09

PPO 알고리즘

1. 기본개념

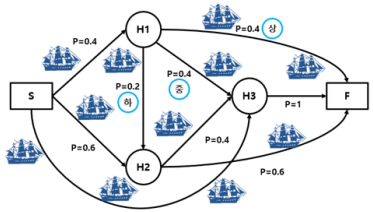
PPO 기본개념 중요도 샘플링

중요도 샘플링(Importance Sampling)

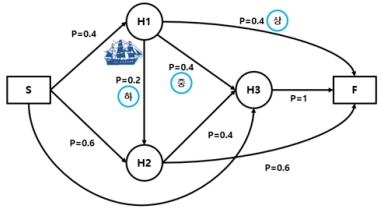
- f(x)의 기댓값을 계산하고자 하는 확률분포 p(x)를 알고 있지만 p에서 샘플을 생성하기 어려울 때, 비교적 샘플을 구하기 쉬운 확률분포 q(x)에서 샘플을 생성하여 확률분포 p(x)에서의 f(x)의 기댓값을 생성하는 것
- 중요도 샘플링을 활용해서 오프 폴리시에서 정책 간 연관성 문제를 해결한다.

$$\sum P(X)f(X) = \sum Q(X) \left[\frac{P(X)}{Q(X)} f(X) \right]$$

P(X)	어떤 환경에서 변수 X의 확률분포 P
Q(X)	다른 환경에서 변수 X의 확률분포 Q
f(X)	X의 함수 어떤 함수도 가능(sin, cos, 2x+1 등)
∑P(X)f(X)	변수X의 함수 f(X)에 대한 확률분포 P의 기댓값

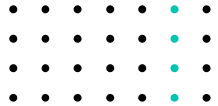


기존 항로 데이터



신규 항로

PPO 기본개념 중요도 샘플링



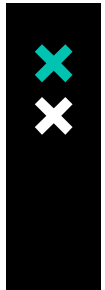
활용방법

- 강화학습에서는 $\pi_{\theta old}$ 정책을 가지고 수집한 데이터 X를 가지고 π_{θ} 정책을 가진 환경에서 기댓값을 구하기 위해서는 π_{θ} 정책의 기댓값을 구하는 공식에다가 각각의 정책 비율을 곱하면 된다

$$E_{X\sim\pi_\theta}[f(X)] = \sum \pi_\theta f(X)$$

$$= \sum \pi_{\theta old} \left[\frac{\pi_\theta}{\pi_{\theta old}} f(X) \right]$$

$$= E_{X\sim\pi_{\theta old}} \left[\frac{\pi_\theta}{\pi_{\theta old}} f(X) \right]$$



PPO 기본개념

A2C 오프 폴리시로 변경

Policy Gradient

$$\nabla_\theta J(\theta) = E_{\pi\theta}[\nabla_\theta log\pi_\theta(a|s)\, R_s^a] \quad ①$$

$$\fallingdotseq E_{\pi\theta}[\nabla_\theta log\pi_\theta(a|s)\, A_s^a] \quad ②$$

미분의 연쇄 법칙 (Chain rule)

$y = \log f(x)$

$y' = \frac{f(x)'}{f(x)}$

$$= E_{\pi\theta}[\frac{\nabla_\theta \pi_\theta(a|s)}{\pi_\theta(a|s)} A_s^a] \quad ③$$

$$= E_{\pi\theta old}[\frac{\cancel{\pi_\theta(a|s)}}{\pi_{\theta old}(a|s)} \frac{\nabla_\theta \pi_\theta(a|s)}{\cancel{\pi_\theta(a|s)}} A_s^a] \quad ④$$

$$= E_{\pi\theta old}[\frac{\nabla_\theta \pi_\theta(a|s)}{\pi_{\theta old}(a|s)} A_s^a] \quad ⑤$$

$$= E_{\pi\theta old}[\nabla_\theta(\frac{\pi_\theta(a|s)}{\pi_{\theta old}(a|s)} A_s^a)] \quad ⑥$$

Cost Function

$$= -\frac{\pi_\theta(a|s)}{\pi_{\theta old}(a|s)} A_s^a \quad ⑦$$

- 경험을 재활용하고 하나의 정책으로 많은 경험을 쌓아 인공신경망을 업데이트 하기 위해 오프 폴리시로 변경

PPO 기본개념 클리핑

빠른 학습속도의 문제점

이미지 출처: http://www.iautocar.co.kr

- 변수의 변화 속도를 제어하면서 학습 효율을 높일 수 있다.
- PPO에서는 학습 속도를 제어하기 위해 클리핑(Clipping) 기법을 사용한다

PPO 기본개념 클리핑

클리핑 개념

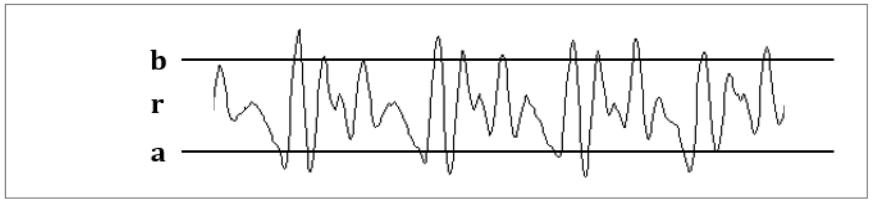
Clipping

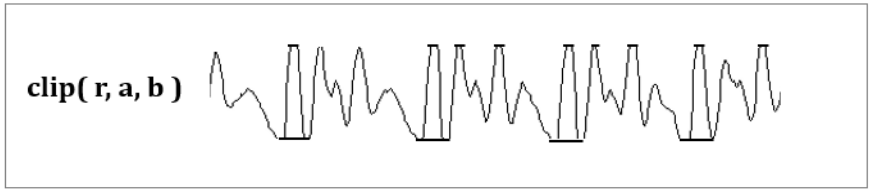
clip(r, a, b)

r < a 이면 a

r > b 이면 b

a ≤ r ≤ b 이면 r





- 클리핑은 데이터의 하한(Lower)과 상한(Upper)을 정해 놓고 입력되는 데이터를 일정 범위 안으로 들어오도록 만드는 기법이다

PPO 기본개념 클리핑

비용함수

- ε 을 통해 클리핑의 상한과 하한을 지정할 수 있다. 강화학습 과정에서 효율적인 ε 값을 선택하느냐는 효율적인 학습을 위한 아주 중요한 문제이다

$$r(\theta) = \frac{\pi_\theta(a|s)}{\pi_{\theta old}(a|s)} \quad ①$$

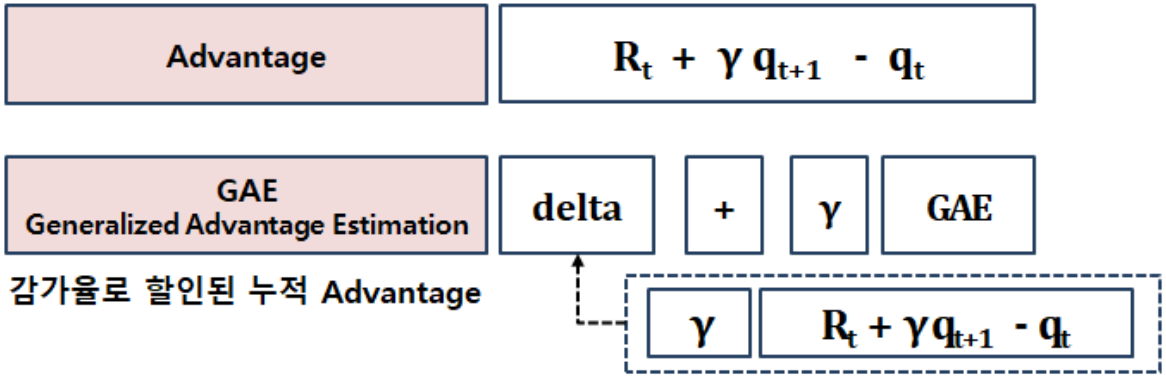
Cost Function

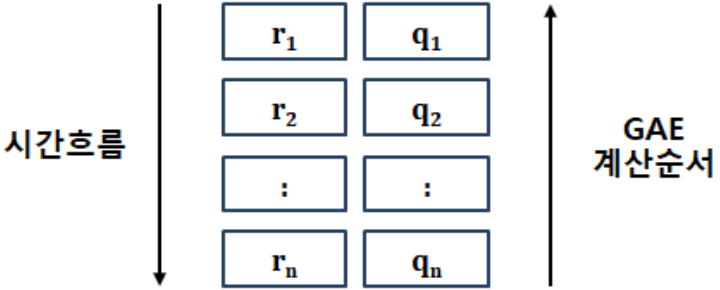
$$\min(\underbrace{r_t(\theta)\,A_t}_{②\text{-}1},\ \underbrace{clip(r_t(\theta), 1-\varepsilon, 1+\varepsilon)\,A_t}_{②\text{-}2}) \quad ②$$

Original Loss

Clipped Loss

PPO 기본개념 GAE





- Generalized Advantage Estimation
- GAE : 감가율로 할인된 누적 어드벤티지
- 에이전트가 경험을 쌓을 때 얻는 보상(r)과 큐함수(q)를 모두 시간 순으로 기록해 두었다가 어드벤티지를 계산할 때 사용