- Expectation over actions :- known — $\pi(a_1|s), \pi(a_2|s), \ldots$
- Expectation over states : not known. — $\sum_a Q(s', a) \cdot \pi(a|s)$

Q4: MDP with 2 states 1 & 2. (π)

$$V^{\pi}(2) = 2 + V^{\pi}(1).$$

sample: $(1, a_1, 1, 2)$ — s, a, r, next state

$(1, a_2, 2, 2)$

$V_0 = 0$; $\alpha = \frac{1}{n+1}$; $\gamma = 1$.

$V_2(1)$ & $V_2(0)$ → timestep

$V_{n+1}(i) = V_n(i) + \alpha(\_\_)$

$$V_{n+1}(i_n) = (1-\alpha_n) V_n(i_n) + \alpha_n [r_n + \gamma V_n(i_{n+1})].$$

(i_n, r_n, i_{n+1})

First sample: $(1, a_1, 1, 2)$; $(1, r_2, 2, 3)$.

$V_0(1) = 0$; $V_0(2) = 0$

$V_1(1) = (1-1) V_0(1) + 1(1 + 1(0)) = 1$

$\rightarrow V_1(2) = 2 + 1 = 2$ (#)

Second sample:-

$V_2(1) = (1 - \frac{1}{2})(1) + \frac{1}{2}(2 + 1 \cdot 2)$

$= \frac{1}{2} + 2 = 5/2$

$V_2(2) = 2 \cdot 5/2 = 5$

$\frac{1}{2} + \frac{1}{2}(2) = 3/2$

$3/2, 0$

$3/2, 3$

$5/2, 5$ →

⑤. C 1, a_1, 1, 2, a_2, 2, 1, a_3, 4, 2). $\gamma = 0.5$.

1st 2 3rd

$V(1)$.

① $+\gamma(0)$. $\quad 1 + 0.5(2) + \gamma^2(0)$; $1 + 0.5(2) + (0.5)^2(4) +$ $(0.5)^3 0$

1 step TD ↓ ①

↳ (+). ②

↳ ③

Off-policy Prediction | control :-

Prediction :- π : given obj: v^π

Model- information is not known.

- samples are obtained from different policy μ.

Monte-Carlo estimation :-

$\tau \rightarrow (S_1, a_1, S_2, a_2, S_3, a_3, \cdots)$ trajectory given.

$Pr_\mu(\tau)$:- $Pr\{S_1, a_1, S_2 \cdots\}$ ↓

$= Pr\{S_1\} \times \mu(a_1|S_1) \times P(S_2|S_1, a_1) \times \mu(a_2|S_2) \times P(S_3|S_2, a_2) \cdots$

$Pr_\pi(\tau)$:-

↓

$Pr(S_1) \times \pi(a_1|S_1) \times P(S_2|S_1, a_1) \times \pi(a_2|S_2) \times \cdots$

$Pr\{x_1, x_2, x_3\} = Pr\{x_1\} \cdot Pr\{x_2|x_1\} \times Pr\{x_3|x_1, x_2\}$

$\rho_\tau = \frac{Pr_\pi(\tau)}{Pr_\mu(\tau)} = \frac{\pi(a_1|S_1) \times \pi(a_2|S_2) \times \cdots}{\mu(a_1|S_1) \times \mu(a_2|S_2) \times \cdots}$

$\Rightarrow \frac{(0.99)^{1000}}{(0.001)^{1000}}$

$(0.001)^{1000}$

- $E_{\mu}[X] = \mu(x_1) \cdot x_1 + \mu(x_2) \cdot x_2$.
- $E_{\pi}[X] = \pi(x_1) \cdot x_1 + \pi(x_2) \cdot x_2$

$= \mu(x_1) \cdot \frac{\pi(x_1)}{\mu(x_1)} \cdot x_1 + \mu(x_2) \cdot \frac{\pi(x_2)}{\mu(x_2)} \cdot x_2$

$= E_{\mu}[\rho(x) \cdot X]$.

$\rho(x) = \frac{\pi(x)}{\mu(x)}$

$v^{\pi}(s) = E[G_{\pi}]$

$\hookrightarrow \sum_{t=0}^{\infty} \gamma^t r_t \rightarrow r(s_t, a_t \sim \pi)$.

$v^{\mu}(s) = E[G_{\mu}]$

$\rightarrow \pi : (s_1, a_1, s_2, a_2, s_3, a_3, \ldots)$ with r_1, r_2, r_3

$v^{\pi}(s_1) \leftarrow$

$r_1 + \gamma r_2 + \gamma^2 r_3$

$\mu :$ (s_1, a_1) (s_2, a_2) (s_3, a_3) $\rightarrow r_1 + \gamma r_2 + \gamma^2 r_3$. ρ_{τ}

$\rho_{\tau}(r_1 + \gamma r_2 + \gamma^2 r_3)$

ordinary I.S.

$V(s) \doteq \frac{\sum_{\tau} \rho_{\tau} G_{\tau}}{|\tau|}$

$\leftarrow$ infinite variance.

certain : $\frac{\sum_{\tau} G_{\tau}}{|\tau|}$

weighted I.S :

$V(s) \doteq \frac{\sum_{\tau} \rho_{\tau} G_{\tau}}{\sum_{\tau} \rho_{\tau}}$

$\frac{\rho_{\tau}}{\sum_{\tau} \rho_{\tau}}$

$=$

Exp1 :-

Exp2.

$E[x] = x^*$

$\rightarrow x_0, x_1, x_2, \ldots$

$x_1, x_2, x_3, \ldots$

$E[x_i] = x^*$

- Incremental implementation :-

$$V_n \doteq \frac{\sum_{k=1}^{n-1} W_k R_k}{\sum_{k=1}^{n-1} W_k}$$

$$V_{n+1} = V_n + \frac{W_n}{C_n}\left[R_n - V_n\right]$$

$$C_{n+1} = C_n + W_{n+1}$$

Prediction

$\frac{a_1 + a_2 + \cdots + a_n}{n}$

$x_{n+1} = (1-\alpha)\, x_n + \alpha\, a_n (r_n)$

$\frac{1}{n+1}$

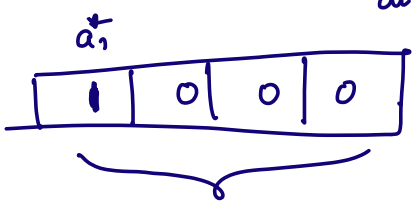
Control :- Prob :- compute the optimal policy, with samples derived from a diff. behavior policy 'b'.

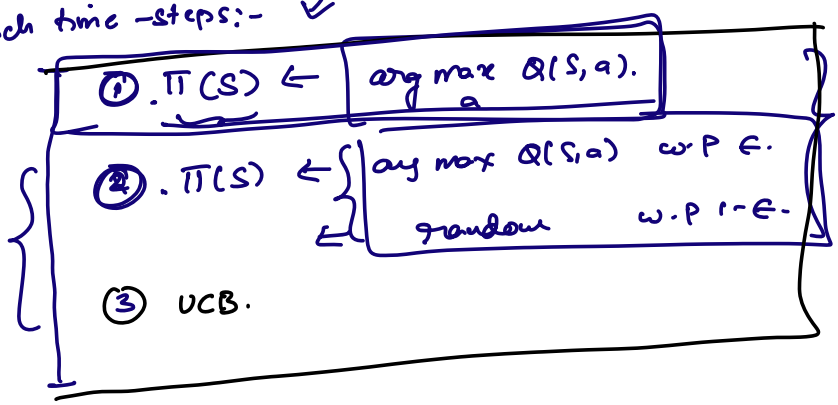
target policy :- fixed π

2) Move it.

$Q(S,a)$ estimates

at each time-steps :-





Summarize:-

(1). Importance Sampling :- (weights trajectory appropriately).

(2). ↳ eliminates unknown & only on known quantities.

(3). weighted I.S : ordinary I.S has infinite variance.

π : V^{π} : $E[G_{\pi}]$

μ : some samples are used

if $\pi(a|s) > 0$ then $\mu(a|s) > 0$. ✓

$\Rightarrow$

$\mu(a^{*}|s) > 0$ but $\pi(a^{*}|s) = 0$

Example 4.3 :- Gamblers problem