

## Stochastic Systems

Publication details, including instructions for authors and subscription information:  
<http://pubsonline.informs.org>

### Minibatch Forward-Backward-Forward Methods for Solving Stochastic Variational Inequalities

Radu Ioan Boț, Panayotis Mertikopoulos, Mathias Staudigl, Phan Tu Vuong

To cite this article:

Radu Ioan Boț, Panayotis Mertikopoulos, Mathias Staudigl, Phan Tu Vuong (2021) Minibatch Forward-Backward-Forward Methods for Solving Stochastic Variational Inequalities. Stochastic Systems

Published online in Articles in Advance 25 Feb 2021

. <https://doi.org/10.1287/stsy.2019.0064>

Full terms and conditions of use: <https://pubsonline.informs.org/Publications/Librarians-Portal/PubsOnLine-Terms-and-Conditions>

This article may be used only for the purposes of research, teaching, and/or private study. Commercial use or systematic downloading (by robots or other automatic processes) is prohibited without explicit Publisher approval, unless otherwise noted. For more information, contact [permissions@informs.org](mailto:permissions@informs.org).

The Publisher does not warrant or guarantee the article's accuracy, completeness, merchantability, fitness for a particular purpose, or non-infringement. Descriptions of, or references to, products or publications, or inclusion of an advertisement in this article, neither constitutes nor implies a guarantee, endorsement, or support of claims made of that product, publication, or service.

Copyright © 2021, The Author(s)

Please scroll down for article—it is on subsequent pages







With 12,500 members from nearly 90 countries, INFORMS is the largest international association of operations research (O.R.) and analytics professionals and students. INFORMS provides unique networking and learning opportunities for individual professionals, and organizations of all types and sizes, to better understand and use O.R. and analytics tools and methods to transform strategic visions and achieve better outcomes.

For more information on INFORMS, its publications, membership, or meetings visit <http://www.informs.org>

# Minibatch Forward-Backward-Forward Methods for Solving Stochastic Variational Inequalities

Radu Ioan Boţ,<sup>a</sup> Panayotis Mertikopoulos,<sup>b,c</sup> Mathias Staudigl,<sup>d</sup> Phan Tu Vuong<sup>e</sup>

<sup>a</sup> Faculty of Mathematics, University of Vienna, 1090 Wien, Austria; <sup>b</sup> University Grenoble Alpes, French National Centre for Scientific Research (CNRS), National Institute for Research in Digital Science and Technology (Inria), Grenoble Institute of Technology (INP), 38000 Grenoble, France; <sup>c</sup> Criteo AI Laboratory, 38130 Echirolles, France; <sup>d</sup> Department of Data Science and Knowledge Engineering, Maastricht University, Maastricht, 6229 GT Maastricht, Netherlands; <sup>e</sup> Mathematical Sciences, University of Southampton, Southampton SO17 1BJ, United Kingdom

**Contact:** radu.bot@univie.ac.at,  <https://orcid.org/0000-0002-4469-314X> (RIB); panayotis.mertikopoulos@imag.fr,  <https://orcid.org/0000-0003-2026-9616> (PM); m.staudigl@maastrichtuniversity.nl,  <https://orcid.org/0000-0003-2481-0019> (MS); t.v.phan@soton.ac.uk,  <https://orcid.org/0000-0002-1474-994X> (PTV)

**Received:** October 29, 2020

**Revised:** June 14, 2020

**Accepted:** August 17, 2020

**Published Online in Articles in Advance:** February 25, 2021

<https://doi.org/10.1287/stsy.2019.0064>

**Copyright:** © 2021 The Author(s)

**Abstract.** We develop a new stochastic algorithm for solving pseudomonotone stochastic variational inequalities. Our method builds on Tseng’s forward-backward-forward 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 pseudomonotone 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 minibatch sampling mechanism and leads to almost sure 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.



**Open Access Statement:** This work is licensed under a Creative Commons Attribution 4.0 International License. You are free to copy, distribute, transmit and adapt this work, but you must attribute this work as “*Stochastic Systems*. Copyright © 2021 The Author(s). <https://doi.org/10.1287/stsy.2019.0064>, used under a Creative Commons Attribution License: <https://creativecommons.org/licenses/by/4.0/>.”

**Funding:** R. I. Boţ and P. T. Vuong acknowledge support from the Austrian Science Fund [Project I2419-N32: “Employing Recent Outcomes in Proximal Theory Outside the Comfort Zone”]. P. M. is grateful for financial support from the French National Research Agency (ANR) in the framework of the Investissements d’avenir program [ANR-15-IDEX-02], the LabEx PERSYVAL [ANR-11-LABX-0025-01], and MIAI@Grenoble Alpes [ANR-19-P3IA-0003]. M. Staudigl and P. Mertikopoulos have been sponsored by the COST Action CA16228 “European Network for Game Theory.”

**Keywords:** variational inequalities • stochastic approximation • forward-backward-forward algorithm • minibatch

## 1. Introduction

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

$$\langle T(x^*), x - x^* \rangle \geq 0, \quad \text{for all } x \in \mathcal{X}. \quad (1)$$

We call  $S(T, \mathcal{X}) \equiv \mathcal{X}_*$  the set of (Stampacchia) solutions of  $VI(T, \mathcal{X})$ . The variational inequality problem (1) arises in many interesting applications in economics, game theory, and engineering (Scutari et al. 2010, Juditsky et al. 2011, Ravat and Shanbhag 2011, Kannan and Shanbhag 2012, Mertikopoulos and Staudigl 2018) 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 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  $\xi : (\Omega, \mathcal{F}) \rightarrow (\Xi, \mathcal{A})$  taking values in a measurable space  $(\Xi, \mathcal{A})$  and inducing a law  $\mathbf{P} = \mathbb{P} \circ \xi^{-1}$ . Given the random element  $\xi$ , consider the measurable mapping  $F : \mathcal{X} \times \Xi \rightarrow \mathbb{R}^d$ , defining an integrable random vector  $F(x, \xi) : \Omega \rightarrow \mathbb{R}^d$  via the composition  $F(x, \xi)(\omega) = F(x, \xi(\omega))$ . The stochastic variational inequality problem on which we focus in this paper is denoted by  $SVI$  and defined as follows.

**Definition 1.** Let the operator  $T : \mathbb{R}^d \rightarrow \mathbb{R}^d$  be defined by

$$T(x) := \mathbb{E}_\xi[F(x, \xi)] := \int_\Omega F(x, \xi(\omega)) d\mathbb{P}(\omega) = \int_\Xi F(x, z) d\mathbf{P}(z). \quad (2)$$

Find  $x^* \in \mathcal{X}$  satisfying (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 (2) is 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 because of 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, two methodologies are currently available. The sample average approximation 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 uses 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. This approach is the standard choice in expected residual minimization problems when  $\mathbf{P}$  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  $\mathbf{P}$  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

$$X_{n+1} = \Pi_{\mathcal{X}}[X_n + \alpha_n F(X_n, \xi_n)] \quad (3)$$

is considered, where  $\Pi_{\mathcal{X}}$  denotes the Euclidean projection onto  $\mathcal{X}$ ,  $(\xi_n)_{n \geq 0}$  is a sample of  $\mathbf{P}$ , 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 that  $T$  is Lipschitz continuous and strongly monotone, and the stochastic error is uniformly bounded in mean square. Relaxing strong monotonicity to plain monotonicity, the paper by Yousefian et al. (2017) incorporated a Tikhonov regularization scheme into the SA algorithm (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 *pseudomonotonicity* of the mean operator is the extragradient approach of Iusem et al. (2017) and Kannan and Shanbhag (2019). The original 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 SO case, we arrive at the Stochastic Extragradient (SEG) method

$$\begin{aligned} Y_n &= \Pi_{\mathcal{X}}[X_n - \alpha_n A_{n+1}], \\ X_{n+1} &= \Pi_{\mathcal{X}}[X_n - \alpha_n B_{n+1}], \end{aligned} \quad (4)$$

where  $(A_n)_{n \geq 1}, (B_n)_{n \geq 1}$  are stochastic estimators of  $T(X_n)$  and  $T(Y_n)$ , respectively. Iusem et al. (2017) construct these estimators by relying on a dynamic sampling strategy, where noise reduction of the estimators is achieved via a *minibatch sampling* of the stochastic operators  $F(X_n, \xi)$  and  $F(Y_n, \xi)$ . Within this minibatch 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.<sup>1</sup>

### 1.1. Our Contribution

We briefly summarize the main contributions of this work. The most costly part of SEG is 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, whereas 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 because 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. Although 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 later in this paper. Additionally, the 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 evidence 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  $\ell_p$  norm on  $\mathbb{R}^d$  is defined for  $x = (x_1, \dots, x_p)$  as  $\|x\|_p := (\sum_{i=1}^d |x_i|^p)^{\frac{1}{p}}$ . For a nonempty, closed, and convex set  $E \subseteq \mathbb{R}^d$ , the Euclidean projector is defined as  $\Pi_E(x) := \arg \min_{y \in E} \|y - x\|$  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  $f : \Omega \rightarrow E$ ; 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  $f_1, \dots, f_k \in L^p(\Omega, \mathcal{F}, \mathbb{P}; E)$ , we denote the sigma-algebra generated by these random variables by  $\sigma(f_1, \dots, f_k)$ ; this is the smallest sigma-algebra measuring the random variables  $f_1, \dots, f_k$ . 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

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

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

**Lemma 1.** Let  $K \subseteq \mathbb{R}^d$  be a nonempty, closed, and convex set. Then

- The Euclidean orthogonal projection  $\Pi_K(x)$  is the unique point of  $K$  satisfying  $\langle x - \Pi_K(x), y - \Pi_K(x) \rangle \leq 0$  for all  $y \in K$ ;
- For all  $x \in \mathbb{R}^d$  and  $y \in K$ , we have  $\|\Pi_K(x) - y\|^2 + \|\Pi_K(x) - x\|^2 \leq \|x - y\|^2$ ;
- For all  $x, y \in \mathbb{R}^d$ ,  $\|\Pi_K(x) - \Pi_K(y)\| \leq \|x - y\|$ ;
- Given  $\alpha > 0$  and  $T : K \rightarrow \mathbb{R}^d$ , the set of solutions of the variational problem  $\text{VI}(T, K)$  can be expressed as  $S(T, K) = \{x \in \mathbb{R}^d \mid x = \Pi_K(x - \alpha T(x))\}$ .

**Remark 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  $\text{VI}(T, K)$ , defined by inequality (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** (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

$$\mathbb{E}[\|f + g\|^p | \mathcal{G}]^{1/p} \leq \mathbb{E}[\|f\|^p | \mathcal{G}]^{1/p} + \mathbb{E}[\|g\|^p | \mathcal{G}]^{1/p}. \quad (5)$$

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

**Lemma 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  $n \geq 0$ ,

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

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 (Stroock 2011).

**Lemma 4.** Let  $(\Omega, \mathcal{F}, (\mathcal{F}_n)_{n \geq 0}, \mathbb{P})$  be a discrete stochastic basis, and let  $(U_n)_{n \geq 0}$  be 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  $N \geq 1$ ,

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

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

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

### 3. 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 nonempty

**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  $F : \mathbb{R}^d \times \Xi \rightarrow \mathbb{R}^d$  is a Carathéodory map.<sup>2</sup>

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

**Assumption 4** (Pseudomonotonicity). The averaged operator  $T(\cdot) = \mathbb{E}_\xi[F(\cdot, \xi)]$  is pseudomonotone on  $\mathbb{R}^d$ , which means that

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

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

$$\hat{T}_{n+1}(x, \xi_{n+1}) \triangleq \frac{1}{m_{n+1}} \sum_{i=1}^{m_{n+1}} F(x, \xi_{n+1}^{(i)}), \quad \text{for } x \in \mathbb{R}^d. \quad (6)$$

The sequence  $(m_n)_{n \geq 1} \subseteq \mathbb{N}$  determines the batch size of the stochastic oracle. The random sequence  $\xi_n = (\xi_n^{(1)}, \dots, \xi_n^{(m_n)})$  is an independent and identically distributed (i.i.d.) draw from  $\mathbb{P}$ . Approximations of the form (6) are very common in Monte Carlo simulation approaches, machine learning, and computational statistics (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  $\mathbb{P}$ . The forward-backward-forward algorithm requires two queries from the SO in which minibatch 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 minibatch consists of two independent stochastic processes  $\xi_n$  and  $\eta_n$  drawn from the law  $\mathbb{P}$  and explicitly given by

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

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** (SFBF)

**Require:** Step-size sequence  $\alpha_n$ ; batch-size sequence  $m_n$ .

```

1: Initialize  $X$  # initialization.
2: for  $n = 1, 2, \dots$ , do
3:   Draw samples  $\xi^i$  and  $\eta^i$  from  $P$  ( $i = 1, \dots, m_n$ ).
4:   Oracle returns  $A \leftarrow \frac{1}{m_n} \sum_{i=1}^{m_n} F(X, \xi^i)$  # first oracle query.
5:   Set  $Y \leftarrow \Pi_{\mathcal{X}}(X - \alpha_n A)$  # forward-backward step
6:   Oracle returns  $B \leftarrow \frac{1}{m_n} \sum_{i=1}^{m_n} F(Y, \eta^i)$  # 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. In the stochastic optimization case, that is, 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 such as the Alternating Direction Method of Multipliers (ADMM) (Boyd et al. 2011, Chen et al. 2018) are infeasible by nature as well. In concrete applications, the infeasibility of Algorithm 1 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 3. Moreover, Theorem 1 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 1 satisfies*

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

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

$$W_{n+1} \triangleq A_{n+1} - T(X_n) \text{ and } Z_{n+1} \triangleq B_{n+1} - T(Y_n), \quad (7)$$

and the sub-sigma-algebras  $(\mathcal{F}_n)_{n \geq 0}, (\hat{\mathcal{F}}_n)_{n \geq 0}$ , defined by  $\mathcal{F}_0 \triangleq \sigma(X_0)$ ;  $\mathcal{F}_n \triangleq \sigma(X_0, \xi_1, \xi_2, \dots, \xi_n, \eta_1, \dots, \eta_n)$ , for all  $n \geq 1$ ; and  $\hat{\mathcal{F}}_n \triangleq \sigma(X_0, \xi_1, \dots, \xi_n, \xi_{n+1}, \eta_1, \dots, \eta_n)$ , 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, and unbiased approximations relative to the respective information structures are provided:

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

**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  $c > 0$  and integer  $n_0 > 0$ , we have

$$m_n = c \cdot (n + n_0)^{1+a} \ln(n + n_0)^{1+b} \quad (8)$$

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  $\sigma : \mathcal{X}_* \rightarrow \mathbb{R}_+$  such that for all  $x \in \mathbb{R}^d$  and all  $x^* \in \mathcal{X}_*$ ,

$$s_p(x) \leq \sigma(x^*) + \sigma_0 \|x - x^*\|. \quad (9)$$

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 a uniformly bounded variance (UBV), that is,

$$\sup_{x \in \mathcal{X}} s_2(x) \leq \sigma. \quad (10)$$

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

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

However, assuming a global variance bound is not realistic in cases where the variance of the SO depends on the position  $x$  (Jofré and Thompson 2019, example 1). Assumption 7 is much weaker than UBV because it exploits the local variance of the SO rather than (potentially hard to estimate) global mean square variance bounds. Recent papers (Iusem et al. 2017, Jofré and Thompson 2019) make similar assumptions on the variance of the SO. 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 in the following example.

**Example 1.** Assume for the Carathéodory map  $F : \mathbb{R}^d \times \Xi \rightarrow \mathbb{R}^d$  that there exists  $\mathcal{L} \in L^p(\Omega, \mathcal{F}, \mathbb{P}; \mathbb{R}_+)$  with

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

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

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

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

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

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 VI( $T, \mathcal{X}$ ). Moreover,  $\{r_a(x); a > 0\}$  is a family of equivalent merit functions for 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

$$\rho_n \triangleq 1 - 2L^2\alpha_n^2, \quad \forall n \geq 0. \quad (11)$$

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

$$\Delta V_n \triangleq V_{n+1} - V_n = (4 + \rho_n)\alpha_n^2 \|W_{n+1}\|^2 + 4\alpha_n^2 \|Z_{n+1}\|^2, \quad \forall n \geq 0. \quad (12)$$

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

$$U_{n+1}(x) \stackrel{\Delta}{=} U_n(x) + 2\alpha_n \langle Z_{n+1}, x - Y_n \rangle, \quad \forall n \geq 1,$$

with corresponding increment

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

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  $n \geq 0$ ; that is, the process  $(U_n(x))_{n \geq 0}$  is a martingale with respect to the filtration  $\hat{\mathbb{F}}$ . Because  $\mathcal{F}_n \subseteq \hat{\mathcal{F}}_n$ , the tower property implies that

$$\mathbb{E}[\Delta U_n(x) | \mathcal{F}_n] = 0, \quad \forall x \in \mathbb{R}^d \quad \forall n \geq 0, \quad (13)$$

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

**Lemma 5.** *For all  $x^* \in \mathcal{X}_*$  and all  $n \geq 0$ , we have*

$$\|X_{n+1} - x^*\|^2 \leq \|X_n - x^*\|^2 - \frac{\rho_n}{2} r_{\alpha_n}(X_n)^2 + \Delta U_n(x^*) + \Delta V_n, \quad \mathbb{P}\text{-a.s.} \quad (14)$$

**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

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

Using that  $\alpha_n > 0$  and the pseudomonotonicity of  $T$ , we see that

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

Using the Doob decomposition in Equation (7), we can rewrite this inequality as

$$\langle \alpha_n B_{n+1}, Y_n - x^* \rangle \geq \alpha_n \langle Z_{n+1}, Y_n - x^* \rangle. \quad (15)$$

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

$$\langle x^* - Y_n, Y_n - X_n + \alpha_n A_{n+1} \rangle \geq 0. \quad (16)$$

Adding (15) and (16) 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

$$\langle x^* - Y_n, X_{n+1} - X_n \rangle \geq \alpha_n \langle Z_{n+1}, Y_n - x^* \rangle. \quad (17)$$

**Step 2.** Using (17), we get

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

where we have used the definition of  $X_{n+1}$  in the preceding equality. The Pythagorean identity Lemma 2 gives us

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



**Step 3.** Using again the definition of  $X_{n+1}$ , we see that

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

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

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

**Step 4.** By the definition of the squared residual function, the definition of  $Y_n$  and Lemma 1(c), 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,

$$-2\|X_n - Y_n\|^2 \leq 2\alpha_n^2 \|W_{n+1}\|^2 - r_{\alpha_n}(X_n)^2. \quad (18)$$

**Step 5.** Combining (18) with the last inequality from Step 3 and recalling Assumption 5, we conclude that

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

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

**Remark 2.** One can notice that in the preceding proof that the pseudomonotonicity of  $T$  is used only in Step 1 in order to obtain relation (15). Thus, as happened in Dang and Lan (2015) and Solodov and Svaiter (1999), the pseudomonotonicity of  $T$  can actually be replaced by the weaker assumption

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

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 the conditional expectations in (14) and using the martingale property (13), we see that for all  $n \geq 0$ ,

$$\mathbb{E}[\|X_{n+1} - x^*\|^2 | \mathcal{F}_n] \leq \|X_n - x^*\|^2 - \frac{\rho_n}{2} r_{\alpha_n}(X_n)^2 + \mathbb{E}[\Delta V_n | \mathcal{F}_n]. \quad (19)$$

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

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

Let  $q \in [1, \infty]$ . 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 the Minkowski inequality,

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

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

**Lemma 6.** *Let  $p' \in [2, p]$ . For all  $n \geq 0$ , we have  $\mathbb{P}$ -a.s.*

$$\mathbb{E}[\|W_{n+1}\|^{p'} | \mathcal{F}_n]^{\frac{1}{p'}} \leq \frac{C_{p'}(\sigma(x^*) + \sigma_0 \|X_n - x^*\|)}{\sqrt{m_{n+1}}} \quad (20)$$

and

$$\mathbb{E}[\|Z_{n+1}\|^{p'} | \mathcal{F}_n]^{\frac{1}{p'}} \leq \frac{C_{p'}}{\sqrt{m_{n+1}}} \left( \sigma(x^*) + \sigma_0 \mathbb{E}[\|Y_n - x^*\|^{p'} | \mathcal{F}_n]^{\frac{1}{p'}} \right). \quad (21)$$

In particular, in case of (10) 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.** See Appendix A.  $\square$

Let  $p' \geq 2$  and  $n \geq 0$ . Then we have

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

Hence, combining this with (20) for  $p' \in [2, p]$  as in Lemma 6, we see that

$$\begin{aligned} \mathbb{E}[\|Y_n - x^*\|^{p'} | \mathcal{F}_n]^{\frac{1}{p'}} &\leq (1 + \alpha_n L) \|X_n - x^*\| + \alpha_n \frac{C_{p'}(\sigma(x^*) + \sigma_0 \|X_n - x^*\|)}{\sqrt{m_{n+1}}} \\ &= \left( 1 + \alpha_n L + \alpha_n \frac{C_{p'} \sigma_0}{\sqrt{m_{n+1}}} \right) \|X_n - x^*\| + \alpha_n \frac{C_{p'} \sigma(x^*)}{\sqrt{m_{n+1}}}. \end{aligned} \quad (22)$$

Plugging this inequality into (21), after rearranging the terms, we see that

$$\begin{aligned} \mathbb{E}[\|Z_{n+1}\|^{p'} | \mathcal{F}_n]^{\frac{1}{p'}} &\leq \frac{C_{p'} \sigma(x^*)}{\sqrt{m_{n+1}}} \left( 1 + \alpha_n \frac{\sigma_0 C_{p'}}{\sqrt{m_{n+1}}} \right) \\ &\quad + \|X_n - x^*\| \frac{C_{p'} \sigma_0}{\sqrt{m_{n+1}}} \left( 1 + \alpha_n L + \alpha_n \frac{C_{p'} \sigma_0}{\sqrt{m_{n+1}}} \right). \end{aligned}$$

We denote

$$G_{n,p} \triangleq \frac{C_p}{\sqrt{m_{n+1}}}, \quad (23)$$

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

$$\mathbb{E}[\|W_{n+1}\|^{p'} | \mathcal{F}_n]^{\frac{1}{p'}} \leq G_{n,p'} (\sigma(x^*) + \sigma_0 \|X_n - x^*\|), \quad (24)$$

$$\mathbb{E}[\|Z_{n+1}\|^{p'} | \mathcal{F}_n]^{\frac{1}{p'}} \leq \sigma(x^*) G_{n,p'} (1 + \alpha_n \sigma_0 G_{n,p'}) + \sigma_0 G_{n,p'} \|X_n - x^*\| (1 + \alpha_n L + \alpha_n \sigma_0 G_{n,p'}), \quad (25)$$

$$\mathbb{E}[\|Y_n - x^*\|^{p'} | \mathcal{F}_n]^{\frac{1}{p'}} \leq (1 + \alpha_n L + \alpha_n \sigma_0 G_{n,p'}) \|X_n - x^*\| + \alpha_n \sigma(x^*) G_{n,p'}. \quad (26)$$

In case of a UBV (10), we obtain from the preceding 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 7.** 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

$$\begin{aligned} \mathbb{E}[|\Delta V_n|^q | \mathcal{F}_n]^{\frac{1}{q}} &\leq \alpha_n^2 G_{n,p'}^2 \sigma(x^*)^2 \left[ 2(4 + \rho_n) + 16 + 16\alpha_n^2 \sigma_0^2 G_{n,p'}^2 \right] \\ &\quad + \alpha_n^2 G_{n,p'}^2 \sigma_0^2 \|X_n - x^*\|^2 \left[ 2(4 + \rho_n) + 8(1 + \alpha_n L + \alpha_n \sigma_0 G_{n,p'})^2 \right] \end{aligned} \quad (27)$$

and

$$\begin{aligned} &\mathbb{E}[|\Delta U_n(x^*)|^q | \mathcal{F}_n]^{\frac{1}{q}} \\ &\leq 2\alpha_n^2 G_{n,p'}^2 \sigma(x^*)^2 (1 + \alpha_n G_{n,p'} \sigma_0) \\ &\quad + 2\alpha_n G_{n,p'} \sigma(x^*) \|X_n - x^*\| \left[ 1 + \alpha_n L + \alpha_n \sigma_0 G_{n,p'} (3 + 2\alpha_n L) + 2\alpha_n^2 \sigma_0^2 G_{n,p'}^2 \right] \\ &\quad + 2\alpha_n G_{n,p'} \sigma_0 \|X_n - x^*\|^2 (1 + \alpha_n L + \alpha_n \sigma_0 G_{n,p'})^2. \end{aligned} \quad (28)$$

If (10) holds with variance bound  $\hat{\sigma}$ , then these upper bounds simplify to

$$\mathbb{E}[|\Delta V_n|^q | \mathcal{F}_n]^{\frac{1}{q}} \leq \alpha_n^2 \hat{\sigma}^2 G_{n,p'}^2 (8 + \rho_n) \quad (29)$$

and, respectively,

$$\mathbb{E}[|\Delta U_n(x^*)|^q | \mathcal{F}_n]^{\frac{1}{q}} \leq 2\alpha_n \hat{\sigma} G_{n,p'} (1 + L\alpha_n) \|X_n - x^*\| + 2\alpha_n^2 \hat{\sigma}^2 G_{n,p'}^2. \quad (30)$$

**Proof.** Let  $n \geq 0$ . 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 (24) and (25) and rearranging terms, we obtain (27). By contrast, we have by definition

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

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

$$\begin{aligned} \mathbb{E}[|\Delta U_n(x^*)|^q | \mathcal{F}_n]^{\frac{1}{q}} &\leq 2\alpha_n G_{n,p'} \sigma(x^*) \mathbb{E}[\|Y_n - x^*\|^q | \mathcal{F}_n]^{\frac{1}{q}} \\ &\quad + 2\alpha_n G_{n,p'} \sigma_0 \mathbb{E}[\|Y_n - x^*\|^{p'} | \mathcal{F}_n]^{\frac{1}{p'}} \\ &\leq 2\alpha_n G_{n,p'} \sigma(x^*) \mathbb{E}[\|Y_n - x^*\|^{p'} | \mathcal{F}_n]^{\frac{1}{p'}} \\ &\quad + 2\alpha_n G_{n,p'} \sigma_0 \mathbb{E}[\|Y_n - x^*\|^{p'} | \mathcal{F}_n]^{\frac{1}{p'}}. \end{aligned}$$

After applying (26) and rearranging terms, we arrive at the (28).

In case UBV (10) holds with uniform variance bound  $\hat{\sigma}$ , the upper bound for  $|\Delta V_{n+1}|^q$  follows immediately from the defining expression (12) 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 (28) by setting  $\sigma_0 = 0$  and replacing  $\sigma(x^*)$  with its uniform upper bound  $\hat{\sigma}$ .  $\square$

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

**Proposition 1.** For all  $x^* \in \mathcal{X}_*$  and all  $n \geq 0$ , we have

$$\mathbb{E}[\|X_{n+1} - x^*\|^2 | \mathcal{F}_n] \leq \|X_n - x^*\|^2 - \frac{\rho_n}{2} r_{\alpha_n}(X_n)^2 + \frac{\kappa_n}{m_{n+1}} \left[ \sigma_0^2 \|X_n - x^*\|^2 + \sigma(x^*)^2 \right], \quad (31)$$

where

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

If (10) holds with uniform variance bound  $\hat{\sigma}$ , then

$$\mathbb{E} \left[ \|X_{n+1} - x^*\|^2 | \mathcal{F}_n \right] \leq \|X_n - x^*\|^2 - \frac{\rho_n}{2} r_{\alpha_n}(X_n)^2 + \frac{\kappa_n \hat{\sigma}^2}{m_{n+1}}, \quad (32)$$

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

**Proof.** Let  $x^* \in \mathcal{X}_*$  and  $n \geq 0$ . Our point of departure is (19), together with (27). From here we derive that

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

In the preceding equality, we 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 (10) holds, we just combine (19) with (29) to obtain the claimed result.  $\square$

**Remark 3.** The scaling factor  $\kappa_n$  only depends on the step size  $\alpha_n$ , the Lipschitz constant  $L$ , and the variance bound on the SO 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  $\rho_n$  in (11), we can bound

$$\begin{aligned} \kappa_n &= \alpha_n^2 C_2^2 \left[ 2(4 + \rho_n) + 16 \left( 1 + \alpha_n L + \frac{\alpha_n \sigma_0 C_2}{\sqrt{m_{n+1}}} \right)^2 \right] \\ &\leq \alpha_n^2 C_2^2 \left[ 10 + 32(1 + \alpha_n L)^2 + 32\alpha_n^2 \sigma_0^2 \frac{C_2^2}{m_{n+1}} \right] \\ &\leq \bar{\alpha}^2 C_2^2 \mathbf{c}_1 \left[ 1 + \frac{\bar{\alpha}^2 \sigma_0^2 C_2^2}{m_{n+1}} \right], \quad \forall n \geq 0, \end{aligned}$$

where  $\mathbf{c}_1 > 1$  is a constant. Combined with the batch-size condition (8), we obtain the existence of constants  $\mathbf{c}_0$  and  $\mathbf{c}_1$  such that

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

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

Next, we 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 theorem (Lemma 3) and recent results on the convergence of stochastic quasi-Féjer monotone sequences (Combettes and Pesquet 2015, proposition 2.3).

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

$$\text{Lim}(f)(\omega) \triangleq \left\{ x \in \mathbb{R}^d \mid (\exists (n_j) \uparrow \infty) : \lim_{n_j \rightarrow \infty} f_{n_j}(\omega) = x \right\}.$$

**Theorem 1.** 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  $n \rightarrow \infty$  a.s. to a limit random variable  $X$  with values in  $\mathcal{X}_*$ , and  $\lim_{n \rightarrow \infty} \mathbb{E}[r_{\alpha_n}(X_n)^2] = 0$ .

**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 (31) can be rewritten for all  $n \geq 0$  as

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

Hence, by Lemma 3, there exists a random variable  $\hat{\delta}(x^*) \in [0, \infty)$  such that  $(\delta_n(x^*))_{n \geq 1} \rightarrow \hat{\delta}(x^*)$  a.s. as  $n \rightarrow \infty$ , and  $\mathbb{P}[\sum_{n \geq 0} u_n < \infty] = 1$ . In particular,  $(X_n)_{n \geq 0}$  is bounded for almost every  $\omega \in \Omega$ . Because  $\sum_{n \geq 0} u_n = \sum_{n \geq 0} \rho_n r_{\alpha_n}(X_n)^2 \geq \hat{\rho} \sum_{n \geq 0} r_{\alpha_n}(X_n)^2$ , where  $\hat{\rho} = 1 - 2\bar{\alpha}^2 L^2 > 0$ , it follows that  $\lim_{n \rightarrow \infty} r_{\alpha_n}(X_n) = 0$ ,  $\mathbb{P}$ -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 a.s. to a random variable  $X$  with values in  $\mathcal{X}_*$ . Let  $\omega \in \Omega$  be such that  $X_n(\omega)$  is bounded. Because  $(\alpha_n)_{n \geq 0}$  is bounded as well, we can construct subsequences  $(\alpha_{n_j})_{j \geq 0}$  and  $(X_{n_j}(\omega))_{j \geq 0}$  such that  $\lim_{j \rightarrow \infty} \alpha_{n_j} = \alpha \in [\underline{\alpha}, \bar{\alpha}]$  and  $\lim_{j \rightarrow \infty} X_{n_j}(\omega) = \chi(\omega)$ . Additionally, we have  $\lim_{j \rightarrow \infty} r_{\alpha_{n_j}}(X_{n_j}(\omega)) = 0$ , so

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

Therefore, by continuity of the projection operator and of the averaged map  $T$ , Lemma 1(d) allows us to conclude that  $\chi(\omega) \in \mathcal{X}_*$ . Because the subsequence is arbitrary, it follows that  $\text{Lim}((X_n)_{n \geq 0})(\omega) \subseteq \mathcal{X}_*$  for  $\mathbb{P}$ -almost all  $\omega \in \Omega$ . Now apply proposition 2.3(iv) of Combettes and Pesquet (2015) to conclude that  $X_n \rightarrow X \in L^0(\Omega, \mathcal{F}, \mathbb{P}; \mathcal{X}_*)$   $\mathbb{P}$ -a.s.

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

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

Let  $z_n = \mathbb{E}[\delta_n(x^*)]$ ,  $u_n = \frac{\rho_n}{2} \mathbb{E}[r_{\alpha_n}(X_n)^2]$  and  $\theta_n$  and  $\beta_n$  be defined as in the preceding paragraph. The deterministic version of Lemma 3 gives  $(u_n)_{n \geq 1} \in \ell_+^1(\mathbb{N})$ . Hence,  $\lim_{n \rightarrow \infty} \mathbb{E}[r_{\alpha_n}(X_n)^2] = 0$ .  $\square$

Theorem 1 considerably strengthens similar results obtained via different splitting techniques. For SEG (10), asymptotic convergence of the iterates in the sense of Theorem 1 is established in theorem 3 of Iusem et al. (2017). However, different from algorithm SFBF, SEG requires two costly projection steps, with the same number of SO calls. This makes algorithm SFBF a potentially more efficient tool, and we will demonstrate that this is actually the case empirically and theoretically. Under strong monotonicity assumptions, a version of Theorem 1 has been established recently for a stochastic version of the classical forward-backward splitting technique in Rosasco et al. (2016), assuming a similar variance structure on the SO as we do. Theorem 1 shows convergence of the SFBF algorithm under the much weaker assumption of pseudomonotonicity 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 2.** Suppose that Assumptions 1–7 hold. For  $x^* \in \mathcal{X}_*$ , set  $\hat{\sigma}(x^*) \triangleq \max\{\sigma(x^*), \sigma_0\}$ , and define  $\text{dist}(x, \mathcal{X}_*) \triangleq \inf_{y \in \mathcal{X}_*} \|y - x\| = \|\Pi_{\mathcal{X}_*}(x) - x\|$ . For all  $n \geq 0$ , it holds that

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

If UBV (10) holds, then we get for all  $n \geq 0$  the uniform bound

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

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

**Proof.** Let  $\pi_n(\omega) = \Pi_{\mathcal{X}_*}(X_n(\omega))$  for all  $n \geq 0$  and all  $\omega \in \Omega$ . Because 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

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

where the second inequality uses Proposition 1.  $\square$



We are now in a 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 3.** 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  $\beta_n \rightarrow 0$  a.s., then  $(Y_n)_n$  converges a.s. to an  $\mathcal{X}_*$ -valued random variable  $Y$ .

**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 (22), we see that

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

Taking expectations on both sides,

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

By Theorem 1,  $X_n \rightarrow X$  a.s., an  $\mathcal{X}_*$ -valued random variable. Therefore, we know that  $\lim_{n \rightarrow \infty} \mathbb{E}[d(X_n)] = 0$ . By hypothesis,  $\beta_n \rightarrow 0$  a.s., so  $\limsup_{n \rightarrow \infty} \mathbb{E}[c_n] \leq 0$ . 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,  $\limsup_{n \rightarrow \infty} \mathbb{E}[d(Y_n)^2] \leq 0$ , and consequently,  $d(Y_n) \rightarrow 0$  a.s. The convergence to an  $\mathcal{X}_*$ -valued limit random variable then follows from proposition 2.3 in Combettes and Pesquet (2015).  $\square$

We remark that the assumption  $\beta_n \rightarrow 0$  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 SO'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 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 that of 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 (10) shows that the constants are consistently smaller. This indicates that the SFBF algorithm should empirically outperform SEG (10). 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 4.** Suppose that Assumptions 1–7 hold. For all  $x^* \in \mathcal{X}_*$ , let

$$\hat{\sigma}(x^*) \triangleq \max\{\sigma(x^*), \sigma_0\}, \quad (33)$$

$$\mathbf{a}(x^*) \triangleq \hat{\sigma}^2(x^*) \bar{\alpha}^2 C_2^2 c_1. \quad (34)$$

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

$$\sum_{n \geq n_0} \frac{1}{m_{n+1}} \leq \gamma \quad (35)$$

and

$$\beta(x^*) \triangleq \gamma \mathbf{a}(x^*) + \gamma^2 \mathbf{a}(x^*)^2 \in (0, 1). \quad (36)$$

Then

$$\sup_{n \geq n_0+1} \mathbb{E}[\|X_n - x^*\|^2] \leq \frac{\mathbb{E}[\|X_{n_0} - x^*\|^2] + 1}{1 - \beta(x^*)}. \quad (37)$$

**Proof.** Because of Assumption 6, for every  $\gamma > 0$ , we can find an index  $n_0 \in \mathbb{N}$  such that (35) holds. Consequently, we fix  $n_0 \in \mathbb{N}$  to be the smallest positive integer so that (35) holds. For all  $n \geq 0$ , we denote  $\psi_n(x^*) \triangleq \mathbb{E}[\|X_n - x^*\|^2]$ . From Proposition 1, we obtain

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

Recall from Remark 3 that

$$\kappa_n \leq \bar{\alpha}^2 C_2^2 c_1 \left( 1 + \frac{\bar{\alpha}^2 \sigma_0^2 C_2^2}{m_{n+1}} \right) \leq \bar{\alpha}^2 C_2^2 c_1 \left( 1 + \frac{\mathbf{a}(x^*)}{c_1 m_{n+1}} \right).$$

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

$$\psi_n(x^*) \leq \psi_{n_0}(x^*) + \sum_{k=n_0}^{n-1} (1 + \psi_k(x^*)) \frac{\mathbf{a}(x^*)}{m_{k+1}} + \sum_{k=n_0}^{n-1} (1 + \psi_k(x^*)) \frac{\mathbf{a}(x^*)^2}{c_1 m_{k+1}^2}.$$

For  $p > \psi_{n_0}(x^*)$ , define  $\tau_p(x^*) \triangleq \inf\{n \geq n_0 + 1 \mid \psi_n(x^*) \geq p\} \in \mathbb{N} \cup \{+\infty\}$ . 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 (35), we get

$$\begin{aligned} p &\leq \psi_{\tau_p(x^*)}(x^*) \leq \psi_{n_0}(x^*) + \sum_{k=n_0}^{\tau_p(x^*)-1} (1 + \psi_k(x^*)) \frac{\mathbf{a}(x^*)}{m_{k+1}} \\ &\quad + \sum_{k=n_0}^{\tau_p(x^*)-1} (1 + \psi_k(x^*)) \frac{1}{c_1} \left( \frac{\mathbf{a}(x^*)}{m_{k+1}} \right)^2 \\ &\leq \psi_{n_0}(x^*) + (1+p)\gamma \mathbf{a}(x^*) + (1+p) \frac{\gamma^2 \mathbf{a}(x^*)^2}{c_1}. \end{aligned}$$

Rearranging and using  $c_1 > 1$  as well as (36), we get

$$p \leq \frac{\psi_{n_0}(x^*) + 1}{1 - \gamma \mathbf{a}(x^*) - \frac{\gamma^2}{c_1} \mathbf{a}(x^*)^2} \leq \frac{\psi_{n_0}(x^*) + 1}{1 - \gamma \mathbf{a}(x^*) - \gamma^2 \mathbf{a}(x^*)^2}.$$

Because  $p > \psi_{n_0}(x^*)$  has been chosen arbitrarily, we can let  $p \rightarrow \infty$  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  $n \geq n_0 + 1$ ,

$$\begin{aligned} \psi_n(x^*) &\leq \psi_{n_0}(x^*) + \sum_{k=n_0}^{n-1} (1 + \psi_k(x^*)) \frac{\mathbf{a}(x^*)}{m_{k+1}} + \sum_{k=n_0}^{n-1} (1 + \psi_k(x^*)) \frac{1}{c_1} \left( \frac{\mathbf{a}(x^*)}{m_{k+1}} \right)^2 \\ &\leq \psi_{n_0}(x^*) + (1 + \bar{p})\gamma \mathbf{a}(x^*) + (1 + \bar{p}) \frac{\gamma^2 \mathbf{a}(x^*)^2}{c_1}. \end{aligned}$$

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 (37).  $\square$

In the 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.** Suppose that Assumptions 1–7 hold. Suppose that 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

$$\mathbf{a} \triangleq \bar{\alpha}^2 \hat{\sigma}^2 C_2^2 c_1. \quad (38)$$

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

$$\sup_{n \geq n_0+1} \mathbb{E}[\text{dist}(X_n, \mathcal{X}_*)^2] \leq \frac{1 + \mathbb{E}[\text{dist}(X_{n_0}, \mathcal{X}_*)^2]}{1 - \phi - \phi^2}. \quad (39)$$

**Proof.** We denote by  $d(x) \triangleq \text{dist}(x, \mathcal{X}_*) : \mathbb{R}^d \rightarrow \mathbb{R}_+$  the distance function of the solution set  $\mathcal{X}_*$ . Because  $\mathcal{X}_*$  is a nonempty, closed, and convex subset of  $\mathbb{R}^d$ , the function  $d(X_n) : \Omega \rightarrow \mathbb{R}_+$  given by  $\omega \mapsto d(X_n)(\omega) \triangleq \text{dist}(X_n(\omega), \mathcal{X}_*)$  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  $n \geq 0$ ,

$$\begin{aligned} \mathbb{E}[d(X_{n+1})^2 | \mathcal{F}_n] &\leq \mathbb{E}[\|X_{n+1} - \pi_n\|^2 | \mathcal{F}_n] \\ &\leq \|X_n - \pi_n\|^2 - \frac{\rho_n}{2} \mathbb{E}[r_{\alpha_n}(X_n)^2] + \frac{\kappa_n}{m_{n+1}} (\sigma_0^2 d(X_n)^2 + \sigma(\pi_n)^2). \end{aligned}$$

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

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

From Remark 3, we know that

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

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

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

From here proceed, mutatis mutandis, as in the proof of Proposition 4.  $\square$

We next give explicit estimates of the rate of convergence and the SO complexity of algorithm 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  $\alpha_n = \alpha \in (0, \frac{1}{\sqrt{2}L})$  for all  $n \geq 0$ .<sup>3</sup> As in Iusem et al. (2017), we can provide nonasymptotic 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  $\phi \in (0, \frac{\sqrt{5}-1}{2})$ , define

$$\begin{aligned} \Gamma_n &\triangleq \sum_{i=0}^n \frac{1}{m_{i+1}}, \quad \Gamma_n^2 \triangleq \sum_{i=0}^n \frac{1}{m_{i+1}^2}, \\ \rho &= 1 - 2\alpha^2 L^2, \quad \delta_n(x^*) \triangleq \|X_n - x^*\|^2, \\ \text{and } H(x^*, n, \phi) &\triangleq \frac{1 + \max_{0 \leq i \leq n} \mathbb{E}[\delta_i(x^*)]}{1 - \phi - \phi^2}. \end{aligned}$$

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

$$\sum_{i \geq n_0} \frac{1}{m_{i+1}} \leq \frac{\phi}{\mathbf{a}(x^*)}, \quad (40)$$

where  $\mathbf{a}(x^*)$  is defined in (34). Let

$$\begin{aligned} \Lambda_n(x^*, \phi) &\triangleq \frac{2}{\rho} \left\{ \mathbb{E}[\delta_0(x^*)] + (1 + H(x^*, n_0, \phi)) (\mathbf{a}(x^*) \Gamma_n + \mathbf{a}(x^*)^2 \Gamma_n^2) \right\}, \\ \Lambda_\infty(x^*, \phi) &\triangleq \sup_{n \geq 0} \Lambda_n(x^*, \phi). \end{aligned}$$

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

$$N_\varepsilon \triangleq \inf\{n \geq 0 \mid \mathbb{E}[r_\alpha(X_n)^2] \leq \varepsilon\}. \quad (41)$$

Then, either  $N_\varepsilon = 0$  or

$$\mathbb{E}[r_\alpha(X_{N_\varepsilon})^2] \leq \varepsilon < \frac{\Lambda_\infty(x^*, \phi)}{N_\varepsilon}. \quad (42)$$

**Proof.** Let  $\gamma = \frac{\phi}{\mathbf{a}(x^*)}$ , with the constant  $\mathbf{a}(x^*)$  defined in (34), and let  $n_0 = n_0(x^*)$ , as required in the statement of the theorem. From Proposition 4, we deduce the bound

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

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

$$\sup_{n \geq 0} \mathbb{E}[\delta_n(x^*)] \leq H(x^*, n_0, \phi). \quad (43)$$

Taking expectations in (31), we get

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

Therefore, for all  $n \geq 0$ ,

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

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 3 the bound

$$\kappa_i \leq \alpha^2 C_2^2 \mathbf{c}_1 \left( 1 + \frac{\alpha^2 C_2^2 \hat{\sigma}(x^*)^2}{m_{i+1}} \right), \quad \forall i \geq 0.$$

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

$$\begin{aligned} \frac{\rho}{2} \sum_{i=0}^n \mathbb{E}[r_\alpha(X_i)^2] &\leq \mathbb{E}[\delta_0(x^*)] + \sum_{i=0}^n \frac{\mathbf{a}(x^*)}{m_{i+1}} (1 + \mathbb{E}[\delta_i(x^*)]) \\ &\quad + \sum_{i=0}^n \frac{1}{\mathbf{c}_1} \left( \frac{\mathbf{a}(x^*)}{m_{i+1}} \right)^2 (1 + \mathbb{E}[\delta_i(x^*)]) \\ &\leq \mathbb{E}[\delta_0(x^*)] + \left( 1 + \max_{0 \leq i \leq n} \mathbb{E}[\delta_i(x^*)] \right) \left( \mathbf{a}(x^*) \Gamma_n + \mathbf{a}(x^*)^2 \Gamma_n^2 \right). \end{aligned}$$

From (43), we conclude that

$$\begin{aligned} \frac{\rho}{2} \sum_{i=0}^n \mathbb{E}[r_\alpha(X_i)^2] &\leq \mathbb{E}[\delta_0(x^*)] + (1 + H(x^*, n_0, \phi)) \left( \mathbf{a}(x^*) \Gamma_n + \mathbf{a}(x^*)^2 \Gamma_n^2 \right) \\ &= \frac{\rho}{2} \Lambda_n(x^*, \phi), \quad \forall n \geq 0. \end{aligned}$$

In conclusion,

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

From Theorem 1, we know that for all  $\varepsilon > 0$  there exists  $M_\varepsilon \in \mathbb{N}$  such that  $\mathbb{E}[r_\alpha(X_n)^2] \leq \varepsilon$  for all  $n \geq M_\varepsilon$ . Hence, the (deterministic) stopping time  $N_\varepsilon$  defined in (41) is either zero or an integer bounded from above. Focusing on the latter case  $N_\varepsilon \geq 1$ , then for every  $0 \leq k \leq N_\varepsilon - 1$ , we have

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

From here, it follows 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,

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

The preceding two cases can be compactly summarized to statement (42).  $\square$

We next turn to the case where the local variance is uniformly bounded over the solution set. In the preceding 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 from  $x^*$ , where  $n_0(x^*)$  and  $\phi$  are chosen such that (40) holds. Assuming a uniform bound on the variance of the SO over the solution set  $\mathcal{X}_*$ , we can obtain much stronger convergence rate estimates, holding uniformly over the solution set.

**Proposition 6.** Assume that  $\sup_{x^* \in \mathcal{X}_*} \hat{\sigma}(x^*) \leq \hat{\sigma}$ , where the function  $\hat{\sigma}(\cdot)$  is defined in (33). Let  $x^* \in \mathcal{X}_*$  be arbitrarily chosen, and consider algorithm SFBF with constant step size  $\alpha \in (0, \frac{1}{\sqrt{2L}})$ . Choose  $\phi \in (0, \frac{\sqrt{5}-1}{2})$  and  $n_0 \triangleq n_0(\hat{\sigma})$  to be the first integer such that

$$\sum_{i \geq n_0} \frac{1}{m_{i+1}} \leq \frac{\phi}{\mathbf{a}}, \quad (44)$$

where  $\mathbf{a} = \hat{\sigma}^2 \alpha^2 C_2^2 c_1$ . Let

$$\begin{aligned} \bar{\Lambda}_n(\hat{\sigma}, \phi) &\triangleq \frac{2}{\rho} \left\{ \mathbb{E}[\text{dist}(X_0, \mathcal{X}_*)^2] + (1 + \bar{H}(\hat{\sigma}, n_0, \phi))(\mathbf{a}\Gamma_n + \mathbf{a}^2\Gamma_n^2) \right\}, \\ \bar{\Lambda}_\infty(\hat{\sigma}, \phi) &= \sup_{n \geq 0} \bar{\Lambda}_n(\hat{\sigma}, \phi), \text{ and} \\ \bar{H}(\hat{\sigma}, n_0, \phi) &\triangleq \frac{1 + \max_{0 \leq i \leq n_0(\hat{\sigma})} \mathbb{E}[\text{dist}(X_i, \mathcal{X}_*)]}{1 - \phi - \phi^2}. \end{aligned}$$

For all  $\varepsilon > 0$ , consider the stopping time defined in (41). Then either  $N_\varepsilon = 0$  or

$$\mathbb{E}[r_\alpha(X_{N_\varepsilon})^2] \leq \varepsilon < \frac{\bar{\Lambda}_\infty(\hat{\sigma}, \phi)}{N_\varepsilon}. \quad (45)$$

**Proof.** The proof is almost identical to the proof of Theorem 2, but now we will use the estimates from Propositions 2 and 5. 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 make this choice, we can repeat all the steps involved in the proof of Theorem 2 verbatim but using Proposition 2 instead of Proposition 1 to conclude that

$$\begin{aligned} \sum_{i=0}^n \mathbb{E}[r_\alpha(X_i)^2] &\leq \mathbb{E}[\text{dist}(X_0, \mathcal{X}_*)^2] + \mathbf{a} \sum_{i=0}^n \frac{1 + \mathbb{E}[\text{dist}(X_i, \mathcal{X}_*)^2]}{m_{i+1}} \\ &\quad + \mathbf{a}^2 \sum_{i=0}^n \frac{1 + \mathbb{E}[\text{dist}(X_i, \mathcal{X}_*)^2]}{m_{i+1}^2}, \quad \forall n \geq 0. \end{aligned}$$

Proposition 5 gives us

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



from which it follows that

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

From here, we conclude just as in the proof of Theorem 2 that

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

Choose  $\varepsilon > 0$  arbitrarily and consider the stopping time (41). Then either  $N_\varepsilon = 0$  or else  $N_\varepsilon \geq 1$ . Focusing on the latter case, we argue just as in the proof of Theorem 2 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  $N_\varepsilon$  not zero, we must have

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

We now turn to the estimate of SO complexity. By this we mean the overall size of the data set that needs 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 (41), 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 from doing so because 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 SO complexity estimate in two parts. First, we show that a specific (although 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 6. Building on this insight, we directly estimate the SO complexity.

As announced, we first establish a bound on the number of iterations we need to meet condition (44).

**Lemma 8.** Let  $\mathbf{a}$  be the constant defined in (38), and let  $\phi \in (0, \frac{\sqrt{5}-1}{2})$ . We choose the sample rate

$$m_i = \lceil \theta(\mu - 1 + i) \ln(\mu + i - 1)^{1+b} \rceil, \quad (46)$$

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

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

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

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

$$\begin{aligned} \sum_{i \geq n_0} \frac{1}{m_{i+1}} &\leq \frac{1}{\theta} \sum_{i \geq n_0} \frac{1}{(i + \mu) \ln(i + \mu)^{1+b}} \\ &\leq \frac{1}{\theta} \int_{n_0-1}^{\infty} \frac{1}{(t + \mu) \ln(t + \mu)^{1+b}} dt \\ &= \frac{1}{\theta b \ln(n_0 - 1 + \mu)^b}. \end{aligned}$$

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

Using the sample rate (46), we will now bound the constant  $\bar{\Lambda}(\hat{\sigma}, \phi)$  and the stopping time  $N_\varepsilon$ . Define the constants

$$\mathcal{A}_{\mu,b} \triangleq \frac{\alpha^2 C_2^2 c_1}{b \ln(\mu-1)^b}, \quad \mathcal{B}_{\mu,b} \triangleq \frac{\alpha^4 C_2^4 c_1^2}{(1+2b)(\mu-1) \ln(\mu-1)^{1+2b}}.$$

Because

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

we conclude that

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

Therefore,

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

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

**Corollary 1.** For all  $\varepsilon > 0$ , the stopping time  $N_\varepsilon$  defined in (41) 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 estimation of the SO complexity. To this end, we have to bound the total number of data points involved in the  $N_\varepsilon$  batches needed to execute algorithm SFBF, that is, we want to upper bound the sum  $2 \sum_{i=0}^{N_\varepsilon} m_i$ . Given the definition of the sample rate in (46), we can perform the following computation:

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

Hence,

$$2 \sum_{i=1}^{N_\varepsilon} m_i \leq \max\{1, \theta\} (N_\varepsilon + 1)(N_\varepsilon + 2\mu) \left[ \ln(N_\varepsilon + 1 + \mu)^{1+b} + \frac{2}{N_\varepsilon + 2\mu} \right]. \quad (47)$$

**Proposition 7.** Let  $\varepsilon \in (0, 1)$  be arbitrarily chosen, and let  $\mu \in (1, 1/\varepsilon)$ . Define

$$\begin{aligned} \mathcal{I}(\hat{\sigma}, \phi) &\triangleq 3 \left( \frac{2}{\rho} \mathbb{E}[\text{dist}(X_0, \mathcal{X}_*)^2] + 2 \right)^2 \\ &\quad + \frac{12}{\rho^2} (1 + \bar{H}(\hat{\sigma}, n_0, \phi))^2 \mathcal{A}_{\mu,b}^2 \hat{\sigma}^4 + \frac{12}{\rho^2} (1 + \bar{H}(\hat{\sigma}, n_0, \phi))^2 \mathcal{B}_{\mu,b}^2 \hat{\sigma}^8, \\ \mathcal{J}(\hat{\sigma}, \phi) &\triangleq \bar{\Lambda}_\infty(\hat{\sigma}, n_0, \phi) + 2. \end{aligned}$$

If the sample rate  $(m_i)_{i \geq 1}$  is given by (46), then we can bound the SO 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.** The proof is patterned after Iusem et al. (2017). Using  $N_\varepsilon < \bar{\Lambda}_\infty(\phi, \hat{\sigma})/\varepsilon$ , we continue from (47) to obtain the bound

$$\begin{aligned} 2 \sum_{i=1}^{N_\varepsilon+1} m_i &\leq \max\{1, \theta\} \frac{(\bar{\Lambda}_\infty(\hat{\sigma}, \phi) + 1)(\bar{\Lambda}_\infty(\hat{\sigma}, \phi) + 2)}{\varepsilon^2} \left[ \ln \left( \frac{\bar{\Lambda}_\infty(\hat{\sigma}, \phi) + 2}{\varepsilon} \right)^{1+b} + \mu^{-1} \right] \\ &\leq \max\{1, \theta\} \frac{(\bar{\Lambda}_\infty(\hat{\sigma}, \phi) + 2)^2}{\varepsilon^2} \left[ \ln(\varepsilon^{-1} \mathcal{J}(\hat{\sigma}, \phi))^{1+b} + \mu^{-1} \right]. \end{aligned}$$

Because

$$\begin{aligned} &(\Lambda_\infty(\hat{\sigma}, \phi) + 2)^2 \\ &\leq \max\{1, \theta^{-4}\} \left\{ \frac{2}{\rho} \mathbb{E}[\text{dist}(X_0, \mathcal{X}_*)^2] + \frac{2}{\rho} (1 + \bar{H}(\hat{\sigma}, n_0, \phi)) [\mathcal{A}_{\mu,b} \hat{\sigma}^2 + \mathcal{B}_{\mu,b} \hat{\sigma}^4] + 2 \right\}^2 \\ &\leq \max\{1, \theta^{-4}\} \left\{ 3 \left( \frac{2}{\rho} \mathbb{E}[\text{dist}(X_0, \mathcal{X}_*)^2] + 2 \right)^2 + \frac{12}{\rho^2} (1 + \bar{H}(\hat{\sigma}, n_0, \phi))^2 \mathcal{A}_{\mu,b}^2 \hat{\sigma}^4 \right. \\ &\quad \left. + \frac{12}{\rho^2} \max\{1, \theta^{-4}\} (1 + \bar{H}(\hat{\sigma}, n_0, \phi))^2 \mathcal{B}_{\mu,b}^2 \hat{\sigma}^8 \right\} \\ &= \max\{1, \theta^{-4}\} \mathcal{I}(\hat{\sigma}, \phi), \end{aligned}$$

the result follows.  $\square$

## 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, besides experiment 2, were generated with MATLAB R2017a on a Linux operating system 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 of memory.

### 6.1. Fractional Programming and Applications to Communication Networks

Because of 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:

$$\begin{aligned} &\text{minimize } f(x) = \mathbb{E} \left[ \frac{G(x; \xi)}{h(x; \xi)} \right], \\ &\text{subject to } x \in \mathcal{X}, \end{aligned} \tag{48}$$

where  $G$  and  $h$  are positive and convex in  $x$  for all  $\xi$ . It is well known that such problems are pseudoconvex (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; that is,

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

for vectors  $a$  and  $b$  of suitable dimension. Solving this problem directly involves the pseudomonotone operator  $T(x) = \nabla f(x)$ . Indeed,  $x^* \in \mathcal{X}$  solves problem (48) if and only if  $x^*$  solves  $\text{VI}(T, \mathcal{X})$ .

**6.1.1. 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 semidefinite. More specifically, the problem data for  $Q$  are randomly generated as follows:

$$Q = M^\top M + I,$$

where  $M$  is a random matrix of size  $d \times d$ , and  $I$  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

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

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

$$c(\xi) := c + c_1(\xi), \quad q(\xi) = q + q_1(\xi), \quad (49)$$

where  $c_1(\xi)$  and  $q_1(\xi)$  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

$$\mathcal{X} = \{x \in \mathbb{R}^d : a_i \leq x_i \leq b_i, \quad i = 1, \dots, d\}, \quad (50)$$

where the lower bound  $a_i$  is a random vector in  $(0, 1)^d$ , and the upper bound is  $b_i = a_i + 10$ . We have implemented the SEG and SFBB algorithms 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 step size of SFBB and SEG as  $\alpha_{\text{FBB}} = 10/d$  and  $\alpha_{\text{EG}} = \alpha_{\text{FBB}}/\sqrt{3}$ . The step size  $\alpha_{\text{EG}}$  is the largest one compatible with the theory developed by Iusem et al. (2017). We choose the batch-size sequence  $m_{n+1} = \lceil \frac{(n+1)^{1.5}}{d} \rceil$  so that Assumption 6 is satisfied. We stop the algorithms when the residual is below a given tolerance  $\varepsilon$ . Specifically, our stopping criterion is

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

Our numerical experiments involve dimension  $d \in \{200, 500, 1,000, 2,000\}$ , and for each value of  $d$ , we perform 10 runs and compare the average number of iterations and central processing unit time. The results are displayed in Table 1 and Figure 1. It can be seen that SFBB is constantly about 1.5 times 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.

**6.1.2. Energy Efficiency in Multiantenna 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 multiantenna devices (i.e., wireless devices equipped with several antennas to multiplex and demultiplex received or transmitted signals).

Following Isheden et al. (2012), Feng et al. (2013), and 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 per second) is given by the familiar Shannon–Telatar capacity formula (Telatar 1999)

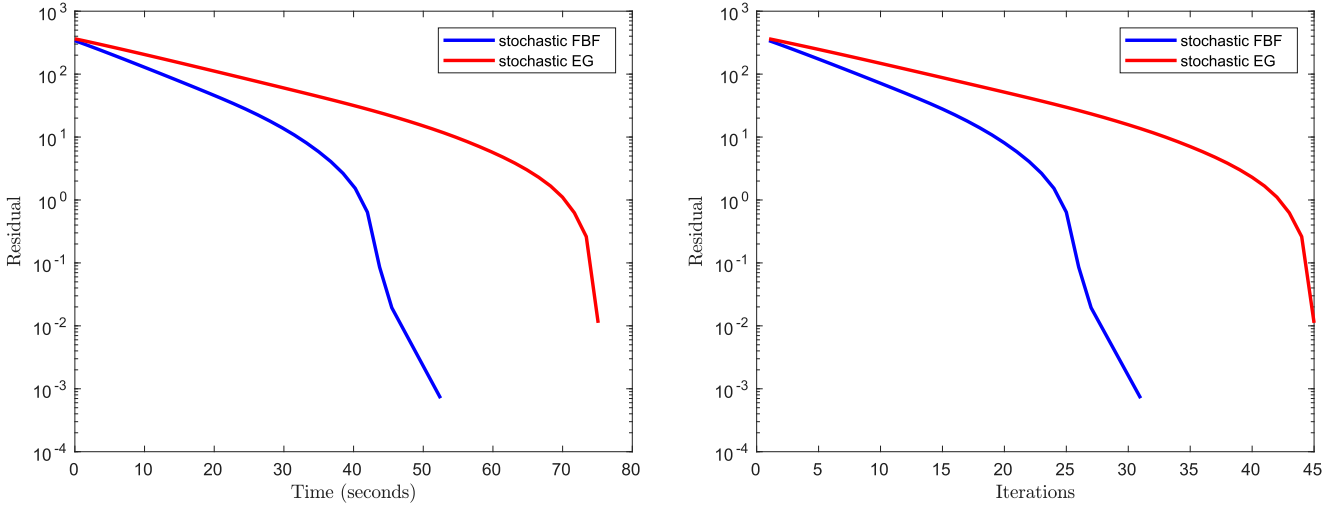
$$R(X; H) = \log \det \left( I + \sum_{k=1}^K H_k X_k H_k^\dagger \right), \quad (51)$$

where

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

**Table 1.** Averaged over 100 Runs for Fractional Problems of Different Size

Dimension: $d$	SFBB		SEG	
	Number of iterations	Time (s)	Number of iterations	Time (s)
200	29.88	0.0473	43.96	0.0835
500	29.84	0.2647	44.49	0.3793
1,000	30.14	1.1650	44.99	1.7017
2,000	30.54	8.0487	45.68	11.4803

**Figure 1.** Comparison Between the SFBF and SEG Algorithms for Solving Fractional Programming

Note. We represent the residual versus running time (left) and number of iterations (right) for one random example ( $n = 5,000$ ).

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.  $I$  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

$$R(X) = \mathbb{E}_H[R(X; H)], \quad (52)$$

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 throughput; that is,

$$EE(X) = \frac{R(X)}{\sum_{k=1}^K [P_k^c + P_k^t]}, \quad (53)$$

where  $P_k^t$  is the transmit power of the  $k$ -th device, and  $P_k^c > 0$  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. By elementary signal processing considerations, it is given by  $P_k^t = \text{tr}(X_k)$ . For concision, we will also write  $P^c = \sum_k P_k^c$  for the total circuit power dissipated 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

$$\text{tr}(X_k) \leq P_{\max} \quad \forall k = 1, \dots, K. \quad (54)$$

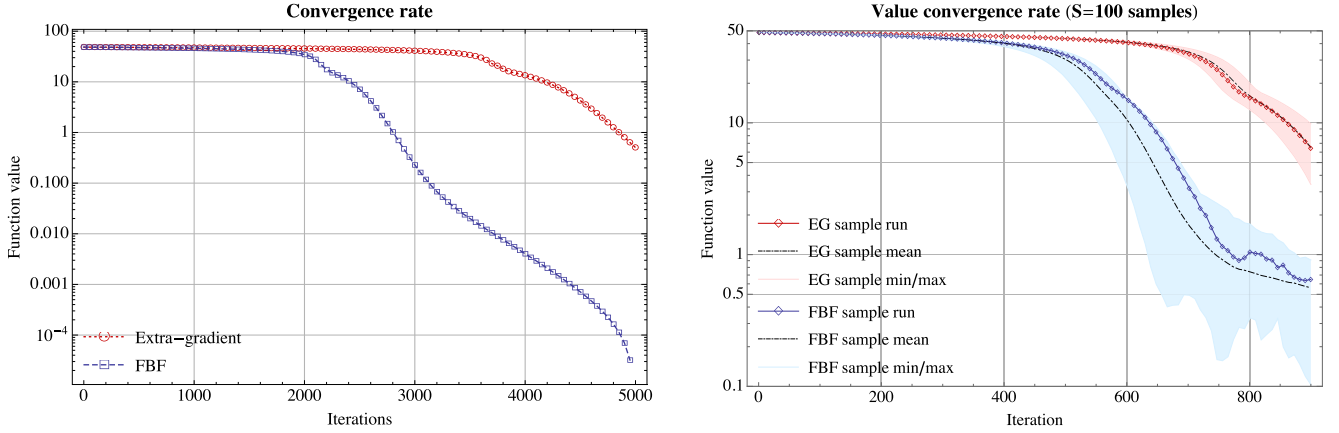
Hence, putting all this together, we obtain the stochastic fractional problem

$$\begin{aligned} \text{maximize} \quad & EE(X) = \frac{\mathbb{E}_H[R(X; H)]}{P^c + \sum_{k=1}^K \text{tr}(X_k)} \\ \text{subject to} \quad & X_k \geq 0, \\ & \text{tr}(X_k) \leq P_{\max} \quad \forall k = 1, 2, \dots, K. \end{aligned} \quad (55)$$

The overall problem dimension is  $d = KM^2$ . The energy-efficiency objective of this problem (which, formally, has units of bits per joule) has been widely studied in the literature (Cui et al. 2004, Isheden et al. 2012), and it captures the fundamental tradeoff between higher spectral efficiency and increased battery life. Importantly, switching from maximization to minimization, we also see that (55) is of the general form (48), 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.



**Figure 2.** Comparison of the Extragradient and Forward–Backward–Forward Methods in the Energy Efficiency Maximization Problem (55)



*Notes.* On the left, we considered static channels, and we ran the SFBF and SEG algorithms 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.

We do so in a series of numerical experiments reported in Figure 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 Cooperation in Science and Technology (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 and 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 a 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 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  $S_I \triangleq \{p \in \mathbb{R}_+^{n_I} \mid \sum_{i=1}^{n_I} p_i = 1\}$ ,  $S_{II} \triangleq \{q \in \mathbb{R}_+^{n_{II}} \mid \sum_{i=1}^{n_{II}} q_i = 1\}$ , and
- 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 consist of 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

$$M \triangleq \begin{bmatrix} \mathbf{0}_{n_I, n_I} & -U_I \\ -U_{II}^\top & \mathbf{0}_{n_{II}, n_{II}} \end{bmatrix}, \quad (56)$$

and consider the set

$$\mathcal{X} \triangleq \{(x_1, x_2) \in \mathbb{R}_+^{n_I} \times \mathbb{R}_+^{n_{II}} \mid U_I x_2 \leq \mathbf{1}_{n_I} \text{ and } U_{II}^\top x_1 \leq \mathbf{1}_{n_{II}}\}. \quad (57)$$

It is a classical fact that a Nash equilibrium  $(p^*, q^*)$  can be computed by finding a pair  $(x_1, x_2) \neq (n_I, n_{II}) \in \mathcal{X}$  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  $(n_I, n_{II})$  is always a solution to the *linear complementarity problem*

$$\begin{cases} x_1^\top (\mathbf{1}_{n_I} - U_I x_2) = 0, \mathbf{1}_{n_I} - U_I x_2 \geq \mathbf{0}_{n_I}, \\ x_2^\top (\mathbf{1}_{n_{II}} - U_{II}^\top x_1) = 0, \mathbf{1}_{n_{II}} - U_{II}^\top x_1 \geq \mathbf{0}_{n_{II}}. \end{cases} \quad (58)$$

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 by Von Stengel (2002). Defining the mapping  $T : \mathbb{R}^d \cong \mathbb{R}^{n_I} \times \mathbb{R}^{n_{II}} \rightarrow \mathbb{R}^d \cong \mathbb{R}^{n_I} \times \mathbb{R}^{n_{II}}$  by

$$T(x) \triangleq \begin{bmatrix} \mathbf{1}_{n_I} \\ \mathbf{1}_{n_{II}} \end{bmatrix} + Mx, \quad (59)$$

we can reformulate the conditions (58) compactly as

$$x^* \geq \mathbf{0}_n \text{ and } T(x^*) \geq \mathbf{0}_n, \langle x^*, T(x^*) \rangle = 0. \quad (60)$$

To turn this into a stochastic complementarity problem, we consider a stochastic Nash game (Kannan and Shanbhag 2012, Duvocelle et al. 2018) where the player set and the set of mixed actions are 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  $(\xi_n)$  is a random process in some set  $\Xi$ , defined on a probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ . For each  $n \geq 1$ , we look at that random operator

$$F(x, \xi_n) \triangleq \begin{bmatrix} \mathbf{1}_{n_I} \\ \mathbf{1}_{n_{II}} \end{bmatrix} + M(\xi_n)x \quad (61)$$

and run algorithm SFBF.

In our experiments,  $M$  is defined as in (56), 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 set up 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  $\sigma = 0.1$ . Because the operator  $T$  is Lipschitz continuous with modulus  $L = \|M\|$ , we run SEG and SFBF with constant step sizes  $\alpha_{\text{SFBF}} = \frac{0.99}{\sqrt{2L}}$  and  $\alpha_{\text{SEG}} = \frac{0.99}{\sqrt{6L}}$ , respectively. We choose the batch-size sequence  $m_{n+1} = \lceil \frac{(n+1)^{1.5}}{d} \rceil$  so that Assumption 6 is satisfied. The same stopping criterion as in the previous experiments in Section 6.1 is used.

From the numerical experiments, we observe that the SFBF algorithm outperforms 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 these results: first, SEG requires two projections per iteration, whereas SFBF only requires one, and more important, the step size of SFBF is  $\sqrt{3}$  times larger than that of SEG.

**6.2.1. Zero-Sum Games.** We compare the performance the SFBF and SEG algorithms for a zero-sum game, that is,  $U_I = -U_{II}^\top$ . 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 Games of Different Size

Dimension: $d = n_I + n_{II}$	SFBF		SEG	
	Iterations	Time (s)	Iterations	Time (s)
$n_I = n_{II} = 100$	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
$n_I = n_{II} = 1,000$	865.67	672.0806	1,508.50	2,535.50

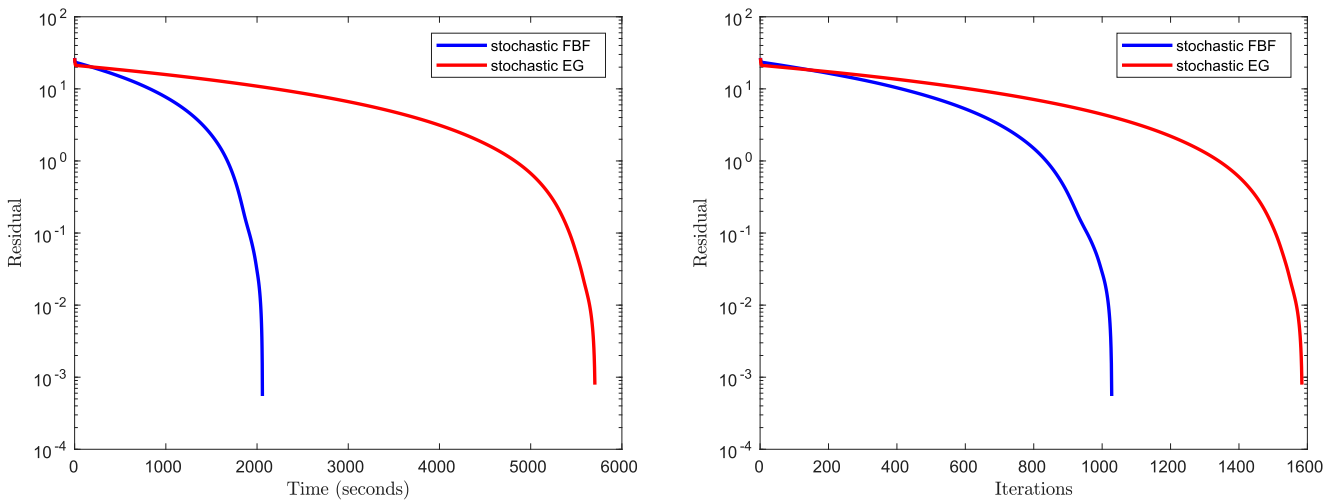
**Table 3.** Averaged over 100 Runs for Symmetric Games of Different Size

Dimension: $d = n_I + n_{II}$	SFBF		SEG	
	Iterations	Time (s)	Iterations	Time (s)
$n_I = n_{II} = 100$	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
$n_I = n_{II} = 1,000$	319.92	92.0417	455.48	191.6854

**Table 4.** Averaged over 100 Runs for Asymmetric Games of Different Size

Dimension: $d = n_I + n_{II}$	SFBF		SEG	
	Iterations	Time (s)	Iterations	Time (s)
$n_I = 100, n_{II} = 200$	100.28	1.9553	155.28	4.8202
$n_I = 300, n_{II} = 600$	293.36	32.3010	466.01	90.2339
$n_I = 500, n_{II} = 1,000$	492.21	136.7019	779.86	394.7606
$n_I = 1,000, n_{II} = 2,000$	992.64	1,597.7266	1,564.12	4,655.9213

**Figure 3.** Comparison Between the SFBF and SEG Algorithms for Solving an Asymmetric Game



Note. We represent the residual versus running time (left) and number of iterations (right) for one random example  $n_I = 1,000, n_{II} = 2,000$ .

**6.2.2. Symmetric Game.** We compare the performance the SFBF and SEG algorithms for a symmetric game; that is,  $U_I, U_{II}$  are symmetric, and  $U_I = U_{II}^T$ . We choose  $n_I = n_{II} \in \{50, 100, 150, \dots, 500\}$  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.

**6.2.3. Bimatrix Games.** We compare the performance the SFBF and SEG algorithms for an asymmetric game. We choose  $n_I \in \{100, 200, \dots, 1,000\}$  and  $n_{II} = 2n_I$ . The results are displayed in Table 4 and Figure 3, showing the advantage of SFBF over SEG.

## 7. Conclusion

In this paper, we developed a stochastic version of Tseng’s forward-backward-forward algorithm for solving stochastic variational inequality problems over nonempty closed and convex sets. As in Iusem et al. (2017), the current analysis can be generalized to Cartesian VI problems, although this has not been done 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 algorithm 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 6

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

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

Setting  $\mathcal{G}_i \triangleq \sigma(\xi^{(1)}, \dots, \xi^{(i)})$ ,  $1 \leq i \leq N$ , 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

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

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

$$\begin{aligned} \mathbb{E}[\|\Delta M_{i-1}^N(x)\|^q]^{1/q} &= \frac{1}{N} \mathbb{E}[\|F(x, \xi^{(i)}) - T(x)\|^q]^{1/q} \\ &\leq \frac{1}{N} \mathbb{E}[\|F(x, \xi^{(i)}) - T(x)\|^p]^{1/p} \\ &\leq \frac{\sigma(x^*) + \sigma_0 \|x - x^*\|}{N}. \end{aligned}$$

Using this, together with Lemma 4, we get

$$\begin{aligned} \mathbb{E}[\|M_N^N(x)\|^q]^{1/q} &\leq C_q \sqrt{\sum_{k=1}^N \mathbb{E} \left[ \left\| \frac{F(x, \xi^{(k)}) - T(x)}{N} \right\|^q \right]^{2/q}} \\ &\leq C_q \sqrt{N^{-2} \sum_{k=1}^N \mathbb{E} (\|F(x, \xi^{(k)}) - T(x)\|^q)^{2/q}} \\ &\leq \frac{C_q (\sigma(x^*) + \sigma_0 \|x - x^*\|)}{\sqrt{N}}. \quad \square \end{aligned}$$

Therefore, to continue the proof of Lemma 6, 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

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

To prove (21), we notice that Lemma A.1 implies that

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

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

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

Finally, by the Minkowski inequality, we get

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

and our proof is complete.  $\square$

## Endnotes

<sup>1</sup>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 *minibatch* instead.

<sup>2</sup>The mapping  $x \mapsto F(x, \xi)$  is continuous almost everywhere for  $\xi \in \Xi$ , and  $\xi \mapsto F(x, \xi)$  is measurable for all  $x \in \mathbb{R}^d$ ;  $\xi$  is a random variable with values in  $\Xi$ , defined on a probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ .

<sup>3</sup>The reason for this is that  $\{r_a(x); a > 0\}$  is a family of equivalent merit functions of  $\text{VI}(T, \mathcal{X})$  (see Facchinei and Pang 2003, proposition 10.3.6, 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.

## References

- Atchadé YF, Fort G, Moulines E (2017) On perturbed proximal gradient algorithms. *J. Machine Learn. Res.* 18(1):310–342.
- Bottou L, Curtis FE, Nocedal J (2018) Optimization methods for large-scale machine learning. *SIAM Rev.* 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 Trends Machine Learning* 3(1):1–122.
- Boyd SP, Vandenberghe L (2004) *Convex Optimization* (Cambridge University Press, Cambridge, UK).
- Chen C, Chen Y, Ouyang Y, Pasiliao E (2018) Stochastic accelerated alternating direction method of multipliers with importance sampling. *J. Optim. Theory Applications* 179(2):676–695.
- Combettes P, Pesquet J (2015) Stochastic quasi-Fejér block-coordinate fixed point iterations with random sweeping. *SIAM J. Optim.* 25(2):1221–1248.
- Cottle RW, Yao JC (1992) Pseudo-monotone complementarity problems in Hilbert space. *J. Optim. Theory Applications* 75(2):281–295.
- Cui S, Goldsmith AJ, Bahai A (2004) Energy-efficiency of MIMO and cooperative MIMO techniques in sensor networks. *IEEE J. Selected Areas Comm.* 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. *Comput. Optim. Applications* 60(2):277–310.
- Duflo M (1996) *Algorithmes Stochastiques* (Springer, New York).
- Duvocelle B, Mertikopoulos P, Staudigl M, Vermeulen D (2018) Learning in time-varying games. Preprint, submitted September 10, 1918, <https://arxiv.org/abs/1809.03066>.
- Facchinei F, Pang JS (2003) *Finite-Dimensional Variational Inequalities and Complementarity Problems*, Springer Series in Operations Research, vols. 1 and 2 (Springer, New York).
- Feng D, Jiang C, Lim G, Cimini LJ Jr, Feng G, Li GY (2013) A survey of energy-efficient wireless communications. *IEEE Comm. Survey and Tutorial* 15(1):167–178.
- Hata M (1980) Empirical formula for propagation loss in land mobile radio services. *IEEE Trans. Vehicular 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 Comm.* 11(8):2946–2957.
- Iusem A, Jofré A, Oliveira RI, Thompson P (2017) Extragradient method with variance reduction for stochastic variational inequalities. *SIAM J. Optim.* 27(2):686–724.
- Jiang H, Xu H (2008) Stochastic approximation approaches to the stochastic variational inequality problem. *IEEE Trans. Automatic Control* 53(6):1462–1475.
- Jofré A, Thompson P (2019) On variance reduction for stochastic smooth convex optimization with multiplicative noise. *Mathematical Programming* 174:253–292.
- 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 J. Optim.* 22(4):1177–1205.
- Kannan A, Shanbhag UV (2019) Optimal stochastic extragradient schemes for pseudomonotone stochastic variational inequality problems and their variants. *Comput. Optim. Applications* 74(3):779–820.
- King AJ, Rockafellar RT (1993) Asymptotic theory for solutions in statistical estimation and stochastic programming. *Math. Oper. Res.* 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* (Springer, New York).
- Mertikopoulos P, Belmega EV (2016) Learning to be green: Robust energy efficiency maximization in dynamic MIMO-OFDM systems. *IEEE J. Selected Areas Comm.* 34(4):743–757.



- Mertikopoulos P, Staudigl M (2018) Stochastic mirror descent dynamics and their convergence in monotone variational inequalities. *J. Optim. Theory Applications* 179(3):838–867.
- Mertikopoulos P, Zhou Z (2018) Learning in games with continuous action sets and unknown payoff functions. *Mathematical Programming* 173:465–507.
- Polyak BT (1987) *Introduction to Optimization* (Optimization Software, New York).
- Ravat U, Shanbhag U (2011) On the characterization of solution sets of smooth and nonsmooth convex stochastic nash games. *SIAM J. Optim.* 21(3):1168–1199.
- Rosasco L, Villa S, Vũ BC (2016) Stochastic forward–backward splitting for monotone inclusions. *J. Optim. Theory 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, Philadelphia).
- Shen K, Yu W (2018) Fractional programming for communication systems: I. Power control and beamforming. *IEEE Trans. Signal Processing* 66(10):2616–2630.
- Solodov M, Svaiter B (1999) A new projection method for variational inequality problems. *SIAM J. Control Optim.* 37(3):765–776.
- Stroock DW (2011) *Probability Theory: An Analytic View*, 2nd ed. (Cambridge University Press, Cambridge, UK).
- Telatar IE (1999) Capacity of multi-antenna Gaussian channels. *Eur. Trans. Telecomm. Related Tech.* 10(6):585–596.
- Tseng P (2000) A modified forward-backward splitting method for maximal monotone mappings. *SIAM J. Control Optim.* 38(2):431–446.
- Von Stengel B (2002) Computing equilibria for two-person games. *Handbook of Game Theory with Economic Applications*, vol. 3 (Elsevier, Amsterdam), 1723–1759.
- Yousefian F, Nedić A, Shanbhag UV (2017) On smoothing, regularization, and averaging in stochastic approximation methods for stochastic variational inequality problems. *Math. Programming* 165(1):391–431.