Posterior Sampling for Competitive RL: Function Approximation and Partial Observation

Shuang Qiu[1*] Ziyu Dai[2*] Han Zhong[3]
Zhaoran Wang[4] Zhuoran Yang[5] Tong Zhang[1]
(*Equal contribution)

[1] HKUST [2] New York University [3] Peking University
[4] Northwestern University [5] Yale University

NeurIPS 2023

Motivation

- Multi-agent reinforcement learning (MARL)
  - Empirical success: autonomous driving, Go, StarCraft, Dota2, Poker
  - Practical scenario: partial observations and function approximation
  - Our focus: the competitive setting
- Posterior sampling
  - A powerful method in practice
  - Extensively studied in single-agent RL
  - Explicit construction of bonus terms is not needed
  - Lacks sufficient theoretical understanding in MARL
- Question

> Can we design provably sample-efficient posterior sampling algorithms for competitive RL with even partial observations under general function approximation?

Contribution

- Propose the two generalized eluder coefficient (GEC) as the complexity measure for MARL with function approximation, named self-play GEC and adversarial GEC
- Propose a model-based posterior sampling algorithm for self-play with general function approximation under both fully and partially observable settings
- Propose a model-based posterior sampling algorithm for adversarial learning with general function approximation under both fully and partially observable settings
- Theoretically prove regret bounds for our proposed algorithms, incorporating the proposed self-play GEC and adversarial GEC.

Problem Setup

- Zero-sum Fully Observable Markov Game(FOMG)
  - State space $\mathcal{S}$, action spaces $\mathcal{A}$ and $\mathcal{B}$, total steps H, and reward function $r_h(s, a, b)$.
  - The state s transitions to s' under an unknown probability distribution $\mathbb{P}_h(s'|s, a, b)$.
  - The state s is observable to agents
- Zero-Sum Partially Observable Markov Games (POMG)
  - An observation space $\mathcal{O}$
  - Only a partial observation $o \in \mathcal{O}$ of state s is observable, sampled from an unknown emission kernel $\mathbb{O}_h(o|s)$
  - Reward function $r_h(o, a, b)$
- Function approximation
  - We use a function f in a function class $\mathcal{F}$ to approximate the environment $f^* \in \mathcal{F}$.
  - f^* represents the true transition kernel $\mathbb{P}$ for FOMG, and the true transition kernel $\mathbb{P}$, emission kernel $\mathbb{O}$, and initial state distribution μ_1 for POMG.

Problem Setup

- Self-play setting
  - The learner can control *both* players to find an approximate Nash equilibrium
  - The objective is designing sample-efficient algorithms to generate a sequence of policy pairs $\{(\pi^t, \nu^t)\}_{t=1}^T$ to minimize the following regret

$$\mathrm{Reg}^{\mathrm{sp}}(T) := \sum_{t=1}^T \left[V_{f^*}^{*,\nu^t} - V_{f^*}^{\pi^t,*}\right].$$

- Adversarial setting
  - Only *single* player is controllable, and the opponent plays *arbitrary* policies.
  - The objective is learning policies $\{\pi^t\}_{t=1}^T$ to maximize the overall cumulative rewards in the presence of an adversary such that the following regret is minimized

$$\mathrm{Reg}^{\mathrm{adv}}(T) := \sum_{t=1}^T \left[V_{f^*}^{*} - V_{f^*}^{\pi^t,\nu^t}\right].$$

Algorithm for the Self-Play Setting

- Self-play algorithm for Max-Player (Player 1) at each step $t \le [T]$
  1. Draw a model $\overline{f}^t \sim p^t(f) \propto p^0(f) \exp[\gamma_1 V_f^* + \sum_{\tau=1}^{t-1} \sum_{h=1}^{H} L_h^\tau(f)]$.
     Compute π^t by letting $(\pi^t, \overline{\nu}^t)$ be the Nash equilibrium of $V_{\overline{f}^t}^{\pi,\nu}$.
  2. Draw a model $\underline{f}^t \sim q^t(f) \propto q^0(f) \exp[-\gamma_2 V_f^{\pi^t,*} + \sum_{\tau=1}^{t-1} \sum_{h=1}^{H} L_h^\tau(f)]$.
     Compute $\underline{\nu}^t$ by letting $\underline{\nu}^t$ be the best response of π^t w.r.t. $V_{\underline{f}^t}^{\pi,\nu}$.
  3. Collect data $\mathcal{D}^t$ via an exploration policy σ^t and calculate $\{L_h^t(f)\}_{h=1}^H$ using $\mathcal{D}^t$.

  Return: $(\pi^1, ..., \pi^T)$.

- Main idea:
  - Optimistic model-based posterior sampling
  - Optimism term + Likelihood function
  - Step 2 aims to assist the learning for the max-player by exploiting her weakness

Algorithm for the Self-Play Setting

- Example setups of data exploration:
  - $\sigma^t = (\pi^t, \underline{\nu}^t)$;
  - FOMG: $\mathcal{D}^t = \{(s_h^t, a_h^t, b_h^t, s_{h+1}^t)\}_{h=1}^H$ and
$$L_h^t(f) = \eta \log \mathbb{P}_{f,h}(s_{h+1}^t \,|\, s_h^t, a_h^t, b_h^t).$$
  - POMG: $\mathcal{D}^t = \{\tau_h^t\}_{h=1}^H$ with $\tau_h^t := (o_1^t, a_1^t, b_1^t \ldots, o_h^t, a_h^t, b_h^t)$ and
$$L_h^t(f) = \eta \log \mathbf{P}_{f,h}(\tau_h^t).$$

    where we define $\mathbf{P}_{f,h}(\tau_h) := \int_{\mathcal{S}^h} \mu_{f,1}(s_1) \prod_{h'=1}^{h-1} [\mathbb{O}_{f,h'}(o_{h'}|s_{h'})\mathbb{P}_{f,h'}(s_{h'+1}|s_{h'}, a_{h'}, b_{h'})]$ $\mathbb{O}_{f,h}(o_h|s_h)\mathrm{d}s_{1:h}$ under an approximation function f.

- The self-play algorithm for Min-Player (Player 2) is symmetric to the above one for Max-Player and returns the policies $(\nu^1, ..., \nu^T)$.

Theoretical Result

Definition 1 (Self-Play GEC)

For any sequences of functions $f^t, g^t \in \mathcal{F}$, suppose that a pair of policies (π^t, ν^t) satisfies: **(a)** $\pi^t = \mathrm{argmax}_\pi \min_\nu V_{f^t}^{\pi,\nu}$ and $\nu^t = \mathrm{argmin}_\nu V_{g^t}^{\pi^t,\nu}$, or **(b)** $\nu^t = \mathrm{argmin}_\nu \max_\pi V_{f^t}^{\pi,\nu}$ and $\pi^t = \mathrm{argmax}_\pi V_{g^t}^{\pi,\nu^t}$. Denoting the joint exploration policy as σ^t depending on f^t and g^t, for any $\rho \in \{f, g\}$ and (π^t, ν^t) following **(a)** and **(b)**, the self-play GEC d_{GEC} is defined as the minimal constant d satisfying

$$\Big|\underbrace{\textstyle\sum_{t=1}^T \big(V_{\rho^t}^{\pi^t,\nu^t} - V_{f^*}^{\pi^t,\nu^t}\big)}_{\text{prediction error}}\Big| \le \Big[d \sum_{h=1}^H \sum_{t=1}^T \underbrace{\Big(\textstyle\sum_{\tau=1}^{t-1} \mathbb{E}_{(\sigma^\tau,h)} \ell(\rho^t, \xi_h^\tau)\Big)}_{\text{training error}}\Big]^{\frac{1}{2}} + \underbrace{2H(dHT)^{\frac{1}{2}} + \epsilon HT}_{\text{burn-in error}},$$

where (σ^τ, h) implies running the joint exploration policy σ^τ to step h to collect a data point ξ_h^τ.

- $\ell(f, \xi_h)$ is determined for FOMGs with $\xi_h = (s_h, a_h, b_h)$ and POMGs with $\xi_h = \tau_h$ as

  FOMG: $D_{\mathrm{He}}^2(\mathbb{P}_{f,h}(\cdot|\xi_h), \mathbb{P}_{f^*,h}(\cdot|\xi_h))$, POMG: $1/2 \cdot \Big(\sqrt{\mathbf{P}_{f,h}(\xi_h)/\mathbf{P}_{f^*,h}(\xi_h)} - 1\Big)^2$.

- Intuition: hypotheses having a small training error on a well-explored dataset imply a small out-of-sample prediction error, characterizing the hardness of exploration.

Theoretical Result

Theorem 2

With proper settings of η, γ_1, γ_2, and ϵ, when the number of rounds T is sufficiently large, for both FOMG and POMG, the proposed self-play algorithm admits a regret of

$$\mathbb{E}[\mathrm{Reg}^{\mathrm{sp}}(T)] \leq 12\sqrt{d_{\mathrm{GEC}} HT \cdot [\omega(4HT, p^0) + \omega(4HT, q^0)]}.$$

- The regret sublinearly depends on T, d_{GEC}, and ω
- ω measures how well the prior distributions cover the optimal model f^*

Definition 3 (Prior around the True Model)

Given $\beta > 0$ and any distribution $p^0 \in \Delta_{\mathcal{F}}$, we define a quantity $\omega(\beta, p^0)$ as $\omega(\beta, p^0) = \inf_{\varepsilon>0}\{\beta\varepsilon - \ln p^0[\mathcal{F}(\varepsilon)]\}$, where we define the classes $\mathcal{F}(\varepsilon) := \{f \in \mathcal{F} \ : \ \sup_{h,s,a,b} \mathrm{KL}^{\frac{1}{2}}(\mathbb{P}_{f^*,h}(\cdot \mid s,a,b) \| \mathbb{P}_{f,h}(\cdot \mid s,a,b)) \leq \varepsilon\}$ for FOMGs and $\mathcal{F}(\varepsilon) := \{f \in \mathcal{F} \ : \ \sup_{\pi,\nu} \mathrm{KL}^{\frac{1}{2}}(\mathbf{P}^{\pi,\nu}_{f^*,H} \| \mathbf{P}^{\pi,\nu}_{f,H}) \leq \varepsilon\}$ for POMGs.

Algorithm for the Adversarial Setting

- Adversarial learning algorithm for the main player at each step $t \leq [T]$
  1. Draw a model $f^t \sim p^t(f) \propto p^0(f) \exp[\gamma V_f^* + \sum_{\tau=1}^{t-1} \sum_{h=1}^{H} L_h^\tau(f)]$.
     Compute π^t by letting $(\pi^t, \overline{\nu}^t)$ be the Nash equilibrium of $V_{f^t}^{\pi,\nu}$.
  2. The opponent picks an arbitrary policy ν^t.
  3. Collect data $\mathcal{D}^t$ by executing an exploration policy $\sigma^t = (\pi^t, \nu^t)$ cnd calculate the likelihood functions $\{L_h^t(f)\}_{h=1}^{H}$.

  Return: $(\pi^1, \ldots, \pi^T)$.

- Differences from the self-play setting:
  - The opponent plays an arbitrary policy ν^t that is uncontrolled by the algorithm
  - The exploration policy σ^t is defined based on the the opponent's arbitrary policy ν^t

Theoretical Results

Definition 4 (Adversarial GEC)

For any sequence of functions $\{f^t\}_{t=1}^T$ with $f^t \in \mathcal{F}$ and any sequence of the opponent's policies $\{\nu^t\}_{t=1}^T$, suppose that the main player's policies $\{\mu^t\}_{t=1}^T$ are generated via $\mu^t = \operatorname{argmax}_\pi \min_\nu V_{f^t}^{\pi,\nu}$. Denoting the joint exploration policy as $\{\sigma^t\}_{t=1}^T$ depending on $\{f^t\}_{t=1}^T$, the adversarial GEC d_{GEC} is defined as the minimal constant d satisfying

$$\sum_{t=1}^T \left(V_{f^t}^{\pi^t,\nu^t} - V_{f^*}^{\pi^t,\nu^t}\right) \le \left[d \sum_{h=1}^H \sum_{t=1}^T \left(\sum_{\tau=1}^{t-1} \mathbb{E}_{(\sigma^\tau,h)} \ell(f^t, \xi_h^\tau)\right)\right]^{\frac{1}{2}} + 2H(dHT)^{\frac{1}{2}} + \epsilon HT.$$

- Difference from self-play GEC: the opponent's policy ν^t is arbitrary and uncontrolled

Theorem 5

With proper settings of η, γ_1, γ_2, and ϵ, when the number of rounds T is sufficiently large, for both FOMG and POMG, the adversarial learning algorithm admits a regret of

$$\mathbb{E}[\mathrm{Reg}^{\mathrm{adv}}(T)] \le 4\sqrt{d_{\mathrm{GEC}} HT \cdot \omega(4HT, p^0)}.$$

- The regret sublinearly depends on T, d_{GEC}, and ω

Examples

- Classes with low self-play/adversarial GEC cover a wide range of known Markov game (MG) classes
- FOMG:
  - **Linear MG.** $r_h(s,a,b) = \mathbf{w}_h^\top \boldsymbol{\phi}(s,a,b)$ and $\mathbb{P}_h(s'|s,a,b) = \boldsymbol{\theta}_h(s')^\top \boldsymbol{\phi}(s,a,b)$ with $\boldsymbol{\phi}(s,a,b) \in \mathbb{R}^d$. We have $d_{\mathrm{GEC}} = \widetilde{O}(H^3 d)$.
  - **Linear Mixture MG.** $\mathbb{P}_h(s'|s,a,b) = \boldsymbol{\theta}_h^\top \boldsymbol{\phi}(s,a,b,s')$ with $\boldsymbol{\phi}(s,a,b,s') \in \mathbb{R}^d$. We have $d_{\mathrm{GEC}} = \widetilde{O}(H^3 d)$.
  - **MG with Low Self-Play Witness Rank.** An inner product of specific vectors in $\mathbb{R}^d$ can lower bound witnessed model misfit and upper bound the Bellman error with a coefficient κ_{wit}. We have $d_{\mathrm{GEC}} = \widetilde{O}(H^3 d/\kappa_{\mathrm{wit}}^2)$.
- POMG:
  - α**-Weakly Revealing POMG.** The matrix by $\mathbb{O}_h(\cdot|\cdot)$ has singular values $\geq \alpha$. We have $d_{\mathrm{GEC}} = \widetilde{O}(H^3|\mathcal{O}|^3|\mathcal{A}|^2|\mathcal{B}|^2|\mathcal{S}|^2/\alpha^2)$.
  - **Decodable POMG.** An unknown decoder ϕ_h recovers states from observations via $\phi_h(o) = s$. We have $d_{\mathrm{GEC}} = \widetilde{O}(H^3|\mathcal{O}|^3|\mathcal{A}|^2|\mathcal{B}|^2)$.

Discussion of $\omega(\beta, p^0)$

- $\mathcal{F}$ is finite
  - $\omega(\beta, p^0) \leq \log|\mathcal{F}|$ with setting $p^0 = \mathrm{Unif}(\mathcal{F})$
- $\mathcal{F}$ is infinite
  - $\omega(\beta, p^0) \leq$ log-covering number of $\mathcal{F}$ w.r.t. the ℓ_1 distance.
- We generalize existing results of $\omega(\beta, p^0)$ for the fully observable setting to the partially observable setting, which is of independent interest

Thank you!