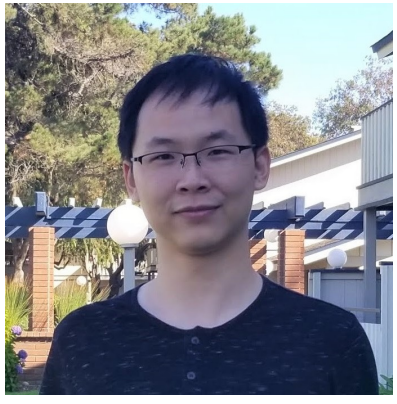
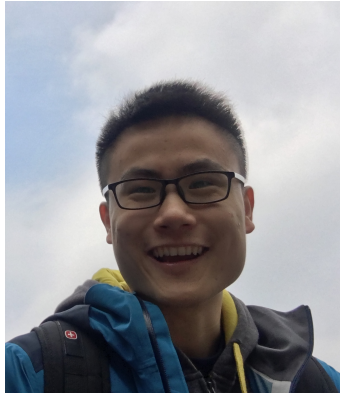
# The Power of Exploiter: Provable Multi-Agent RL in Large State Spaces

Chi Jin, Qinghua Liu, Tiancheng Yu

# Authors

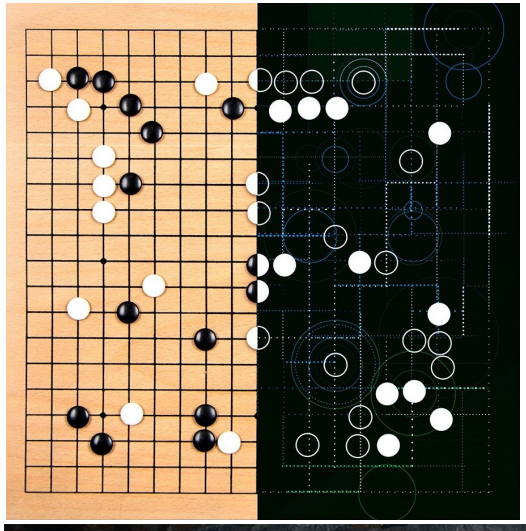
Chi Jin
Princeton University

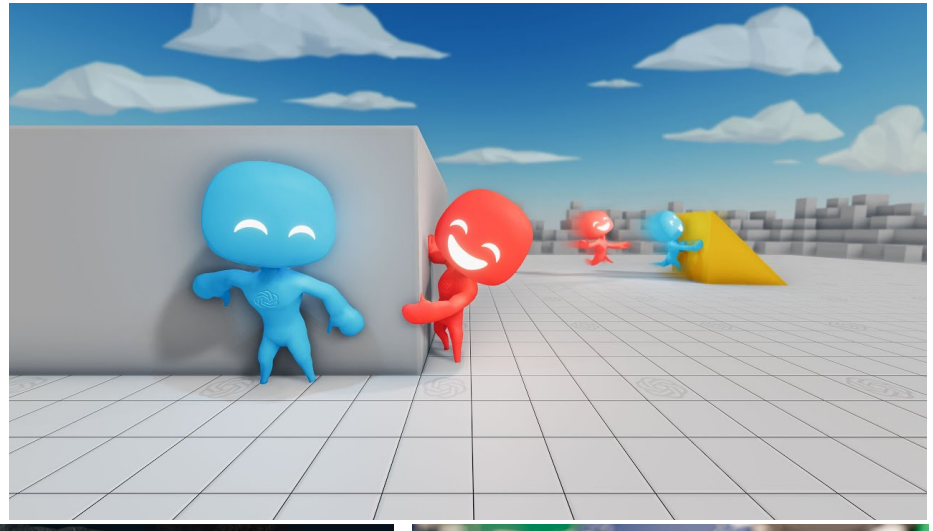


Qinghua Liu
Princeton University

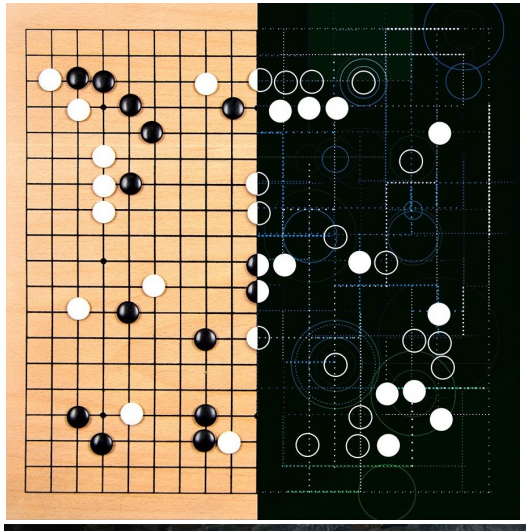
Tiancheng Yu
MIT

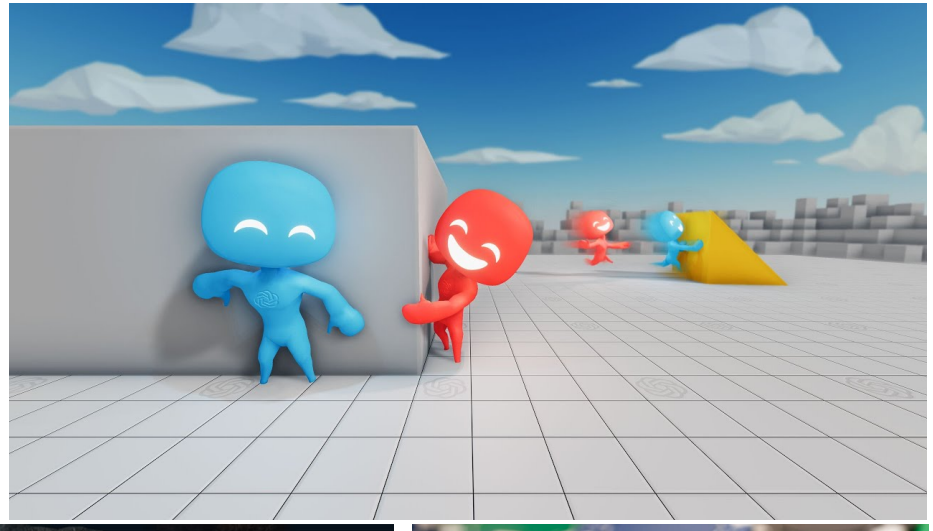
# Multiagent Reinforcement Learning





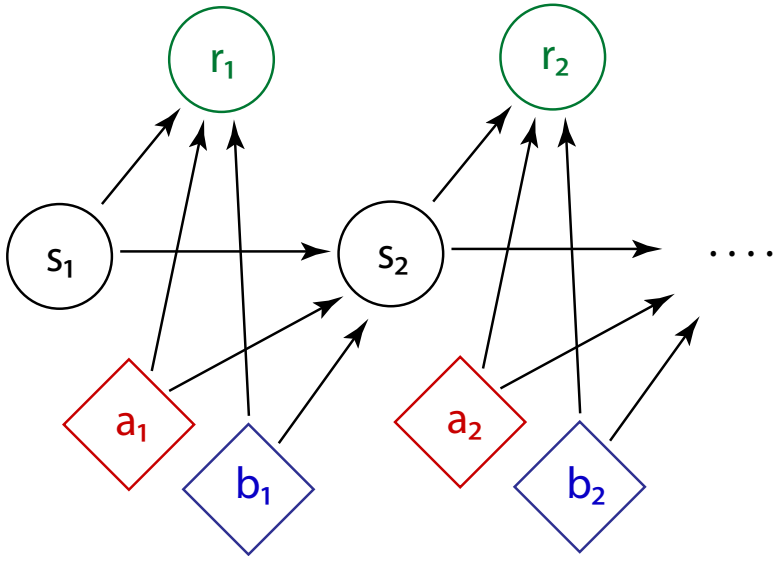

# Multiagent Reinforcement Learning





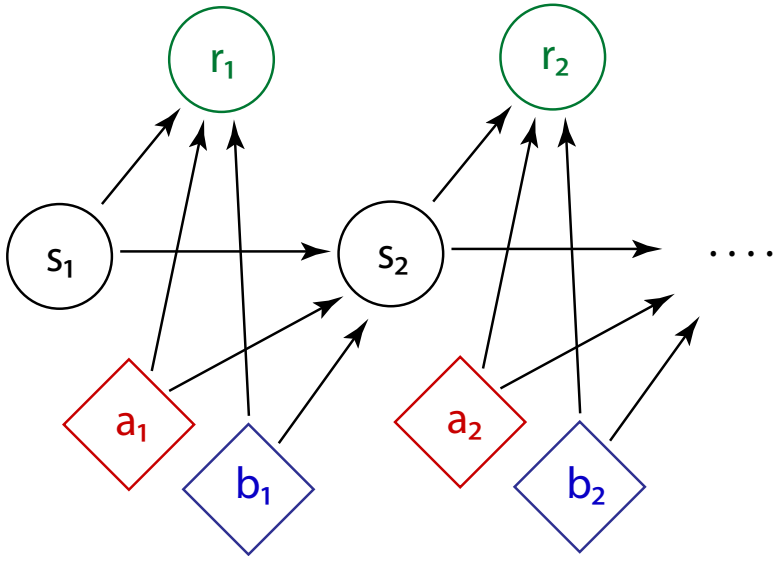

The opponents are not fixed, and can be adaptive.

# Markov Games


Markov Game $(\mathcal{S}, \{\mathcal{A}_i\}_{i=1}^m, \mathbb{P}, \{r_i\}_{i=1}^m, H)$.

# Markov Games


Markov Game $(\mathcal{S}, \{\mathcal{A}_i\}_{i=1}^m, \mathbb{P}, \{r_i\}_{i=1}^m, H)$.

- Transition $\mathbb{P}_h(s_{h+1}|s_h, \boldsymbol{a}_h)$, reward for $i^{\text{th}}$ player $r_{i,h}(s_h, \boldsymbol{a}_h)$.
- $\boldsymbol{a}_h$ is the joint action of all players $\boldsymbol{a} = (a^{(1)}, \ldots, a^{(m)})$.

# Learning objectives

- Policy for $i^{\text{th}}$ player $\pi_i : \mathcal{S} \times [H] \to \Delta_{\mathcal{A}_i}$.

# Learning objectives

- Policy for $i^{\text{th}}$ player $\pi_i : \mathcal{S} \times [H] \to \Delta_{\mathcal{A}_i}$.
- Goal: finding optimal policy in the sense of **Nash equilibrium**: a *product* policy $\pi$, where no player can gain by deviating from her own policy while fixing other players' policies.

# Learning objectives

- Policy for $i^{\text{th}}$ player $\pi_i : \mathcal{S} \times [H] \to \Delta_{\mathcal{A}_i}$.
- Goal: finding optimal policy in the sense of **Nash equilibrium**: a *product* policy $\pi$, where no player can gain by deviating from her own policy while fixing other players' policies.
- We focus on two-player zero-sum game in this work, where it remains to achieve sub-linear **Regret**:

$$\text{Regret}(K) := \sum_{k=1}^{K} \left[ V_1^{\star}(s_1) - V_1^{\mu^k,\dagger}(s_1) \right].$$

# Unique Challenge: Large State Space

- Classical RL: **Tabular case**

# Unique Challenge: Large State Space

- Classical RL: **Tabular case**
- The numbers of states & actions are finite and small.

# Unique Challenge: Large State Space

- Classical RL: **Tabular case**
- The numbers of states & actions are finite and small.
- Strategy: visit all "reachable" states, and learn directly.

# Unique Challenge: Large State Space


- Modern RL:**Function Approximation**

# Unique Challenge: Large State Space


- Modern RL:**Function Approximation**
- The number of states in practice is typically $\geq 10^{100}$. Most states are not visited even once.

# Unique Challenge: Large State Space


- Modern RL:**Function Approximation**
- The number of states in practice is typically $\geq 10^{100}$. Most states are not visited even once.
- Strategy: approximate "value" or "policy" by functions in a parameteric class $\mathcal{F}$ (eg. Deep NN).

# Main results

In this work, we propose the first provably sample-efficient RL algorithm with general function approximation, *GOLF_with_Exploiter*.

# Main results

In this work, we propose the first provably sample-efficient RL algorithm with general function approximation, *GOLF_with_Exploiter*.

**Theorem**

For zero-sum MGs equipped with a $Q$-function class $\mathcal{F}$ whose multiagent Bellman-Eluder dimension is $d$, *GOLF_with_Exploiter* learns an $\epsilon$-Nash policy within $\tilde{O}(H^2 d \log(|\mathcal{F}|)/\epsilon^2)$ episodes.

# Main results

In this work, we propose the first provably sample-efficient RL algorithm with general function approximation, *GOLF_with_Exploiter*.

**Theorem**

For zero-sum MGs equipped with a $Q$-function class $\mathcal{F}$ whose multiagent Bellman-Eluder dimension is $d$, *GOLF_with_Exploiter* learns an $\epsilon$-Nash policy within $\tilde{O}(H^2 d \log(|\mathcal{F}|)/\epsilon^2)$ episodes.

Exploiter style of exploration:

- Main agent: play optimistic Nash policy.
- Exploiter: play optimistic best response to the main agent.

# Main results

In this work, we propose the first provably sample-efficient RL algorithm with general function approximation, *GOLF_with_Exploiter*.

**Theorem**

For zero-sum MGs equipped with a $Q$-function class $\mathcal{F}$ whose multiagent Bellman-Eluder dimension is $d$, *GOLF_with_Exploiter* learns an $\epsilon$-Nash policy within $\tilde{O}(H^2 d \log(|\mathcal{F}|)/\epsilon^2)$ episodes.

Exploiter style of exploration:

- Main agent: play optimistic Nash policy.
- Exploiter: play optimistic best response to the main agent.

Applies to a rich class of models including tabular MGs, MGs with linear or kernel function approximation, and MGs with rich observations.