Convergent Noisy forward-backward-forward algorithms in non-monotone variational inequalities
Mathias Staudigl, Panayotis Mertikopoulos

To cite this version:
Mathias Staudigl, Panayotis Mertikopoulos. Convergent Noisy forward-backward-forward algorithms in non-monotone variational inequalities. LSS 2019 - 15th IFAC Symposium on Large Scale Complex Systems: Theory and Applications, May 2019, Delft, Netherlands. pp.120-125, 10.1016/j.ifacol.2019.06.021 . hal-02405750

HAL Id: hal-02405750
https://hal.inria.fr/hal-02405750
Submitted on 11 Dec 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers. L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Convergent Noisy forward-backward-forward algorithms in non-monotone variational inequalities

Mathias Staudigl * Panayotis Mertikopoulos **

* Department of Quantitative Economics, Maastricht University, Maastricht, The Netherlands (e-mail: m.staudigl@maastrichtuniversity.nl)
** Univ. Grenoble Alpes, CNRS, INRIA, LIG, Grenoble, France (e-mail: panayotis.mertikopoulos@imag.fr).

Abstract: We develop a new stochastic algorithm with variance reduction for solving pseudo-monotone 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 with pseudo-monotone and 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 variance reduction mechanism, and leads to a.s. convergence to solutions of a merely pseudo-monotone stochastic variational inequality problem. To the best of our knowledge, this is the first stochastic algorithm achieving this by using only a single projection at each iteration.

Keywords: Variational inequalities; Forward-Backward-Forward Algorithm; Stochastic Approximation, Variance Reduction

1. INTRODUCTION

Several applications in engineering, science, finance and economics lead to a broad range of optimization and equilibrium problems. Under suitable convexity assumptions, the equilibrium conditions of such problems can be compactly formulated as variational inequalities (Facchinei and Pang, 2003). The standard deterministic variational inequality problem, denoted as VI(T, X) (or simply VI), is defined as follows: given a closed convex set $X \subset \mathbb{R}^d$ and a single valued map $T : \mathbb{R}^d \to \mathbb{R}^d$, find $x^* \in X$ such that

$$\langle T(x^*), x - x^* \rangle \geq 0.$$  \hspace{1cm} \text{(VI)}

Call $X$ the set of solutions of VI(T, X). The variational inequality problem includes many important applications in economics, game theory and engineering (see e.g. Scutari et al. (2010); Ravat and Shanbhag (2011); Kannan and Shanbhag (2012); Juditsky et al. (2011); Mertikopoulos and Staudigl (2018)). If $X$ is unbounded it also can be used to model complementarity problems, systems of equations and saddle point problems.

In practice the evaluation of the map $T(x)$ is corrupted by (numerical or random) noise, or it is derived from some other stochastic model, calling for a stochastic analysis of (VI). In the stochastic VI, we start with a measurable set $(\Xi, \mathcal{A})$, a measurable function $F : \mathbb{R}^d \times \Omega \to \mathbb{R}^d$, and a random variable $\xi : (\Omega, \mathcal{F}) \to (\Xi, \mathcal{A})$, defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, such that $F(x, \xi) \in L^1(\Omega; \mathbb{R}^d)$. We let $\mathbb{P} \triangleq \mathbb{P} \circ \xi^{-1}$ be the law of the random variable $\xi$, and define

$$T(x) \triangleq \mathbb{E}_{\xi}[F(x, \xi)] = \int_{\Xi} F(x, z) \, d\mathbb{P}(z).$$  \hspace{1cm} \text{(1)}

The expected value formulation (EV) of the stochastic variational inequality problem, is to find $x^* \in X$ s.t. $(T(x^*), x - x^*) \geq 0 \ \forall x \in X$. (EV)

Since computing the expected value $T(x)$ is rarely possible in practice, advanced stochastic methods for solving EV are formulated without recourse to the mean operator $T$, but rather directly involve the random variable $F(x, \xi)$. Stochastic approximation (SA) theory is the mathematical tool to use in such settings. Recent advances have been made in deriving low complexity schemes for solving stochastic VIs using SA with variance reduction to solve EV even under weak pseudo-monotonicity assumptions on the operator $T$. These advances have been made via stochastic versions of Korpelevich’s extragradient method (Korpelevich, 1976), which read as

$$Y_n = \Pi_X[X_n - \alpha_n A_{n+1}], \ \ X_{n+1} = \Pi_X[X_n - \alpha_n B_{n+1}],$$

where $A_{n+1}$ and $B_{n+1}$ are random estimators of $T(X_n)$ and $T(Y_n)$, respectively. Convergence and computational complexity of this scheme has been thoroughly studied in Yousefi et al. (2014); Iusem et al. (2017). In this paper we present another competitive scheme, based on the forward-backward-forward (FBF) scheme of Tseng (2000). We illustrate the practical advantage of our FBF scheme by tackling an energy efficiency problem in multi-antenna...
communication networks. A more detailed analysis is provided in Boț et al. (2019).

2. THE STOCHASTIC FORWARD-BACKWARD-FORWARD ALGORITHM

The standing hypothesis used in our analysis are summarized as follows.

**Hypothesis 1.** (Consistency). $\mathcal{X} \neq \emptyset$.

**Hypothesis 2.** (Stochastic Model). $\mathcal{X} \subset \mathbb{R}^d$ is closed convex, $(\mathcal{X}, \mathcal{A})$ is a measurable space, with Borel $\sigma$-algebra $\mathcal{A}$, and $F: \mathcal{X} \times \mathbb{R}^d \to \mathbb{R}^d$ is a Carathéodory map (i.e. continuous in $x$, measurable in $\xi$), $\xi$ is a random variable with values in $\mathcal{X}$, defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

**Hypothesis 3.** (Lipschitz continuity). The averaged map $T: \mathcal{X} \to \mathbb{R}^d$ is Lipschitz continuous with modulus $L > 0$.

**Hypothesis 4.** (Pseudo-Monotonicity). The map $T(x) \triangleq \mathbb{E}_\xi[F(x, \xi)]$ is pseudo-monotone on $\mathbb{R}^d$.

$(T(x), y - x) \geq 0 \Rightarrow \langle y, y - x \rangle \geq 0$.

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

$$A_{n+1}(x) \triangleq \frac{1}{m_{n+1}} \sum_{i=1}^{m_{n+1}} F(x, \xi^{(i)}_{n+1}) \quad x \in \mathbb{R}^d,$$

where $\xi_i = (\xi^{(1)}_i, \ldots, \xi^{(m_{n+1})}_i)$ is an i.i.d draw from $\mathbb{P}$. The sequence $(m_{n})_{n \geq 1} \subset \mathbb{N}$ determines the sample rate, or batch size, of the SO.

**Hypothesis 5.** (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 $(m_{n})_{n \geq 1}$ to cope with Assumption 5 is that for some constant $\zeta > 0$ and integer $n_0 > 0$, we have $m_n = \zeta \cdot (n + n_0)^{1+\alpha} \ln(n + m_0)^{1+\beta}$, for $a > 0$ and $b \geq -1$, or $a = 0$ and $b > 0$. Approximations of the form (2) have received considerable interest in machine learning and computational statistics (see e.g. Atchadé et al. (2017)).

**Hypothesis 6.** (Stepsize choice). The stepsize $(\alpha_n)_{n \geq 0}$ in Algorithm 1 satisfies

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

For $n \geq 0$, we introduce the approximation error $\xi_{n+1} \triangleq A_{n+1} - T(X_n)$, and $Z_{n+1} \triangleq B_{n+1} - T(Y_{n})$. (6) The next hypothesis imposes a control on the SO’s variance.

**Hypothesis 7.** (Variance Control). There exists $p \geq 2$, $x^* \in \mathcal{X}$, and $\sigma(x^*) > 0$ such that for all $x \in \mathbb{R}^d$

$$\mathbb{E}[\|F(x, \xi) - T(x)\|^p]^{1/p} \leq \sigma(x^*) + \sigma_0 \|x - x^*\|.$$  

This Hypothesis considerably weakens the standard assumption in stochastic optimization of uniformly bounded oracle variance (see Boț et al. (2019) for a thorough discussion). Given the batch size sequence $(m_{n})_{n \geq 1}$, introduce two stochastic processes $\xi_n, \eta_n$ such that $\xi_n := (\xi^{(1)}_n, \ldots, \xi^{(m_{n})}_n)$, and $\eta_n := (\eta^{(1)}_n, \ldots, \eta^{(m_{n})}_n)$. Define the filtration $(\mathcal{F}_n)_{n \geq 0}$, by $\mathcal{F}_0 = \sigma(X_0)$, and $\mathcal{F}_n = \sigma(X_0, \xi_1, \xi_2, \ldots, \xi_n, \eta_1, \ldots, \eta_n)$.

### Algorithm 1 Stochastic Tseng-Forward-Backward-Forward method (SFBF)

**Require:** step-size sequence $(\alpha_n)_{n \geq 0}$; batch size sequence $(m_{n})_{n \geq 1}$; probability measure $\mathbb{P}$

1. initialize $X^0 \sim \mu$ \hspace{1cm} \# initialization
2. for $n \geq 0$ do
3. Given $X_n$, draw $\xi_{n+1} = (\xi^{(i)}_n)_{1 \leq i \leq m_{n+1}}$ and $\eta_{n+1} = (\eta^{(i)}_n)_{1 \leq i \leq m_{n+1}} \sim \mathbb{P}$
4. Oracle returns $A_{n+1} = \frac{1}{m_{n+1}} \sum_{i=1}^{m_{n+1}} F(X_n, \xi^{(i)}_{n+1})$.
5. \# First Oracle query
6. Oracle returns $B_{n+1} = \frac{1}{m_{n+1}} \sum_{i=1}^{m_{n+1}} F(Y_n, \eta^{(i)}_{n+1})$.
7. Compute $Y_{n+1} = \Pi_T(X_n - \alpha_n A_{n+1})$ \hspace{1cm} \# Forward step
8. Compute $X_{n+1} = Y_{n+1} + \alpha_n (A_{n+1} - B_{n+1})$ \hspace{1cm} \# Backward step
9. $n \leftarrow n + 1$ \hspace{1cm} \# next stage
end for

Hypothesis (7), coupled with eqs. (3) and (4), imply an online variance reduction scheme, as illustrated in the following Lemma.

**Lemma 8.** Let $p \geq 2$ be as in Hypothesis 7. For all $n \geq 0, p' \in [2, p]$ we have

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

and

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

for universal constants $C_{p'} > 0$.

**Proof.** We prove both statements via the verification of a general result. Let $N \in \mathbb{N}$ and $\xi^{(1)}, \ldots, \xi^{(N)}$ be an i.i.d sample from the measure $\mathbb{P}$. Define the process $(M^N_i(x))_{i=0}^N$ by $M_0(x) \triangleq 0$, and for $1 \leq i \leq N$

$$M^N_i(x) \triangleq \frac{1}{N} \sum_{n=1}^{i} \left(F(x, \xi^{(n)}) - T(x)\right).$$

We claim that, for all $1 \leq q \leq p, N \in \mathbb{N}$, and $x \in \mathbb{R}^d$, we have

$$\mathbb{E}[\|M^N_i(x)||q\|^{q/2}] \leq \frac{C_q}{\sqrt{N}} (\sigma(x^*) + \sigma_0 ||x - x^*||).$$

For $i \in \{1, 2, \ldots, N\}$, the monotonicity of $L^p(\Omega, \mathcal{F}, \mathbb{P})$ and (7) implies that

$$\mathbb{E}[\|\Delta M^N_{i-1}(x)||q\|^{q/2}] \leq \frac{1}{N}\mathbb{E}[\|F(x, \xi^{(i)}) - T(x)||q\|^{q/2}] \leq \frac{1}{N}\mathbb{E}[\|F(x, \xi^{(i)}) - T(x)||q\|^{q/2}] \leq \frac{\sigma(x^*) + \sigma_0 ||x - x^*||}{N}.$$  

Using this, together with the Burkholder-Davis-Gundy inequality (Stroock, 2011), there exists constants $C_q > 0$ such that
\[ \mathbb{E} \left[ |M_N^N(x)|^q \right]^{1/q} \leq C_q \sum_{k=1}^N \mathbb{E} \left( \left| \frac{F(x, \xi(k)) - T(x)}{N} \right|^q \right)^{2/q} \]
\[ \leq C_q (\sigma(x^*) + \sigma_0 \|x - x^*\|) \right) \sqrt{N}. \]

3. CONVERGENCE ANALYSIS

We can give a full convergence proof of the stochastic process \( \{X_k, Y_k; k \in \mathbb{N}\} \) generated by SFBBF (Algorithm 1). To measure the progress of the algorithm, we need to introduce a merit function. For our purposes, the most convenient choice for a merit function is the residual function
\[ r_\alpha(x) : = \|x - \Pi_X(x - \alpha T(x))\| \quad \forall x \in \mathbb{R}^d. \] (8)
Define \( \rho_\alpha \equiv 1 - 2L^2\alpha^2 \) for all \( \alpha \geq 0 \). Our analysis starts by verifying a stochastic quasi Fejér property of the sequence \( \{\|X_n - x^*\|^2\}_{k=0}^\infty \).

Lemma 9. For all \( x^* \in X_0 \), we have
\[ \mathbb{E}[\|X_{n+1} - x^*\|^2 | F_n] \leq \|X_n - x^*\|^2 - \rho_n^2 r_\alpha(x)^2 + \frac{\kappa_n}{m+1} \left[ \sigma_0^2 + \|X_n - x^*\|^2 + \sigma(x)^2 \right], \]
where
\[ \kappa_n = \sigma_0^2 c_2^2 [2(4 + \rho_n) + 16(1 + \alpha_n L + \sigma_0 \alpha_n c_2^2 / \sqrt{m+1})^2], \] and \( C_2 > 0 \) is a constant.

Proof. Since the proof is quite long and tedious, we only outline the main steps. The full proof can be found in Bot et al. (2019). We start with verifying the recursion
\[ \|X_{n+1} - x^*\|^2 \leq \|X_n - x^*\|^2 - \frac{\rho_n}{2} r_\alpha(x)^2 + \Delta U_n(x^*) + \Delta V_n, \]
where
\[ \Delta V_n \equiv V_{n+1} - V_n = (4 + \rho_n)\alpha_n^2 \|W_{n+1}\|^2 + 4\alpha_n^2 \|Z_{n+1}\|^2, \] and
\[ \Delta U_n(x) \equiv 2\alpha_n \langle Z_{n+1}, x - Y_n \rangle. \] The proof of this recursive relation proceeds via several algebraic steps.

Step 1: By definition of a point \( x^* \in X_0 \), we have
\[ \langle T(x^*), y - x^* \rangle \geq 0 \quad \forall y \in X. \]
Set \( y = Y_n \), and using \( \alpha_n > 0 \) as well as pseudo-monotonicity (Hypothesis 4), we see
\[ \langle \alpha_n T(Y_n), Y_n - x^* \rangle \geq 0 \quad \forall n \geq 0. \]

Using the Doob decomposition eq. (6), we can rewrite this inequality as
\[ \langle \alpha_n B_{n+1}, Y_n - x^* \rangle \geq \alpha_n \langle Z_{n+1}, Y_n - x^* \rangle. \] (9)
Since \( Y_n = \Pi_X(x_n - \alpha_n A_{n+1}) \), properties of the euclidean projection tell us that
\[ \langle x - Y_n, Y_n - X_n + \alpha_n A_{n+1} \rangle \geq 0. \] (10)
Adding equations (9) and (10) and using the definition of the iterate \( X_{n+1} \), gives
\[ \|x^* - Y_n\| \geq \alpha_n \langle Z_{n+1}, Y_n - x^* \rangle. \] (11)

Step 2: For all \( n \geq 0 \), using (11) and the definition of \( X_{n+1} \) in the last equality, we get
\[ \|X_{n+1} - x^*\|^2 = \|X_n - x^*\|^2 - \rho_n^2 r_\alpha(x)^2 + \frac{\kappa_n}{m+1} \left[ \sigma_0^2 + \|X_n - x^*\|^2 + \sigma(x^*) \right], \]
This gives
\[ \|X_{n+1} - x^*\|^2 = \|X_n - x^*\|^2 - \rho_n^2 r_\alpha(x)^2 + \frac{\kappa_n}{m+1} \left[ \sigma_0^2 + \|X_n - x^*\|^2 + \sigma(x^*) \right]. \]

The first inequality is the Cauchy-Schwarz inequality. The second inequality uses the L-Lipschitz continuity of the operator \( T \) (Hypothesis 3), as well as the fact that \( \|a - b\|^2 \leq 2\|a\|^2 + 2\|b\|^2 \). Combining this with the last inequality obtained in Step 2, we see that
\[ \|X_{n+1} - x^*\|^2 \leq \|X_n - x^*\|^2 - (1 - 2L^2\alpha^2)\|Y_n - Y_n\|^2 + 4\alpha_n^2 \|W_{n+1}\|^2 + 4\alpha_n^2 \|Z_{n+1}\|^2 + 2\alpha_n \|A_{n+1} - B_{n+1}, Y_n - Y_n\|^2. \]

Step 4: By definition of the squared residual function, the definition of \( Y_n \), and the non-expansiveness of the euclidean projection, we have
\[ r_\alpha(x) = \|x - \Pi_X(x - \alpha T(x))\|^2 \leq 2\|x - Y_n\|^2 + 2\|Y_n - \Pi_X(x - \alpha T(x))\|^2 \leq 2\|x - Y_n\|^2 + 2\|\alpha_n W_{n+1}\|^2. \]

Step 5: Combining the last inequality from Step 4 with the last inequality from Step 3 (recalling the Step-size condition Hypothesis 6), we conclude
\[ \|X_{n+1} - x^*\|^2 \leq \|X_n - x^*\|^2 - \frac{1}{2} \|1 - 2L^2\alpha^2 r_\alpha(x)^2 + \|W_{n+1}\|^2 + 4\alpha_n^2 \|W_{n+1}\|^2 + 4\alpha_n^2 \|Z_{n+1}\|^2 + 2\alpha_n \|Z_{n+1}, x^* - Y_n\|^2 \]
\[ = \|X_n - x^*\|^2 - \frac{\rho_n}{2} r_\alpha(x)^2 + \frac{\kappa_n}{m+1} \left[ \sigma_0^2 + \|X_n - x^*\|^2 + \sigma(x^*) \right] \]
which can be verified using Lemma 8.

Lemma 9 allows us to proof that the process \( \{X_n\}_{n \geq 0} \) converges a.s. to a random variable \( X^* \) with values in \( \mathcal{X} \), as a consequence of the Robbins-Siegmund Lemma, and general facts due to Combettes and Pesquet (2015).

The next proposition provides explicit norm bounds on the iterates \( \{X_n\}_{n \geq 0} \) in \( L^2(\mathbb{P}) \). These bounds are going to be crucial to assess the convergence rate and the per-iteration complexity of the method.
Proposition 10. Consider Hypothesis 1-7. For all $x^* \in X$, let
\[
\hat{\sigma}(x^*) \triangleq \max\{\gamma(x^*), \sigma_0\}, \quad a(x^*) \triangleq \hat{\sigma}(x^*) \alpha^2 C_2^2 c_1.
\]
Choose $n_0 \in \mathbb{N}$ and $\gamma > 0$ such that $\frac{1}{n_0} \leq \gamma$, and
\[
\beta(x^*) \triangleq \gamma a(x^*) + \gamma^2 a(x^*)^2 \in (0, 1).
\]
Then
\[
\sup_{n \geq n_0} E[\|X_n - x^*\|^2] \leq \frac{E[\|X_{n_0} - x^*\|^2] + 1}{1 - \beta(x^*)}.
\]

Proof. We first remark that for every $\gamma > 0$ we can find an index $n_0 \in \mathbb{N}$ as required, thanks to Hypothesis 5. Call $\psi_n(x^*) = E[\|X_n - x^*\|^2]$. From Proposition 9, we obtain
\[
\psi_n(x^*) \leq \psi_{n_0}(x^*) + \sum_{i=n_0}^{n-1} (1 + \psi_i(x^*)) a(x^*) m_{i+1} + \sum_{i=n_0}^{n-1} (1 + \psi_i(x^*)) a(x^*)^2 c_1 m_{i+1} + \frac{K_1}{m_{i+1}}.
\]
For $p > \psi_{n_0}(x^*)$, define $\tau_p(x^*) \triangleq \inf\{n \geq n_0 | \psi_n(x^*) \geq p\}$. We claim that there exists $\tilde{p} > \psi_{n_0}(x^*)$ such that $\tau_{\tilde{p}}(x^*) = \infty$. Suppose not. Then $\psi_n(x^*) < \infty$ for all $p > \psi_{n_0}(x^*)$. Therefore, by definition of $\tau_p(x^*)$, and the definition of $n_0$, we get
\[
p \leq \psi_{\tau_p}(x^*) \leq \psi_{n_0}(x^*) + \sum_{k=n_0}^{\tau_p(x^*)-1} (1 + \psi_k(x^*)) a(x^*) m_{k+1} + \sum_{k=n_0}^{\tau_p(x^*)-1} (1 + \psi_k(x^*)) \frac{a(x^*)}{c_1 m_{k+1}} + \frac{1}{1 - \gamma a(x^*) - \gamma^2 a(x^*)^2} c_1.
\]
Rearranging, and using $c_1 > 1$ as well as (13), gives
\[
p \leq \frac{\psi_{n_0}(x^*) + 1}{1 - \gamma a(x^*) - \gamma^2 a(x^*)^2} \leq \frac{\psi_{n_0}(x^*) + 1}{1 - \gamma a(x^*) - \gamma^2 a(x^*)^2}.
\]
Since $p > \psi_{n_0}(x^*)$ has been chosen arbitrarily, we can let $p \to \infty$, to arrive at a contradiction. Therefore, there exits $\tilde{p} > \psi_{n_0}(x^*)$ such that $\tilde{p} \triangleq \sup_{n \geq n_0} \psi_n(x^*) \leq \tilde{p} < \infty$. Therefore, for all $n \geq n_0$ we get
\[
\psi_n(x^*) \leq \psi_{n_0}(x^*) + \sum_{k=n_0}^{n-1} (1 + \psi_k(x^*)) a(x^*) m_{k+1} + \sum_{k=n_0}^{n-1} (1 + \psi_k(x^*)) \frac{a(x^*)}{c_1 m_{k+1}} + \frac{1}{1 - \gamma a(x^*) - \gamma^2 a(x^*)^2} c_1.
\]
Taking the supremum over $n \geq n_0$, and shifting back to the original expressions of the involved data, we get
\[
\tilde{p} = \sup_{n \geq n_0} E[\|X_n - x^*\|^2] \leq \frac{E[\|X_{n_0} - x^*\|^2] + 1}{1 - \beta(x^*)}.
\]

We now give explicit estimates on the convergence rate, exhibiting an optimal $O(1/K)$ convergence rate in the mean-square residual function. This result is in line with the stochastic extragradient method (SEG) of Iusem et al. (2017). Without loss of generality, we can assume a constant step size $\alpha \in (0, 1/\sqrt{2L})$. For $n \in \mathbb{N}, \phi \in \mathbb{R}, x^* \in X$, define
\[
\rho = 1 - 2L^2 \alpha^2, \quad \Gamma_n \triangleq \sum_{i=0}^{n} \frac{1}{m_{i+1}}, \quad \Gamma_n^2 \triangleq \sum_{i=0}^{n} \frac{m_{i+1}}{2}, \quad \delta_n(x^*) \triangleq \|X_n - x^*\|^2, \quad H(x^*, n, \phi) \triangleq \frac{1 + \max_{0 \leq i \leq n} E[\delta_i(x^*)]}{1 - \phi - \phi^2}.
\]

Theorem 11. Consider Assumptions 1-7. Let $x^* \in X$, be arbitrarily. Choose $\phi \in (0, \frac{\sqrt{2\rho}}{2L})$ and $n_0 = n_0(x^*)$ to be the first integer such that $\sum_{i=n_0}^{\infty} \frac{1}{m_{i+1}} \leq \frac{\phi}{\alpha a(x^*)}$. Then, for all $\varepsilon > 0$, there exists $N_\varepsilon \in \mathbb{N}$ such that
\[
E[\delta_0(X_{N_\varepsilon})] \leq \frac{\Lambda_\varepsilon(x^*, \phi)}{N_\varepsilon},
\]
where, for all $n \geq 1$,
\[
\Lambda_n(x^*, \phi) \triangleq \frac{1}{\rho} E[\delta_0(x^*)] + \frac{2}{\rho} (1 + H(x^*, n_0(x^*), \phi)) (a(x^*) \Gamma_n + a(x^*)^2 \Gamma_n^2).
\]

Proof. Choosing $\gamma = \frac{\phi}{a(x^*)}$ and $n_0 = n_0(x^*)$ as required in the statement of the Theorem. We use Proposition 10, to get
\[
\sup_{n \geq n_0} E[\delta_n(x^*)] \leq \frac{1}{1 - \phi - \phi^2} \leq H(x^*, n_0(x^*), \phi).
\]
Calling $H(x^*, n_0(x^*), \phi) \equiv H(x^*, \phi)$, it follows
\[
\sup_{n \geq 0} E[\delta_n(x^*)] \leq H(x^*, \phi).
\]
Taking expectation and summing we get from Proposition 9
\[
\rho \sum_{i=0}^{n} E[\delta_0(X_i)] \leq \frac{E[\delta_0(x^*)]}{\rho} + \sum_{i=0}^{n} \frac{K_i}{m_{i+1}} (\sigma(x^*)^2 + \sigma_0^2 E[\delta_i(x^*)]).
\]
First, using the variance bound $\hat{\sigma}(x^*) = \max\{\gamma(x^*), \sigma_0\}$, we get $\kappa_i \leq \alpha^2 C_2^2 c_1 \left(1 + \alpha^2 C_2^2 \frac{a(x^*)^2}{a(x^*)} \right)$. Second, calling $a(x^*) = \alpha^2 \hat{\sigma}(x^*) \alpha^2 C_2^2 c_1$, we get
\[
\rho \sum_{i=0}^{n} E[\delta_0(X_i)] \leq \frac{E[\delta_0(x^*)]}{\rho} + \sum_{i=0}^{n} \frac{a(x^*)}{m_{i+1}} (1 + E[\delta_i(x^*)]) + \sum_{i=0}^{n} \frac{1}{c_1} (\frac{a(x^*)}{m_{i+1}})^2 (1 + E[\delta_i(x^*)]) \leq \left(1 + \max_{0 \leq i \leq n} E[\delta_i(x^*)]\right) (a(x^*) \Gamma_n + a(x^*)^2 \Gamma_n^2),
\]
From (16), and $c_1 > 1$, we conclude
\[
\frac{\rho}{2} \sum_{i=0}^{n} \mathbb{E}[r_{\alpha}(X_i)^2] \leq \mathbb{E}[d_{0}(x^*)] \\
+ (1 + \mathcal{H}(x^*)) \left( \mathbf{a}(x^*) \Gamma_n + \mathbf{a}(x^*) \Gamma_n^2 \right) \\
= \frac{\rho}{2} \Lambda_n(x^*, \phi).
\]
Hence, $\sum_{i=0}^{n} \mathbb{E}[r_{\alpha}(X_i)^2] \leq \Lambda_n(x^*, \phi)$ for all $n \geq 1$, $x^* \in \mathcal{X}$. For all $\varepsilon > 0$, we define the stopping time
\[
N_\varepsilon \triangleq \inf \{n \geq 0 | \mathbb{E}[r_{\alpha}(X_n)^2] \leq \varepsilon \}.
\]
Choose $n = \min\{N_\varepsilon, k\} - 1$ for $k \in \mathbb{N}$, we know that $\sum_{i=0}^{\min\{N_\varepsilon, k\}} \mathbb{E}[r_{\alpha}(X_i)^2] > \varepsilon \min\{N_\varepsilon, k\}$. Since $k \in \mathbb{N}$ is chosen arbitrarily, we can let $k \uparrow \infty$, so that
\[
\varepsilon N_\varepsilon \leq \sum_{i=0}^{N_\varepsilon-1} \mathbb{E}[r_{\alpha}(X_i)^2] \leq \Lambda_{N_\varepsilon-1}(x^*, \phi) \leq \Lambda_\infty(x^*, \phi).
\]
Since this bound holds for all $x^* \in \mathcal{X}$, we conclude
\[
N_\varepsilon \leq \frac{1}{\varepsilon} \inf_{x^* \in \mathcal{X}} \Lambda_{N_\varepsilon}(x^*, \phi) \leq \inf_{x^* \in \mathcal{X}} \Lambda_\infty(x^*, \phi).
\]
Using the definition of the stopping time $N_\varepsilon$ in the above gives the desired result.

4. ENERGY EFFICIENCY IN MULTI-ANTENNA COMMUNICATIONS

Energy efficiency is one of the most important requirements for mobile systems, and it plays a crucial role in preserving battery life and reducing the carbon footprint of multi-antenna devices (i.e., wireless devices equipped with several antennas to multiplex and demultiplex received or transmitted signals). Following Isheden et al. (2012); Feng et al. (2013); Mertikopoulos and Belmega (2016), we consider $K$ wireless devices (e.g., mobile phones), each equipped with $M$ transmit antennas and seeking to connect to a common base-station with $N$ receiver antennas.

In this case, the users’ achievable throughput (received bits/sec) is given by the familiar Shannon-Telatar capacity formula Telatar (1999):
\[
R(X; H) = \log \det \left( \text{Id} + \sum_{k=1}^{K} H_k X_k H_k^H \right)
\]
where:

1. $X_k \in \mathbb{C}^M$ is the input signal covariance matrix of user $k$ and $X = (X_1, \ldots, X_K)$ denotes their aggregate covariance profile (hence Hermitian positive semi-definite).
2. $H_k \in \mathbb{C}^{M \times N}$ is the channel matrix of user $k$, representing the quality of the wireless medium between user $k$ and the receiver.
3. Id is the $N \times N$ identity matrix.

In practice, because of fading and other signal attenuation factors, the channel matrices $H_k$ are random variables, so the users’ achievable throughput is given by
\[
R(X) = \mathbb{E}_H[R(X; H)]
\]
where the expectation is taken over the 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 achieve, i.e.,
\[
EE(X) = \frac{R(X)}{\sum_{k=1}^{K} [P_{k}^c + P_{k}^t]}
\]
where

1. $P_{k}^t$ is the transmit power of the $k$-th device; by elementary signal processing considerations, it is given by $P_{k}^t = \text{tr}(X_k)$.
2. $P_{k}^c > 0$ is a constant representing the total power dissipated in all circuit components of the $k$-th device (mixer, frequency synthesizer, digital-to-analog converter, etc.), except for transmission. For concision, we will also write $P = \sum_{k=1}^{K} P_{k}^c$ 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
\[
\text{tr}(X_k) \leq P_{\text{max}} \quad \text{for all } k = 1, \ldots, K.
\]

Hence, putting all this together, we obtain the stochastic fractional problem:
\[
\begin{align*}
\text{maximize} & \quad EE(X) = \frac{R(X)}{P^c + \sum_{k=1}^{K} \text{tr}(X_k)} \\
\text{subject to} & \quad X_k \succeq 0, \\
& \quad \text{tr}(X_K) \leq P_{\text{max}}.
\end{align*}
\]

The EE objective of this problem (which, formally, has units of bits/Joule) has been widely studied in the literature Cui et al. (2004); Isheden et al. (2012) and it captures the fundamental trade-off between higher spectral efficiency and increased battery life. Importantly, switching from maximization to minimization, we also see that (23) is a fractional programming program of the form quadratic over linear, and hence quasi-convex. Therefore, it can be solved by applying SFBF; in fact, given the costly projection step to the problem’s feasible region, SFBF seems ideally suited to the task.

We do so in a series of numerical experiments illustrated in Figure 1. Specifically, we consider a network consisting of $K = 16$ users, each with $M = 4$ transmit antennas, and a common receiver with $N = 128$ receive antennas. To simulate realistic network conditions, the users’ channel matrices are drawn at each update cycle from a COST Hata radio propagation model with Rayleigh fading Hata (1980); to establish a baseline, we also ran an experiment with static, deterministic channels. For comparison purposes, we ran both SFBF and the SEG of Iusem et al. (2017) with the same variance reduction schedule, the same number of iterations, and step-sizes for both methods as $\alpha_{SEG} = \alpha_{FBF} / \sqrt{3}$, and $\alpha_{FBF} = 10 / N$. Also, to reduce statistical error, we performed $S = 100$ sample runs for each algorithm. We observe that SFBF performs consistently better than SEG, converging to a given target value between 1.5 and 3 times faster.

5. CONCLUSION

In a forthcoming publication Bot et al. (2019) we derive many more characteristics of the algorithm, including explicit bounds on the iterates, and error bounds. We also plan to extend the analysis to settings involving set-valued operators, to capture applications to Generalized Nash equilibrium problems, as illustrated in Grammatico (2017).
REFERENCES

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

Bot, R.I., Staudigl, M., Mertikopoulos, P., and Vuong, P.T. (2019). On the convergence of stochastic forward-backward-forward methods with variance reduction for stochastic variational inequalities. *arXiv preprint arXiv:1902.03355*.

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

Cui, S., Goldsmith, A.J., and Bahai, A. (2004). Energy-efficiency of MIMO and cooperative MIMO techniques in sensor networks. *IEEE Journal on selected areas in communications*, 22(6), 1089–1098.

Facchinei, F. and Pang, J.S. (2003). *Finite-Dimensional Variational Inequalities and Complementarity Problems*. Springer Series in Operations Research. Springer.

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

Grammatico, S. (2017). Proximal dynamics in multi-agent network games. *IEEE Transactions on Control of Network Systems*.

Hata, M. (1980). Empirical formula for propagation loss in land mobile radio services. *IEEE Transactions on Vehicular Technology*, 29(3), 317–325.

Ishedon, C., Chong, Z., Jorswieck, E., and Fettweis, G. (2012). Framework for link-level energy efficiency optimization with informed transmitter. *IEEE Transactions on Wireless Communications*, 11(8), 2946–2957.

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

Juditsky, A., Nemirovski, A., and Tauvel, C. (2011). Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 17–58. doi:10.1214/10-SSY011. URL http://projecteuclid.org/euclid.ss/1393252123.

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

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

Mertikopoulos, P. and Belmega, E.V. (2016). Learning to be green: Robust energy efficiency maximization in dynamic MIMO-OFDM systems. *IEEE Journal on Selected Areas in Communications*, 34(4), 743 – 757.

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

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

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

Stroock, D.W. (2011). *Probability Theory: An Analytic View*. Cambridge University Press, Cambridge, 2nd edition.

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

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

Yousefiyan, F., Nedic, A., and Shanbhag, U.V. (2014). Optimal robust smoothing extragradient algorithms for stochastic variational inequality problems. In *Decision and Control (CDC), 2014 IEEE 53rd Annual Conference on*, 5831–5836. IEEE.