Explore Aggressively, Update Conservatively: Stochastic Extragradient Methods with Variable Stepsize Scaling

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

NeurIPS 2020

Outline

1. Background: Saddle-point optimization
2. Literature review: Convergence of extragradient
3. Contributions: Explore aggressively, update conservatively

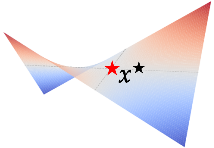
Outline

1. Background: Saddle-point optimization
2. Literature review: Convergence of extragradient
3. Contributions: Explore aggressively, update conservatively

Saddle-point problem

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

$\mathcal{L}(\theta^\star, \phi) \leq \mathcal{L}(\theta^\star, \phi^\star) \leq \mathcal{L}(\theta, \phi^\star)$ for all $\theta \in \mathbb{R}^{d_1}$ and all $\phi \in \mathbb{R}^{d_2}$.

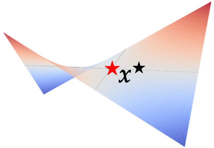


$\mathcal{L} : \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \to \mathbb{R}$ is a differentiable function.

Saddle-point problem

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

$$\mathcal{L}(\theta^\star, \phi) \le \mathcal{L}(\theta^\star, \phi^\star) \le \mathcal{L}(\theta, \phi^\star) \quad \text{for all } \theta \in \mathbb{R}^{d_1} \text{ and all } \phi \in \mathbb{R}^{d_2}.$$



$\mathcal{L} : \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \to \mathbb{R}$ is a differentiable function.

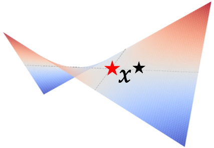
- Arising from: • GANs • adversarial training • robust optimization • self-play in RL . . .
  Caveat: saddle-point problem versus minimax optimization

Saddle-point problem

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

$$\mathcal{L}(\theta^\star, \phi) \le \mathcal{L}(\theta^\star, \phi^\star) \le \mathcal{L}(\theta, \phi^\star) \quad \text{for all } \theta \in \mathbb{R}^{d_1} \text{ and all } \phi \in \mathbb{R}^{d_2}.$$

$\mathcal{L} : \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \to \mathbb{R}$ is a differentiable function.



- Arising from: • GANs • adversarial training • robust optimization • self-play in RL . . .
  Caveat: saddle-point problem versus minimax optimization
- Associated vector field: $V(\theta, \phi) = (\nabla_\theta \mathcal{L}(\theta, \phi), -\nabla_\phi \mathcal{L}(\theta, \phi))$
  First order optimality condition: $V(x^\star) = 0$

The failure of gradient descent/ascent in bilinear games

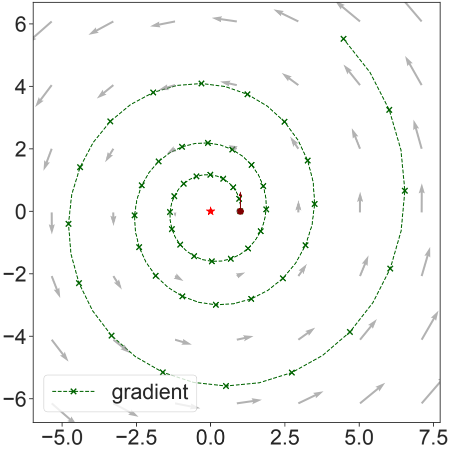
Algorithm

Gradient descent/ascent

$$\theta_{t+1} = \theta_t - \gamma_t \nabla_\theta \mathcal{L}(\theta_t, \phi_t)$$
$$\phi_{t+1} = \phi_t + \gamma_t \nabla_\phi \mathcal{L}(\theta_t, \phi_t)$$

Equivalently, $X_{t+1} = X_t - \gamma_t V(X_t)$.

$$\min_{\theta\in\mathbb{R}} \max_{\phi\in\mathbb{R}} \theta\phi; \quad V(\theta,\phi) = (\phi, -\theta)$$



Remedy: Extragradient [Korpelevich 1976]

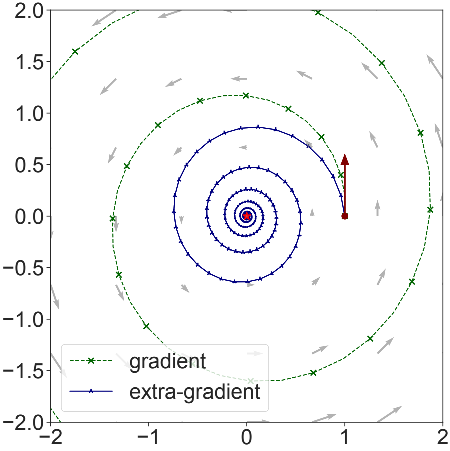
Algorithm

Extra-gradient

$$X_{t+\frac{1}{2}} = X_t - \gamma_t V(X_t) \quad \text{(leading state)}$$

$$X_{t+1} = X_t - \gamma_t V(X_{t+\frac{1}{2}}) \quad \text{(base state)}$$

$$\min_{\theta\in\mathbb{R}} \max_{\phi\in\mathbb{R}} \theta\phi; \qquad V(\theta,\phi) = (\phi, -\theta)$$



Remedy: Extragradient [Korpelevich 1976]

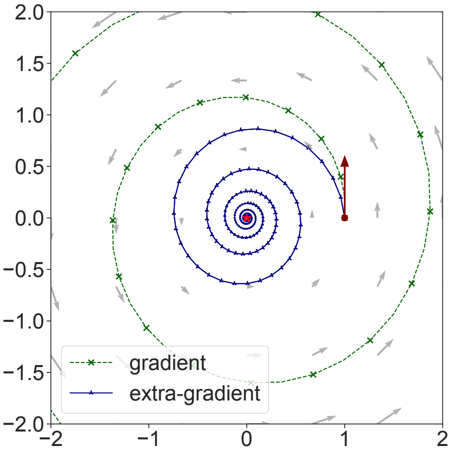
Algorithm

Extra-gradient

$$X_{t+\frac{1}{2}} = X_t - \gamma_t V(X_t) \quad \text{(leading state)}$$

$$X_{t+1} = X_t - \gamma_t V(X_{t+\frac{1}{2}}) \quad \text{(base state)}$$

$$\min_{\theta \in \mathbb{R}} \max_{\phi \in \mathbb{R}} \theta\phi; \qquad V(\theta, \phi) = (\phi, -\theta)$$



Outline

1. Background: Saddle-point optimization
2. Literature review: Convergence of extragradient
3. Contributions: Explore aggressively, update conservatively

Extragradient in the deterministic setting

Blanket assumption: V is β-Lipschitz continuous

Deterministic	Additional Hypothesis	Convergence type	Rate
Korpelevich 1976	Monotone	Last iterate	-
Tseng 1995	Monotone + Error bound	Last iterate	Geometric
Nemirovski 2004	Monotone	Ergodic	$\mathcal{O}(1/t)$

Extragradient in the deterministic setting

Blanket assumption: V is β-Lipschitz continuous

Deterministic	Additional Hypothesis	Convergence type	Rate
Korpelevich 1976	Monotone	Last iterate	-
Tseng 1995	Monotone + Error bound	Last iterate	Geometric
Nemirovski 2004	Monotone	Ergodic	$\mathcal{O}(1/t)$

Extensive literature: • Different convergence metrics and assumptions • Adaptive and universal methods • Dealing with non-smoothness • More efficient variants . . .

Extragradient in the stochastic setting

Stochastic oracle ($s \in \mathbb{N}/2$)

$$\hat{V}_s = V(X_s) + Z_s \qquad (i)\ \mathbb{E}[Z_s \mid \mathcal{F}_s] = 0 \qquad (ii)\ \mathbb{E}[\|Z_s\|^2 \mid \mathcal{F}_s] \leq \sigma^2$$

Extragradient in the stochastic setting

Stochastic oracle ($s \in \mathbb{N}/2$)

$$\hat{V}_s = V(X_s) + Z_s \qquad (i)\ \mathbb{E}[Z_s \,|\, \mathcal{F}_s] = 0 \qquad (ii)\ \mathbb{E}[\|Z_s\|^2 \,|\, \mathcal{F}_s] \le \sigma^2$$

V is β-Lipschitz continuous

Stochastic	Additional Hypothesis	Convergence type	rate
Juditsky et al. 2011	Monotone	Ergodic	$\mathcal{O}(1/\sqrt{t})$
Kannan and Shanbhag 2019	Strongly monotone	Last iterate	$\mathcal{O}(1/t)$
Mertikopoulos et al. 2019	Strictly coherent	Last iterate	-

Extragradient in the stochastic setting

Stochastic oracle ($s \in \mathbb{N}/2$)

$$\hat{V}_s = V(X_s) + Z_s \qquad (i)\ \mathbb{E}[Z_s \mid \mathcal{F}_s] = 0 \qquad (ii)\ \mathbb{E}[\|Z_s\|^2 \mid \mathcal{F}_s] \leq \sigma^2$$

V is β-Lipschitz continuous

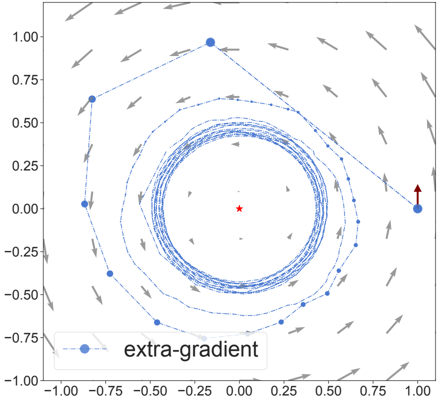
Stochastic	Additional Hypothesis	Convergence type	rate
Juditsky et al. 2011	Monotone	Ergodic	$\mathcal{O}(1/\sqrt{t})$
Kannan and Shanbhag 2019	Strongly monotone	Last iterate	$\mathcal{O}(1/t)$
Mertikopoulos et al. 2019	Strictly coherent	Last iterate	-

Last-iterate convergence for stochastic monotone operators?

$$\min_{\theta\in\mathbb{R}} \max_{\phi\in\mathbb{R}} \theta\phi; \quad \hat{V}_t = (\phi_t + \xi_t, -\theta_t)$$

$$\mathbb{E}[\xi_t] = 0, \quad \mathbb{E}[\xi_t^2] = \sigma^2 > 0.$$

Non-convergence: Solutions?



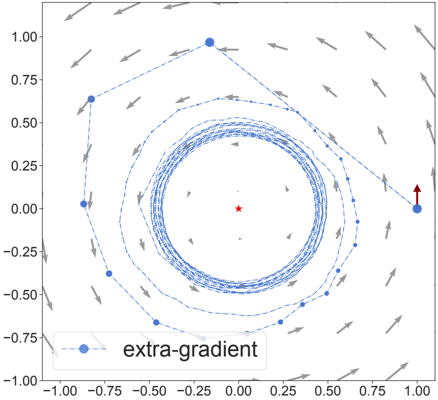
Last-iterate convergence for stochastic monotone operators?

$$\min_{\theta \in \mathbb{R}} \max_{\phi \in \mathbb{R}} \theta\phi; \quad \hat{V}_t = (\phi_t + \xi_t, -\theta_t)$$

$$\mathbb{E}[\xi_t] = 0, \quad \mathbb{E}[\xi_t^2] = \sigma^2 > 0.$$

Candidate solutions:

- Regularization with vanishing weight

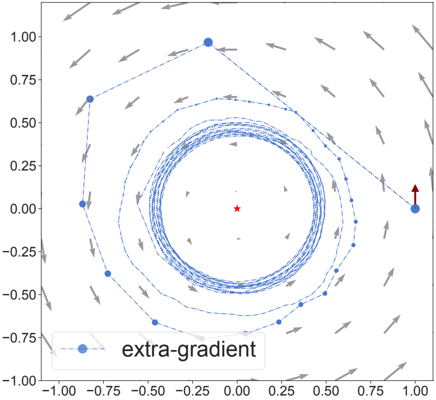


Last-iterate convergence for stochastic monotone operators?

$$\min_{\theta \in \mathbb{R}} \max_{\phi \in \mathbb{R}} \theta\phi; \quad \hat{V}_t = (\phi_t + \xi_t, -\theta_t)$$
$$\mathbb{E}[\xi_t] = 0, \quad \mathbb{E}[\xi_t^2] = \sigma^2 > 0.$$

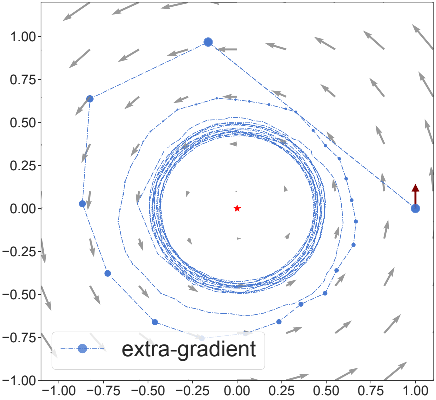
Candidate solutions:

- Regularization with vanishing weight
- Variance reduction with increasing batch size



Last-iterate convergence for stochastic monotone operators?

$$\min_{\theta\in\mathbb{R}} \max_{\phi\in\mathbb{R}} \theta\phi; \quad \hat{V}_t = (\phi_t + \xi_t, -\theta_t)$$
$$\mathbb{E}[\xi_t] = 0, \quad \mathbb{E}[\xi_t^2] = \sigma^2 > 0.$$



Candidate solutions:

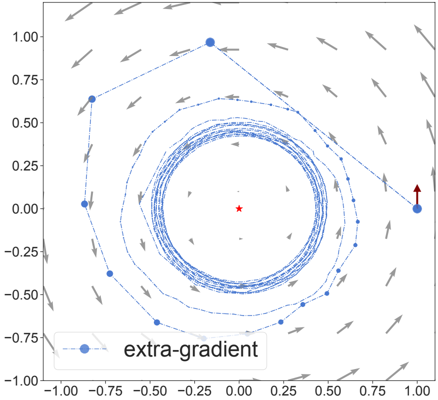
- Regularization with vanishing weight
- Variance reduction with increasing batch size
- Finite sum: SVRG-like variance reduction

Last-iterate convergence for stochastic monotone operators?

$$\min_{\theta\in\mathbb{R}} \max_{\phi\in\mathbb{R}} \theta\phi; \quad \hat{V}_t = (\phi_t + \xi_t, -\theta_t)$$
$$\mathbb{E}[\xi_t] = 0, \quad \mathbb{E}[\xi_t^2] = \sigma^2 > 0.$$

Candidate solutions:

- Regularization with vanishing weight
- Variance reduction with increasing batch size
- Finite sum: SVRG-like variance reduction
- Second-order: stochastic Hamiltonian descent



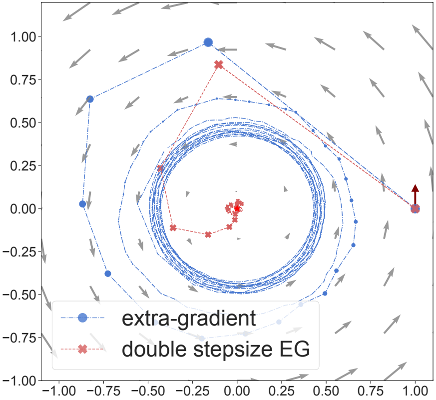
Last-iterate convergence for stochastic monotone operators?

$$\min_{\theta\in\mathbb{R}} \max_{\phi\in\mathbb{R}} \theta\phi; \quad \hat{V}_t = (\phi_t + \xi_t, -\theta_t)$$

$$\mathbb{E}[\xi_t] = 0, \quad \mathbb{E}[\xi_t^2] = \sigma^2 > 0.$$

Candidate solutions:

- Regularization with vanishing weight
- Variance reduction with increasing batch size
- Finite sum: SVRG-like variance reduction
- Second-order: stochastic Hamiltonian descent
- Different stepsizes for the two steps of EG!



Outline

1. Background: Saddle-point optimization
2. Literature review: Convergence of extragradient
3. Contributions: Explore aggressively, update conservatively

Our proposal: Double stepsize extragradient

$$\begin{aligned} X_{t+\frac{1}{2}} &= X_t - \gamma_t \hat{V}_t \\ X_{t+1} &= X_t - \gamma_t \hat{V}_{t+\frac{1}{2}} \end{aligned} \quad \longrightarrow \quad \begin{aligned} X_{t+\frac{1}{2}} &= X_t - \gamma_t \hat{V}_t \\ X_{t+1} &= X_t - \eta_t \hat{V}_{t+\frac{1}{2}} \end{aligned}$$

- Explore aggressively, update conservatively: $\eta_t \leq \gamma_t$ (frequently $\eta_t/\gamma_t \to 0$)

Our proposal: Double stepsize extragradient

$$\begin{aligned} X_{t+\frac{1}{2}} &= X_t - \gamma_t \hat{V}_t \\ X_{t+1} &= X_t - \gamma_t \hat{V}_{t+\frac{1}{2}} \end{aligned} \longrightarrow \begin{aligned} X_{t+\frac{1}{2}} &= X_t - \gamma_t \hat{V}_t \\ X_{t+1} &= X_t - \eta_t \hat{V}_{t+\frac{1}{2}} \end{aligned}$$

- Explore aggressively, update conservatively: $\eta_t \le \gamma_t$ (frequently $\eta_t / \gamma_t \to 0$)
- Stochastic oracle $\hat{V}_s = V(X_s) + Z_s$

  (i) $\mathbb{E}[Z_s \mid \mathcal{F}_s] = 0$

  (ii) $\mathbb{E}[\|Z_s\|^2 \mid \mathcal{F}_s] \le (\sigma + \kappa \|X_s - x^\star\|)^2, \forall x^\star \in \mathcal{X}$

Our proposal: Double stepsize extragradient

$$\begin{aligned} X_{t+\frac{1}{2}} &= X_t - \gamma_t \hat{V}_t \\ X_{t+1} &= X_t - \gamma_t \hat{V}_{t+\frac{1}{2}} \end{aligned} \longrightarrow \begin{aligned} X_{t+\frac{1}{2}} &= X_t - \gamma_t \hat{V}_t \\ X_{t+1} &= X_t - \eta_t \hat{V}_{t+\frac{1}{2}} \end{aligned}$$

- Explore aggressively, update conservatively: $\eta_t \leq \gamma_t$ (frequently $\eta_t / \gamma_t \to 0$)
- Stochastic oracle $\hat{V}_s = V(X_s) + Z_s$

  (i) $\mathbb{E}[Z_s \,|\, \mathcal{F}_s] = 0$

  (ii) $\mathbb{E}[\|Z_s\|^2 \,|\, \mathcal{F}_s] \leq (\sigma + \kappa \|X_s - x^\star\|)^2, \forall x^\star \in \mathcal{X}$

  (ii') $\hat{V}_s = \hat{V}(\xi, X_s)$; $\hat{V}(\xi, \cdot)$ is $(\kappa/2)$-Lipschitz; $\hat{V}$ has bounded variance on $\mathcal{X}^\star$

Descent inequality and assumptions

Descent Lemma ($\kappa = 0$)

$$\mathbb{E}[\|X_{t+1} - x^\star\|^2 \mid \mathcal{F}_t] \le \|X_t - x^\star\|^2 - 2\eta_t \, \mathbb{E}_t[\langle V(X_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - x^\star\rangle] - (\gamma_t\eta_t - \gamma_t^3\eta_t\beta^2)\|V(X_t)\|^2 + (2\gamma_t^2\eta_t\beta + \gamma_t^3\eta_t\beta^2 + \eta_t^2)\sigma^2.$$

1. Variational stability: $\langle V(x), x - x^\star\rangle \ge 0$ for all $x \in \mathbb{R}^d$, $x^\star \in \mathcal{X}^\star$.
   Bilinear ⊂ Convex-concave ⊂ Monotone ⊂ Pseudo-monotone ⊂ Variationally stable

Descent inequality and assumptions

Descent Lemma ($\kappa = 0$)

$$\mathbb{E}[\|X_{t+1} - x^\star\|^2 \mid \mathcal{F}_t] \le \|X_t - x^\star\|^2 - 2\eta_t \mathbb{E}_t[\langle V(X_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - x^\star\rangle] - (\gamma_t\eta_t - \gamma_t^3\eta_t\beta^2)\|V(X_t)\|^2 + (2\gamma_t^2\eta_t\beta + \gamma_t^3\eta_t\beta^2 + \eta_t^2)\sigma^2.$$

1. Variational stability: $\langle V(x), x - x^\star\rangle \ge 0$ for all $x \in \mathbb{R}^d$, $x^\star \in \mathcal{X}^\star$.
   Bilinear ⊂ Convex-concave ⊂ Monotone ⊂ Pseudo-monotone ⊂ Variationally stable
2. Error bound: For some $\tau > 0$ and all $x \in \mathbb{R}^d$, we have $\|V(x)\| \ge \tau \operatorname{dist}(x, \mathcal{X}^\star)$.
   e.g., Affine, strongly monotone operators...

Descent inequality and assumptions

Descent Lemma ($\kappa = 0$)

$$\mathbb{E}[\|X_{t+1} - x^\star\|^2 \mid \mathcal{F}_t] \leq \|X_t - x^\star\|^2 - 2\eta_t \, \mathbb{E}_t[\langle V(X_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - x^\star \rangle] - (\gamma_t\eta_t - \gamma_t^3\eta_t\beta^2)\|V(X_t)\|^2 + (2\gamma_t^2\eta_t\beta + \gamma_t^3\eta_t\beta^2 + \eta_t^2)\sigma^2.$$

1. Variational stability: $\langle V(x), x - x^\star \rangle \geq 0$ for all $x \in \mathbb{R}^d$, $x^\star \in \mathcal{X}^\star$.
   Bilinear $\subset$ Convex-concave $\subset$ Monotone $\subset$ Pseudo-monotone $\subset$ Variationally stable
2. Error bound: For some $\tau > 0$ and all $x \in \mathbb{R}^d$, we have $\|V(x)\| \geq \tau \operatorname{dist}(x, \mathcal{X}^\star)$.
   e.g., Affine, strongly monotone operators...

Descent inequality and assumptions

Descent Lemma ($\kappa = 0$)

$$\mathbb{E}[\|X_{t+1} - x^\star\|^2 \mid \mathcal{F}_t] \leq \|X_t - x^\star\|^2 - 2\eta_t \mathbb{E}_t[\langle V(X_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - x^\star\rangle] - (\gamma_t\eta_t - \gamma_t^3\eta_t\beta^2)\|V(X_t)\|^2 + (2\gamma_t^2\eta_t\beta + \gamma_t^3\eta_t\beta^2 + \eta_t^2)\sigma^2.$$

1. Variational stability: $\langle V(x), x - x^\star\rangle \geq 0$ for all $x \in \mathbb{R}^d$, $x^\star \in \mathcal{X}^\star$.
   Bilinear ⊂ Convex-concave ⊂ Monotone ⊂ Pseudo-monotone ⊂ Variationally stable
2. Error bound: For some $\tau > 0$ and all $x \in \mathbb{R}^d$, we have $\|V(x)\| \geq \tau\,\mathrm{dist}(x, \mathcal{X}^\star)$.
   e.g., Affine, strongly monotone operators...

Descent inequality and assumptions

Descent Lemma ($\kappa = 0$)

$$\mathbb{E}[\|X_{t+1} - x^\star\|^2 \mid \mathcal{F}_t] \le \|X_t - x^\star\|^2 - 2\eta_t \mathbb{E}_t[\langle V(X_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - x^\star\rangle] - (\gamma_t\eta_t - \gamma_t^3\eta_t\beta^2)\|V(X_t)\|^2 + (2\gamma_t^2\eta_t\beta + \gamma_t^3\eta_t\beta^2 + \eta_t^2)\sigma^2.$$

$$\eta_t < \gamma_t$$

Convergence results

Theorem [Main result]

1. Let V be variationally stable. Assume that $\sum_t \gamma_t \eta_t = \infty$, $\sum_t \eta_t^2 < \infty$, $\sum_t \gamma_t^2 \eta_t < \infty$, $\gamma_t \leq c/\beta$ with $c < 1$. Then $(X_t)_{t \in \mathbb{N}}$ converges to a point $x^\star \in \mathcal{X}^\star$ almost surely.

Convergence results

Theorem [Main result]

1. Let V be variationally stable. Assume that $\sum_t \gamma_t \eta_t = \infty$, $\sum_t \eta_t^2 < \infty$, $\sum_t \gamma_t^2 \eta_t < \infty$, $\gamma_t \leq c/\beta$ with $c < 1$. Then $(X_t)_{t\in\mathbb{N}}$ converges to a point $x^\star \in \mathcal{X}^\star$ almost surely.

2. Let V be monotone and affine. With stepsizes $\gamma_t \equiv \gamma$ and $\eta_t = \Theta(1/t)$,

$$\mathbb{E}[\operatorname{dist}(X_t, \mathcal{X}^\star)^2] = \mathcal{O}\left(1/t\right)$$

Convergence results

Theorem [Main result]

1. Let V be variationally stable. Assume that $\sum_t \gamma_t \eta_t = \infty$, $\sum_t \eta_t^2 < \infty$, $\sum_t \gamma_t^2 \eta_t < \infty$, $\gamma_t \leq c/\beta$ with $c < 1$. Then $(X_t)_{t \in \mathbb{N}}$ converges to a point $x^\star \in \mathcal{X}^\star$ almost surely.

2. Let V be monotone and affine. With stepsizes $\gamma_t \equiv \gamma$ and $\eta_t = \Theta(1/t)$,

$$\mathbb{E}[\operatorname{dist}(X_t, \mathcal{X}^\star)^2] = \mathcal{O}(1/t)$$

3. Let V be variationally stable and satisfy the error bound condition. With stepsizes of the form $\gamma_t = \gamma/(t+b)^{1/3}$ and $\eta_t = \eta/(t+b)^{2/3}$,

$$\mathbb{E}[\operatorname{dist}(X_t, \mathcal{X}^\star)^2] = \mathcal{O}\left(1/\sqrt[3]{t}\right)$$

Convergence results

Theorem [Main result]

1. Let V be variationally stable. Assume that $\sum_t \gamma_t \eta_t = \infty$, $\sum_t \eta_t^2 < \infty$, $\sum_t \gamma_t^2 \eta_t < \infty$, $\gamma_t \leq c/\beta$ with $c < 1$. Then $(X_t)_{t \in \mathbb{N}}$ converges to a point $x^\star \in \mathcal{X}^\star$ almost surely.

2. Let V be monotone and affine. With stepsizes $\gamma_t \equiv \gamma$ and $\eta_t = \Theta(1/t)$,

$$\mathbb{E}[\operatorname{dist}(X_t, \mathcal{X}^\star)^2] = \mathcal{O}(1/t)$$

3. Let V be variationally stable and satisfy the error bound condition. Further suppose that the noise vanishes on the solution set (i.e., $\sigma = 0$). With suitable constant stepsizes,

$$\mathbb{E}[\operatorname{dist}(X_t, \mathcal{X}^\star)^2] = \mathcal{O}\left(e^{-\rho t}\right)$$

Convergence results

$\gamma_t = \eta_t$	Does not converge in bilinear game
$\sum_t \gamma_t \eta_t = \infty$, $\sum_t \eta_t^2 < \infty$, $\sum_t \gamma_t^2 \eta_t < \infty$	a.s. convergence for monotone/VS operators
$\gamma_t = \eta_t = \gamma/(t+b)$	$\mathcal{O}(1/t)$ for strongly monotone operators
$\gamma_t = \gamma/(t+b)^{1/3}$, $\eta_t = \eta/(t+b)^{2/3}$	$\mathcal{O}(1/\sqrt[3]{t})$ under error bound condition + VS
$\gamma_t \equiv \gamma$, $\eta_t = \eta/(t+b)$	$\mathcal{O}(1/t)$ for affine and monotone operators

Beyond monotonicity: Local convergence

Theorem

Assumptions:

(i) Locally variational stable and locally Lipschitz around a soultion $x^\star$.
(ii) V is differentiable at $x^\star$ and $\mathrm{Jac}_V(x^\star)$ is invertible.

Beyond monotonicity: Local convergence

Theorem

Assumptions:

(i) Locally variational stable and locally Lipschitz around a soultion $x^\star$.
(ii) V is differentiable at $x^\star$ and $\mathrm{Jac}_V(x^\star)$ is invertible.

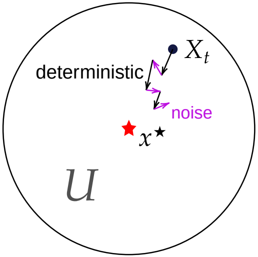
Guarantee:

For any tolerance level $\delta > 0$, there exists a stepsize policy for double stepsize extra-gradient such that if the algorithm is initialized close enough to $x^\star$, there exists an event with probability at least $1 - \delta$ and, conditioned on this event:

- Under (i), the iterates converge to $x^\star$.
- Under (i) and (ii), X_t converges to $x^\star$ at a rate $\mathcal{O}\left(1/\sqrt[3]{t}\right)$ in mean square error.

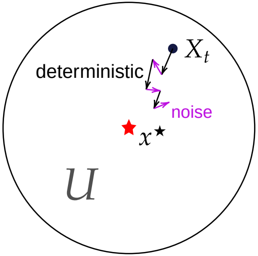
Proof sketch

- **Stability of the algorithm**
  Control the probability of escaping from the neighborhood at each step: thanks to the use of the specific stepsize policy, we prove the summability of these probabilities and that this sum can be made arbitrarily small.

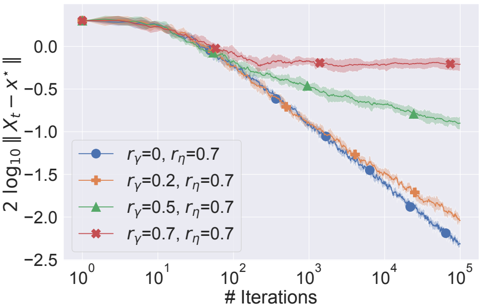


Proof sketch

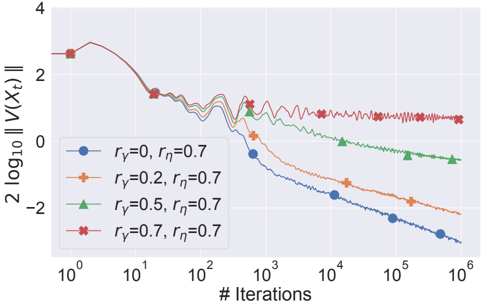
- **Stability of the algorithm**
  Control the probability of escaping from the neighborhood at each step: thanks to the use of the specific stepsize policy, we prove the summability of these probabilities and that this sum can be made arbitrarily small.
- **Conditional convergence rate**
  Caveat. The unbiasedness is not maintained after conditioning.
  Solution. Work directly with the indicator function of the probability event. Precisely, we prove recurrent bounds for $\mathbb{E}[\|X_t - x^\star\|^2 \mathbb{1}_{E_{t-1}} \mid \mathcal{F}_{t-1}]$.



Numerical illustrations



(a) Bilinear zero-sum game



(b) Linear quadratic gaussian GAN

$$X_{t+\frac{1}{2}} = X_t - \gamma_t \hat{V}_t \quad [\gamma_t = \gamma/(t+b)^{r_\gamma}] \qquad X_{t+1} = X_t - \eta_t \hat{V}_{t+\frac{1}{2}} \quad [\eta_t = \eta/(t+b)^{r_\eta}]$$

Conclusion

- We propose a simple modification of the stochastic extragradient scheme to make its last iterate converge in a large spectrum of problems including all monotone games.

Conclusion

- We propose a simple modification of the stochastic extragradient scheme to make its last iterate converge in a large spectrum of problems including all monotone games.
- Explicit convergence rate under additional assumptions and local convergence results are derived.

Conclusion

- We propose a simple modification of the stochastic extragradient scheme to make its last iterate converge in a large spectrum of problems including all monotone games.
- Explicit convergence rate under additional assumptions and local convergence results are derived.

Thanks for your attention!

Bibliography I

- 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.
- Kannan, Aswin and Uday V Shanbhag (2019). "Optimal stochastic extragradient schemes for pseudomonotone stochastic variational inequality problems and their variants". In: *Computational Optimization and Applications* 74.3, pp. 779–820.
- Korpelevich, G. M. (1976). "The extragradient method for finding saddle points and other problems". In: *Èkonom. i Mat. Metody* 12, pp. 747–756.
- Mertikopoulos, Panayotis et al. (2019). "Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile". In: *ICLR '19: Proceedings of the 2019 International Conference on Learning Representations.*
- 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.
- Tseng, Paul (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.