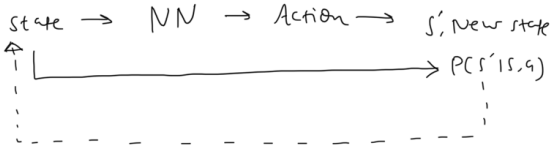
Introduction

Policy, $\pi_\theta(a|s)$, parametrized using 'θ'

→ We can use Deep NN to represent $\pi_\theta(a|s)$.

↓

'θ' denotes the policy parameters which will be the weights of Deep Neural Network



Model-free : 1. Do not know $P(s_{t+1}|s_t, a_t)$
2. Do not know the initial state probability

Trajectory distribution

$$\underbrace{p_\theta(s_1, a_1, s_2, a_2 \cdots s_T, a_T)}_{p_\theta(\tau)} = p(s_1) \prod_{t=1}^{T} \pi_\theta(a_t|s_t)\, p(s_{t+1}|s_t, a_t)$$

→ Trajectory distribution depends on the parameter 'θ'

REINFORCEMENT LEARNING OBJECTIVE

$$\theta^* = \arg\max_\theta \underbrace{E_{\tau \sim p_\theta(\tau)}\left[\sum_t r(s_t, a_t)\right]}_{J(\theta)}$$

Find the parameters θ which maximise the Expectation

$$J(\theta) = E_{\tau \sim p_\theta(\tau)}\underbrace{[r(\tau)]}_{\sum_{t=1}^{T} r(s_t, a_t)}$$

$$= \int p_\theta(\tau)\, r(\tau)\, d\tau$$

$$\nabla_\theta J(\theta) = \nabla_\theta \int p_\theta(\tau)\, r(\tau)\, d\tau = \int \nabla_\theta p_\theta(\tau) r(\tau)\, d\tau$$

differential operator is linear

→ Cannot differentiate as transition probabilities and initial state distribution is unknown.

→ What we do know are the trajectory samples and the rewards

→ $p_\theta(\tau) \nabla_\theta \log p_\theta(\tau) = p_\theta(\tau) \frac{\nabla_\theta p_\theta(\tau)}{p_\theta(\tau)} = \nabla_\theta p_\theta(\tau)$

$$\nabla_\theta J(\theta) = \int p_\theta(\tau) \nabla_\theta \log p_\theta(\tau)\, r(\tau)\, d\tau$$

$$= E_{\tau \sim p_\theta(\tau)}\left[\nabla_\theta \log p_\theta(\tau)\, r(\tau)\right]$$

samples from trajectory

→ $p_\theta(s_1, a_1, \cdots s_T, a_T) = p(s_1) \prod_{t=1}^{T} \pi_\theta(a_t|s_t)\, p(s_{t+1}|s_t, a_t)$

→ $\log p_\theta(\tau) = \log p(s_1) + \sum_{t=1}^{T} \log \pi_\theta(a_t|s_t) + \log p(s_{t+1}|s_t, a_t)$

→ $\nabla_\theta \log p_\theta(\tau) = \underbrace{0}_{\text{independent of } \theta} + \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t) + 0$

$$\nabla_\theta J(\theta) = E_{\tau \sim p_\theta(\tau)}\left[\nabla_\theta \log p_\theta(\tau)\, r(\tau)\right]$$

$$= E_{\tau \sim p_\theta(\tau)}\left[\left(\sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t)\right)\left(\sum_{t=1}^{T} r(s_t, a_t)\right)\right]$$

$$\nabla_\theta J(\theta) \approx \frac{1}{N} \sum_{n=1}^{N} (\quad)(\quad)$$

$$\theta \leftarrow \theta + \alpha \nabla_\theta J_\theta$$