Investigating Ratio Clipping in Multi-agent Reinforcement Learning

Benjamin Ellis, Mingfei Sun and Shimon Whiteson

University of Oxford

Introduction

Multi-agent reinforcement learning (MARL) is a promising approach to address challenges such as autonomous driving. It has three primary settings: purely competitive, where one agent's loss is another's gain, purely cooperative, where agents work together to maximise a shared reward, and mixed, where agents sometimes would be incentivised to cooperate and sometimes not. Here we focus on the cooperative setting.

The simplest approach to multi-agent learning is *independent learning*, where each agent treats the others as part of the environment. Here we focus on independent actor-critic methods. Although the resulting non-stationarity can cause problems (see [1]), there has been some recent success [2] demonstrating independent PPO's (IPPO) ability to perform well on SMAC. Despite this success, little work has been done to understand the success of PPO in a multi-agent setting. This is the topic of this paper.

Objectives

1. Demonstrate that clipping the actor ratios is important to PPO's success in multi-agent tasks.
2. Show that clipping the actor reduces critic bias.
3. Prove that constraining the actor update constrains the distribution of the critic
4. Investigate the effect of this value clipping on task performance.

Independent PPO (IPPO) and Multi-Agent PPO (MAPPO)

Independent PPO and Multi-agent PPO (and similarly independent actor-critic [IAC] and multi-agent actor-critic [MAAC]) are extensions of PPO to the multi-agent setting. IPPO and MAPPO differ in their estimation of the advantage function. IPPO uses a critic per-agent of the form $A_i^\pi(h_{t,i}, a_{t,i})$. By contrast, in MAPPO, the critic receives as input the *joint* observation history and *joint* action, i.e. the critic has the form $A^\pi(\mathbf{h}_t, \mathbf{a}_t)$. IAC and MAAC have their critics defined analogously.

Ratio Clipping in IPPO and MAPPO

Both IPPO and MAPPO optimise decentralised policies with independently maintained clipping ratios per agent and share parameters between agents. This means that they both optimise the below objective.

$$\max_\theta \mathbb{E}_{\pi^i_{\theta_{\text{old}}}} \left[\min \left(\frac{\pi^i_\theta(a_i|h_i)}{\pi^i_{\theta_{\text{old}}}(a_i|h_i)} A^{\pi^i_{\theta_{\text{old}}}}(h_i, a_i), \right.\right.$$
$$\left.\left. \text{clip}\left(1-\epsilon, 1+\epsilon, \frac{\pi^i_\theta(a_i|h_i)}{\pi^i_{\theta_{\text{old}}}(a_i|h_i)}\right) A^{\pi^i_{\theta_{\text{old}}}}(h_i, a_i) \right) \right]$$

PPO and IAC Performance

Although PPO usually improves upon actor-critic methods in the single-agent setting, in this section we demonstrate a significant difference in performance in the multi-agent setting, even on relatively simple tasks. We therefore consider both methods with a wide range of critic forms.

We consider 3 varieties of critic in both centralised and decentralised form to demonstrate this. These critics differ primarily in their advantage estimation; we train all V^π critics using TD error (using TD(0) when advantages are estimated with TD error and TD(λ) when advantages are estimated with GAE), and Q^π using the SARSA error. For V^π critics we estimate the advantage using both GAE and TD error. For Q critics we estimate the advantage using

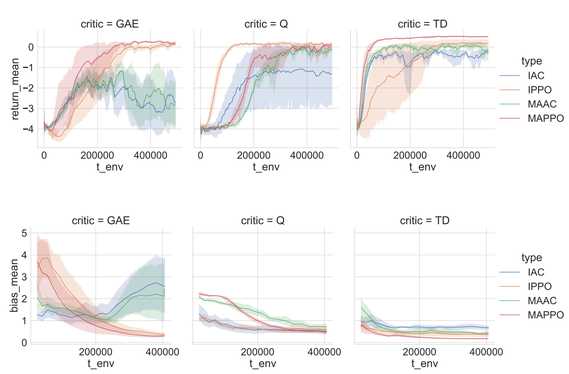
$$A^\pi(\mathbf{h}, \mathbf{a}) = Q^\pi(\mathbf{h}, \mathbf{a}) - \sum_{\mathbf{a}' \in_i \mathcal{A}_i} Q^\pi(\mathbf{h}, \mathbf{a}')\, \pi(\mathbf{a}'|\mathbf{h})$$

Although this can be computed exactly in the decentralised case, it is not feasible to sum over the joint action space. We solve this by making a Monte Carlo approximation.

To evaluate these critics, we consider a predator-prey task. Four predators must catch a single prey by surrounding it on a 4x4 grid. The prey moves randomly and there is a reward of -0.1 at each step and 1.0 when the prey is caught. Each episode is stopped after 50 steps if the prey has not been caught. The results of this experiment are shown in the next column.

Although most policies converge to a similar quality of policy, it is noticeable that, even on this relatively simple task, PPO provides an improvement in performance for every type of critic.

PPO and IAC Performance



The plot shows that, for all critics, the bias is lower for the PPO version of the critic than the actor-critic version. A possible explanation for this is that clipping the actor makes the returns distribution more predictable and therefore easier to learn.

Bound on Value Function Update

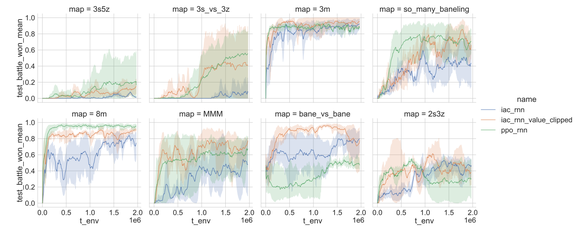
We now show in the finite-horizon setting that if the KL-divergence between the individual agent policies is bounded, then so is the ∞-norm of the difference between their value functions, and hence any required update of the critic is bounded as well.

Theorem. *Suppose that the reward at each step is bounded such that $r(s_t, a_t) \in [0, 1]$. Further suppose that the Dec-POMDP has a finite horizon of length T. Then $\forall i. D_{KL}(\pi_i(\cdot|h_i) \parallel \pi'_i(\cdot|h_i)) \leq \delta$ implies that $\left\|V^\pi - V^{\pi'}\right\|_\infty \leq (T+1)\sqrt{2\delta(T+1)}N$. Further, this bound is non-trivial when $\delta < \frac{1}{2N(T+1)}$.*

This result shows that if an algorithm bounds the amount it changes each actor's policy, for example by PPO's clipping mechanism, then this naturally leads to a bound on how much the critic needs to be updated, which could lead to reduced bias. This is because there are two sources of critic bias: first, the critic does not see a complete representation of the outcomes from each state, and secondly, the critic might not converge to the best value function for the data represented within the limited computational budget available. This bound on the value function update would mitigate the second source of bias.

Value Clipping

We now investigate whether the previous theoretical result of the bound on the value function is of practical benefit. We do this by clipping the value function. The below figure compares IPPO, IAC and IAC with value clipping on some easy SMAC maps and one hard map (3s5z).



Value clipping seems to improve the performance of IAC significantly, having a positive impact on most of the maps.

Conclusion

We studied PPO in a simple setting and showed that clipping the actor has two effects on the critic. First it reduces the critic bias. This might be because it makes the returns more stationary and therefore easier to fit. Secondly, bounding the actor bounds the value update, which is and hence might reduce the bias. We tested whether this had a positive effect in practice and found that it improved the performance of IAC.

References

[1] Laetitia Matignon, Guillaume Laurent, and Nadine Fort-Piat. Independent reinforcement learners in cooperative markov games: A survey regarding coordination problems. *The Knowledge Engineering Review*, 27:1 – 31, 03 2012.

[2] Christian Schröder de Witt, Tarun Gupta, Denys Makoviichuk, Viktor Makoviychuk, Philip H. S. Torr, Mingfei Sun, and Shimon Whiteson. Is independent learning all you need in the starcraft multi-agent challenge? *CoRR*, abs/2011.09533, 2020.

AIMS EPSRC Centre for Doctoral Training in Autonomous Intelligent Machines & Systems

UKRI Engineering and Physical Sciences Research Council

WhiRL Whiteson Research Lab