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

Yulai Zhao, Yuandong Tian, Jason D. Lee, Simon S. Du

Tsinghua University, Meta AI Research, Princeton University, University of Washington

Abstract

- Policy-based methods with function approximation are widely used for solving two-player zero-sum games with large state and/or action spaces.
- However, it remains elusive how to obtain optimization and statistical guarantees for such algorithms.
- We present a new policy optimization algorithm with function approximation and prove that under standard regularity conditions on the Markov game and the function approximation class, our algorithm finds a near-optimal policy within a polynomial number of samples and iterations.
- To our knowledge, this is the first provably efficient policy optimization algorithm with function approximation that solves 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?

We answer this question affirmatively!

- Two-Player zero-sum Markov Games
  - a tuple $M = (\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma)$: A set of states $\mathcal{S}$, a set of actions $\mathcal{A}$, a transition probability $\mathcal{P}: \mathcal{S} \times \mathcal{A} \times \mathcal{A} \to \Delta(\mathcal{S})$, a reward function $r: \mathcal{S} \times \mathcal{A} \times \mathcal{A} \to [0,1]$, a discounted factor $\gamma \in [0,1)$.
  - define policies as probability distributions over action space: $x, f \in \mathcal{S} \to \Delta(\mathcal{A})$, max player x seeks to maximize the reward while min player f seeks to minimize.

value function

$$V^{x,f}(s) = E_{\substack{a_t \sim x \\ b_t \sim f}}\left[\sum_{t=0}^{\infty} \gamma^t\, r(s_t, a_t, b_t) \mid s_0 = s\right]$$

$$V^{x,f}(\rho) = E_{s \sim \rho} V^{x,f}(s)$$

Setup

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

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

Our goal: find an approximate pair of Nash equilibrium, which means output x should make $V^*(\rho) - \inf_f V^{x,f}(\rho)$ small

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*

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

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

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

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

Results

Population Algorithm for Tabular case

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

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:

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

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:

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

Conclusions

- This paper gave the first quantitative analysis of policy gradient methods for general two-player zero-sum Markov games with function approximation.
- We quantified the performance gap of the output policy in terms of the number of iterations, number of samples, concentrability coefficients, and approximation error.
- An interesting direction is to extend our results to more advanced PG methods such as PPO.

References

Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. Optimality and approximation with policy gradient methods in Markov decision processes. In Conference on Learning Theory, pages 64–66. PMLR, 2020.

Yu Bai and Chi Jin. Provable self-play algorithms for competitive reinforcement learning. In International Conference on Machine Learning, pages 551–560. PMLR, 2020

Sham M Kakade. A natural policy gradient. In Advances in neural information processing systems, pages 1531–1538, 2002.

Julien Perolat, Bruno Scherrer, Bilal Piot, and Olivier Pietquin. Approximate dynamic programming for two-player zero-sum Markov games. In International Conference on Machine Learning, pages 1321–1329, 2015.

Sasha Rakhlin and Karthik Sridharan. Optimization, learning, and games with predictable sequences. In Advances in Neural Information Processing Systems, pages 3066–3074, 2013.

Yuandong Tian, Jerry Ma, Qucheng Gong, Shubho Sengupta, Zhuoyuan Chen, James Pinkerton, and Larry Zitnick. Elf opengo: An analysis and open reimplementation of alphazero. In International Conference on Machine Learning, pages 6244–6253. PMLR, 2019.

Kaiqing Zhang, Zhuoran Yang, and Tamer Basar. Policy optimization provably converges to Nash equilibria in zero-sum linear quadratic games. In Advances in Neural Information Processing Systems, pages 11602–11614, 2019.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347, 2017.

Acknowledgements

JDL acknowledges support of the ARO under MURI Award W911NF-11-1-0304, the Sloan Research Fellowship, NSF CCF 2002272, NSF IIS 2107304, and an ONR Young Investigator Award.

SSD acknowledges funding from NSF Award's IIS-2110170 and DMS-2134106.