Available online at www.sciencedirect.com

ScienceDirect

IFAC PapersOnline CONFERENCE PAPER ARCHIVE

IFAC PapersOnLine 52-3 (2019) 120–125

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.

© 2019, IFAC (International Federation of Automatic Control) Hosting by Elsevier Ltd. All rights reserved.

Keywords: Variational inequalites; 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$(T, \mathcal{X})$ (or simply VI), is defined as follows: given a closed convex set $\mathcal{X} \subset \mathbb{R}^d$ and a single valued map $T : \mathbb{R}^d \to \mathbb{R}^d$, find $x^* \in \mathcal{X}$ such that

$$\langle T(x^*), x - x^* \rangle \geq 0. \tag{VI}$$

Call $\mathcal{X}_*$ the set of solutions of VI$(T, \mathcal{X})$. 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 $\mathcal{X}$ 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 $\mathsf{P} \triangleq \mathbb{P} \circ \xi^{-1}$ be the law of the random variable ξ, and define

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

The *expected value formulation* (EV) of the stochastic variational inequality problem, is to

$$\text{find } x^* \in \mathcal{X} \text{ s.t } \langle T(x^*), x - x^* \rangle \geq 0 \quad \forall x \in \mathcal{X}. \tag{EV}$$

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

$$Y_n = \Pi_{\mathcal{X}}[X_n - \alpha_n A_{n+1}], \; X_{n+1} = \Pi_{\mathcal{X}}[X_n - \alpha_n B_{n+1}],$$

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 Yousefian 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

⋆ The authors acknowledge financial support from the COST Action CA16228 "European Network for Game Theory". P. Mertikopoulos was partially supported by the French National Research Agency (ANR) grant ORACLESS (ANR16CE33000401).

2405-8963 © 2019, IFAC (International Federation of Automatic Control) Hosting by Elsevier Ltd. All rights reserved.
Peer review under responsibility of International Federation of Automatic Control.
10.1016/j.ifacol.2019.06.021

communication networks. A more detailed analysis is provided in Boţ et al. (2019).

2. THE STOCHASTIC FORWARD-BACKWARD-FORWARD ALGORITHM

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

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

Hypothesis 2. (Stochastic Model). $\mathcal{X} \subset \mathbb{R}^d$ is closed convex, $(\Xi, \mathcal{A})$ is a measurable space, with Borel σ-algebra $\mathcal{A}$, and $F : \mathcal{X} \times \Xi \to \mathbb{R}^d$ is a Carathéodory map (i.e. continuous in x, measurable in ξ). ξ is a random variable with values in Ξ, 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 $\mathbb{R}^d$:

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

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

$$A_{n+1}(x) \triangleq \frac{1}{m_{n+1}} \sum_{i=1}^{m_{n+1}} F(x, \xi_{n+1}^{(i)}) \qquad x \in \mathbb{R}^d, \quad (2)$$

where $\xi_n = (\xi_n^{(1)}, \dots, \xi_n^{(m_n)})$ is an i.i.d draw from 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 $\mathsf{c} > 0$ and integer $n_0 > 0$, we have $m_n = \mathsf{c} \cdot (n + n_0)^{1+a} \ln(n + n_0)^{1+b}$, 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

$$0 < \inf_{n\geq 0} \alpha_n \leq \bar{\alpha} = \sup_{n\geq 1} \alpha_n < \frac{1}{\sqrt{2}L}. \quad (5)$$

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

$$, W_{n+1} \triangleq A_{n+1} - T(X_n), \text{ and } Z_{n+1} \triangleq B_{n+1} - T(Y_n). \quad (6)$$

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 $x \in \mathbb{R}^d$

$$\mathbb{E}[\|F(x,\xi) - T(x)\|^p]^{1/p} \leq \sigma(x^*) + \sigma_0\|x - x^*\|. \quad (7)$$

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

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

```
Require: step-size sequence (α_n)_{n≥0}; batch size
   sequence (m_n)_{n≥1}; probability measure μ
initialize X^0 ~ μ                          # initialization
for n ≥ 0 do
   Given X_n, draw ξ_{n+1} = (ξ_{n+1}^(i))_{1≤i≤m_{n+1}} and
   η_{n+1} = (η_{n+1}^(i))_{1≤i≤m_{n+1}} ~ P
   Oracle returns
        A_{n+1} = (1/m_{n+1}) Σ_{i=1}^{m_{n+1}} F(X_n, ξ_{n+1}^(i)).   (3)
                                    # First Oracle query
   Compute Y_n = Π_X(X_n − α_n A_{n+1})     # Forward
   step
   Oracle returns
        B_{n+1} = (1/m_{n+1}) Σ_{i=1}^{m_{n+1}} F(Y_n, η_{n+1}^(i)).   (4)
                                    # Second Oracle query
   Compute X_{n+1} = Y_n + α_n(A_{n+1} − B_{n+1})
   # Backward step
   n ← n + 1                                # 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

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

and

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

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

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

$$M_i^N(x) \triangleq \frac{1}{N}\sum_{n=1}^{i}\Big(F(x, \xi^{(n)}) - T(x)\Big).$$

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

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

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

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

Using this, together with the Burkholder-Davis-Gundy inequality (Stroock, 2011), there exists constants $C_q > 0$ such that

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

3. CONVERGENCE ANALYSIS

We can give a full convergence proof of the stochastic process $\{(X_k, Y_k); k \in \mathbb{N}\}$ generated by SFBF (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*

$$r_\alpha(x) :\triangleq \|x - \Pi_{\mathcal{X}}(x - \alpha T(x))\| \qquad \forall x \in \mathbb{R}^d. \tag{8}$$

Define $\rho_n \triangleq 1-2L^2\alpha_n^2$ for all $n \geq 0$. Our analysis starts by verifying a stochastic quasi Fejér property of the sequence $\left(\|X_n - x^*\|^2\right)_{k\geq 0}$.

Lemma 9. For all $x^* \in \mathcal{X}_*$, we have

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

where
$\kappa_n = \alpha_n^2 C_2^2[2(4+\rho_n) + 16(1+\alpha_n L + \sigma_0\alpha_n C_2^2/\sqrt{m_{n+1}})^2]$, 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 Boţ et al. (2019). We start with verifying the recursion

$$\|X_{n+1}-x^*\|^2 \leq \|X_n-x^*\|^2 - \frac{\rho_n}{2} r_{\alpha_n}(X_n)^2 + \Delta U_n(x^*) + \Delta V_n,$$

where
$\Delta V_n \triangleq V_{n+1} - V_n = (4+\rho_n)\alpha_n^2\|W_{n+1}\|^2 + 4\alpha_n^2\|Z_{n+1}\|^2$, and $\Delta U_n(x) \triangleq 2\alpha_n\langle Z_{n+1}, x - Y_n\rangle$. The proof of this recursive relation proceeds via several algebraic steps.
Step 1: By definition of a point $x^* \in \mathcal{X}_*$, we have

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

Set $y = Y_n$, and using $\alpha_n > 0$ as well as pseudo-monotonicity (Hypothesis 4), we see

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

Using the Doob decomposition eq. (6), we can rewrite this inequality as

$$\langle \alpha_n B_{n+1}, Y_n - x^*\rangle \geq \alpha_n\langle Z_{n+1}, Y_n - x^*\rangle. \tag{9}$$

Since $Y_n = \Pi_{\mathcal{X}}(X_n - \alpha_n A_{n+1})$, properties of the euclidean projection tell us that

$$\langle x^* - Y_n, Y_n - X_n + \alpha_n A_{n+1}\rangle \geq 0. \tag{10}$$

Adding equations (9) and (10) and using the definition of the iterate X_{n+1}, gives

$$\langle x^* - Y_n, X_{n+1} - X_n\rangle \geq \alpha_n\langle Z_{n+1}, Y_n - x^*\rangle. \tag{11}$$

Step 2: For all $n \geq 0$, using (11) and the definition of X_{n+1} in the last equality, we get

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

This gives

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

Step 3: Using again the definition of X_{n+1}, we see

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

The first inequality is the Cauchy-Schwarz inequality. The second inequality 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

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

Step 4: By definition of the squared residual function, the definition of Y_n, and the non-expansiveness of the euclidean projection, we have

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

Hence,

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

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

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

Now take conditional expectations on both sides, taking into account that $\mathbb{E}[\Delta U_n(x)|\mathcal{F}_n] = 0$, as well as the bound

$$\mathbb{E}[\Delta V_n|\mathcal{F}_n] \leq \frac{\kappa_n}{m_{n+1}}[\sigma_0^2\|X_n - x^*\|^2 + \sigma(x^*)^2]$$

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 $x^* \in \mathcal{X}_*$ let

$$\hat{\sigma}(x^*) := \max\{\sigma(x^*), \sigma_0\},\ \mathtt{a}(x^*) := \hat{\sigma}^2(x^*)\bar{\alpha}^2 C_2^2 \mathtt{c}_1. \quad (12)$$

Choose $n_0 \in \mathbb{N}$ and $\gamma > 0$ such that $\sum_{n\geq n_0} \frac{1}{m_{n+1}} \leq \gamma$, and

$$\beta(x^*) \triangleq \gamma \mathtt{a}(x^*) + \gamma^2 \mathtt{a}(x^*)^2 \in (0,1). \quad (13)$$

Then

$$\sup_{n\geq n_0} \mathbb{E}[\|X_n - x^*\|^2] \leq \frac{\mathbb{E}[\|X_{n_0} - x^*\|^2] + 1}{1-\beta(x^*)}. \quad (14)$$

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 $\psi_n(x^*) = \mathbb{E}[\|X_n - x^*\|^2]$. From Proposition 9, we obtain

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

It is easy to see that there exists a constant $\mathtt{c}_1 > 1$ such that

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

Using this bound, the previous display telescopes to

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

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 $\hat{p} > \psi_{n_0}(x^*)$ such that $\tau_{\hat{p}}(x^*) = \infty$. Suppose not. Then $\tau_p(x^*) < \infty$ for all $p > \psi_{n_0}(x^*)$. Therefore, by definition of $\tau_p(x^*)$, and the definition of n_0, we get

$$\begin{aligned} p \leq \psi_{\tau_p(x^*)}(x^*) &\leq \psi_{n_0}(x^*) + \sum_{k=n_0}^{\tau_p(x^*)-1}(1+\psi_k(x^*))\frac{\mathtt{a}(x^*)}{m_{k+1}} \\ &+ \sum_{k=n_0}^{\tau_p(x^*)-1}(1+\psi_k(x^*))\frac{1}{\mathtt{c}_1}\left(\frac{\mathtt{a}(x^*)}{m_{k+1}}\right)^2 \\ &\leq \psi_{n_0}(x^*) + (1+p)\gamma\mathtt{a}(x^*) + (1+p)\frac{\mathtt{a}(x^*)^2\gamma^2}{\mathtt{c}_1}. \end{aligned}$$

Rearranging, and using $\mathtt{c}_1 > 1$ as well as (13), gives

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

Since $p > \psi_{n_0}(x^*)$ has been chosen arbitrarily, we can let $p \to \infty$, to arrive at a contradiction. Therefore, there exits $\hat{p} > \psi_{n_0}(x^*)$ such that $\bar{p} \triangleq \sup_{n\geq n_0}\psi_n(x^*) \leq \hat{p} < \infty$. Therefore, for all $n \geq n_0$ we get

$$\begin{aligned} \psi_n(x^*) &\leq \psi_{n_0}(x^*) + \sum_{k=n_0}^{n-1}(1+\psi_k(x^*))\frac{\mathtt{a}(x^*)}{m_{k+1}} \\ &+ \sum_{k=n_0}^{n-1}(1+\psi_k(x^*))\frac{1}{\mathtt{c}_1}\left(\frac{\mathtt{a}(x^*)}{m_{k+1}}\right)^2 \\ &\leq \psi_{n_0}(x^*) + (1+\bar{p})\gamma\mathtt{a}(x^*) + (1+\bar{p})\frac{\mathtt{a}(x^*)^2\gamma^2}{\mathtt{c}_1}. \end{aligned}$$

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

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

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 $\alpha \in (0, 1/\sqrt{2}L)$. For $n \in \mathbb{N}, \phi \in \mathbb{R}, x^* \in \mathcal{X}_*$, define

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

Theorem 11. Consider Assumptions 1-7. Let $x^* \in \mathcal{X}_*$ be arbitrarily. Choose $\phi \in (0, \frac{\sqrt{5}-1}{2})$ and $n_0 = n_0(x^*)$ to be the first integer such that $\sum_{i\geq n_0}\frac{1}{m_{i+1}} \leq \frac{\phi}{\mathtt{a}(x^*)}$. Then, for all $\varepsilon > 0$, there exists $N_\varepsilon \in \mathbb{N}$ such that

$$\mathbb{E}[r_\alpha(X_{N_\varepsilon})^2] \leq \frac{\Lambda_\infty(x^*,\phi)}{N_\varepsilon}, \quad (15)$$

where, for all $n \geq 1$,

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

Proof. Choosing $\gamma = \frac{\phi}{\mathtt{a}(x^*)}$ and $n_0 = n_0(x^*)$ as required in the statement of the Theorem. We use Proposition 10, to get

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

Calling $H(x^*, n_0(x^*), \phi) \equiv \mathcal{H}(x^*, \phi)$, it follows

$$\sup_{n\geq 0}\mathbb{E}[\delta_n(x^*)] \leq \mathcal{H}(x^*,\phi). \quad (16)$$

Taking expectation and summing we get from Proposition 9

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

First, using the variance bound $\hat{\sigma}(x^*) = \max\{\sigma(x^*), \sigma_0\}$, we get $\kappa_i \leq \alpha^2 C_2^2 \mathtt{c}_1\left(1 + \frac{\alpha^2 C_2^2 \hat{\sigma}(x^*)^2}{m_{i+1}}\right)$. Second, calling $\mathtt{a}(x^*) = \alpha^2\hat{\sigma}(x^*)^2 C_2^2 \mathtt{c}_1$, we get

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

From (16), and $\mathtt{c}_1 > 1$, we conclude

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

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

$$N_\varepsilon \triangleq \inf\{n \geq 0 | \mathbb{E}[r_\alpha(X_n)^2] \leq \varepsilon\}. \tag{17}$$

Choose $n = \min\{N_\varepsilon, k\} - 1$ for $k \in \mathbb{N}$, we know that $\sum_{i=0}^{\min\{N_\varepsilon,k\}-1}\mathbb{E}[r_\alpha(X_i)^2] > \varepsilon\min\{N_\varepsilon, k\}$. 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

$$N_\varepsilon \leq \frac{1}{\varepsilon}\inf_{x^*\in\mathcal{X}_*}\Lambda_{N_\varepsilon}(x^*,\phi) \leq \inf_{x^*\in\mathcal{X}_*}\Lambda_\infty(x^*,\phi). \tag{18}$$

Using the definition of the stopping time N_ε 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):

$$R(X;H) = \log\det\left(\mathrm{Id} + \textstyle\sum_{k=1}^{K} H_k X_k H_k^\dagger\right) \tag{19}$$

where:

(1) $X_k \in \mathbb{C}^M$ is the *input signal covariance matrix* of user k and $X = (X_1, \dots, X_K)$ 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

$$R(X) = \mathbb{E}_H[R(X;H)] \tag{20}$$

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 achieved, i.e.,

$$\mathrm{EE}(X) = \frac{R(X)}{\sum_{k=1}^{K}[P_k^c + P_k^t]} \tag{21}$$

where

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

$$\mathrm{tr}(X_k) \leq P_{\max} \quad \text{for all } k = 1, \dots, K. \tag{22}$$

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

$$\begin{aligned}\text{maximize} \quad & \mathrm{EE}(X) = \frac{R(X)}{P^c + \sum_{k=1}^{K}\mathrm{tr}(X_k)}\\ \text{subject to} \quad & X_k \succcurlyeq 0,\\ & \mathrm{tr}(X_K) \leq P_{\max}.\end{aligned} \tag{23}$$

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 $\alpha_{EG} = \alpha_{FBF}/\sqrt{3}$, 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 Boţ 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-valuled operators, to capture applications to Generalized Nash equilibrium problems, as illustrated in Grammatico (2017).

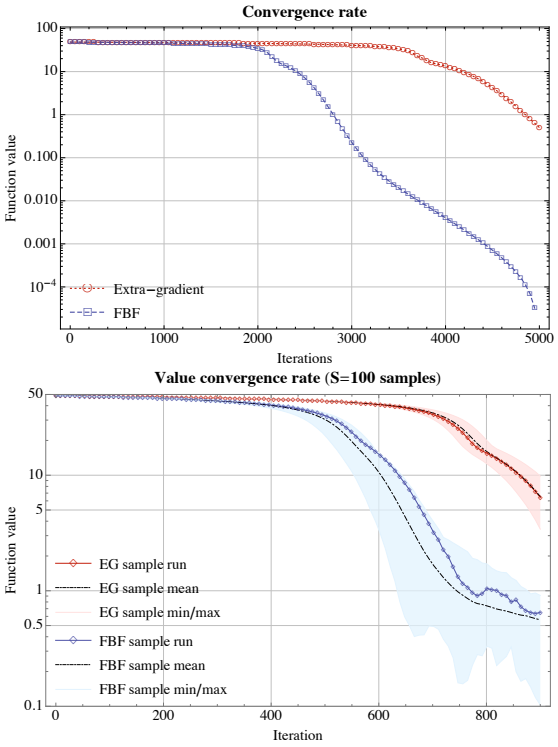

Fig. 1. Comparison of the extra-gradient and FBF methods in problem (23). The top plot, considers static channels, and we ran SFBF and SEG with the same initialization. The plot below, illustrates ergodic channels following a Rayleigh fading model and we performed 100 sample runs for each algorithm; we then plotted a sample run, the sample mean, and the best and worst values at each iteration for each algorithm.

REFERENCES

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

Boţ, 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-fejér block-coordinate fixed point iterations with random sweeping. *SIAM Journal on Optimization*, 25(2), 1221–1248. doi:10.1137/140971233. URL `https://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.

Isheden, 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.ssy/1393252123`.

Kannan, 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`.

Yousefian, F., Nedić, 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.