|e_k^1\|^\alpha. \tag{92}$$

Case $\alpha = 1$. Based on the alternative characterization of α-symmetric operators from Proposition 2.2(b) (as given in (16)), when $\alpha = 1$, the following inequality holds for any $k \geq 0$,

$$\|F(h_k) - F(u_k)\| \leq \|h_k - u_k\|(L_0 + L_1\|F(h_k)\|)\exp(L_1\|h_k - u_k\|). \tag{93}$$

We upperbound $\|F(h_k)\|$ using equation (88) and get

$$\|F(h_k) - F(u_k)\| \leq \|h_k - u_k\|(L_0 + L_1\|\Phi(h_k, \xi_k^1)\| + L_1\|e_k^1\|)\exp(L_1\|h_k - u_k\|). \tag{94}$$

Note that relation in (90) holds irrespective of the value of α. Thus, since $\|h_k - u_k\| \leq 1$, we have $\exp(L_1\|h_k - u_k\|) \leq \exp(L_1\beta_k)$, and we obtain

$$\begin{aligned}
\gamma_k\|F(h_k) - F(u_k)\| &\leq \gamma_k(L_0 + L_1\|\Phi(h_k, \xi_k^1)\| + L_1\|e_k^1\|)\exp(L_1\beta_k)\|h_k - u_k\| \\
&= \exp(L_1\beta_k)L_0\beta_k \min\{1, \frac{1}{\|\Phi(h_k, \xi_k^1)\|}\}\|h_k - u_k\| \\
&\quad + \exp(L_1\beta_k)L_1\beta_k \min\{1, \frac{1}{\|\Phi(h_k, \xi_k^1)\|}\}\|\Phi(h_k, \xi_k^1)\|\|h_k - u_k\| \\
&\quad + \exp(L_1\beta_k)L_1\beta_k \min\{1, \frac{1}{\|\Phi(h_k, \xi_k^1)\|}\}\|e_k^1\|\|h_k - u_k\| \\
&\leq \exp(L_1\beta_k)\beta_k(L_0 + L_1 + L_1\|e_k^1\|)\|h_k - u_k\|.
\end{aligned} \tag{95}$$

By inequality (90), we have $\|h_k - u_k\| \leq 1$. Using this estimate in (95), we further obtain

$$\gamma_k \|F(h_k) - F(u_k)\| \leq \exp(L_1\beta_k)\beta_k(L_0 + L_1)\|h_k - u_k\| + \exp(L_1\beta_k)\beta_k\|e_k^1\|. \tag{96}$$

Now, we are done with the cases of α values. Let

$$C_a(\beta_k, \alpha) = \begin{cases} (K_0 + K_1 + K_2), & \text{when } \alpha \in (0,1), \\ \exp(L_1\beta_k)(L_0 + L_1), & \text{when } \alpha = 1. \end{cases} \tag{97}$$

Also, define

$$C_e(\beta_k, \alpha) = \begin{cases} K_1, & \text{when } \alpha \in (0,1), \\ \exp(L_1\beta_k), & \text{when } \alpha = 1. \end{cases} \tag{98}$$

Then, by inequality $(\sum_{i=1}^m a_i)^2 \leq m \sum_{i=1}^m a_i^2$, for both cases we have

$$\gamma_k^2 \|F(h_k) - F(u_k)\|^2 \leq 2\beta_k^2 C_a(\beta_k, \alpha)^2 \|h_k - u_k\|^2 + 2\beta_k^2 C_e(\beta_k, \alpha)^2 \|e_k^1\|^{2\alpha}. \tag{99}$$

Combining preceding inequality with (86) we obtain that for any $k \geq 0$,

$$\begin{aligned} \|h_{k+1} - y\|^2 \leq \|h_k - y\|^2 &- (1 - 6\beta_k^2 C_a(\beta_k, \alpha)^2)\|h_k - u_k\|^2 - 2\gamma_k \langle F(u_k), u_k - y\rangle \\ &- 2\gamma_k \langle e_k^2, u_k - y\rangle + 6\beta_k^2 C_e(\beta_k, \alpha)^2 \|e_k^1\|^{2\alpha} + 3\gamma_k^2(\|e_k^2\|^2 + \|e_k^1\|^2). \end{aligned} \tag{100}$$

Next, we plug $y = u^*$, where $u^* \in U^*$ is an arbitrary solution and use p-quasi sharpness of the operator F to obtain

$$\begin{aligned} \|h_{k+1} - u^*\|^2 \leq \|h_k - u^*\|^2 &- (1 - 6\beta_k^2 C_a(\beta_k, \alpha)^2)\|h_k - u_k\|^2 - 2\gamma_k \mu \text{dist}^p(u_k, U^*) \\ &- 2\gamma_k \langle e_k^2, u_k - u^*\rangle + 6\beta_k^2 C_e(\beta_k, \alpha)^2 \|e_k^1\|^{2\alpha} + 3\gamma_k^2(\|e_k^2\|^2 + \|e_k^1\|^2). \end{aligned} \tag{101}$$

By the stepsize choice $\gamma_k \leq \beta_k$, thus

$$\begin{aligned} \|h_{k+1} - u^*\|^2 \leq \|h_k - u^*\|^2 &- (1 - 6\beta_k^2 C_a(\beta_k, \alpha)^2)\|h_k - u_k\|^2 - 2\gamma_k \mu \text{dist}^p(u_k, U^*) \\ &- 2\gamma_k \langle e_k^2, u_k - u^*\rangle + 6\beta_k^2 C_e(\beta_k, \alpha)^2 \|e_k^1\|^{2\alpha} + 3\beta_k^2(\|e_k^2\|^2 + \|e_k^1\|^2). \end{aligned} \tag{102}$$

Since $\sum_{k=0}^\infty \beta_k^2 < \infty$, it follows that $\beta_k \to 0$. By definitions of $C_a(\beta_k, \alpha)$ and $C_e(\beta_k, \alpha)$ in (97) and (98), respectively, there exists $N \geq 0$ such that the stepsizes satisfy $1 - 6\beta_k^2 C_a(\beta_k, \alpha)^2 \geq \frac{1}{2}$ and $C_e(\beta_k, \alpha)^2 \leq \max\{K_1, \exp(L_1\beta_k)\} \leq \max\{K_1, 2\}$. Thus, the following inequality holds for any $k \geq N$,

$$\begin{aligned} \|h_{k+1} - u^*\|^2 \leq \|h_k - u^*\|^2 &- \frac{1}{2}\|h_k - u_k\|^2 - 2\gamma_k \mu \text{dist}^p(u_k, U^*) \\ &- 2\gamma_k \langle e_k^2, u_k - u^*\rangle + 3\beta_k^2(\|e_k^2\|^2 + \|e_k^1\|^2 + 2\max\{K_1, 2\}\|e_k^1\|^{2\alpha}). \end{aligned} \tag{103}$$

By rearranging the terms, defining $v_k = \|h_k - u^*\|^2 + \frac{1}{2}\|h_{k-1} - u_{k-1}\|^2 + 2\gamma_k \mu \text{dist}^p(u_k, U^*)$ and adding, and substituting $\frac{1}{2}\|h_{k-1} - u_{k-1}\|^2 + 2\gamma_{k-1}\mu \text{dist}^p(u_{k-1}, U^*)$ on the RHS we obtain

$$\begin{aligned} v_{k+1} \leq v_k &- \frac{1}{2}\|h_{k-1} - u_{k-1}\|^2 - 2\gamma_{k-1}\mu \text{dist}^p(u_{k-1}, U^*) \\ &- 2\gamma_k \langle e_k^2, u_k - u^*\rangle + 3\beta_k^2(\|e_k^2\|^2 + \|e_k^1\|^2 + 2\max\{K_1, 2\}\|e_k^1\|^{2\alpha}). \end{aligned} \tag{104}$$

Recalling that $e_k^1 = \Phi(h_k, \xi_k^1) - F(h_k)$, $e_k^2 = \Phi(u_k, \xi_k^2) - F(u_k)$ and using the stochastic properties of ξ_k^1, ξ_k^2 imposed by Assumption 2.5 and method's updates, we have

$$\mathbb{E}[\gamma_k \langle e_k^2, u_k - u^*\rangle \mid \mathcal{F}_{k-1}] = \mathbb{E}[\gamma_k \langle \mathbb{E}[e_k^2 \mid \mathcal{F}_{k-1} \cup \{\xi_k^1\}], u_k - u^*\rangle \mid \mathcal{F}_{k-1}] = 0,$$

since stepsizes γ_k is measurable in $\mathcal{F}_{k-1} \cup \{\xi_k^1\}$. Also, it holds that for all $k \geq 0$,

$$\mathbb{E}[\mathbb{E}[\|e_k^2\|^2 \mid \mathcal{F}_{k-1} \cup \{\xi_k^1\}] | \mathcal{F}_{k-1}] \leq \sigma^2, \quad \text{and} \quad \mathbb{E}[\|e_k^1\|^2 \mid \mathcal{F}_{k-1}] \leq \sigma^2.$$

Moreover, since $\alpha \leq 1$, the conditional expectation $\mathbb{E}[\|e_k^1\|^{2\alpha}|\mathcal{F}_{k-1}]$ is finite, and by Jensen inequality, it follows that for all $k \geq 0$,

$$\mathbb{E}[\|e_k^1\|^{2\alpha}|\mathcal{F}_{k-1}] \leq \sigma^{2\alpha}.$$

Therefore, by taking the conditional expectation on $\mathcal{F}_{k-1}$ in relation (104), we obtain for all $u^* \in U^*$ and for all $k \geq N$,

$$\begin{aligned}\mathbb{E}[v_{k+1} \mid \mathcal{F}_{k-1}] \leq v_k - \frac{1}{2}\|h_k - u_k\|^2 - 2\mu\gamma_{k-1}\mathrm{dist}^p(u_k, U^*) \\ + 6\beta_k^2(\sigma^2 + \max\{K_1, 2\}\sigma^{2\alpha}).\end{aligned} \tag{105}$$

By Lemma A.2, it follows that the sequence $\{\|h_k - u^*\|^2\}$ converges *a.s.* to a non-negative scalar for any $u^* \in U^*$, and almost surely we have

$$\sum_{k=0}^{\infty} \gamma_k \mathrm{dist}^p(u_k, U^*) < \infty, \quad \sum_{k=0}^{\infty} \|h_k - u_k\|^2 < \infty. \tag{106}$$

Since the sequence $\{\|h_k - u^*\|^2\}$ converges *a.s.* for all $u^* \in U^*$, it follows that the sequence $\{\|h_k - u^*\|\}$ is bounded *a.s.* for all $u^* \in U^*$. ■

C.2 Proof of Theoreom 4.2

Lemma C.2. *The stepsizes γ_k are given by (16) are nonsummable almost surely,*

$$\sum_{k=0}^{\infty} \gamma_k = \infty \quad \text{a.s.} \tag{107}$$

Proof. We will show that $\sum_{k=0}^{\infty} \beta_k \min\left\{1, \frac{1}{\|\Phi(h_k, \xi_k^1)\|}\right\} = \infty$ almost surely by the sequences of lower bound on this series. Consider the following event:

$$A_k = \{\|e_k^1\| \leq 2\sigma\},$$

where $e_k^1 = \Phi(h_k, \xi_k^1) - F(h_k)$ is a stochastic error from the sample for the clipping stepsize γ_k. Define $x_k = \min\left\{1, \frac{1}{\|\Phi(h_k, \xi_k^1)\|}\right\}$, then,

$$x_k = x_k \mathbb{I}(A_k) + x_k \mathbb{I}(\overline{A}_k) \geq x_k \mathbb{I}(A_k), \tag{108}$$

where the random variable $\mathbb{I}(A_k) =$ is an indicator function $\mathbb{I}(A_k) = \begin{cases} 1, & \omega \in A_k, \\ 0, & \omega \notin A_k \end{cases}$.

By the definition of x_k, the triangle inequality and definition of $\mathbb{I}(A_k)$, we have

$$\begin{aligned} x_k \mathbb{I}(A_k) &= \min\left\{1, \frac{1}{\|\Phi(h_k, \xi_k^1)\|}\right\} \mathbb{I}(A_k) \\ &\geq \min\left\{1, \frac{1}{\|F(h_k)\| + \|e_k^1\|}\right\} \mathbb{I}(A_k) \\ &\geq \min\left\{1, \frac{1}{\|F(h_k)\| + 2\sigma}\right\} \mathbb{I}(A_k). \end{aligned} \tag{109}$$

Adding and subtracting $\mathbb{E}[\mathbb{I}(A_k)|\mathcal{F}_{k-1}]$ and combining (108), (109) we have the lowerbound

$$\begin{aligned} \sum_{k=0}^{\infty} \beta_k x_k \geq \sum_{k=0}^{\infty} \beta_k \min\left\{1, \frac{1}{\|F(h_k)\| + 2\sigma}\right\} \mathbb{E}[\mathbb{I}(A_k)|\mathcal{F}_{k-1}] \\ + \sum_{k=0}^{\infty} \beta_k \min\left\{1, \frac{1}{\|F(h_k)\| + 2\sigma}\right\} (\mathbb{I}(A_k) - \mathbb{E}[\mathbb{I}(A_k)|\mathcal{F}_{k-1}]). \end{aligned} \tag{110}$$

To bound $p_k := \mathbb{E}[\mathbb{I}(A_k)|\mathcal{F}_{k-1}] = \mathbb{P}(A_k \mid \mathcal{F}_{k-1})$ we provide an upperbound on $\mathbb{P}(\overline{A}_k \mid \mathcal{F}_{k-1})$ using Markov's inequality and Assumption 2.5:

$$\mathbb{P}(\overline{A}_k \mid \mathcal{F}_{k-1}) = \mathbb{P}(\|e_k^1\| > 2\mathbb{E}[\|e_k^1\| \mid \mathcal{F}_{k-1}]\}) \leq \frac{\mathbb{E}[\|e_k^1\| \mid \mathcal{F}_{k-1}]}{2\mathbb{E}[\|e_k^1\| \mid \mathcal{F}_{k-1}])} = \frac{1}{2}. \tag{111}$$

This implies $\mathbb{E}[\mathbb{I}(A_k)|\mathcal{F}_{k-1}] \geq \frac{1}{2}$. Define $S_n = \sum_{k=0}^{n} \beta_k(\mathbb{I}(A_k) - \mathbb{E}[\mathbb{I}(A_k)|\mathcal{F}_{k-1}])$, by construction, S_n is a martingale. We want to show that $\lim_{n\to\infty} S_n \to S < \infty$ almost surely. We provide an upper bound for $\mathbb{E}[S_n^2]$:

$$\mathbb{E}[S_n^2] = \sum_{k=0}^{n} \beta_k^2 \mathbb{E}[(\mathbb{I}(A_k) - p_k)^2] + 2 \sum_{0\leq k<i\leq n} \beta_k^2 \mathbb{E}[(\mathbb{I}(A_k) - p_k)(\mathbb{I}(A_i) - p_i)] \tag{112}$$

By the law of total expectation, for all $0 \leq k < i \leq n$,

$$\mathbb{E}[(\mathbb{I}(A_k) - p_k)(\mathbb{I}(A_i) - p_i)] = \mathbb{E}[(\mathbb{I}(A_k) - p_k)\mathbb{E}[(\mathbb{I}(A_i) - p_i) \mid \mathcal{F}_{i-1}]] = 0. \tag{113}$$

Observe that $\mathbb{E}[\mathbb{I}(A_k) - p_k \mid \mathcal{F}_{k-1}] = 0$, then $\mathbb{E}[(\mathbb{I}(A_k) - p_k)^2 \mid \mathcal{F}_{k-1}] = Var(I(A_k) \mid \mathcal{F}_{k-1})$. A R.V. $\mathbb{I}(A_k)$ is a Bernoulli given $\mathcal{F}_{k-1}$ with mean p_k, then

$$\mathbb{E}[(\mathbb{I}(A_k) - p_k)^2 \mid \mathcal{F}_{k-1}] = \mathrm{Var}(\mathbb{I}(\mathrm{A_k}) \mid \mathcal{F}_{\mathrm{k}-1}) \leq \frac{1}{4}$$

By taking the total expectation we get $\mathbb{E}[(\mathbb{I}(A_k) - p_k)^2] \leq \frac{1}{4}$, and combining the previous results we have,

$$\mathbb{E}[S_n^2] \leq \frac{1}{4} \sum_{k=0}^{n} \beta_k^2 \leq \infty \quad \text{a.s.}$$

From Theorem 4.4.6. in Durrett (2019) it follows that S_n converges to $S < \infty$ almost surely (and in L_2).

To further lower bound $x_k\mathbb{I}(A_k)$ we show *a.s.* boundedness of $\|F(h_k)\|$ for all $k \geq 0$, using property of α-symmetric operators. To estimate $\|F(h_k)\|$, we add and subtract $F(v^*)$, where $v^* \in U^*$ is an arbitrary but fixed solution, and get

$$\|F(h_k)\| = \|F(h_k) - F(v^*) + F(v^*)\| \leq \|F(h_k) - F(v^*)\| + \|F(v^*)\|.$$

Define the following event:

$$A = \{\omega \in \Omega : \exists\, C(\omega) \in \mathbb{R} \text{ s.t.} \|h_k(\omega) - v^*\| < C(\omega)\ \forall\, k \geq 0\}.$$

Based on Lemma 4.1, the sequence $\{\|h_k - v^*\|\}$ is bounded *a.s.*, and thus $\mathbb{P}(A) = 1$. Let $\omega \in A$, now we can estimate $\|F(h_k(\omega))\|$ using the α-symmetric assumption on the operator.
Case $\alpha \in (0, 1)$.

$$\|F(h_k(\omega)) - F(u^*)\| \leq \|h_k(\omega) - v^*\|(K_0 + K_1\|F(v^*)\|^\alpha + K_2\|h_k(\omega) - v^*\|^{\alpha/(1-\alpha)}). \tag{114}$$

Since $\omega \in A$, it follows that $\|h_k(\omega) - v^*\| \leq C(\omega)$ for all $k \geq 0$. Using this fact and (114) we obtain that for all $k \geq 0$,

$$\|F(h_k(\omega))\| \leq C(\omega)(K_0 + K_1\|F(v^*)\|^\alpha + K_2 C(\omega)^{\alpha/(1-\alpha)}) + \|F(v^*)\|. \tag{115}$$

Therefore, the sequence $\{\|F(h_k(\omega))\|\}$ is upper bounded by $C_1(\omega) = C(\omega)(K_0 + K_1\|F(v^*)\|^\alpha + K_2 C(\omega)^{\alpha/(1-\alpha)}) + \|F(v^*)\|$.
Case $\alpha = 1$.
For $\alpha = 1$ by Proposition 2.2 we have

$$\|F(h_k(\omega)) - F(v^*)\| \leq \|h_k(\omega) - v^*\|(L_0 + L_1\|F(v^*)\|)\exp(L_1\|h_k(\omega) - v^*\|). \tag{116}$$

Therefore, for all $k \geq 0$,

$$\begin{aligned}\|F(h_k(\omega))\| &\leq \|F(h_k(\omega)) - F(v^*)\| + \|F(v^*)\| \\ &\leq \|h_k(\omega) - v^*\|(L_0 + L_1\|F(v^*)\|)\exp(L_1\|h_k(\omega) - v^*\|) + \|F(v^*)\|. \end{aligned} \tag{117}$$

Since $\omega \in A$, we have $\|h_k(\omega) - v^*\| \leq C(\omega)$ for all $k \geq 0$, which when used in (117), implies that for all $k \geq 0$,

$$\begin{aligned}\|F(h_k(\omega))\| &\leq \|h_k(\omega) - v^*\|(L_0 + L_1\|F(v^*)\|)\exp(L_1\|h_k(\omega) - v^*\|) + \|F(v^*)\| \\ &\leq C(\omega)(L_0 + L_1\|F(v^*)\|)\exp(L_1 C(\omega)) + \|F(v^*)\|. \end{aligned} \tag{118}$$

Hence, the sequence $\{\|F(h_k(\omega))\|\}$ is upper bounded by $\overline{C}_1(\omega)$, where $\overline{C}_1(\omega) = C(\omega)(L_0 + L_1\|F(v^*)\|)\exp(L_1 C(\omega)) + \|F(v^*)\|$. Now, for both cases $\alpha \in (0,1)$ and $\alpha = 1$ in (115) and (118), respectively, we have that $\|F(h_k(\omega))\|$ is upper bounded by $\max\{C_1(\omega), \overline{C}_1(\omega)\}$. Thus

$$\mathbb{P}(F(h_k) \text{ is bounded}) = 1.$$

Then almost surely we have (i) $F(h_k)$ is bounded (ii) $\sum_k^\infty \beta_k(\mathbb{I}(A_k) - \mathbb{E}[\mathbb{I}(A_k) \mid \mathcal{F}_{k-1}])$ converges to $S < \infty$ (iii) $\mathbb{E}[\mathbb{I}(A_k)|\mathcal{F}_{k-1}] \geq \frac{1}{2}$. Consider $\omega \in \Omega$ such that (i), (ii), (iii) holds, then in a view of (110) we have

$$\begin{aligned}\sum_{k=0}^{\infty} \beta_k x_k(\omega) &\geq \frac{1}{2}\min\left\{1, \frac{1}{\overline{C}_1(\omega) + 2\sigma}\right\}\sum_{k=0}^{\infty}\beta_k \\ &\quad + \min\left\{1, \frac{1}{\overline{C}_1(\omega) + 2\sigma}\right\} S(\omega) \\ &= \infty \end{aligned} \tag{119}$$

where the last bound comes from $\sum_{k=0}^\infty \beta_k = \infty$, which concludes the proof. ■

Proof of Theorem 4.2

Proof. By Lemma 4.1, we almost surely have

$$\sum_{k=0}^{\infty} \gamma_k \text{dist}^p(u_k, U^*) < \infty. \tag{120}$$

By Lemma C.2, we have $\sum_{k=0}^\infty \gamma_k = \infty$ almost surely, than from (120) it follows that

$$\liminf_{k\to\infty} \text{dist}^p(u_k, U^*) = 0 \quad a.s. \tag{121}$$

By Lemma 4.1, the sequence $\{\|h_k - u^*\|\}$ converges *a.s.* for any given $u^* \in U^*$. Thus, the sequence $\{h_k\}$ is bounded *a.s.* and, consequently, it has accumulation points *a.s.* In view of relation (20) in Lemma 4.1, it follows that

$$\lim_{k\to\infty} \|h_k - u_k\| = 0 \qquad \text{a.s.} \tag{122}$$

Therefore, the sequences $\{u_k\}$ and $\{h_k\}$ have the same accumulation points a.s.

Now, let $\{k_i \mid i \geq 1\}$ be a (random) index sequence such that

$$\lim_{i\to\infty} \text{dist}^p(u_{k_i}, U^*) = \liminf_{k\to\infty} \text{dist}^p(u_k, U^*) = 0 \quad a.s. \tag{123}$$

Without loss of generality, we may assume that $\{h_{k_i}\}$ is a convergent sequence (for otherwise, we will select a convergent subsequence), and let $\bar{u}$ be its (random) limit point, i.e.,

$$\lim_{i\to\infty} \|h_{k_i} - \bar{u}\| = 0 \qquad a.s. \tag{124}$$

By relation (20), it follows that $\lim_{k\to\infty}\|h_k - u_k\| = 0$ a.s., which in view of the preceding relation implies that

$$\lim_{i\to\infty}\|u_{k_i} - \bar{u}\| = 0 \qquad a.s.$$

By continuity of the distance function $\text{dist}(\cdot, U^*)$, from relation (123) we conclude that $\text{dist}(\bar{u}, U^*) = 0$ *a.s.*, which implies that $\bar{u} \in U^*$ almost surely since the set U^* is closed. Since the sequence $\{\|h_k - u^*\|^2\}$ converges *a.s.* for any $u^* \in U^*$, it follows that $\lim_{k\to\infty}\|h_k - \bar{u}\| = 0$ *a.s.* By relation (122) we conclude that $\lim_{k\to\infty}\|u_k - \bar{u}\| = 0$ a.s. ■

C.3 Proof of Lemma 4.3

Proof. The choice of parameters β_k, ensures that $1 - 6\beta_k^2(K_0 + K_1 + K_3)^2 \geq 1/2$. Then, by taking the expectation in (19) of Lemma 4.1 and using Assumption 2.5, and definition of $C_e(\beta_k, \alpha) = K_1$ for $\alpha \in (0, 1)$, we obtain

$$\begin{aligned}\mathbb{E}[\|h_{k+1} - u^*\|^2] \leq \mathbb{E}[\|h_k - u^*\|^2] - \frac{1}{2}\mathbb{E}[\|h_k - u_k\|^2] - 2\mathbb{E}[\gamma_k\mu\text{dist}^p(u_k, U^*)] \\ + 6\beta_k^2(\sigma^2 + K_1\sigma^{2\alpha}).\end{aligned} \tag{125}$$

The equation (125) satisfies the condition of Lemma A.3 with

$$\begin{aligned}\bar{v}_k = \mathbb{E}[\|u_k - u^*\|^2], \quad \bar{a}_k = 0, \quad \bar{b}_k = 6\beta_k^2(\sigma^2 + K_1\sigma^{2\alpha}), \\ \bar{z}_k = 2\mu\mathbb{E}[\gamma_k\,\text{dist}^p(u_k, U^*)] + \frac{1}{2}\mathbb{E}[\|h_k - u_k\|^2].\end{aligned} \tag{126}$$

By Lemma A.2, it follows that the sequence $\mathbb{E}[\|h_k - u^*\|^2]$ converges to a non-negative scalar for any $u^* \in U^*$. Since the sequence $\{\mathbb{E}[\|h_k - u^*\|^2]\}$ converges for all $u^* \in U^*$, it follows that the sequence $\{\mathbb{E}[\|h_k - u^*\|^2]\}$ is bounded for all $u^* \in U^*$. Next, using property of α-symmetric operators, we show that $\mathbb{E}[\|F(h_k)\|]$ is bounded for all $k \geq 0$. Let $v^* \in U^*$ be an arbitrary but fixed solution. Since $\alpha \leq 1/2$, it holds that

$$\begin{aligned}\|F(h_k)\| &\leq \|F(h_k) - F(v^*)\| + \|F(v^*)\| \\ &\leq \|h_k - v^*\|(K_0 + K_1\|F(v^*)\|^\alpha + K_2\|h_k - v^*\|^{\alpha/(1-\alpha)}) + \|F(v^*)\|.\end{aligned} \tag{127}$$

Taking the expectation, we obtain

$$\mathbb{E}[\|F(h_k)\|] \leq (K_0 + K_1\|F(v^*)\|^\alpha)\mathbb{E}[\|h_k - v^*\|] + K_2\mathbb{E}[\|h_k - v^*\|^{1+\alpha/(1-\alpha)})] + \|F(v^*)\|. \tag{128}$$

Notice that $\mathbb{E}[\|h_k - v^*\|^{1+\alpha/(1-\alpha)})] = \mathbb{E}[(\|h_k - v^*\|^2)^{1/2(1-\alpha)})]$ and, for $\alpha \leq 1/2$, the quantity $1/2(1-\alpha) \leq 1$. Thus, we can apply Jensen inequality for concave function

$$\mathbb{E}[(\|h_k - v^*\|^2)^{1/2(1-\alpha)})] \leq \mathbb{E}[\|h_k - v^*\|^2]^{1/2(1-\alpha)}.$$

Therefore, using the preceding relation and Jensen inequality for the first term on the RHS of equation (128), we obtain

$$\mathbb{E}[\|F(h_k)\|] \leq (K_0 + K_1\|F(v^*)\|^\alpha)\mathbb{E}[\|h_k - v^*\|^2]^{1/2} + K_2\mathbb{E}[\|h_k - v^*\|^2]^{1/2(1-\alpha)} + \|F(v^*)\|. \tag{129}$$

Since $\mathbb{E}[\|h_k - v^*\|^2]$ is bounded, it follows that $\mathbb{E}[\|F(h_k)\|]$ is bounded by some constant $C_F > 0$ for all $k \geq 0$. ■

C.4 Proof of Theorem 4.4

Proof. The choice of the parameters β_k ensures that $1 - 6\beta_k^2(K_0 + K_1 + K_2)^2 \geq \frac{1}{2}$, then by letting $u^* = P_{U^*}(h_k)$ in (103) in the proof of Lemma 4.1, with $C_e(\beta_k, \alpha) = K_1$, we get

$$\begin{aligned}\|h_{k+1} - P_{U^*}(h_k)\|^2 \leq \text{dist}^2(h_k, U^*) - \frac{1}{2}\|h_k - u_k\|^2 - 2\gamma_k\mu\text{dist}^p(u_k, U^*) \\ - 2\gamma_k\langle e_k^2, u_k - u^*\rangle + 3\beta_k^2(\|e_k^2\|^2 + \|e_k^1\|^2 + 2K_1\|e_k^1\|^{2\alpha}).\end{aligned} \tag{130}$$

By the definition of the distance function, we have

$$\mathrm{dist}^2(h_{k+1}, U^*) \leq \|h_{k+1} - P_{U^*}(h_k)\|^2.$$

Thus,

$$\begin{aligned}\mathrm{dist}^2(h_{k+1}, U^*) \leq \mathrm{dist}^2(h_k, U^*) - \frac{1}{2}\|h_k - u_k\|^2 - 2\gamma_k\mu\mathrm{dist}^p(u_k, U^*) \\ - 2\gamma_k\langle e_k^2, u_k - u^*\rangle + 3\beta_k^2(\|e_k^2\|^2 + \|e_k^1\|^2 + 2K_1\|e_k^1\|^{2\alpha}).\end{aligned} \tag{131}$$

Next, we estimate the term $\mathrm{dist}^p(u_k, U^*)$ in (131). By the triangle inequality, we have

$$\|h_k - u^*\| \leq \|u_k - h_k\| + \|u_k - u^*\| \qquad \text{for all } u^* \in U^*,$$

and by taking the minimum over $u^* \in U^*$ on both sides of the preceding relation, we obtain

$$\mathrm{dist}(h_k, U^*) \leq \|u_k - h_k\| + \mathrm{dist}(u_k, U^*). \tag{132}$$

Applying Lemma A.5 with $p > 0$ in equation (132) yields

$$\begin{aligned}\mathrm{dist}^p(h_k, U^*) &\leq (\|u_k - h_k\| + \mathrm{dist}(u_k, U^*))^p \\ &\leq 2^{p-1}\|u_k - h_k\|^p + 2^{p-1}\,\mathrm{dist}^p(u_k, U^*).\end{aligned} \tag{133}$$

Using projection inequality (27), and stepsizes choice (16), we obtain

$$\|u_k - h_k\| \leq \|\gamma_k\Phi(h_k, \xi_k^1)\| \leq 1.$$

Combining this result with equation (133), with $p \geq 2$, we get

$$\begin{aligned}\mathrm{dist}^p(h_k, U^*) &\leq 2^{p-1}\|u_k - h_k\|^{2+(p-2)} + 2^{p-1}\mathrm{dist}^p(u_k, U^*) \\ &\leq 2^{p-1}\|u_k - h_k\|^2 + 2^{p-1}\,\mathrm{dist}^p(u_k, U^*).\end{aligned} \tag{134}$$

By dividing the relation in (134) with 2^{p-1} and by rearranging the terms, we obtain the following relation

$$-\mathrm{dist}^p(u_k, U^*) \leq \|u_k - h_k\|^2 - 2^{1-p}\,\mathrm{dist}^p(h_k, U^*). \tag{135}$$

Combining the preceding inequality with (131), we find that for any $k \geq 0$,

$$\begin{aligned}\mathrm{dist}^2(h_{k+1}, U^*) \leq \mathrm{dist}^2(h_k, U^*) - 2^{2-p}\mu\gamma_k\mathrm{dist}^p(h_k, U^*) - \frac{1}{2}\|u_k - h_k\|^2 + 2\mu\gamma_k\,\|u_k - h_k\|^2 \\ - 2\gamma_k\langle e_k^2, u_k - u^*\rangle + 3\beta_k^2(\|e_k^2\|^2 + \|e_k^1\|^2 + 2K_1\|e_k^1\|^{2\alpha}).\end{aligned} \tag{136}$$

By the choice of β_k, we have $\beta_k = \dfrac{2}{a(\frac{2d}{a} + k)}$, where $a = \mu\min\left\{1, \frac{1}{2(C_F+\sigma)}\right\}$ and $d \geq 4\mu$. Thus, for all $k \geq 0$,

$$\beta_k \leq \frac{1}{d} \leq \frac{1}{4\mu} \qquad \Longrightarrow \qquad 2\mu\beta_k \leq \frac{1}{2}.$$

By the definition of the stepsize γ_k, we always have $\gamma_k \leq \beta_k$. Therefore, $2\mu\gamma_k \leq 2\mu\beta_k \leq \frac{1}{2}$ for all $k \geq 0$, thus implying that

$$-\frac{1}{2}\|u_k - h_k\|^2 + 2\mu\gamma_k\,\|u_k - h_k\|^2 \leq 0. \tag{137}$$

Using the stochastic properties of ξ_k imposed by Assumption 2.5, we have for all $k \geq 0$,

$$\begin{aligned}&\mathbb{E}[\mathbb{E}[\gamma_k\mathbb{E}[\langle e_k^2, u_k - u^*\rangle \mid \mathcal{F}_{k-1} \cup \{\xi_k^1\}] \mid \mathcal{F}_{k-1}]] = 0, \\ &\mathbb{E}[\mathbb{E}[\|e_k^2\|^2 \mid \mathcal{F}_{k-1} \cup \{\xi_k^1\}]] \leq \sigma^2, \qquad \mathbb{E}[\mathbb{E}[\|e_k^1\|^2 \mid \mathcal{F}_{k-1}]] \leq \sigma^2.\end{aligned} \tag{138}$$

Moreover, since $\alpha \leq 1$ then the conditional expectation $\mathbb{E}[\|e_k^1\|^{2\alpha}|\mathcal{F}_{k-1}]$ is defined, and by Jensen inequality $\mathbb{E}[\|e_k^1\|^{2\alpha}|\mathcal{F}_{k-1}] \leq \sigma^{2\alpha}$ for all $k \geq 0$. Thus, by taking the total expectation in relation (136) and using an estimate from (137), we obtain for all $u^* \in U^*$ and for all $k \geq 0$,

$$\mathbb{E}[\text{dist}^2(h_{k+1}, U^*)] \leq \mathbb{E}[\text{dist}^2(h_k, U^*)] - 2^{2-p}\mu\mathbb{E}[\gamma_k \text{dist}^p(h_k, U^*)] + 6\beta_k^2(\sigma^2 + K_1\sigma^{2\alpha}). \tag{139}$$

The equation (139) is similar to equation (69) in the proof of Theorem 3.4, with the same stepsize structure. Thus, by following the same arguments from equations (69) to equation (75) in the proof of Theorem 3.4, we arrive at

$$\begin{aligned}\mathbb{E}[\text{dist}^2(h_{k+1}, U^*)] \leq \mathbb{E}[\text{dist}^2(h_k, U^*)] - 2^{2-p}\mu\beta_k \min\left\{1, \frac{1}{2(C_F+\sigma)}\right\}\mathbb{E}[\text{dist}^p(h_k, U^*)] \\ + 6\beta_k^2(\sigma^2 + K_1\sigma^{2\alpha}),\end{aligned} \tag{140}$$

where C_F is an upperbound on $\mathbb{E}[\|F(h_k)\|]$ from the statement of Lemma 4.3. Now let $D_k = \mathbb{E}[\text{dist}^2(h_k, U^*)]$, and consider two cases $p = 2$ and $p > 2$.

Case $p = 2$. We note that by the definition of $a = \mu \min\left\{1, \frac{1}{2(C_F+\sigma)}\right\}$ and d, we have that $d \geq 4\mu$ and $\mu \geq a$, implying that $d \geq a$. Hence, for $p = 2$, relation (140) satisfies the conditions of Lemma A.6 with the following identification

$$r_k = D_k, \quad a = \mu \min\left\{1, \frac{1}{2(C_F+\sigma)}\right\}, \quad \alpha_k = \beta_k, \quad s_k = 0, \quad c = 6(\sigma^2 + K_1\sigma^{2\alpha}). \tag{141}$$

Therefore, for the choice $\beta_k = \dfrac{2}{a(\frac{2d}{a} + k)}$, we get the following convergence rate for all $k \geq 1$,

$$D_{k+1} \leq \frac{8d^2 D_0}{a^2k^2} + \frac{12(\sigma^2 + K_1\sigma^{2\alpha})}{a^2 k}. \tag{142}$$

Case $p \geq 2$. When $p \geq 2$, by applying telescoping sum to inequality (140) and rearranging the terms we obtain

$$\mathbb{E}[2^{2-p}a\sum_{t=0}^{k}\beta_k \text{dist}^p(h_k, U^*)] \leq D_0 - D_{k+1} + 6(\sigma^2 + K_1\sigma^{2\alpha})\sum_{t=0}^{k}\beta_k^2. \tag{143}$$

Since $p \geq 2$, the function $\text{dist}^p(\cdot, U^*)$ is convex, thus by defining $\bar{u}_k = (\sum_{t=0}^k \beta_k)^{-1}\sum_{t=0}^k \beta_k h_t$ and applying Jensen inequality be obtain

$$(\sum_{t=0}^{k}\beta_k)\mathbb{E}[\text{dist}^p(\bar{h}_k, U^*)] \leq \mathbb{E}[\sum_{t=0}^{k}\beta_k \text{dist}^p(h_k, U^*)].$$

Since $p \geq 2$, by applying Jensen inequality one more time, we obtain

$$(\bar{D}_k)^{p/2} = \left(\mathbb{E}[\text{dist}^2(\bar{h}_k, U^*)]\right)^{p/2} \leq \mathbb{E}\left[\left(\text{dist}^2(\bar{h}_k, U^*)\right)^{p/2}\right] = \mathbb{E}[\text{dist}^p(\bar{h}_k, U^*)].$$

$$(\bar{D}_k)^{p/2}\sum_{t=0}^{k}\beta_t \leq \sum_{t=0}^{k}\beta_t(D_t)^{p/2} \leq \frac{D_0 - D_{k+1} + 6(\sigma^2 + K_1\sigma^{2\alpha})\sum_{t=0}^k \beta_t^2}{2^{2-p}a}. \tag{144}$$

Now, we use the choice for β_k, i.e., $\beta_k = \frac{b}{(k+1)^q}$, where $0 < b < \frac{1}{2\sqrt{3}(K_0+K_1+K_2)}$ and $1/2 < q < 1$. Then, the sequence $\{\beta_k\}$ satisfies the conditions of Lemma 4.3. Furthermore, by Lemma A.7, we have that for all $k \geq 1$,

$$\sum_{t=0}^{k}\beta_t \geq \frac{b}{1-q}((k+1)^{1-q} - 2^{1-q}), \qquad \sum_{t=0}^{k}\beta_t^2 \leq \frac{b^2}{2q-1}. \tag{145}$$

Combining equations (144) and (145), and omitting D_{k+1}, we obtain for all $k \geq 1$,

$$(\bar{D}_k)^{p/2} \leq \frac{2^{p-2}(1-q)\left(D_0 + 6b^2(\sigma^2 + K_1\sigma^{2\alpha})(2\sigma^2+1)/(2q-1)\right)}{ab\left((k+1)^{1-q} - 2^{1-q}\right)}. \tag{146}$$

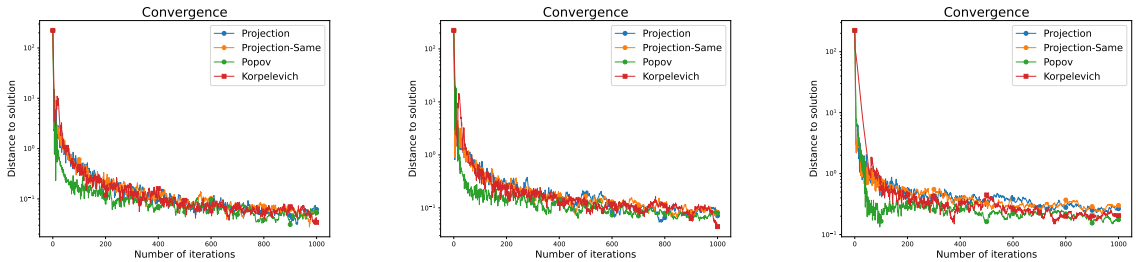
Raising both sides of the preceding inequality in power $2/p$, we have that for all $k \geq 1$,

$$\bar{D}_k \leq \frac{2^{2(p-2)/p}(1-q)^{2/p}\left(D_0 + 6b^2(\sigma^2 + K_1\sigma^{2\alpha})(2\sigma^2+1)/(2q-1)\right))^{2/p}}{(ab)^{2/p}\ \left((k+1)^{1-q} - 2^{1-q}\right)^{2/p}}. \tag{147}$$

■

D Additional Experiments

We investigate the robustness of the methods for a larger choice of the initial parameter value β_0. In Figure 5, we set $q = 1 - \epsilon$, and corresponding $\beta_k = \frac{50}{10+k^{1-\epsilon}}$, so the initial stepsize $\beta_0 \approx 5$ and run all four methods for the same problem parameter choice. We observe that for a generalized smooth SVI when we increase stepsizes, the performance of clipped stochastic Popov and Korpelevich is comparable to that of both clipped stochastic versions. While in smooth SVI, the stepsizes for stochastic Korplevich and Popov methods can be much larger than for stochastic projection methods, improving the convergence performance of stochastic Korplevich and Popov methods.



(a) $(\alpha \approx 0.09, p = 2.5)$ (b) $(\alpha \approx 0.5, p = 3.0)$ (c) $(\alpha \approx 0.8, p = 6.0)$

Figure 5: Comparison of the clipped stochastic projection, same-sample projection, Korpelevich, and Popov methods with $\beta = 50/(10 + k^{1-\epsilon})$.