Mid Term Exam - II solutions

1 a) $p(\text{head}) = p$

$p(\text{head second time on } 5^{th} \text{ toss})$

$p(\text{head on } 5^{th} \text{ toss}) = p$ — (1)

For head to occur second time on 5^{th} toss, it should have occured only once on 1st 4 times

$p(\text{One head in first 4 tosses})$

$= {}^4C_1\, p(1-p)^3$ — (2)

- ${}^4C_1$: Choosing 1 out of 4
- p: head once
- $(1-p)^3$: tail 3 times

$\therefore p(\text{head second times on } 5^{th} \text{ toss}) = {}^4C_1\, p(1-p)^3 \times p$

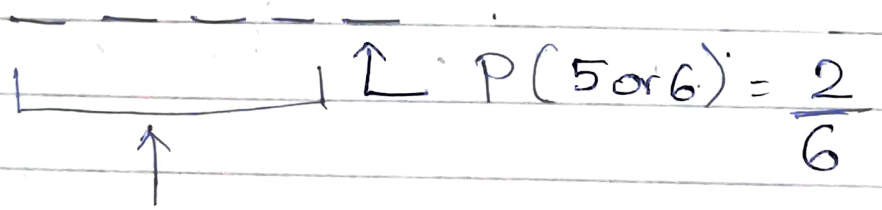
[from (1) & (2)] $= \boxed{4p^2(1-p)^3}$

1b) Dice (fair) rolled repeatedly till atleast 1-5 and One-6

P(5 rolls) = ?

The roll stops if 5 and 6 are rolled once.

∴ 5 (or) 6 should be rolled on 5th roll
and 6 (or) 5 should be rolled in 1st 4 rolls respectively



One place can be 5 or 6 = $\frac{1}{6}$

Chossing 1 place out of 4 = $^4C_1$

Other 3 places can have, 5 values except one in 5th place.

= $\frac{5}{6}$

$$\therefore p(5 \text{ rolls}) = {}^{4}C_1 \left(\frac{1}{6}\right)\left(\frac{5}{6}\right)^3 \times \frac{2}{6}$$

$$= 4 \times (0.167) \times (0.83)^3 \times (0.33)$$

$$= 0.126.$$

2) 5 Arms

Arm 1 $p_1 = 1/4$ $\quad R_{t+1} = \begin{cases} 1 - 1/4 = 3/4 & 1/4 : \text{Prob} \\ 0 & 3/4 \end{cases}$

Arm 2 $p_2 = 1/3$ $\quad R_{t+1} = \begin{cases} 1 - 1/3 = 2/3 & 1/3 : \text{prob} \\ 0 & 2/3 \end{cases}$

Arm 3 $p_3 = 1/2$ $\quad R_{t+1} = \begin{cases} 1 - 1/2 = 1/2 & 1/2 \text{ prob} \\ 0 & 1/2 \end{cases}$

Arm 4 $p_4 = 2/3$ $\quad R_{t+1} = \begin{cases} 1 - 2/3 = 1/3 & 2/3 \text{ Prob} \\ 0 & 1/3 \end{cases}$

Arm 5 $P_5 = 3/4$ $R_{t+1} = \begin{cases} 1-3/4 = 1/4 & 3/4 \text{ Prob} \\ 0 & 1/4 \end{cases}$

a) $E(\text{Arm } 1) = \frac{3}{4} \times \frac{1}{4} + 0 = \frac{3}{16} = 0.19$

$E(\text{Arm } 2) = \frac{2}{3} \times \frac{1}{3} = \frac{2}{9} = 0.22$

$E(\text{Arm } 3) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} = 0.25$

$E(\text{Arm } 4) = \frac{1}{3} \times \frac{2}{3} = \frac{2}{9} = 0.22$

$E(\text{Arm } 5) = \frac{1}{4} \times \frac{3}{4} = \frac{3}{16} = 0.19$

Arm 3 gives maximum expected record

b) Maximize $E[R^i_{t+1}]$, find value p_i

$$E[R^i_{t+1}] = p_i(1-p_i)$$

$$\because R^i_{t+1} = \begin{cases} 1-p_i & \text{with prob } p_i \\ 0 & \text{otherwise} \end{cases}$$

$$\frac{\partial}{dp_i} E[R^i_{t+1}] = \frac{d}{dp_i}(p_i - p_i^2) = 1 - 2p_i = 0$$

$$\Rightarrow p_i = \frac{1}{2}$$

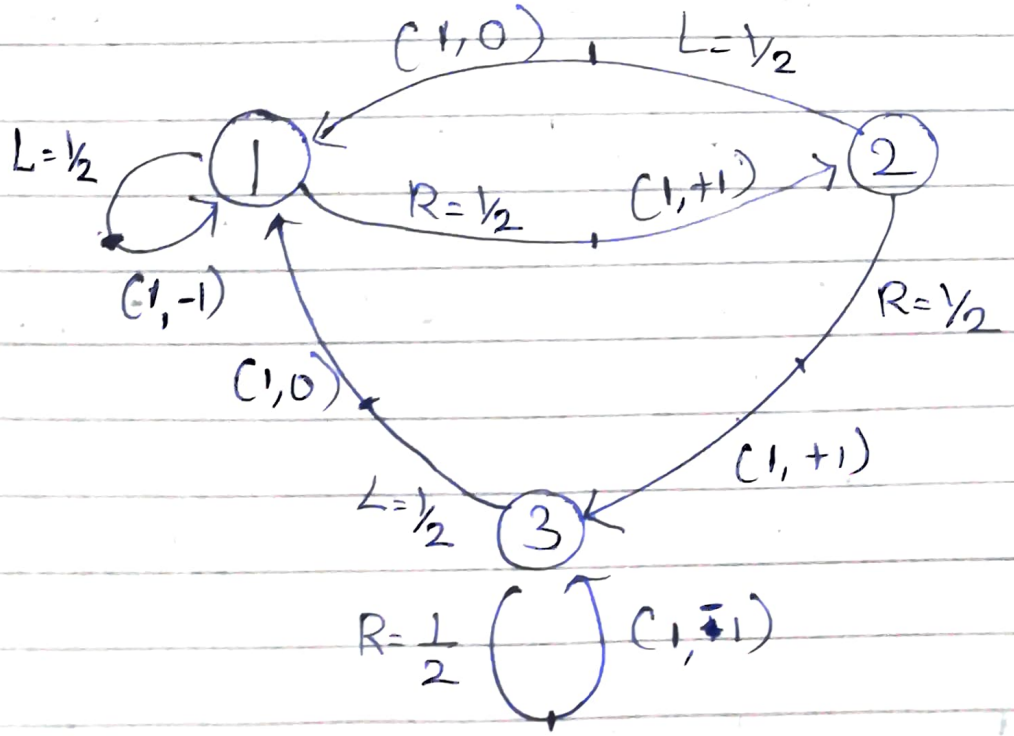
Also, $\frac{d^2}{dp} E[R^i_{t+1}] = -2 < 0 \Rightarrow$ Maximum

$\therefore p_i = \frac{1}{2}$ that maximises expected reward

3) a)

Current State (s)	Action (a)	Reward (r)	Next state (s')
1	Left	-1	1
1	Right	+1	2
2	Left	0	1
2	Right	+1	3
3	Left	0	1
3	Right	-1	3

b)



c) $V_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s', r | s, a) [r + \gamma V(s')]$

$V_\pi(1) = \frac{1}{2}$ [Value from action Left + Value from action Right] $\gamma = 0.9$

$V_\pi(1) = \frac{1}{2} [1(-1 + 0.9 V_\pi(1)) + 1(+1 + 0.9 V_\pi(2))]$

$2 V_\pi(1) = 0.9 V_\pi(1) + 0.9 V_\pi(2)$

$2.22 V_\pi(1) = V_\pi(1) + V_\pi(2)$

$\boxed{1.22 V_\pi(1) = V_\pi(2)} \rightarrow (1)$

$V_\pi(2) = \frac{1}{2}$ [Value from action Left + Value from action Right]

$V_\pi(2) = \frac{1}{2} [(0 + 0.9 V_\pi(1)) + (+1 + 0.9 V_\pi(3))]$

$$2V_{\pi}(2) = 0.9V_{\pi}(1) + 0.9V_{\pi}(3) + 1$$

From (1), $2.44\,V_{\pi}(1) - 0.9V_{\pi}(1) = 0.9V_{\pi}(3) + 1$

$$(1.54V_{\pi}(1) - 1)/0.9 = V_{\pi}(3)$$

$$\boxed{1.71\,V_{\pi}(1) - 1.11 = V_{\pi}(3)} \rightarrow (2)$$

$V_{\pi}(3) = \frac{1}{2}$ [Value from action left + Value from action Right]

$$V_{\pi}(3) = \frac{1}{2}[(0 + 0.9V_{\pi}(1)) + (-1 + 0.9V_{\pi}(3))]$$

$$2V_{\pi}(3) = 0.9V_{\pi}(1) + 0.9V_{\pi}(3) - 1$$

$$1.1V_{\pi}(3) = 0.9V_{\pi}(1) - 1$$

From (2), $1.1(1.71V_{\pi}(1) - 1.11) = 0.9V_{\pi}(1) - 1$

$$1.881V_{\pi}(1) - 1.221 = 0.9V_{\pi}(1) - 1$$

$V_\pi(1) = 0.221$

In ① $V_\pi(2) = 0.270$

In ② $V_\pi(3) = -0.732$

4) a) $Q_\pi(s,a) = E_\pi[G_t \mid S_t = s, A_t = a]$

$Q_\pi(s,a) = E_\pi[R_{(t+1)} + \gamma R_{(t+2)} + \ldots \mid S_t = s, A_t = a]$

$\tilde{Q}_\pi(s,a) = E_\pi[R_{(t+1)} + C + \gamma[R_{(t+2)} + C] + \ldots \mid S_t = s, A_t = a]$

$\tilde{Q}_\pi(s,a) = E_\pi[(R_{(t+1)} + \gamma R_{(t+2)} + \gamma^2 R_{(t+3)} + \ldots$

$+ C + \gamma C + \gamma^2 C + \ldots) \mid S_t = s, A_t = a]$

$\tilde{Q}_\pi(s,a) = E_\pi\left[\left(G_t + \frac{C}{1-\gamma}\right) \mid S_t = s, A_t = a\right]$

$\because \gamma \in (0,1)$

$$\tilde{Q}_\pi(s,a) = Q_\pi(s,a) + \frac{C}{1-\gamma} \qquad \left[\because \frac{C}{1-\gamma} \text{ is constant}\right]$$

$\therefore$ Q value is changed by $\left(\frac{+C}{1-\gamma}\right)$

4b) If ϵ-greedy used to update policy.

Actions not optimal under π (suggested) will be chosen. These actions rewards will not be translated by '+C' as they are not suggested by π.

Thus, actions obtained will differ between two cases, due to partial advantage to selected actions.

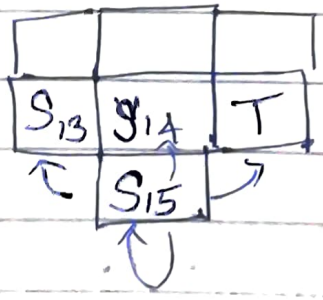
4c) If task is discounted and episodic

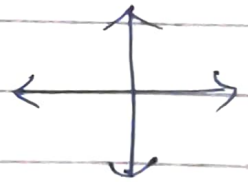
$$\tilde{Q}(s,a) = E\Big(R_{(t+1)} + \gamma R_{(t+2)} + \gamma^2 R_{(t+3)} + \dots \gamma^{T-1} R_T + C + \gamma C + \gamma^2 C + \dots + \gamma^{T-1} C \;\Big|\; S_t = s,\ A_t = a\Big)$$

$$\tilde{Q}(s,a) = E\left[G_t + \frac{C(1-\gamma^T)}{1-\gamma}\right]$$

$$\tilde{Q}(s,a) = \tilde{Q}(s,a) + \frac{C(1-\gamma^T)}{(1-\gamma)}$$

5)





Equiprobable

$R_t = -1$ ∀ transitions

$\gamma = 1$

5a) $V_{\pi}(15) = \frac{1}{4}$ [Value from action Left +
" " " Up +
" " " Right +
" " " Down]

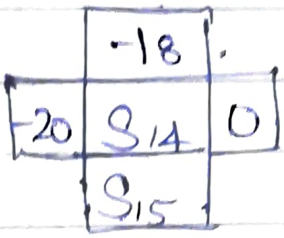
$$V_{\pi}(15) = \frac{1}{4}\left[-1 + \gamma(-20) - 1 + \gamma(-14) + \gamma(0) - 1 \right.$$
$$\left. -1 + \gamma V_{\pi}(15) \right]$$

$$4V_{\pi}(15) = -4 - .34 + V_{\pi}(15)$$

$$3V_{\pi}(15) = -38 \Rightarrow V_{\pi}(15) = -12.67$$

5b) If dynamics of S_{14} change, other states values will also be affected.

For simplicity, we consider only value changes of state: 14 & 15



$$V(S_{14}) = \frac{1}{4}[-1-18+0-1-1+V(S_{15})-1-20]$$

$$4V(S_{14}) = V(S_{15}) - 4 - 38$$

$$\boxed{4V(S_{14}) - V(S_{15}) = -42} \rightarrow (1)$$

$$V(S_{15}) = \frac{1}{4}[-1+V(S_{14})-1+0+V(S_{15})-1-1-20]$$

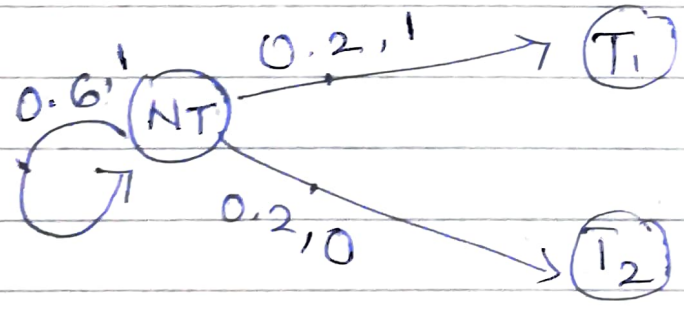
$$4V(S_{15}) = V(S_{14}) + V(S_{15}) - 4 - 20$$

$$\boxed{3V(S_{15}) - V(S_{14}) = -24} \rightarrow (2)$$

Solving (1), (2)

$$V(S_{14}) = -13.64 \; ; \quad V(S_{15}) = -12.55$$

(6)



Ep.1 = NT $\rightarrow$ ... $\rightarrow T_1$ R = 10

Ep.2 = NT $\rightarrow$ ... $\rightarrow T_2$ R = 4

a) Episode 1.

NT $\xrightarrow{1}$ NT $\xrightarrow{1}$ NT $\xrightarrow{1}$ NT $\xrightarrow{1}$ NT $\xrightarrow{1}$

NT $\xrightarrow{1}$ NT $\xrightarrow{1}$ NT $\xrightarrow{1}$ NT $\xrightarrow{1}$ NT $\xrightarrow{1}$ T_1

Episode 2,

NT $\xrightarrow{1}$ NT $\xrightarrow{1}$ NT $\xrightarrow{1}$ NT $\xrightarrow{1}$ NT $\xrightarrow{0}$ T2

b) First Visit $\quad \frac{G_{Episode1} + G_{Episode2}}{2}$

Every Visit $\quad G_{ij}$: i^{th} Episode j^{th} Occurance of NT

$(G_{11}+G_{12}+G_{13}+G_{14}+G_{15}+G_{16}+G_{17}+G_{18}+G_{19}+G_{10}$
$+ G_{21}+G_{22}+G_{23}+G_{24}+G_{25}) / 15$

c) First Visit : 2 Estimates ; Every Visit - 15 estimates

d) i) First visit Estimator : $(G_{epi1} + G_{epi2}) / 2$

$= (10+4)/2 = 7$

ii) Every Visit Estimator

$(10+9+\ldots+1) + (4+3+2+1+0) / 15$

$$= \left(\frac{10 \times 11}{7} + \frac{4 \times 5}{2} \right) / 15 = 65/15 = 4.33$$

7) G_i has weight e^{W_i}

$$V_n = \frac{\sum_{k=1}^{n} e^{W_k} G_k}{\sum_{k=1}^{n} e^{W_k}}, \quad n \geq 1$$

a)

$$V_{n+1} = \frac{\sum_{k=1}^{n+1} e^{W_k} G_k}{\sum_{k=1}^{n+1} e^{W_k}}$$

$$= \frac{\sum_{k=1}^{n} e^{W_k} G_k + e^{W_{n+1}} G_{n+1}}{\sum_{k=1}^{n+1} e^{W_k}}$$

$$V_{n+1} = \frac{\sum_{k=1}^{n} e^{\mu_k} G_k}{\sum_{k=1}^{n+1} e^{\mu_k}} + \frac{e^{\mu_{n+1}} G_{n+1}}{\sum_{k=1}^{n+1} e^{\mu_k}}$$

$$V_{n+1} = \frac{\sum_{k=1}^{n} e^{\mu_k} G_k}{\sum_{k=1}^{n} e^{\mu_k} \cdot \sum_{k=1}^{n+1} e^{\mu_k}} \cdot \sum_{k=1}^{n} e^{\mu_k} + \frac{e^{\mu_{n+1}} G_{n+1}}{\sum_{k=1}^{n+1} e^{\mu_k}}$$

$$V_{n+1} = V_n \frac{\sum_{k=1}^{n+1} e^{\mu_k} - e^{\mu_{n+1}}}{\sum_{k=1}^{-n+1} e^{\mu_k}} + \frac{e^{\mu_{n+1}} G_{n+1}}{\sum_{k=1}^{n+1} e^{\mu_k}}$$

$$V_{n+1} = V_n \left[1 - \frac{e^{\mu_{n+1}}}{\sum_{k=1}^{n+1} e^{\mu_k}} \right] + \frac{e^{\mu_{n+1}} G_{n+1}}{\sum_{k=1}^{n+1} e^{\mu_k}}$$

$$V_{n+1} = V_n + \frac{e^{W_{n+1}}}{\sum_{k=1}^{n+1} e^{W_k}} \left[G_{n+1} - V_n \right]$$

b) Off-policy MC control Algorithm

Initialize $\forall s \in S, a \in A(s)$

→ $Q(s,a) \in \mathbb{R}$ arbitarily

→ $C(s,a) \leftarrow 0$

→ $\pi(s) \leftarrow \underset{a}{\text{argmax}}\, Q(s,a)$

Loop for each episode:

$b \leftarrow$ any soft policy

Generate an episode using b: $S_0, A_0, R_1 \ldots S_{T-1}, A_{T-1}, R_T$

$G \leftarrow 0$

$W \leftarrow 0$

Loop for each step in episode $t = T-1, T-2 \ldots 0$

$G \leftarrow \gamma G + R_{t+1}$

$C(S_t, A_t) \leftarrow C(S_t, A_t) + W$

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{e^{W}}{C(S_t, A_t)} \left[G - Q(S_t, A_t) \right]$$

$$\pi(S_t) \leftarrow \arg\max_a Q(S_t, a)$$

If $A_t \neq \pi(S_t)$ then exit innerloop (Go to next episode)

$$W \leftarrow W + \log\left(\frac{\pi(A_t|S_t)}{b(A_t|S_t} \right)$$

(c)

$$e^{W_i} = e_{t_i : T(t)-1}$$

$$W_i = \ln e_{t_i : T(t)-1}$$

$$W_i = \ln \left[\prod_{k=t_i}^{T_i - 1} \frac{\pi(A_k|S_k)}{b(A_k|S_k)} \right]$$

$$W_i = \sum_{k=t_i}^{T_i - 1} \ln \left(\frac{\pi(A_k|S_k)}{b(A_k|S_t)} \right)$$