A New Horizon on PPO Algorithm

Bowen Xu, Jun Zhang, Yuxiang Ying

Shanghaitech

December 19, 2022

Outline

1. PPO Review
   - TRPO
   - PPO
2. Motivation and Process
   - Modify the critic function
   - Use mean to estimate the Value function
3. Algorithm
   - Idea: Penalty Method
   - Algorithm Design
4. Evaluation
   - Environment
   - Improvements Experienced

Outline

1. PPO Review
   - TRPO
   - PPO
2. Motivation and Process
   - Modify the critic function
   - Use mean to estimate the Value function
3. Algorithm
   - Idea: Penalty Method
   - Algorithm Design
4. Evaluation
   - Environment
   - Improvements Experienced

PPO Review - TRPO

- The initial problem we want to solve is to maximize the expected discounted reward:

$$\eta(\pi) = \mathop{\mathbb{E}}_{s_0, a_0, \cdots \sim \tilde{\pi}} \left[\sum_{t=0}^{\infty} \gamma^t r_t \right].$$

PPO Review - TRPO

- The initial problem we want to solve is to maximize the expected discounted reward:

$$\eta(\pi) = \mathop{\mathbb{E}}_{s_0, a_0, \cdots \sim \tilde{\pi}} \left[\sum_{t=0}^{\infty} \gamma^t r_t \right].$$

- The following useful identity expresses the expected return of another policy $\tilde{\pi}$ in terms of the advantage over π:

$$\eta(\tilde{\pi}) = \eta(\pi) + \mathop{\mathbb{E}}_{s_0, a_0, \cdots \sim \tilde{\pi}} \left[\sum_{t=0}^{\infty} \gamma^t \mathcal{A}_\pi(s_t, a_t) \right].$$

  - $\mathcal{A}_\pi(s_t, a_t)$ is the advantage function:

$$\mathcal{A}_\pi(s_t, a_t) = \mathcal{Q}_\pi(s_t, a_t) - \mathcal{V}_\pi(s_t).$$

PPO Review - TRPO

- Rewrite the equation and we will get the following:

$$\eta(\tilde{\pi}) = \eta(\pi) + \sum_{s} \rho_{\tilde{\pi}}(s) \sum_{a} \tilde{\pi} \mathcal{A}_{\pi}(s, a).$$

PPO Review - TRPO

- Rewrite the equation and we will get the following:

$$\eta(\tilde{\pi}) = \eta(\pi) + \sum_{s} \rho_{\tilde{\pi}}(s) \sum_{a} \tilde{\pi} \mathcal{A}_{\pi}(s, a).$$

- In TRPO (Trust Region Policy Optimization) algorithm, the problem is transformed into the following form:

$$\begin{aligned} \underset{\theta}{\text{maximize}} \quad & \mathbb{E}_{s \sim \rho_{\theta_{\text{old}}}, a \sim q}\left[\frac{\pi_\theta(a|s)}{q(a|s)} \mathcal{A}_{\theta_{\text{old}}}(s, a)\right] \\ \text{subject to} \quad & \mathbb{E}_{s \sim \rho_{\theta_{\text{old}}}}\left[D_{\text{KL}}\left(\pi_{\theta_{\text{old}}}(\cdot|s) \,||\, \pi_\theta(\cdot|s)\right)\right] \leq \delta. \end{aligned}$$

  - ρ_θ is the stationary distributions of states under policy θ.

PPO review - PPO

- Let $r_t(\theta)$ denote the ratio $r_t(\theta) = \frac{\pi_\theta(a|s)}{q(a|s)}$.

PPO review - PPO

- Let $r_t(\theta)$ denote the ratio $r_t(\theta) = \frac{\pi_\theta(a|s)}{q(a|s)}$.
- Then the objective function in TRPO becomes

$$L(\theta) = \mathbb{E}_t \left[r_t(\theta) \mathcal{A}_{\theta_{\text{old}}}(t) \right].$$

PPO review - PPO

- Let $r_t(\theta)$ denote the ratio $r_t(\theta) = \frac{\pi_\theta(a|s)}{q(a|s)}$.
- Then the objective function in TRPO becomes

$$L(\theta) = \mathbb{E}_t \left[r_t(\theta) \mathcal{A}_{\theta_{\text{old}}}(t) \right].$$

- In PPO (Proximal Policy Optimization) algorithm, the objective is modified as:

$$L^{CLIP}(\theta) = \mathbb{E}_t \left[\min(r_t(\theta) \mathcal{A}_{\theta_{\text{old}}}(t), \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon) \mathcal{A}_{\theta_{\text{old}}}(t)) \right].$$

to penalize changes to the policy that move $r_t(\theta)$ away from 1.

PPO review - PPO, AC Style

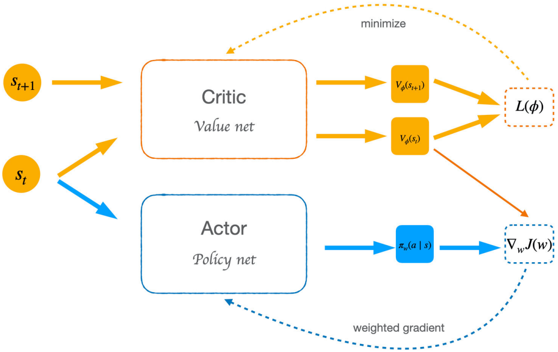
Algorithm 1 PPO, Actor-Critic Style

```
for iteration = 1, 2, ... do
    for iteration = 1, 2, ..., N do
        Run policy π_{θ_old} in environment for T timesteps
        Compute advantage estimates Â_1, ..., Â_T
    end for
    Optimize surrogate L wrt θ, with K epochs and minibatch size M ≤ NT
    θ_old ← θ
end for
```

Outline

1. PPO Review
   - TRPO
   - PPO
2. Motivation and Process
   - Modify the critic function
   - Use mean to estimate the Value function
3. Algorithm
   - Idea: Penalty Method
   - Algorithm Design
4. Evaluation
   - Environment
   - Improvements Experienced

Motivation 1 - Modify the critic function



Motivation 1 - Modify the critic function

- Update Actor:

$$\underset{\theta}{\text{maximize}} \quad \mathbb{E}_t \left[\min(r_t(\theta)\mathcal{A}_{\theta_{\text{old}}}(t), \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\mathcal{A}_{\theta_{\text{old}}}(t))\right],$$

where $\mathcal{A}_{\theta_{old}}(t) = \mathcal{Q}_{\theta_{old}}(s_t, a_t) - V_{w_{old}}(s_t)$.

- Update Critic:

$$\underset{w}{\text{minimize}} \quad \mathbb{E}_{s\sim\rho_{\theta_{old}}}\left[\left(V_{\theta_{old}}(s) - V_w(s)\right)^2\right].$$

Motivation 1 - Modify the critic function

$$\begin{aligned} \underset{\theta, w}{\text{maximize}} \quad & \mathbb{E}_t \left[\min(r_t(\theta)\mathcal{A}_{\theta_{\text{old}}}(t), \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\mathcal{A}_{\theta_{\text{old}}}(t))\right] \\ \text{subject to} \quad & \mathbb{E}_{s\sim\rho_{\theta_{old}}}\left[(V_{\theta_{old}}(s) - V_w(s))^2\right] = 0, \end{aligned}$$

$$\text{where } \mathcal{A}_{\theta_{old}}(t) = \mathcal{Q}_{\theta_{old}}(s_t, a_t) - V_w(s_t).$$

Motivation 2 - Use mean to estimate the Value function

- Sample-Based Estimation

$$Q_{\theta_{old}}(s_t, a_t) = \sum_{t'=t}^{\infty} \gamma^{t'-t} r_{t'},$$

$$V_{\theta_{old}}(s_t) = \sum_{t'=t}^{\infty} \gamma^{t'-t} r_{t'}.$$

Motivation 2 - Use mean to estimate the Value function

$$\begin{aligned} \underset{\theta, w}{\text{maximize}} \quad & \mathbb{E}_t \left[\min(r_t(\theta)\mathcal{A}_{\theta_{\text{old}}}(t), \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\mathcal{A}_{\theta_{\text{old}}}(t))\right] \\ \text{subject to} \quad & \mathbb{E}_{s\sim\rho_{\theta_{old}}, a\sim q}\left[(\mathcal{Q}_{\theta_{old}}(s,a) - \mathcal{Q}_w(s,a))^2\right] = 0, \end{aligned}$$

$$\text{where } \mathcal{A}_{\theta_{old}}(t) = \mathcal{Q}_{\theta_{old}}(s_t, a_t) - \mathbb{E}_{a\sim q}[\mathcal{Q}_w(s_t, a_t)].$$

Outline

Penalty Method

Reformulate the constrained problem:

$$\begin{aligned} max_x \quad & f(x) \\ s.t. \quad & c(x) = 0 \end{aligned}$$

as the unconstrained quadratic penalty subproblem:

$$max_x \quad f(x) - \lambda ||c(x)||_2^2$$

where λ is the penalty parameter.

Penalty Method

What can Penalty Method Do?

- Avoid difficulties caused by constraints .
- Balance objective maximization and constraint violation.

Main Innovation

- Maximize objective function (Combine Actor and Critic):

$$L^{CLIP}(\theta, w) - \lambda \mathbb{E}_{s\sim\rho_{\theta_{old}}, a\sim q}\left[\left(\mathcal{Q}_{\theta_{old}}(s,a) - \mathcal{Q}_w(s,a)\right)^2\right]$$

- Change Critic Networks to approximate Q value and sample to approximate V value.

Main innovation

Algorithm 2

for iteration $= 1, 2, \ldots$ **do**
  for iteration $= 1, 2, \ldots, N$ **do**
    Run policy $\pi_{\theta_{\text{old}}}$ in environment for T timesteps
    Compute state value estimates $\hat{V}(s_1), \ldots, \hat{V}(s_T)$
    Compute advantage estimates $\hat{\mathcal{A}}_1, \ldots, \hat{\mathcal{A}}_T$
  end for
  Optimize objective function wrt θ, w, with K epochs and minibatch size $M \leq NT$
  $\theta_{\text{old}} \leftarrow \theta$
  $w_{\text{old}} \leftarrow w$
end for

Outline

Pendulum

Figure: pendulum

- This is a pendulum swingup problem. The system consists of a pendulum attached at one end to a fixed point, and the other end being free. The pendulum starts in a random position and the goal is to apply torque on the free end to swing it into an upright position.

Pendulum

Figure: pendulum

- This is a pendulum swingup problem. The system consists of a pendulum attached at one end to a fixed point, and the other end being free. The pendulum starts in a random position and the goal is to apply torque on the free end to swing it into an upright position.
- Rewards:
  $r = -(\omega^2 + 0.1 * \omega_{dt}^2 + 0.001 * \text{torque}^2)$

Improvements Experienced

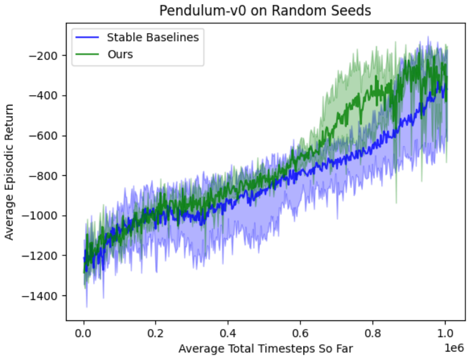


Figure: Comparison of stable baseline PPO

Improvements Experienced

- Where our model outperforms the PPO model:
  - Better highest rewards
    - Our model: about -190.
    - Baselines: about -370.
  - Converges faster to suboptimal
    - Our model reaches -400 in about $0.8 \times 1e6$ step.
    - Baseline reaches -400 in about $1.0 \times 1e6$ step.

Thank you

Thank you!

Q&A

Q&A