Derivation of Likelihood Ratio Policy Gradient Step

Monday, December 28, 2020 10:13 PM

$$\nabla_\theta U(\theta) \approx \hat{g} = \frac{1}{m}\sum_{i=1}^{m}\sum_{t=0}^{H}\nabla_\theta \log \pi_\theta(a_t^{(i)} \mid s_t^{(i)}) R(\tau^{(i)})$$

Likelihood Ratio Policy Gradient

Maximize the return w/ respect to θ $\longrightarrow$ (Probability of Traj. x Reward of Traj.)

$$\nabla_\theta U(\theta) = \nabla_\theta \sum_\tau P(\tau;\theta) R(\tau) \quad (1)$$

$$= \sum_\tau \nabla_\theta P(\tau;\theta) R(\tau) \quad (2)$$

$$= \sum_\tau \frac{P(\tau;\theta)}{P(\tau;\theta)} \nabla_\theta P(\tau;\theta) R(\tau) \quad (3)$$

$$= \sum_\tau P(\tau;\theta) \frac{\nabla_\theta P(\tau;\theta)}{P(\tau;\theta)} R(\tau) \quad (4)$$

$$= \sum_\tau P(\tau;\theta) \nabla_\theta \log P(\tau;\theta) R(\tau) \quad (5)$$

↳ Recall $\nabla_x \log f(x) = \frac{\nabla_x f(x)}{f(x)}$ thus

$$\nabla_\theta \log P(\tau;\theta) = \frac{\nabla_\theta P(\tau;\theta)}{P(\tau;\theta)}$$ (likelihood ratio trick)

Simple-Based Estimate:

$$\nabla_\theta U(\theta) = \sum_\tau P(\tau;\theta) \underbrace{\nabla_\theta \log P(\tau;\theta)} R(\tau)$$

↳

$$\nabla_\theta \log P(\tau;\theta)$$ ↘ expand

$$= \nabla_\theta \log \left[\prod^{H} P(s_{t+1}^{(i)} \mid s_t^{(i)}, a_t^{(i)}) \, \pi_\theta(a_t^{(i)} \mid s_t^{(i)}) \right] \quad (1)$$

(Probability of Trajectory)

$$= \nabla_\theta \left[\sum^{H} \log P(s \ldots \right.$$

$$\nabla_\theta \left[\sum_{t=0} \log P(s_{t+1}^{(i)} \mid s_t^{(i)}, a_t^{(i)}) + \sum_{t=0} \log \pi_\theta (a_t^{(i)} \mid s_t^{(i)}) \right] \quad (2)$$

(sum of logs)

$$= \underbrace{\nabla_\theta \sum_{t=0}^{H} \log P(s_{t+1}^{(i)} \mid s_t^{(i)}, a_t^{(i)})}_{\text{No dependance on } \theta, \therefore 0} + \nabla_\theta \sum_{t=0}^{H} \log \pi_\theta (a_t^{(i)} \mid s_t^{(i)}) \quad (3)$$

$$= \nabla_\theta \sum_{t=0}^{H} \log \pi_\theta (a_t^{(i)} \mid s_t^{(i)}) \quad (4)$$

$$= \sum_{t=0}^{H} \nabla_\theta \log \pi_\theta (a_t^{(i)} \mid s_t^{(i)}) \quad (5)$$

(Rewrite)

Overall:

$$\nabla_\theta U(\theta) \approx \hat{g} = \frac{1}{m} \sum_{i=1}^{m} \sum_{t=0}^{H} \nabla_\theta \log \pi_\theta (a_t^{(i)} \mid s_t^{(i)}) R(\tau^{(i)})$$