MINI-BATCH FORWARD-BACKWARD-FORWARD METHODS FOR SOLVING STOCHASTIC VARIATIONAL INEQUALITIES

RADU IOAN BOŢ*, PANAYOTIS MERTIKOPOULOS°,♯, MATHIAS STAUDIGL°,
AND PHAN TU VUONG§

ABSTRACT. We develop a new stochastic algorithm for solving pseudo-monotone stochastic variational inequalities. Our method builds on Tseng's forward-backward-forward (FBF) algorithm, which is known in the deterministic literature to be a valuable alternative to Korpelevich's extragradient method when solving variational inequalities over a convex and closed set governed by pseudo-monotone, Lipschitz continuous operators. The main computational advantage of Tseng's algorithm is that it relies only on a single projection step and two independent queries of a stochastic oracle. Our algorithm incorporates a mini-batch sampling mechanism and leads to almost sure (a.s.) convergence to an optimal solution. To the best of our knowledge, this is the first stochastic look-ahead algorithm achieving this by using only a single projection at each iteration.

1. INTRODUCTION

In this paper we consider the following variational inequality problem, denoted as $\text{VI}(T, \mathcal{X})$, or simply VI: given a nonempty closed and convex set $\mathcal{X} \subseteq \mathbb{R}^d$ and a single valued map $T : \mathbb{R}^d \to \mathbb{R}^d$, find $x^* \in \mathcal{X}$ such that

$$\langle T(x^*), x - x^* \rangle \geq 0 \quad \text{for all } x \in \mathcal{X}. \tag{1.1}$$

We call $S(T, \mathcal{X}) \equiv \mathcal{X}_*$ the set of (Stampacchia) solutions of $\text{VI}(T, \mathcal{X})$. The variational inequality problem (1.1) arises in many interesting applications in economics, game theory and engineering Juditsky et al. (2011), Kannan and Shanbhag (2012), Mertikopoulos and Staudigl (2018), Ravat and Shanbhag (2011), Scutari et al. (2010), and includes as a special case first-order optimality conditions for nonlinear optimization, by choosing $T = \nabla f$ for some smooth function f. If $\mathcal{X}$ is unbounded, it can also be used to formulate complementarity problems, systems of equations, saddle point problems and many

*FACULTY OF MATHEMATICS, UNIVERSITY OF VIENNA, OSKAR-MORGENSTERN-PLATZ 1, A-1090 VIENNA, AUSTRIA.

°UNIV. GRENOBLE ALPES, CNRS, INRIA, GRENOBLE INP, LIG 38000 GRENOBLE, FRANCE.

♯CRITEO AI LAB

°MAASTRICHT UNIVERSITY, DEPARTMENT OF DATA SCIENCE AND KNOWLEDGE ENGINEERING, P.O. BOX 616, NL–6200 MD MAASTRICHT, THE NETHERLANDS.

§MATHEMATICAL SCIENCES, UNIVERSITY OF SOUTHAMPTON, HIGHFIELD SOUTHAMPTON SO17 1BJ, UNITED KINGDOM.

E-mail addresses: radu.bot@univie.ac.at, panayotis.mertikopoulos@imag.fr, m.staudigl@maastrichtuniversity.nl, T.V.phan@soton.ac.uk.

equilibrium problems. We refer the reader to Facchinei and Pang (2003) for an extensive review of applications in engineering and economics.

In many instances the problem VI arises as the expected value of an underlying stochastic optimization problem whose primitives are defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ carrying a random variable $\xi : (\Omega, \mathcal{F}) \to (\Xi, \mathcal{A})$ taking values in a measurable space $(\Xi, \mathcal{A})$ and inducing a law $\mathsf{P} = \mathbb{P} \circ \xi^{-1}$. Given the random element ξ, consider the measurable mapping $F : \mathcal{X} \times \Xi \to \mathbb{R}^d$, defining an integrable random vector $F(x, \xi) : \Omega \to \mathbb{R}^d$ via the composition $F(x, \xi)(\omega) = F(x, \xi(\omega))$. The stochastic variational inequality problem on which we will focus in this paper is denoted by SVI and defined as follows:

Definition 1.1. Let the operator $T : \mathbb{R}^d \to \mathbb{R}^d$ be defined by

$$T(x) := \mathbb{E}_\xi[F(x, \xi)] := \int_\Omega F(x, \xi(\omega)) \, d\mathbb{P}(\omega) = \int_\Xi F(x, z) \, d\mathsf{P}(z). \qquad (1.2)$$

Find $x^* \in \mathcal{X}$ satisfying (1.1).

This definition is known as the *expected value formulation of the stochastic variational inequality problem.* The expected value formulation goes back to the seminal work of King and Rockafellar (1993). By its very definition, if the operator T defined in (1.2) would be known, then the expected value formulation can be solved by any standard solution technique for deterministic variational inequalities. However, in practice, the operator T is usually not directly accessible, either due to excessive computations involved in performing the integral, or because T itself is the solution of an embedded subproblem. Hence, in most situations of interest, the solution of SVI relies on random samples of the operator $F(x, \xi)$. In this context, there are two current methodologies available; the sample average approximation (SAA) approach replaces the expected value formulation with an empirical estimator of the form

$$\hat{T}^N(x) = \frac{1}{N} \sum_{j=1}^N F(x, \xi_j),$$

and use the resulting deterministic map $\hat{T}^N$ as the input in one existing algorithm of choice. We refer to Shapiro et al. (2009) for this solution approach in connection with Monte Carlo simulation. We note that this approach is the standard choice in expected residual minimization problems, when P is unknown but accessible via a Monte Carlo approach.

A different methodology is the stochastic approximation (SA) approach, where samples are obtained in an online fashion, and key terms in a deterministic algorithm, such as gradients, are replaced by unbiased estimators by drawing a fresh random variable whenever needed.The mechanism to draw a fresh sample from P is usually named a stochastic oracle (SO), which report generates a stochastic error $F(x, \xi) - T(x)$.

Until very recently, the SA approach has only been used for the expected value formulation under very restrictive assumptions. To the best of our knowledge, the first formulation of an SA approach for a stochastic VI problem was made by Jiang and Xu (2008), under the assumption of strong monotonicity and continuity of the operator T. There, a proximal point algorithm of the form

$$X_{n+1} = \Pi_\mathcal{X}[X_n + \alpha_n F(X_n, \xi_n)] \qquad (1.3)$$

is considered, where $\Pi_{\mathcal{X}}$ denotes the Euclidean projection onto $\mathcal{X}$, $(\xi_n)_{n \geq 0}$ is a sample of P, and $(\alpha_n)_{n \geq 0}$ is a sequence of positive step sizes. Almost sure convergence of the iterates is proven for small step sizes, assuming T is Lipschitz continuous and strongly monotone, and the stochastic error is uniformly bounded in mean square. Relaxing strong monotonicity to plain monotonicity, the recent paper Yousefian et al. (2017) incorporated a Tikhonov regularization scheme into the stochastic approximation algorithm (1.3) and proved almost sure convergence of the generated stochastic process. The only established method guaranteeing almost sure convergence under the significantly weaker assumption of *pseudo-monotonicty* of the mean operator is the extragradient approach of Iusem et al. (2017) and Kannan and Shanbhag (2019). The original Korpelevich extragradient scheme of Korpelevich (1976) consists of two projection steps using two evaluations of the deterministic map T at generated test points y_n and x_n. Extending this to the stochastic oracle case, we arrive at the stochastic extra-gradient (SEG) method

$$
\begin{aligned}
Y_n &= \Pi_{\mathcal{X}}[X_n - \alpha_n A_{n+1}] \\
X_{n+1} &= \Pi_{\mathcal{X}}[X_n - \alpha_n B_{n+1}]
\end{aligned}
\tag{SEG}
$$

where $(A_n)_{n \geq 1}, (B_n)_{n \geq 1}$ are stochastic estimators of $T(X_n)$, and $T(Y_n)$, respectively. The paper Iusem et al. (2017) constructs these estimators by relying on a dynamic sampling strategy, where noise reduction of the estimators is achieved via a *mini-batch sampling* of the stochastic operators $F(X_n, \xi)$ and $F(Y_n, \xi)$. Within this mini-batch formulation, almost sure convergence of the stochastic process $(X_n)_{n \in \mathbb{N}}$ to the solution set can be proven even with constant step size implementations of SEG. In addition, optimal convergence rates of $O(1/N)$ in terms of the mean squared residual of the VI are obtained.[1]

1.1. **Our Contribution.** We briefly summarize the main contributions of this work. The most costly part of SEG are the two separate projection steps performed at each single iteration of the method. We show in this paper that a stochastic version of Tseng's forward-backward-forward method Tseng (2000), which we call the stochastic forward-backward-forward (SFBF) algorithm, preserves the strong trajectory-based convergence results, while the saving of one projection step allows us to beat SEG significantly in terms of computational overhead and runtime. In terms of convergence properties the SFBF algorithm developed in this paper has the same good properties as SEG. However, SFBF is potentially more efficient than SEG in each iteration since it relies only on a single Euclidean projection step. The price to pay for this is that we obtain an *infeasible method* (as is typical for primal-dual schemes) with a lower computational complexity count at the positive side. With infeasibility of a method we mean that parts of the algorithm's outputs may not satisfy state-space constraints present in the underlying optimization problem. While feasibility is a big concern in many applications (in particular in engineering and economics, where such constraints may represent technological constraints), it is not really a big problem for our method. Our numerical scheme will always provide one sequence respecting state-space constraints, and we will show that this feasible "shadow sequence" is an equally good proposal for an approximate solution. We will make this somewhat loose statement precise in the paper. Additionally, the

---

[1]Iusem et al. (2017), as well as our working paper, called this sampling process a variance reduction strategy. We follow the suggestion of the Associate Editor, and do not use this potentially confusing terminology anymore, and simply use the term *mini-batch* instead.

theoretically allowed range for step sizes is by the constant factor $\sqrt{3}$ times larger than the theoretically allowed largest step size in SEG. This constant factor gain results in significant improvements in terms of the convergence speed. This will be illustrated with extensive numerical evidences reported in Section 6.

2. Preliminaries

2.1. **Notation.** For $x, y \in \mathbb{R}^d$, we denote by $\langle x, y \rangle$ the standard inner product, and by $\|x\| \equiv \|x\|_2 := \langle x, x \rangle^{\frac{1}{2}}$ the corresponding norm. For $p \in [1, \infty]$, the ℓ_p norm on $\mathbb{R}^d$ is defined for $x = (x_1, \cdot, x_p)$ as $\|x\|_p := (\sum_{i=1}^n |x_i|^p)^{\frac{1}{p}}$. For a nonempty, closed and convex set $E \subseteq \mathbb{R}^d$, the Euclidean projector is defined as $\Pi_E(x) := \operatorname{argmin}_{y \in E} \|y - x\|$ for $x \in \mathbb{R}^d$. All random elements are defined on a given probability space $(\Omega, \mathcal{F}, \mathbb{P})$. An E-valued random variable is a $(\mathcal{F}, \mathcal{E})$-measurable mapping $f : \Omega \to E$; we write $f \in L^0(\Omega, \mathcal{F}, \mathbb{P}; E)$. For every $p \in [1, \infty]$, define the equivalence class of random variables $f \in L^0(\Omega, \mathcal{F}, \mathbb{P}; E)$ with $\mathbb{E}(\|f\|^p)^{1/p} < \infty$ as $L^p(\Omega, \mathcal{F}, \mathbb{P}; E)$. If $\mathcal{G} \subseteq \mathcal{F}$, the conditional expectation of the random variable $f \in L^p(\Omega, \mathcal{F}, \mathbb{P}; E)$ is denoted by $\mathbb{E}[f|\mathcal{G}]$. For $f_1, \ldots, f_k \in L^p(\Omega, \mathcal{F}, \mathbb{P}; E)$ we denote the sigma-algebra generated by these random variables by $\sigma(f_1, \ldots, f_k)$, this is the smallest sigma-algebra measuring the random variables $f_1, \ldots, f_k$. Let $(\Omega, \mathcal{F}, \mathbb{F} = (\mathcal{F}_n)_{n \geq 0}, \mathbb{P})$ be a complete stochastic basis. We denote by $\ell^0(\mathbb{F})$ the set of random sequences $(\xi_n)_{n \geq 1}$ such for each $n \in \mathbb{N}$, $\xi_n \in L^0(\Omega, \mathcal{F}_n, \mathbb{P}; \mathbb{R})$. For $p \in [1, \infty]$, we set

$$\ell^p(\mathbb{F}) \triangleq \{(\xi_n)_{n \geq 1} \in \ell^0(\mathbb{F}) | \sum_{n \geq 1} |\xi_n|^p < \infty \quad \mathbb{P}\text{-a.s.}\}.$$

The following properties of the Euclidean projection onto a closed convex set are well known.

Lemma 2.1. *Let $K \subseteq \mathbb{R}^d$ be a nonempty, closed and convex set. Then:*
- *(i) $\Pi_K(x)$ is the unique point of K satisfying $\langle x - \Pi_K(x), y - \Pi_K(x) \rangle \leq 0$ for all $y \in K$;*
- *(ii) for all $x \in \mathbb{R}^d$ and $y \in K$, we have $\|\Pi_K(x) - y\|^2 + \|\Pi_K(x) - x\|^2 \leq \|x - y\|^2$;*
- *(iii) for all $x, y \in \mathbb{R}^d$, $\|\Pi_K(x) - \Pi_K(y)\| \leq \|x - y\|$;*
- *(iv) given $\alpha > 0$ and $T : K \to \mathbb{R}^d$, the set of solutions of the variational problem $\mathrm{VI}(T, K)$ can be expressed as $S(T, K) = \{x \in \mathbb{R}^d | x = \Pi_K(x - \alpha T(x))\}$.*

Remark 2.1. In the literature on variational inequalities, there exists an alternative solution concept known as *weak*, or *Minty*, solutions. In this paper we are only interested in *strong*, or *Stampacchia*, solutions of $\mathrm{VI}(T, K)$, defined by inequality (1.1). For the problems of interest in this paper, Minty and Stampacchia solutions coincide Cottle and Yao (1992).

Another useful fact we use in this paper is the following elementary identity.

Lemma 2.2 (Pythagorean identity). *For all $x, x_n, x_{n+1} \in \mathbb{R}^d$ we have*

$$\|x_{n+1} - x\|^2 + \|x_{n+1} - x_n\|^2 - \|x_n - x\|^2 = 2\langle x_{n+1} - x_n, x_{n+1} - x \rangle.$$

2.2. **Probabilistic Tools.** We recall the Minkowski inequality: for $f, g \in L^p(\Omega, \mathcal{F}, \mathbb{P}; E), \mathcal{G} \subseteq \mathcal{F}$ and $p \in [1, \infty]$, we have

$$\mathbb{E}[\|f + g\|^p |\mathcal{G}]^{1/p} \leq \mathbb{E}[\|f\|^p |\mathcal{G}]^{1/p} + \mathbb{E}[\|g\|^p |\mathcal{G}]^{1/p}. \tag{2.1}$$

For the convergence analysis we will make use of the following classical lemma (Polyak 1987, Lemma 11, page 50).

Lemma 2.3 (Robbins-Siegmund). *Let $(\Omega, \mathcal{F}, \mathbb{F} = (\mathcal{F}_n)_{n \geq 0}, \mathbb{P})$ be a discrete stochastic basis. Let $(v_n)_{n \geq 1}, (u_n)_{n \geq 1} \in \ell_+^0(\mathbb{F})$ and $(\theta_n)_{n \geq 1}, (\beta_n)_{n \geq 1} \in \ell_+^1(\mathbb{F})$ be such that for all $n \geq 0$*

$$\mathbb{E}[v_{n+1}|\mathcal{F}_n] \leq (1 + \theta_n)v_n - u_n + \beta_n \qquad \mathbb{P} - a.s. \ .$$

Then $(v_n)_{n \geq 0}$ converges a.s. to a random variable v, and $(u_n)_{n \geq 1} \in \ell_+^1(\mathbb{F})$.

Finally, we need the celebrated Burkholder-Davis-Gundy inequality (see e.g. Stroock (2011)).

Lemma 2.4. *Let $(\Omega, \mathcal{F}, (\mathcal{F}_n)_{n \geq 0}, \mathbb{P})$ be a discrete stochastic basis and $(U_n)_{n \geq 0}$ a vector-valued martingale relative to this basis. Then, for all $p \in [1, \infty)$, there exists a universal constant $C_p > 0$ such that for every $N \geq 1$*

$$\mathbb{E}\left[\left(\sup_{0 \leq i \leq N} \|U_i\|\right)^p\right]^{1/p} \leq C_p \mathbb{E}\left[\left(\sum_{i=1}^N \|U_i - U_{i-1}\|^2\right)^{p/2}\right]^{\frac{1}{p}}.$$

When combined with Minkowski inequality, we obtain for all $p \geq 2$ a constant $C_p > 0$ such that for every $N \geq 1$

$$\mathbb{E}\left[\left(\sup_{0 \leq i \leq N} \|U_i\|\right)^p\right]^{1/p} \leq C_p \sqrt{\sum_{i=1}^N \mathbb{E}\left(\|U_i - U_{i-1}\|^p\right)^{2/p}}$$

3. THE STOCHASTIC FORWARD-BACKWARD-FORWARD ALGORITHM

In this paper we study a forward-backward-forward algorithm of Tseng type under weak monotonicity assumptions. The blanket hypotheses we consider throughout our analysis are summarized here:

Assumption 1 (Consistency). The solution set $\mathcal{X}_* \equiv S(T, \mathcal{X})$ is non-emtpy.

Assumption 2 (Stochastic Model). The set $\mathcal{X} \subseteq \mathbb{R}^d$ is nonempty, closed and convex, $(\Xi, \mathcal{A})$ is a measurable space and $F : \mathbb{R}^d \times \Xi \to \mathbb{R}^d$ is a Carathéodory map.[2]

Assumption 3 (Lipschitz continuity). The averaged operator $T(\cdot) = \mathbb{E}_\xi[F(\cdot, \xi)] : \mathbb{R}^d \to \mathbb{R}^d$ is Lipschitz continuous with modulus $L > 0$.

Assumption 4 (Pseudo-Monotonicity). The averaged operator $T(\cdot) = \mathbb{E}_\xi[F(\cdot, \xi)]$ is pseudo-monotone on $\mathbb{R}^d$, which means

$$\forall x, y \in \mathbb{R}^d : \langle T(x), y - x \rangle \geq 0 \Rightarrow \langle T(y), y - x \rangle \geq 0.$$

At each iteration, the decision maker has access to a stochastic oracle, reporting an approximation of $T(x)$ of the form

$$\hat{T}_{n+1}(x, \xi_{n+1}) \triangleq \frac{1}{m_{n+1}} \sum_{i=1}^{m_{n+1}} F(x, \xi_{n+1}^{(i)}) \text{ for } x \in \mathbb{R}^d. \tag{3.1}$$

---

[2]The mapping $x \mapsto F(x, \xi)$ is continuous for a. e. $\xi \in \Xi$, and $\xi \mapsto F(x, \xi)$ is measurable for all $x \in \mathbb{R}^d$; ξ is a random variable with values in Ξ, defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

The sequence $(m_n)_{n\geq 1} \subseteq \mathbb{N}$ determines the batch size of the stochastic oracle. The random sequence $\xi_n = (\xi_n^{(1)}, \ldots, \xi_n^{(m_n)})$ is an i.i.d draw from P. Approximations of the form (3.1) are very common in Monte-Carlo simulation approaches, machine learning and computational statistics (see e.g. Atchadé et al. (2017), Bottou et al. (2018), and references therein); they are easy to obtain in case we are able to sample from the measure P. The forward-backward-forward algorithm requires two queries from the stochastic oracle in which mini-batch estimators of the averaged map T are revealed. This dynamic sampling strategy requires a sequence of integers $(m_n)_{n\geq 1}$ (the *batch size*) determining the size of the data set to be processed at each iteration. The random sample on each mini-batch consists of two independent stochastic processes ξ_n and η_n drawn from the law P, and explicitly given by

$$\xi_n \triangleq (\xi_n^{(1)}, \ldots, \xi_n^{(m_n)}) \text{ and } \eta_n \triangleq (\eta_n^{(1)}, \ldots, \eta_n^{(m_n)}) \qquad \forall n \geq 1.$$

Given the current position X_n, Algorithm SFBF queries the SO once, to obtain the estimator $A_{n+1} \triangleq \hat{T}_{n+1}(X_n, \xi_{n+1})$, and then constructs the random variable $Y_n = \Pi_{\mathcal{X}}(X_n - \alpha_n A_{n+1})$. Next, a second query to SO is made to obtain the estimator $B_{n+1} \triangleq \hat{T}_{n+1}(Y_n, \eta_{n+1})$, followed by the update $X_{n+1} = Y_n + \alpha_n(A_{n+1} - B_{n+1})$. The pseudocode for SFBF is given in Algorithm 1.

Algorithm 1 Stochastic forward-backward-forward (SFBF)

Require: step-size sequence α_n; batch size sequence m_n

1: Initialize X # initialization
2: **for** $n = 1, 2, \ldots$ **do**
3:     Draw samples ξ^i and η^i from P $(i = 1, \ldots, m_n)$

4:     Oracle returns $A \leftarrow \dfrac{1}{m_n}\sum_{i=1}^{m_n} F(X, \xi^i)$ # first oracle query

5:     Set $Y \leftarrow \Pi_{\mathcal{X}}(X - \alpha_n A)$ # forward-backward step

6:     Oracle returns $B \leftarrow \dfrac{1}{m_n}\sum_{i=1}^{m_n} F(Y, \eta^i)$ # second oracle query

7:     Set $X \leftarrow Y + \alpha_n(A - B)$ # second forward step
8: **end for**

Observe that Algorithm SFBF is an infeasible method: the iterates $(X_n)_{n\geq 0}$ are not necessarily elements of the admissible set $\mathcal{X}$, but the "shadow sequence" $(Y_n)_{n\geq 0}$ is by construction so. In the stochastic optimization case, i.e. for instances where A_{n+1} is an unbiased estimator of the gradient of a real-valued function, the process $(Y_n)_{n\geq 0}$ is seen to be a projected gradient step, where A_{n+1} acts as an unbiased estimator for the stochastic gradient. This gradient step is used in an extrapolation step to generate the iterate X_{n+1}. We just mention that related popular primal-dual splitting schemes like ADMM (Boyd et al. 2011, Chen et al. 2018) are infeasible by nature as well. In concrete applications, the infeasibility of Algorithm SFBF is not really a big problem. First, if feasibility is a strict requirement, we can always propose the shadow sequence $(Y_n)_{n\geq 0}$ as an approximate solution. This is justified by Proposition 4.7. Moreover, Theorem 4.5 shows that the random process $(X_n)_{n\geq 0}$ will converge to a solution almost

surely. Hence, SFBF is for sure "asymptotically feasible", and always contains feasible approximate solutions in case of early stopping.

Assumption 5 (Step-size choice). The step-size sequence $(\alpha_n)_{n\geq 0}$ in Algorithm SFBF satisfies

$$0 < \underline{\alpha} \triangleq \inf_{n\geq 0} \alpha_n \leq \bar{\alpha} \triangleq \sup_{n\geq 1} \alpha_n < \frac{1}{\sqrt{2}L}.$$

For $n \geq 0$, we introduce the *approximation error*

$$W_{n+1} \triangleq A_{n+1} - T(X_n), \text{ and } Z_{n+1} \triangleq B_{n+1} - T(Y_n), \tag{3.2}$$

and the sub-sigma algebras $(\mathcal{F}_n)_{n\geq 0}, (\hat{\mathcal{F}}_n)_{n\geq 0}$, defined by $\mathcal{F}_0 \triangleq \sigma(X_0)$, and $\mathcal{F}_n \triangleq \sigma(X_0, \xi_1, \xi_2, \ldots, \xi_n, \eta_1, \ldots, \eta_n)$ for all $n \geq 1$, and $\hat{\mathcal{F}}_n \triangleq \sigma(X_0, \xi_1, \ldots, \xi_n, \xi_{n+1}, \eta_1, \ldots, \eta_n)$ for all $n \geq 0$, respectively. Observe that $\mathcal{F}_n \subseteq \hat{\mathcal{F}}_n$ for all $n \geq 0$. We also define the filtrations $\mathbb{F} \triangleq (\mathcal{F}_n)_{n\geq 0}$ and $\hat{\mathbb{F}} \triangleq (\hat{\mathcal{F}}_n)_{n\geq 0}$. The introduction of these two different sub-sigma algebras is important for many reasons. First, observe that they embody the information the learner has about the optimization problem. Indeed, the sub-sigma algebra $(\mathcal{F}_n)_{n\geq 0}$ corresponds to the information the decision maker has at the beginning the n-th iteration, whereas $(\hat{\mathcal{F}}_n)_{n\geq 0}$ is the information the decision maker has after the first (projection)-step of the iteration. Therefore, $(Y_n)_{n\geq 0}$ is measurable with respect to the sub-sigma algebra $(\hat{\mathcal{F}}_n)_{n\geq 0}$ and $(X_n)_{n\geq 0}$ is measurable with respect to the sub-sigma algebra $(\mathcal{F}_n)_{n\geq 0}$. Second, we see that the process $(W_n)_{n\geq 1}$ is $\mathbb{F}$-adapted, whereas the process $(Z_n)_{n\geq 1}$ is $\hat{\mathbb{F}}$-adapted, unbiased approximations relative to the respective information structures are provided:

$$\mathbb{E}[W_{n+1}|\mathcal{F}_n] = 0 \text{ and } \mathbb{E}[Z_{n+1}|\hat{\mathcal{F}}_n] = 0 \qquad \forall n \geq 0.$$

Assumption 6 (Batch Size). The batch size sequence $(m_n)_{n\geq 1}$ satisfies $\sum_{n=1}^{\infty} \frac{1}{m_n} < \infty$.

A sufficient condition on the sequence $(m_n)_{n\geq 1}$ is that for some constant $\mathsf{c} > 0$ and integer $n_0 > 0$, we have

$$m_n = \mathsf{c} \cdot (n + n_0)^{1+a} \ln(n + n_0)^{1+b} \tag{3.3}$$

for $a > 0$ and $b \geq -1$, or $a = 0$ and $b > 0$.

The next assumption is essentially the same as the variance control assumption in Iusem et al. (2017).

Assumption 7 (Variance Control). For all $x \in \mathbb{R}^d$ and $p \geq 1$, let

$$s_p(x) \triangleq \mathbb{E}_\xi[\|F(x, \xi) - T(x)\|^p]^{1/p}.$$

There exist $p \geq 2, \sigma_0 \geq 0$, and a measurable locally bounded function $\sigma : \mathcal{X}_* \to \mathbb{R}_+$ such that for all $x \in \mathbb{R}^d$ and all $x^* \in \mathcal{X}_*$

$$s_p(x) \leq \sigma(x^*) + \sigma_0\|x - x^*\|. \tag{3.4}$$

Before we proceed with the convergence analysis, we want to make some clarifying remarks on this assumption. The most frequently used assumption on the SO's approximation error, which dates back to the seminal work of Robbins and Monro (Duflo 1996, Kushner and Yin 1997), asks for uniformly bounded variance (UBV), i.e.

$$\sup_{x\in\mathcal{X}} s_2(x) \leq \sigma. \tag{UBV}$$

UBV is covered by Assumption 7 when $\sigma_0 = 0$ and $\sup_{x \in \mathcal{X}_*} \sigma(x^*) \leq \sigma$. UBV is for instance valid when additive noise with finite p-th moment is assumed, that is, for some random variable ξ with $\mathbb{E}[\|\xi\|^p]^{1/p} \leq \sigma < \infty$, we have

$$F(x, \xi) = T(x) + \xi \qquad \mathbb{P}\text{-a.s.}$$

However, assuming a global variance bound is not realistic in cases where the variance of the stochastic oracle depends on the position x (see e.g. Example 1 in Jofré and Thompson (2018)). Assumption 7 is much weaker than UBV, as it exploits the local variance of the stochastic oracle rather than, potentially hard to estimate, global mean square variance bounds. The recent papers Iusem et al. (2017), Jofré and Thompson (2018) make similar assumptions on the variance of the stochastic oracle. It is shown there that Assumption (7) is most natural in cases where the feasible set $\mathcal{X}$ is unbounded, and it is always satisfied when the Carathéodory functions $F(\cdot, \xi)$ are random Lipschitz, as illustrated with the example below.

Example 3.1. Assume for the Carathéodory map $F : \mathbb{R}^d \times \Xi \to \mathbb{R}^d$ that there exists $\mathcal{L} \in L^p(\Omega, \mathcal{F}, \mathbb{P}; \mathbb{R}_+)$ with

$$\|F(x, \xi) - F(y, \xi)\| \leq \mathcal{L}(\xi)\|x - y\| \qquad \forall x, y \in \mathbb{R}^d.$$

Call L the Lipschitz constant of the map $x \mapsto T(x) = \mathbb{E}_\xi[F(x, \xi)]$. Then, a repeated application of the Minkowski inequality, shows that for all $x \in \mathbb{R}^d$ and all $x^* \in \mathcal{X}_*$ we have

$$s_p(x) \leq \mathbb{E}_\xi[\|F(x, \xi) - F(x^*, \xi)\|^p]^{1/p} + s_p(x^*) + \|T(x) - T(x^*)\|$$
$$\leq (\mathbb{E}_\xi[\mathcal{L}(\xi)^p]^{1/p} + L)\|x - x^*\| + s_p(x^*).$$

Let $\sigma(x^*)$ denote a bound on $s_p(x^*)$ and set $\sigma_0 \triangleq L + \mathbb{E}_\xi[\mathcal{L}(\xi)^p]^{1/p}$, to get a variance bound as required in Assumption 7.

4. Convergence Analysis

We consider the quadratic residual function defined by

$$r_a(x)^2 \triangleq \|x - \Pi_{\mathcal{X}}(x - aT(x))\|^2 \ \forall x \in \mathbb{R}^d.$$

The reader familiar with the literature on finite-dimensional variational inequalities will recognize this immediately as the energy defined by the natural map $F_a^{\text{nat}}(x) \triangleq x - \Pi_{\mathcal{X}}(x - aT(x))$ (Facchinei and Pang 2003, chapter 10). It is well known that $r_a(x)$ is a merit function for $\text{VI}(T, \mathcal{X})$. Moreover, $\{r_a(x); a > 0\}$ is a family of equivalent merit functions for $\text{VI}(T, \mathcal{X})$, in the sense that $r_b(x) \geq r_a(x)$ for all $b > a > 0$ (Facchinei and Pang 2003, Proposition 10.3.6). Denote

$$\rho_n \triangleq 1 - 2L^2\alpha_n^2 \qquad \forall n \geq 0. \tag{4.1}$$

We define recursively the process $(V_n)_{n \geq 0}$ by $V_0 \triangleq 0$ and, for all $n \geq 1$,

$$V_{n+1} \triangleq V_n + (4 + \rho_n)\alpha_n^2\|W_{n+1}\|^2 + 4\alpha_n^2\|Z_{n+1}\|^2,$$

so that

$$\Delta V_n \triangleq V_{n+1} - V_n = (4 + \rho_n)\alpha_n^2\|W_{n+1}\|^2 + 4\alpha_n^2\|Z_{n+1}\|^2 \qquad \forall n \geq 0. \tag{4.2}$$

Additionally, we define for all $x \in \mathbb{R}^d$ the process $(U_n(x))_{n \geq 0}$ given by $U_0(x) \triangleq 0$, and

$$U_{n+1}(x) \triangleq U_n(x) + 2\alpha_n\langle Z_{n+1}, x - Y_n\rangle \qquad \forall n \geq 1$$

with corresponding increment

$$\Delta U_n(x) \triangleq 2\alpha_n \langle Z_{n+1}, x - Y_n \rangle \qquad \forall n \geq 0.$$

For any reference point $x \in \mathbb{R}^d$ we see that $\mathbb{E}[\Delta U_n(x)|\hat{\mathcal{F}}_n] = 0$ for all $n \geq 0$, i.e. the process $(U_n(x))_{n \geq 0}$ is a martingale with respect to the filtration $\hat{\mathbb{F}}$. Since $\mathcal{F}_n \subseteq \hat{\mathcal{F}}_n$, the tower property implies that

$$\mathbb{E}[\Delta U_n(x)|\mathcal{F}_n] = 0 \qquad \forall x \in \mathbb{R}^d \ \forall n \geq 0, \tag{4.3}$$

showing that it is also a $\mathbb{F}$-martingale. $(V_n)_{n \geq 0}$ is an increasing process, with increments ΔV_n whose expected value is determined by the variance of the approximation error of the stochastic oracle feedback. In terms of these increment processes, we establish the following fundamental recursion.

Lemma 4.1. *For all $x^* \in \mathcal{X}_*$ and all $n \geq 0$ we have*

$$\|X_{n+1} - x^*\|^2 \leq \|X_n - x^*\|^2 - \frac{\rho_n}{2} r_{\alpha_n}(X_n)^2 + \Delta U_n(x^*) + \Delta V_n \quad \mathbb{P} - a.s.. \tag{4.4}$$

Proof. Proof. This recursive relation follows via several simple algebraic steps. Let $x^* \in \mathcal{X}_*$ and $n \geq 0$ be fixed.

Step 1.. We have

$$\langle T(x^*), y - x^* \rangle \geq 0 \qquad \forall y \in \mathcal{X}.$$

Using that $\alpha_n > 0$ as well as the pseudo-monotonicity of T, we see

$$\langle \alpha_n T(Y_n), Y_n - x^* \rangle \geq 0.$$

Using the Doob decomposition in equation (3.2), we can rewrite this inequality as

$$\langle \alpha_n B_{n+1}, Y_n - x^* \rangle \geq \alpha_n \langle Z_{n+1}, Y_n - x^* \rangle. \tag{4.5}$$

Since $Y_n = \Pi_{\mathcal{X}}(X_n - \alpha_n A_{n+1})$, from Lemma 2.1(i) we conclude that

$$\langle x^* - Y_n, Y_n - X_n + \alpha_n A_{n+1} \rangle \geq 0. \tag{4.6}$$

Adding (4.5) and (4.6) gives

$$\langle \alpha_n(A_{n+1} - B_{n+1}) - X_n + Y_n, x^* - Y_n \rangle \geq \alpha_n \langle Z_{n+1}, Y_n - x^* \rangle,$$

which is equivalent to

$$\langle x^* - Y_n, X_{n+1} - X_n \rangle \geq \alpha_n \langle Z_{n+1}, Y_n - x^* \rangle. \tag{4.7}$$

Step 2.. Using (4.7), we get

$$\begin{aligned}
\langle X_{n+1} - X_n, X_{n+1} - x^* \rangle =& \langle X_{n+1} - X_n, Y_n - x^* \rangle + \langle X_{n+1} - X_n, X_{n+1} - Y_n \rangle \\
\leq& \langle \alpha_n Z_{n+1}, x^* - Y_n \rangle + \|X_{n+1} - X_n\|^2 \\
& + \langle X_{n+1} - X_n, X_n - Y_n \rangle \\
=& \langle \alpha_n Z_{n+1}, x^* - Y_n \rangle + \|X_{n+1} - X_n\|^2 - \|X_n - Y_n\|^2 \\
& + \alpha_n \langle A_{n+1} - B_{n+1}, X_n - Y_n \rangle
\end{aligned}$$

where we have used the definition of X_{n+1} in the last equality. The Pythagorean identity (Lemma 2.2) gives us

$$\begin{aligned}
\|X_{n+1} - x^*\|^2 =& \|X_n - x^*\|^2 - \|X_{n+1} - X_n\|^2 + 2\langle X_{n+1} - X_n, X_{n+1} - x^* \rangle \\
\leq& \|X_n - x^*\|^2 + \|X_{n+1} - X_n\|^2 - 2\|X_n - Y_n\|^2
\end{aligned}$$

$$+ 2\langle \alpha_n Z_{n+1}, x^* - Y_n \rangle + 2\alpha_n \langle A_{n+1} - B_{n+1}, X_n - Y_n \rangle.$$

<u>Step 3</u>. Using again the definition of X_{n+1}, we see

$$\begin{aligned}
\|X_{n+1} - X_n\|^2 =& \|Y_n + \alpha_n(A_{n+1} - B_{n+1}) - X_n\|^2 \\
=& \|X_n - Y_n\|^2 + \alpha_n^2 \|A_{n+1} - B_{n+1}\|^2 + 2\alpha_n \langle A_{n+1} - B_{n+1}, Y_n - X_n \rangle \\
\leq& \|X_n - Y_n\|^2 + 2\alpha_n^2 \|T(X_n) - T(Y_n)\|^2 + 2\alpha_n^2 \|W_{n+1} - Z_{n+1}\|^2 \\
&+ 2\alpha_n \langle A_{n+1} - B_{n+1}, Y_n - X_n \rangle \\
\leq& \|X_n - Y_n\|^2 + 2L^2\alpha_n^2 \|X_n - Y_n\|^2 + 4\alpha_n^2 \|W_{n+1}\|^2 + 4\alpha_n^2 \|Z_{n+1}\|^2 \\
&+ 2\alpha_n \langle A_{n+1} - B_{n+1}, Y_n - X_n \rangle.
\end{aligned}$$

The first inequality is the Cauchy-Schwarz inequality. The second inequality follows from the L-Lipschitz continuity of the averaged operator T (Assumption 3), and again the Cauchy-Schwarz inequality. Combining this with the last inequality obtained in Step 2, we see that

$$\begin{aligned}
\|X_{n+1} - x^*\|^2 \leq& \|X_n - x^*\|^2 - (1 - 2L^2\alpha_n^2)\|X_n - Y_n\|^2 + 4\alpha_n^2 \|W_{n+1}\|^2 \\
&+ 4\alpha_n^2 \|Z_{n+1}\|^2 + 2\langle \alpha_n Z_{n+1}, x^* - Y_n \rangle.
\end{aligned}$$

<u>Step 4.</u>. By the definition of the squared residual function, the definition of Y_n and Lemma 2.1(iii), we have

$$\begin{aligned}
r_{\alpha_n}(X_n)^2 =& \|X_n - \Pi_{\mathcal{X}}(X_n - \alpha_n T(X_n))\|^2 \\
\leq& 2\|X_n - Y_n\|^2 + 2\|Y_n - \Pi_{\mathcal{X}}(X_n - \alpha_n T(X_n))\|^2 \\
=& 2\|X_n - Y_n\|^2 + 2\|\Pi_{\mathcal{X}}(X_n - \alpha_n A_{n+1}) - \Pi_{\mathcal{X}}(X_n - \alpha_n T(X_n))\|^2 \\
\leq& 2\|X_n - Y_n\|^2 + 2\|\alpha_n W_{n+1}\|^2.
\end{aligned}$$

Hence,

$$-2\|X_n - Y_n\|^2 \leq 2\alpha_n^2 \|W_{n+1}\|^2 - r_{\alpha_n}(X_n)^2. \tag{4.8}$$

<u>Step 5.</u>. Combining (4.8) with the last inequality from Step 3 and recalling Assumption 5, we conclude

$$\begin{aligned}
\|X_{n+1} - x^*\|^2 \leq& \|X_n - x^*\|^2 - \frac{1}{2}(1 - 2L^2\alpha_n^2)r_{\alpha_n}(X_n)^2 + (1 - 2L^2\alpha_n^2)\alpha_n^2 \|W_{n+1}\|^2 \\
&+ 4\alpha_n^2 \|W_{n+1}\|^2 + 4\alpha_n \|Z_{n+1}\|^2 + 2\langle \alpha_n Z_{n+1}, x^* - Y_n \rangle \\
=& \|X_n - x^*\|^2 - \frac{\rho_n}{2}r_{\alpha_n}(X_n)^2 + (4 + \rho_n)(\alpha_n)^2 \|W_{n+1}\|^2 + 4\alpha_n^2 \|Z_{n+1}\|^2 \\
&+ 2\langle \alpha_n Z_{n+1}, x^* - Y_n \rangle.
\end{aligned}$$

The definitions of the increments associated with the martingale $(U_n(x^*))_{n \geq 0}$ and the non-decreasing process $(V_n)_{n \geq 0}$ give the claimed result. ∎

Remark 4.1. One can notice that in the above proof the pseudo-monotonicity of T is used only in Step 1, in order to obtain relation (4.5). Thus, as happened in Dang and Lan (2015), Solodov and Svaiter (1999), the pseudo-monotonicity of T can actually be replaced by the weaker assumption

$$\langle T(x), x - x^* \rangle \geq 0 \qquad \forall x \in \mathcal{X}, x^* \in \mathcal{X}_*.$$

See also Mertikopoulos and Zhou (2018) for a similar condition.

In the following, we let $p \geq 2$ be the exponent as specified in Assumption 7. Taking conditional expectations in equation (4.4) and using the martingale property (4.3), we see for all $n \geq 0$ that

$$\mathbb{E}[\|X_{n+1} - x^*\|^2 | \mathcal{F}_n] \leq \|X_n - x^*\|^2 - \frac{\rho_n}{2} r_{\alpha_n}(X_n)^2 + \mathbb{E}[\Delta V_n | \mathcal{F}_n]. \tag{4.9}$$

In order to prove convergence of the process $(X_n)_{n \geq 0}$, we aim to deduce a stochastic quasi-Fejér relation. For that, we need to understand the properties of the conditional expectation

$$\mathbb{E}[\Delta V_n | \mathcal{F}_n] = (4 + \rho_n) \alpha_n^2 \mathbb{E}[\|W_{n+1}\|^2 | \mathcal{F}_n] + 4\alpha_n^2 \mathbb{E}[\|Z_{n+1}\|^2 | \mathcal{F}_n] \ \forall n \geq 0.$$

Let be $q \in [1, \infty]$. The monotonicity of $L^q(\mathbb{P}) \triangleq L^q(\Omega, \mathcal{F}, \mathbb{P}; \mathbb{R})$ norms gives $\mathbb{E}[\Delta V_n | \mathcal{F}_n] \leq \mathbb{E}[|\Delta V_n|^q | \mathcal{F}_n]^{\frac{1}{q}}$ for all $n \geq 0$. By Minkowski inequality,

$$\mathbb{E}[|\Delta V_n|^q | \mathcal{F}_n]^{\frac{1}{q}} \leq (4 + \rho_n) \alpha_n^2 \mathbb{E}[\|W_{n+1}\|^{2q} | \mathcal{F}_n]^{1/q} + 4\alpha_n^2 \mathbb{E}[\|Z_{n+1}\|^{2q} | \mathcal{F}_n]^{1/q} \ \forall n \geq 0.$$

The next lemma provides the required bounds for these expressions, and also highlights the implicit variance reduction of our method.

Lemma 4.2. *Let be $p' \in [2, p]$. For all $n \geq 0$ we have $\mathbb{P}$-a.s.*

$$\mathbb{E}[\|W_{n+1}\|^{p'} | \mathcal{F}_n]^{\frac{1}{p'}} \leq \frac{C_{p'} (\sigma(x^*) + \sigma_0 \|X_n - x^*\|)}{\sqrt{m_{n+1}}} \tag{4.10}$$

and

$$\mathbb{E}[\|Z_{n+1}\|^{p'} | \mathcal{F}_n]^{\frac{1}{p'}} \leq \frac{C_{p'}}{\sqrt{m_{n+1}}} \left(\sigma(x^*) + \sigma_0 \mathbb{E}[\|Y_n - x^*\|^{p'} | \mathcal{F}_n]^{\frac{1}{p'}} \right). \tag{4.11}$$

In particular, in case of (UBV) with $\sigma_0 = 0$ and $\sup_{x \in \mathcal{X}_*} \sigma(x^*) \leq \hat{\sigma}$, both approximation errors are bounded in $L^{p'}(\mathbb{P})$ by the common factor $\frac{C_{p'} \hat{\sigma}}{\sqrt{m_{n+1}}}$.

Proof. Proof. See Appendix A of the Appendix. ∎

Let be $p' \geq 2$ and $n \geq 0$. We have

$$\mathbb{E}[\|Y_n - x^*\|^{p'} | \mathcal{F}_n]^{\frac{1}{p'}} \leq (1 + \alpha_n L) \|X_n - x^*\| + \alpha_n \mathbb{E}[\|W_{n+1}\|^{p'} | \mathcal{F}_n]^{1/p'}.$$

Hence, combining this with (4.10) for $p' \in [2, p]$ as in Lemma 4.2, we see that

$$\mathbb{E}[\|Y_n - x^*\|^{p'} | \mathcal{F}_n]^{\frac{1}{p'}} \leq (1 + \alpha_n L) \|X_n - x^*\| + \alpha_n \frac{C_{p'} (\sigma(x^*) + \sigma_0 \|X_n - x^*\|)}{\sqrt{m_{n+1}}}$$

$$= \left(1 + \alpha_n L + \alpha_n \frac{C_{p'} \sigma_0}{\sqrt{m_{n+1}}} \right) \|X_n - x^*\| + \alpha_n \frac{C_{p'} \sigma(x^*)}{\sqrt{m_{n+1}}}. \tag{4.12}$$

Plugging this inequality into (4.11), after rearranging the terms we see that

$$\mathbb{E}[\|Z_{n+1}\|^{p'} | \mathcal{F}_n]^{\frac{1}{p'}} \leq \frac{C_{p'} \sigma(x^*)}{\sqrt{m_{n+1}}} \left(1 + \alpha_n \frac{\sigma_0 C_{p'}}{\sqrt{m_{n+1}}} \right)$$

$$+ \|X_n - x^*\| \frac{C_{p'} \sigma_0}{\sqrt{m_{n+1}}} \left(1 + \alpha_n L + \alpha_n \frac{C_{p'} \sigma_0}{\sqrt{m_{n+1}}} \right).$$

We denote

$$G_{n,p} \triangleq \frac{C_p}{\sqrt{m_{n+1}}}, \tag{4.13}$$

such that, for all $n \geq 0$ and $p' \in [2, p]$ we obtain the expressions

$$\mathbb{E}[\|W_{n+1}\|^{p'}|\mathcal{F}_n]^{\frac{1}{p'}} \leq G_{n,p'} \left(\sigma(x^*) + \sigma_0 \|X_n - x^*\| \right), \tag{4.14}$$

$$\mathbb{E}[\|Z_{n+1}\|^{p'}|\mathcal{F}_n]^{\frac{1}{p'}} \leq \sigma(x^*)G_{n,p'}(1 + \alpha_n \sigma_0 G_{n,p'}) \tag{4.15}$$
$$+ \sigma_0 G_{n,p'}\|X_n - x^*\|(1 + \alpha_n L + \alpha_n \sigma_0 G_{n,p'}),$$

$$\mathbb{E}[\|Y_n - x^*\|^{p'}|\mathcal{F}_n]^{\frac{1}{p'}} \leq (1 + \alpha_n L + \alpha_n \sigma_0 G_{n,p'})\|X_n - x^*\| + \alpha_n \sigma(x^*)G_{n,p'}. \tag{4.16}$$

In case of a (UBV), we obtain from the above estimates simple upper bounds, by setting $\sigma_0 = 0$, and replacing $\sigma(x^*)$ with the uniform upper bound $\hat{\sigma}$. We next use these derived expressions to obtain $L^q(\mathbb{P})$ bounds for the error increments $(\Delta U_n(x^*))_{n \geq 1}$ and $(\Delta V_n)_{n \geq 1}$, when $q \in [1, p/2]$.

Lemma 4.3. *Let Assumption 7 be fulfilled with $p \geq 2$. For $p' \in [2, p]$, $q = \frac{p'}{2} \geq 1$ and all $n \geq 0$ we have*

$$\mathbb{E}[|\Delta V_n|^q|\mathcal{F}_n]^{\frac{1}{q}} \leq \alpha_n^2 G_{n,p'}^2 \sigma(x^*)^2 [2(4 + \rho_n) + 16 + 16\alpha_n^2 \sigma_0^2 G_{n,p'}^2]$$
$$+ \alpha_n^2 G_{n,p'}^2 \sigma_0^2 \|X_n - x^*\|^2$$
$$\times [2(4 + \rho_n) + 8(1 + \alpha_n L + \alpha_n \sigma_0 G_{n,p'})^2] \tag{4.17}$$

and

$$\mathbb{E}[|\Delta U_n(x^*)|^q|\mathcal{F}_n]^{\frac{1}{q}}$$
$$\leq 2\alpha_n^2 G_{n,p'}^2 \sigma(x^*)^2 (1 + \alpha_n G_{n,p'}\sigma_0)$$
$$+ 2\alpha_n G_{n,p'}\sigma(x^*)\|X_n - x^*\|[1 + \alpha_n L + \alpha_n \sigma_0 G_{n,p'}(3 + 2\alpha_n L) + 2\alpha_n^2 \sigma_0^2 G_{n,p'}^2]$$
$$+ 2\alpha_n G_{n,p'}\sigma_0 \|X_n - x^*\|^2 (1 + \alpha_n L + \alpha_n \sigma_0 G_{n,p'})^2. \tag{4.18}$$

If (UBV) holds with variance bound $\hat{\sigma}$, then these upper bounds simplify to

$$\mathbb{E}[|\Delta V_n|^q|\mathcal{F}_n]^{\frac{1}{q}} \leq \alpha_n^2 \hat{\sigma}^2 G_{n,p'}^2 (8 + \rho_n) \tag{4.19}$$

and, respectively,

$$\mathbb{E}[|\Delta U_n(x^*)|^q|\mathcal{F}_n]^{\frac{1}{q}} \leq 2\alpha_n \hat{\sigma} G_{n,p'}(1 + L\alpha_n)\|X_n - x^*\| + 2\alpha_n^2 \hat{\sigma}^2 G_{n,p'}^2. \tag{4.20}$$

Proof. Proof. Let be $n \geq 0$. For $q \geq 1$, we know that

$$\mathbb{E}[|\Delta V_n|^q|\mathcal{F}_n]^{\frac{1}{q}} \leq (4 + \rho_n)\alpha_n^2 \mathbb{E}[\|W_{n+1}\|^{p'}|\mathcal{F}_n]^{\frac{2}{p'}} + 4\alpha_n^2 \mathbb{E}[\|Z_{n+1}\|^{p'}|\mathcal{F}_n]^{\frac{2}{p'}}.$$

Using (4.14) and (4.15), and rearranging terms, we obtain (4.17). On the other hand, we have by definition

$$\mathbb{E}[|\Delta U_n(x^*)|^q|\hat{\mathcal{F}}_n]^{\frac{1}{q}} \leq 2\alpha_n \|Y_n - x^*\| \cdot \mathbb{E}[\|Z_{n+1}\|^q|\hat{\mathcal{F}}_n]^{\frac{1}{q}}$$
$$\leq 2\alpha_n \|Y_n - x^*\| \cdot \mathbb{E}[\|Z_{n+1}\|^{p'}|\hat{\mathcal{F}}_n]^{\frac{1}{p'}}$$
$$\leq 2\alpha_n \|Y_n - x^*\|G_{n,p'}\sigma(x^*) + 2\alpha_n G_{n,p'}\sigma_0 \|Y_n - x^*\|^2,$$

where the first estimate follows from the Cauchy-Schwarz inequality, the second one uses the monotonicity of $L^q(\mathbb{P})$ norms, and the third one uses eq. (A.4). Applying the operator $\mathbb{E}[\cdot|\mathcal{F}_n]$ on both sides, and using again the monotonicity of $L^q(\mathbb{P})$ norms, we obtain

$$\mathbb{E}[|\Delta U_n(x^*)|^q|\mathcal{F}_n]^{\frac{1}{q}} \leq 2\alpha_n G_{n,p'}\sigma(x^*)\mathbb{E}[\|Y_n - x^*\|^q|\mathcal{F}_n]^{\frac{1}{q}}$$

$$+ 2\alpha_n G_{n,p'} \sigma_0 \mathbb{E}[\|Y_n - x^*\|^{p'} |\mathcal{F}_n]^{\frac{2}{p'}}$$

$$\leq 2\alpha_n G_{n,p'} \sigma(x^*) \mathbb{E}[\|Y_n - x^*\|^{p'} |\mathcal{F}_n]^{\frac{1}{p'}}$$

$$+ 2\alpha_n G_{n,p'} \sigma_0 \mathbb{E}[\|Y_n - x^*\|^{p'} |\mathcal{F}_n]^{\frac{2}{p'}}.$$

After applying (4.16) and rearranging terms we arrive at the expression (4.18).

In case UBV holds with uniform variance bound $\hat{\sigma}$, the upper bound for $|\Delta V_{n+1}|^q$ follows immediately from the defining expression (4.2) by using the uniform bounds $\frac{C_{p'}\hat{\sigma}}{\sqrt{m_{n+1}}} = G_{n,p'}\hat{\sigma}$ for the quadratic error terms $\|W_{n+1}\|^2$ and $\|Z_{n+1}\|^2$. The corresponding bound for $|\Delta U_n(x^*)|^q$ is obtained from (4.18) by setting $\sigma_0 = 0$ and replacing $\sigma(x^*)$ by its uniform upper bound $\hat{\sigma}$. ∎

Based on the previous estimates, we can now derive the announced stochastic quasi-Fejér inequality for the sequence $\left(\|X_n - x^*\|^2\right)_{n \geq 0}$.

Proposition 4.4. *For all $x^* \in \mathcal{X}_*$ and all $n \geq 0$, we have*

$$\mathbb{E}[\|X_{n+1} - x^*\|^2 |\mathcal{F}_n] \leq \|X_n - x^*\|^2 - \frac{\rho_n}{2} r_{\alpha_n}(X_n)^2 + \frac{\kappa_n}{m_{n+1}} \left[\sigma_0^2 \|X_n - x^*\|^2 + \sigma(x^*)^2\right],$$
(4.21)

where

$$\kappa_n \triangleq \alpha_n^2 C_2^2 [2(4 + \rho_n) + 16(1 + \alpha_n L + \alpha_n \sigma_0 G_{n,2})^2].$$

If (UBV) *holds with uniform variance bound $\hat{\sigma}$, then*

$$\mathbb{E}[\|X_{n+1} - x^*\|^2 |\mathcal{F}_n] \leq \|X_n - x^*\|^2 - \frac{\rho_n}{2} r_{\alpha_n}(X_n)^2 + \frac{\kappa_n \hat{\sigma}^2}{m_{n+1}},$$
(4.22)

where now $\kappa_n = \alpha_n^2 C_2^2 (8 + \rho_n)$.

Proof. Proof. Let be $x^* \in \mathcal{X}_*$ and $n \geq 0$. Our point of departure is (4.9), together with (4.17). From here we derive that

$$\mathbb{E}[\|X_{n+1} - x^*\|^2 |\mathcal{F}_n]$$
$$\leq \|X_n - x^*\|^2 - \frac{\rho_n}{2} r_{\alpha_n}(X_n)^2$$
$$+ \alpha_n^2 G_{n,2}^2 \sigma(x^*)^2 [2(4 + \rho_n) + 16 + 16\alpha_n^2 \sigma_0^2 G_{n,2}^2]$$
$$+ \alpha_n^2 G_{n,2}^2 \sigma_0^2 \|X_n - x^*\|^2 [2(4 + \rho_n) + 8(1 + \alpha_n L + \alpha_n \sigma_0 G_{n,p2})^2]$$
$$\leq \|X_n - x^*\|^2 - \frac{\rho_n}{2} r_{\alpha_n}(X_n)^2$$
$$+ \left(\sigma_0^2 \|X_n - x^*\|^2 + \sigma(x^*)^2\right) \left[2(4 + \rho_n) + 16(1 + \alpha_n L + \alpha_n \sigma_0 G_{n,2})^2\right] \alpha_n^2 G_{n,2}^2.$$

In the last equality, we have used that $2(4 + \rho_n) + 8(1 + \alpha_n L + \alpha_n \sigma_0 G_{n,2})^2 \leq 2(4 + \rho_n) + 16(1 + \alpha_n L + \alpha_n \sigma_0 G_{n,2})^2$, and that $2(4 + \rho_n) + 16 + 16\alpha_n^2 \sigma_0^2 G_{n,2}^2 \leq 2(4 + \rho_n) + 16(1 + \alpha_n L + \alpha_n \sigma_0 G_{n,2})^2$. Recalling that $G_{n,2} = C_2/\sqrt{m_{n+1}}$, the proof is complete.

In the case where (UBV) holds, we just have to combine (4.9) with (4.19) to obtain the claimed result. ∎

Remark 4.2. The scaling factor κ_n only depends on the step size α_n, the Lipschitz constant L, and the variance bound on the stochastic oracle. Let $\bar{\alpha} \triangleq \sup_{n \geq 0} \alpha_n$ and

$\underline{\alpha} \triangleq \inf_{n \geq 0} \alpha_n$ (both finite and positive according to Assumption 5). Using the definition of ρ_n in (4.1), we can bound

$$\kappa_n = \alpha_n^2 C_2^2 \left[2(4 + \rho_n) + 16(1 + \alpha_n L + \frac{\alpha_n \sigma_0 C_2}{\sqrt{m_{n+1}}})^2 \right]$$

$$\leq \alpha_n^2 C_2^2 \left[10 + 32(1 + \alpha_n L)^2 + 32 \alpha_n^2 \sigma_0^2 \frac{C_2^2}{m_{n+1}} \right]$$

$$\leq \bar{\alpha}^2 C_2^2 \mathsf{c}_1 \left[1 + \frac{\bar{\alpha}^2 \sigma_0^2 C_2^2}{m_{n+1}} \right] \quad \forall n \geq 0,$$

where $\mathsf{c}_1 > 1$ is a constant. Combined with the batch size condition (3.3), we obtain the existence of constants c_0 and c_1 such that

$$\kappa_n \leq \mathsf{c}_1 \left(1 + \frac{\bar{\alpha}^2 \sigma_0^2 C_2^2}{\mathsf{c}_0 (n + n_0)^{1+a} \ln(n + n_0)^{1+b}} \right)$$

for all $n \gg n_0$. Such non-asymptotic bounds will be used in the estimation of the rate of convergence of the algorithm.

Next we will prove that the process $(X_n)_{n \geq 0}$ converges a.s. to a random variable X with values in the set $\mathcal{X}_*$. This will be obtained as a consequence of the classical Robbins-Siegmund Lemma 2.3, and recent results on the convergence of stochastic quasi-Féjer monotone sequences (Proposition 2.3 in Combettes and Pesquet (2015)).

Given a stochastic process $(f_n)_{n \geq 0} \subseteq L^0(\Omega, \mathcal{F}, \mathbb{P}; \mathbb{R}^d)$, we define the (random) set of cluster points

$$\mathsf{Lim}(f)(\omega) \triangleq \{x \in \mathbb{R}^d | (\exists (n_j) \uparrow \infty) : \lim_{n_j \to \infty} f_{n_j}(\omega) = x\}$$

Theorem 4.5. *Consider the stochastic process $(X_n, Y_n)_{n \geq 0}$ generated by Algorithm SFBF under Assumptions (1)-(7). Then, $(X_n)_{n \geq 0}$ converges as $n \to \infty$ almost surely to a limit random variable X with values in $\mathcal{X}_*$, and $\lim_{n \to \infty} \mathbb{E}[r_{\alpha_n}(X_n)^2] = 0$.*

Proof. Proof. We fix an element $x^* \in \mathcal{X}_*$. Let $\delta_n(x^*) \triangleq \|X_n - x^*\|^2$, $u_n \triangleq \frac{\rho_n}{2} r_{\alpha_n}(X_n)^2$, $\theta_n \triangleq \frac{\kappa_n \sigma_0^2}{m_{n+1}}$, and $\beta_n = \frac{\kappa_n \sigma(x^*)^2}{m_{n+1}}$, so that (4.21) can be rewritten for all $n \geq 0$ as

$$\mathbb{E}[\delta_{n+1}(x^*)|\mathcal{F}_n] \leq (1 + \theta_n)\delta_n(x^*) - u_n + \beta_n \qquad \mathbb{P} - \text{a.s.} \; .$$

Hence, by Lemma 2.3, there exists a random variable $\hat{\delta}(x^*) \in [0, \infty)$ such that $(\delta_n(x^*))_{n \geq 1} \to \hat{\delta}(x^*)$ a.s. as $n \to \infty$, and $\mathbb{P}\left[\sum_{n \geq 0} u_n < \infty\right] = 1$. In particular, $(X_n)_{n \geq 0}$ is bounded for almost every $\omega \in \Omega$. Since $\sum_{n \geq 0} u_n = \sum_{n \geq 0} \rho_n r_{\alpha_n}(X_n)^2 \geq \hat{\rho} \sum_{n \geq 0} r_{\alpha_n}(X_n)^2$, where $\hat{\rho} = 1 - 2\bar{\alpha}^2 L^2 > 0$, it follows that $\lim_{n \to \infty} r_{\alpha_n}(X_n) = 0$ $\mathbb{P}$−a.s.

We next show that for all $\omega \in \Omega$ all limit points of $(X_n(\omega))_{n \geq 0}$ are points in $\mathcal{X}_*$, and then apply Proposition 2.3(iii) in Combettes and Pesquet (2015) to conclude that $(X_n)_n$ converges almost surely to a random variable X with values in $\mathcal{X}_*$. Let $\omega \in \Omega$ be such that $X_n(\omega)$ is bounded. Since $(\alpha_n)_{n \geq 0}$ is bounded as well, we can construct subsequences $(\alpha_{n_j})_{j \geq 0}$ and $(X_{n_j}(\omega))_{j \geq 0}$ such that $\lim_{j \to \infty} \alpha_{n_j} = \alpha \in [\underline{\alpha}, \bar{\alpha}]$ and $\lim_{j \to \infty} X_{n_j}(\omega) = \chi(\omega)$. Additionally, we have $\lim_{j \to \infty} r_{\alpha_{n_j}}(X_{n_j}(\omega)) = 0$, so that

$$\lim_{j \to \infty} X_{n_j}(\omega) = \lim_{j \to \infty} \Pi_{\mathcal{X}}(X_{n_j}(\omega) - \alpha_{n_j} T(X_{n_j}(\omega))).$$

Therefore, by continuity of the projection operator and of the averaged map T, Lemma 2.1(iv) allows us to conclude that $\chi(\omega) \in \mathcal{X}_*$. Since the subsequence is arbitrary, it follows that $\mathsf{Lim}((X_n)_{n\geq 0})(\omega) \subseteq \mathcal{X}_*$ for $\mathbb{P}$-almost all $\omega \in \Omega$. Now apply Proposition 2.3(iv) of Combettes and Pesquet (2015) to conclude that $X_n \to X \in L^0(\Omega, \mathcal{F}, \mathbb{P}; \mathcal{X}_*)$ $\mathbb{P}$−a.s.

To prove that $r_{\alpha_n}(X_n)$ converges to 0 in mean square as $n \to \infty$, observe first that

$$\mathbb{E}[\delta_{n+1}(x^*)] \leq \mathbb{E}[\delta_n(x^*)] - \frac{\rho_n}{2}\mathbb{E}[r_{\alpha_n}(X_n)^2] + \frac{\kappa_n}{m_{n+1}}\left(\sigma_0^2 \mathbb{E}[\delta_n(x^*)] + \sigma(x^*)^2\right) \ \forall n \geq 0.$$

Let $z_n = \mathbb{E}[\delta_n(x^*)]$, $u_n = \frac{\rho_n}{2}\mathbb{E}[r_{\alpha_n}(X_n)^2]$ and θ_n and β_n be defined as in the previous paragraph. The deterministic version of the Robbins-Siegmund Lemma (Lemma 2.3) gives $(u_n)_{n\geq 1} \in \ell_+^1(\mathbb{N})$. Hence, $\lim_{n\to\infty} \mathbb{E}[r_{\alpha_n}(X_n)^2] = 0$. ∎

Theorem 4.5 considerably strengthens similar results obtained via different splitting techniques. For SEG, asymptotic convergence of the iterates in the sense of Theorem 4.5 is established in Theorem 3 of Iusem et al. (2017). However, different to SFBF, SEG requires two costly projection steps, with the same number of oracle calls. This makes Algorithm SFBF a potentially more efficient tool, and we will demonstrate that this is actually the case empirically, as well as theoretically. Under strong monotonicity assumptions, a version of Theorem 4.5 has been recently established for a stochastic version of the classical forward-backward splitting technique in Rosasco et al. (2016), assuming a similar variance structure on the stochastic oracle as we do. Theorem 4.5 shows convergence of SFBF under the much weaker assumption of pseudo-monotonicity of the mean operator T.

We close this section by reporting an improved stochastic quasi-Fejér property in terms of the distance to the solution set $\mathcal{X}_*$.

Proposition 4.6. *Suppose that Assumptions (1)-(7) hold. For $x^* \in \mathcal{X}_*$, set $\hat{\sigma}(x^*) \triangleq \max\{\sigma(x^*), \sigma_0\}$, and define $\mathrm{dist}(x, \mathcal{X}_*) \triangleq \inf_{y\in\mathcal{X}_*}\|y - x\| = \|\Pi_{\mathcal{X}_*}(x) - x\|$. For all $n \geq 0$, it holds*

$$\mathbb{E}[\mathrm{dist}(X_{n+1}, \mathcal{X}_*)^2 | \mathcal{F}_n] \leq \mathrm{dist}(X_n, \mathcal{X}_*)^2 - \frac{\rho_n}{2}r_{\alpha_n}(X_n)^2$$
$$+ \frac{\kappa_n\hat{\sigma}(\Pi_{\mathcal{X}_*}(X_n))^2}{m_{n+1}}[1 + \mathrm{dist}(X_n, \mathcal{X}_*)^2].$$

If (UBV) holds, then we get for all $n \geq 0$ the uniform bound

$$\mathbb{E}[\mathrm{dist}(X_{n+1}, \mathcal{X}_*)^2 | \mathcal{F}_n] \leq \mathrm{dist}(X_n, \mathcal{X}_*)^2 - \frac{\rho_n}{2}r_{\alpha_n}(X_n)^2 + \frac{\kappa_n\hat{\sigma}^2}{m_{n+1}},$$

with $\kappa_n = \alpha_n^2 C_2^2(8 + \rho_n)$.

Proof. Proof. Let be $\pi_n(\omega) = \Pi_{\mathcal{X}_*}(X_n(\omega))$ for all $n \geq 0$ and all $\omega \in \Omega$. Since the projection operator onto the closed and convex set $\mathcal{X}_*$ is nonexpansive, we have $(\pi_n)_{n\geq 0} \in \ell^0(\mathbb{F})$. For all $n \geq 0$, we have

$$\mathbb{E}[\mathrm{dist}(X_{n+1}, \mathcal{X}_*)^2 | \mathcal{F}_n] \leq \mathbb{E}[\|X_{n+1} - \pi_n\|^2 | \mathcal{F}_n]$$
$$\leq \|X_n - \pi_n\|^2 - \frac{\rho_n}{2}r_{\alpha_n}(X_n) + \frac{\kappa_n}{m_{n+1}}[\sigma_0^2\|X_n - \pi_n\|^2 + \sigma(\pi_n)^2]$$
$$\leq \mathrm{dist}(X_n, \mathcal{X}_*)^2 - \frac{\rho_n}{2}r_{\alpha_n}(X_n) + \frac{\kappa_n\hat{\sigma}^2(\pi_n)}{m_{n+1}}[\mathrm{dist}(X_n, \mathcal{X}_*)^2 + 1],$$

where the second inequality uses Proposition 4.4. ∎

We are now in the position to make our claim precise that we can always provide the current iterate of the shadow sequence $(Y_n)_{n \geq 0}$ as the approximate solution of our SVI.

Proposition 4.7. *Consider the stochastic process $(X_n, Y_n)_{n \geq 0}$ generated by Algorithm SFBF under Assumptions (1)-(7). Let $(\pi_n)_{n \geq 0}$ be the $\mathcal{X}_*$-valued $\mathbb{F}$-adapted stochastic process defined by $\pi_n = \Pi_{\mathcal{X}_*}(X_n)$ for all $n \geq 0$. Define the sequences $(\gamma_n)_{n \geq 0}, (\beta_n)_{n \geq 0}$ by $\gamma_n = 1 + \alpha_n L + \alpha_n \frac{C_2 \sigma_0}{\sqrt{m_{n+1}}}$, and $\beta_n = \alpha_n \frac{C_2 \sigma(\pi_n)}{\sqrt{m_{n+1}}}$. If $\beta_n \to 0$ almost surely, then $(Y_n)_n$ converges almost surely to a $\mathcal{X}_*$ valued random variable Y.*

Proof. Proof. Let $d(x) \triangleq \text{dist}(x, \mathcal{X}_*)$, and consider the stochastic process $(c_n)_{n \geq 0}$ defined as $c_n \triangleq \mathbb{E}[d(Y_n)^2 | \mathcal{F}_n]^{1/2}$. Furthermore, we define the $\mathcal{X}_*$-valued random process by $\pi_n \triangleq \Pi_{\mathcal{X}_*}(X_n)$. From (4.12), we see

$$c_n \leq \mathbb{E}[\|Y_n - \pi_n\|^2 | \mathcal{F}_n]^{\frac{1}{2}}$$
$$\leq \left(1 + \alpha_n L + \alpha_n \frac{C_2 \sigma_0}{\sqrt{m_{n+1}}} \right) d(X_n) + \alpha_n \frac{C_2 \sigma(\pi_n)}{\sqrt{m_{n+1}}}$$
$$= \gamma_n d(X_n) + \beta_n.$$

Taking expectations on both sides,

$$\mathbb{E}[c_n] \leq \gamma_n \mathbb{E}[d(X_n)] + \mathbb{E}[\beta_n]$$

By Theorem 4.5, $X_n \to X$ almost surely, an $\mathcal{X}_*$-valued random variable. Therefore, we know that $\lim_{n \to \infty} \mathbb{E}[d(X_n)] = 0$. By hypothesis $\beta_n \to 0$ almost surely, so that $\limsup_{n \to \infty} \mathbb{E}[c_n] \leq 0$. Now, $c_n^2 = \mathbb{E}[d(Y_n)^2 | \mathcal{F}_n]$, and by Jensen's inequality

$$\mathbb{E}[c_n^2] = \mathbb{E}[d(Y_n)^2] \leq \mathbb{E}[c_n]^2.$$

Hence, $\limsup_{n \to \infty} \mathbb{E}[d(Y_n)^2] \leq 0$, and consequently $d(Y_n) \to 0$ almost surely. The convergence to an $\mathcal{X}_*$-valued limit random variable then follows from Proposition 2.3 in Combettes and Pesquet (2015). ∎

We remark that the assumption $\beta_n \to 0$ a.s. is rather mild. A sufficient condition is that $(\beta_n)_{n \geq 0} \in \ell_+^1(\mathbb{F})$. It trivially holds under the (UBV) assumption on the stochastic oracle's variance.

5. COMPLEXITY ANALYSIS AND RATES

The next two propositions provide explicit norm bounds on the iterates $(X_n)_{n \geq 0}$. These bounds are going to be crucial to assess the convergence rate and the per-iteration complexity of the proposed algorithm. To be sure, the formal appearance of the complexity estimates derived in this section is naturally similar to the corresponding bounds derived in Iusem et al. (2017). However, the key observation we would like to emphasize here is that an explicit comparison between the constants involved in the upper bounds obtained for Algorithm SFBF with those appearing in SEG shows that the constants are consistently smaller. This indicates that SFBF should empirically outperform SEG. This fact is consistently observed in all our numerical experiments, and, as we show in Section 6, actually this promised gain can be quite significant.

Proposition 5.1. *Suppose that Assumptions (1)-(7) hold. For all $x^* \in \mathcal{X}_*$ let*

$$\hat{\sigma}(x^*) \triangleq \max\{\sigma(x^*), \sigma_0\}, \tag{5.1}$$

$$\mathsf{a}(x^*) \triangleq \hat{\sigma}^2(x^*)\bar{\alpha}^2 C_2^2 \mathsf{c}_1. \tag{5.2}$$

Choose $n_0 \in \mathbb{N}$ and $\gamma > 0$ such that

$$\sum_{n \geq n_0} \frac{1}{m_{n+1}} \leq \gamma \tag{5.3}$$

and

$$\beta(x^*) \triangleq \gamma \mathsf{a}(x^*) + \gamma^2 \mathsf{a}(x^*)^2 \in (0, 1). \tag{5.4}$$

Then

$$\sup_{n \geq n_0+1} \mathbb{E}[\|X_n - x^*\|^2] \leq \frac{\mathbb{E}[\|X_{n_0} - x^*\|^2] + 1}{1 - \beta(x^*)}. \tag{5.5}$$

Proof. Proof. Thanks to Assumption 6, for every $\gamma > 0$, we can find an index $n_0 \in \mathbb{N}$ such that (5.3) holds. Consequently, we fix $n_0 \in \mathbb{N}$ to be the smallest positive integer so that (5.3) holds. For all $n \geq 0$ we denote $\psi_n(x^*) \triangleq \mathbb{E}[\|X_n - x^*\|^2]$. From Proposition 4.4, we obtain

$$\psi_{n+1}(x^*) \leq \psi_n(x^*) - \frac{\rho_n}{2}\mathbb{E}[r_{\alpha_n}(X_n)^2] + \frac{\kappa_n}{m_{n+1}}[\sigma_0^2 \psi_n(x^*) + \sigma(x^*)^2] \ \forall n \geq 0.$$

Recall from Remark (4.2) that

$$\kappa_n \leq \bar{\alpha}^2 C_2^2 \mathsf{c}_1 \left(1 + \frac{\bar{\alpha}^2 \sigma_0^2 C_2^2}{m_{n+1}}\right) \leq \bar{\alpha}^2 C_2^2 \mathsf{c}_1 \left(1 + \frac{\mathsf{a}(x^*)}{\mathsf{c}_1 m_{n+1}}\right).$$

Using this bound, for all $n \geq n_0 + 1$ the previous display telescopes to

$$\psi_n(x^*) \leq \psi_{n_0}(x^*) + \sum_{k=n_0}^{n-1}(1 + \psi_k(x^*))\frac{\mathsf{a}(x^*)}{m_{k+1}} + \sum_{k=n_0}^{n-1}(1 + \psi_k(x^*))\frac{\mathsf{a}(x^*)^2}{\mathsf{c}_1 m_{k+1}^2}.$$

For $p > \psi_{n_0}(x^*)$, define $\tau_p(x^*) \triangleq \inf\{n \geq n_0 + 1 | \psi_n(x^*) \geq p\} \in \mathbb{N} \cup \{+\infty\}$. We claim that there exists $\hat{p} > \psi_{n_0}(x^*)$ such that $\tau_{\hat{p}}(x^*) = \infty$. Assuming that this is not the case, then we must have that $\tau_p(x^*) < \infty$ for all $p > \psi_{n_0}(x^*)$. Therefore, by definition of $\tau_p(x^*)$ and (5.3), we get

$$p \leq \psi_{\tau_p(x^*)}(x^*) \leq \psi_{n_0}(x^*) + \sum_{k=n_0}^{\tau_p(x^*)-1}(1 + \psi_k(x^*))\frac{\mathsf{a}(x^*)}{m_{k+1}}$$

$$+ \sum_{k=n_0}^{\tau_p(x^*)-1}(1 + \psi_k(x^*))\frac{1}{\mathsf{c}_1}\left(\frac{\mathsf{a}(x^*)}{m_{k+1}}\right)^2$$

$$\leq \psi_{n_0}(x^*) + (1 + p)\gamma \mathsf{a}(x^*) + (1 + p)\frac{\gamma^2 \mathsf{a}(x^*)^2}{\mathsf{c}_1}.$$

Rearranging, and using $\mathsf{c}_1 > 1$ as well as (5.4), gives

$$p \leq \frac{\psi_{n_0}(x^*) + 1}{1 - \gamma \mathsf{a}(x^*) - \frac{\gamma^2}{\mathsf{c}_1}\mathsf{a}(x^*)^2} \leq \frac{\psi_{n_0}(x^*) + 1}{1 - \gamma \mathsf{a}(x^*) - \gamma^2 \mathsf{a}(x^*)^2}.$$

Since $p > \psi_{n_0}(x^*)$ has been chosen arbitrarily, we can let $p \to \infty$ and obtain a contradiction. Therefore, there exists $\hat{p} > \psi_{n_0}(x^*)$ such that $\bar{p} \triangleq \sup_{n \geq n_0+1} \psi_n(x^*) \leq \hat{p} < \infty$.

From here we get for all $n \geq n_0 + 1$

$$\psi_n(x^*) \leq \psi_{n_0}(x^*) + \sum_{k=n_0}^{n-1} (1 + \psi_k(x^*)) \frac{\mathsf{a}(x^*)}{m_{k+1}} + \sum_{k=n_0}^{n-1} (1 + \psi_k(x^*)) \frac{1}{\mathsf{c}_1} \left(\frac{\mathsf{a}(x^*)}{m_{k+1}} \right)^2$$

$$\leq \psi_{n_0}(x^*) + (1 + \bar{p}) \gamma \mathsf{a}(x^*) + (1 + \bar{p}) \frac{\gamma^2 \mathsf{a}(x^*)^2}{\mathsf{c}_1}.$$

Taking the supremum over $n \geq n_0 + 1$, and shifting back to the original expressions of the involved data, we get

$$\bar{p} = \sup_{n \geq n_0+1} \mathbb{E}[\|X_n - x^*\|^2] \leq \frac{\mathbb{E}[\|X_{n_0} - x^*\|^2] + 1}{1 - \beta(x^*)},$$

which further leads to (5.5). ∎

In case where the local variance of the SO is uniformly bounded over the solution set $\mathcal{X}_*$, we obtain much sharper results, allowing us to bound the distance of the iterates away from the solution set.

Proposition 5.2. *Suppose that Assumptions* (1)-(7) *hold. Suppose the variance over the solution set* $\mathcal{X}_*$ *is bounded:* $\hat{\sigma}(x^*) \triangleq \max\{\sigma(x^*), \sigma_0\} \leq \hat{\sigma}$ *for all* $x^* \in \mathcal{X}_*$. *Define*

$$\mathsf{a} \triangleq \bar{\alpha}^2 \hat{\sigma}^2 C_2^2 \mathsf{c}_1. \tag{5.6}$$

Let $\phi \in (0, \frac{\sqrt{5}-1}{2})$ *and choose* $n_0 \geq 1$ *such that* $\sum_{i \geq n_0} \frac{1}{m_{i+1}} \leq \frac{\phi}{\mathsf{a}}$. *Then*

$$\sup_{n \geq n_0+1} \mathbb{E}[\operatorname{dist}(X_n, \mathcal{X}_*)^2] \leq \frac{1 + \mathbb{E}[\operatorname{dist}(X_{n_0}, \mathcal{X}_*)^2]}{1 - \phi - \phi^2}. \tag{5.7}$$

Proof. Proof. We denote by $d(x) \triangleq \operatorname{dist}(x, \mathcal{X}_*) : \mathbb{R}^d \to \mathbb{R}_+$ the distance function of the solution set $\mathcal{X}_*$. Since $\mathcal{X}_*$ is a nonempty, closed and convex subset of $\mathbb{R}^d$, the function $d(X_n) : \Omega \to \mathbb{R}_+$ given by $\omega \mapsto d(X_n)(\omega) \triangleq \operatorname{dist}(X_n(\omega), \mathcal{X}_*)$ is $\mathcal{F}_n$-measurable for all $n \geq 0$. Indeed, letting $\pi_n(\omega) \triangleq \Pi_{\mathcal{X}_*}(X_n(\omega))$ for all $n \geq 0$, then first, $(\pi_n)_{n \geq 0} \in \ell_+^0(\mathbb{F})$, and second $d(X_n)(\omega) = \|X_n(\omega) - \pi_n(\omega)\|$ is a well-defined random process in $\ell_+^0(\mathbb{F})$, being a composition of continuous and measurable functions. Therefore, for all $n \geq 0$,

$$\mathbb{E}[d(X_{n+1})^2 | \mathcal{F}_n] \leq \mathbb{E}[\|X_{n+1} - \pi_n\|^2 | \mathcal{F}_n]$$

$$\leq \|X_n - \pi_n\|^2 - \frac{\rho_n}{2} \mathbb{E}[r_{\alpha_n}(X_n)^2] + \frac{\kappa_n}{m_{n+1}} \left(\sigma_0^2 d(X_n)^2 + \sigma(\pi_n)^2 \right).$$

Call $\psi_n \triangleq \sqrt{\mathbb{E}[d(X_n)^2]}$ for all $n \geq 0$. Taking expectations in the previous display, and using the assumed uniform bound of the variance, we arrive at

$$\psi_{n+1}^2 \leq \psi_n^2 - \frac{\rho_n}{2} \mathbb{E}[r_{\alpha_n}(X_n)^2] + \frac{\hat{\sigma}^2 \kappa_n}{m_{n+1}} \left(1 + \psi_n^2 \right) \quad \forall n \geq 0.$$

From Remark (4.2), we know that

$$\kappa_n \leq \bar{\alpha}^2 C_2^2 \mathsf{c}_1 \left(1 + \frac{\bar{\alpha}^2 \hat{\sigma}^2 C_2^2}{m_{n+1}} \right),$$

so that $\hat{\sigma}^2 \kappa_n \leq \mathsf{a}(1 + \frac{\mathsf{a}}{m_{n+1}\mathsf{c}_1})$ for all $n \geq 0$. Hence, for all $n \geq n_0 + 1$

$$\psi_n^2 \leq \psi_{n_0}^2 + \sum_{k=n_0}^{n-1} (1 + \psi_k^2) \frac{\mathsf{a}}{m_{k+1}} + \sum_{k=n_0}^{n-1} (1 + \psi_k^2) \frac{\mathsf{a}^2}{\mathsf{c}_1 m_{k+1}^2}.$$

From here proceed, mutatis mutandis, as in the proof of Proposition 5.1. ∎

We next give explicit estimates of the rate of convergence and the oracle complexity of SFBF. The reported results are very similar to the extragradient method, with the important remark that all numerical constants can be improved under our forward-backward-forward scheme. For that purpose, it is sufficient to consider Algorithm SFBF with a constant step size $\alpha_n = \alpha \in \left(0, \frac{1}{\sqrt{2}L}\right)$ for all $n \geq 0$.[3] As in Iusem et al. (2017), we can provide non-asymptotic convergence rates for the sequence $(\mathbb{E}[r_\alpha(X_n)^2])_{n \geq 0}$.

For all $n \geq 0, x^* \in \mathcal{X}_*$ and $\phi \in \left(0, \frac{\sqrt{5}-1}{2}\right)$, define

$$\Gamma_n \triangleq \sum_{i=0}^{n} \frac{1}{m_{i+1}}, \ \Gamma_n^2 \triangleq \sum_{i=0}^{n} \frac{1}{m_{i+1}^2},$$

$$\rho = 1 - 2\alpha^2 L^2, \ \delta_n(x^*) \triangleq \|X_n - x^*\|^2,$$

$$\text{and } H(x^*, n, \phi) \triangleq \frac{1 + \max_{0 \leq i \leq n} \mathbb{E}[\delta_i(x^*)]}{1 - \phi - \phi^2}.$$

Theorem 5.3. *Suppose that Assumptions* (1)-(7) *hold. Let* $x^* \in \mathcal{X}_*$ *be arbitrarily chosen, and consider Algorithm SFBF with constant step size* $\alpha \in \left(0, \frac{1}{\sqrt{2}L}\right)$. *Choose* $\phi \in \left(0, \frac{\sqrt{5}-1}{2}\right)$ *and* $n_0 \triangleq n_0(x^*)$ *to be the first integer such that*

$$\sum_{i \geq n_0} \frac{1}{m_{i+1}} \leq \frac{\phi}{\mathsf{a}(x^*)}, \tag{5.8}$$

where $\mathsf{a}(x^*)$ *is defined in* (5.2). *Let*

$$\Lambda_n(x^*, \phi) \triangleq \frac{2}{\rho} \left\{ \mathbb{E}[\delta_0(x^*)] + (1 + H(x^*, n_0, \phi)) \left(\mathsf{a}(x^*)\Gamma_n + \mathsf{a}(x^*)^2 \Gamma_n^2 \right) \right\},$$

$$\Lambda_\infty(x^*, \phi) \triangleq \sup_{n \geq 0} \Lambda_n(x^*, \phi).$$

For all $\varepsilon > 0$, *define the stopping time*

$$N_\varepsilon \triangleq \inf\{n \geq 0 | \mathbb{E}[r_\alpha(X_n)^2] \leq \varepsilon\}. \tag{5.9}$$

Then, either $N_\varepsilon = 0$, *or*

$$\mathbb{E}[r_\alpha(X_{N_\varepsilon})^2] \leq \varepsilon < \frac{\Lambda_\infty(x^*, \phi)}{N_\varepsilon}. \tag{5.10}$$

Proof. Proof. Let $\gamma = \frac{\phi}{\mathsf{a}(x^*)}$, with the constant $\mathsf{a}(x^*)$ defined in (5.2), and $n_0 = n_0(x^*)$ as required in the statement of the theorem. From Proposition 5.1, we deduce the bound

$$\sup_{n \geq n_0 + 1} \mathbb{E}[\delta_n(x^*)] \leq \frac{1 + \mathbb{E}[\delta_{n_0}(x^*)]}{1 - \phi - \phi^2} \leq H(x^*, n_0, \phi).$$

Since $1 - \phi - \phi^2 \in (0, 1)$, $\sup_{0 \leq i \leq n_0} \mathbb{E}[\delta_i(x^*)] \leq H(x^*, n_0, \phi)$. Therefore,

$$\sup_{n \geq 0} \mathbb{E}[\delta_n(x^*)] \leq H(x^*, n_0, \phi). \tag{5.11}$$

---

[3]The reason for this is that $\{r_a(x); a > 0\}$ is a family of equivalent merit functions of VI$(T, \mathcal{X})$ (see Proposition 10.3.6 in Facchinei and Pang (2003), and the opening to Section 4). Hence, as long as the step size policy $(\alpha_n)_{n \geq 0}$ obeys Assumption 5, we obtain the same rate estimates.

Taking expectations in equation (4.21), we get

$$\frac{\rho}{2}\mathbb{E}[r_\alpha(X_n)^2] \le \mathbb{E}[\delta_n(x^*)] - \mathbb{E}[\delta_{n+1}(x^*)] + \frac{\kappa_n}{m_{n+1}}\left(\sigma(x^*)^2 + \sigma_0^2\mathbb{E}[\delta_n(x^*)]\right) \ \ \forall n \ge 0.$$

Therefore, for all $n \ge 0$,

$$\frac{\rho}{2}\sum_{i=0}^n \mathbb{E}[r_\alpha(X_i)^2] \le \mathbb{E}[\delta_0(x^*)] + \sum_{i=0}^n \frac{\kappa_i}{m_{i+1}}\left(\sigma(x^*)^2 + \sigma_0^2\mathbb{E}[\delta_i(x^*)]\right).$$

Using the variance bound $\hat\sigma(x^*) = \max\{\sigma(x^*), \sigma_0\}$, which is well defined given the local boundedness of the variance, we get first from Remark (4.2) the bound

$$\kappa_i \le \alpha^2 C_2^2 \mathsf{c}_1 \left(1 + \frac{\alpha^2 C_2^2 \hat\sigma(x^*)^2}{m_{i+1}}\right) \ \ \forall i \ge 0.$$

Second, recalling that $\mathsf{a}(x^*) = \alpha^2\hat\sigma(x^*)^2 C_2^2 \mathsf{c}_1$, it yields for all $n \ge 0$

$$\frac{\rho}{2}\sum_{i=0}^n \mathbb{E}[r_\alpha(X_i)^2] \le \mathbb{E}[\delta_0(x^*)] + \sum_{i=0}^n \frac{\mathsf{a}(x^*)}{m_{i+1}}(1 + \mathbb{E}[\delta_i(x^*)])$$

$$+ \sum_{i=0}^n \frac{1}{\mathsf{c}_1}\left(\frac{\mathsf{a}(x^*)}{m_{i+1}}\right)^2 (1 + \mathbb{E}[\delta_i(x^*)])$$

$$\le \mathbb{E}[\delta_0(x^*)] + \left(1 + \max_{0 \le i \le n}\mathbb{E}[\delta_i(x^*)]\right)\left(\mathsf{a}(x^*)\Gamma_n + \mathsf{a}(x^*)^2\Gamma_n^2\right).$$

From (5.11), we conclude

$$\frac{\rho}{2}\sum_{i=0}^n \mathbb{E}[r_\alpha(X_i)^2] \le \mathbb{E}[\delta_0(x^*)] + (1 + H(x^*, n_0, \phi))\left(\mathsf{a}(x^*)\Gamma_n + \mathsf{a}(x^*)^2\Gamma_n^2\right)$$

$$= \frac{\rho}{2}\Lambda_n(x^*, \phi) \ \forall n \ge 0.$$

In conclusion,

$$\sum_{i=0}^n \mathbb{E}[r_\alpha(X_i)^2] \le \Lambda_n(x^*, \phi) \qquad \forall n \ge 0.$$

From Theorem 4.5, we know that for all $\varepsilon > 0$ there exists $M_\varepsilon \in \mathbb{N}$ such that $\mathbb{E}[r_\alpha(X_n)^2] \le \varepsilon$ for all $n \ge M_\varepsilon$. Hence, the (deterministic) stopping time N_ε defined in (5.9) is either 0, or an integer bounded from above. Focussing on the latter case $N_\varepsilon \ge 1$, then for every $0 \le k \le N_\varepsilon - 1$, we have

$$\varepsilon < \mathbb{E}[r_\alpha(X_i)^2].$$

From here, it follows

$$\varepsilon N_\varepsilon < \sum_{i=0}^{N_\varepsilon - 1} \mathbb{E}[r_\alpha(X_i)^2] \le \Lambda_{N_\varepsilon - 1}(x^*, \phi).$$

Hence,

$$\mathbb{E}[r_\alpha(X_{N_\varepsilon})^2] \le \varepsilon < \frac{\Lambda_\infty(x^*, \phi)}{N_\varepsilon}.$$

The two cases above can be compactly summarized to statement (5.10).   ∎

We next turn to the case where the local variance is uniformly bounded over the solution set. In the previous theorem, given $x^* \in \mathcal{X}_*$, the constant $\Lambda_\infty(x^*, n_0(x^*), \phi)$ in the convergence rate depends on the variance and on the distance of the $n_0(x^*)$ initial iterates to x^*, where $n_0(x^*)$ and ϕ are chosen such that (5.8) holds. Assuming a uniformly bound on the variance of SO over the solution set $\mathcal{X}_*$, we can obtain much stronger convergence rate estimates, holding uniformly over the solution set.

Proposition 5.4. *Assume that* $\sup_{x^* \in \mathcal{X}_*} \hat{\sigma}(x^*) \leq \hat{\sigma}$, *where the function* $\hat{\sigma}(\cdot)$ *is defined in* (5.1). *Let* $x^* \in \mathcal{X}_*$ *be arbitrarily chosen, and consider Algorithm SFBF with constant step size* $\alpha \in \left(0, \frac{1}{\sqrt{2}L}\right)$. *Choose* $\phi \in \left(0, \frac{\sqrt{5}-1}{2}\right)$ *and* $n_0 \triangleq n_0(\hat{\sigma})$ *to be the first integer such that*

$$\sum_{i \geq n_0} \frac{1}{m_{i+1}} \leq \frac{\phi}{\mathsf{a}}, \tag{5.12}$$

where $\mathsf{a} = \hat{\sigma}^2 \alpha^2 C_2^2 \mathsf{c}_1$. *Let*

$$\bar{\Lambda}_n(\hat{\sigma}, \phi) \triangleq \frac{2}{\rho} \left\{ \mathbb{E}[\mathrm{dist}(X_0, \mathcal{X}_*)^2] + (1 + \bar{H}(\hat{\sigma}, n_0, \phi))(\mathsf{a}\Gamma_n + \mathsf{a}^2\Gamma_n^2) \right\},$$

$$\bar{\Lambda}_\infty(\hat{\sigma}, \phi) = \sup_{n \geq 0} \bar{\Lambda}_n(\phi, \hat{\sigma}), \text{ and}$$

$$\bar{H}(\hat{\sigma}, n_0, \phi)) \triangleq \frac{1 + \max_{0 \leq i \leq n_0(\hat{\sigma})} \mathbb{E}[\mathrm{dist}(X_i, \mathcal{X}_*)]}{1 - \phi - \phi^2}.$$

For all $\varepsilon > 0$ *consider the stopping time defined in* (5.9). *Then, either* $N_\varepsilon = 0$, *or*

$$\mathbb{E}[r_\alpha(X_{N_\varepsilon})^2] \leq \varepsilon < \frac{\bar{\Lambda}_\infty(\hat{\sigma}, \phi)}{N_\varepsilon}, \tag{5.13}$$

Proof. Proof. The proof is almost identical to the proof of Theorem 5.3, but now we will use the estimates from Proposition 4.6 and Proposition 5.2 . We first remark that the upper variance bound $\hat{\sigma}$ is the only parameter in this statement; hence, the threshold index $n_0 = n_0(\hat{\sigma})$ depends on this parameter only. Once we made this choice, we can repeat all the steps involved in the proof of Theorem 5.3 verbatim, but by using Proposition 4.6 instead of Proposition 4.4, to conclude that

$$\sum_{i=0}^{n} \mathbb{E}[r_\alpha(X_i)^2] \leq \mathbb{E}[\mathrm{dist}(X_0, \mathcal{X}_*)^2] + \mathsf{a} \sum_{i=0}^{n} \frac{1 + \mathbb{E}[\mathrm{dist}(X_i, \mathcal{X}_*)^2]}{m_{i+1}}$$

$$+ \mathsf{a}^2 \sum_{i=0}^{n} \frac{1 + \mathbb{E}[\mathrm{dist}(X_i, \mathcal{X}_*)^2]}{m_{i+1}^2} \ \forall n \geq 0.$$

Proposition 5.2 gives us

$$\sup_{n \geq n_0+1} \mathbb{E}[\mathrm{dist}(X_n, \mathcal{X}_*)^2] \leq \frac{1 + \mathbb{E}[\delta(X_{n_0}, \mathcal{X}_*)^2]}{1 - \phi - \phi^2} \leq \bar{H}(\hat{\sigma}, n_0, \phi),$$

from which it follows

$$\sup_{n \geq 0} \mathbb{E}[\mathrm{dist}(X_n, \mathcal{X}_*)^2] \leq \bar{H}(\hat{\sigma}, n_0, \phi).$$

From here, we conclude just as in the proof of Theorem 5.3 that

$$\sum_{i=0}^{n} \mathbb{E}[r_\alpha(X_i)^2] \leq \bar{\Lambda}_n(\hat{\sigma}, \phi) \leq \bar{\Lambda}_\infty(\hat{\sigma}, \phi) \ \forall n \geq 0.$$

Choose $\varepsilon > 0$ arbitrary, and consider the stopping time (5.9). Then, either $N_\varepsilon = 0$, or else $N_\varepsilon \geq 1$. Focussing on the latter case, we argue just as in the proof of Theorem 5.3, that

$$\varepsilon N_\varepsilon < \sum_{i=0}^{N_\varepsilon - 1} \mathbb{E}[r_\alpha(X_i)^2] \leq \Lambda_{N_\varepsilon - 1}(x^*, \phi).$$

Hence, if N_ε not zero, we must have

$$\mathbb{E}[r_\alpha(X_{N_\varepsilon})^2] \leq \varepsilon < \frac{\Lambda_\infty(\hat{\sigma}, \phi)}{N_\varepsilon}.$$

∎

We now turn to the estimate of the oracle complexity. By this we mean the overall size of the data set needed to be processed in order to make the natural residual function smaller than a given tolerance level $\varepsilon > 0$, in mean square. Hence, using the stopping time (5.9), we would like to estimate the number $\sum_{i=0}^{N_\varepsilon} 2m_{i+1}$.

For simplicity, we will assume that the local variance function $\sigma(x^*)$ is uniformly bounded over the solution set $\mathcal{X}_*$. That is, we assume that there exists $\hat{\sigma} \in (0, \infty)$ such that $\sup_{x \in \mathcal{X}_*} \hat{\sigma}(x) \leq \hat{\sigma}$. A more complete argument, without making this strong assumption can be given similar to Proposition 3.23 in Iusem et al. (2017). We refrain doing so, since our main aim in this paper is to illustrate the improvement in the convergence rate when using Algorithm SFBF instead of SEG, and the simplest setting is enough for this purpose. We organize the derivation of an oracle complexity estimate in two parts. First, we will show that a specific (though admissible) choice of the sample rate, allows us to give an explicit bound on the number of preliminary iterates $n_0 \triangleq n_0(\hat{\sigma})$ needed to apply the general bounds reported in Proposition 5.4. Building on this insight, we directly estimate the oracle complexity.

As announced, we first establish a bound on the number of iterations we need to meet condition (5.12).

Lemma 5.5. *Let* $\mathtt{a}$ *be the constant defined in* (5.6), *and* $\phi \in (0, \frac{\sqrt{5}-1}{2})$. *We choose the sample rate*

$$m_i = \lceil \theta(\mu - 1 + i) \ln(\mu + i - 1)^{1+b} \rceil, \tag{5.14}$$

for $i \geq 1, \theta > 0, \mu > 1$ *and* $b > 0$. *Then, if* n_0 *is an integer satisfying*

$$n_0 \geq 1 - \mu + e^{\left(\frac{\mathtt{a}}{\phi \theta b}\right)^{1/b}},$$

we have $\sum_{i \geq n_0} \frac{1}{m_{i+1}} \leq \frac{\phi}{\mathtt{a}}$.

Proof. Proof. For $n_0 \geq 1$, we compute

$$\sum_{i \geq n_0} \frac{1}{m_{i+1}} \leq \frac{1}{\theta} \sum_{i \geq n_0} \frac{1}{(i + \mu) \ln(i + \mu)^{1+b}}$$

$$\leq \frac{1}{\theta} \int_{n_0 - 1}^{\infty} \frac{1}{(t + \mu) \ln(t + \mu)^{1+b}} \, dt$$

$$= \frac{1}{\theta b \ln(n_0 - 1 + \mu)^b}.$$

Therefore, if $\frac{1}{\theta b \ln(n_0 - 1 + \mu)^b} \leq \frac{\phi}{\mathtt{a}}$, we obtain the desired bound. Solving the latter inequality for n_0 gives the claimed result. ∎

Using the sample rate (5.14), we will now bound the constant $\bar{\Lambda}(\hat{\sigma}, \phi)$, and the stopping time N_ε. Define the constants

$$\mathcal{A}_{\mu,b} \triangleq \frac{\alpha^2 C_2^2 \mathsf{c}_1}{b \ln(\mu - 1)^b}, \ \mathcal{B}_{\mu,b} \triangleq \frac{\alpha^4 C_2^4 \mathsf{c}_1^2}{(1 + 2b)(\mu - 1) \ln(\mu - 1)^{1+2b}}.$$

Since,

$$\Gamma_\infty \leq \frac{1}{\theta b} \frac{1}{\ln(\mu - 1)^b}, \ \text{and} \ \Gamma_\infty^2 \leq \frac{1}{\theta^2} \frac{1}{(2b + 1)(\mu - 1) \ln(\mu - 1)^{1+2b}},$$

we conclude

$$\mathsf{a}\Gamma_\infty + \mathsf{a}^2\Gamma_\infty^2 \leq \max\{1, \theta^{-2}\}(\mathcal{A}_{\mu,b}\hat{\sigma}^2 + \mathcal{B}_{\mu,b}\hat{\sigma}^4).$$

Therefore,

$$\bar{\Lambda}(\hat{\sigma}, \phi) \leq \max\{1, \theta^{-2}\} \left\{ \frac{2}{\rho}\mathbb{E}[\text{dist}(X_0, \mathcal{X}_*)^2] + \frac{2}{\rho}(1 + \bar{H}(\hat{\sigma}, n_0, \phi)) \left[\mathcal{A}_{\mu,b}\hat{\sigma}^2 + \mathcal{B}_{\mu,b}\hat{\sigma}^4 \right] \right\}$$

$$\triangleq \max\{1, \theta^{-2}\}\mathcal{Q}(\phi, \hat{\sigma}).$$

This yields the following refined uniform bound on the squared residual function.

Corollary 5.6. *For all $\varepsilon > 0$, the stopping time N_ε defined in (5.9) is either zero, or*

$$\mathbb{E}[r_\alpha(X_{N_\varepsilon})^2] \leq \varepsilon < \frac{\max\{1, \theta^{-2}\}\mathcal{Q}(\phi, \hat{\sigma})}{N_\varepsilon}.$$

We now turn to the estimation of the oracle complexity. To this end, we have to bound the total number of data points involved in the N_ε batches needed to execute Algorithm SFBF, i.e. we want to upper bound the sum $2\sum_{i=0}^{N_\varepsilon} m_i$. Given the definition of the sample rate in (5.14), we can perform the following computation:

$$\sum_{i=1}^{N_\varepsilon+1} m_i \leq \max\{1, \theta\} \left[\ln(N_\varepsilon + \mu + 1)^{1+b} \sum_{i=1}^{N_\varepsilon+1} (i - 1 + \mu) + (N_\varepsilon + 1) \right]$$

$$\leq \max\{1, \theta\} \left[\ln(N_\varepsilon + 1 + \mu)^{1+b} \frac{(N_\varepsilon + 1)}{2} (N_\varepsilon + 2\mu) + (N_\varepsilon + 1) \right].$$

Hence,

$$2\sum_{i=1}^{N_\varepsilon} m_i \leq \max\{1, \theta\}(N_\varepsilon + 1)(N_\varepsilon + 2\mu) \left[\ln(N_\varepsilon + 1 + \mu)^{1+b} + \frac{2}{N_\varepsilon + 2\mu} \right]. \quad (5.15)$$

Proposition 5.7. *Let $\varepsilon \in (0, 1)$ be arbitrarily chosen, and $\mu \in (1, 1/\varepsilon)$. Define*

$$\mathcal{I}(\hat{\sigma}, \phi) \triangleq 3 \left(\frac{2}{\rho}\mathbb{E}[\text{dist}(X_0, \mathcal{X}_*)^2] + 2 \right)^2$$

$$+ \frac{12}{\rho^2}(1 + \bar{H}(\hat{\sigma}, n_0, \phi))^2 \mathcal{A}_{\mu,b}^2 \hat{\sigma}^4 + \frac{12}{\rho^2}(1 + \bar{H}(\hat{\sigma}, n_0, \phi))^2 \mathcal{B}_{\mu,b}^2 \hat{\sigma}^8,$$

$$\mathcal{J}(\hat{\sigma}, \phi) \triangleq \bar{\Lambda}_\infty(\hat{\sigma}, n_0, \phi) + 2.$$

If the sample rate $(m_i)_{i \geq 1}$ is given by (5.14), then we can bound the oracle complexity by

$$2\sum_{i=1}^{N_\varepsilon+1} m_i \leq \frac{2\max\{1, \theta\}\max\{1, \theta^{-4}\}\mathcal{I}(\hat{\sigma}, \phi)\left(\ln(\mathcal{J}(\hat{\sigma}, \phi)/\varepsilon)^{1+b} + \mu^{-1}\right)}{\varepsilon^2}.$$

Proof. Proof. The proof is patterned after Iusem et al. (2017). Using $N_\varepsilon < \bar{\Lambda}_\infty(\phi, \hat{\sigma})/\varepsilon$, we continue from (5.15) to obtain the bound

$$2 \sum_{i=1}^{N_\varepsilon+1} m_i \le \max\{1, \theta\} \frac{(\bar{\Lambda}_\infty(\sigma, \hat{\phi}) + 1)(\bar{\Lambda}_\infty(\hat{\sigma}, \phi) + 2)}{\varepsilon^2} \left[\ln\left(\frac{\bar{\Lambda}_\infty(\hat{\sigma}, \phi) + 2}{\varepsilon} \right)^{1+b} + \mu^{-1} \right]$$

$$\le \max\{1, \theta\} \frac{(\bar{\Lambda}_\infty(\hat{\sigma}, \phi) + 2)^2}{\varepsilon^2} \left[\ln(\varepsilon^{-1} \mathcal{J}(\hat{\sigma}, \phi))^{1+b} + \mu^{-1} \right].$$

Since,

$$(\Lambda_\infty(\hat{\sigma}, \phi) + 2)^2$$

$$\le \max\{1, \theta^{-4}\} \left\{ \frac{2}{\rho} \mathbb{E}[\text{dist}(X_0, \mathcal{X}_*)^2] + \frac{2}{\rho}(1 + \bar{H}(\hat{\sigma}, n_0, \phi)) \left[\mathcal{A}_{\mu,b} \hat{\sigma}^2 + \mathcal{B}_{\mu,b} \hat{\sigma}^4 \right] + 2 \right\}^2$$

$$\le \max\{1, \theta^{-4}\} 3 \left(\frac{2}{\rho} \mathbb{E}[\text{dist}(X_0, \mathcal{X}_*)^2] + 2 \right)^2 + \frac{12}{\rho^2}(1 + \bar{H}(\hat{\sigma}, n_0, \phi))^2 \mathcal{A}_{\mu,b}^2 \hat{\sigma}^4$$

$$+ \frac{12}{\rho^2} \max\{1, \theta^{-4}\}(1 + \bar{H}(\hat{\sigma}, n_0, \phi))^2 \mathcal{B}_{\mu,b}^2 \hat{\sigma}^8$$

$$= \max\{1, \theta^{-4}\} \mathcal{I}(\hat{\sigma}, \phi),$$

the result follows. ∎

6. Computational Experiments

We provide three examples to verify our theoretical results and compare our methods with the SEG proposed in Iusem et al. (2017). All experiments, beside Experiment 2, were generated with Matlab R2017a on a Linux OS with a 2.39 Ghz processor and 16 GB of memory. Experiment 2 was generated with Mathematica 11 on a MacBook Pro with a 2.9 Ghz processor and 16 GB memory.

6.1. **Fractional programming and applications to communication networks.** Due to its widespread use and applications, fractional programming is instrumental to operations research and engineering, ranging from network science to signal processing, wireless communications and many other related fields (Shen and Yu 2018). The standard form of a stochastic fractional program is as follows:

$$\begin{aligned} \text{minimize} \quad & f(x) = \mathbb{E}\left[\frac{G(x; \xi)}{h(x; \xi)} \right], \\ \text{subject to} \quad & x \in \mathcal{X} \end{aligned} \tag{6.1}$$

where G and h are positive and convex in x for all ξ. It is well known that such problems are pseudo-convex Boyd and Vandenberghe (2004), so they fall within the general framework of this paper. In particular, one of the cases most commonly encountered in practice is when h is linear in x and deterministic, i.e.,

$$h(x; \xi) \triangleq h(x) = a^\top x + b$$

for vectors a and b of suitable dimension. Solving this problem directly involves the pseudo-monotone operator $T(x) = \nabla f(x)$. Indeed, $x^* \in \mathcal{X}$ solves problem (6.1) if and only if x^* solves $\text{VI}(T, \mathcal{X})$.

Experiment 1 (Quadratic fractional programming). In our first experiment, we consider functions G of the form

$$G(x, \xi) = \frac{1}{2} x^\top Q(\xi) x + c(\xi)^\top x + q(\xi),$$

where $Q(\xi) \in \mathbb{R}^{d \times d}$, $c(\xi) \in \mathbb{R}^d$ and $q(\xi) \in \mathbb{R}$ are randomly generated, and Q is further assumed to be positive semi-definite. More specifically, the problem data for Q is randomly generated as follows:

$$Q = M^\top M + \text{Id},$$

where M is a random matrix of size $d \times d$ and Id is the $d \times d$ identity matrix. Finally, the vectors a and c are drawn uniformly at random from $(0, 2)^d$, q is a random number in $(1, 2)$, and $b = 1 + 4d$.

At each sample of the methods, we generate a sample matrix as

$$Q(\xi) = Q + \frac{1}{2} \left(V(\xi) + V(\xi)^T \right),$$

where $V(\xi)$ is a $d \times d$ random matrix with iid entries drawn from a normal distribution with zero mean and standard derivation $\sigma = 0.1$. Similarly,

$$c(\xi) := c + c_1(\xi), \quad q(\xi) = q + q_1(\xi), \tag{6.2}$$

where $c_1(\xi)$ and $q(\xi)$ are a random vector and a random number with zero mean and normal distribution with derivation $\sigma = 0.1$, respectively. Also, for the problem's feasible region, we consider box constraints of the form

$$\mathcal{X} = \{x \in \mathbb{R}^d : a_i \leq x_i \leq b_i \quad i = 1, ..., d\}, \tag{6.3}$$

where the lower bound a_i is a random vector in $(0, 1)^d$ and the upper bound $b_i = a_i + 10$. We have implemented SEG and SFBF for this problem, using the random operator $F(x, \xi) = \nabla_x \left(\frac{G(x, \xi)}{h(x)} \right)$. The starting point x_0 is randomly chosen in $(1, 10)^d$. Both algorithms are run with a constant step-size policy. We fix the stepsize of SFBF and SEG as $\alpha_{FBF} = 10/d$ and $\alpha_{EG} = \alpha_{FBF}/\sqrt{3}$. The step-size α_{EG} is the largest one compatible with the theory developed in Iusem et al. (2017). We choose the batch size sequence $m_{n+1} = \left\lceil \frac{(n+1)^{1.5}}{d} \right\rceil$, so that Assumption 6 is satisfied. We stop the algorithms when the residual is below a given tolerance ε. Specifically, our stopping criterion is

$$r_n \triangleq \|x_n - \Pi_{\mathcal{X}}(x_n - T(x_n))\| \leq \varepsilon = 10^{-3}.$$

Our numerical experiments involve dimension $d \in \{200, 500, 1000, 2000\}$, and for each value of d we perform 10 runs and compare the average number of iterations and CPU time. The results are displayed in Table 1 and Fig. 1. It can be seen that SFBF is constantly about 1.5 faster than SEG in both computational time and number of iterations. An interesting observation is that the number of iterations seems not to depend on the problem dimension.

Experiment 2 (Energy efficiency in multi-antenna communications). Energy efficiency is one of the most important requirements for mobile systems, and it plays a crucial role in preserving battery life and reducing the carbon footprint of multi-antenna devices (i.e., wireless devices equipped with several antennas to multiplex and demultiplex received or transmitted signals).

Table 1. Averaged over 100 runs for fractional problems of different size

d	SFBF		SEG	
	number of iterations	time(sec.)	number of iterations	time(sec.)
200	29.88	0.0473	43.96	0.0835
500	29.84	0.2647	44.49	0.3793
1000	30.14	1.1650	44.99	1.7017
2000	30.54	8.0487	45.68	11.4803
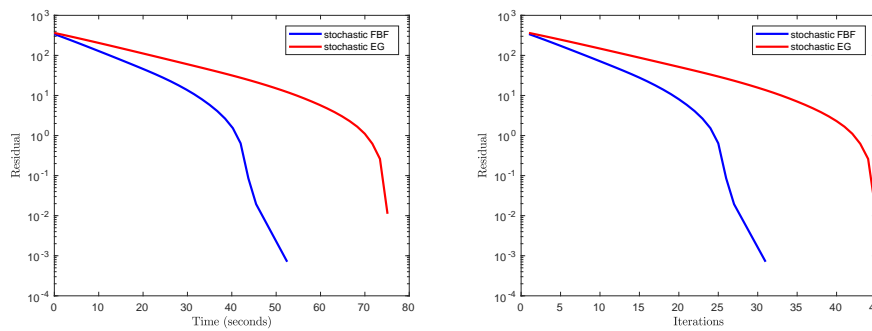
Figure 1. Comparison between SFBF and SEG for solving the fractional programming. We represent the residual vs. running time (left) and number of iterations (right) for one random example $n = 5000$.

Following Feng et al. (2013), Isheden et al. (2012), Mertikopoulos and Belmega (2016), the problem can be formulated as follows: consider K wireless devices (e.g., mobile phones), each equipped with M transmit antennas and seeking to connect to a common base-station with N receiver antennas. In this case, the users' achievable throughput (received bits/sec) is given by the familiar Shannon–Telatar capacity formula Telatar (1999):

$$R(X; H) = \log \det \left(\mathrm{Id} + \sum_{k=1}^{K} H_k X_k H_k^\dagger \right), \tag{6.4}$$

where:

(1) X_k is the $M \times M$ Hermitian *input signal covariance matrix* of user k and $X = (X_1, \ldots, X_K)$ denotes their aggregate covariance profile. As a covariance matrix, each X_k is Hermitian positive semi-definite.

(2) H_k is the $N \times M$ *channel matrix* of user k, representing the quality of the wireless medium between user k and the receiver.

(3) Id is the $N \times N$ identity matrix.

In practice, because of fading and other signal attenuation factors, the channel matrices H_k are random variables, so the users' achievable throughput is given by

$$R(X) = \mathbb{E}_H[R(X; H)], \tag{6.5}$$

where the expectation is taken over the (often unknown) law of H. The system's *energy efficiency* (EE) is then defined as the ratio of the users' achievable throughput per the unit of power consumed to achieved, i.e.,

$$\text{EE}(X) = \frac{R(X)}{\sum_{k=1}^{K}[P_k^c + P_k^t]}, \tag{6.6}$$

where

(1) P_k^t is the transmit power of the k-th device; by elementary signal processing considerations, it is given by $P_k^t = \text{tr}(X_k)$.

(2) $P_k^c > 0$ is a constant representing the total power dissipated in all circuit components of the k-th device (mixer, frequency synthesizer, digital-to-analog converter, etc.), *except* for transmission. For concision, we will also write $P^c = \sum_k P_k^c$ for the total circuit power dissipitated by the system.

The users' transmit power is further constrained by the maximum output of the transmitting device, corresponding to a trace constraint of the form

$$\text{tr}(X_k) \leq P_{\max} \quad \forall k = 1, \ldots, K. \tag{6.7}$$

Hence, putting all this together, we obtain the stochastic fractional problem:

$$\begin{aligned}
\text{maximize} \quad & \text{EE}(X) = \frac{\mathbb{E}_H[R(X; H)]}{P^c + \sum_{k=1}^{K} \text{tr}(X_k)} \\
\text{subject to} \quad & X_k \succcurlyeq 0, \\
& \text{tr}(X_k) \leq P_{\max} \quad \forall k = 1, 2, \ldots, K.
\end{aligned} \tag{6.8}$$

Note that the overall problem dimension is $d = KM^2$. The energy efficiency objective of this problem (which, formally, has units of bits/Joule) has been widely studied in the literature Cui et al. (2004), Isheden et al. (2012) and it captures the fundamental trade-off between higher spectral efficiency and increased battery life. Importantly, switching from maximization to minimization, we also see that (6.8) is of the general form (6.1), so it can be solved by applying the SFBF algorithm: in fact, given the costly projection step to the problem's feasible region, SFBF seems ideally suited to the task.

We do so in a series of numerical experiments reported in Fig. 2. Specifically, we consider a network consisting of $K = 16$ users, each with $M = 4$ transmit antennas, and a common receiver with $N = 128$ receive antennas. To simulate realistic network conditions, the users' channel matrices are drawn at each update cycle from a COST Hata radio propagation model with Rayleigh fading Hata (1980); to establish a baseline, we also ran an experiment with static, deterministic channels. For comparison purposes, we ran both SFBF and SEG with the same variance reduction schedule, the same number of iterations, and step-sizes chosen as in Experiment 1; also, to reduce statistical error, we performed $S = 100$ sample runs for each algorithm. As in the case of Experiment 1, the SFBF algorithm performs consistently better than SEG, converging to a given target value between 1.5 and 3 times faster.

6.2. **Matrix Games.** As numerical illustration we investigate the performance of the algorithm to compute Nash equilibria in random matrix games. To be specific, we revisit in this experiment the problem of computing one Nash equilibrium in random two-player
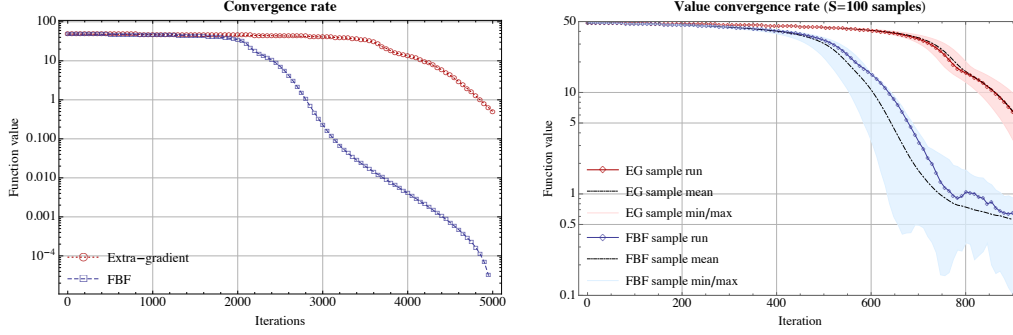
Figure 2. Comparison of the extra-gradient and forward-backward-forward (FBF) methods in the energy efficiency maximization problem (6.8). On the left, we considered static channels, and we ran SFBF and SEG with the same initialization. On the right, we considered ergodic channels following a Rayleigh fading model and we performed $S = 100$ sample runs for each algorithm; we then plotted a sample run, the sample mean, and the best and worst values at each iteration for each algorithm. In all cases, SFBF exhibits significant performance gains over SEG.

bimatrix games. A bimatrix game presented in its mixed extension consists of a tuple $\mathcal{G} = (\{I, II\}, (u_I, u_{II}), (S_I, S_{II}))$, defined by

- the set of players $\{I, II\}$;
- strategy sets $S_I \triangleq \{p \in \mathbb{R}_+^{n_I} \mid \sum_{i=1}^{n_I} p_i = 1\}, S_{II} \triangleq \{q \in \mathbb{R}_+^{n_{II}} \mid \sum_{i=1}^{n_{II}} q_i = 1\}$;
- real valued utility functions $u_I(p, y) \triangleq p^\top U_I q, u_{II}(p, q) \triangleq p^\top U_{II} q$, defined by the matrices (U_I, U_{II}), both of which are real matrices of dimension $n_I \times n_{II}$.

Recall that a pair of mixed actions (p^*, q^*) is called a Nash equilibrium of the bimatrix game (U_I, U_{II}), if

$$p_i^* > 0 \Rightarrow (U_I q)_i = \max_{1 \leq j \leq n_I} (U_I q)_j \text{ and}$$

$$q_i^* > 0 \Rightarrow (U_{II}^\top p)_i = \max_{1 \leq j \leq n_{II}} (U_{II}^\top p)_j.$$

The bimatrix game $\mathcal{G}$ is symmetric if $n_I = n_{II}$ and $U_I = U_{II}$. In symmetric games, it is natural to focus on symmetric Nash equilibria, which is a Nash equilibrium (p^*, q^*) with $p^* = q^*$.

Let $d \triangleq n_I + n_{II}$, and note that $\mathbb{R}^d \cong \mathbb{R}^{n_I} \times \mathbb{R}^{n_{II}}$, via the usual embedding of a pair (p, q) to a stacked vector in $\mathbb{R}^d$. Define the $d \times d$ matrix

$$M \triangleq \begin{bmatrix} \mathbf{0}_{n_I, n_I} & -U_I \\ -U_{II}^\top & \mathbf{0}_{n_{II}, n_{II}} \end{bmatrix}, \tag{6.9}$$

and consider the set

$$\mathcal{X} \triangleq \{(x_1, x_2) \in \mathbb{R}_+^{n_I} \times \mathbb{R}_+^{n_{II}} \mid U_I x_2 \leq \mathbf{1}_{n_I} \text{ and } U_{II}^\top x_1 \leq \mathbf{1}_{n_{II}}\}. \tag{6.10}$$

It is a classical fact that a Nash equilibrium (p^*, q^*) can be computed by finding a pair $(x_1, x_2) \neq (\mathbf{0}_{n_I}, \mathbf{0}_{n_{II}}) \in \mathcal{X}$ such that

$$x_1^\top (\mathbf{1}_{n_I} - U_I x_2) = 0, \text{ and } x_2^\top (\mathbf{1}_{n_{II}} - U_{II}^\top x_1) = 0.$$

The payoffs of the players in equilibrium can be recovered by looking at $v = \frac{1}{\sum_{j=1}^{n_I} x_{1,j}}, u = \frac{1}{\sum_{i=1}^{n_{II}} x_{2,i}}$, and the mixed actions defining equilibrium play are recovered by $p = x_1 \cdot v, q = x_2 \cdot u$. It is clear that $(\mathbf{0}_{n_I}, \mathbf{0}_{n_{II}})$ is always a solution to the *linear complementarity problem*

$$\begin{cases} x_1^\top(\mathbf{1}_{n_I} - U_I x_2) = 0, \mathbf{1}_{n_I} - U_I x_2 \geq \mathbf{0}_{n_I}, \\ x_2^\top(\mathbf{1}_{n_{II}} - U_{II}^\top x_1) = 0, \mathbf{1}_{n_{II}} - U_{II}^\top x_1 \geq \mathbf{0}_{n_{II}}. \end{cases} \tag{6.11}$$

This the so-called *artificial equilibrium* of the game, and serves as the initial point in the most used algorithm for computing Nash equilibria in bimatrix games, the Lemke-Howson algorithm, as masterly surveyed in Von Stengel (2002). Defining the mapping $T : \mathbb{R}^d \cong \mathbb{R}^{n_I} \times \mathbb{R}^{n_{II}} \to \mathbb{R}^d \cong \mathbb{R}^{n_I} \times \mathbb{R}^{n_{II}}$, by

$$T(x) \triangleq \begin{bmatrix} \mathbf{1}_{n_I} \\ \mathbf{1}_{n_{II}} \end{bmatrix} + Mx \tag{6.12}$$

we can reformulate the conditions (6.11) compactly as

$$x^* \geq \mathbf{0}_n \text{ and } T(x^*) \geq \mathbf{0}_n, \langle x^*, T(x^*) \rangle = 0. \tag{6.13}$$

To turn this into a stochastic complementarity problem, we consider a stochastic Nash game Duvocelle et al. (2018), Kannan and Shanbhag (2012), where the player set and the set of mixed actions if fixed, but the payoff functions are realizations of random matrices

$$U_I^n = U_I(\xi_n), U_{II}^n = U_{II}(\xi_n)$$

and (ξ_n) is a random process in some set Ξ, defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. For each $n \geq 1$, we look at that random operator

$$F(x, \xi_n) \triangleq \begin{bmatrix} \mathbf{1}_{n_I} \\ \mathbf{1}_{n_{II}} \end{bmatrix} + M(\xi_n)x, \tag{6.14}$$

and run Algorithm SFBF.

In our experiments, M is defined as in (6.9) and $d = n_I + n_{II}$. Each element of the matrices U_I, U_{II} is generated randomly with uniform distribution in $(0, 1)$. To setup the experiments, we generate random matrices $M(\xi) := M + V(\xi)$, where $V(\xi)$ is a $d \times d$ random matrix with zero mean and normal distribution with derivation $\sigma = 0.1$. Since the operator T is Lipschitz continuous with modulus $L = \|M\|$, we run SEG and SFBF with constant stepsizes $\alpha_{FBF} = \frac{0.99}{\sqrt{2}L}$, and $\alpha_{EG} = \frac{0.99}{\sqrt{6}L}$, respectively. We choose the batch size sequence $m_{n+1} = \left\lceil \frac{(n+1)^{1.5}}{d} \right\rceil$ so that Assumption 6 is satisfied. The same stopping criterion as in the previous experiments of Section 6.1 is used.

From the numerical experiments, we observe that the SFBF outperforms the SEG, being on average 1.7 times faster in computational time and 1.5 times faster in number of iterations. The difference becomes larger as the problem dimension increases. There are two reasons for results: firstly, SEG requires two projections per iteration while SFBF only requires one and more importantly, the stepsize of SFBF is $\sqrt{3}$ times larger than that of SEG.

Experiment 3 (Zero-Sum games). We compare the performance SFBF and SEG for zero sum game, i.e., $U_I = -U_{II}^T$. The results are displayed in Table 2, showing the advantage of SFBF over SEG. On average, SFBF is 1.7 times faster in computational time and 3.4 times faster in number of iterations than SEG.

Table 2. Averaged over 100 runs for zero sum game of different size

Dimension	SFBF		SEG	
$d = n_I + n_{II}$	Iterations	time(sec.)	Iterations	time(sec.)
$n_I = n_{II} = 100$	84.38	0.4421	172.42	1.4768
$n_I = n_{II} = 250$	214.09	9.2088	372.80	32.4321
$n_I = n_{II} = 500$	430.18	73.9068	749.65	270.5911
$n_I = n_{II} = 1000$	865.67	672.0806	1508.50	2535.50

Table 3. Averaged over 100 runs for symmetric game of different size

Dimension	SFBF		SEG	
$d = n_I + n_{II}$	Iterations	time(sec.)	Iterations	time(sec.)
$n_I = n_{II} = 100$	52.00	0.3882	68.68	0.6293
$n_I = n_{II} = 250$	97.96	2.589	142.55	5.1276
$n_I = n_{II} = 500$	173.30	10.5297	247.30	21.0797
$n_I = n_{II} = 1000$	319.92	92.0417	455.48	191.6854

Table 4. Averaged over 100 runs for asymmetric game of different size

	SFBF		SEG	
$d = n_I + n_{II}$	Iterations	time(sec.)	Iterations	time(sec.)
$n_I = 100, n_{II} = 200$	100.28	1.9553	155.28	4.8202
$n_I = 300, n_{II} = 600$	293.36	32.3010	466.01	90.2339
$n_I = 500, n_{II} = 1000$	492.21	136.7019	779.86	394.7606
$n_I = 1000, n_{II} = 2000$	992.64	1597.7266	1564.12	4655.9213

Experiment 4 (Symmetric game). We compare the performance SFBF and SEG for symmetric game, i.e., U_I, U_{II} are symmetric and $U_I = U_{II}^T$. We choose $n_I = n_{II} \in \{50, 100, 150, \ldots, 500\}$ and $d = n_I + n_{II}$. The results are displayed in Table 3 showing the advantage of SFBF over SEG. On average, sFBF is 1.4 times faster in computational time and 1.8 times faster in number of iterations.

Experiment 5 (Bimatrix Games). We compare the performance SFBF and SEG for asymmetric game. We choose $n_I \in \{100, 200, \ldots, 1000\}$ and $n_{II} = 2n_I$. The results are displayed in Table 4 and Fig. 3, showing the advantage of SFBF over SEG.

7. Conclusion

In this paper we have developed a stochastic version of Tseng's forward-backward-forward algorithm for solving stochastic variational inequality problems over nonempty
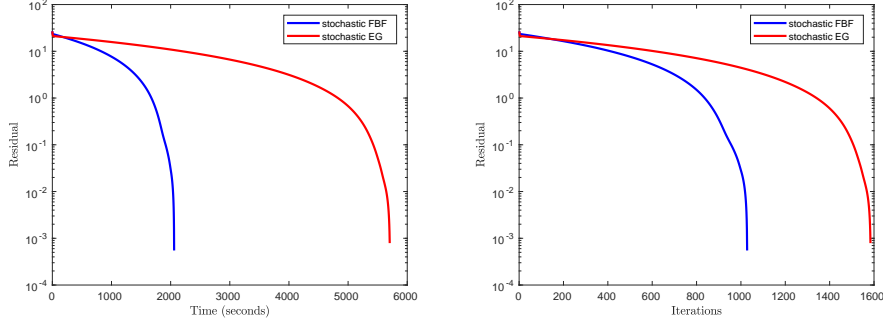
Figure 3. Comparison between SFBF and SEG for solving the asymmetric game. We represent the Residual vs. running time (left) and number of iterations (right) for one random example $n_I = 1000, n_{II} = 2000$.

closed and convex sets. As in Iusem et al. (2017), the current analysis can be generalized to Cartesian VI problems, though have not done this explicitly. We show that the known theoretical convergence guarantees of SEG carry over to this setting, but our method consistently outperforms SEG in terms of convergence rate and complexity. We therefore believe that SFBF is a serious competitor to SEG in typical primal-dual settings, where feasibility is a minor issue. Interesting directions for the future are to test the performance of the method in other instances where variance reduction is of importance, such as in composite optimization involving a large but finite sum of functions. Another possible extension would be to develop an infinite-dimensional Hilbert space version of the algorithm, and modify the basic SFBF scheme to allow for inertial and relaxation effects. We will investigate these, and other issues, in the future.

APPENDIX A. PROOF OF LEMMA 4.2

We start with a general result. Let $N \in \mathbb{N}$ and $\xi^{(1)}, \ldots, \xi^{(N)}$ be an i.i.d sample from the measure P. Define the process $\left(M_i^N(x)\right)_{i=0}^N$ by $M_0(x) \triangleq 0$, and for $1 \leq i \leq N$, by

$$M_i^N(x) \triangleq \frac{1}{N} \sum_{n=1}^i \left(F(x, \xi^{(n)}) - T(x) \right) \qquad \forall x \in \mathbb{R}^d. \tag{A.1}$$

Setting $\mathcal{G}_i \triangleq \sigma(\xi^{(1)}, \ldots, \xi^{(i)}), 1 \leq i \leq N$, we see that the process $\{(M_i^N(x), \mathcal{G}_i), 1 \leq i \leq N\}$ is a martingale starting at zero.

Lemma A.1. *Let $p \geq 2$ be as specified in Assumption 7. For all $1 \leq q \leq p, N \in \mathbb{N}$ and $x \in \mathbb{R}^d$, we have*

$$\mathbb{E}\left[\|M_N^N(x)\|^q \right]^{\frac{1}{q}} \leq \frac{C_q}{\sqrt{N}} (\sigma(x^*) + \sigma_0 \|x - x^*\|). \tag{A.2}$$

Proof. Proof. For $i \in \{1, 2, \ldots, N\}$, the monotonicity of $L^p(\mathbb{P})$ norms implies that

$$\mathbb{E}\left[\|\Delta M_{i-1}^N(x)\|^q \right]^{\frac{1}{q}} = \frac{1}{N} \mathbb{E}\left[\|F(x, \xi^{(i)}) - T(x)\|^q \right]^{\frac{1}{q}}$$

$$\leq \frac{1}{N} \mathbb{E}\left[\|F(x, \xi^{(i)}) - T(x)\|^p \right]^{\frac{1}{p}}$$

$$\leq \frac{\sigma(x^*) + \sigma_0 \|x - x^*\|}{N}.$$

Using this, together with Lemma 2.4, we get

$$\mathbb{E}\left[\|M_N^N(x)\|^q\right]^{1/q} \leq C_q \sqrt{\sum_{k=1}^{N} \mathbb{E}\left(\left\|\frac{F(x, \xi^{(k)}) - T(x)}{N}\right\|^q\right)^{2/q}}$$

$$\leq C_q \sqrt{N^{-2} \sum_{k=1}^{N} \mathbb{E}\left(\|F(x, \xi^{(k)}) - T(x)\|^q\right)^{2/q}}$$

$$\leq \frac{C_q(\sigma(x^*) + \sigma_0 \|x - x^*\|)}{\sqrt{N}}.$$

∎

Proof. Proof of Lemma (4.2). Observe that $M_{m_{n+1}}^{m_{n+1}}(X_n) = W_{n+1}$ and $M_{m_{n+1}}^{m_{n+1}}(Y_n) = Z_{n+1}$. Hence, we immediately obtain from Lemma A.1 that

$$\mathbb{E}\left[\|W_{n+1}\|^{p'}|\mathcal{F}_n\right]^{1/p'} \leq \frac{C_{p'}(\sigma(x^*) + \sigma_0\|X_n - x^*\|)}{\sqrt{m_{n+1}}}. \tag{A.3}$$

To prove (4.11), we notice that Lemma A.1 implies that

$$\mathbb{E}\left[\|Z_{n+1}\|^{p'}|\hat{\mathcal{F}}_n\right]^{1/p'} \leq \frac{C_{p'}(\sigma(x^*) + \sigma_0\|Y_n - x^*\|)}{\sqrt{m_{n+1}}}. \tag{A.4}$$

The tower property of conditional expectations (recall that $\mathcal{F}_n \subseteq \hat{\mathcal{F}}_n$) gives

$$\mathbb{E}\left[\|Z_{n+1}\|^{p'}|\mathcal{F}_n\right] = \mathbb{E}\left\{\mathbb{E}[\|Z_{n+1}\|^{p'}|\hat{\mathcal{F}}_n]|\mathcal{F}_n\right\}$$

$$\leq \left(\frac{C_{p'}}{\sqrt{m_{n+1}}}\right)^{p'} \mathbb{E}\left[(\sigma(x^*) + \sigma_0\|Y_n - x^*\|)^{p'}|\mathcal{F}_n\right].$$

Finally, by the Minkowski inequality, we get

$$\mathbb{E}\left[\|Z_{n+1}\|^{p'}|\mathcal{F}_n\right]^{1/p'} \leq \frac{C_{p'}}{\sqrt{m_{n+1}}}\left(\sigma(x^*) + \sigma_0\mathbb{E}\left[\|Y_n - x^*\|^{p'}|\mathcal{F}_n\right]^{1/p'}\right),$$

and our proof is complete. ∎

Acknowledgments

References

Atchadé YF, Fort G, Moulines E (2017) On perturbed proximal gradient algorithms. *J. Mach. Learn. Res* 18(1):310–342.

Bottou L, Curtis FE, Nocedal J (2018) Optimization methods for large-scale machine learning. *SIAM Review* 60(2):223–311.

Boyd S, Parikh N, Chu E, Peleato B, Eckstein J (2011) Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning* 3(1):1–122.

Boyd SP, Vandenberghe L (2004) *Convex Optimization* (Cambridge, UK: Cambridge University Press).

Chen C, Chen Y, Ouyang Y, Pasiliao E (2018) Stochastic accelerated alternating direction method of multipliers with importance sampling. *Journal of Optimization Theory and Applications* 179(2):676–695, URL http://dx.doi.org/10.1007/s10957-018-1270-0.

Combettes P, Pesquet J (2015) Stochastic quasi-fejér block-coordinate fixed point iterations with random sweeping. *SIAM Journal on Optimization* 25(2):1221–1248, URL http://dx.doi.org/10.1137/140971233.

Cottle RW, Yao JC (1992) Pseudo-monotone complementarity problems in hilbert space. *Journal of Optimization Theory and Applications* 75(2):281–295, ISSN 1573-2878, URL https://doi.org/10.1007/BF00941468.

Cui S, Goldsmith AJ, Bahai A (2004) Energy-efficiency of MIMO and cooperative MIMO techniques in sensor networks. *IEEE J. Sel. Areas Commun.* 22(6):1089–1098.

Dang CD, Lan G (2015) On the convergence properties of non-euclidean extragradient methods for variational inequalities with generalized monotone operators. *Computational Optimization and Applications* 60(2):277–310, URL http://dx.doi.org/10.1007/s10589-014-9673-9.

Duflo M (1996) *Algorithmes Stochastiques* (New York: Springer).

Duvocelle B, Mertikopoulos P, Staudigl M, Vermeulen D (2018) Learning in time-varying games. *arXiv preprint arXiv:1809.03066* .

Facchinei F, Pang Js (2003) *Finite-Dimensional Variational Inequalities and Complementarity Problems - Volume I and Volume II* (Springer Series in Operations Research).

Feng D, Jiang C, Lim G, Cimini Jr LJ, Feng G, Li GY (2013) A survey of energy-efficient wireless communications. *IEEE Communications Surveys & Tutorials* 15(1):167–178.

Hata M (1980) Empirical formula for propagation loss in land mobile radio services. *IEEE Trans. Veh. Technol.* 29(3):317–325.

Isheden C, Chong Z, Jorswieck E, Fettweis G (2012) Framework for link-level energy efficiency optimization with informed transmitter. *IEEE Trans. Wireless Commun.* 11(8):2946–2957.

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

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

Jofré A, Thompson P (2018) On variance reduction for stochastic smooth convex optimization with multiplicative noise. *Mathematical Programming* URL http://dx.doi.org/10.1007/s10107-018-1297-x.

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

Kannan A, Shanbhag U (2012) Distributed computation of equilibria in monotone nash games via iterative regularization techniques. *SIAM Journal on Optimization* 22(4):1177–1205, URL http://dx.doi.org/10.1137/110825352.

Kannan A, Shanbhag UV (2019) Optimal stochastic extragradient schemes for pseudomonotone stochastic variational inequality problems and their variants. *Computational Optimization and Applications* 74(3):779–820, ISSN 1573-2894, URL https://doi.org/10.1007/s10589-019-00120-x.

King AJ, Rockafellar RT (1993) Asymptotic theory for solutions in statistical estimation and stochastic programming. *Mathematics of Operations Research* 18(1):148–162.

Korpelevich GM (1976) The extragradient method for finding saddle points and other problems. *Èkonom. i Mat. Metody* 12:747–756.

Kushner HJ, Yin GG (1997) *Stochastic Approximation Algorithms and Applications* (New York: Springer).

Mertikopoulos P, Belmega EV (2016) Learning to be green: Robust energy efficiency maximization in dynamic MIMO-OFDM systems. *IEEE J. Sel. Areas Commun.* 34(4):743 – 757.

Mertikopoulos P, Staudigl M (2018) Stochastic mirror descent dynamics and their convergence in monotone variational inequalities. *Journal of Optimization Theory and Applications* 179(3):838–867.

Mertikopoulos P, Zhou Z (2018) Learning in games with continuous action sets and unknown payoff functions. *Mathematical Programming* URL http://dx.doi.org/10.1007/s10107-018-1254-8.

Polyak BT (1987) *Introduction to Optimization* (New York, NY, USA: Optimization Software).

Ravat U, Shanbhag U (2011) On the characterization of solution sets of smooth and nonsmooth convex stochastic nash games. *SIAM Journal on Optimization* 21(3):1168–1199, URL http://dx.doi.org/10.1137/100792644.

Rosasco L, Villa S, Vũ BC (2016) Stochastic forward–backward splitting for monotone inclusions. *Journal of Optimization Theory and Applications* 169(2):388–406.

Scutari G, Palomar DP, Facchinei F, Pang Js (2010) Convex optimization, game theory, and variational inequality theory. *IEEE Signal Processing Magazine* 27(3):35–49

Shapiro A, Dentcheva D, Ruszczyński AX (2009) *Lectures on stochastic programming: modeling and theory* (SIAM).

Shen K, Yu W (2018) Fractional programming for communication systems—part i: Power control and beamforming. *IEEE Transactions on Signal Processing* 66(10):2616–2630.

Solodov M, Svaiter B (1999) A new projection method for variational inequality problems. *SIAM Journal on Control and Optimization* 37(3):765–776, URL http://dx.doi.org/10.1137/S0363012997317475.

Stroock DW (2011) *Probability Theory: An Analytic View* (Cambridge: Cambridge University Press), 2nd edition, ISBN 0-521-43123-9.

Telatar IE (1999) Capacity of multi-antenna Gaussian channels. *European Transactions on Telecommunications and Related Technologies* 10(6):585–596.

Tseng P (2000) A modified forward-backward splitting method for maximal monotone mappings. *SIAM Journal on Control and Optimization* 38(2):431–446, URL http://dx.doi.org/10.1137/S0363012998338806.

Von Stengel B (2002) Computing equilibria for two-person games. *Handbook of game theory with economic applications* 3:1723–1759.

Yousefian F, Nedić A, Shanbhag UV (2017) On smoothing, regularization, and averaging in stochastic approximation methods for stochastic variational inequality problems. *Mathematical Programming* 165(1):391–431, URL http://dx.doi.org/10.1007/s10107-017-1175-y.