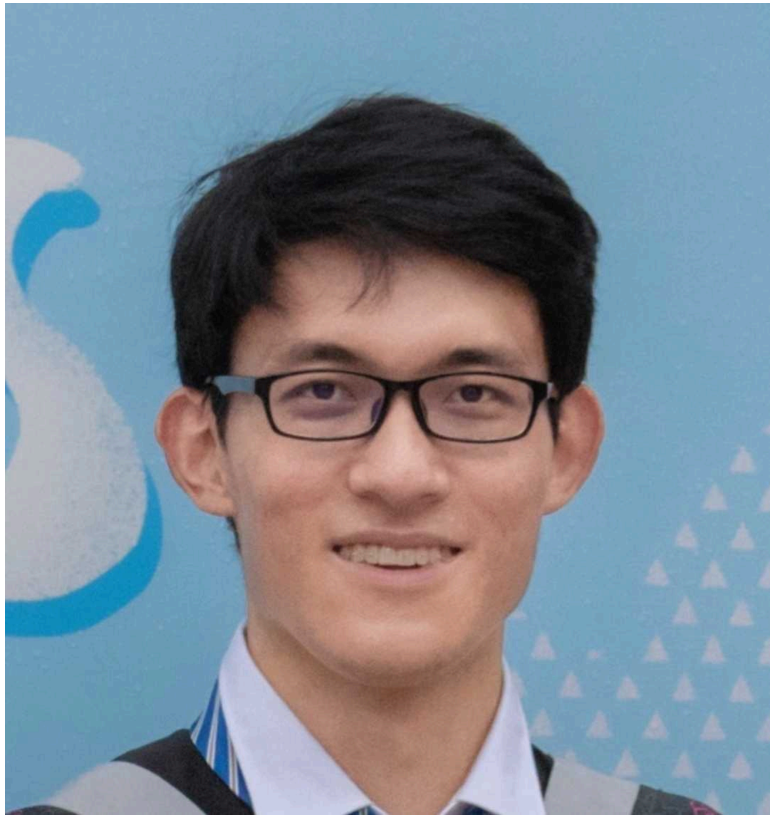
When is an Offline Two-Player Zero-Sum Markov Game Solvable?

Simon S. Du

05-04-2022

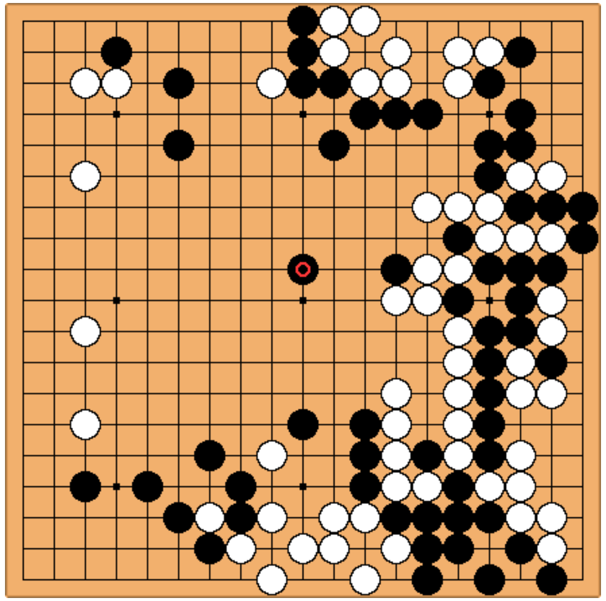
Simons Institute

Acknowledgement



Qiwen Cui
University of Washington

Two-Player Zero-Sum Markov Games



- Two players compete against each other. Each has a strategy.
- **Goal:** find a Nash Equilibrium
- Nash Equilibrium: a pair of strategies that no player can do better by unilaterally changing the policy.
- Applications: poker, Go, chess, computer games, investment, …..

Offline Reinforcement Learning

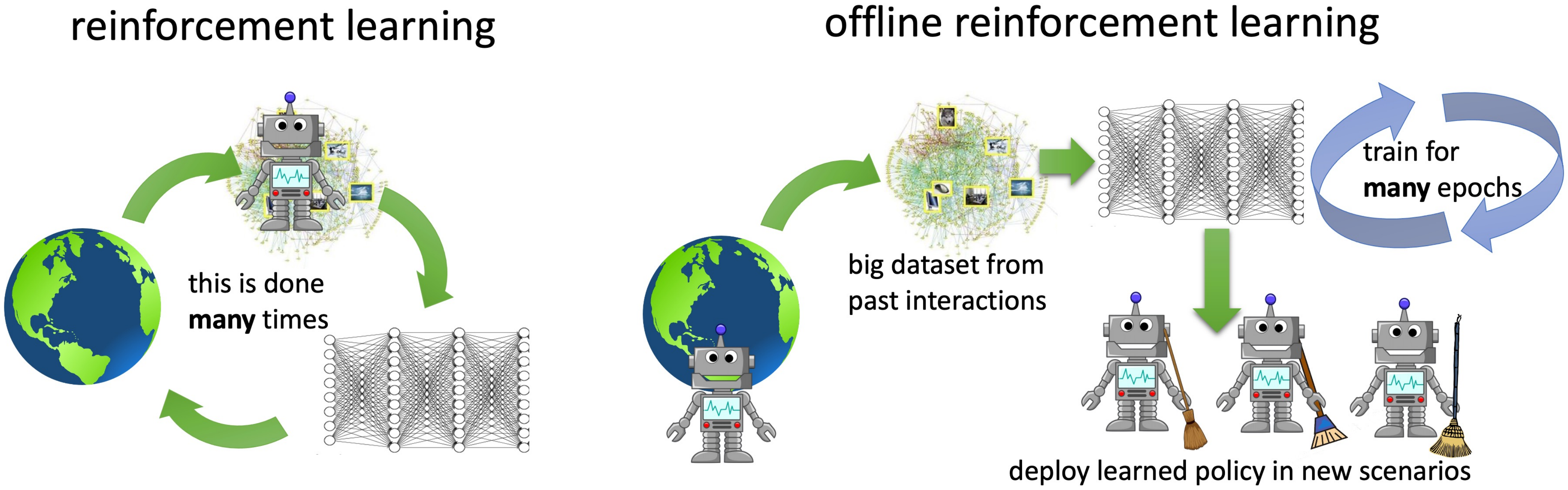
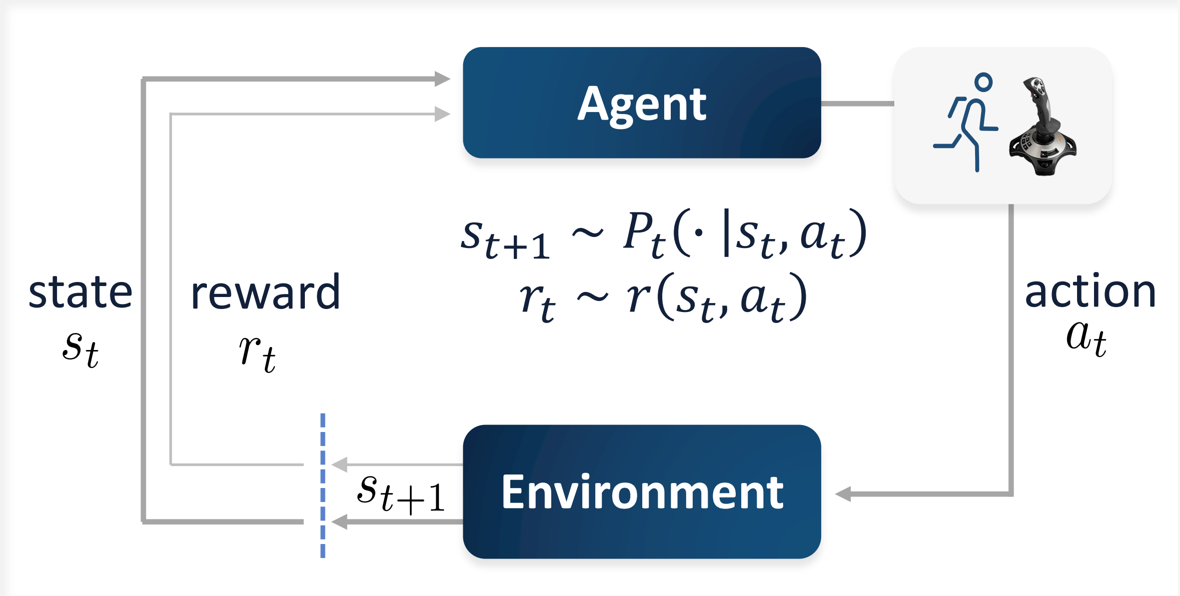


Figure credit: Berkeley AI Research Blog

- Lots of available **offline data** from prior experience. Fresh samples are expensive
- **This Talk:** When can we learn a Nash Equilibrium in offline two-player zero-sum Markov games?

Single-Agent Reinforcement Learning



Agent

$s_{t+1} \sim P_t(\cdot\,|s_t, a_t)$

$r_t \sim r(s_t, a_t)$

state s_t

reward r_t

action a_t

s_{t+1}

Environment

Repeat **H** times

H: planning horizon / Episode length

A policy π :

π: States(S) → Actions (A), $a = \pi(s)$

Goal: maximize value function

$$V^{\pi}(s_1) = \mathbb{E}[r_1 + r_2 + \cdots r_H]$$

Near-optimal policy:

$$V^*(s_1) - V^{\pi}(s_1) \leq \epsilon$$

$V^* = V^{\pi^*}$: value function of opt policy

Tabular Markov Decision Process

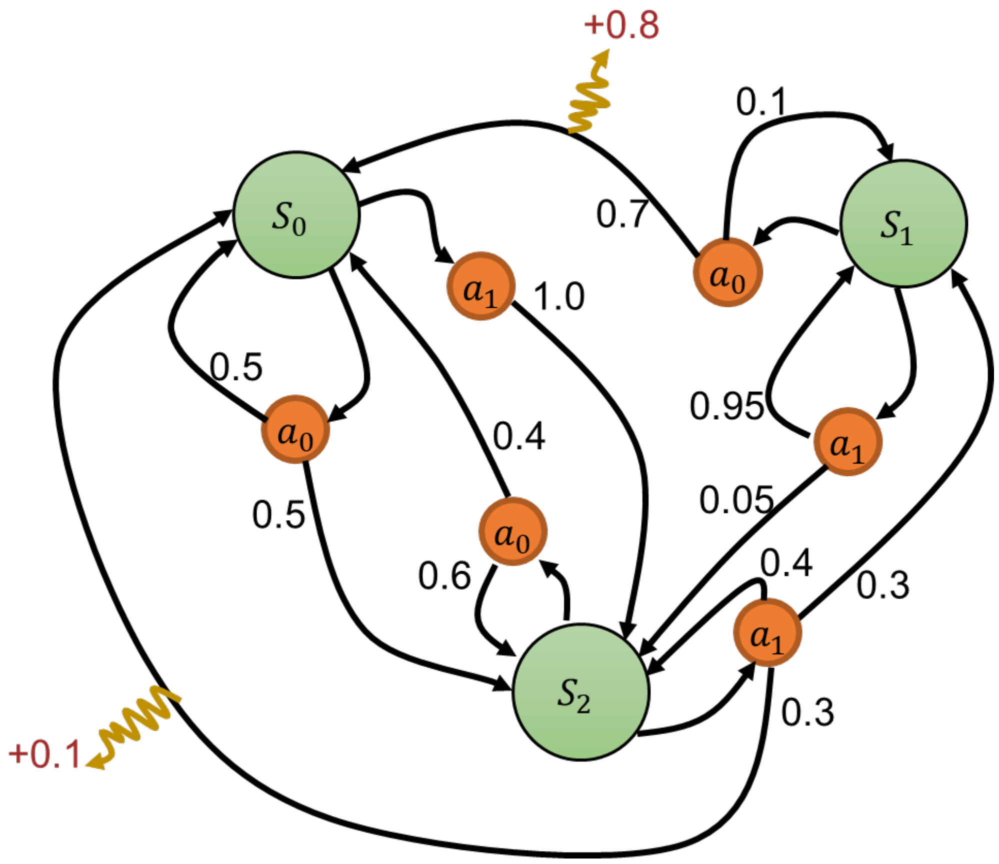
Assumptions:

1. # of States $\boldsymbol{S} < \infty$
2. # of actions $\mathbf{A} < \infty$
3. Bounded rewards:

   $0 \leq r_h \leq 1, h = 1, \dots, H$

Sample complexity depends on

$(S, A, H, 1/\epsilon)$



Offline Single-Agent Reinforcement Learning

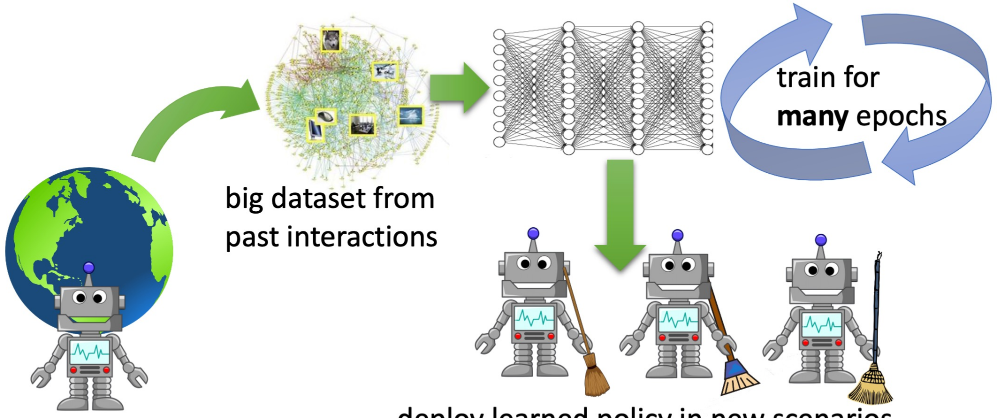
Offline Data: n (state, action, reward, next state) tuples:

$$D = \{(s_h^i, a_h^i, \mathrm{r}_h^i, s_{h+1}^i)\}_{h\in[H]}^{i\in[n]} \overset{i.i.d.}{\sim} d^\rho$$

- ρ is the data-collection / behavior policy
- $d_h^\rho(s,a)$ is the state-action distribution induced by ρ and transition P.
- Goal: learn a policy π from D such that

$$V^*(s_1) - V^\pi(s_1) \le \epsilon$$

offline reinforcement learning



deploy learned policy in new scenarios

Q: Under what conditions on d^ρ we can learn a near-optimal policy?

Dataset Coverage and Results

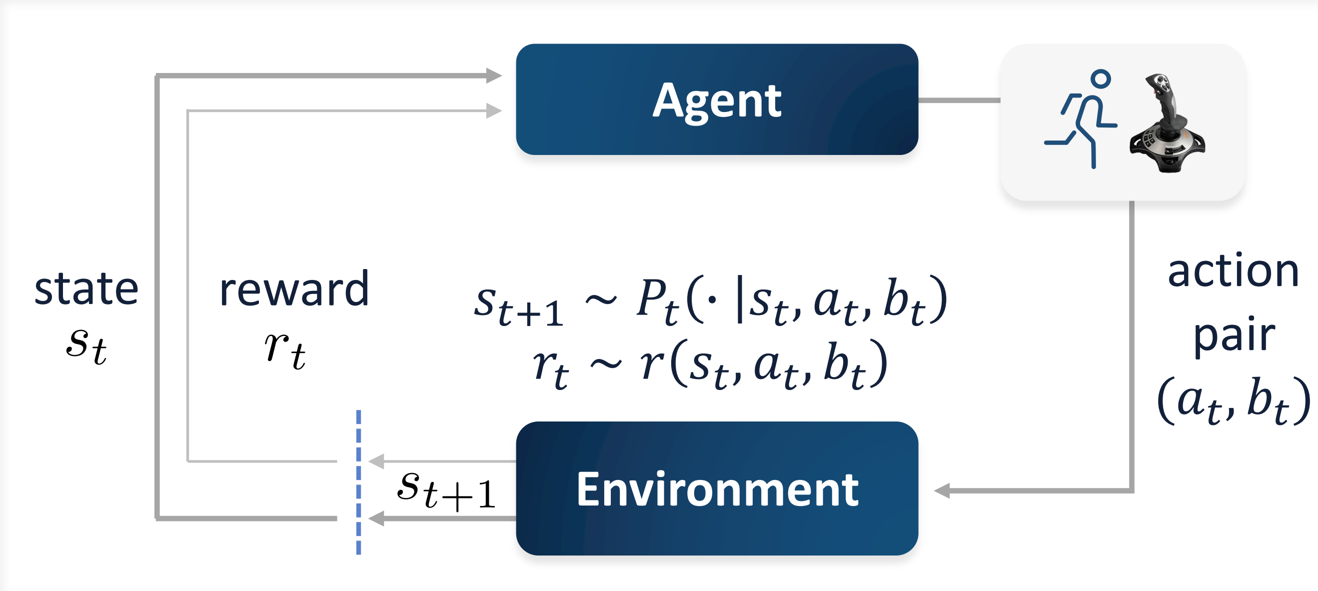
Single Policy Coverage Assumption

Necessary and Sufficient

- The behavior policy **only covers a single optimal policy**.
- There exists some constant $\mathbf{C}_{\text{single}}$ such that $\frac{d_h^{\pi^*}(s,a)}{d_h^{\rho}(s,a)} \leq \mathbf{C}_{\text{single}}$ for every (s, a) [LSAB19,JYW20].
- $1 \leq \mathbf{C}_{\text{single}} \leq \infty$
- Algorithmic idea: **Pessimism**. Penalize uncertain policies [JYW20,RZMJR21]. More later.
- Near-optimal bounds: $\widetilde{\Theta}(\frac{SH^3\mathbf{C}_{\text{single}}}{\epsilon^2})$ [XJWXB21].

Two-Player Zero-Sum Markov Games

Zero-Sum Markov Games



Agent

state s_t

reward r_t

$$s_{t+1} \sim P_t(\cdot \mid s_t, a_t, b_t)$$
$$r_t \sim r(s_t, a_t, b_t)$$

action pair (a_t, b_t)

s_{t+1}

Environment

Repeat H times, H: planning horizon

Max player $(a_1, a_2, \ldots, a_H)$: max $\mathbb{E}[r_1 + \cdots r_H]$

Min player $(b_1, b_2, \ldots, b_H)$: min $\mathbb{E}[r_1 + \cdots r_H]$

Zero-Sum Bandits

Agent

$r(a, b)$

action pair (a, b)

Environment

Special case of Markov games with H = 1 and a fixed state.

Only reward $r(a, b)$ matters.

Tabular Two-Player Zero-Sum Markov Games

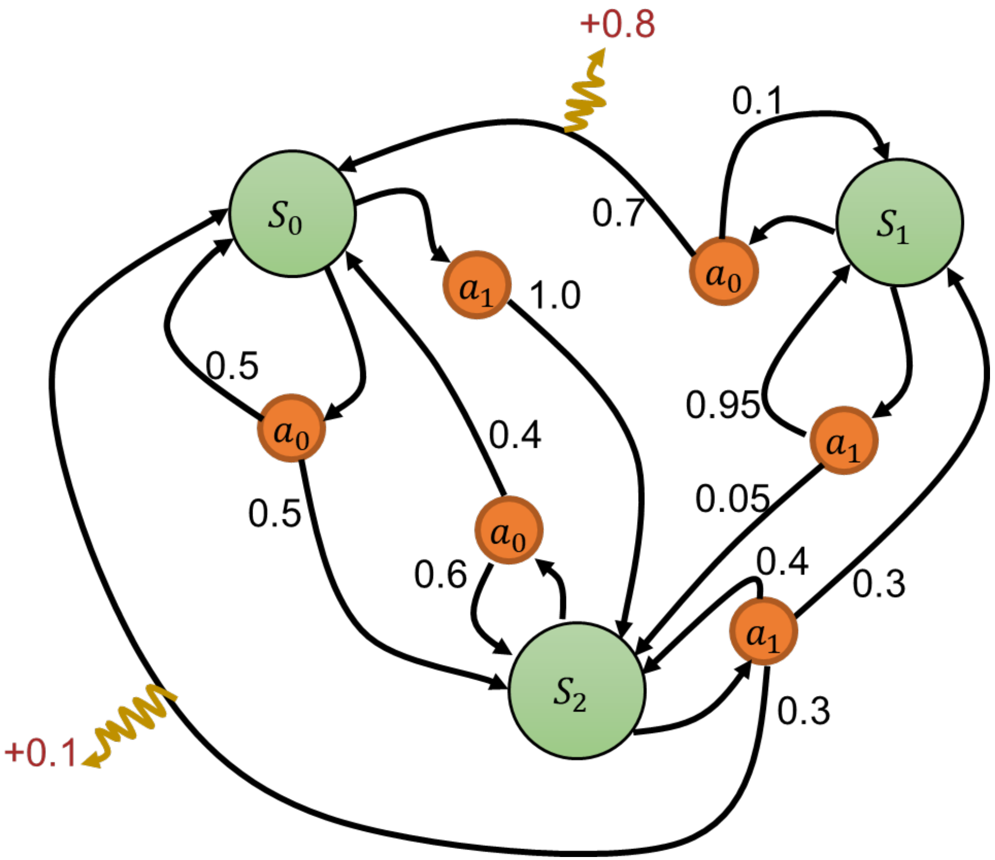
Assumptions:

1. # of States $\boldsymbol{S} < \infty$
2. Max player # of actions $\mathbf{A} < \infty$
3. Min player # of actions $\mathbf{B} < \infty$
4. Bounded rewards:

$$0 \leq r_h \leq 1, h = 1, \dots, H$$

Sample complexity depends on

$$(S, A, B, H, 1/\epsilon)$$



Value Function, Best Response and Duality Gap

- **Policy pair:** $(\boldsymbol{\mu}, \boldsymbol{\nu})$

  Max player policy $\boldsymbol{\mu}$ and min player policy $\boldsymbol{\nu}$. $\mu: \boldsymbol{S} \to \Delta(\boldsymbol{A}), \nu: \boldsymbol{S} \to \Delta(\boldsymbol{B})$.
- **Q-function and Value Function:**

$$Q_h^{\mu,\nu}(s,a,b) = \mathbb{E}\left[r_h + r_{h+1} + \cdots r_H \,\middle|\, s_h = s, a_h = a, b_h = b, \mu, \nu\right]$$
$$V_h^{\mu,\nu}(s) = \mathbb{E}\left[r_h + r_{h+1} + \cdots r_H \,\middle|\, s_h = s, \mu, \nu\right]$$

- **Best response** value for Max-player: Given μ, $V_h^{\mu,*}(s_h) = \min_{\nu} V_h^{\mu,\nu}(s_h)$
- **Best response** value for Min-player: Given ν, $V_h^{*,\nu}(s_h) = \max_{\mu} V_h^{\mu,\nu}(s_h)$
- **Nash Equilibrium** $(\boldsymbol{\mu}^*, \boldsymbol{\nu}^*)$: $V_h^{\mu^*,\nu^*}(s_h) = V_h^{\mu^*,*}(s_h) = V_h^{*,\nu^*}(s_h)$ [Shapley, 53].
- **Duality gap:** $\mathrm{Gap}(\mu, \nu) = V_1^{*,\nu}(s_1) - V_1^{\mu,*}(s_1)$

Goal: find $(\boldsymbol{\mu}, \boldsymbol{\nu})$ such that $\mathrm{Gap}(\mu, \nu) \leq \epsilon$

Offline Two-Player Zero-Sum Markov Game

Offline Data: *n* (state, action, reward, next state) tuples:

$$D = \{(s_h^i, a_h^i, b_h^i, \mathrm{r}_h^i, s_{h+1}^i)\}_{h\in[H]}^{i\in[n]} \overset{i.i.d.}{\sim} d^\rho$$

- ρ: data-collection /behavior policy pair
- $d_h^\rho(s,a,b)$ is the state-action distribution induced by ρ and transition P.
- Goal: learn a policy pair (μ, ν) from D:

$$\mathrm{Gap}(\mu,\nu) \le \epsilon$$

Q: Under what conditions on d^ρ we can learn a near Nash Equilibrium?

What about single policy-pair coverage?

$$\frac{d_h^{(\mu^*,\nu^*)}(s,a,b)}{d_h^\rho(s,a,b)} \le \mathbf{C}_{\text{single}}$$

😭 NO

Counter Example for Single Strategy Coverage

Min Player

Max Player

	b_1	b_2
a_1	**0.5**	1
a_2	0	0.5

Game 1

Min Player

Max Player

	b_1	b_2
a_1	0.5	0
a_2	1	**0.5**

Game 2

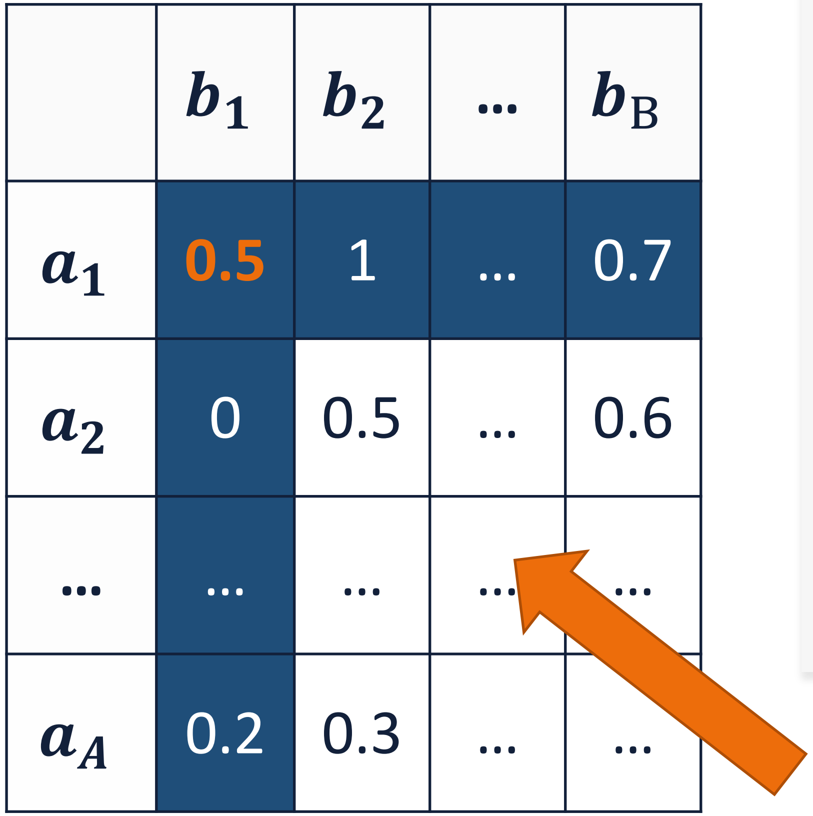
Need to cover (a_1, b_2) , (a_2, b_1)

- NE for Game 1: $(\boldsymbol{a_1}, \boldsymbol{b_1})$, NE for Game 2: $(\boldsymbol{a_2}, \boldsymbol{b_2})$
- Covers $(\boldsymbol{a_1}, \boldsymbol{b_1})$ and $(\boldsymbol{a_2}, \boldsymbol{b_2})$ with $d^\rho(a_1, b_1)$= $d^\rho(a_2, b_2) = 0.5 \Rightarrow \mathbf{C}_{\text{single}} = \mathbf{2}$.
- We cannot differentiate Game 1 or Game 2!

Unilateral Coverage Assumption

Min Player

Max Player



	$\boldsymbol{b}_1$	$\boldsymbol{b}_2$	…	$\boldsymbol{b}_\mathrm{B}$
$\boldsymbol{a}_1$	0.5	1	…	0.7
$\boldsymbol{a}_2$	0	0.5	…	0.6
…	…	…	…	…
$\boldsymbol{a}_A$	0.2	0.3	…	…

Nash Equilibrium: (a_1, b_1)

- For a Nash Equilibrium (μ^*, ν^*), the behavior policy covers **$(\boldsymbol{\mu}^*, \boldsymbol{\nu})$ and $(\boldsymbol{\mu}, \boldsymbol{\nu}^*)$ for all $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$.**
- There exists some constant $\mathbf{C}_{\text{unilateral}}$ such that $\frac{d_h^{\mu^*,\nu}(s,a,b)}{d_h^{\rho}(s,a,b)}, \frac{d_h^{\mu,\nu^*}(s,a,b)}{d_h^{\rho}(s,a,b)} \leq \mathbf{C}_{\text{unilateral}}$ for every (s, a, b) and (μ, ν).
- $A + B \leq \mathbf{C}_{\text{unilateral}} \leq \infty$

Covered or not doesn't matter.

A Weaker Assumption Than Unilateral Coverage?

Min Player

Max Player

	b_1	b_2
a_1	**0.25**	0.5
a_2	0	0.75

Game 1

(a_2, b_1) not covered

Min Player

Max Player

	b_1	b_2
a_1	0.25	0.5
a_2	1	**0.75**

Game 2

Not Sufficient!

- A slightly weaker assumption: there exists **at most one** deterministic μ or ν such that the behavior policy ρ **does not cover** (μ^*, ν) or (μ, ν^*).
- We cannot differentiate Game 1 or Game 2 without information of (a_2, b_1) .

Algorithm for Two-Player Zero-Sum Bandits

Min Player

Max Player

	$\boldsymbol{b_1}$	$\boldsymbol{b_2}$	…	$\boldsymbol{b}_{\mathrm{B}}$
$\boldsymbol{a_1}$	**[0.4, 0.6]**	[0.8, 1]	…	[0.7, 0.8]
$\boldsymbol{a_2}$	[0, 0.1]	[0.4, 0.7]	…	[0.6, 0.7]
…	…	…	…	…
$\boldsymbol{a_A}$	[0.1, 0.3]	[0.2, 0.4]	…	…

- Estimate $r(a,b) \in \left[\underline{r}(a,b), \overline{r}(s,a)\right] \forall\ (a,b)$.
- Computer NE $\left(\underline{\mu}, \underline{\nu}\right)$ for $\underline{r}(\cdot,\cdot)$.
- Computer NE $(\overline{\mu}, \overline{\nu})$ for $\overline{r}(\cdot,\cdot)$.
- Output $(\underline{\mu}, \overline{\nu})$.

Pessimism

Result for Two-Player Zero-Sum Bandits

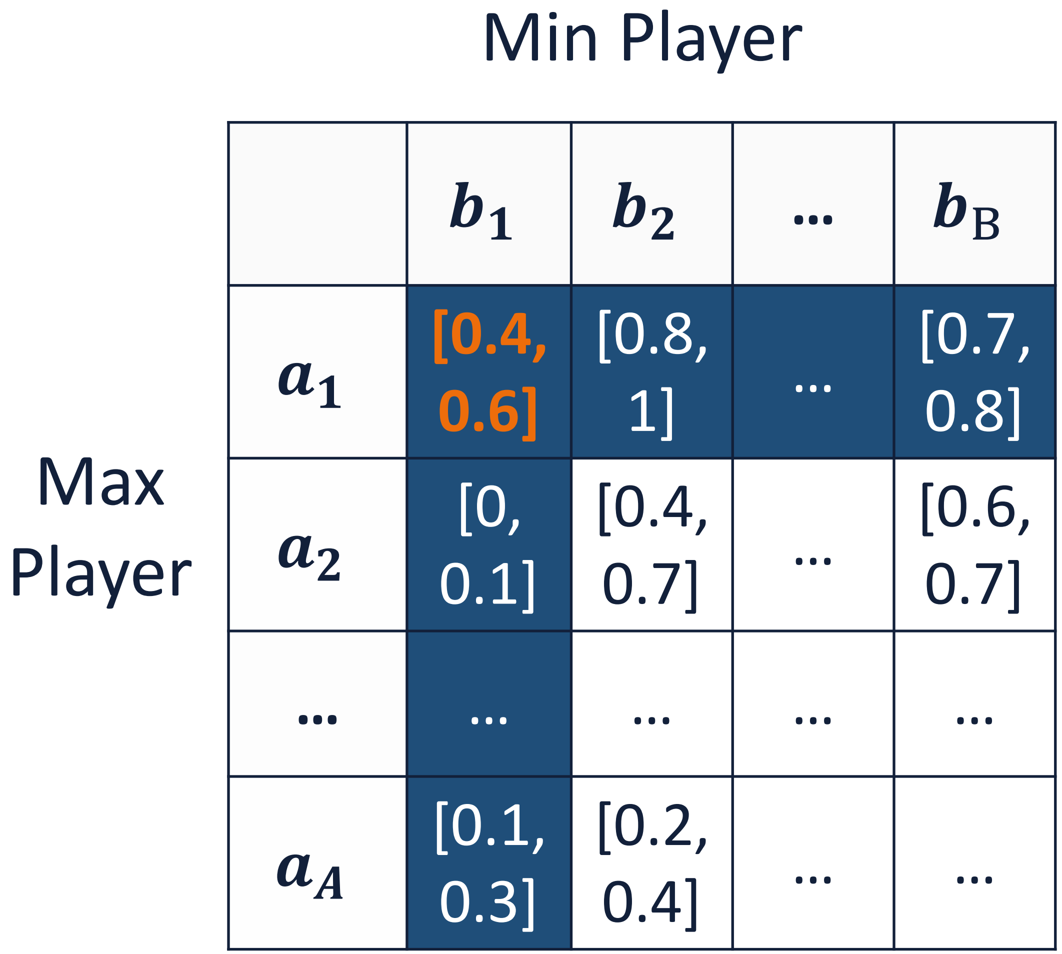
Theorem

- Sample complexity with unilateral coverage: $\tilde{O}(\frac{AB\mathbf{C}_{\mathrm{unilateral}}}{\epsilon^2})$
- Sample complexity with uniform coverage: $\tilde{O}(\frac{\mathbf{C}_{\mathrm{unif}}}{\epsilon^2})$
- Sample complexity for turn-based game with unilateral coverage: $\tilde{O}(\frac{\mathbf{C}_{\mathrm{unilateral}}}{\epsilon^2})$

- **Unilateral assumption is sufficient.**
- Lower bounds (from single-agent bandits)
  - Sample complexity with unilateral coverage: $\Omega(\frac{\mathbf{C}_{\mathrm{unilateral}}}{\epsilon^2})$
  - Sample complexity with uniform coverage: $\Omega(\frac{\mathbf{C}_{\mathrm{unif}}}{\epsilon^2})$
  - Sample complexity for turn-based game with unilateral coverage: $\Omega(\frac{\mathbf{C}_{\mathrm{unilateral}}}{\epsilon^2})$

Match

Algorithm for Markov Games



Min Player

Max Player

	b_1	b_2	…	b_B
a_1	**[0.4, 0.6]**	[0.8, 1]	…	[0.7, 0.8]
a_2	[0, 0.1]	[0.4, 0.7]	…	[0.6, 0.7]
…	…	…	…	…
a_A	[0.1, 0.3]	[0.2, 0.4]	…	…

Confidence for one state $\boldsymbol{s}$ at one step $\boldsymbol{h}$

- Estimate transition and reward using the dataset: $\widehat{P_h}(s'|s,a,b), \hat{r}(s,a,b)$
- Set $\underline{V}_{H+1}(s) = \overline{V}_{H+1}(s) \leftarrow 0, \forall s.$
- For h= H,H-1, ….,1:
  - $\underline{Q_h}(s,a,b) \leftarrow \hat{r}(s,a,b)$ $+\langle \widehat{P_h}(\cdot\,|s,a,b), \underline{V}_{h+1}(\cdot)\rangle - \mathbf{bonus}_{\boldsymbol{h}}\,(\boldsymbol{s},\boldsymbol{a},\boldsymbol{b})$
  - Computer NE $\left(\underline{\mu_h}, \underline{\nu_h}\right)$ for $\underline{Q}_h(\cdot,\cdot,\cdot)$.
  - $\underline{V}_h(s) \leftarrow \mathbb{E}_{(a,b)\sim\left(\underline{\mu_h},\underline{\nu_h}\right)}[\underline{Q}_h(s,a,b)]$
  - Similarly get $\overline{Q}_h$ with $+\mathbf{bonus}_{\boldsymbol{h}}$, $\overline{V}_h$, $(\overline{\mu_h}, \overline{\nu_h})$
- Output $(\underline{\mu}, \overline{\nu})$.

DP Step

Result for Two-Player Zero-Sum Markov Games

Theorem

If the bonus is constructed using a **reference function** and **Bernstein bound**:

- with unilateral coverage: $\tilde{O}(\frac{SABH^3\mathbf{C}_{\text{unilateral}}}{\epsilon^2})$
- with uniform coverage: $\tilde{O}(\frac{\mathrm{SH}^3\mathbf{C}_{\text{unif}}}{\epsilon^2})$
- for turn-based game with unilateral coverage: $\tilde{O}(\frac{SH^3\mathbf{C}_{\text{unilateral}}}{\epsilon^2})$

- **Unilateral assumption is sufficient for Markov games.**
- Lower bounds (from single-agent RL)
  - with unilateral coverage: $\Omega(\frac{\mathrm{SH}^3\mathbf{C}_{\text{unilateral}}}{\epsilon^2})$
  - with uniform coverage: $\Omega(\frac{\mathrm{SH}^3\mathbf{C}_{\text{unif}}}{\epsilon^2})$
  - for turn-based game with unilateral coverage: $\Omega(\frac{\mathrm{SH}^3\mathbf{C}_{\text{unilateral}}}{\epsilon^2})$

Match

Summary and Open Problems

First theoretical study on two-player zero-sum Markov games

- Single-policy coverage not sufficient: **separation** between single-agent and two-player
- **Unilateral coverage**: sufficient and cannot be weakened.
- Algorithms based on pessimism for both players
  - Polynomial bound for unilateral coverage.
  - Near-optimal bounds for (1) uniform coverage, (2) unilateral coverage + turn-based games.
- Concurrent work also studied linear MDP [ZXTWZWY22].

Upcoming Work!

Future Directions

- Improve bound under unilateral coverage (now $\boldsymbol{AB}$ factor gap).
- General sum in multi-agent games (online setting [ZMB21, JLWY21, …]).

Thank You

Analysis

- Confidence interval length: bonus$(a, b) \approx \sqrt{\frac{1}{n(a,b)}} \approx \sqrt{\frac{1}{nd^{\rho}(a,b)}}$.
- $r(\mu^*, \nu^*) \le r(\mu^*, \underline{\nu})$ (by the defn of ν^*)
- $r\left(\underline{\mu}, *\right) \ge \underline{r}\left(\underline{\mu}, *\right) \ge \underline{r}\left(\underline{\mu}, \underline{\nu}\right) \ge \underline{r}(\mu^*, \underline{\nu})$ (by the defns of of $\underline{r}$ and $\underline{\nu}$)
- $r(\mu^*, \nu^*) - r\left(\underline{\mu}, *\right) \le r\left(\mu^*, \underline{\nu}\right) - \underline{r}\left(\mu^*, \underline{\nu}\right) \le \mathbb{E}_{(a,b)\sim(\mu^*, \underline{\nu})}[\mathrm{bonus(a, b)}]$
- Similarly, $r(*, \overline{\nu}) - r(\mu^*, \nu^*) \le \mathbb{E}_{(a,b)\sim(\overline{\mu}, *)}[\mathrm{bonus(a, b)}]$
- Gap $\left(\underline{\mu}, \overline{\nu}\right) \le \mathbb{E}_{(a,b)\sim(\mu^*, \underline{\nu})}[\mathrm{bonus(a, b)}] + \mathbb{E}_{(a,b)\sim(\overline{\mu}, *)}[\mathrm{bonus(a, b)}]$
- Then use Cauchy-Schwartz