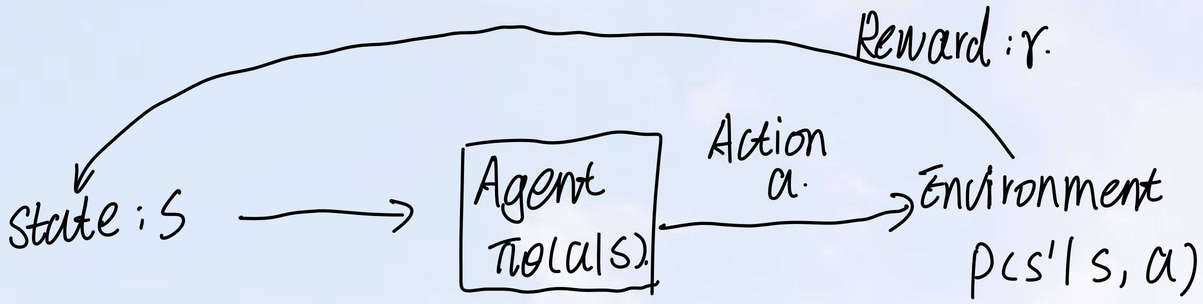
中国科学院大学
University of Chinese Academy of Sciences

1. The goal of reinforcement learning



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

$p_\theta(\tau)$, where τ is trajetory

The expected value of reward is:

$$E_{\tau\sim p_\theta(\tau)}\Big[\sum_t r(s_t, a_t)\Big] \quad (2)$$

In RL, we want to obtain max reward, so the goal of RL is maxmizing (2)

$$\theta^* = \arg\max_\theta E_{\tau\sim p_\theta(\tau)}\Big[\sum_t r(s_t, a_t)\Big] \quad (3)$$

For infinite horizon case: $\theta^* = \arg\max_\theta E_{(s,a)\sim p_\theta(s,a)}[r(s,a)]$

For finite horizon case: $\theta^* = \arg\max_\theta \sum_{t=1}^{T} E_{(s_t,a_t)\sim p_\theta(s_t,a_t)}[r(s_t,a_t)]$

2. Evaluating the objective.

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

$$J(\theta) = E_{\tau\sim p_\theta(\tau)}\Big[\sum_t r(s_t,a_t)\Big] \simeq \frac{1}{N}\sum_i \sum_t r(s_t, a_t)$$

sum over samples from π_θ.

3. Direct policy differentiation

$$J(\theta) = E_{\tau \sim p_\theta(\tau)}[r(\tau)] = \int p_\theta(\tau) r(\tau) d\tau \text{ , where } r(\tau) \text{ is } \sum_{t=1}^{T} r(s_t, a_t)$$

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

$$= E_{\tau \sim p_\theta(\tau)}[\nabla_\theta \log p_\theta(\tau) r(\tau)] \quad (4)$$

Where, $\nabla_\theta p_\theta(\tau) = p_\theta(\tau) \dfrac{\nabla_\theta p_\theta(\tau)}{p_\theta(\tau)} = p_\theta(\tau) \nabla_\theta \log p_\theta(\tau)$.

According to (1),

$$\log p_\theta(\tau) = \log p(s_1) + \sum_{t=1}^{T} \log \pi_\theta(a_t|s_t) + \log p(s_{t+1}|s_t, a_t) \quad (5)$$

So, we can rewrite $\nabla_\theta \log p_\theta(\tau)$:

$$\nabla_\theta [\cancel{\log p(s_1)} + \sum_{t=1}^{T} \log \pi_\theta(a_t|s_t) + \log p(\cancel{s_{t+1}}|s_t, a_t)] \quad (6)$$

where the first and third terms arn't function of θ.

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

$$\simeq \frac{1}{N} \sum_{i=1}^{N} \left(\sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t) \cdot \left(\sum_{t=1}^{T} r(s_t, a_t)\right)\right)$$

$$\theta \longleftarrow \theta + \alpha \cdot \nabla_\theta J(\theta)$$

REINFORCE ALGORITHM:

step 1. sample $\{\tau^i\}$ from $\pi_\theta(a_t|s_t)$ (run the policy)

step 2. $\nabla_\theta J(\theta) \simeq \sum_i \left(\sum_t \nabla_\theta \log \pi_\theta(a_t^i|s_t^i)\right)\left(\sum_t r(s_t^i, a_t^i)\right)$

step 3. $\theta \leftarrow \theta + \alpha \cdot \pi_\theta J(\theta)$

4. what is wrong with the policy gradient? High Variance

$$\nabla_\theta J(\theta) \simeq \frac{1}{N}\sum_{i=1}^{N}\left(\sum_{t=1}^{T}\nabla_\theta \log \pi_\theta(a_t|s_t)\cdot\left(\sum_{t=1}^{T} r(s_t,a_t)\right)\right) \quad (7)$$

I. Causality: policy at time t' can't affect the reward at time t, where $t < t'$

So, we rewrite (7):

$$\nabla_\theta J(\theta) \simeq \frac{1}{N}\sum_{i=1}^{N}\left(\sum_{t=1}^{T}\nabla_\theta \log \pi_\theta(a_t|s_t)\cdot\underbrace{\left(\sum_{t'=t}^{T} r(s_t,a_t)\right)}_{\text{reward to go}}\right) \quad (8)$$

II. Baselines

$$\nabla_\theta J(\theta) \simeq \frac{1}{N}\sum_{i=1}^{N}\nabla_\theta \log P_\theta(\tau)[r(\tau)-b]$$

$$= \frac{1}{N}\sum_{i=1}^{N}[\nabla_\theta \log P_\theta(\tau)r(\tau)-\nabla_\theta \log P_\theta(\tau)b]$$

where, $b=\frac{1}{N}\sum_{i=1}^{N} r(\tau)$

$$E[\nabla_\theta \log P_\theta(\tau)b] = \int P_\theta(\tau)\nabla_\theta \log P_\theta(\tau)\, b\, d\tau = \int \nabla_\theta P_\theta(\tau)\, b\, d\tau$$

$$= b\nabla_\theta \int P_\theta(\tau)\, d\tau = b\nabla_\theta 1 = 0. \quad (*)$$

So, subtracting a baseline is unbiased in expectation!

Tip: Average reward is not the best baseline, but it's pretty good!