(I). MC estimation → Prediction
↳ Estimate value function

↳ control; * how to opt. make use of trajectory
↳ find optimal policy.

(II) off policy estimation:-

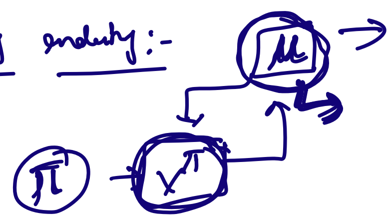
Prediction :- (π): V^{π} ← value fn corre policy π.

generate trajectory : 1). $S_0, a \sim \pi, S_1, a \sim \pi, \ldots$ (τ_1, τ_2 ...)

2)

off-policy:- samples will be generated from different policy 'μ'

1. gaming industry :-



• Both are stochastic policies :-

$\mu(a|s) > 0$ whenever $\pi(a|s) > 0$. $\forall s$

off-policy estimation :-

① $(\Omega, \mathcal{F}, \mu)$ $\quad X : \Omega \to \mathbb{R}$

② $(\Omega, \mathcal{F}, \pi)$

Q1 : $E_{\mu}[X] = \sum_{x \in X} \mu(x) \cdot x$

$E_{\pi}\left[\frac{\pi(x)}{\mu(x)} \cdot X\right] \stackrel{?}{=} \sum_{x \in X} \mu(x) \cdot \frac{\pi(x)}{\mu(x)} \cdot x$

$$= \sum_{x \in X} \pi(x) \cdot x$$

$$= \mathbb{E}_{\pi}[X]$$

$$f(x) = \frac{\pi(x)}{\mu(x)} \; ; \quad f(x) \cdot x$$

(*). $V^{\pi}(s) = \mathbb{E}_{(s_0, s_1, s_2, \ldots)}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t)\right]$. $\rightarrow$ $X =$

$$V^{\pi}(s) = \mathbb{E}_{\mu}\left[f(s_0, s_1, s_2 \ldots)\right]$$

$$X \quad \sum_{t=0}^{\infty} \gamma^t r(s_t)$$

$$= \sum_{s_0, s_1 \ldots} P_{\pi}(s_0, s_1, s_2, \ldots) \underbrace{\sum_{t=0}^{\infty} \gamma^t r(s_t)}_{X}$$

$$V^{\mu}(s) = \sum_{s_0, s_1 \ldots} P_{\mu}(s_0, s_1, s_2 \ldots) \sum_{t=0}^{\infty} \gamma^t r(s_t).$$

Q :- If you get trajectories from μ, can you compute value function of π. ?

$$f(s_0, s_1, \ldots) = \left[\frac{P_{\pi}(s_0, s_1 \ldots)}{P_{\mu}(s_0, s_1, \ldots)}\right]$$

$$P_{\pi}(s_0, a_0, s_1, a_1, s_2).$$

$$P(s_0) \times \pi(a_0 | s_0) \times P(s_1 | s_0, a_0)$$
$$\times \pi(a_1 | s_1)$$
$$\times P(s_2 | s_1, a_1)$$

$$P(x_1, x_2, x_3) = P(x_1) \cdot P(x_2 | x_1) \cdot P(x_3 | x_1, x_2).$$

π, μ

$P_\mu(s_0, a_0, s_1, a_1, s_2)$:-

$P(s_0) * \mu(a_0|s_0) * P(s_1|s_0, a_0) * \mu(a_1|s_1) * P(s_2|s_1, a_1)$.

$P_\pi(s_0, a_0, s_1, a_1, s_2)$ / $P_\mu(s_0, a_0, s_1, a_1, s_2)$ $= \dfrac{\pi(a_0|s_0) * \pi(a_1|s_1)}{\mu(a_0|s_0) * \mu(a_1|s_1)}$

↳ we know

$\mu : S \to A$ det

$\mu : S \times A \to [0,1]$ → sto dest

environment:

$\pi : S, a \to s'$

(*). I know policy μ ⇒ $(s_0, a_0 \sim \mu, s_1, a_1 \sim \mu, s_2, \dots)$

π ⇒

on policy :-

π :-

1. $s_0, a_0, r_0, s_1, a_1, r_1 \dots$ $R_1 = \sum_{t=0}^{\infty} \gamma^t r_t$.
2. $s_0, a_1, r_0^1 \dots$ R_2
3. R_3
4. ..

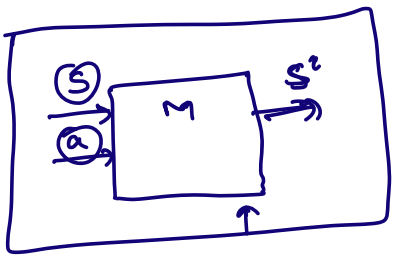
$V^\pi = \dfrac{\sum R_n}{n} = \left(\dfrac{R_1}{n} + \dfrac{R_2}{n} + \dfrac{R_3}{n} + \dots\right)$

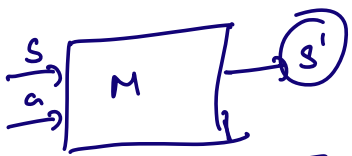
Generating

Offpolicy: μ:

1. R_1^{μ} $(s_0, a \sim \mu, r_0, s_1, a_1 \sim \mu \ldots)$
2. R_2^{μ}
3. R_3^{μ}

V^{π} ← ? ← $\frac{R_1^{\mu} + R_2^{\mu} + \ldots}{n}$





↳ available

V^{π} ← $\frac{\rho_1 R_1^{\mu} + \rho_2 R_2^{\mu} + \rho_3 R_3^{\mu} + \ldots}{n}$