A Note on Truly Proximal Policy Optimization

Weichen Liao
Department of Computer Science
National Yang Ming Chiao Tung University
wcl.cs07@nycu.edu.tw

1 Introduction

Trust-region policy optimization (TRPO, Schulman et al. [2015]) and proximal policy optimization (PPO, Schulman et al. [2017]) are two of the most successful reinforcement learning algorithms. However, for some theoretical points, it still needs more research in order to improve its stability. This paper (Wang et al. [2020]) raises two questions to PPO: (1) Could PPO enforce a trust-region constraint? (2) Could PPO bound the likelihood ratio within the clipping range as it attempts to do? and the authors proposed three variant methods to handle these issues.

2 Problem Formulation

2.1 MDP Formulation and Notations

We consider the reinforcement learning framework in which a Markov Decision Process (MDP) is described by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{T}, c, \rho_1, \gamma)$.

- $\mathcal{S}$: state space
- $\mathcal{A}$: action space
- $\mathcal{T}$:$\mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$: transition probability function
- c: reward function
- ρ_1: distibution of initial state
- γ: discout factor

The performance of a policy π is defined as $\eta(\pi)$.

- $\eta(\pi) = \mathbb{E}_{s\sim\rho^\pi, a\sim\pi}[c(s,a)]$
- $\rho^\pi(s) = (1-\gamma)\sum_{t=1}^{\infty}\gamma^{t-1}\rho_t^\pi(s)$
- $\rho_t^\pi(s)$ is the density function of state at time t.

Surrogate performance objective fucntion:

$$L_{\pi_{\text{old}}}^{\text{PG}}(\pi) = \mathbb{E}_{s,a}[r_{s,a}^{\pi_{\text{old}}}(\pi)A_{s,a}^{\pi_{\text{old}}}] + \eta(\pi_{\text{old}}) \tag{1}$$

- $r_{s,a}^{\pi_{\text{old}}}(\pi) = \frac{\pi(a|s)}{\pi_{\text{old}}(a|s)}$: likelihood ratio between new policy π and old policy π_{old}
- $A_{s,a}^{\pi_{\text{old}}} = \mathbb{E}[R_t^\gamma | s_t = s, a_t = a; \pi_{\text{old}}] - E[R_t^\gamma | s_t = s; \pi_{\text{old}}]$

2.2 Trust Region Policy Optimization

The following theorem depicts the performance bound in Trust Region Policy Optimization (TRPO), which implies that maximizing $M(\pi)$ guarantee non-decreasing of the performance of the new policy π.

Theorem 1. *Let:*

- $C = (\max_{s,a} |A_{s,a}|) \times \frac{4\gamma}{(1-\gamma)^2}$
- $D_{KL}(\pi_{old}, \pi) = D_{KL}(\pi_{old}(\cdot|s)||\pi(\cdot|s))$
- $M(\pi) = L^{PG}(\pi) - C \max_{s \in \mathcal{S}} D^s_{KL}(\pi_{old}, \pi)$

, then we have:

- $\eta(\pi) \geq M(\pi)$
- $\eta(\pi_{old}) = M(\pi_{old})$

Besides, TRPO imposed a trust region-based constraint on the KL divergence:

$$\max_{\pi} L^{\text{PG}}(\pi) \; s.t. \; \max_{s \in \mathcal{S}} D^s_{\text{KL}}(\pi_{\text{old}}, \pi) \leq \delta \tag{2}$$

2.3 Proximal Policy Optimization

Proximal policy optimization (PPO) attempts to restrict the policy by a clipping function.

$$L^{\text{CLIP}}(\pi) = \mathbb{E}[\min r_{s,a}(\pi A_{s,a}, \mathcal{F}^{\text{FLIP}}(r_{s,a}(\pi), \epsilon) A_{s,a})] \tag{3}$$

$$\mathcal{F}^{\text{CLIP}}(r_{s,a}(\pi), \epsilon) = \begin{cases} 1-\epsilon, & r_{s,a}(\pi) \leq 1-\epsilon \\ 1+\epsilon, & r_{s,a}(\pi) \geq 1+\epsilon \\ r_{s,a}(\pi), & \text{otherwise} \end{cases} \tag{4}$$

, where ϵ is parameter of the clipping range ($0 < \epsilon < 1$)

For simplicity, we use t to denote the corresponding value for sample (s_t, a_t):

- $r_t(\pi_\theta) \triangleq r_{s_t,a_t}(\pi_\theta)$
- $A_t \triangleq A_{s_t,a_t}$

For more simplicity, we may use neural network parameter θ to represent π_θ alternatively. For example:

- $r_t(\theta) \triangleq r_t(\pi_\theta)$
- $L^{\text{CLIP}}(\theta) \triangleq L^{\text{CLIP}}(\pi_\theta)$
- $D^s_{\text{KL}}(\theta_{\text{old}}, \theta) \triangleq D^s_{\text{KL}}(\pi_{\theta_{\text{old}}}, \pi_\theta)$
- Let $\hat{L}^{\text{CLIP}}(\theta)$ denotes the empirical version of the objective function.

Eqation 3 and 4 can be combined and re-written as:

$$L^{\text{CLIP}}(\theta) = \begin{cases} (1-\epsilon)A_t, & r_t(\theta) \leq 1-\epsilon \text{ and } A_t < 0 \\ (1+\epsilon)A_t, & r_t(\theta) \geq 1+\epsilon \text{ and } A_t > 0 \\ r_t(\theta)A_t, & \text{otherwise} \end{cases} \tag{5}$$

2.4 Questions of the "Proximal" property of PPO

PPO attempts to restrict the policy by clipping the likelihood ratio between the new policy and the old one. However, does this clipping mechanism really restrict the policy ? The authors investigate the following 2 questions of PPO:

1. Could PPO enforce a trust region constraint?
2. Could PPO bound the likelihood ratio within the clipping range as it attempts to do?

The first one is wheither it enforces a trust region constraint, which is a theoretical indicator related to the performance guarantee (theorem 1). The second one is that whether PPO could bound the likelihood ratio as it attempts to do.

2.4.1 Question 1: Could PPO enforce a trust region constraint?

PPO does not explicitly attempt to impose a trust region constraint. In the authors' previous workWang et al. [2019b], they revealed that a different scale of the clipping range can affect the scale of the KL divergence. Even the likelihood ratio $rt(\theta)$ is bounded, the corresponding KL divergence is not necessarily bounded. The following theorem shows this property.

Theorem 2. *Assume that for discrete action space tasks where $|\mathcal{A}| \geq 3$, the policy is parametrized by $\pi_\theta(s_t) = p_t \in \mathbb{R}^{+|\mathcal{A}|}$, where $\sum_d^{|\mathcal{A}|} p_t^{(d)} = 1$; for continous action space tasks, the policy is parametrized by $\pi_\theta(a|s_t) = \mathcal{N}(a|\mu_t, \Sigma_t)$. Let $\Theta = \theta|1-\epsilon \leq r_t(\theta) \leq 1+\epsilon$. We have $\max_{\theta\in\Theta} D_{KL}^{s_t}(\theta_{old}, \theta) = +\infty$ for both discrete and continous action space tasks.*

2.4.2 Question 2: Could PPO bound the likelihood ratio within the clipping range as it attempts to do?

PPO generates an effect of preventing the likelihood ratio from exceeding the clipping range too much, but it could not strictly bound the likelihood ratio. Let's see the following theorm:

Theorem 3. *Given θ_0 that $r_t(\theta_0)$ satisfies the clipping condition (either $r_t(\theta) \leq 1-\epsilon$ and $A_t < 0$ or $r_t(\theta) \geq 1+\epsilon$ and $A_t > 0$). Let $\nabla \hat{L}^{CLIP}$ denote the gradient of $\hat{L}^{CLIP}$ at θ_0, and similarly $\nabla r_t(\theta_0)$. Let $\theta_1 = \theta_0 + \beta \nabla \hat{L}^{CLIP}(\theta_0)$, where β is the step size. If*

$$\langle \nabla \hat{L}^{CLIP}(\theta_0), \nabla r_t(\theta_0) A_t \rangle > 0 \tag{6}$$

then there exists some $\bar{\beta} > 0$ s.t. for any $\beta \in (0, \bar{\beta})$, we have

$$|r_t(\theta_1) - 1| > |r_t(\theta_0) - 1| > \epsilon \tag{7}$$

Proof. Let $\phi(\beta) = r_t(\theta_0 + \beta \nabla \hat{L}^{\text{CLIP}}(\theta_0)) = r_t(\theta_1)$.

$$\begin{aligned}
\phi'(\beta)|_{\beta=0} &= (\nabla_{\theta_1} r_t(\theta_1))^\intercal (\frac{\partial \theta_1}{\partial \beta})|_{\beta=0} \\
&= (\nabla_{\theta_1} r_t(\theta_1))^\intercal (\nabla_{\theta_0} \hat{L}^{\text{CLIP}}(\theta_0))|_{\beta=0} \\
&= (\nabla_{\theta_0} r_t(\theta_0))^\intercal (\nabla_{\theta_0} \hat{L}^{\text{CLIP}}(\theta_0))|_{\beta=0} \\
&= \langle \nabla r_t(\theta_0), \nabla \hat{L}^{\text{CLIP}}(\theta_0) \rangle
\end{aligned} \tag{8}$$

For the case of clipping condition: $r_t(\theta) \geq 1+\epsilon$ and $A_t > 0$,

$$\begin{aligned}
&\Rightarrow \phi'(0) > 0 \\
&\Rightarrow \text{Exits } \bar{\beta} > 0 \text{ s.t.}
\end{aligned} \tag{9}$$

$$\begin{aligned}
\text{for any } \beta \in (0, \bar{\beta}), &\phi(\beta) > \phi(0) \\
&\Rightarrow r_t(\theta_1) > r_t(\theta_0) \geq 1+\epsilon \\
&\Rightarrow |r_t(\theta_1) - 1| > |r_t(\theta_0) - 1|
\end{aligned} \tag{10}$$

For the other case of clipping condition: $r_t(\theta) \leq 1-\epsilon$ and $A_t < 0$, we can use similarly way to proof.

■

From equation 6, we can see that the condition requires the two gradients to be similar in direction to each other. This theorem implies, even the likelihood ratio $rt(\theta_0)$ is already out of the clipping range, it could go farther beyond the range (eq. 7).

3 Theoretical Analysis

3.1 Trust Region-based PPO (TR-PPO)

Method 1 (Trust Region-based PPO)**.**

$$\mathcal{F}^{TR}(r_{s,a}(\pi), \delta) = \begin{cases} r_{s,a}(\pi_{old}) & D^{s}_{KL}(\pi_{old}, \pi) \geq \delta \\ r_{s,a}(\pi) & otherwise \end{cases} \tag{11}$$

, where δ is a parameter, $r_{s,a}(\pi_{old}) = 1$.

It proposed a new region-based (by KL-divergence) verion of clipping function. It seems to improve stability since bounding KL-divergence is a theoretical indicator on the performance guarantee. However, the paper didn't provide any rigorous explanation about this part, so whether it's always useful is doubtful. From the experiments in this paper, we could see it's better than purely PPO but in some tasks it's not that powerful and even worse.

3.2 PPO with Rollback (PPO-RB)

In 2.4.2, we discussed a doubt about PPO: "Could PPO bound the likelihood ratio within the clipping range as it attempts to do?" In this part, we know that even though the r is beyond the clipping range, it still has chance to get next r even worse. The authors propose a rollback mechanism, which allowd L to rollback when it goes out the clipping range.

Method 2 (PPO with Rollback)**.**

$$\mathcal{F}^{RB}(r_{s,a}(\pi), \epsilon, \alpha) = \begin{cases} -\alpha r_{s,a}(\pi) + (1+\alpha)(1-\epsilon), & r_{s,a}(\pi) \leq 1-\epsilon \\ -\alpha r_{s,a}(\pi) + (1+\alpha)(1+\epsilon), & r_{s,a}(\pi) \geq 1+\epsilon \\ r_{s,a}(\pi), & otherwise \end{cases} \tag{12}$$

, where $\alpha > 0$ is a hyper-parameter.

We can re-wirte L^{RB} as:

$$L^{\text{RB}}(\theta) = \begin{cases} (-\alpha r_{s,a}(\theta) + (1+\alpha)(1-\epsilon))A_t, & r_t(\theta) \leq 1-\epsilon \text{ and } A_t < 0 \\ (-\alpha r_{s,a}(\theta) + (1+\alpha)(1+\epsilon))A_t, & r_t(\theta) \geq 1+\epsilon \text{ and } A_t > 0 \\ r_t(\theta)A_t, & \text{otherwise} \end{cases} \tag{13}$$

The following theorem shows that after adding rollback mechanism, its ability in preventing the out-of-the-range ratio can be improved.

Theorem 4. *Given θ_0 that $r_t(\theta_0)$, let $\theta_1^{CLIP} = \theta_0 + \beta\nabla\hat{L}^{CLIP}(\theta_0)$, $\theta_1^{RB} = \theta_0 + \beta\nabla\hat{L}^{RB}(\theta_0)$. The set of indices of the samples which satisfy the clipping condition is denoted as $\Omega = \{t | 1 \leq t \leq T, |r_t(\theta_0)| - 1 \geq \epsilon$, and $r_t(\theta_0)A_t \geq r_t(\theta_{old})A_t\}$. If $t \in \Omega$ and $r_t(\theta_0)$ satisfies $\sum_{t' \in \Omega} \langle \nabla r_t(\theta_0), \nabla r_{t'}(\theta_0) \rangle A_t A_{t'} > 0$, then for any $\beta \in (0, \beta)$, we have:*

$$|r_t(\theta_1^{RB}) - 1| < |r_t(\theta_1^{CLIP}) - 1| \tag{14}$$

Proof. Let $\phi(\beta) = r_t(\theta_0 + \beta\nabla\hat{L}^{\text{RB}}(\theta_0)) - r_t(\theta_0 + \beta\nabla\hat{L}^{\text{CLIP}}(\theta_0)) = r_t(\theta_1)$. Similarly to the proof of 3,

$$\begin{aligned} \phi'(\beta)|_{\beta=0} &= (\nabla_{\theta_0} r_t(\theta_0))^\intercal (\nabla_{\theta_0} \hat{L}^{\text{RB}}(\theta_0))|_{\beta=0} - (\nabla_{\theta_0} r_t(\theta_0))^\intercal (\nabla_{\theta_0} \hat{L}^{\text{CLIP}}(\theta_0))|_{\beta=0} \\ &= (\nabla r_t(\theta_0))^\intercal (\nabla \hat{L}^{\text{RB}}(\theta_0) - \nabla \hat{L}^{\text{CLIP}}(\theta_0)) \\ &= -\alpha \sum_{t' \in \Omega} \langle \nabla r_t(\theta_0), r_{t'}(\theta_0) \rangle A_{t'} \end{aligned} \tag{15}$$

For the case of clipping condition: $r_t(\theta) \geq 1+\epsilon$ and $A_t < 0$,

$$\begin{aligned} &\Rightarrow \phi'(0) < 0 \\ &\Rightarrow \text{Exits } \bar{\beta} > 0 \text{ s.t.} \end{aligned} \tag{16}$$

$$\begin{aligned}\text{for any } \beta \in (0, \bar{\beta}), \phi(\beta) < \phi(0) \\ &\Rightarrow r_t(\theta_1^{\text{RB}}) < r_t(\theta_1^{\text{CLIP}}) \\ &\Rightarrow |r_t(\theta_1^{\text{RB}}) - 1| < |r_t(\theta_1^{\text{CLIP}}) - 1|\end{aligned} \tag{17}$$

For the other case of clipping condition: $r_t(\theta) \leq 1 - \epsilon$ and $A_t < 0$, we can use similarly way to proof. ■

This shows the improvement after adding the rollback functionin. Ω indicates the indices of samples that are beyond the clipping range but are really improved.

This method also remind us that a proper tuning of step size is indeed signicant. Not all step size gurantee the improvement by this trick, and it didn't show how to well derive the step size. Therefore, these results just show it "may" improve and seems to be more stable.

In addition, extra hyper-parameter α is introduced here. The new variable α for tunning the force of rollback and the step size β tell us tuning hyper-parameter is still a essential issue in reinforcement learning.

The authors said if α is idealy sufficiently large, then large, then the new policy are guaranteed to be confined within the clipping range. But, is it practical to pick up an extremly large α? I saw the experiments uses small α such as 0.02 and 0.3. It's interesting.

3.3 Combining TR-PPO with Rollback (Truly PPO)

The trust region-based clipping still possibly suffers from the unbounded KL divergence problem, since it is not enforced any negative incentive when the policy is out of the trust region. Thus the authors integrate the trust region-based clipping with the rollback operation on KL divergence. The new method is named Truly PPO. It uses the "case" form to formulte L rather than design an extra clipping function. Besides, the other difference between this and the previous two methods, it uses rollback operation on KL divergence instead of likelihood ratio.

Method 3 (Truly PPO)**.**

$$L_{s,a}^{truly}(\pi) = r_{s,a}(\pi)A_{s,a} - \begin{cases} \alpha D_{KL}^{s}(\pi_{old}, \pi), \textit{ if } D_{KL}^{s}(\pi_{old}, \pi) > \delta \textit{ and } r_{s,a}(\pi)A_{s,a} \geq r_{s,a}(\pi_{old})A_{s,a} \\ \delta, \quad \textit{otherwise} \end{cases} \tag{18}$$

, where $\alpha > 0$ is a hyper-parameter.

To analyse the monotonic improvement property, we use the below maximum KL divergence instead. Such objective function also has the theoretical property of the guaranteed monotonic improvement.

$$L_{s,a}^{\text{truly}}(\pi) = r_{s,a}(\pi)A_{s,a} - \begin{cases} \alpha \max_{s' \in \mathcal{S}} D_{\text{KL}}^{s'}(\pi_{\text{old}}, \pi), \text{ if } \max_{s' \in \mathcal{S}} D_{\text{KL}}^{s}(\pi_{\text{old}}, \pi) \geq \delta \text{ and} \\ \exists a', r_{s,a'}(\pi)A_{s,a'} \geq r_{s,a'}(\pi_{\text{old}})A_{s,a'} \\ \delta, \quad \text{otherwise} \end{cases} \tag{19}$$

Theorem 5. *Let $\pi_{new}^{truly} = argmax_{\pi} L^{truly}(\pi)$, $\pi_{new}^{TRPO} = argmax_{\pi} L^{truly}(\pi)$. If $\alpha = C \triangleq \max_{s,a} |A_{s,a}| 4\gamma/(1-\gamma)^2$, and $\delta \leq \max_{s \in \mathcal{S}} D_{KL}^{s}(\pi_{old}, \pi_{new}^{TRPO})$, then $\eta(\pi_{new}^{truly}) > \eta(\pi_{old})$*

Proof.

$$\begin{aligned} L^{\text{truly}}(\pi') + \eta(\pi_{\text{old}}) &= L^{\text{PG}}(\pi') + \alpha \max_{s \in \mathcal{S}} (\pi_t extold, \pi') \\ &\leq L^{\text{PG}}(\pi_{\text{new}}^{\text{TRPO}}) - \alpha \max_{s' \in \mathcal{S}} D_{\text{KL}}^{s}(\pi_{\text{old}}, \pi_{\text{new}}^{TRPO}) \\ &= L^{\text{truly}}(\pi_{\text{new}}^{\text{TRPO}}) + \eta(\pi_{\text{old}}) \end{aligned} \tag{20}$$

By theorem 1, we have $\eta(\pi_{\text{new}}^{truly}) = eta(\pi_{\text{new}}^{\text{TRPO}}) \geq M(\pi_{\text{new}}^{\text{TRPO}}) = M(\pi_{\text{old}}) \geq \eta(\pi_{\text{old}})$. ■

4 Conclusion

This paper discusses two questions of PPO theoretically, and it provides three variant methods. These methods mitigate the problems and improve the stability to some extent. To sum up, the methods proposed in this paper has their value.

References

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR, 2015.

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

Yuhui Wang, Hao He, and Xiaoyang Tan. Truly proximal policy optimization. In *Uncertainty in Artificial Intelligence*, pages 113–122. PMLR, 2020.

Yuhui Wang, Hao He, Xiaoyang Tan, and Yaozhong Gan. Trust region-guided proximal policy optimization. In *Advances in Neural Information Processing Systems*, 2019b.