Last Iterate is Slower than Averaged Iterate in Smooth Convex-Concave Saddle Point Problems

Noah Golowich Sarath Pattathil Constantinos Daskalakis Asuman Ozdaglar

Massachusetts Institute of Technology

Min-max optimization

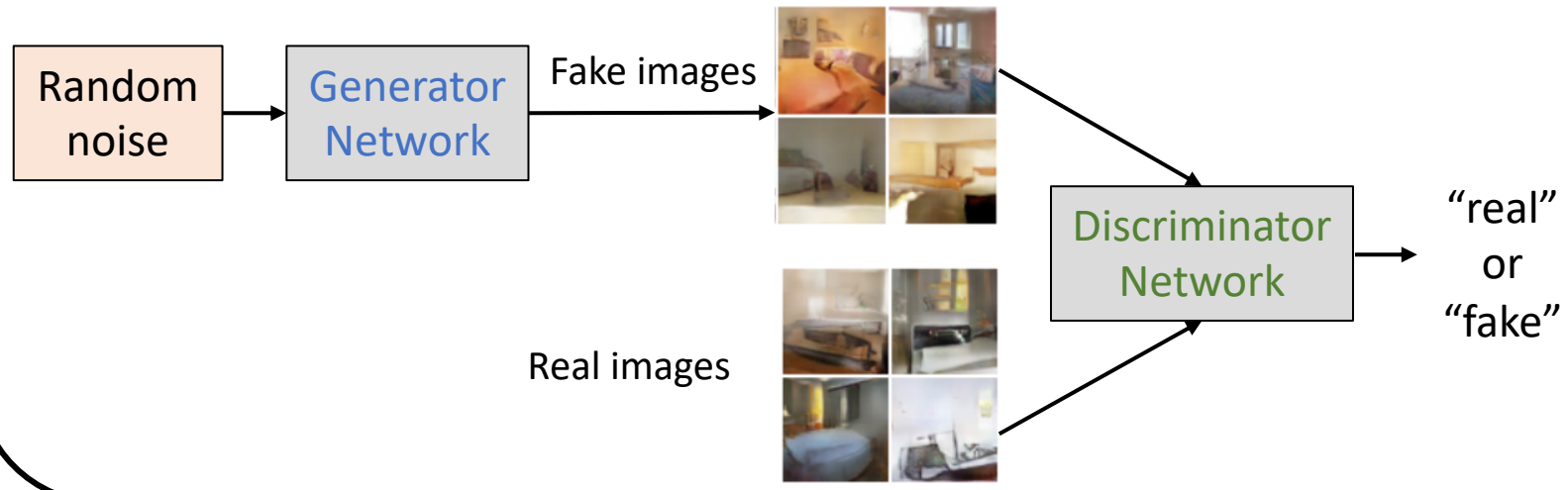
- Two agents ('min player' and 'max player') choose inputs $x, y \in \mathbb{R}^n$ to a function f:

$$\min_x \max_y f(x, y)$$

Example: Wasserstein GANs *[Arjovsky et al., '17]*

- Min player x: parameters of generator network G_x.
- Max player y: parameters of discriminator network D_y.

$$f(x, y) = \mathbb{E}[D_y(\text{real images})] - \mathbb{E}[D_y(G_x(\text{noise}))].$$



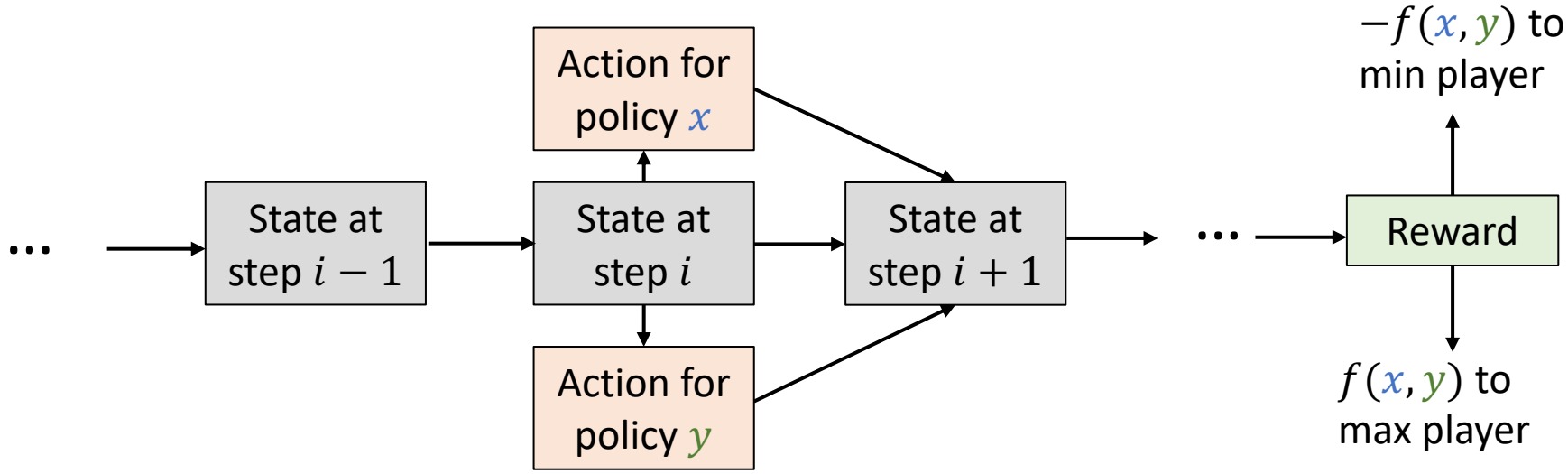
Min-max optimization

- Two agents ('min player' and 'max player') choose inputs $x, y \in \mathbb{R}^n$ to a function f:

$$\min_x \max_y f(x, y)$$

Example: multi-agent RL *[Shapley, '53]*

- $f(x, y)$ is value of a zero-sum **Markov game** with policies x, y:



Convex-concave saddle point problem

$$\min_x \max_y f(x, y)$$

- We assume $f(x, y)$ is convex in x and concave in y (**convex-concave**).
  - Also assume L-smooth (i.e., L-Lipschitz gradients).
- Under mild additional assumptions *[Sion, '58]*, exists **Nash equilibrium** (x^*, y^*), i.e.,

$$f(x^*, y) \leq f(x^*, y^*) \leq f(x, y^*) \quad \forall x, y$$

- Goal: find algorithms for **approximate Nash equilibrium** $(\hat{x}, \hat{y})$:

$$\text{Gap}(\hat{x}, \hat{y}) = \max_{y \in \mathcal{Y}} f(\hat{x}, y) - \min_{x \in \mathcal{X}} f(x, \hat{y})$$

- Note that $\text{Gap}(x^*, y^*) = 0$.

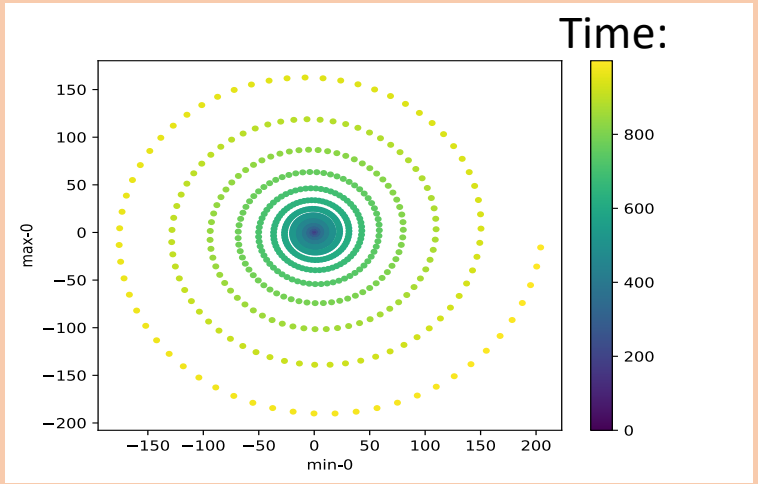
First attempt: GDA

$$\min_x \max_y f(x, y)$$

- Unconstrained setting, i.e., assume $f(x, y)$ defined for all $x, y \in \mathbb{R}^n$.
- **Gradient descent-ascent (GDA):**
  - Initialize $x_0, y_0 \in \mathbb{R}^n$, fix step size $\eta > 0$.
  - For $t \geq 0$, update:

$$x_{t+1} \leftarrow x_t - \eta \nabla_x f(x_t, y_t), \qquad y_{t+1} \leftarrow y_t + \eta \nabla_y f(x_t, y_t)$$

Negative result: iterates of GDA, x_t, y_t diverge:



$$\mathrm{Gap}(x_t, y_t) = \max_{y \in \mathcal{Y}} f(x_t, y) - \min_{x \in \mathcal{X}} f(x, y_t)$$

Extragradient (EG) algorithm

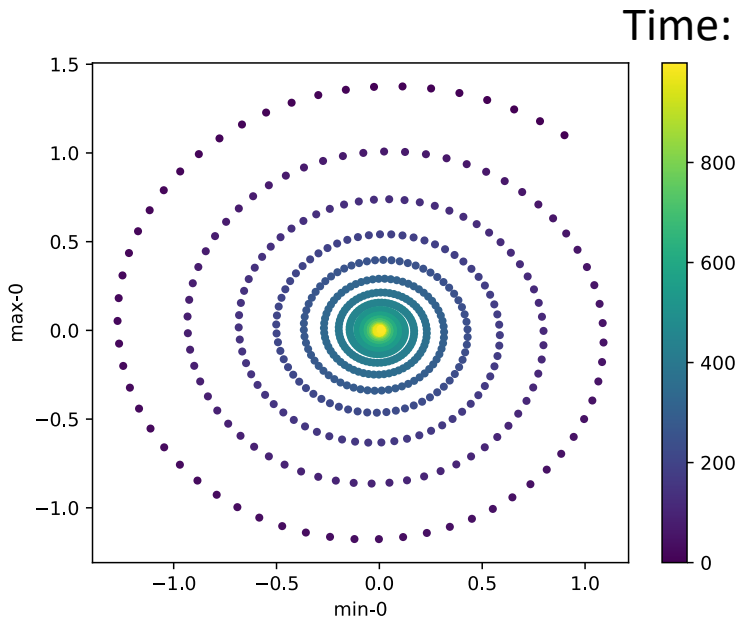
- **Extragradient** algorithm *[Korpelevich, '76];* extra gradient at each time t:

$$x_{t+1/2} = x_t - \eta \nabla_x f(x_t, y_t), \qquad y_{t+1/2} = y_t + \eta \nabla_y f(x_t, y_t)$$

$$x_{t+1} = x_t - \eta \nabla_x f\left(x_{t+1/2}, y_{t+1/2}\right), \qquad y_{t+1} = y_t + \eta \nabla_y f(x_{t+1/2}, y_{t+1/2})$$

- **Theorem** *[Korpelevich, '76]:* Iterates (x_t, y_t) of Extragradient algorithm converge to (x^*, y^*) as $t \to \infty$.

Main question: what are rates of convergence? I.e., how fast does $\mathrm{Gap}(x_t, y_t) \to 0$?



Prior convergence rates for convex-concave saddle point problem

- **Averaged iterate** convergence for EG algorithm *[Nemirovski, '04]:*
  - Let $\bar{x}_t = \frac{1}{t}\sum_{1\le s\le t} x_s$ and $\bar{y}_t = \frac{1}{t}\sum_{1\le s\le t} y_s$ be averaged iterates up to time t
  - Then $\mathrm{Gap}(\bar{x}_t, \bar{y}_t) \le O(1/t)$.
- **Best iterate** convergence for EG algorithm *[Monteiro & Svaiter, '10]:*
  - $\min_{1\le s\le t} \mathrm{Gap}(x_s, y_s) \le O(1/\sqrt{t})$.
- What about $\mathrm{Gap}(x_t, y_t)$?

Last-iterate convergence

Goal: prove explicit convergence rate for iterates (x_t, y_t) of some algorithm (e.g., extragradient) with constant step size.

- Motivation:
  - *No theoretical guarantees for averaging* in nonconvex-nonconcave case.
  - **Multi-agent learning**: averaged and best iterates *do not describe game dynamics (e.g., [Mertikopoulos et al., '18]).*
  - **Why constant step size?** Step-size decay uses *new information with decreasing weight* (implausible in practice) *[Lin et al., '20].*

Last-iterate convergence rate for EG

- **Smoothness assumption**: Convex-concave $f(x, y)$ satisfies:
  - L-Lipschitz gradient
  - Λ-Lipschitz Hessian
- **Initialization of EG**: at (x_0, y_0) of distance $\leq D$ from equilibrium (x^*, y^*).
- **Step size of EG**: $\eta \leq \mathrm{O}(\min\left\{\frac{1}{\Lambda \mathrm{D}}, \frac{1}{L}\right\})$

Theorem (this paper): If $f(x, y)$ satisfies above assumptions, then iterates (x_t, y_t) of EG satisfy:

$$\mathrm{Gap}(x_t, y_t) \leq O(\frac{D^2}{\eta\sqrt{t}})$$

- Compare to rate for averaged iterates: $O(D^2/\eta t)$ – **this gap is necessary!**

Last-iterate convergence no faster than $1/\sqrt{T}$

- How to formally define "last-iterate algorithm"?
- We consider any "degree-k 1-SCLI algorithm" $\mathcal{A}$ *[Azizian et al., '19]*; includes, e.g., k-step extrapolation methods (includes EG for $k = 2$):

$$x_{t+1/k} = x_t - \eta_1 \nabla_x f(x_t, y_t)$$
$$x_{t+2/k} = x_t - \eta_2 \nabla_x f(x_{t+1/k}, y_{t+1/k})$$
$$\vdots$$
$$x_{t+1} = x_t - \eta_k \nabla_x f(x_{t+(k-1)/k}, y_{t+(k-1)/k})$$

$y_{t+1/k}, \dots, y_{t+(k-1)/k}, y_{t+1}$ defined similarly

Theorem (this paper): for any $\mathcal{A}$ as above, there is some smooth convex-concave f so that iterates (x_t, y_t) of $\mathcal{A}$ satisfy:

$$Gap(x_t, y_t) \geq \Omega\left(\frac{1}{k\sqrt{t}}\right)$$

Dependence on k is tight.

Proof of upper bound on EG convergence rate

Theorem: $\mathrm{Gap}(x_t, y_t) \leq O(\frac{D^2}{\eta\sqrt{t}})$

- Define $G(x,y) = \begin{pmatrix} \nabla_x f(x,y) \\ -\nabla_y f(x,y) \end{pmatrix}$.
- Since f is convex-concave, suffices to show $||G(x_t, y_t)|| \leq O(1/\sqrt{t})$.
- Best-iterate convergence: $\exists t^* \leq t$ so that $||G(x_{t^*}, y_{t^*})|| \leq O(1/\sqrt{t})$.

Lemma 1 (main lemma): For all t,

$$||G(x_{t+1}, y_{t+1})|| \leq \left(1 + O(||G(x_t, y_t)||^2)\right) \cdot ||G(x_t, y_t)||$$

- Using lemma & induction:

$$||G(x_t, y_t)|| \leq \underbrace{\left(1 + O(1/t)\right)^{t-t^*}}_{O(1)} \cdot ||G(x_{t^*}, y_{t^*})|| \leq O(1/\sqrt{t})$$

Proof of main lemma

Lemma 1 (main lemma): For all t,

$$||G(x_{t+1}, y_{t+1})|| \leq \left(1 + O(||G(x_t, y_t)||^2)\right) \cdot ||G(x_t, y_t)||$$

Taylor's theorem & Lipschitzness of Hessian of f:

$$G(x_{t+1}, y_{t+1}) = (I - A + AB) \cdot G(x_t, y_t)$$

for matrices $A, B \approx \eta \cdot \partial G(x_t, y_t)$ with

$$||A - B||_\sigma \leq O(||G(x_t, y_t)||)$$

Lemma 2: In above setting,

$$||I - A + AB||_\sigma \leq 1 + O(||A - B||_\sigma^2) \leq 1 + O(||G(x_t, y_t)||^2)$$

Conclusion / Future work

- **This paper**: provable quadratic speedup for averaging iterates of EG in convex-concave case.
- **Future work**: nonconvex-nonconcave case:
  - Empirical studies have indicated averaging does help convergence in GANs *[Yazici et al., ICLR'19]*.
  - Averaging is used in large-scale GANs [Brock et al., ICLR'19].
- **Question: why does averaging help in nonconvex case?**
  - Hard to show that averaging doesn't hurt (i.e., can't use Jensen's inequality).

Our paper: https://arxiv.org/abs/2002.00057

Thank you for listening!