Extragradient Method with Variance Reduction for Stochastic Variational Inequalities

A. Iusem, A. Jofré, R. Imbuzeiro, P. Thompson

June 14, 2016
1. Formulation of the Stochastic VIP

2. Methods for VI and SVI

3. An extragradient method for SVI
The standard (deterministic) variational inequality problem, which we will denote as $\text{VI}(T, X)$, is defined as follows: given a closed and convex set $X \subset \mathbb{R}^n$ and an operator $T : \mathbb{R}^n \to \mathbb{R}^n$, find $x^* \in X$ such that $\langle T(x^*), x - x^* \rangle \geq 0$ for all $x \in X$. 
The standard (deterministic) variational inequality problem, which we will denote as \( \text{VI}(T, X) \), is defined as follows: given a closed and convex set \( X \subset \mathbb{R}^n \) and an operator \( T : \mathbb{R}^n \to \mathbb{R}^n \), find \( x^* \in X \) such that \( \langle T(x^*), x - x^* \rangle \geq 0 \) for all \( x \in X \).

The variational inequality problem includes many interesting special instances with applications in economics, game theory and engineering. The basic prototype is the smooth convex optimization problem \( \min f(x) \) s.t. \( x \in X \), for some \( f : \mathbb{R}^n \to \mathbb{R} \), with \( T = \nabla f \).
The standard (deterministic) variational inequality problem, which we will denote as $\text{VI}(T, X)$, is defined as follows: given a closed and convex set $X \subset \mathbb{R}^n$ and an operator $T : \mathbb{R}^n \to \mathbb{R}^n$, find $x^* \in X$ such that $\langle T(x^*), x - x^* \rangle \geq 0$ for all $x \in X$.

The variational inequality problem includes many interesting special instances with applications in economics, game theory and engineering. The basic prototype is the smooth convex optimization problem $\min f(x)$ s.t. $x \in X$, for some $f : \mathbb{R}^n \to \mathbb{R}$, with $T = \nabla f$.

Other particular cases of variational inequalities problems are complementarity problems, minimax problems and many equilibrium problems, e.g., Nash Equilibrium problems and traffic network Wardrop equilibrium problems.
In the stochastic case, we take a measurable space $(\Xi, \mathcal{G})$, a measurable (random) operator $F : \Xi \times \mathbb{R}^n \to \mathbb{R}^n$ and a random variable $\xi : \Omega \to \Xi$ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, which induces an expectation $\mathbb{E}$ and the distribution $\mathbb{P}_\xi$ of $\xi$. 
In the stochastic case, we take a measurable space $(\Xi, G)$, a measurable (random) operator $F : \Xi \times \mathbb{R}^n \to \mathbb{R}^n$ and a random variable $\xi : \Omega \to \Xi$ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, which induces an expectation $\mathbb{E}$ and the distribution $\mathbb{P}_{\xi}$ of $\xi$.

We assume that for every $x \in \mathbb{R}^n$, $F(\xi, x) : \Omega \to \mathbb{R}^n$ is an integrable random vector.
In the stochastic case, we take a measurable space \((\Xi, \mathcal{G})\), a measurable (random) operator \(F : \Xi \times \mathbb{R}^n \to \mathbb{R}^n\) and a random variable \(\xi : \Omega \to \Xi\) defined on a probability space \((\Omega, \mathcal{F}, \mathbb{P})\), which induces an expectation \(\mathbb{E}\) and the distribution \(\mathbb{P}_\xi\) of \(\xi\).

We assume that for every \(x \in \mathbb{R}^n\), \(F(\xi, x) : \Omega \to \mathbb{R}^n\) is an integrable random vector.

The solution criterion analyzed in this paper consists of solving the deterministic VI\((T, X)\), where \(T : \mathbb{R}^n \to \mathbb{R}^n\) is the expected value of \(F(\xi, \cdot)\), i.e., for any \(x \in \mathbb{R}^n\), \(T(x) = \mathbb{E}[F(\xi, x)]\).
Precisely, the definition of stochastic variational inequality problem (SVI) is:
Precisely, the definition of stochastic variational inequality problem (SVI) is:

**Definition**

Find $x^* : \Omega \rightarrow X$, such that $\langle T(x^*), x - x^* \rangle \geq 0$ holds almost surely for any $x \in X$, with $T(x) = \mathbb{E}[F(\xi, x)]$. 
Precisely, the definition of stochastic variational inequality problem (SVI) is:

**Definition**

Find $x^* : \Omega \rightarrow X$, such that $\langle T(x^*), x - x^* \rangle \geq 0$ holds almost surely for any $x \in X$, with $T(x) = \mathbb{E}[F(\xi, x)]$.

We remark that in the above definition, the solution is a random variable $x^* : \Omega \rightarrow X$ that satisfies the (deterministic) VI($T, C$) with probability 1. This formulation of SVI is called the expected value formulation. It was first proposed by S. Robinson in 1996.
Methods for the deterministic VI($T, X$) are well known. If $T$ is available then SVI can be solved by these methods, but in most cases $T$ is not available, because:

a) $P_{\xi}$ is known but the integration required for the computation of $T$ is computationally expensive,

b) $P_{\xi}$ is not known, so that the information on $\xi$ can be only obtained using past data or sampling,

c) $F$ has no closed form and can be only ‘accessed’ by an auxiliary problem (often called stochastic oracle), which gives values $F(\eta, x)$ of $F$ given a sample $\eta$ of $\xi$. 
We will use the stochastic approximation (SA) approach for removing this obstacle. It consists of replacing the exact mean operator $T$ in each iteration by a random sample of $F$. 
We will use the **stochastic approximation (SA)** approach for removing this obstacle. It consists of replacing the exact mean operator $T$ in each iteration by a random sample of $F$.

This approach induces an stochastic error $F(\xi, x) - T(x)$ along the trajectory of the method.
We will use the stochastic approximation (SA) approach for removing this obstacle. It consists of replacing the exact mean operator $T$ in each iteration by a random sample of $F$.

This approach induces an stochastic error $F(\xi, x) - T(x)$ along the trajectory of the method.

The SA methodology was first proposed by Robbins and Monro in 1951 for stochastic optimization problems and extended to SVI by Jiang and Xu in 2008.
An additional difficulty is the possibly complicated structure of the feasible set $X$. Often, the feasible set takes the form $X = \cap_{i=1}^{m} X_i$, where each $X_i$ is a closed convex set.
An additional difficulty is the possibly complicated structure of the feasible set $X$. Often, the feasible set takes the form $X = \cap_{i=1}^{m} X_i$, where each $X_i$ is a closed convex set.

Frequently, the orthogonal projection $P_i$ onto each $X_i$ is relatively easy to compute, but the projection $P$ onto $X$ is hard (e.g., when the $X_i$’s are hyperplanes). In this case, an interesting alternative consists of replacing in each iteration $P$ by some of the $P_i$’s, chosen according to a certain rule. This is the idea behind the so called row action methods which we will use here for SVI’s.
The classical projection method for $\text{VI}(T, X)$, akin to the projected gradient method for convex optimization, is the projection algorithm:

$$x^{k+1} = \Pi[x^k - \alpha_k T(x^k)]$$

where $\Pi$ is the orthogonal projection onto $X$ and $\{\alpha_k\}$ is a sequence of positive exogenous stepsizes.
The classical projection method for $\text{VI}(T, X)$, akin to the projected gradient method for convex optimization, is the projection algorithm:

$$x^{k+1} = \Pi[x^k - \alpha_k T(x^k)],$$

where $\Pi$ is the orthogonal projection onto $X$ and $\{\alpha_k\}$ is a sequence of positive exogenous stepsizes.

Convergence of the sequence $\{x^k\}$ to a solution of $\text{VI}(T, X)$ has been proved when $T$ is strongly monotone and Lipschitz continuous, and the stepsizes satisfy $\alpha_k \in (0, 2\sigma/L^2)$, $\inf_k \alpha_k > 0$, where $\sigma > 0$ is the modulus of strong monotonicity and $L$ is the Lipschitz constant.
The strong monotonicity assumption is quite demanding, and the projection method may diverge when $T$ is merely monotone. For this case, Korpelevich proposed in 1976 the extragradient algorithm:

$$z^k = \Pi [x^k - \alpha_k T(x^k)]$$

$$x^{k+1} = \Pi [x^k - \alpha_k T(z^k)],$$

in which an additional auxiliary projection step is introduced. The method converges to a solution when $T$ is just monotone and the stepsizes satisfy $\alpha_k \in (\nu, 1/L)$ for some $\nu > 0$. 
The strong monotonicity assumption is quite demanding, and the projection method may diverge when $T$ is merely monotone. For this case, Korpelevich proposed in 1976 the extragradient algorithm:

$$z^k = \Pi[x^k - \alpha_k T(x^k)]$$

$$x^{k+1} = \Pi[x^k - \alpha_k T(z^k)],$$

in which an additional auxiliary projection step is introduced. The method converges to a solution when $T$ is just monotone and the stepsizes satisfy $\alpha_k \in (\nu, 1/L)$ for some $\nu > 0$.

For the cases in which $T$ is not Lipschitz continuous, or the Lipschitz constant $L$ is either unknown or too large, variants of the extragradient methods were introduced in which the stepsizes are determined through an Armijo-like line-search.
In 1987 Khobotov proposed a line-search requiring an orthogonal projection onto $X$ in each step of the inner search. In 1997 Iusem and Svaiter introduced a line-search without projections, keeping just the two orthogonal projections per outer iteration of the extragradient method.
In 1987 Khobotov proposed a line-search requiring an orthogonal projection onto $X$ in each step of the inner search. In 1997 Iusem and Svaiter introduced a line-search without projections, keeping just the two orthogonal projections per outer iteration of the extragradient method.

Observe that both the projection method and the extra-gradient method are explicit, i.e., the formula for obtaining $x^{k+1}$ is an explicit one, up to the computation of the orthogonal projection $\Pi$. 
In 1987 Khobotov proposed a line-search requiring an orthogonal projection onto $X$ in each step of the inner search. In 1997 Iusem and Svaiter introduced a line-search without projections, keeping just the two orthogonal projections per outer iteration of the extragradient method.

Observe that both the projection method and the extra-gradient method are explicit, i.e., the formula for obtaining $x^{k+1}$ is an explicit one, up to the computation of the orthogonal projection $\Pi$.

The method presented in this talk is an extension to SVI of Korpelevich’s extragradient method.
SA methods for SVI I

The first SA method for SVI extends the projection method to SVI and was proposed by Jiang and Xu and in 2008. It is given by:

$$x^{k+1} = \Pi[x^k - \alpha_k F(\xi^k, x^k)]$$

where $\Pi$ is the orthogonal projection onto $X$, $\{\xi^k\}$ is a sample of $\xi$ and $\alpha_k$ is the stepsize.
The first SA method for SVI extends the projection method to SVI and was proposed by Jiang and Xu and in 2008. It is given by:

\[ x^{k+1} = \Pi[x^k - \alpha_k F(\xi^k, x^k)] , \]

where \( \Pi \) is the orthogonal projection onto \( X \), \( \{\xi^k\} \) is a sample of \( \xi \) and \( \alpha_k \) is the stepsize.

The sequence \( \{x^k\} \) converges almost surely to a solution of SVI(\( T, X \)) when \( T \) is strictly monotone and Lipschitz continuous and the step-sizes are small, i.e.

\[ \sum_k \alpha_k = \infty, \quad \sum_k \alpha_k^2 < \infty. \]
In the stochastic setting, the following assumption on the sampling must also be imposed:

there exists $\sigma > 0$ such that, for all $k \in \mathbb{N}$,

$$
\mathbb{E} \left[ F(\xi^k, x^k) | \mathcal{F}_k \right] = T(x^k),
$$

$$
\mathbb{E} \left[ \| F(\xi^k, x^k) - T(x^k) \|^2 | \mathcal{F}_k \right] \leq \sigma^2,
$$

where $\mathcal{F}_k = \sigma(x^0, \xi^0, \ldots, \xi^{k-1})$ denotes the natural filtration associated to the sample $\{\xi_i\}$ and the initial iterate $x^0$. 
In the stochastic setting, the following assumption on the sampling must also be imposed:
there exists $\sigma > 0$ such that, for all $k \in \mathbb{N}$,

$$\mathbb{E} \left[ F(\xi^k, x^k) | \mathcal{F}_k \right] = T(x^k),$$

$$\mathbb{E} \left[ \| F(\xi^k, x^k) - T(x^k) \|^2 | \mathcal{F}_k \right] \leq \sigma^2,$$

where $\mathcal{F}_k = \sigma(x^0, \xi^0, \ldots, \xi^{k-1})$ denotes the natural filtration associated to the sample $\{\xi_i\}$ and the initial iterate $x^0$.

They mean that the stochastic error is unbiased, and that its variance is uniform over the generated sequence.
Wang and Bertsekas in 2015 considered the case in which $X = \bigcap_{i=1}^{m} X_i$, and improved upon the method of Jiang and Xu by using projections onto a certain $X_i$ as an approximation of the projection onto $X$. The index $\omega_k$ of the set be used at iteration $k$ is chosen randomly. The algorithm is as follows:
Wang and Bertsekas in 2015 considered the case in which $X = \cap_{i=1}^{m} X_i$, and improved upon the method of Jiang and Xu by using projections onto a certain $X_i$ as an approximation of the projection onto $X$. The index $\omega_k$ of the set be used at iteration $k$ is chosen randomly. The algorithm is as follows:

$$y^k = x^k - \alpha_k F(\xi^k, x^k)$$

$$x^{k+1} = y^k - \beta_k (y^k - \Pi_{\omega_k}(y^k)).$$
Wang and Bertsekas in 2015 considered the case in which $X = \cap_{i=1}^{m} X_i$, and improved upon the method of Jiang and Xu by using projections onto a certain $X_i$ as an approximation of the projection onto $X$. The index $\omega_k$ of the set be used at iteration $k$ is chosen randomly. The algorithm is as follows:

$$y^k = x^k - \alpha_k F(\xi^k, x^k)$$

$$x^{k+1} = y^k - \beta_k (y^k - \Pi_{\omega_k}(y^k)).$$

Note that the second step is just a relaxed projection onto $X_{\omega_k}$. We emphasize that both the random map $F$ and the control sequence $\{\omega_k\}$ are sampled.
Wang and Bertsekas in 2015 considered the case in which $X = \bigcap_{i=1}^{m} X_i$, and improved upon the method of Jiang and Xu by using projections onto a certain $X_i$ as an approximation of the projection onto $X$. The index $\omega_k$ of the set be used at iteration $k$ is chosen randomly. The algorithm is as follows:

$$y^k = x^k - \alpha_k F(\xi^k, x^k)$$

$$x^{k+1} = y^k - \beta_k (y^k - \Pi_{\omega_k}(y^k)).$$

Note that the second step is just a relaxed projection onto $X_{\omega_k}$. We emphasize that both the random map $F$ and the control sequence $\{\omega_k\}$ are sampled.

The operator is assumed to be strongly monotone and Lipschitz-continuous and the sampling must satisfy the same assumptions as in the method of Jiang and Xu.
Statement of the method

Now we consider an extension to SVI of Korpelevich’s extragradient algorithm, which converges under just pseudomonotonicity of $T$. The novelty is that, instead of taking at iteration $k$ a sample of $F(\xi, x^k)$, we take $N_k$ samples and use the average of the values of $F$ at the sampled points.
Statement of the method

- Now we consider an extension to SVI of Korpelevich’s extragradient algorithm, which converges under just pseudomonotonicity of $T$. The novelty is that, instead of taking at iteration $k$ a sample of $F(\xi, x^k)$, we take $N_k$ samples and use the average of the values of $F$ at the sampled points.

- Our extragradient method for SVI is as follows:
  
  
  \[
  \begin{align*}
  z^k &= \Pi[x^k - \frac{\alpha_k}{N_k} \sum_{j=1}^{N_k} F(\xi^k_j, x^k)] \\
  x^{k+1} &= \Pi[x^k - \frac{\alpha_k}{N_k} \sum_{j=1}^{N_k} F(\eta^k_j, z^k)]
  \end{align*}
  \]

  where $\Pi$ is the orthogonal projection onto $X$, $\{N_k\} \subset \mathbb{N}$ is a nondecreasing sequence, called the sample rate sequence, $\alpha_k$ is the stepsize, and $\{\xi^k_j, \eta^k_j : k \in \mathbb{N}, j = 1, \ldots, N_k\}$ are independent identically distributed samples of $\xi$. 

A. Iusem, A. Jofré, R. Imbuzeiro, P. Thompson
Extragradient Method with Variance Reduction for Stochastic Variational Inequalities
Define the natural residual function of the problem $r_\alpha(x)$ as

$$r_\alpha(x) := \|x - \Pi(x - \alpha T(x))\|.$$ 

The set of zeroes of $r_\alpha$ coincides with $X^*$. 
Define the natural residual function of the problem $r_\alpha(x)$ as

$$r_\alpha(x) := \|x - \Pi(x - \alpha T(x))\|.$$ 

The set of zeroes of $r_\alpha$ coincides with $X^*$. 

Given $\epsilon > 0$, we prove that there exists $K = K_\epsilon$ such that

$$\mathbb{E}[r_\alpha(x^K)^2] < \epsilon.$$ 

$\mathbb{E}[r_\alpha(x^K)^2]$ is our non-asymptotic convergence rate.
Define the natural residual function of the problem $r_\alpha(x)$ as
\[ r_\alpha(x) := \|x - \Pi(x - \alpha T(x))\|. \]
The set of zeroes of $r_\alpha$ coincides with $X^*$.

Given $\epsilon > 0$, we prove that there exists $K = K_\epsilon$ such that
\[ \mathbb{E}[r_\alpha(x^K)^2] < \epsilon. \]
$\mathbb{E}[r_\alpha(x^K)^2]$ is our non-asymptotic convergence rate.

If $\mathbb{E}[r_\alpha(x^K)^2] \leq Q/K$ for some constant $Q > 0$ then we have an $O(1/K)$ convergence rate. The oracle complexity will be defined as the total number of oracle calls needed for $\mathbb{E}[r_\alpha(x^K)^2] < \epsilon$ to hold, i.e., $\sum_{k=1}^{K} 2N_k$. 
i) We employ an iterative variance reduction procedure. This means that, instead of calling the oracle once per iteration (as in previous SA methods for SVI studied so far), our method calls the oracle $N_k$ times at iteration $k$ and uses the empirical average of the values of the random operator $F$ at the current iterates $x^k$ and $z^k$. 
i) We employ an iterative variance reduction procedure. This means that, instead of calling the oracle once per iteration (as in previous SA methods for SVI studied so far), our method calls the oracle $N_k$ times at iteration $k$ and uses the empirical average of the values of the random operator $F$ at the current iterates $x^k$ and $z^k$.

This variance reduction procedure is the mechanism that allows our extragradient method to converge in an unbounded setting with stepsizes bounded away from zero, and to achieve an accelerated rate in terms of the natural residual.
ii) Assuming just pseudo-monotonicity of the operator (relevant e.g. for stochastic fractional programming, stochastic optional pricing and stochastic economic equilibria), and using an extragradient scheme without regularization, we prove that, almost surely, the generated sequence is bounded, its distance to the solution set converges to zero and its natural residual value converges to zero almost surely and in $L^2$. 
ii) Assuming just pseudo-monotonicity of the operator (relevant e.g. for stochastic fractional programming, stochastic optional pricing and stochastic economic equilibria), and using an extragradient scheme without regularization, we prove that, almost surely, the generated sequence is bounded, its distance to the solution set converges to zero and its natural residual value converges to zero almost surely and in $L^2$.

iii) Also, for any $p \geq 4$, if the random operator has finite $p$-moment then the sequence is bounded in $L^p$, and we are able to provide explicit upper bounds in terms of the problem parameters. Previous work required a bounded monotone operator, specific forms of monotonicity or regularization procedures, which induce a suboptimal performance in terms of rate and complexity.
iv) Our work is the first SA method for SVI with stepsizes bounded away from zero. Such feature allows our method to achieve an accelerated convergence rate of $O(1/K)$ under plain pseudo-monotonicity, with no regularization requirements. In previous works, methods with vanishing (small) stepsizes were used, achieving a $O(1/K)$ rate only under more demanding monotonicity assumptions (strong monotonicity, or monotonicity and weak sharpness), and a rate of just $O(1/\sqrt{k})$ for bounded monotone operators.
iv) Our work is the first SA method for SVI with stepsizes bounded away from zero. Such feature allows our method to achieve an accelerated convergence rate of $O(1/K)$ under plain pseudo-monotonicity, with no regularization requirements. In previous works, methods with vanishing (small) stepsizes were used, achieving a $O(1/K)$ rate only under more demanding monotonicity assumptions (strong monotonicity, or monotonicity and weak sharpness), and a rate of just $O(1/\sqrt{k})$ for bounded monotone operators.

v) Our method preserves the optimal oracle complexity $O(\epsilon^{-2})$ up to a first order logarithmic term. By accelerating the rate, we reduce the computational complexity. We provide explicit upper bounds for the rate and complexity in terms of the problem parameters.
vi) The results in items (i)-(ii) are valid for an **unbounded feasible set.**
vi) The results in items (i)-(ii) are valid for an unbounded feasible set.

vii) Our sampling procedure for this method is robust in the sense of Nemirovsky.
vi) The results in items (i)-(ii) are valid for an unbounded feasible set.

vii) Our sampling procedure for this method is robust in the sense of Nemirovsky.

viii) A fundamental tool used in our convergence analysis is the Non-negative Almost Supermartingale Convergence Theorem of Robbins and Siegmund (1971), which can be seen as a stochastic version of Ermoliev’s Quasi-Fejér Convergence Theorem (1967).


