1.4 & 1.5 Sampling Models & The Performance Difference Lemma

1. The episodic setting.

Learners acts for finite steps, they start from a fixed starting state $s_0 \sim \mu$, obeserve the trajectory and then reset to $s_0 \sim \mu$.

(i) Finite Horizon MDPs:

each episode lasts for H steps.

(ii) Infinite Horizon MDPs:

☆1: agents can terminate episodes after fixed steps;

☆2: each step has a probability of $1-\gamma$ to terminate.

this leads to an unbiased estimate of V^π.

Interests:

(i) number of episodes to find a near optimal policy

(ii) regret guarantee.

(iii) the strategy for the agents' exploration.

2. The generative model setting

Input a state-action pair (s,a);

Return a sample $s' \sim P(\cdot|s,a)$ and $r(s,a)$

3. The offline RL setting.

Agents has access to an offline dataset generated under certain policy under certain policy.

we assume the dataset is of the form $\{(s,a,s',r)\}$

$s' \sim P(\cdot|s,a)$, $r \sim r(s,a)$. $(s,a) \sim \Delta(S \times A)$ i.i.d.

1.5 The performance difference lemma.

[Notations]

① $V^\pi(\mu) := E_{s\sim\mu}[V^\pi(s)]$

② advantage : $A^\pi(s,a) = Q^\pi(s,a) - V^\pi(s)$

$A^*(s,a) := A^{\pi^*}(s,a) \le 0$

[Def] visitation measure over states :

$$d_{s_0}^\pi(s) = (1-\gamma)\sum_{t=0}^\infty \gamma^t \Pr^\pi(s_t = s|s_0)$$

$$d_\mu^\pi(s) = E_{s_0\sim\mu}[d_{s_0}^\pi(s)]$$

note : $\sum_s d_\mu^\pi(s) = E_{s_0\sim\mu}[\sum_s d_{s_0}^\pi(s)] = 1$. d_μ^π is a distribution.

> **Lemma 1.16 [The performance lemma]**
> For all policies π, π' and distributions μ over S:
> $$V^\pi(s) - V^{\pi'}(s) = \frac{1}{1-\gamma} E_{s'\sim d_\mu^\pi} E_{a'\sim\pi(s')}[A^{\pi'}(s',a')]$$

Pf: Let $\Pr^\pi(\tau|s_0=s)$ denote the probability of obseving a trajectory τ when starting in state s and following π.

① first we show that:

$$E_{\tau\sim \Pr^\pi(s_0)}[\sum_{t=0}^\infty \gamma^t f(s_t,a_t)] = \frac{1}{1-\gamma} E_{s\sim d_{s_0}^\pi} E_{a\sim\pi(s)}[f(s,a)]$$

In fact,

$$
\begin{aligned}
& E_{\tau\sim p_r^{\pi}(s_0)}\left[\sum_{t=0}^{\infty}\gamma^t f(s_t,a_t)\right] \\
&= \sum_{\tau}\left[\sum_{t=0}^{\infty}\gamma^t f(s_t,a_t)\right]p_r^{\pi}(\tau|s_0) \\
&= \sum_{t=0}^{\infty}\gamma^t\left[\sum_{\tau} f(s_t,a_t)\,p_r^{\pi}(\tau|s_0)\right] \\
&= \sum_{t=0}^{\infty}\gamma^t\left[\sum_{\tau_t} f(s_t,a_t)\,p_r^{\pi}(\tau_t|s_0)\right] \\
&= \sum_{t=0}^{\infty}\gamma^t\left[\sum_{s,a} f(s,a)\,p_r^{\pi}(s_t=s|s_0)\,\pi(a|s)\right] \\
&= \sum_{s,a} f(s,a)\,\pi(a|s)\,\frac{1}{1-\gamma}\left[(1-\gamma)\sum_{t=0}^{\infty}\gamma^t p_r^{\pi}(s_t=s|s_0)\right] \\
&= \frac{1}{1-\gamma}\sum_{s} d_{s_0}^{\pi}(s)\sum_{a} f(s,a)\,\pi(a|s) \\
&= \frac{1}{1-\gamma}E_{s\sim d_{s_0}^{\pi}}E_{a\sim\pi(s)}[f(s,a)]
\end{aligned}
$$

② now we prove this lemma.

$$
\begin{aligned}
V^{\pi}(s)-V^{\pi'}(s) &= E_{\tau\sim p_r^{\pi}(s)}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\right]-V^{\pi'}(s) \\
&= E_{\tau\sim p_r^{\pi}(s)}\left[\sum_{t=0}^{\infty}\gamma^t\left(r(s_t,a_t)+V^{\pi'}(s_t)-V^{\pi'}(s_t)\right)\right]-V^{\pi'}(s) \\
&= E_{\tau\sim p_r^{\pi}(s)}\left[\sum_{t=0}^{\infty}\gamma^t\left(r(s_t,a_t)+\gamma V^{\pi'}(s_{t+1})-V^{\pi'}(s_t)\right)\right] \\
\text{(minus } E_{\tau}[V^{\pi'}(s_0)]\text{)} &= E_{\tau\sim p_r^{\pi}(s)}\left[\sum_{t=0}^{\infty}\gamma^t\left(r(s_t,a_t)+\gamma E[V^{\pi'}(s_{t+1})|s_t,a_t]-V^{\pi'}(s_t)\right)\right] \\
&= E_{\tau\sim p_r^{\pi}(s)}\left[\sum_{t=0}^{\infty}\gamma^t\left(Q^{\pi'}(s_t,a_t)-V^{\pi'}(s_t)\right)\right] \\
&= E_{\tau\sim p_r^{\pi}(s)}\left[\sum_{t=0}^{\infty}\gamma^t A^{\pi'}(s_t,a_t)\right] \\
\text{(by ①)} &= \frac{1}{1-\gamma}E_{s\sim d_{s_0}^{\pi}}E_{a\sim\pi(s)}[f(s,a)] \qquad \square
\end{aligned}
$$