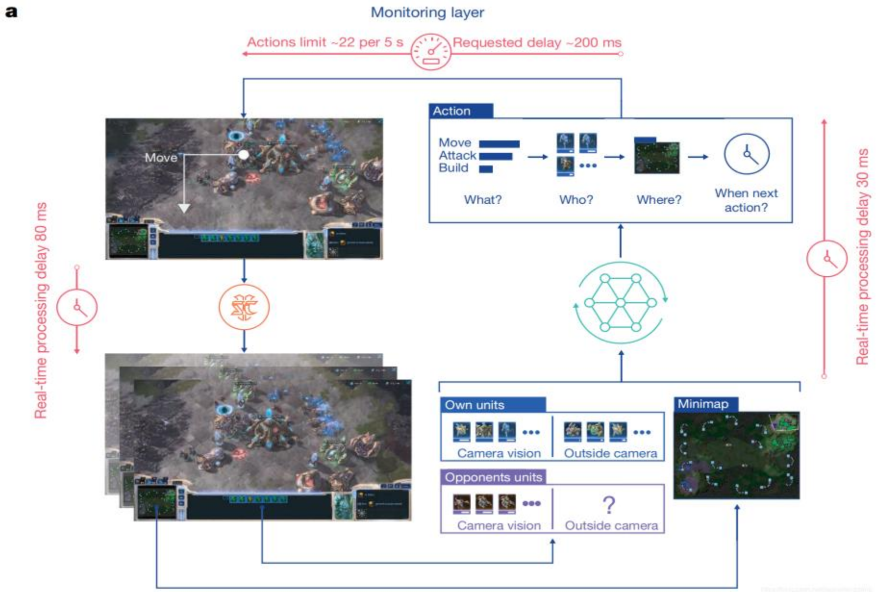
Provably Efficient Policy Optimization for Two-Player Zero-Sum Markov Games

Yulai Zhao · Yuandong Tian · Jason Lee · Simon Du

Provably Efficient Policy Optimization for Two-Player Zero-Sum Markov Games -- Backgrounds

- Two-player zero-sum game is a widely used setting with applications (Go, StarCraft Ⅱ …)
- Policy optimization methods are widely used in solving zero-sum games (AlphaGo, LOLA…)



Provably Efficient Policy Optimization for Two-Player Zero-Sum Markov Games -- Problem

- Despite the large body of empirical work on using policy optimization methods for two-player zero-sum Markov games, theoretical studies are very limited.

Can we design a provably efficient policy optimization algorithm with function approximation for two-player zero-sum Markov games with a large state-action space?

Provably Efficient Policy Optimization for Two-Player Zero-Sum Markov Games -- Setup

- Two-Player zero-sum Markov Games
  - a tuple $M = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathrm{r}, \gamma)$: A set of states $\mathcal{S}$, a set of actions $\mathcal{A}$, a transition probability $\mathcal{P}: \mathcal{S} \times \mathcal{A} \times \mathcal{A} \rightarrow \Delta(\mathcal{S})$, a reward function $\mathrm{r}: \mathcal{S} \times \mathcal{A} \times \mathcal{A} \rightarrow [0, 1]$, a discounted factor $\gamma \in [0, 1)$.
  - define policies as probability distributions over action space: $x, f \in \mathcal{S} \rightarrow \Delta(\mathcal{A})$, max player x seeks to maximize the reward while min player f seeks to minimize.
- value function
  - $V^{x,f}(s) = E_{\substack{a_t \sim x \\ b_t \sim f}} \left[\sum_{t=0}^{\infty} \gamma^t \, r(s_t, a_t, b_t) \middle| \, s_0 = s \right]$
  - $V^{x,f}(\rho) = E_{s \sim \rho} V^{x,f}(s)$

Provably Efficient Policy Optimization for Two-Player Zero-Sum Markov Games -- Setup

- (x^*, f^*) is a pair of **Nash equilibrium (NE)** if the following inequalities hold for any distribution ρ and policy pair (x, f):

$$V^{x,f^*}(\rho) \leq V^{x^*,f^*}(\rho) = V^*(\rho) \leq V^{x^*,f}(\rho)$$

- Our goal: find an approximate pair of Nash equilibrium, which means output x should make the following metric small

$$V^*(\rho) - \inf_f V^{x,f}(\rho)$$

- We use **concentrability coefficients** as in the previous work [Perolat et al., 2015].

Definition 1 (Concentrability Coefficients). *Given two distributions over states: ρ and σ. When σ is element-wise non-negative, define*

$$c_{\rho,\sigma}(j) = \sup_{x^1, f^1, \cdots x^j, f^j \in \mathcal{S} \to \Delta(\mathcal{A})} \left\| \frac{\rho \mathcal{P}_{x^1,f^1} \cdots \mathcal{P}_{x^j,f^j}}{\sigma} \right\|_\infty,$$

$$\mathcal{C}'_{\rho,\sigma} = (1-\gamma)^2 \sum_{m \geq 1} m\gamma^{m-1} c_{\rho,\sigma}(m-1),$$

$$\mathcal{C}^{l,k,d}_{\rho,\sigma} = \frac{(1-\gamma)^2}{\gamma^l - \gamma^k} \sum_{i=l}^{k-1} \sum_{j=i}^{\infty} \gamma^j c_{\rho,\sigma}(j+d).$$

- **σ** is the optimization measure we use to train the policy.
- **ρ** is the performance measure of our interest.

Population Algorithm for Tabular case

- We divide each outer loop into two steps.
  I. In Greedy Step, we intend to find approximate solution (x, f) for Bellman operator $\mathcal{T}$ onto current value function V_{k-1} with T' updates. (towards V^*)
  II. In Iteration Step, we run T NPG updates to solve $\arg\min_f V^{x,f}$ which is known as finding the best response of min player when fixing $x = x^k$.

- Theorem 1 (informal): For this setting, after K outer loops

$$V^*(\rho) - \inf_f V^{x^K,f}(\rho) = \tilde{O}\left(\frac{\mathcal{C}_{\rho,\sigma}^{1,K,0}}{(1-\gamma)^4 T} + \frac{\mathcal{C}_{\rho,\sigma}^{0,K,0}}{(1-\gamma)^4 T'} \log T' + \frac{\gamma^K}{1-\gamma} \mathcal{C}_{\rho,\sigma}^{K,K+1,0}\right).$$

$$\mathcal{T}_{x,f} v = r_{x,f} + \gamma\, \mathcal{P}_{x,f} v$$
$$\mathcal{T} v = \sup_x \inf_f \mathcal{T}_{x,f}\ v$$

Online Algorithm with Function Approximation

- We still divide each outer loop into two steps.
- Assume **Episodic Sampling Oracle** to provide unbiased estimates or a fixed state-action distribution ν_0, we can start from $s_0, a_0, b_0 \sim \nu_0$, then act according to any policy x, f, and terminate it when desired.
  I. In Greedy Step, our goal is still to obtain a near-optimal x^k with respect to V_{k-1}. Different from tabular case, we use sample-based NPG updates.
  II. After obtaining x^k from Greedy Step, we run T sample-based NPG updates (each with N samples) to find best response of min player.

- Theorem 2 (informal): For this setting, after K outer loops

$$E\left[V^*(\rho) - \inf_f V^{x,f}(\rho)\right] = \tilde{O}\left(\frac{1}{\sqrt{T}} + \frac{1}{N^{1/4}}\right)$$

Provably Efficient Policy Optimization for Two-Player Zero-Sum Markov Games -- Contributions

I. In Greedy Step, design a subroutine that found minmax solutions to a matrix game without prior knowledge of model parameters.

II. In Iteration Step, leverage NPG methods to update policies.

III. Finally, develop new perturbation analyses which may be of independent interest for provable multi-agent RL.

Thank you for your time
and consideration!