First-order Methods for Stochastic Variational Inequality Problems with Function Constraints

Informs Annual Meeting, 2024

Digvijay Boob, Qi Deng, Mohammad Khalafi

Southern Methodist Univesity

October 21, 2024

Outline

Function constrained VI problem

$$\text{Find } x^* \in \widetilde{X} : \quad \langle F(x^*), x - x^* \rangle \geq 0, \quad \forall x \in \widetilde{X}, \tag{1}$$

- $\widetilde{X} := X \cap \{x : g_j(x) \leq 0, j = 1, \ldots, m\}$
- X is convex compact set with easy projection operator
- F is monotone operator, i.e.,

$$\langle F(x_1) - F(x_2), x_1 - x_2, \rangle \geq 0, \quad \forall x_1, x_2 \in X$$

- $g_j, j = 1, \ldots, m$ are continuous convex functions
- The main challenge: feasible set $\widetilde{X}$ does not have an easy projection oracle (due to presence of function constraints)

Assumptions: Lipschitz Continuity and Smoothness

We consider composite problems satisfying for all $x_1, x_2 \in X$:

- $$\langle F(x_1) - F(x_2), x_1 - x_2, \rangle \leq L\|x_1 - x_2\| + H,$$
- $$g(x_1) - g(x_2) - \langle \nabla g(x_2), x_1 - x_2, \rangle \leq \frac{L_g}{2}\|x_1 - x_2\|^2 + H_g\|x_1 - x_2\|,$$
- $$|g(x_1) - g(x_2)| \leq M_g\|x_1 - x_2\|.$$
- Smooth problems when $H = H_g = 0$
- Nonsmooth problem when either H or H_g is nonzero.

Assumptions: Stochastic and Fully-Stochastic Settings

- Stochastic setting when operator F is stochastic
  $F(x) = \mathbb{E}[\mathfrak{F}(x,\xi)], \quad \|\mathfrak{F}(x,\xi) - F(x)\|^2 \le \sigma^2$
- Fully stochastic setting when both operator F and function constraints g are stochastic

$$\begin{aligned} \mathbb{E}[\mathfrak{g}(x,\xi)] &= g(x), \\ \mathbb{E}[\mathfrak{G}(x,\xi)] &= \nabla g(x), \\ \mathbb{E}[\|\mathfrak{G}(x,\xi) - \nabla g(x)\|^2] &\le \sigma_{\mathfrak{G}}^2, \\ \mathbb{E}[(\mathfrak{g}(x,\xi) - g(x))^2] &\le \sigma_{\mathfrak{g}}^2. \end{aligned}$$

Solution criterion

We say that $\widetilde{x}$ is an ϵ-solution of (1) if

$$\max_{x \in \widetilde{X}} \langle F(x), \widetilde{x} - x \rangle \leq \epsilon, \quad \|[g(\widetilde{x})]_+\| \leq \epsilon.$$

Similarly, for the stochastic case, $\widetilde{x}$ is an ϵ-solution of (1) if the above bounds hold under expectation, i.e.,

$$\mathbb{E}[\max_{x \in \widetilde{X}} \langle F(x), \widetilde{x} - x \rangle] \leq \epsilon, \quad \mathbb{E}[\|[g(\widetilde{x})]_+\|] \leq \epsilon.$$

Motivation

- Consider FCVI problem as a KKT system where $x \in X$ is the primal variable and $\lambda_j (\geq 0)$ denotes the dual multipliers associated with constraints.
- This systems is a VI for operator $\widetilde{F}(x, \lambda)$ where $\widetilde{F}$ is jointly monotone in (x, λ) over the feasible set $X \times \{\lambda \geq \mathbf{0}\}$.
- The current methods converge for Lipschitz continuous operators which can be violated even for smooth FCVIs since the dual set $\{\lambda \geq 0\}$ is unbounded.
- Second challenge is the gradient w.r.t. x is not computable since x^* is involved in the formulation.
- Adaptive Operator Extrapolation (`AdOpEx`) method that converges for smooth deterministic FCVIs.
- In the stochastic case, the noise in g is magnified by possibly unconstrained λ which can lead to noise level that is difficult to bound apriori leading to propose better methods.

Literature Review

- VI is a classical topic with various works showing algorithms that converge asymptotically.
- Nemirovski showed a Mirror Prox method that converges to ϵ-solution in $O(\frac{1}{\epsilon})$ iterations.
- This rate significantly improved over convergence rate of $O(\frac{1}{\epsilon^2})$ of standard projected gradient method.
- Malitsky proposed an adoptive and optimal method.
- Kotsalis et al. proposed a Operator Extrapolation method that maintains a single sequence and requires a single projection in every iteration.

Reflections on the State of the Art

- Without exception, all of these methods require projection onto the feasible set.
- If a function constraint $g(x) \leq 0$ is involved, projection may not be evaluated efficiently.
- If g is data driven function then it may not be possible to evaluate a projection.

Adaptive Operator Extrapolation (AdOpEx) method

Algorithm `AdOpEx` method

1: **Input:** $x^0 \in X, \lambda^0 = \mathbf{0}$.
2: Set $x^{-1} = x^0$.
3: **for** $t = 1, \ldots, T-1$ **do**
4: $s^t \leftarrow (1+\theta_t) g(x^t) - \theta_t g(x^{t-1})$
5: $\lambda^{t+1} \leftarrow \mathrm{argmin}_{\lambda \geq \mathbf{0}} \langle -s^t, \lambda \rangle + \frac{\tau_t}{2} \|\lambda - \lambda^t\|^2$
6: $u^t \leftarrow (1+\theta_t)[F(x^t) + \nabla g(x^t)\lambda^t] - \theta_t[F(x^{t-1}) + \nabla g(x^{t-1})\lambda^{t-1}]$
7: $x^{t+1} \leftarrow \mathrm{argmin}_{x \in X} \langle u^t, x \rangle + \frac{\eta_t}{2} \|x - x^t\|^2$
8: **end for**
9: **Output:** $\bar{x}^T := (\sum_{t=0}^{T-1} \gamma_t x^{t+1}) / (\sum_{t=0}^{T-1} \gamma_t)$

AdOpEx Method's Convergence

Theorem

Let $\gamma_t = \frac{\eta_0}{\eta_t}$ *for* $t \geq 0$, $\theta_t = \frac{\gamma_{t-1}}{\gamma_t}$, $\tau_t = \frac{M_g^2}{3L^2}\eta_t$ *and* $\eta_t = 6(L + L_g \max_{i\in[t]} \|\lambda^i\|)$. *Then, we have,*

$$\|\lambda^{t+1}\| \leq B := \frac{\sqrt{6}L}{M_g}\|x^0 - x^*\| + (\sqrt{2}+1)\|\lambda^*\|, \forall t \geq 0,$$

and obtain an $O(1/T)$ *convergence rate for* AdOpEx *method.*

Partial Operator Constraint Extrapolation Method

Algorithm Partial Operator Extrapolation Method for fully-stochastic FCVI

1: **Input:** $x^0 \in X, \lambda^0 = \mathbf{0}$.
2: Set $x^{-1} = x^0$.
3: **for** $t = 0, 1, 2, \dots, T-1$ **do**
4: $\quad \lambda^{t+1} \leftarrow \operatorname{argmin}_{\lambda \geq 0} -\langle \mathfrak{g}(x^t, \bar{\xi}^t), \lambda \rangle + \frac{\tau_t}{2} \|\lambda - \lambda^t\|^2$
5: $\quad x^{t+1} \leftarrow \operatorname{argmin}_{x \in X} \langle (1+\theta_t)\mathfrak{F}^t - \theta_t \mathfrak{F}^{t-1} + \mathfrak{G}(x^t, \xi^t)\lambda^{t+1}, x \rangle + \frac{\eta_t}{2} \|x - x^t\|^2$
6: **end for**
7: **Output:** $\bar{x}_T := \sum_{t=0}^{T-1} x^{t+1} / T$

OpConEx Method's Convergence

Theorem

Suppose Algorithm 2 generates $\{x^{t+1}, \lambda^{t+1}, v^{t+1}\}$ *by setting* $\theta_t = 1, \eta_t = \eta$ *and* $\tau_t = \tau$ *such that*

$$\eta = \frac{\sqrt{2T}[B(3M_g+4\|\sigma_{\mathfrak{G}}\|(1+\sigma_{\mathfrak{g}}))+H+\sigma]}{D_X} + \frac{25L}{3},$$

$$\tau = \frac{2\sqrt{2T}(2M_g+5\|\sigma_{\mathfrak{G}}\|+\sigma_{\mathfrak{g}})D_X}{B}, \quad \text{for } B \geq 1,$$

OpConEx Method's Convergence

Theorem

Then we have

$$\mathbb{E}[\|\lambda^* - \lambda^t\|^2] \le 2R_f e,$$
$$R_f = \tfrac{\eta}{\tau}\|x^* - x^0\|^2 + 3\|\lambda^*\|^2 + 2\sigma_{\mathfrak{g}}^2 + 9(H^2 + 2\sigma^2).$$

and we obtain expected convergence rate of $O\big(\frac{LD_x^2}{T} + \frac{M_g D_X}{\sqrt{T}}(B + \frac{(\|\sigma_{\mathfrak{G}}\|+\sigma_{\mathfrak{g}})(\|\lambda^*\|+1)^2}{B}) + \frac{(H+\sigma)D_X}{\sqrt{T}} + \frac{BD_X\|\sigma_{\mathfrak{G}}\|\sigma_{\mathfrak{g}}}{\sqrt{T}}\big)$ *where* $(L\backslash H)$ *is the (smoothness\nonsmoothness) constant of* F*. Also* $\sigma, \sigma_{\mathfrak{g}}$*, and* $\|\sigma_{\mathfrak{G}}\|$ *are the variance of approximation errors in* F, g*, and* ∇g *respectively.*

- Observation: OpConEx is slow $(O(1/\sqrt{T}))$ in terms of M_g and $\|\sigma_{\mathfrak{G}}\|\sigma_{\mathfrak{g}}$.

Operator Extrapolation Method

Algorithm Operator Constraint Extrapolation (`OpConEx`) method

1: **Input:** $x^0 \in X$, $\lambda^0 = \mathbf{0}$, $\mathfrak{g}(x^0, \bar{\xi}^0)$ and $\mathfrak{F}^0$.

2: Set $x^{-1} = x^0$, $\mathfrak{F}^{-1} = \mathfrak{F}^0$ and $\ell_{\mathfrak{g}}^0(x^0) = \ell_{\mathfrak{g}}^0(x^{-1}) = \mathfrak{g}(x^0, \bar{\xi}^0)$.

3: **for** $t = 0, 1, 2, \ldots, T-1$ **do**

4: $\quad \mathfrak{s}^t \leftarrow (1+\theta_t)\ell_{\mathfrak{g}}^t(x^t) - \theta_t \ell_{\mathfrak{g}}^t(x^{t-1})$

5: $\quad \lambda^{t+1} \leftarrow \operatorname{argmin}_{\lambda \geq \mathbf{0}} \langle -\mathfrak{s}^t, \lambda \rangle + \frac{\tau_t}{2}\|\lambda - \lambda^t\|^2$

6: $\quad x^{t+1} \leftarrow \operatorname{argmin}_{x \in X} \langle (1+\theta_t)\mathfrak{F}^t - \theta_t \mathfrak{F}^{t-1} + \sum_{i=1}^m \lambda_i^{t+1} \mathfrak{G}_i(x^t, \xi^t), x \rangle + \frac{\eta_t}{2}\|x - x^t\|^2$

7: **end for**

8: **Output:** $\bar{x}^T := (\sum_{t=0}^{T-1} \gamma_t x^{t+1}) / (\sum_{t=0}^{T-1} \gamma_t)$

OpConEx Method's Convergence

Theorem

Let $B \geq 1$ *and* $\sigma_{X,g} := \sqrt{\sigma_{\mathfrak{g}}^2 + D_X^2 \|\sigma_{\mathfrak{G}}\|^2}$. *Suppose we set*

$$\gamma_t = \theta_t = 1, \qquad \eta_t = L_g B + \eta, \qquad \tau_t = \tau,$$

where $\eta = 8L + \frac{8M_g B}{D_X} + \frac{2(H + H_g B + \sqrt{2}\sigma + 4B\|\sigma_{\mathfrak{G}}\|)}{D_X}\sqrt{T}$ *and* $\tau = \frac{9D_X}{B}\max\{M_g, \|\sigma_{\mathfrak{G}}\|\} + \frac{8\sigma_{X,g}}{B}\sqrt{T}$. *Then we improve the dependence on* M_g *from* $O(M_g/\sqrt{T})$ *to* $O(M_g/T)$ *and the dependence on* $\sigma_{\mathfrak{g}}$ *from* $O(\sigma_{\mathfrak{g}}^2/\sqrt{T})$ *to* $O(\sigma_{\mathfrak{g}}/\sqrt{T})$.

Saddle Point Problem with Coupled Constraint

consider a general class of saddle point problems as follows

$$\begin{aligned} \min_{u \in U} \max_{v \in V} \; & f(u, v) \\ \text{s.t. } & g(u, v) \leq \mathbf{0}, \end{aligned} \tag{2}$$

where $f(\cdot, v)$ is a convex function of u for all $v \in V$, $f(u, \cdot)$ is a concave function of v for all $u \in U$, $U \subset \mathbb{R}^{n_u}$ and $V \subset \mathbb{R}^{n_v}$ are feasible set constraints, and $g(u, v)$ is a coupling constraint that is jointly convex in u and v.

Notation: The usual "Lagrangian" $f(u, v) + \lambda g(u, v)$ is neither convex nor concave in v.

Saddle Point Problem with Coupled Constraint

Define $\bar{w} \in W := U \times V$ as an ϵ-approximate saddle-point of (2) if

$$\max_{w \in \widetilde{W}} G(\widehat{w}; w) \leq \epsilon, \qquad g(\widehat{w}) \leq \epsilon.$$

where $\widetilde{W} := W \cap \{(u, v) : g(u, v) \leq \mathbf{0}\}$.

Reformulation of Saddle Point problem to FCVI

Let the set constraint $X = W$ and the variable $x = w$. We define $F(x) = F(w) := \begin{bmatrix} \nabla_u f(u,v) \\ -\nabla_v f(u,v) \end{bmatrix}$ as the operator throughout this section. It is clear that since f is a convex-concave function, the resulting operator $F(w)$ is monotone. As per this reformulation, convex coupled constraint $\{g(u,v) \le \mathbf{0}\}$ in (2) is equivalent to $g(x) \le \mathbf{0}$ in (1). Then, $\widetilde{W}$ corresponds to $\widetilde{X}$.

Conclusion

Table: Convergence rate of the proposed methods for solving different FCVIs

Algorithm	Deterministic smooth	Nonsmooth	Stochastic	Fully stochastic
AdOpEx	$\mathcal{O}(\frac{1}{T})$	–	–	–
OpConEx	–	$\mathcal{O}(\frac{M_g}{\sqrt{T}})$	$\mathcal{O}(\frac{M_g+\sigma}{\sqrt{T}})$	$\mathcal{O}(\frac{M_g+\sigma}{\sqrt{T}} + \frac{\sigma_g^2}{\sqrt{T}})$
OpConEx	–	$\mathcal{O}(\frac{M_g}{T} + \frac{H_g}{\sqrt{T}})$	$\mathcal{O}(\frac{M_g}{T} + \frac{H_g+\sigma}{\sqrt{T}})$	$\mathcal{O}(\frac{M_g}{T} + \frac{H_g+\sigma+\sigma_g}{\sqrt{T}})$

M_g: Lipschitz constant of g; H_g: Lipschitz constant of only the nonsmooth component of g, σ: Standard deviation of stochastic oracle for F, σ_g: standard deviation of stochastic oracles associated with g and ∇g.

- **Thanks!**
- **Question?**