(*). Monte-carlo estimation :-

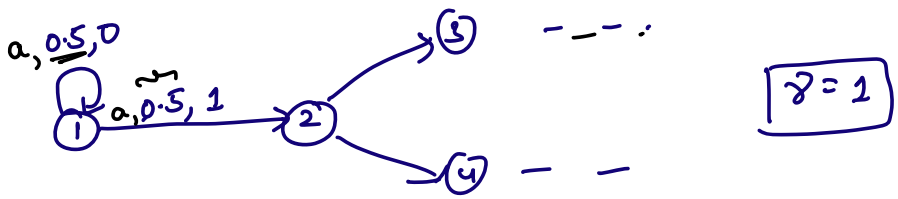
1). given samples { First visit / Every visit }

2). Prediction, control.
↳ optimal policy.

(2). off-policy :- target policy, behavior policy.

{ Prediction :
control → optimal policy.

Can we still improve the sample efficiency :-



given policy π, estimate value functions $V^{\pi}(1)$?

$V^{\pi}(2) = 2$. ✓ → (I)

Actual estimate of $V^{\pi}(1)$?

(*) $V^{\pi}(1) = \frac{1}{2}[0 + V^{\pi}(1)] + \frac{1}{2}[1 + 2]$

$\Rightarrow \frac{V^{\pi}(1)}{2} = \frac{3}{2}$

$\Rightarrow V^{\pi}(1) = 3.$

one Sample :- (1st, next, next state)

1, 1, 2, 0, 3, 1, 4, ..

(*) → (II)

M.C. :- $R_1 = 1 + 0 + 1 = 2$

TD(0) :- $1 + \gamma \cdot V^{\pi}(2)$

$= 1 + 1.2 = 3$

MDP + policy (π) → Markov chain.

↳ Stationary distn.

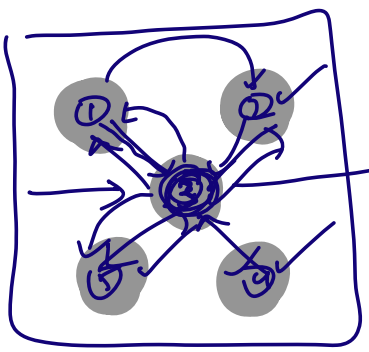
$d^\pi(1)$ $d^\pi(2)$ $\cdots$ $d^\pi(n)$

0.7 0.2 0.1.

3, 0, 2, 1, 5, ...

$V(3) = 0 + \gamma \cdot 1 + \gamma^2 \cdot c \cdots$

$V(3) = 0 + \gamma \cdot V(2)$



$V(3) \approx V^\pi(3)$

1, a, r, 3, a_2, r_2, 4, ...

(*) $V(1) = r + \gamma \cdot V(3)$ →

TD(0) update:

$x_n = E[c]$ ✓

$c_1, c_2, c_3, \ldots$ samples ✓

$x_{n+1} = (1-\alpha_n) x_n + \alpha_n (c_n)$.

$x_n \to x$

Prediction :-

$\pi : V^\pi$

$(s, a \sim \pi, r, s')$

$V^\pi(s) = E_{s'}[r + \gamma V^\pi(s')]$

↳ c

$V_{n+1}(s) = (1-\alpha_n) V_n(s) + \alpha_n [r + \gamma V_n(s')]$.

$[V_1(s_1), V_1(s_2), \ldots]$

$$V_n(s) \longrightarrow V^{\pi}(s), \quad \forall s \in S.$$

$n \to \infty$

Q-learning:- control :- optimal policy.

optimal Q-Bellman eqn:-

$$Q(s,a) = \underset{s'}{E}\left[r + \gamma \max_b Q(s',b)\right].$$

(s, a, r, s') ↓ given.

$Q(s,a)$ → x ; $\max_b Q(s',b)$ → C

(*) $$Q_{n+1}(s,a) = (1-\alpha_n) Q_n(s,a) + \alpha_n \left[r + \gamma \max_b Q_n(s',b)\right].$$

α_n → step-size

$\sum \alpha_n = \infty \; ; \; \sum \alpha_n^2 < \infty$

$\left\{\frac{1}{n}\right\}$.

$1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots$

(*) please derive the same for SARSA & expected SARSA.

(*) TD(0) prediction
(*) Q-learning.

Mountain car :-