KL divergence, Clipping,

중요도 샘플링, GAE

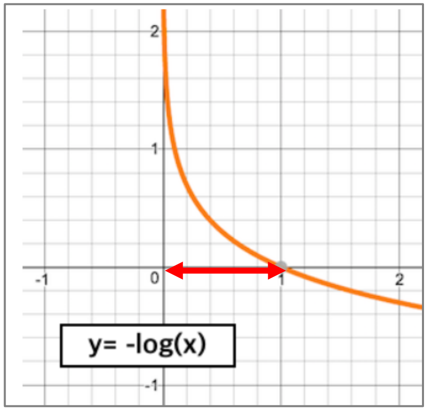
최우진

2022/02/08

Information theory

- 확률이 낮은 사건일수록 더욱 놀랍고 정보량이 크다.

$$h(x) = -log_2 P(x)$$



- Entropy: 랜덤 변수 x가 가질 수 있는 모든 값(사건)에 대해 정보량의 평균

$$H(x) = -\sum_{x} P(x) log_2 P(x)$$

$$= -\int_{-\infty}^{\infty} p(x) log_2 p(x)\, dx$$

Entropy 최대(균등분포): $\log_2 n$
Entropy 최소(한 경우만 1, 나머지 0인 경우):0

Kullback-Leibler divergence

- 두 확률분포의 차이를 계산하는 데 사용하는 함수

- 수식

$$KL(P\,||\,Q) = \sum_{i=0}^{n} p(x_i)\log_2\left(\frac{p(x_i)}{q(x_i)}\right) = \underbrace{\boxed{\sum_{i=0}^{n} p(x_i)\log_2 p(x_i)}}_{\text{Entropy}} - \underbrace{\boxed{\sum_{i=0}^{n} p(x_i)\log_2 q(x_i)}}_{\text{Cross - Entropy}}$$

Kullback-Leibler divergence 예제

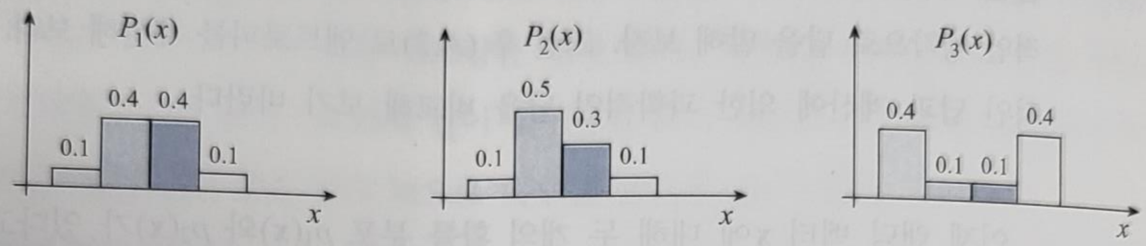


그림 A.2 서로 다른 세 확률 분포

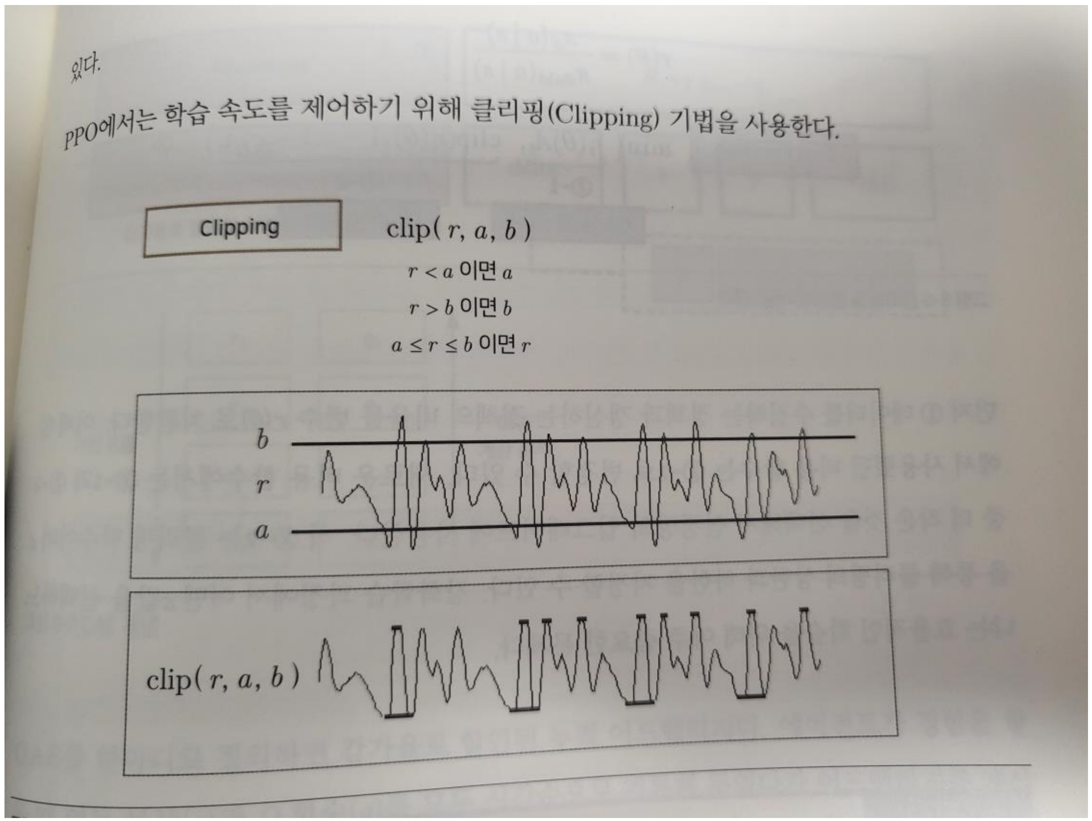
$$KL(P_1(\mathbf{x}), P_2(\mathbf{x})) = 0.1\log_2\frac{0.1}{0.1} + 0.4\log_2\frac{0.4}{0.5} + 0.4\log_2\frac{0.4}{0.3} + 0.1\log_2\frac{0.1}{0.1} = 0.037$$

$$KL(P_1(\mathbf{x}), P_3(\mathbf{x})) = 0.1\log_2\frac{0.1}{0.4} + 0.4\log_2\frac{0.4}{0.1} + 0.4\log_2\frac{0.4}{0.1} + 0.1\log_2\frac{0.1}{0.4} = 1.200$$

실제 KL 다이버전스를 계산해 본 결과 $P_1(x)$과 $P_2(x)$의 거리는 0.037로 매우 가깝고 $P_1(x)$와 $P_3(x)$는 1.200이라는 먼 거리를 가짐을 확인할 수 있다. $KL(P_2(x), P_1(x))$를 계산하여 $KL(P_1(x), P_2(x))$와 같은지 확인해 보자. ■■■

Clipping function

- **Clipping 기법**



있다.

PPO에서는 학습 속도를 제어하기 위해 클리핑(Clipping) 기법을 사용한다.

Importance Sampling

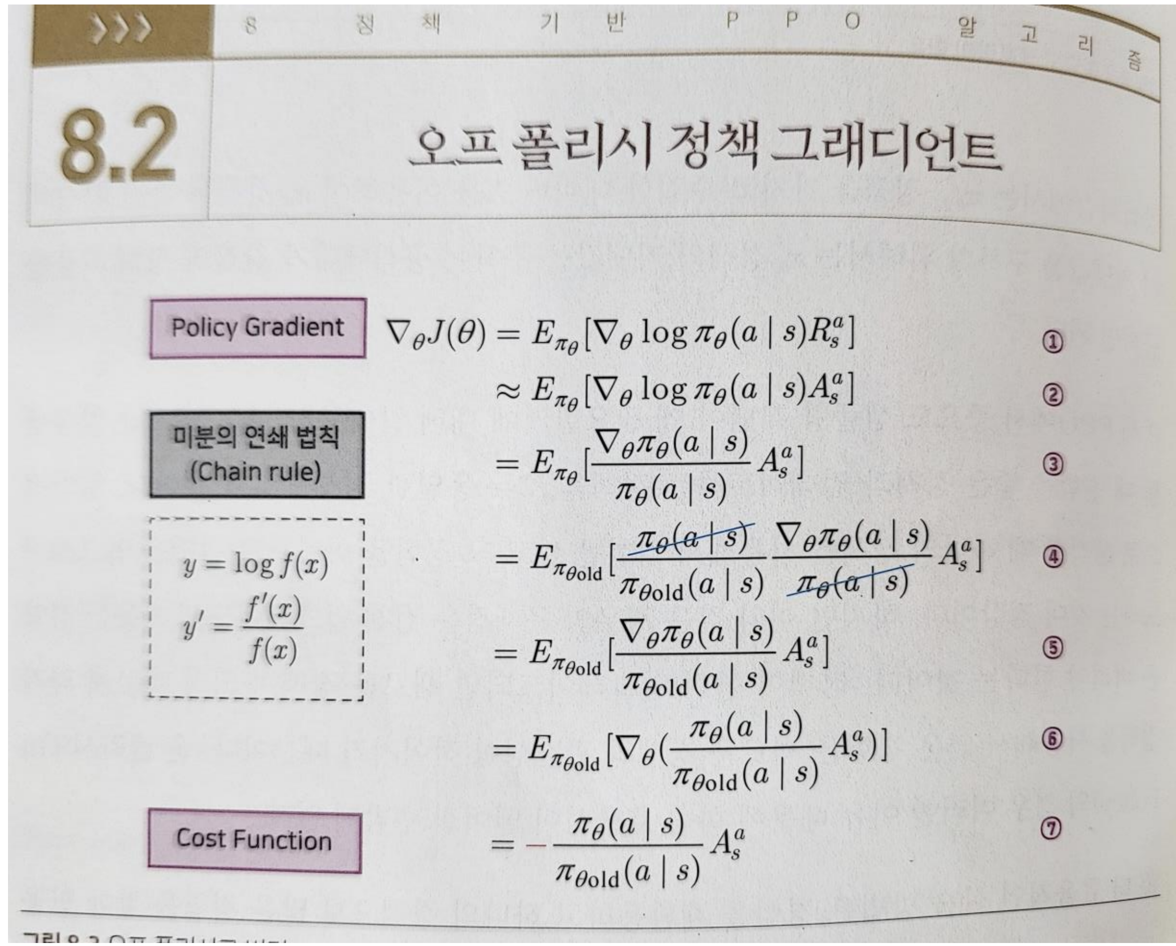
- **중요도 샘플링**

$$\mathbb{E}_{X\sim P}[f(X)] = \sum P(X)f(X)$$

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

$$= \mathbb{E}_{X\sim Q}\left[\frac{P(X)}{Q(X)}f(X)\right]$$

$$\approx \frac{1}{n}\sum_{i=1}^{n}\frac{P(x_i)}{Q(x_i)}f(x_i)$$



8 정책 기반 PPO 알고리즘

8.2 오프 폴리시 정책 그래디언트

Policy Gradient

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

$$\approx E_{\pi_\theta}[\nabla_\theta \log \pi_\theta(a \mid s) A_s^a] \quad ②$$

미분의 연쇄 법칙 (Chain rule)

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

$y = \log f(x)$

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

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

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

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

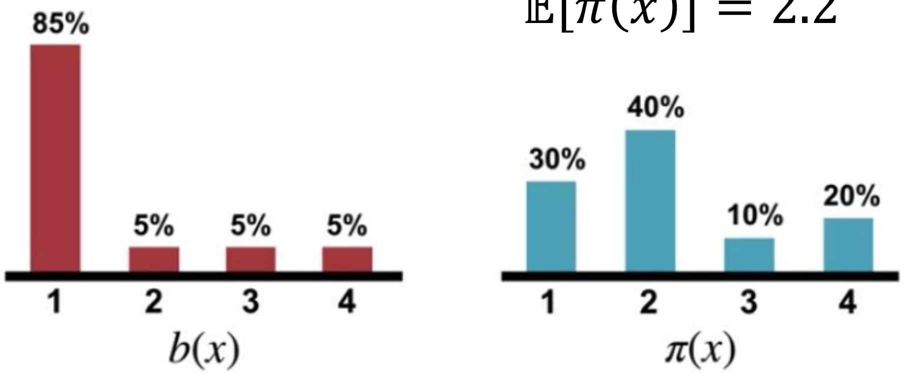
Cost Function

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

Importance Sampling

- 중요도 샘플링 예제

$$\mathbb{E}[\pi(x)] = 2.2$$



- 분포 b에서 샘플링 예시: x = [1, 3, 1]

$\pi(x)$: 가우시안 분포
b(x): 균등 분포

- b(x): [0.85, 0.05, 0.85]

- $\boldsymbol{\pi(x)}$: [0.3, 0.1, 0.3]

$$\mathbb{E}_{X\sim P}[f(X)] \approx \frac{1}{n}\sum_{i=1}^{n}\frac{P(x_i)}{Q(x_i)}f(x_i) = \frac{\left(1\times\frac{0.3}{0.85}\right)+\left(3\times\frac{0.1}{0.05}\right)+\left(1\times\frac{0.3}{0.85}\right)}{3} = 2.24$$

Generalized Advantage Estimation

- **Advantage를 비율로 곱해서 사용**
- **즉, GAE는 감가율로 할인된 누적 Advantage이다.**

- **TD(λ)**

- Consider the following n-step returns for $n = 1, 2, \infty$:

$$
\begin{array}{lll}
n = 1 & (TD) & G_t^{(1)} = R_{t+1} + \gamma V(S_{t+1}) \\
n = 2 & & G_t^{(2)} = R_{t+1} + \gamma R_{t+2} + \gamma^2 V(S_{t+2}) \\
\vdots & & \vdots \\
n = \infty & (MC) & G_t^{(\infty)} = R_{t+1} + \gamma R_{t+2} + \dots + \gamma^{T-1} R_T
\end{array}
$$

- Define the n-step return

$$G_t^{(n)} = R_{t+1} + \gamma R_{t+2} + \dots + \gamma^{n-1} R_{t+n} + \gamma^n V(S_{t+n})$$

- The λ-return G_t^λ combines all n-step returns $G_t^{(n)}$
- Using weight $(1-\lambda)\lambda^{n-1}$

$$G_t^\lambda = (1-\lambda) \sum_{n=1}^{\infty} \lambda^{n-1} G_t^{(n)}$$

Generalized Advantage Estimation

$$\hat{A}_t^{(1)} := \delta_t^V = -V(s_t) + r_t + \gamma V(s_{t+1}) \tag{11}$$

$$\hat{A}_t^{(2)} := \delta_t^V + \gamma\delta_{t+1}^V = -V(s_t) + r_t + \gamma r_{t+1} + \gamma^2 V(s_{t+2}) \tag{12}$$

$$\hat{A}_t^{(3)} := \delta_t^V + \gamma\delta_{t+1}^V + \gamma^2\delta_{t+2}^V = -V(s_t) + r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \gamma^3 V(s_{t+3}) \tag{13}$$

$$\hat{A}_t^{(k)} := \sum_{l=0}^{k-1} \gamma^l \delta_{t+l}^V = -V(s_t) + r_t + \gamma r_{t+1} + \cdots + \gamma^{k-1} r_{t+k-1} + \gamma^k V(s_{t+k}) \tag{14}$$

$$\begin{aligned}
\hat{A}_t^{\mathrm{GAE}(\gamma,\lambda)} &:= (1-\lambda)\left(\hat{A}_t^{(1)} + \lambda\hat{A}_t^{(2)} + \lambda^2\hat{A}_t^{(3)} + \ldots\right) \\
&= (1-\lambda)\left(\delta_t^V + \lambda(\delta_t^V + \gamma\delta_{t+1}^V) + \lambda^2(\delta_t^V + \gamma\delta_{t+1}^V + \gamma^2\delta_{t+2}^V) + \ldots\right) \\
&= (1-\lambda)(\delta_t^V(1+\lambda+\lambda^2+\ldots) + \gamma\delta_{t+1}^V(\lambda+\lambda^2+\lambda^3+\ldots) \\
&\qquad + \gamma^2\delta_{t+2}^V(\lambda^2+\lambda^3+\lambda^4+\ldots) + \ldots) \\
&= (1-\lambda)\left(\delta_t^V\left(\frac{1}{1-\lambda}\right) + \gamma\delta_{t+1}^V\left(\frac{\lambda}{1-\lambda}\right) + \gamma^2\delta_{t+2}^V\left(\frac{\lambda^2}{1-\lambda}\right) + \ldots\right) \\
&= \sum_{l=0}^{\infty}(\gamma\lambda)^l \delta_{t+l}^V
\end{aligned} \tag{16}$$

Reference

- Hands-On Reinforcement Learning for Games, Ch 9.
- 엔트로피 최대: https://ichi.pro/ko/jeong-gyu-bunpo-e-daehan-ihae-161890053490130
- 정보이론: https://hoya012.github.io/blog/cross_entropy_vs_kl_divergence/
- 패턴인식 부록
- 중요도 샘플링: https://jyoondev.tistory.com/150
- 중요도 샘플링: https://pasus.tistory.com/52

KL divergence, Clipping,

중요도 샘플링, GAE

감사합니다