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

Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, Panayotis Mertikopoulos (Univ. Grenoble Alpes)

NEURAL INFORMATION PROCESSING SYSTEMS

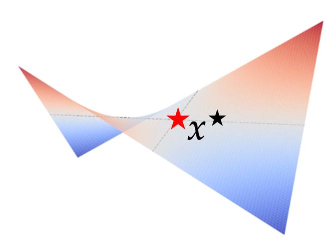
Saddle-point Optimization

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

$$\forall \theta \in \mathbb{R}^{d_1},\ \forall \phi \in \mathbb{R}^{d_2},\ \mathcal{L}(\theta^\star, \phi) \le \mathcal{L}(\theta^\star, \phi^\star) \le \mathcal{L}(\theta, \phi^\star)$$

Associated vector field:

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



First order optimality condition: $V(x^\star) = 0$

Applications. • Generative adversarial networks • Adversarial training • Self-play • Robust optimization

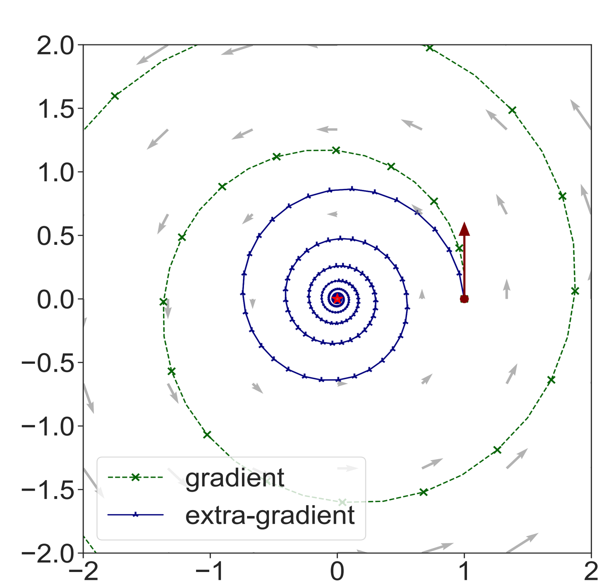
Extragradient and its Failure

From gradient to extragradient

$$\mathcal{L} : (\theta, \phi) \in \mathbb{R} \times \mathbb{R} \mapsto \theta \cdot \phi; \quad V(\theta, \phi) = (\phi, -\theta); \quad x^\star = (0, 0)$$

Algorithms

1. Gradient method:
$$X_{t+1} = X_t - \gamma_t V(X_t)$$
2. Extragradient (EG):
$$X_{t+\frac{1}{2}} = X_t - \gamma_t V(X_t)$$
$$X_{t+1} = X_t - \gamma_t V(X_{t+\frac{1}{2}})$$

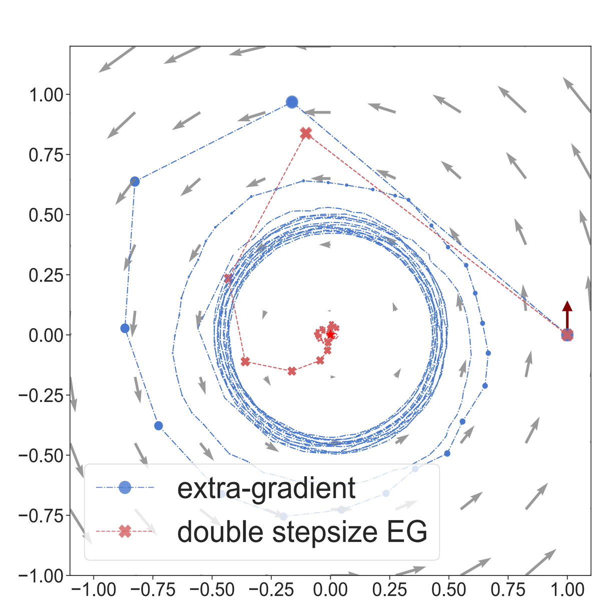


The non-convergence of stochastic EG

$$\hat{V}_t = (\phi_t + \xi_t, -\theta_t); \quad \mathbb{E}[\xi_t] = 0,\ \mathbb{E}[\xi_t^2] \ge \sigma^2 > 0$$

Propostion. Whatever stepsize is used, running EG with oracle feedback $\hat{V}_t$ leads to $\liminf_{t\to\infty} \mathbb{E}[\theta_t^2 + \phi_t^2] > 0$.

N.B. EG is known to converge ergodically in $\mathcal{O}(1/\sqrt{t})$ in all stochastic monotone problems. However, in this work, we are interested in its last-iterate convergence.



A Remedy with Double Stepsize

DSEG

$$X_{t+\frac{1}{2}} = X_t - \gamma_t \hat{V}_t, \qquad X_{t+1} = X_t - \eta_t \hat{V}_{t+\frac{1}{2}}$$

Explore aggressively, update conservatively: $\eta_t \le \gamma_t$

Assumptions

On the **operator**

1. β-Lipschitz continuity (L): $\|V(x) - V(x')\| \le \beta \|x - x'\|$
2. Variational stability (VS): $\langle V(x), x - x^\star \rangle \ge 0$
3. Error bound (EB): $\exists \tau > 0, \forall x, \|V(x)\| \ge \tau \operatorname{dist}(x, \mathcal{X}^\star)$ — solution set

On the **noise** ($\forall s \in \mathbb{N}/2, \hat{V}_s = V(X_s) + Z_s$)

1. Unbiasedness: $\mathbb{E}[Z_s \mid \mathcal{F}_s] = 0$
2. Variance control: $\mathbb{E}[\|Z_s\|^2 \mid \mathcal{F}_s] \le (\sigma + \kappa \|X_s - x^\star\|)^2, \forall x^\star \in \mathcal{X}^\star$

A descent lemma

Assume (L). Let $C_t = 4\gamma_t^2 \eta_t \beta + 2\gamma_t^3 \eta_t \beta^2 + 4\eta_t^2 + 16\gamma_t^2 \eta_t^2 \kappa^2$. Then

$$\mathbb{E}[\|X_{t+1} - x^\star\|^2 \mid \mathcal{F}_t] \le \underbrace{(1 + C_t \kappa^2)}_{\to 1} \|X_t - x^\star\|^2 \underbrace{- 2\eta_t \mathbb{E}[\langle V(X_{t+\frac{1}{2}}), X_{t+\frac{1}{2}} - x^\star \rangle \mid \mathcal{F}_t]}_{\le 0 \text{ (VS)}} \underbrace{- \gamma_t \eta_t (1 - \gamma_t^2 \beta^2 - 8\gamma_t \eta_t \kappa^2) \|V(X_t)\|^2}_{<0 \text{ possible to use (EB)}} + \underbrace{C_t \sigma^2}_{\ge 0}$$

The decrement term is in $\Theta(\gamma_t \eta_t)$ while the noise term is in $\Theta(\eta_t^2)$

Convergence Result

Asymptotic convergence

Stepsize condition (SC). $\sum_t \gamma_t \eta_t = \infty$, $\sum_t \eta_t^2 < \infty$, $\sum_t \gamma_t^2 \eta_t < \infty$

Theorem. Assume (L) + (VS). DSEG with (SC) and $\gamma_t \le c/\beta$ for some $c < 1$ converges to a solution $x^\star$ almost surely.

Convergence rate

Consider $\gamma_t = \gamma/(t+b)^{r_\gamma}$ and $\eta_t = \eta/(t+b)^{r_\eta}$

$r_\gamma = r_\eta = 1$	Strongly Monotone	$\mathcal{O}(1/t)$
$r_\gamma = 1/3, r_\eta = 2/3$	(EB) + (VS)	$\mathcal{O}(1/t^{1/3})$
$r_\gamma = r_\eta = 0$ $(\eta < \gamma)$	(EB) + (VS) + $(\sigma = 0)$	$\mathcal{O}(e^{-\rho t})$
$r_\gamma = 0, r_\eta = 1$	Affine Monotone	$\mathcal{O}(1/t)$

Local convergence

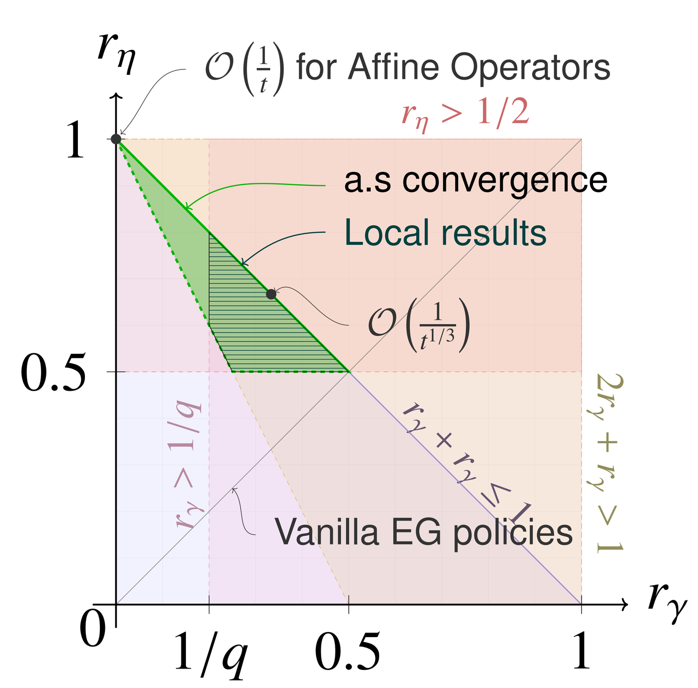
New!

(a) Around solution $x^\star$: (L) + (VS) + q-th moment control for Z_t

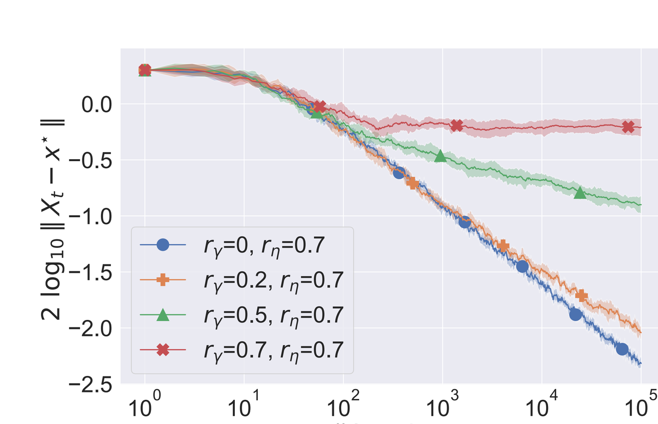
(b) $\operatorname{Jac}_V(x^\star)$ is defined and invertible $\Rightarrow$ local (EB)

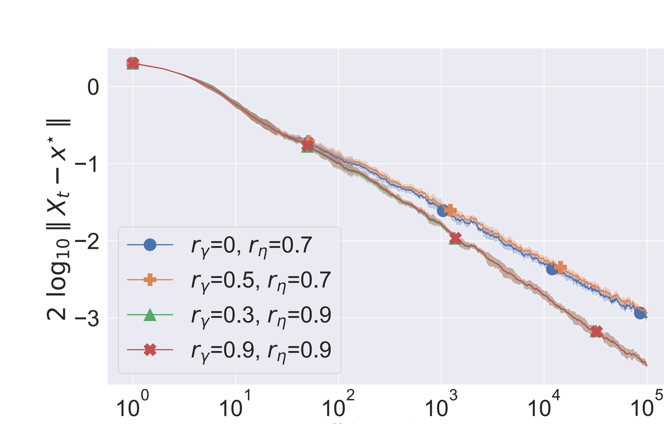
Theorem. Fix tolerance level $\delta \in (0, 1)$. DSEG with close enough initialization and suitable stepsizes guarantees:

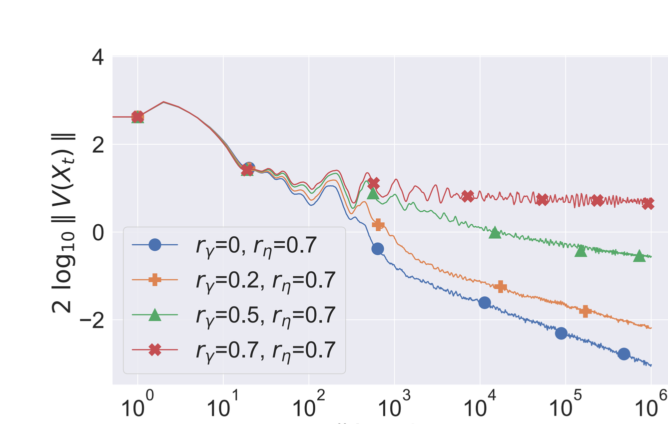
- If (a), then $\mathbb{P}(X_t \to x^\star) \ge 1 - \delta$
- If (a)+(b), then there exists event E such that $\mathbb{P}(E) \ge 1 - \delta$ and
$$\mathbb{E}[\|X_t - x^\star\|^2 \mid E] = \mathcal{O}\left(1/t^{1/3}\right)$$



Numerical Illustrations







Left Bilinear zero-sum game

Middle Strongly monotone (biquadratic)

Right Linear quadratic Gaussian WGAN

$x \sim \mathcal{N}(0, \Sigma)$

$G(z) = Yz, D(x) = x^\top W x$

Read the paper