PPO

- 기존에 샘플 데이터를 조금쓰고 버리는 방식에서 여러번 사용하게 만들었다.
  - On-Policy 알고리즘 이지만 데이터를 한번만쓰고 버리지 않는다
- PPO는 TRPO를 발전시켜모델이다.

TRPO

- 기존의 Policy Gradient는 $\hat{\mathbb{E}}_t[\nabla_\theta \log \pi_\theta(a_t|s_t)\hat{A}_t]$ 값으로 정책을 업데이트 했다. (Â = estimate 측정)
- 기존의 Loss 또한 $\hat{\mathbb{E}}_t[\log \pi_\theta(a_t|s_t)\hat{A}_t]$ 값으로 Loss를 만들었다.
- 만약 이전θ값과 oldθ의 차이가 심하지 않다면 Loss를 $\hat{\mathbb{E}}_t\left[\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}\hat{A}_t\right]$로 만들수있다. (Importance sampling 방법이다.)

Chain rule을 이용해서 미분식을 만들고 미분을 뺀다.

$$\nabla_\theta \log \pi_\theta(a_t|s_t)|_{\theta_{old}} = \frac{\nabla_\theta \pi_\theta(a_t|s_t)|_{\theta_{old}}}{\pi_{\theta_{old}}(a_t|s_t)} = \nabla_\theta \left(\frac{f(\theta)}{f(\theta_{old})}\right)\Big|_{\theta_{old}}$$

- 위의 조건을 만족하도록 식을 작성하면 아래와같다.

$$\underset{\theta}{\text{maximize}} \quad \hat{\mathbb{E}}_t\left[\frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{\text{old}}}(a_t \mid s_t)}\hat{A}_t\right]$$
$$\text{subject to} \quad \hat{\mathbb{E}}_t[\mathrm{KL}[\pi_{\theta_{\text{old}}}(\cdot \mid s_t), \pi_\theta(\cdot \mid s_t)]] \le \delta.$$

식을 하나로 만들면 아래와 같다.

$$\underset{\theta}{\text{maximize}} \quad \hat{\mathbb{E}}_t\left[\frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{\text{old}}}(a_t \mid s_t)}\hat{A}_t\right] - \beta\hat{\mathbb{E}}_t[\mathrm{KL}[\pi_{\theta_{\text{old}}}(\cdot \mid s_t), \pi_\theta(\cdot \mid s_t)]]$$

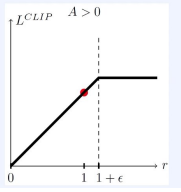
PPO

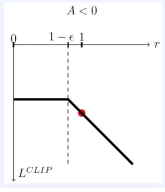
- TRPO에서 변형된 형태이다.
- 2가지 방식중에 clip 방식이 성능이 좋다.
- clip방식의 Loss는 다음과 같다.
  - 기존에 original Loss에 clip 함수를 추가하여 최대 최소 값을 정한다.
  - 마지막에 2개중에 작은 값을 사용한다.

$$L^{CLIP}(\theta) = \hat{\mathbb{E}}_t\left[\min(r_t(\theta)\hat{A}_t, \mathrm{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t)\right]$$

Original Loss — Clipped Loss

Lower bound of unclipped objective = L^{CPI}





- 이 식은
여 고안