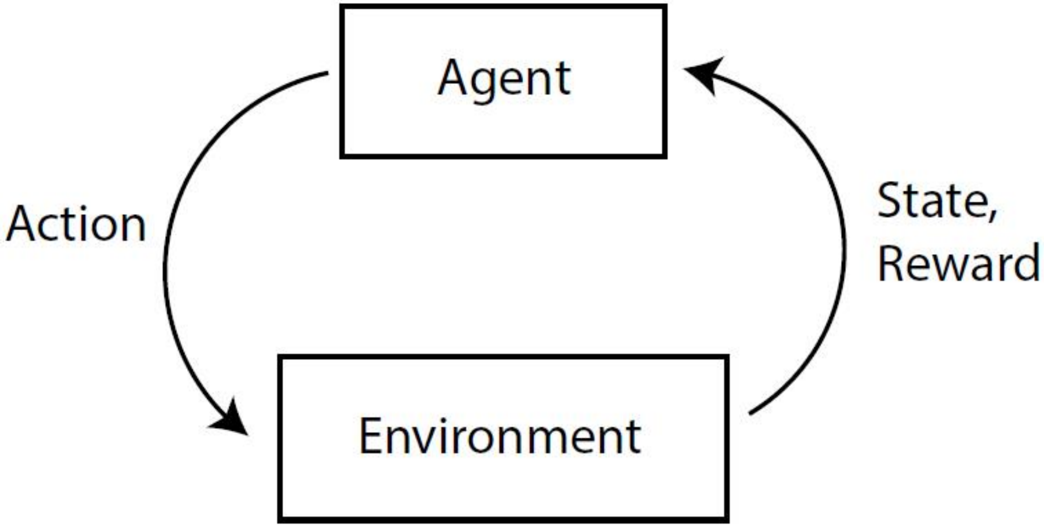
Learning Markov Games with Adversarial Opponents: Efficient Algorithms and Fundamental Limits

Qinghua Liu, **Yuanhao Wang** and Chi Jin

Princeton University

Sequential Decision Making and RL



- Goal: maximize rewards in a fixed environment through learning

RL in Games

- Environment defined by opponent behavior
- Opponent can play adaptively and adversarially
- Will focus on two-player zero-sum adversarial opponents

Markov Game (MG)

- Generalization of MDP for games
- State Space $\mathcal{S}$, $|\mathcal{S}| = S$
- Two-player zero-sum game.
- Action space $\mathcal{A} = \mathcal{A}_{\max} \times \mathcal{A}_{\min}$, $|\mathcal{A}| = A$
- Reward: $r_h(s, \boldsymbol{a}) \in [-1,1]$
- Transition probability: $P_h(\cdot\,|s, \boldsymbol{a}) \in \Delta_S$
- Horizon: H
- Episodic: $\{s_1, \boldsymbol{a}_1, r_1, s_2, \ldots, s_H, \boldsymbol{a}_H, r_H\}$, K episodes

Policies in Markov Game (MG)

Markov Policy	General (history dependent) policy
$\mu_h: S \to \Delta_{\mathcal{A}_{\max}}$	$\mu_h: (S \times \mathcal{A})^{h-1} \times S \to \Delta_{\mathcal{A}_{\max}}$

- Best response to changing series of Markov policy is general policy (in general)
- Max player policy $\mu \in \Phi$, min player policy $\nu \in \Psi$
- Algorithm picks μ to maximize $V_1^{\mu \times \nu}(s) = \mathbb{E}[\sum_{h' \geq 1} r_{h'} | s_1 = s]$
- $\{\mu^1, \nu^1\}, \{\mu^2, \nu^2\}, \cdots, \{\mu^K, \nu^K\}$

What is a reasonable performance metric?

What is a reasonable performance metric?

- For single player MDP: optimal value function
- Standard notion in online learning:

$$\text{Regret}_\Phi = \max_{\mu \in \Phi} \sum_{k=1}^{K} \left(V_1^{\mu \times \nu^k} - V_1^{\mu^k \times \nu^k} \right)(s_1)$$

- Defined against **best policy in hindsight**

What is a reasonable performance metric?

- For single player MDP: optimal value function
- Standard notion in online learning:

$$\text{Regret}_{\Phi} = \max_{\mu \in \Phi} \sum_{k=1}^{K} \left(V_1^{\mu \times \nu^k} - V_1^{\mu^k \times \nu^k} \right)(s_1)$$

- Defined against **best policy in hindsight**

Can we achieve no-regret in Markov games?

- Unclear even for 2-player zero-sum games

Standard setting: Only observes min-player's actions

Lower Bound I. Exists MG with $|S| = O(1), |\mathcal{A}| = O(1)$, such that when Φ is the set of all Markov policies, $|\Psi| = 1$ (fixed general policy) regret is $\Omega(\min\{K, 2^H\})$

Standard setting: Only observes min-player's actions

Lower Bound I. Exists MG with $|S| = O(1), |\mathcal{A}| = O(1)$, such that when Φ is the set of all Markov policies, $|\Psi| = 1$ (fixed general policy) regret is $\Omega(\min\{K, 2^H\})$

Lower Bound II. Exists MG with $|S| = O(H), |\mathcal{A}| = O(H)$, such that when Φ is the set of all Markov policies, $|\Psi| = H$ (Markov policies), regret is $\Omega(\min\{K, 2^H\})$

Standard setting: Only observes min-player's actions

Lower Bound I. Exists MG with $|S| = O(1), |\mathcal{A}| = O(1)$, such that when Φ is the set of all Markov policies, $|\Psi| = 1$ (fixed general policy) regret is $\Omega(\min\{K, 2^H\})$

Lower Bound II. Exists MG with $|S| = O(H), |\mathcal{A}| = O(H)$, such that when Φ is the set of all Markov policies, $|\Psi| = H$ (Markov policies), regret is $\Omega(\min\{K, 2^H\})$

Key idea: MG adversarial opponent is general enough to simulate POMDP (Lower bound I) or latent MDPs (Lower bound II)

Standard setting: Only observes min-player's actions

Lower Bound I. Exists MG with $|S| = O(1), |\mathcal{A}| = O(1)$, such that when Φ is the set of all Markov policies, $|\Psi| = 1$ (fixed general policy) regret is $\Omega(\min\{K, 2^H\})$

Lower Bound II. Exists MG with $|S| = O(H), |\mathcal{A}| = O(H)$, such that when Φ is the set of all Markov policies, $|\Psi| = H$ (Markov policies), regret is $\Omega(\min\{K, 2^H\})$

If Ψ = {Single Markov Policy}, becomes standard RL ($\sqrt{\text{poly}(S, A, H)K}$ regret)

If $H = 1$, contextual bandit algorithm solves the problem ($\sqrt{\text{poly}(S, A, H)K}$ regret)

Statistical hardness of MG stems from both adversarial opponents AND sequential nature

Opponent's policy contains much information its action doesn't reveal

Assume: Observes ν^k after episode k

May occur in self-play scenario

Algorithm I: Optimistic Policy EXP3

- Maintain model of MG transitions
- Optimistically evaluate values of all policies in Φ with model
- Run EXP3 on Φ using optimistic values

Assume: Observes min-player's policies

Upper Bound I. Regret of Algorithm I is $\tilde{O}\left(\sqrt{K(H^2 \log|\Phi| + S^2 AH)}\right)$

- If Φ = All Markov policies, Regret = $\tilde{O}(\sqrt{KS^2AH^4})$

Assume: Observes min-player's policies

Upper Bound I. Regret of Algorithm I is $\tilde{O}\left(\sqrt{K(H^2 \log|\Phi| + S^2AH)}\right)$

- If $\Phi =$ All Markov policies, Regret $= \tilde{O}(\sqrt{KS^2AH^4})$
- Independent of the size of Ψ

Assume: Observes min-player's policies

Upper Bound I. Regret of Algorithm I is $\tilde{O}\left(\sqrt{K(H^2\log|\Phi| + S^2AH)}\right)$

- If $\Phi =$ All Markov policies, Regret $= \tilde{O}(\sqrt{KS^2AH^4})$
- Independent of the size of Ψ
- Might be too large if Φ is all general policies, when $|\Phi| = \Omega\left(A^{S^H}\right)$
- Requires knowledge of Φ

Assume: Observes min-player's policies

To compete against general policies:

Algorithm II: Adaptive Optimistic Policy EXP3

Algorithm I +

- Update model sparsely (when visitation count doubles)
- Maintain candidate set of best responses of all possible mixtures of seen opponent policies
- Run EXP3 on candidate set. Reset whenever it's updated

Upper Bound II. Regret of Algorithm II is $\tilde{O}\left(\sqrt{K(S^2AH^4 + |\Psi|SAH^3 + |\Psi|^2H^2)}\right)$

- Compares against **best general policy in hindsight**
- Sublinear if $|\Psi| = o(\sqrt{K})$
- When opponent's strategy lacks diversity or changes infrequently

Drawbacks of Algorithms

- No guarantee when $|\Phi|$ is super-exponentially large AND $|\Psi|$ is exponentially large

Drawbacks of Algorithms

- No guarantee when $|\Phi|$ is super-exponentially large AND $|\Psi|$ is exponentially large

→ Unavoidable in general.

Lower Bound III. Exists MG with $|S| = O(1), |\mathcal{A}| = O(1)$, Φ is the set of all general policies, $|\Psi| = 2^H$, where regret is $\Omega(\min\{K, 2^H\})$ even if opponent reveals policy.

- Can't have polynomial regret in this regime (doubly-exponential $|\Phi|$, exponential $|\Psi|$)

Drawbacks of Algorithms

- No guarantee when $|\Phi|$ is super-exponentially large AND $|\Psi|$ is exponentially large

→ Unavoidable in general.

- Algorithm I & II have exponential runtime (linear in $|\Phi|$, exponential in $|\Psi|$)

Drawbacks of Algorithms

- No guarantee when $|\Phi|$ is super-exponentially large AND $|\Psi|$ is exponentially large

→ Unavoidable in general.

- Algorithm I & II have exponential runtime (linear in $|\Phi|$, exponential in $|\Psi|$)

→ Unavoidable in general

Computational Lower Bound. A polynomial time algorithm with poly$(S, A, H) \cdot K^{1-c}$ regret for a MG can be used to solve 3-SAT in polynomial time.

This holds even if the MG dynamics is known, the set Ψ is known, and policies are revealed.

Summary

Can we achieve low regret in Markov games?

Baseline Policy Φ	**Opponent's Policy Ψ**	**Only Action Revealed**	**Full Policy Revealed**
Markov Policies	General Policies	NO	$\tilde{O}\left(\sqrt{KS^2AH^4}\right)$
General Policies	Small Finite Class		$\tilde{O}\left(\sqrt{K\mathrm{poly}(\lvert\Psi\rvert, S, A, H)}\right)$
	General Policies		NO

For details, join us at **Poster #1111**