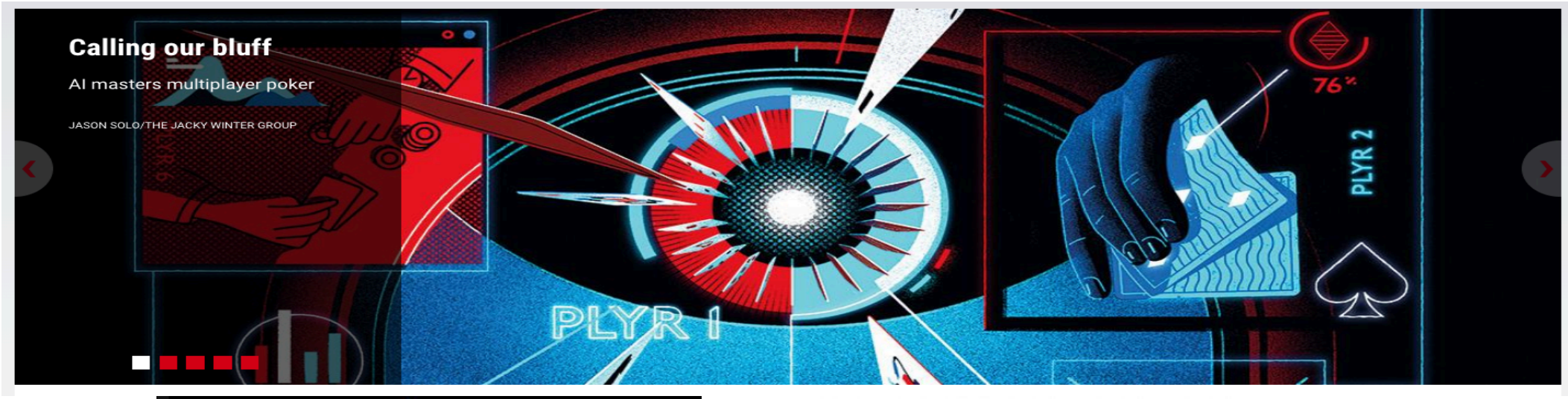
Near-Independent Policy Gradient Methods for Competitive Reinforcement Learning

Constantinos Daskalakis Dylan Foster Noah Golowich

Massachusetts Institute of Technology

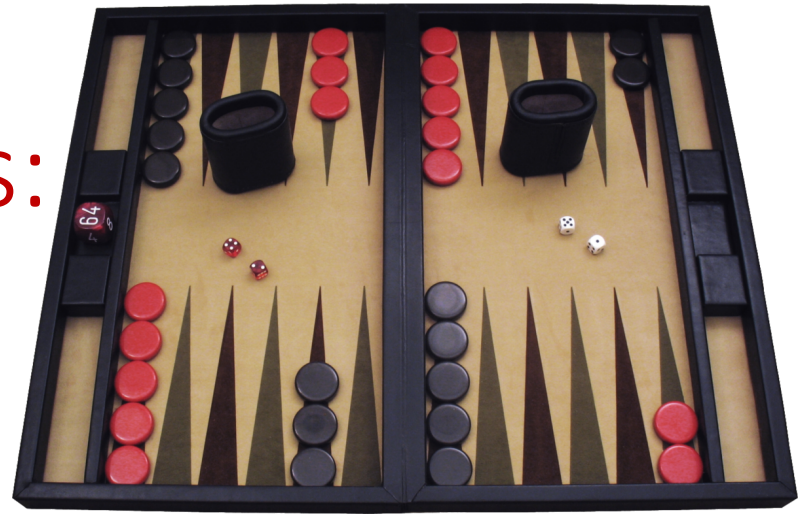
Multi-agent reinforcement learning



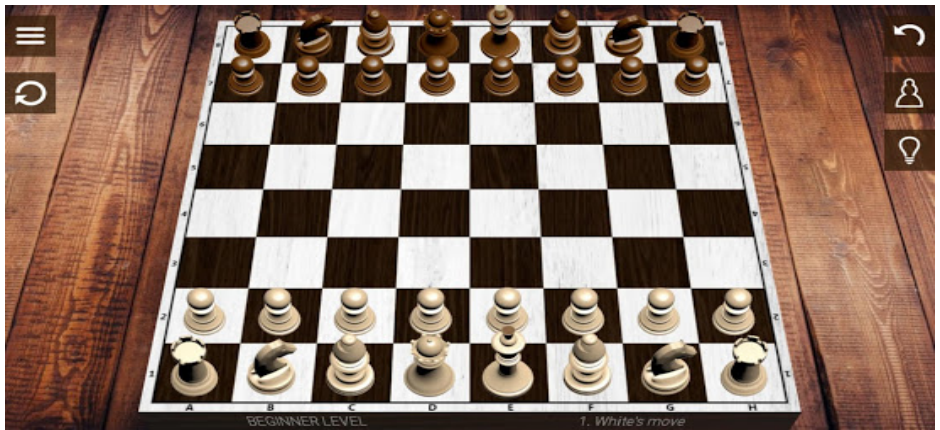
Two-player zero-sum Markov games

- 2-player zero-sum Markov game: [Shapley, '53; Littman, '94]
  - $G = (S, A, B, P, r, \zeta, \rho)$.
  - S is a finite set of **states**.
  - A, B are finite sets of **actions** for each player.
  - $\rho \in \Delta(S)$ is **initial state distribution** at time $t = 0$.
  - $r(s, a, b)$ is **reward** function (A player wants to minimize reward, B player wants to maximize).
  - $P(s'|s, a, b)$ is the **transition probability matrix**: given in state s and players take actions a, b, gives distribution for next state s'.
  - **Stopping probability**: $\boldsymbol{\zeta_{\{s,a,b\}}} := 1 - \sum_{s' \in S} P(s'|s, a, b) > 0$.
  - Assume $\boldsymbol{\zeta} := \min_{s,a,b} \boldsymbol{\zeta_{\{s,a,b\}}} > 0$; i.e., game stops after exp. $\leq 1/\zeta$ steps.

(Fun) examples of Markov games:



- Chess, Go *(not stochastic, $\zeta = 0$)*
- Backgammon *($\zeta = 0$)*
- Above are **turn-based**, i.e., at each state only 1 agent can take actions that influence next state's distribution.
- *Example* of non-turn-based Markov game (still $\zeta = 0$):
  - Chess, except both players move simultaneously at each step:
  - If two players move a piece to the same square, choose one randomly to remove from the board.



Policies, value function

- Fix **policies** $\pi_1 : S \to \Delta(A), \pi_2 : S \to \Delta(B)$.
- Induced distribution of **trajectories** $(s_t, a_t, b_t, r_t)_{0 \le t \le T}$, where:
  - $s_0 \sim \rho$
  - $a_t \sim \pi_1(\cdot|s_t), b_t \sim \pi_2(\cdot|s_t)$
  - $r_t = r(s_t, a_t, b_t)$
  - $T \geq 0$ is last time step before game steps (T is random).
- **Value function**:

$$V_\rho(\pi_1, \pi_2) := E_{\pi_1, \pi_2, \rho}\left[\sum_{0 \le t \le T} r(s_t, a_t, b_t) | \, s_0 \sim \rho\right]$$

Shapley's min-max theorem

- **Theorem (Shapley, '53):** There exists a Nash equilibrium in any Markov game, i.e., a policy tuple (π_1^*, π_2^*) so that:

$$V_{\rho(\pi_1^*,\pi_2)} \leq V_\rho(\pi_1^*, \pi_2^*) \leq V_\rho(\pi_1, \pi_2^*) \qquad \forall \pi_1, \pi_2$$

- In particular:

$$V_\rho^* := \min_{\pi_1} \max_{\pi_2} V_\rho(\pi_1, \pi_2) = \max_{\pi_1} \min_{\pi_2} V_\rho(\pi_1, \pi_2)$$

Problem: given ability to play policies π_1, π_2 to sample trajectories, find $\hat{\pi}_1, \hat{\pi}_2$ so that

$$\max_{\pi_2} V_\rho(\hat{\pi}_1, \pi_2) - V_\rho(\pi_1^*, \pi_2^*) \leq \epsilon$$

$$V_\rho(\pi_1^*, \pi_1^*) - \min_{\pi_1} V_\rho(\pi_1, \hat{\pi}_2) \leq \epsilon$$

Prior work: centralized/coordinated protocols

- Most previous works take algos from single-agent setting and ``replace the maximum (of reward) with computation of Nash'':
  - VI-ULCB [Bai & Jin, '20]: take UCBVI [Azar et al., '17], compute upper + lower estimates of $Q(s,a,b)$, find Nash eq. of each game $Q(s,\cdot,\cdot)$.
  - VI-Explore [Bai & Jin, '20]: exploration phase to build a model of the game, then value iteration on the empirical model.
  - OMNI-VI [Qie et al., '20]: similar to LSVI-UCB [Jin et al., '19], except find a CCE of a game for each s as opposed to max of (optimistic) Q-values.
  - Nash Q-learning [Bai et al., '20]: similar to proof of Q-learning [Jin et al., '18], except replace max over Q values with computation of a CCE.
  - *One exception: Nash V-learning [Bai et al., '20]*

Our goal: independent learning

i.e., Min player observes $\left\{\left(s_t^{(i)}, a_t^{(i)}, r_t^{(i)}\right)\right\}_{0\le t\le T}$; similarly for Max player.

- **Independent protocol**:
  - **Independent game oracle**: Each episode i, players propose policies $\pi_1^{(i)}: S \to \Delta(A), \pi_2^{(i)}: S \to \Delta(B)$, executed in game G, players observe states, their own actions, rewards.
  - **Limited private storage**: can only store policy parameter vector, $O(1)$ past trajectories.
- Not meant to be a formal definition (though our protocols clearly satisfy above requirements).

Our protocol: independent policy gradient method

1. Players choose **policy parametrizations** $x \mapsto \pi_x, y \mapsto \pi_y$ ($x \in X, y \in Y$ are parameter vectors).

We use ε-greedy (direct) parametrization: e.g., $X = \Delta(A)^S$, $\pi_x(a\,|s) := (1 - \varepsilon_x)x_{s,a} + \frac{\varepsilon_x}{|A|}$

2. Treat finding equilibrium as optimization problem: i.e., do **stochastic gradient descent-ascent (SGDA)**:

$$x^{(i+1)} \leftarrow \Pi_X\left(x^{(i)} - \eta_x g_x^{(i)}\right), \qquad y^{(i+1)} \leftarrow \Pi_Y(y^{(i)} + \eta_y g_y^{(i)})$$

Take $\eta_x \ll \eta_y$ (**2-timescale algorithm**)

Where $g_x^{(i)}, g_y^{(i)}$ are unbiased gradient estimators:

$$E\left[g_x^{(i)}\right] = \nabla_x V_\rho\left(\pi_{x^{(i)}}, \pi_{y^{(i)}}\right), \qquad E\left[g_y^{(i)}\right] = \nabla_y V_\rho(\pi_{x^{(i)}}, \pi_{y^{(i)}})$$

We use REINFORCE estimator: e.g., $g_x^{(i)} := \left(\sum_{0 \le t \le T} r_t^{(i)}\right) \cdot \sum_t \nabla_x \log \pi_x(a_t^{(i)} \mid s_t^{(i)})$,

(Recall: trajectory $\left(s_t^{(i)}, a_t^{(i)}, r_t^{(i)}\right)_{0 \le t \le T}$ observed by min player)

Main theorem: polynomial sample complexity

Theorem [DFG '20]: Let $\epsilon > 0$ be given. Suppose both players follow SGDA with learning rates $\eta_x \asymp \epsilon^{10.5}, \eta_y \asymp \epsilon^6$; then we have ``on-average convergence'' for min player:

$$E\left[\frac{1}{N} \cdot \sum_{1 \le i \le N} \max_{\pi_2} V_\rho\left(\pi_{x^{(i)}}, \pi_2\right)\right] - \min_{\pi_1} \max_{\pi_2} V_\rho(\pi_1, \pi_2) \le \epsilon \qquad (*)$$

for $N \le poly(\epsilon^{-1}, C_G, |S|, |A|, |B|, \zeta^{-1})$.

$$N = O\left(\frac{(|A| \vee |B|)^{10.75} |S|^{1.25} C_G^{17.5}}{\epsilon^{12.5} \zeta^{48.5}}\right)$$

- C_G is **distribution mismatch coefficient** (occurs in 1-player setting too).
- We *do not* get guarantee $(*)$ for max player $\pi_{y^{(i)}}$ when learning rates are $\eta_x \asymp \epsilon^{10.5}, \eta_y \asymp \epsilon^6$ (due to asymmetric nature).
  - *Indeed: in experiments, $(*)$ does not hold for max player.*
- Note: greedy parametrizations tuned to ϵ as well: $\varepsilon_x \asymp \epsilon, \varepsilon_y \asymp \epsilon^2$.

Distribution mismatch coefficient

- For policy π_1, let $\Pi_2^*(\pi_1)$ be the set of best responses for agent 2; similarly define $\Pi_1^*(\pi_2)$.
- **State visitation distribution:**

$$d_\rho^{\pi_1,\pi_2}(s) \propto \sum_{t\geq 0} \Pr_{\pi_1,\pi_2}(s_t = s \,|s_0 \sim \rho)$$

- $C_G := \max\{\max_{\pi_2} \min_{\pi_1 \in \Pi_1^*(\pi_2)} \left|\frac{d_\rho^{\pi_1,\pi_2}}{\rho}\right|_\infty, \max_{\pi_1} \min_{\pi_2 \in \Pi_2^*(\pi_1)} \left|\frac{d_\rho^{\pi_1,\pi_2}}{\rho}\right|_\infty\}$.
- Compare with typical C_M for a (1-agent) MDP [AKLM, '19]:
- $C_M := \left|\frac{d_\rho^{\pi^*}}{\rho}\right|_\infty$

Proof overview of main theorem

1. Show that $V_\rho(\pi_x, \pi_y)$ satisfies a 2-sided gradient domination (GD) condition.
   - I.e., for all x, function $y \mapsto V_\rho(\pi_x, \pi_y)$ satisfies a GD condition; and:
   - For all y, function $x \mapsto V_\rho(\pi_x, \pi_y)$ satisfies a GD condition.
2. Show that 2-timescale SGDA converges for any objective $f(x, y)$ satisfying such a 2-sided GD condition.

Some prior work for item 2:

- [YKH, '20]: (deterministic) 2-timescale GDA converges under a 2-sided PL condition (much stronger, allows linear rates)
- [LJJ, '20]: 2-timescale SGDA converges if $f(x, y)$ is concave in y.

Last-iterate convergence: extragradient

- Korpelevich's **extragradient (EG)** method (i.e., **mirror-prox** *[Nemirovskii, '04]*):
- Given $f : X \times Y \to \mathbb{R}$; goal is to solve $\min_x \max_y f(x, y)$:

$$x^{(i+\frac{1}{2})} \leftarrow \Pi_X\left(x^{(i)} - \eta f\left(x^{(i)}, y^{(i)}\right)\right), \qquad y^{(i+\frac{1}{2})} \leftarrow \Pi_Y(y^{(i)} + \eta f\left(x^{(i)}, y^{(i)}\right))$$

$$x^{(i+1)} \leftarrow \Pi_X\left(x^{(i)} - \eta f\left(x^{(i+\frac{1}{2})}, y^{(i+\frac{1}{2})}\right)\right), y^{(i+1)} \leftarrow \Pi_Y(y^{(i)} + \eta f\left(x^{(i+\frac{1}{2})}, y^{(i+\frac{1}{2})}\right))$$

Theorem [Korpelevich, '76; FP, '01]: If $f(x, y)$ satisfies the **MVI property**, then the iterates $(x^{(i)}, y^{(i)})$ of EG converge to a Nash equilibrium(x^*, y^*).

Definition: MVI property means that for all Nash equilibria (x^*, y^*) of f,

$$\langle \nabla_x f(x, y), x - x^* \rangle + \langle -\nabla_y f(x, y), y - y^* \rangle \geq 0 \quad \forall (x, y) \in X \times Y$$

Back to Markov games…

- Focus on **single-state** case with direct parametrization:

$$x = \pi_x \in X := \Delta(A), \qquad y = \pi_y \in Y := \Delta(B)$$

- Game value is that of a **ratio game**: *[von Neumann, '45]*

$$V(\pi_x, \pi_y) = \frac{E_{a\sim\pi_x, b\sim\pi_y}[r(a,b)]}{E_{a\sim\pi_x, b\sim\pi_y}[1 - P(s_0|s_0,a,b)]}$$

~~$\frac{\langle x, Ry\rangle}{\langle x, Sy\rangle}$, $R_{a,b} = r(a,b)$, $S_{a,b} \geq \zeta > 0$~~

~~$\max_{\pi_x}\min_{\pi_y}$ [game stops]~~

Proposition [DFG, '20]: There exists an objective $f(x,y) = \frac{\langle x,Ry\rangle}{\langle x,Sy\rangle}$, for $x, y \in \Delta(\{1,2\})$, which does not satisfy the MVI property.

Conjecture [DFG, '20]: For any $R \in [-1,1]^{n\times m}, S \in [\zeta, 1]^{n\times m}$, EG applied to the function $f(x,y) = \frac{\langle x,Ry\rangle}{\langle x,Sy\rangle}$ $(x \in \Delta([n]), y \in \Delta([m]))$ converges to Nash eq.

Other open problems: better rates, exploration

- Get better rates (ours are quite bad)
  - Incorporate optimism (would also get rid of dependence on C_G), e.g., [ESRM, ‘20] for single agent setting.
  - Issue with above: hard to make that approach independent.
- Other directions in [AKLM, ‘20]: natural policy gradient, linear function approximation.
- Multi-agent/non-zero-sum games.

Thank you!