On the convergence of stochastic forward-backward-forward algorithms with variance reduction in pseudo-monotone variational inequalities

Radu Bot, Panayotis Mertikopoulos, Mathias Staudigl, Phan Vuong

On the convergence of stochastic forward-backward-forward algorithms with variance reduction in pseudo-monotone variational inequalities

Radu Ioan Boţ
University of Vienna
Faculty of Mathematics
r.bot@univie.ac.at

Panayotis Mertikopoulos
Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP
LIG, 38000, Grenoble, France.
panayotis.mertikopoulos@imag.fr

Mathias Staudigl
Maastricht University
Department of Quantitative Economics
m.staudigl@maastrichtuniversity.nl

Phan Tu Vuong
University of Vienna
Faculty of Mathematics
p.vuong@univie.ac.at

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.

1 Introduction

The standard deterministic variational inequality problem, which we will denote as $\mathrm{VI}(T, \mathcal{X})$, or simply VI, is defined as follows: given a closed convex set $\mathcal{X} \subset \mathbb{R}^n$ and a single valued map $T : \mathbb{R}^n \to \mathbb{R}^n$, find $x^* \in \mathcal{X}$ such that

$$\langle T(x^*), x - x^* \rangle \geq 0. \tag{1.1}$$

Call $S(T, \mathcal{X}) \equiv \mathcal{X}_*$ the set of solutions of $\mathrm{VI}(T, \mathcal{X})$. The variational inequality problem includes many interesting applications in economics, game theory and engineering (see e.g. Juditsky et al. [2011], Kannan and Shanbhag [2012], Mertikopoulos and Staudigl [2018], Ravat and Shanbhag [2011], Scutari et al. [2010]). If $\mathcal{X}$ is unbounded it also can 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, physical sciences and economics.

In the stochastic VI problem, we start with a measurable set $(\Xi, \mathcal{A})$, and measurable function $F : \mathbb{R}^n \times \Omega \to \mathbb{R}^n$ and a random variable $\xi : (\Omega, \mathcal{F}) \to (\Xi, \mathcal{A})$, defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that $F(x, \xi) \in L^1(\Omega; \mathbb{R}^n)$. We let $\mathsf{P} = \mathbb{P} \circ \xi^{-1}$ be the law of the random variable ξ

Abstract for: Smooth Games Optimization and Machine Learning Workshop (NIPS 2018), Montréal, Canada.

on $(\Xi, \mathcal{A})$, and define

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

The *expected value formulation* (EV) of the stochastic variational inequality problem, is to find $x^* \in \mathcal{X}$ such that $\langle T(x^*), x - x^* \rangle \geq 0$ for all $x \in \mathcal{X}$.

2 Setup and preliminaries

A map $H : C \to \mathbb{R}^n$ is pseudo-monotone if

$$\langle H(x), y - x \rangle \geq 0 \Rightarrow \langle H(y), y - x \rangle \geq 0 \tag{2.1}$$

Pseudo-monotonicity is a weakened notion of monotonicity in variational analysis.[1] If $T = \nabla f$ then T is pseudo-monotone whenever f is quasi-convex. The *Minty Lemma* implies that

$$S(H, C) = \{x \in C | (\forall p \in C) : \langle H(p), p - x \rangle \geq 0\}.$$

3 The stochastic forward-backward-forward algorithm

The standing hypothesis we use in our analysis are summarized in the following paragraph.

Assumption 1 (Consistency)**.** The solution set $\mathcal{X}_* = S(T, \mathcal{X})$ is nonemtpy.

Assumption 2 (Stochastic Model)**.** $\mathcal{X} \subset \mathbb{R}^n$ is closed convex, $(\Xi, \mathcal{A})$ is a measurable space and $F : \mathcal{X} \times \Xi \to \mathbb{R}^n$ 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 $(\Omega, \mathcal{F}, \mathbb{P})$.

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

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

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

$$A_{n+1}(x) \triangleq \frac{1}{m_{n+1}} \sum_{i=1}^{m_{n+1}} F(x, \xi_{n+1}^{(i)}) \qquad x \in \mathbb{R}^n. \tag{3.1}$$

The sequence $(m_n)_{n\geq 1} \subset \mathbb{N}$ determines the *sample rate*, or batch size, of the SO. The random sequence $\xi_n = (\xi_n^{(1)}, \ldots, \xi_n^{(m_n)})$ is an i.i.d draw from P. Approximations of the form (3.1) have received some considerable in machine learning and computational statistics (see e.g. Atchadé et al. [2017], and references therein). The implicit assumption on the SO standing behind (3.1) is that it is possible to obtain i.i.d samples from the measure P. As in the extragradient method (**EG**), the stochastic forward-backward-forward (**SFBF**) method of Tseng type requires two queries from the SO. Its pseudocode is given in Algorithm 1. In particular, given the batch size sequence $(m_n)_{n\geq 1}$, introduce two stochastic processes ξ_n, η_n such that

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

and define the sub-sigma algebras $(\mathcal{F}_n)_{n\geq 0}, (\hat{\mathcal{F}}_n)_{n\geq 0}$ by $\mathcal{F}_0 = \sigma(X_0)$, and $\mathcal{F}_n = \sigma(X_0, \xi_1, \xi_2, \ldots, \xi_n, \eta_1, \ldots, \eta_n)$ and $\hat{\mathcal{F}}_n = \sigma(X_0, \xi_1, \xi_2, \ldots, \xi_{n+1}, \eta_1, \ldots, \eta_n)$.

Assumption 5 (Stepsize choice)**.** The stepsize sequence $(\alpha_n)_{n\geq 0}$ in Algorithm 1 satisfies

$$0 < \inf_{n\geq 0} \alpha_n \leq \bar{\alpha} = \sup_{n\geq 1} \alpha_n < \frac{1}{\sqrt{2}L}. \tag{3.4}$$

[1] The strongest, and most used assumption is *strong monotonicity*: $\langle H(y) - H(x), x - y \rangle \geq \lambda \|x - y\|^2$ for some $\lambda \geq 0$. This clearly implies *monotonicity*: $\langle H(y) - H(x), x - y \rangle \geq 0$, which in turn implies pseudo-monotonicity. None of the reverse implications is true.

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 μ

initialize $X^0 \sim \mu$ # initialization
for $n \geq 0$ **do**
  Given X_n, draw $\xi_{n+1} = (\xi_{n+1}^{(i)})_{1\leq i\leq m_{n+1}}$ and $\eta_{n+1} = (\eta_{n+1}^{(i)})_{1\leq i\leq m_{n+1}} \sim \mathsf{P}$
  Oracle returns

$$A_{n+1} = \frac{1}{m_{n+1}} \sum_{i=1}^{m_{n+1}} F(X_n, \xi_{n+1}^{(i)}). \tag{3.2}$$

  # First Oracle query
  Compute $Y_n = \Pi_{\mathcal{X}}(X_n - \alpha_n A_{n+1})$ # Forward step
  Oracle returns

$$B_{n+1} = \frac{1}{m_{n+1}} \sum_{i=1}^{m_{n+1}} F(Y_n, \eta_{n+1}^{(i)}). \tag{3.3}$$

  # Second Oracle query
  Compute $X_{n+1} = Y_n + \alpha_n(A_{n+1} - B_{n+1})$ # Backward step
  $n \leftarrow n+1$ # next stage
end for

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

$$W_{n+1} \triangleq A_{n+1} - T(X_n), \text{ and } Z_{n+1} \triangleq B_{n+1} - T(Y_n), \tag{3.5}$$

One can check that $(Y_n)_{n\in\mathbb{N}_0}$ is measurable with respect to the sub-sigma algebra $(\hat{\mathcal{F}}_n)_{n\in\mathbb{N}_0}$ and $(X_n)_{n\in\mathbb{N}_0}$ is measurable with respect to the sub-sigma algebra $(\mathcal{F}_n)_{n\in\mathbb{N}_0}$. The next assumption is essentially the same as the variance control assumption in Iusem et al. [2017].

Assumption 6 (Variance Control)**.** There exists $p \geq 2$ and $x^* \in \mathcal{X}_*$ and $\sigma(x^*) > 0$ such that for all $x \in \mathbb{R}^n$

$$\mathbb{E}[\|F(x,\xi) - T(x)\|^p]^{1/p} \leq \sigma(x^*) + \sigma_0\|x - x^*\|. \tag{3.6}$$

This Assumption considerably weakens the standard assumption in stochastic optimization of uniformly bounded oracle variance (UVB). Since SFBF is an infeasible method, we have to make assumption on the SO variance on the full domain $\mathbb{R}^n$, which makes (UVB) an extremely restrictive assumption. In fact, if $\mathcal{X}$ is unbounded, estimates of the form (3.6) are the most natural ones, as also argued in Iusem et al. [2017]. It can be shown that estimate (3.6) holds if the map $x \mapsto F(x,\xi)$ is random Lipschitz with a Lipschitz modulus $\mathcal{L}(\xi) \in L^p(\mathbb{P})$. Assumption (6), coupled with eqs. (3.2) and (3.3), imply an online variance reduction scheme, as made precise in the following Lemma.

Lemma 3.1. *Let $p \geq 2$ be as in Assumption 6. For all $n \geq 0, p' \in [2,p]$ we have*

$$\mathbb{E}[\|W_{n+1}\|^{p'}|\mathcal{F}_n]^{\frac{1}{p'}} \leq \frac{C_{p'}\left(\sigma(x^*) + \sigma_0\|X_n - x^*\|\right)}{\sqrt{m_{n+1}}} \tag{3.7}$$

and

$$\mathbb{E}[\|Z_{n+1}\|^{p'}|\mathcal{F}_n]^{\frac{1}{p'}} \leq \frac{C_{p'}}{\sqrt{m_{n+1}}}\left(\sigma(x^*) + \sigma_0\mathbb{E}[\|Y_n - x^*\|^{p'}|\mathcal{F}_n]^{\frac{1}{p'}}\right). \tag{3.8}$$

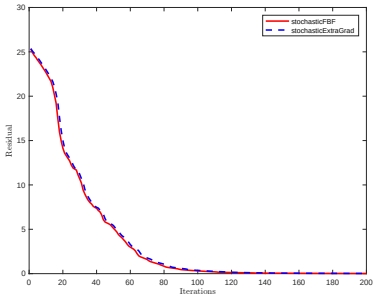
Variance reduction techniques as the above have been successfully used in stochastic variational problems in Palaniappan and Bach [2016] and Shi et al. [2017] in the context of convex-concave saddle-point problems, and Iusem et al. [2017] in the context of variational inequalities. The latter paper gives a thorough discussion when Assumption 6 is appropriate.

4 Convergence Analysis

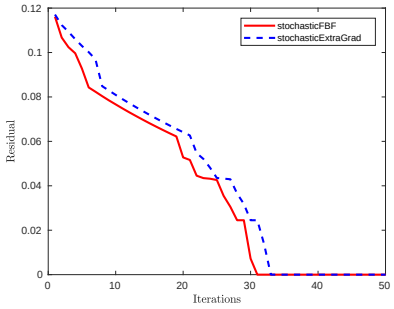
We can give a full convergence proof of the stochastic process $\{(X_k, Y_k); k \in \mathbb{N}\}$ generated by Algorithm 1.[2] To measure the progress of SFBF, we need to introduce a merit function. For our purposes, the most convenient choice for a merit function is the *residual function*

$$r_\alpha(x) \triangleq \|x - \Pi_{\mathcal{X}}(x - \alpha T(x))\| \qquad \forall x \in \mathbb{R}^n. \tag{4.1}$$

[2] All proofs of the announced results are available upon request.



(a) SFBF vs. EG for fractional program (5.1) [$n = 50$].



(b) SFBF vs. EG in bimatrix games.

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

Proposition 4.1. *For all $x^* \in \mathcal{X}_*$, we have*

$$\mathbb{E}[\|X_{n+1} - x^*\|^2|\mathcal{F}_n] \leq \|X_n - x^*\|^2 - \frac{\rho_n}{2} r_{\alpha_n}(X_n)^2 + \frac{\kappa_n}{m_{n+1}} \left[\sigma_0^2 \|X_n - x^*\|^2 + \sigma(x^*)^2\right], \quad (4.2)$$

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.

Proposition 4.1 allows us to deduce that the process $(X_n)_{n\geq 0}$ converges a.s. to a random variable X with values in the set $\mathcal{X}_*$ as a consequence of the classical Robbins-Siegmund Lemma, and general facts due to Combettes and Pesquet [2015]. More precisely, define the random set of cluster points $\mathsf{Lim}(X)(\omega) \triangleq \{x \in \mathbb{R}^n | (\exists (n_j) \uparrow \infty) : \lim_{n_j \to \infty} X_{n_j}(\omega) = x\}$. Then we can show that $\mathsf{Lim}(X)(\omega) \subset \mathcal{X}_*$ for almost all $\omega \in \Omega$. In particular, this result holds for constant step size policies $\alpha_n \equiv \alpha \in [\underline{\alpha}, \bar{\alpha}]$, with step size bounds as specified in Assumption 5.

The theoretical complexity and the provable rate of convergence is very similar to stochastic EG. Our constants in the estimates are, however, always smaller. As shown in the numerical experiments we performed, this means that SFBF will never be slower than stochastic EG, but asymptotically its behavior is very similar. Hence, we have developed a method with (i) lower per-iteration complexity, and (ii) similar convergence rate than EG, with smaller constant factors in the estimates.

Proposition 4.2. *Consider Assumption 1-6. Let $\phi \in (0, \frac{\sqrt{5}-1}{2})$, and $\varepsilon > 0$ be given, and assume that the step-size is constant $\alpha_n \equiv \alpha \in (\underline{\alpha}, \bar{\alpha})$. Define $N_\varepsilon = \inf\{n \geq 0 | \mathbb{E}[r_\alpha(X_{N_\varepsilon})^2] \leq \varepsilon\}$. For every $x^* \in \mathcal{X}_*$ as guaranteed in Assumption 6. There is an integer $n_0 = n_0(x^*)$ and a constant $\mathsf{Q}(x^*, \phi, n_0)$ such that*

$$\mathbb{E}[r_\alpha(X_{N_\varepsilon})^2] \leq \frac{\mathsf{Q}(x^*, \phi, n_0)}{N_\varepsilon}. \quad (4.3)$$

5 Numerical experiments

We have tested SFBF on a stochastic fractional program problem of the form

$$\min_{x \in \mathcal{X}} \left\{ f(x) = \mathbb{E}\left[\frac{G(x,\xi)}{h(x)}\right] \right\} \quad (5.1)$$

where $G(x,\xi) = x^\top Q(\xi) x + c(\xi)^\top x + q(\xi)$ with $Q(\xi)$ a positive semi-definite random matrix with positive semi-definite mean Q, and $h(x) = a^\top x + b > 0$ for $x \in \mathcal{X}$. We compared the average performance of SFBF with the stochastic EG method of Iusem et al. [2017] using a *constant step size*. Figure 1a shows the numerical comparison. As a second numerical test, we have solved random bi-matrix games with payoff matrices (U_I, U_{II}) with SFBF, using the formulation of Nash equilibrium as a complementarity problem, as described in Von Stengel [2002]. Figure 1b shows the results obtained, showing the clear superiority of the SFBF compared to EG. In our opinion these results are very supportive for our approach.

References

Yves F Atchadé, Gersende Fort, and Eric Moulines. On perturbed proximal gradient algorithms. *J. Mach. Learn. Res*, 18(1):310–342, 2017.

P. Combettes and J. Pesquet. Stochastic quasi-fejér block-coordinate fixed point iterations with random sweeping. *SIAM Journal on Optimization*, 25(2):1221–1248, 2018/09/20 2015. doi: 10.1137/140971233. URL https://doi.org/10.1137/140971233.

Francisco Facchinei and Jong-shi Pang. *Finite-Dimensional Variational Inequalities and Complementarity Problems - Volume I and Volume II*. Springer Series in Operations Research, 2003.

AN Iusem, Alejandro Jofré, Roberto I Oliveira, and Philip Thompson. Extragradient method with variance reduction for stochastic variational inequalities. *SIAM Journal on Optimization*, 27(2): 686–724, 2017.

Anatoli Juditsky, Arkadi Nemirovski, and Claire Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. pages 17–58, 2011. doi: 10.1214/10-SSY011. URL http://projecteuclid.org/euclid.ssy/1393252123.

A. Kannan and U. Shanbhag. Distributed computation of equilibria in monotone nash games via iterative regularization techniques. *SIAM Journal on Optimization*, 22(4):1177–1205, 2017/12/28 2012. doi: 10.1137/110825352. URL https://doi.org/10.1137/110825352.

Panayotis Mertikopoulos and Mathias Staudigl. Stochastic mirror descent dynamics and their convergence in monotone variational inequalities. *Journal of Optimization Theory and Applications*, (to appear), 2018.

Balamurugan Palaniappan and Francis Bach. Stochastic variance reduction methods for saddle-point problems. In *Advances in Neural Information Processing Systems*, pages 1416–1424, 2016.

U. Ravat and U. Shanbhag. On the characterization of solution sets of smooth and nonsmooth convex stochastic nash games. *SIAM Journal on Optimization*, 21(3):1168–1199, 2017/12/29 2011. doi: 10.1137/100792644. URL https://doi.org/10.1137/100792644.

Gesualdo Scutari, Daniel P Palomar, Francisco Facchinei, and Jong-shi Pang. Convex optimization, game theory, and variational inequality theory. *IEEE Signal Processing Magazine*, 27(3):35–49

Zhan Shi, Xinhua Zhang, and Yaoliang Yu. Bregman divergence for stochastic variance reduction: saddle-point and adversarial prediction. In *Advances in Neural Information Processing Systems*, pages 6031–6041, 2017.

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