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

Yu-Guan Hsieh[1,2] Franck Iutzeler[1,2] Jérôme Malick[1,2,4] Panayotis Mertikopoulos[1,3,4,5]

[1]Univ. Grenoble Alpes [2]LJK [3]LIG [4]CNRS [5]Inria

Beyond Minimization

- Generative adversarial network (GAN)

$$\min_{\theta}\max_{\phi} \mathbb{E}_{x\sim p_{\mathcal{D}}}[f(D_\phi(x))] + \mathbb{E}_{z\sim p_Z}[g(D_\phi(G_\theta(z)))].$$

More min-max: distributionally robust, primal-dual, ...

- Seach of equilibrium: games, multi-agent RL, ...

Variational Inequalities

Definition and Setup

Closed convex set $\mathcal{X} \subseteq \mathbb{R}^d$; Vector field $V : \mathbb{R}^d \to \mathbb{R}^d$

Stampacchia variational inequality

Find $x^\star \in \mathcal{X}$ s.t. $\forall x \in \mathcal{X}$, $\langle V(x^\star), x - x^\star\rangle \geq 0$. (VI)

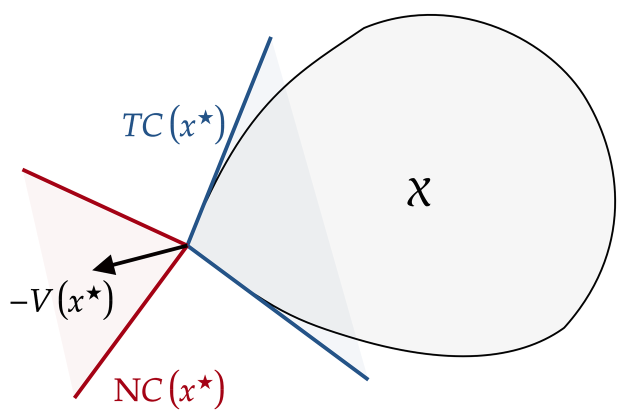
Monoticity

$$\forall x, x' \in \mathbb{R}^d,\ \langle V(x') - V(x), x' - x\rangle \geq \alpha\|x' - x\|^2$$

constant $\alpha \geq 0$; strongly monotone: $\alpha > 0$.

Assumptions:

- Lipschitz continuous V.
- Noisy unbiased oracle $\hat{V}$.
- Finite-variance noise.



Example of Saddle Point Problem

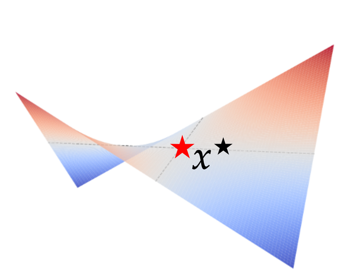
Find $x^\star = (\theta^\star, \phi^\star)$ such that

$$\forall\theta \in \Theta,\ \forall\phi \in \Phi,\ \mathcal{L}(\theta^\star, \phi) \leq \mathcal{L}(\theta^\star, \phi^\star) \leq \mathcal{L}(\theta, \phi^\star).$$

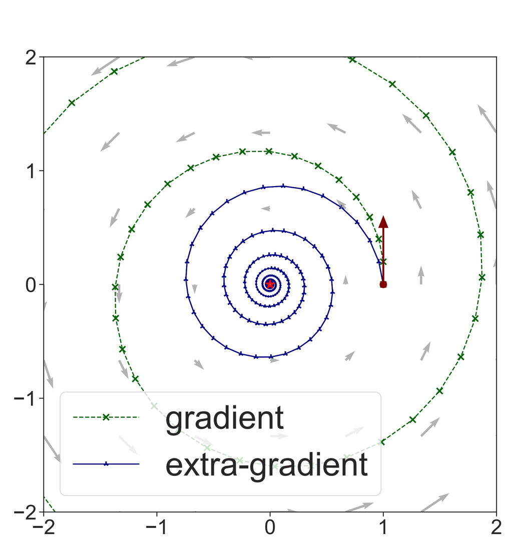
Let $\mathcal{X} := \Theta \times \Phi$, $V := (\nabla_\theta\mathcal{L}, -\nabla_\phi\mathcal{L})$. (VI) gives

$$\forall(\theta, \phi) \in \mathcal{X},\ \langle\nabla_\theta\mathcal{L}(x^\star), \theta - \theta^\star\rangle - \langle\nabla_\phi\mathcal{L}(x^\star), \phi - \phi^\star\rangle \geq 0.$$

- Stationary condition.
- If $\mathcal{L}$ is convex-concave, it solves the original problem.



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



TL;DR

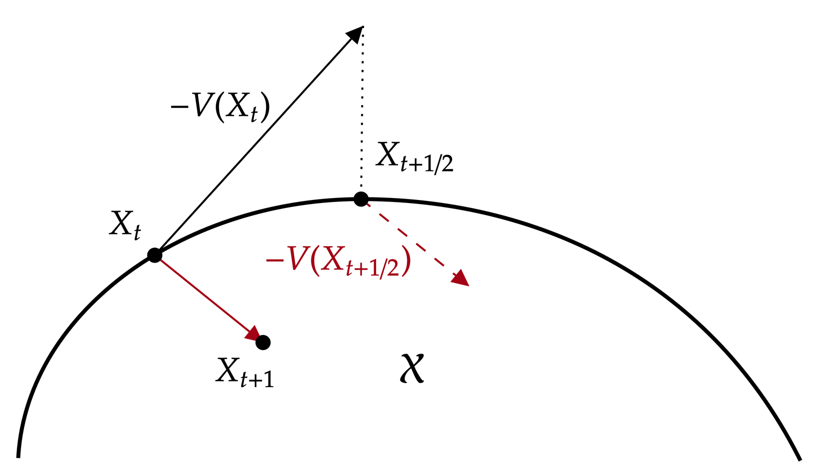
1. *The most widely used single-call variants of Extra-Gradient (EG) are equivalent for unconstrained problems.*
2. *Such single-call EG methods enjoy similar convergence guarantees as EG.*
3. *First local convergence rate analysis for stochastic non-monotone VIs.*

Extra-Gradient (EG)

Extra-Gradient [Korpelevich 1976]

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

The first step anticipates the landscape to achieve better convergence.



But it requires two gradient evaluations per iteration!

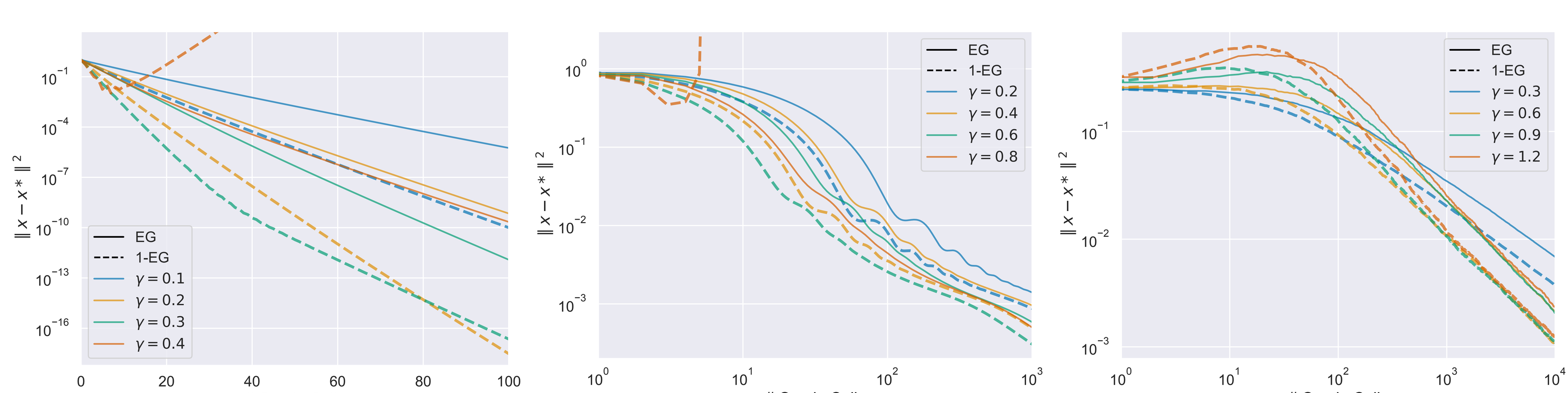
Single-call Extra-Gradient (1-EG)

Past Extra-Gradient [2]	**Reflected Gradient** [3]	**Optimistic Gradient** [4]
$X_{t+\frac{1}{2}} = \Pi_{\mathcal{X}}(X_t - \gamma_t\hat{V}_{t-\frac{1}{2}})$	$X_{t+\frac{1}{2}} = X_t - (X_{t-1} - X_t)$	$X_{t+\frac{1}{2}} = \Pi_{\mathcal{X}}(X_t - \gamma_t\hat{V}_{t-\frac{1}{2}})$
$X_{t+1} = \Pi_{\mathcal{X}}(X_t - \gamma_t\hat{V}_{t+\frac{1}{2}})$	$X_{t+1} = \Pi_{\mathcal{X}}(X_t - \gamma_t\hat{V}_{t+\frac{1}{2}})$	$X_{t+1} = X_{t+\frac{1}{2}} + \gamma_t\hat{V}_{t-\frac{1}{2}} - \gamma_t\hat{V}_{t+\frac{1}{2}}$

Proposition. The above three methods are equivalent in the unconstrained setting.

Illustrative Experiments

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



(a) Strongly monotone; deterministic; last iterate. (b) Monotone; deterministic; ergodic. (c) Non monotone; stochastic; last iterate; local convergence.

Convergence Analysis

The following results hold for all the three variants of 1-EG.

Global Convergence

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

Local Convergence

Definition [Regular Solution $x^\star$].

$$\forall z \in \mathrm{TC}(x^\star),\ z^\top \mathrm{Jac}_V(x^\star)z = \sum_{i,j=1}^{d} z_i\frac{\partial V_i}{\partial x_j}(x^\star)z_j > 0.$$

Theorem. If 1-EG is initialized sufficiently close to $x^\star$ and run with sufficiently small step-sizes, then:

- Deterministic: geometrical convergence of iterates.
- Stochastic:
  (a) The iterates are guaranteed to stay in a neighborhood of $x^\star$ with probability arbitrarily close to 1.
  (b) $\mathbb{E}\left[\|X_t - x^\star\|^2 \mid \text{the above happens}\right] = \mathcal{O}(1/t)$.

Proof Ingredients

Deterministic

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

Stochastic + strongly monotone

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

References

[1] G. M. Korpelevich, *Ekonomika i Matematicheskie Metody* **1976**.
[2] L. D. Popov, *Mathematical Notes* **1980**.
[3] Y. Malitsky, *SIAM Journal on Optimization* **2015**.
[4] C. Daskalakis, A. Ilyas, V. Syrgkanis, H. Zeng, *ICLR* **2018**.
[5] G. Gidel, H. Berard, G. Vignoud, P. Vincent, S. Lacoste-Julien, *ICLR* **2019**.

Read the paper