Local Optimization Achieves Global Optimality in Multi-Agent Reinforcement Learning

Yulai Zhao, Princeton University
Zhuoran Yang, Yale University
Zhaoran Wang, Northwestern University
Jason Lee, Princeton University

ABSTRACT

Policy optimization methods with function approximation are widely used in multi-agent reinforcement learning. However, it remains elusive how to design such algorithms with statistical guarantees. Leveraging a multi-agent performance difference lemma that characterizes the landscape of multi-agent policy optimization, we find that the localized action value function serves as an ideal descent direction for each local policy. Motivated by the observation, we present a multi-agent PPO algorithm in which the local policy of each agent is updated similarly to vanilla PPO. We prove that with standard regularity conditions on the Markov game and problem-dependent quantities, our algorithm converges to the globally optimal policy at a sublinear rate. We extend our algorithm to the off-policy setting and introduce pessimism to policy evaluation, which aligns with experiments. To our knowledge, this is the first provably convergent multi-agent PPO algorithm in cooperative Markov games.

MOTIVATION

Multi-agent reinforcement learning (MARL) has demonstrated many empirical successes, e.g. strategic games (Go, StarCraft II…)

- **Main Challenges in MARL** (Zhang 2021)

1. non-stationarity: each action taken by one agent affects the total reward and the transition of state.

2. scalability: taking other agents into consideration, each individual agent would face the joint action space, whose dimension increases exponentially with the number of agents

3. function approximation: closely related to the scalability issue, the state space and joint action space are often immense in MARL

Despite the empirical successes, theoretical studies of policy optimization in MARL are very limited. Even for the cooperative setting where the agents share a common goal: maximizing the total reward function. In this paper, we aim to answer the following fundamental question:

Can we design a provably convergent multi-agent policy optimization algorithm in the cooperative setting with function approximation?

We answer this question affirmatively!

PROBLEM SETUP

- **Fully-cooperative Markov Games**
  - a tuple $M = (N, \mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma)$: A party of participants N, 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: $\pi \in \mathcal{S} \to \Delta(\mathcal{A})$. We shall use bold $\boldsymbol{\pi}$ to represent the joint policy.

Value function: $V^{\pi}(s) = E_{a\sim\pi}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, b_t) | s_0 = s\right]$.

- **Multi-agent Notations**

We write index k on superscript when we refer to the specific k-th agent. When bold symbols are used without any superscript (e.g., $\boldsymbol{a}$), they consider all agents. For simplicity, let $(m: m')$ be shorthand for set: $\{i | m \le i \le m',\ i \in N\}$.

Definition 3.1. Let P be a subset in N . The multi-agent action value function associated with agents in P is: $Q_{\pi}^{P}(s,\ \boldsymbol{a}^{P}) = E_{\tilde{\boldsymbol{a}}\sim\tilde{\boldsymbol{\pi}}}[Q_{\pi}(s,\ \boldsymbol{a}^{P},\ \tilde{\boldsymbol{a}})]$, we use a tilde over symbols to refer to the complement agents, namely $\tilde{a} = \{a^{i} | i \notin P,\ i \in N\}$.

Let $P,\ P' \subset N$ be two disjoint subsets of agents. The multiagent advantage function is defined below. Essentially, it accounts for the improvements of setting agents $a^{P'}$ upon setting agents a^{P}, while all other agents follow $\boldsymbol{\pi}$: $A_{\pi}^{P'}(s,\ \boldsymbol{a}^{P},\ \boldsymbol{a}^{P'}) = Q_{\pi}^{P\cup P'}(s, \boldsymbol{a}^{P}, \boldsymbol{a}^{P'}) - Q_{\pi}^{P}(s,\ \boldsymbol{a}^{P})$.

RESULTS

- **Multi-agent PPO for online setting**

For this setting, we adopt a useful log-linear parametrization to build conditional dependency structure. Such a conditional structure enables us to implement local optimization for every agent while still aiming at global optimality, because action value functions are ideal descent directions for local policies, a result from the multi-agent advantage function decomposition (Kuba 2022).

At the k-th iteration, we have the current policy $\boldsymbol{\pi}_{\theta_k}$, and we need to:

(1) perform **policy evaluation** to obtain the action value function estimates $\widehat{\boldsymbol{Q}}_{\boldsymbol{\pi}_{\theta_k}}$ for determining the quality of $\boldsymbol{\pi}_{\theta_k}$.

(2) perform **policy improvement** to update policy to $\boldsymbol{\pi}_{\theta_{k+1}}$. In practice, we approximate such policy within a certain parametrization class, which is achieved by implementing SGD updates.

Algorithm 1 Multi-Agent PPO

Input: Markov game $(\mathcal{N}, \mathcal{S}, \boldsymbol{\mathcal{A}}, \mathcal{P}, r, \gamma)$, penalty parameter β, stepsize η for sub-problem, number of SGD iterations T, number of iterations K.
Output: Uniformly sample k from $0, 1, \cdots K-1$, return $\bar{\boldsymbol{\pi}} = \boldsymbol{\pi}_{\theta_k}$.
1: Initialize $\theta_0^m = 0$ for every $m \in \mathcal{N}$.
2: **for** $k = 0, 1, \ldots, K-1$ **do**
3: Set parameter $\beta_k \leftarrow \beta\sqrt{K}$
4: **for** $m = 1, \cdots, N$ **do**
5: Sample $\{s_t, \mathbf{a}_t^{1:m-1}, a_t^m\}_{t=0}^{T-1}$ from $\sigma_k = \nu_k\boldsymbol{\pi}_{\theta_k}$.
6: Obtain $\hat{Q}_{\boldsymbol{\pi}_{\theta_k}}^{1:m}(s, \mathbf{a}^{1:m-1}, a^m)$ for each sample .
7: Feed samples into Algorithm 3, obtain θ_{k+1}^m.
8: **end for**
9: **end for**

Parametrization For the m-th agent ($m \in \mathcal{N}$), its conditional policy depends on all prior ordered agents $\mathbf{a}^{1:m-1}$. Given a coefficient vector $\theta^m \in \Theta$, where $\Theta = \{\|\theta\| \le R | \theta \in \mathbb{R}^d\}$ is a convex, norm-constrained set. The probability of choosing action a^m under state s is

$$\pi_{\theta^m}(a^m | s, \mathbf{a}^{1:m-1}) = \frac{\exp\left(\phi^{\top}(s, \mathbf{a}^{1:m-1}, a^m)\theta^m\right)}{\sum_{a^m \in \mathcal{A}} \exp\left(\phi^{\top}(s, \mathbf{a}^{1:m-1}, a^m)\theta^m\right)} \quad (2)$$

where ϕ is a set of feature vector representations. Without loss of generality, we impose a regularity condition such that every $\|\phi\|_2 \le 1$. This parametrization has been widely used in RL literature [Branavan et al., 2009, Gimpel and Smith, 2010, Heess et al., 2013, Agarwal et al., 2020, Zhao et al., 2022].[3]

Algorithm 3 Policy Improvement Solver for MA-PPO

Input: MG $(\mathcal{N}, \mathcal{S}, \boldsymbol{\mathcal{A}}, \mathcal{P}, r, \gamma)$, iterations T, stepsize η, samples $\{s_t, \mathbf{a}_t^{1:m-1}, a_t^m\}_{t=0}^{T-1}$.
Output: Policy update θ.
1: Initialize $\theta_0 = 0$.
2: **for** $t = 0, 1, \ldots, T-1$ **do**
3: Let $(s, \mathbf{a}^{1:m-1}, a) \leftarrow (s_t, \mathbf{a}_t^{1:m-1}, a_t^m)$.
4: $\theta(t+\frac{1}{2}) \leftarrow \theta(t) - 2\eta\phi(s, \mathbf{a}^{1:m-1}, a)\left(\left(\theta(t) - \theta_k^m\right)^{\top}\phi(s, \mathbf{a}^{1:m-1}, a^m) - \beta_k^{-1}\hat{Q}_{\boldsymbol{\pi}_k}^{1:m}(s, \mathbf{a}^{1:m-1}, a^m)\right)$
5: $\theta(t+1) \leftarrow \Pi_{\Theta}\theta(t+\frac{1}{2})$
6: **end for**
7: Calculate average: $\bar{\theta} \leftarrow \frac{1}{T}\sum_{t=1}^{T}\theta_t$.

RESULTS (CONTINUE)

Theorem 1 (informal): For this setting, after K iterations, we have $J(\pi^*) - J(\bar{\pi})$ upper bounded by $O\left(\frac{N}{1-\gamma}\sqrt{\frac{\log(|A|)}{K}}\right)$.

- **Pessimistic MA-PPO with Linear Function Approximation**

We perform pessimistic policy evaluation via regularization to reduce such overestimation aligning with experimental works.

Theorem 2 (informal): For this setting, after K iterations, we have $J(\pi^*) - J(\bar{\pi})$ upper bounded by $O\left(\frac{N}{(1-\gamma)^2}\sqrt{\frac{\log(|A|)}{K}} + \frac{C}{(1-\gamma)^2}\sqrt[4]{\frac{d\log\frac{NK}{\alpha}}{n}}\right)$.

REFERENCES

Kaiqing Zhang, Zhuoran Yang, and Tamer Basar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. Handbook of Reinforcement Learning and Control, pp. 321–384, 2021.

Jakub Grudzien Kuba, Ruiqing Chen, Muning Wen, Ying Wen, Fanglei Sun, Jun Wang, and Yaodong Yang. Trust region policy optimisation in multiagent reinforcement learning. In International Conference on Learning Representations, 2022.

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

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.

CONCLUSIONS

1.We propose a multi-agent PPO algorithm in which the local policy of each agent is updated sequentially in a similar fashion as vanilla PPO algorithm (Schulman et al., 2017).

2.We adopt the log-linear function approximation for the policies. We prove that multi-agent PPO converges at a sublinear $O\left(\frac{N}{1-\gamma}\sqrt{\frac{\log(|A|)}{K}}\right)$ rate up to some statistical errors incurred in evaluating/improving policies.

3.Moreover, we propose an off-policy variant of the multi-agent PPO algorithm and introduce pessimism into policy evaluation.

ACKNOWLEDGEMENTS

JDL acknowledges support of the ARO under MURI Award W911NF-11-1-0304, the Sloan Research Fellowship, NSF CCF 2002272, NSF IIS 2107304, NSF CIF 2212262, ONR Young Investigator Award, and NSF CAREER Award 2144994