On the Convergence of Single-Call Stochastic Extra-Gradient Methods

Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, Panayotis Mertikopoulos

NeurIPS, December 2019

Outline:

Variational Inequality

Introduction: Variational Inequalities in Machine Learning

- Generative adversarial network (GAN)

$$\min_{\theta}\max_{\phi}\mathbb{E}_{x\sim p_{\mathrm{data}}}[\log(D_\phi(x))]+\mathbb{E}_{z\sim p_{\mathcal{Z}}}[\log(1-D_\phi(G_\theta(z)))].$$

  More min-max (saddle point) problems: distributionally robust learning, primal-dual formulation in optimization, . . .

- Seach of equilibrium: games, multi-agent reinforcement learning, . . .

Definition

Stampacchia variational inequality

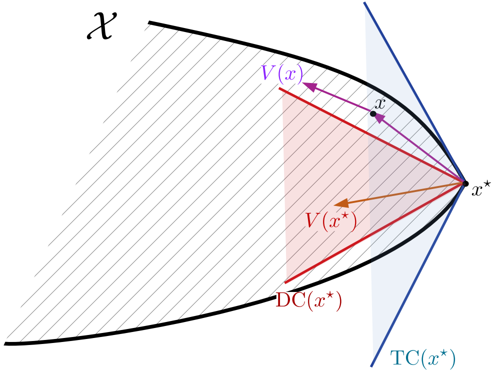
$$\text{Find } x^\star \in \mathcal{X} \text{ such that } \langle V(x^\star), x - x^\star \rangle \geq 0 \text{ for all } x \in \mathcal{X}. \tag{SVI}$$

Minty variational inequality

$$\text{Find } x^\star \in \mathcal{X} \text{ such that } \langle V(x), x - x^\star \rangle \geq 0 \text{ for all } x \in \mathcal{X}. \tag{MVI}$$

With closed convex set $\mathcal{X} \subseteq \mathbb{R}^d$ and vector field $V : \mathbb{R}^d \to \mathbb{R}^d$.

Illustration



SVI: $V(x^\star)$ belongs to the dual cone $\mathrm{DC}(x^\star)$ of $\mathcal{X}$ at $x^\star$ [local]

MVI: $V(x)$ forms an acute angle with tangent vector $x - x^\star \in \mathrm{TC}(x^\star)$ [global]

Example: Function Minimization

$$\min_{x} f(x)$$

$$\text{subject to } x \in \mathcal{X}$$

$f : \mathcal{X} \to \mathbb{R}$ differentiable function to minimize.

Let $V = \nabla f$.

(SVI) $\quad \forall x \in \mathcal{X}, \langle \nabla f(x^\star), x - x^\star \rangle \geq 0 \quad$ [first-order optimality]

(MVI) $\quad \forall x \in \mathcal{X}, \langle \nabla f(x), x - x^\star \rangle \geq 0 \quad$ [$x^\star$ is a minimizer of f]

If f is convex, (SVI) and (MVI) are equivalent.

Example: Saddle point Problem

Find $x^\star = (\theta^\star, \phi^\star)$ such that

$$\mathcal{L}(\theta^\star, \phi) \leq \mathcal{L}(\theta^\star, \phi^\star) \leq \mathcal{L}(\theta, \phi^\star) \quad \text{for all } \theta \in \Theta \text{ and all } \phi \in \Phi.$$

$\mathcal{X} \equiv \Theta \times \Phi$ and $\mathcal{L} : \mathcal{X} \to \mathbb{R}$ differentiable function.

Let $V = (\nabla_\theta \mathcal{L}, -\nabla_\phi \mathcal{L})$.

(SVI) $\quad \forall (\theta, \phi) \in \mathcal{X}, \langle \nabla_\theta \mathcal{L}(x^\star), \theta - \theta^\star \rangle - \langle \nabla_\phi \mathcal{L}(x^\star), \phi - \phi^\star \rangle \geq 0$ [stationary]

(MVI) $\quad \forall (\theta, \phi) \in \mathcal{X}, \langle \nabla_\theta \mathcal{L}(x), \theta - \theta^\star \rangle - \langle \nabla_\phi \mathcal{L}(x), \phi - \phi^\star \rangle \geq 0$ [saddle point]

If $\mathcal{L}$ is convex-concave, (SVI) and (MVI) are equivalent.

Monoticity

The solutions of (SVI) and (MVI) coincide when V is continuous and **monotone**, i.e.,

$$\langle V(x') - V(x), x' - x\rangle \geq 0 \quad \text{for all } x, x' \in \mathbb{R}^d.$$

In the above two examples, this corresponds to either f being convex or $\mathcal{L}$ being convex-concave.

The operator analogue of strong convexity is **strong monoticity**

$$\langle V(x') - V(x), x' - x\rangle \geq \alpha \|x' - x\|^2 \quad \text{for some } \alpha > 0 \text{ and all } x, x' \in \mathbb{R}^d.$$

Extra-Gradient

From Forward-backward to Extra-Gradient

Forward-backward

$$X_{t+1} = \Pi_{\mathcal{X}}(X_t - \gamma_t V(X_t)) \tag{FB}$$

Extra-Gradient [Korpelevich 1976]

$$\begin{aligned} X_{t+\frac{1}{2}} &= \Pi_{\mathcal{X}}(X_t - \gamma_t V(X_t)) \\ X_{t+1} &= \Pi_{\mathcal{X}}(X_t - \gamma_t V(X_{t+\frac{1}{2}})) \end{aligned} \tag{EG}$$

The Extra-Gradient method anticipates the landscape of V by taking an extrapolation step to reach the leading state $X_{t+\frac{1}{2}}$.

From Forward-backward to Extra-Gradient

Forward-backward does not converge in bilinear games, while Extra-Gradient does.

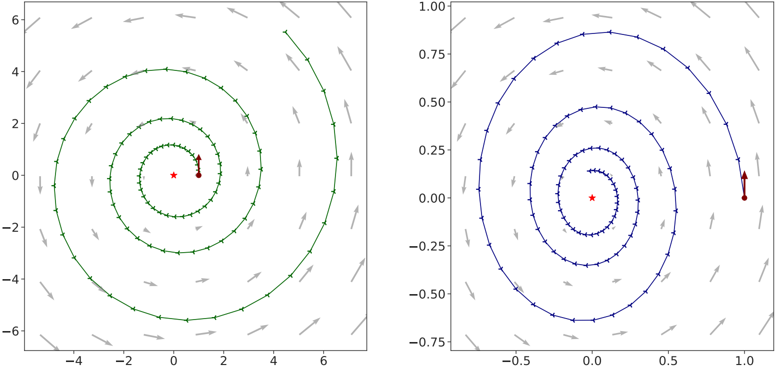
$$\min_{\theta \in \mathbb{R}} \max_{\phi \in \mathbb{R}} \theta\phi$$

Left:

Forward-backward

Right:

Extra-Gradient



Stochastic Oracle

If a stochastic oracle is involved:

$$X_{t+\frac{1}{2}} = \Pi_{\mathcal{X}}(X_t - \gamma_t \hat{V}_t)$$
$$X_{t+1} = \Pi_{\mathcal{X}}(X_t - \gamma_t \hat{V}_{t+\frac{1}{2}})$$

With $\hat{V}_t = V(X_t) + Z_t$ satisfying (and same for $\hat{V}_{t+\frac{1}{2}}$)

a) Zero-mean: $\mathbb{E}[Z_t \mid \mathcal{F}_t] = 0.$

b) Bounded variance: $\mathbb{E}[\|Z_t\|^2 \mid \mathcal{F}_t] \le \sigma^2.$

$(\mathcal{F}_t)_{t\in\mathbb{N}/2}$ is the natural filtration associated to the stochastic process $(X_t)_{t\in\mathbb{N}/2}$.

Convergence Metrics

- Ergodic convergence: restricted error function

$$\mathrm{Err}_R(\hat{x}) = \max_{x \in \mathcal{X}_R} \langle V(x), \hat{x} - x \rangle,$$

where $\mathcal{X}_R \equiv \mathcal{X} \cap \mathbb{B}_R(0) = \{x \in \mathcal{X} : \|x\| \leq R\}$.

- Last iterate convergence: squared distance $\mathrm{dist}(\hat{x}, \mathcal{X}^\star)^2$.

Lemma [Nesterov 2007]

Assume V is monotone. If $x^\star$ is a solution of (SVI), we have $\mathrm{Err}_R(x^\star) = 0$ for all sufficiently large R. Conversely, if $\mathrm{Err}_R(\hat{x}) = 0$ for large enough $R > 0$ and some $\hat{x} \in \mathcal{X}_R$, then $\hat{x}$ is a solution of (SVI).

Literature Review

We further suppose that V is β-Lipschitz.

	Convergence type	Hypothesis
Korpelevich 1976	Last iterate asymptotic	Pseudo monotone
Tseng 1995	Last iterate geometric	Monotone + error bound (e.g., strongly monotone, affine)
Nemirovski 2004	Ergodic in $\mathcal{O}(1/t)$	Monotone
Juditsky et al. 2011	Ergodic in $\mathcal{O}(1/\sqrt{t})$	Stochastic monotone

In Deep Learning

Extra-Gradient (EG) needs two oracle calls per iteration, while gradient computations can be very costly for deep models:

And if we drop one oracle call per iteration?

Single-call Extra-Gradient

Algorithms

1. **Past Extra-Gradient** [Popov 1980]

$$\begin{aligned} X_{t+\frac{1}{2}} &= \Pi_{\mathcal{X}}(X_t - \gamma_t \hat{V}_{t-\frac{1}{2}}) \\ X_{t+1} &= \Pi_{\mathcal{X}}(X_t - \gamma_t \hat{V}_{t+\frac{1}{2}}) \end{aligned} \tag{PEG}$$

2. **Reflected Gradient** [Malitsky 2015]

$$\begin{aligned} X_{t+\frac{1}{2}} &= X_t - (X_{t-1} - X_t) \\ X_{t+1} &= \Pi_{\mathcal{X}}(X_t - \gamma_t \hat{V}_{t+\frac{1}{2}}) \end{aligned} \tag{RG}$$

3. **Optimistic Gradient** [Daskalakis et al. 2018]

$$\begin{aligned} X_{t+\frac{1}{2}} &= \Pi_{\mathcal{X}}(X_t - \gamma_t \hat{V}_{t-\frac{1}{2}}) \\ X_{t+1} &= X_{t+\frac{1}{2}} + \gamma_t \hat{V}_{t-\frac{1}{2}} - \gamma_t \hat{V}_{t+\frac{1}{2}} \end{aligned} \tag{OG}$$

A First Result

$$X_{t+\frac{1}{2}} = \Pi_{\mathcal{X}}(X_t - \gamma_t \hat{V}_t)$$
$$X_{t+1} = \Pi_{\mathcal{X}}(X_t - \gamma_t \hat{V}_{t+\frac{1}{2}})$$

Proxy

- PEG: [Step 1] $\hat{V}_t \leftarrow \hat{V}_{t-\frac{1}{2}}$
- RG: [Step 1] $\hat{V}_t \leftarrow (X_{t-1} - X_t)/\gamma_t$; no projection
- OG: [Step 1] $\hat{V}_t \leftarrow \hat{V}_{t-\frac{1}{2}}$

  [Step 2] $X_t \leftarrow X_{t+\frac{1}{2}} + \gamma_t \hat{V}_{t-\frac{1}{2}}$; no projection

Proposition

Suppose that the Single-call Extra-Gradient (1-EG) methods presented above share the same initialization, $X_0 = X_1 \in \mathcal{X}$, $\hat{V}_{1/2} = 0$ and a same constant step-size $(\gamma_t)_{t\in\mathbb{N}} \equiv \gamma$. If $\mathcal{X} = \mathbb{R}^d$, the generated iterates X_t coincide for all $t \geq 1$.

Global Convergence Rate

- Always with Lipschitz continuity.
- Stochastic strongly monotone: step size in $\mathcal{O}(1/t)$.
- New results!

	Monotone		Strongly Monotone	
	Ergodic	Last Iterate	Ergodic	Last Iterate
Deterministic	$\boxed{1/t}$	Unknown	$1/t$	$e^{-\rho t}$
Stochastic	$1/\sqrt{t}$	Unknown	$\boxed{1/t}$	$\boxed{1/t}$

Proof Ingredients

Descent Lemma [Deterministic + Monotone]

There exists $(\mu_t)_{t\in\mathbb{N}} \in \mathbb{R}_+^{\mathbb{N}}$ such that for all $p \in \mathcal{X}$,

$$\|X_{t+1} - p\|^2 + \mu_{t+1} \le \|X_t - p\|^2 - 2\gamma\langle V(X_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - p\rangle + \mu_t.$$

Descent Lemma [Stochastic + Strongly Monotone]

Let $x^\star$ be the unique solution of (SVI). There exists $(\mu_t)_{t\in\mathbb{N}} \in \mathbb{R}_+^{\mathbb{N}}, M \in \mathbb{R}_+$ such that

$$\mathbb{E}[\|X_{t+1} - x^\star\|^2] + \mu_{t+1} \le (1 - \alpha\gamma_t)(\mathbb{E}[\|X_t - x^\star\|^2] + \mu_t) + M\gamma_t^2\sigma^2.$$

Regular Solution

Definition [Regular Solution]

We say that $x^\star$ is a **regular solution** of (SVI) if V is C^1-smooth in a neighborhood of $x^\star$ and the Jacobian $\mathrm{Jac}_V(x^\star)$ is positive-definite along rays emanating from $x^\star$, i.e.,

$$z^\top \mathrm{Jac}_V(x^\star) z \equiv \sum_{i,j=1}^{d} z_i \frac{\partial V_i}{\partial x_j}(x^\star) z_j > 0 \quad \text{for all } z \in \mathbb{R}^d \setminus \{0\} \text{ that are tangent to } \mathcal{X} \text{ at } x^\star.$$

- To be compared with
  - positive definiteness of the Hessian along qualified constraints in minimization;
  - differential equilibrium in games.
- Localization of strong monoticity.

Local Convergence

Theorem [Local convergence for stochastic non-monotone operators]

Let $x^\star$ be a regular solution of (SVI) and fix a tolerance level $\delta > 0$. Suppose (PEG) is run with step-sizes of the form $\gamma_t = \gamma/(t+b)$ for large enough γ and b. Then:

a. There are neighborhoods U and U_1 of $x^\star$ in $\mathcal{X}$ such that, if $X_{1/2} \in U, X_1 \in U_1$, the event

$$E_\infty = \{X_{t+\frac{1}{2}} \in U \text{ for all } t = 1, 2, \dots\}$$

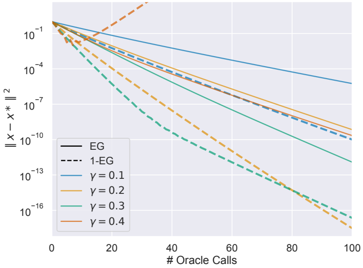
occurs with probability at least $1 - \delta$.

b. Conditioning on the above, we have:

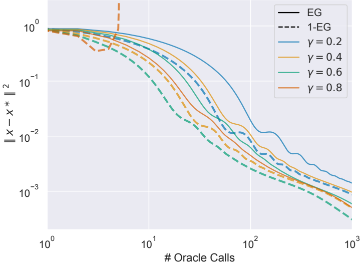
$$\mathbb{E}[\|X_t - x^\star\|^2 \mid E_\infty] = \mathcal{O}\left(\frac{1}{t}\right).$$

Experiments

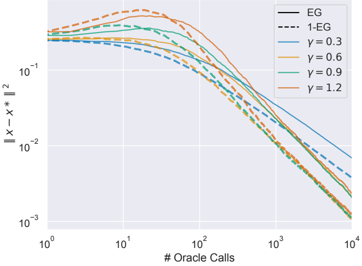
$$\mathcal{L}(\theta,\phi) = 2\epsilon_1\theta^\top A_1\theta + \epsilon_2\left(\theta^\top A_2\theta\right)^2 - 2\epsilon_1\phi^\top B_1\phi - \epsilon_2\left(\phi^\top B_2\phi\right)^2 + 4\theta^\top C\phi$$



(a) Strongly monotone
$(\epsilon_1 = 1,\ \epsilon_2 = 0)$
Deterministic
Last iterate



(b) Monotone $(\epsilon_1 = 0,\ \epsilon_2 = 1)$
Deterministic
Ergodic



(c) Non monotone $(\epsilon_1 = 1,\ \epsilon_2 = -1)$
Stochastic $Z_t \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2 = .01)$
Last iterate $(b = 15)$

Conclusion and Perspectives

Conclusion

- Single-call rates $\sim$ Two-call rates.
- Localization of stochastic guarantee.
- Last iterate convergence: a first step to the non-monotone world.
- Some research directions: Bregman, universal, ...

Bibliography

Daskalakis, Constantinos et al. (2018). "Training GANs with optimism". In: *ICLR '18: Proceedings of the 2018 International Conference on Learning Representations*.

Juditsky, Anatoli, Arkadi Semen Nemirovski, and Claire Tauvel (2011). "Solving variational inequalities with stochastic mirror-prox algorithm". In: *Stochastic Systems* 1.1, pp. 17–58.

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

Malitsky, Yura (2015). "Projected reflected gradient methods for monotone variational inequalities". In: *SIAM Journal on Optimization* 25.1, pp. 502–520.

Nemirovski, Arkadi Semen (2004). "Prox-method with rate of convergence $O(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems". In: *SIAM Journal on Optimization* 15.1, pp. 229–251.

Nesterov, Yurii (2007). "Dual extrapolation and its applications to solving variational inequalities and related problems". In: *Mathematical Programming* 109.2, pp. 319–344.

Popov, Leonid Denisovich (1980). "A modification of the Arrow–Hurwicz method for search of saddle points". In: *Mathematical Notes of the Academy of Sciences of the USSR* 28.5, pp. 845–848.

Tseng, Paul (June 1995). "On linear convergence of iterative methods for the variational inequality problem". In: *Journal of Computational and Applied Mathematics* 60.1-2, pp. 237–252.