Assignment - 02

Reinforcement Learning

K. Surya Prakash

EE18BTECH11026

Q1) a) Evaluate V(s) : first visit MC:

* State A: $\frac{14+15+17+16+15}{5} = 15.4$

* State B: $\frac{13+14+16+15+14}{5} = 14.4$

* State C: $\frac{12+13+15+14+13}{5} = 13.4$

* State D: $\frac{12+12+12+11}{4} = 11.75$

* State E: $\frac{11+11+11+10+9}{5} = 10.4$

State F: $\frac{11+10+10+10+9}{5} = 8$

State G: (Terminal State) : 0

b) States {A, D} will remain the same

{B, C, E, F} are likely to change.

Reason: {A, D} only occur only once

in the sequence in all cases when they appear. ~~mat~~

* Rest other states are occuring more than once in an episode (atleast once).

(c) * π_f : moves right or jump

$$V^{\pi_f}(s) = E_{\pi_f}\left(\sum \gamma^k r_{t+k+1}\right) \quad : \gamma = 1$$

* G : terminal state : $\boxed{V^{\pi_f}(G) = 0}$

* $V^{\pi_f}(F) = \frac{10}{1} = 10$

* $V^{\pi_f}(E) = \frac{(10+1)}{1} = 11$

* $V^{\pi_f}(D) = \frac{(1+1+10)}{1} = 12$

* $V^{\pi_f}(C) = \frac{((1+12)+(4+11))}{2} = 14$

* $V^{\pi_f}(B) = \frac{1+14}{1} = 15$

* $V^{\pi_f}(A) = \frac{1+15}{1} = 16$

d) Consider trajectories $\{2, 3\}$

$V^{\pi_f}(s)$: from MLE estimates

$$\hat{P}(s'|s,a) = \frac{1}{N(s,a)} \sum_{k=1}^{K} \sum_{t=1}^{l_k - 1} \mathbb{1}(s_{k,t} = s, a_{k,t} = a, s_{k,t+1} = s')$$

$$\hat{R}(s'|s,a) = \frac{1}{N(s,a,s')} \sum_{k} \sum_{t} \mathbb{1}(s_{k,t} = s, a_{k,t} = a, s_{k,t+1} = s') \cdot r_{t,k}$$

* State-G : Terminal State :

$$V^{\pi_f}(G) = 0$$

* State-F : $\hat{P}^{A}_{s_1 s_2} \Rightarrow \hat{P}(s_2 | s_1, A)$: A: R → right, J → jump

$$\hat{P}^{R}_{GF} = \frac{2}{2} = 1$$

$$\hat{R}^{R}_{GF} = \frac{10}{2} + \frac{10}{2} = 10$$

$$V^{\pi_f}(F) = \hat{P}^{R}_{GF}\left(R^{R}_{GF} + \gamma V^{\pi_f}(G)\right)$$

$$= 1(10 + \gamma(0))$$

$$= 10$$

$$\boxed{V^{\pi_f}(F) = 10}$$

Similarly:

* State-E:

$$V^{\pi_f}(E) = \hat{P}^{R}_{EF}\left(\hat{R}^{R}_{EF} + \gamma V^{\pi_f}(F)\right)$$

$$= 1(1+10) = \underline{\underline{11}}$$

* State-D:

$$V^{\pi_f}(D) = \hat{P}^{R}_{DE}\left(\hat{R}^{R}_{DE} + V^{\pi_f}(E)\right)$$

$$= \frac{1}{1}\left(\frac{1}{1} + 11\right) = \underline{\underline{12}}$$

* State-C: (Both ways Right & Jump

$$V^{\pi_f}(C) = \hat{P}^{R}_{CD}\left(\hat{R}^{R}_{CD} + V^{\pi_f}(D)\right) +$$

$$\hat{P}^{J}_{CE}\left(\hat{R}^{J}_{CE} + V^{\pi_f}(E)\right)$$

$$\hat{P}^{R}_{CD} = \frac{1}{2}, \quad \hat{P}^{J}_{CE} = \frac{1}{2}$$

$\hat{P}^R_{CD} = \frac{1}{1} = 1$, $\hat{R}_{CE} = \frac{4}{4} = 4$

$v^{\pi_f}(C) = \frac{1}{2}(1+12) + \frac{1}{2}(4+11) = \cancel{16}\ \underline{\underline{14}}$

* State-B:

$\hat{P}^R_{BC} = \frac{2}{2} = 1$; $\hat{R}^R_{BC} = \frac{2}{2} = 1$

$v^{\pi_f}(B) = 1(1 + \cancel{15}^{14}) = \cancel{16}\ \underline{\underline{15}}$

* State-A:

$v^{\pi_f}(A) = \hat{P}^R_{AB}\left(\hat{R}^R_{AB} + v^{\pi_f}(B)\right)$

$= \frac{2}{2}\left(\frac{2}{2} + 15\right) = \underline{\underline{16}}$

(e) Since $v^{\pi_f}(s)$ is a unique value func.

with infinite trajectories

$\rightarrow$ MC converges to $v^{\pi_f}(s)$

while

By satisfying Robin Munoe's convergence

$\rightarrow$ TD(0) $\longrightarrow v^{\pi_f}(s)$

Both will converge to same value.

(f) Q-learning.

$$Q(s,a) \leftarrow Q(s,a) + \alpha\left(r + \gamma\left(\max_{a'} Q(s',a')\right) - Q(s,a)\right)$$

$s \leftarrow s'$

C jump 4: $Q(C,J) = 0 + \frac{1}{2}(4 + 1(0) - 0) = 2$

E right 1: $Q(E,R) = 0 + \frac{1}{2}(1 + 1(0) - 0) = 0.5$

F left −2: $Q(F,L) = 0 + \frac{1}{2}(-2 + 0.5) = -0.75$

E right +1: $Q(E,R) = 0.5 + \frac{1}{2}(1 + 0 - 0.5) = 0.75$

Q-table:

~~Q(C,L) Q(C,J) Q(E,L) Q(E,R)~~

	Q(C,L)	Q(C,J)	Q(E,L)	Q(E,R)	Q(F,L)	Q(F,R)
Init	0	0	0	0	0	0
T1	0	2	0	0	0	0
T2	0	2	0	0.5	0	0
T3	0	2	0	0.5	−0.75	0
T4	0	2	0	0.75	−0.75	0

(g) Greedy policy. $\pi(s) = \arg\max_a Q(s,a)$

$\pi(C) = \text{Jump}$

$\pi(E) = \text{Right}$

$\pi(F) = \text{Right}$

Q2) a) $\sum_{t=0}^{\infty} \alpha_t = \infty$

b) $\sum_{t=0}^{\infty} \alpha_t^2 < \infty$

1) $\alpha_t = 1/t$

a. $\sum 1/t = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8} + \dots$

$> 1 + \frac{1}{2} + \underbrace{\frac{1}{4} + \frac{1}{4}} + \underbrace{\frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8}} + \dots$

$> 1 + \frac{1}{2} + 1/2 + 1/2 + \dots \rightarrow \infty$

$> \infty$

$\therefore \boxed{\sum \alpha_t = \infty}$

b) $\sum 1/t^2 = \frac{1}{1} + \frac{1}{2^2} + \dots$

$= \frac{\pi^2}{6} < \infty$

$\therefore \alpha_t = \frac{1}{t}$ will converge

(b) $\alpha_t = 1/t^2$

$\sum \alpha_t = \sum \frac{1}{t^2} = \frac{\pi^2}{6} < \infty \neq \infty$

Does not satisfy conditions

$\boxed{\text{\# Does not converge}}$

(c) $\alpha_t = \frac{1}{t^{2/3}}$; If

If $\sum 1/t^p \rightarrow$ converges if $p > 1$

$\rightarrow$ diverges if $p \leq 1$

$\therefore \because p = 2/3 < 1 \rightarrow$ diverges

$\therefore \sum \alpha_t = \infty$ ✓

$\sum \alpha_t^2 = \sum \frac{1}{t^{4/3}} \Rightarrow p = 4/3 \rightarrow$ converges

$\sum \alpha_t^2 < \infty$ ✓

$\boxed{\text{\# } \alpha_t = \frac{1}{t^{2/3}} \text{ converges}}$

(4) $\alpha_t = \frac{1}{t^{1/2}}$

$\Rightarrow \sum \alpha_t^2 = \sum \frac{1}{t} = \infty$ } proved in (1).

Does not satisfy conditions

$\therefore$ $\boxed{\text{Does not converge}}$

$\Rightarrow$ By observation:

let $\alpha_t = \frac{1}{t^p}$ to converge

~~if $p > 1$ & $2p < 1$~~

~~$p > 1$ & $p < 1/2$~~

~~$p \in [1/2, 1)$~~

if $p \leq 1$, $2p > 1$

$p \leq 1$; $p > 1/2$

$\boxed{p \in (1/2, 1]}$

Q4) $|A| = K$

horizon-length $= 1$

$a \in A$; $R^a(r) \to$ distribution of reward $[0,1]$

$a \sim \pi_b$ } behaviour

π } Target

(a) $V^{\pi}(s) = \underset{\pi}{E}(r \mid a \sim \pi)$

$V^{\pi_b}(s) = \underset{\pi_b}{E}(r \mid a \sim \pi_b) = r$ } from data.

$= \pi_b(a) \cdot r = 1 \cdot r = r$

$V^{\pi}(s) = \frac{\pi(a)}{\pi_b(a)} \cdot V^{\pi_b} = \frac{\pi(a) \cdot r}{\pi_b(a)}$

$$\boxed{\hat{V}^{\pi} = \frac{\pi(a)}{\pi_b(a)} \cdot r}$$

$\because$ Only one observation : $\pi_b(a) = 1$

$$\boxed{\hat{V}^{\pi} = \pi(a) \cdot r}$$

* Consider

$$E_\pi(\hat{V}^\pi) = E_{\pi_b}\left[\frac{\pi(a)}{\pi_b(a)}\cdot r\right]$$

$$= \sum_a \sum_r \pi_b(a)\,\frac{\pi(a)}{\pi_b(a)} \cdot R^a(r)\cdot r$$

$$= \sum_a \sum_r \pi(a)\, R^a(r)\cdot r$$

$$= \sum_a \pi(a) \sum_r R^a(r)\cdot r$$

$$= E_\pi(r \mid a)$$

$$= V^\pi$$

$$\boxed{\therefore\ E_\pi(\hat{V}^\pi) = V^\pi} \Big\}\ \text{unbiased estimate}$$

(b) $$E_{\pi_b}\left[\frac{\pi(a|\cdot)}{\pi_b(a|\cdot)}\right] = \sum_a \pi_b(a|\cdot)\cdot\frac{\pi(a|\cdot)}{\pi_b(a|\cdot)} = \sum_a \pi(a|\cdot)$$

$$= \underline{\underline{1}}$$

(c) π_b : uniform : $\pi_b(a) = \frac{1}{K}$

importance-sampling ratio : $\frac{\pi(a|\cdot)}{\pi_b(a|\cdot)}$

Since π: deterministic policy

let $\pi(a') = 1$ if $a = a'$

$= 0$ else

$\therefore \cancel{\frac{\pi(a)}{\pi_b(a)}} = \frac{\pi(a|.)}{\pi_b(a|.)} = \frac{1}{1/K} = K$ if $a = a'$

$= 0/{1/K} = 0$ else

(d) $\underset{\pi}{\mathrm{Var}}\left[V^{\pi}\right] = \underset{\pi_b}{\mathrm{Var}}\left[\frac{\pi(a|.)}{\pi_b(a|.)} \cdot r\right]$

as r is deterministic

$\sum^{a} R(r') \cdot r' = \underline{\underline{r}}$

$\frac{\pi(a|.)}{\pi_b(a|.)} = IS = K$ if $a = a'$

0 else

$$= \underset{\pi_b}{E}\left[\left(\frac{\pi(a|.)}{\pi_b(a|.)} \cdot r\right)^2\right] - \left(\underset{\pi_b}{E}\left(\frac{\pi(a|.)}{\pi_b(a|.)} \cdot r\right)\right)^2$$

$$= \sum_{a \in A} \pi_b(a) \cdot \left(\frac{\pi(a|.)}{\pi_b(a|.)} \cdot r\right)^2 - \left(\sum_{a \in A} \pi_b(a)\left(\frac{\pi(a|.)}{\pi_b(a|.)} \cdot r\right)\right)^2$$

$$= \pi_b(a'|.)\left(\frac{\pi(a'|.)}{\pi_b(a'|.)} \cdot r\right)^2 - \left(\pi_b(a') \cdot \left(\frac{\pi(a'|.)}{\pi_b(a'|.)} \cdot r\right)\right)^2$$

$$= \frac{1}{K}\left(K \cdot r \right)^2 - \left(\frac{1}{K}\left(K \cdot r \right) \right)^2$$

$$= K r^2 - r^2$$

$$= (K-1) r^2$$

$$\boxed{\operatorname{Var}_{\pi}\left(V^{\pi} \right) = (K-1) r^2}$$

(c) $$\operatorname{Var}_{q}\left(\hat{\mu}_{Q} \right) = \operatorname{Var}_{q}\left[\frac{P(x)}{Q(x)} \cdot f(x) \right]$$

$$= E_p\left[\left(\frac{P(x)}{Q(x)} \cdot f(x) \right)^2 \right] - E_p\left(f(x) \right)^2$$

$$\operatorname{Var}_{\pi_b}\left(\hat{V}(\pi) \right) = E_{\pi_b}\left(\frac{\pi(a)}{\pi_b(a)} \cdot r^2 \right] - \left(E_{\pi}(r) \right)^2$$

$$= \sum_a \pi(a) \left(\frac{\pi(a)}{\pi_b(a)} r^2 \right)$$

$$- \left(\sum_a \left(\frac{1}{\pi_b(a)} \cdot r \right) \right)^2$$

$$\operatorname{Var}_{\pi_b}\left(\hat{V}_{\pi} \right) = \left(\sum_a \pi(a) \left(\frac{\pi(a)}{\pi_b(a)} - 1 \right) \right) \left(\sum_r r^2 R^a(r) \right)$$

~~max boi~~

$$= \left(\sum_a \frac{\pi^2(a)}{\pi_b(a)} - \sum_a \pi(a) \right) \left(\sum_r r^2 R^q(r) \right)$$

$$= \left(\sum_a \frac{\pi^2(a)}{\pi_b(a)} - 1 \right) \left(\sum_r r^2 R^q(r) \right)$$

$\sum \pi(a) = 1$ ~~max value of $\sum_a \pi^2$~~

~~Max bound occurs when $R^q(r)$ is deterministic~~

Max bound.

$$= \left(\sum_a \frac{1}{\pi_b(a)} - 1 \right) \left[\sum_r r^2 R^q(r) \right]$$

(f) If π completely diverges from π_b

Then both trajectories does not match

Given $P(T)$; $Q(T) = 0$

$$\Rightarrow \frac{P(T)}{Q(T)} = \infty$$

Else if follows

$P(T) = P(S_0).$

$P(S_0, a_0, S_1 \ldots) = \cancel{P(S_0) \cdot \frac{1}{K} \cdot P^{a_0}_{S_0 S_1}}$

$= P(S_0) \cdot P(a_0) \cdot P^{a_0}_{S_0 S_1} \cdot \prod_{k=1}^{N} \left(P_{a_k} \cdot P^{a_k}_{S_k S_{k+1}} \right)$

$Q(S_0, a_0 \ldots) = Q\, P(S_0) \cdot P(a_0) \cdot P^{a_0}_{S_0 S_1} \ldots$

$\because$ P is followed by π_b which is uniform random, $\Rightarrow$ $P(a_0) = 1/K$

$\Rightarrow$ Q is deterministic. $P(a_0) = 1$

$\Rightarrow \frac{P(T)}{Q(T)} = \frac{1}{K^N}$ $\quad$: N : ~~no~~ horizon length

$\Rightarrow$ if $N \to \infty \Rightarrow \frac{P(T)}{Q(T)} \to 0$

$$\frac{P(T)}{Q(T)} = \begin{cases} 0 & \text{if } P \text{ matches } Q \\ \infty & \text{else} \end{cases}$$

Q3) Q-learning

$S = \{s\}$, $A = \{a_1, a_2\}$

$\gamma = 1$

$E(r|a_1) = E(r|a_2) = c$

$r \sim R^{a_i}$ $i \in \{1, 2\}$

a) $Q(s, a_1) = E(r/s, a_1) = E(r/a_1) = c$

$Q(s, a_2) = E(r/s, a_2) = E(r/a_2) = c$

$V^*(s) = \max(Q(s, a_1), Q(s, a_2)) = \underline{\underline{c}}$

(b) Prove that estimated value of $\hat{\pi} \rightarrow \hat{V}^{\hat{\pi}}$ is a biased estimate of $V^*(s)$

$\{r_1, r_2, \ldots r_{n/2}\} \rightarrow$ Rewards obtained by choosing a_1

$\{r_{\frac{n}{2}+1}, \ldots r_n\} \rightarrow$ By choosing a_2

$\hat{V}^{\hat{\pi}} = \max(\hat{Q}(s, a_1), \hat{Q}(s, a_2))$

$$\hat{Q}(s, a_1) = \frac{r_1 + \ldots r_{n/2}}{n/2}$$

$$\hat{Q}(s, a_2) = \frac{r_{\frac{n}{2}+1} \ldots + r_n}{n/2}$$

$\hat{V}^{\hat{\pi}}$, $\max f_i$

<u>Bias of Estimator:</u>

$$\text{Bias}(\hat{V}^{\hat{\pi}}) = E(\hat{V}^{\hat{\pi}}) - V^*$$

$$= E(\max(\hat{Q}(s,a_1), \hat{Q}(s,a_2)) - V^*$$

Since (max.) is a convex operator

$$\boxed{E(\max(Q)) \geq \max E(Q)}$$

$$\geq \max[E(\hat{Q}(s,a_1)), E(\hat{Q}(s,a_2))] - V^*$$

For $n \to \infty$

$$\geq \max[c, c] - V^*$$

$$\geq c - V^*$$

$$\geq c - c$$

$$\geq 0$$

$$\therefore \boxed{\text{Bias}(\hat{V}^{\hat{\pi}}) \geq 0} \quad (= 0 \text{ when } n \to \infty)$$

$\therefore$ $\hat{V}^{\hat{\pi}}$ is a biased estimate of V^*

(c) $a_1 \rightarrow$ constant reward c

$a_2 \rightarrow c + N(-0.2, 1)$

~~E[a1]~~ $E(r|a_1) = c$; $E(r|a_2) = E\left(\frac{c + N(-0.2,1)}{a_2}\right)$

$E(r|a_2) = \underline{\underline{c - 0.2}}$

$\because E(r|a_1) > E(r|a_2)$

$\therefore a_1$ is a better action

b) No. TD algorithms might not always favour the best in expectation especially when trained on finite samples.

→ It can happen that in every episode trained we might earn a reward of $c + \varepsilon$; $(\varepsilon > 0)$ for action 'a_2'

∴ ~~Here~~ ~~the~~

Although the ~~change~~ chance for this to happen in all episodes is less but is not impossible

→ In this case $Q(s, a_1) = c$

$Q(s, a_2) = c + \varepsilon > c$

The algorithm might prefer '$\underline{\underline{a_2}}$'.

— x —