CS5500: Reinforcement Learning

Assignment - 1

5

- Abburi Venkata Sai Mahesh

- CS18BTECH11001

1. Given that

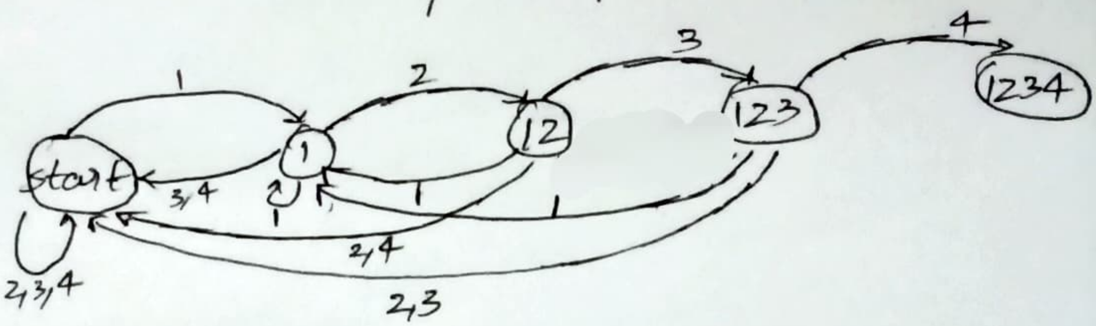
dice has 4 faces marked {'1', '2', '3', '4'}

The required pattern is '1234'.

a) <u>state space</u> : {start, 1, 12, 123, 1234}

Here start is starting state of process

1, 12, 123, 1234 are states where these respective patterns are observed



<u>Transition probabilities:</u>

	start	1	12	123	1234
start	3/4	1/4	0	0	0
1	2/4	1/4	1/4	0	0
12	2/4	1/4	0	1/4	0
123	2/4	1/4	0	0	1/4
1234	0	0	0	0	1

<u>Terminal states:</u>

1234 is the terminal state of the MRP

remaining are non-terminal states.

b) Reward function

This is similar to the counting of no. of tosses required for a given pattern. So the reward function can be given as

$$R = \begin{bmatrix} -1 \\ -1 \\ -1 \\ -1 \\ 0 \end{bmatrix}$$ [∵ i.e., -1 for non-terminal state, 0 for terminal state]

Discount factor:

The discount factor is 1.

Average number of tosses:

For calculating the no. of tosses required for the final pattern we only consider the non-terminal states.

So $$P = \begin{bmatrix} 3/4 & 1/4 & 0 & 0 \\ 2/4 & 1/4 & 1/4 & 0 \\ 2/4 & 1/4 & 0 & 1/4 \\ 2/4 & 1/4 & 0 & 0 \end{bmatrix} \quad R = \begin{bmatrix} -1 \\ -1 \\ -1 \\ -1 \end{bmatrix}$$

from the Bellman equation:

$$V = (I - \gamma P^{-1})R$$

$$= \begin{bmatrix} 1/4 & -1/4 & 0 & 0 \\ -2/4 & 3/4 & -1/4 & 0 \\ -2/4 & -1/4 & 1 & -1/4 \\ -2/4 & -1/4 & 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} -1 \\ -1 \\ -1 \\ -1 \end{bmatrix}$$

$$= \begin{bmatrix} 172 & 64 & 16 & 4 \\ 168 & 64 & 16 & 4 \\ 160 & 60 & 16 & 4 \\ 128 & 48 & 12 & 4 \end{bmatrix} \begin{bmatrix} -1 \\ -1 \\ -1 \\ -1 \end{bmatrix} = \begin{bmatrix} -256 \\ -252 \\ -240 \\ -192 \end{bmatrix}$$

∴ The average no of days for pattern 1234 = 256

2. Given that

$V^n(s)$ = state function at time step n

$Q^n(s)$ = action function at time step n

a) To evaluate the function $V^N(s)$

as N is the last time step, the value function will be the immediate reward.

$V^N(1) = 3(1)^2 + 5 = 8$

$V^N(2) = 3(2)^2 + 5 = 17$

$V^N(3) = 3(3)^2 + 5 = 32$

$V^N(4) = 3(4)^2 + 5 = 53$.

b) From the Bellman Equation

$$Q^{N-1}(s,a) = E\left[R(s,a,s') + \gamma V^N(s')\right]$$

$$= \sum_{s' \in S} P^a_{ss'} \left[R(s,a,s') + V^N(s')\right] \quad [\because \gamma = 1]$$

But given that the dice is a fair dice and the Reward function as $3s^2+5$ which does not depend on previous state

$$R(s,a,s') = R(s) \text{ and } P_{ss'} = \frac{1}{4}$$

$$\Rightarrow Q^{N-1}(s,a) = R(s) + \sum_{s' \in S} \frac{1}{4} \cdot V^N(s')$$

$$= R(s) + \frac{1}{4} \sum_{s' \in S} V^N(s').$$

if a = Quit, then there will be no s' as we get the immediate Reward

$$Q^{N-1}(s,a) = 3s^2 + 5 + 0 = 3s^2 + 5$$

i.e $\boxed{Q^{N-1}(s, \text{'Quit'}) = 3s^2 + 5 \quad \forall s \in \{1,2,3,4\}}$

When a = 'Continue', the immediate reward is zero

$$Q^{N-1}(s,a) = 0 + \sum_{s' \in S} \frac{1}{4} V^{N}(s')$$

$$= \frac{1}{4}\left[V^{N}(1) + V^{N}(2) + V^{N}(3) + V^{N}(4)\right]$$

$$= \frac{1}{4}\left[8+17+32+53\right]$$ [∵ calculated in previous subquestion]

$$= 27.5$$

$$\boxed{\therefore Q^{N-1}(s, \text{'Continue'}) = 27.5 \quad \forall\, s \in \{1,2,3,4\}}$$

c) It is mentioned that the value of a state at any intermediate time is equal to the best action value possible for that state at the time.

So $V^{N-1}(s) = \max_{a \in A} Q^{N-1}(s,a)$

$V^{N-1}(1) = \max(Q^{N-1}(1, \text{'Continue'}), Q^{N-1}(1, \text{'Quit'})) = \max(27.5, 8) = 27.5$

$V^{N-1}(2) = \max(Q^{N-1}(2, \text{'Continue'}), Q^{N-1}(2, \text{'Quit'})) = \max(27.5, 17) = 27.5$

$V^{N-1}(3) = \max(Q^{N-1}(3, \text{'Continue'}), Q^{N-1}(3, \text{'Quit'})) = \max(27.5, 32) = 32$

$V^{N-1}(4) = \max(Q^{N-1}(4, \text{'Continue'}), Q^{N-1}(4, \text{'Quit'})) = \max(27.5, 53) = 53.$

d) $V^{n-1}(s) = \max_{a \in A} Q^{n-1}(s,a)$

$$= \max_{a \in A}\left[\sum_{s' \in S} P^{a}_{ss'}\left[R(s,a,s') + V^{n}(s')\right]\right]$$

$$= \max\Big(\underbrace{3s^2+5}_{\text{Quit}}, \underbrace{\sum_{s' \in S} \frac{1}{4} V^{n}(s')}_{\text{Continue}}\Big)$$

$$\boxed{\therefore V^{n-1}(s) = \max\left(3s^2+5, \frac{1}{4}\sum_{s' \in S} V^{n}(s')\right)}$$

e) $Q^{n-1}(s, \text{'Continue'}) = \frac{1}{4} \sum_{s' \in S} V^{n}(s')$ [$\because$ from subquestion c]

$= \frac{1}{4} \sum_{s' \in S} \left[\max_{a \in A} Q^{n}(s', a) \right]$

$= \frac{1}{4} \sum_{s' \in S} \left[\max \left(Q^{n}(s', \text{'Continue'}), Q^{n}(s', \text{'Quit'}) \right) \right]$

$$\boxed{\therefore Q^{n-1}(s, \text{'Continue'}) = \frac{1}{4} \sum_{s' \in S} \left[\max \left(Q^{n}(s, \text{'Continue'}), 3s'^2 + 5 \right) \right]}$$

f) Let's observe the values of $V^{n}(s)$ $\quad 2 \leq n \leq N$

$V^{N}(s) = \{8, 17, 32, 53\}$

$V^{N-1}(s) = \max(3s^2+5, 27.5) = \{27.5, 27.5, 32, 53\}$

$V^{N-2}(s) = \max(3s^2+5, \frac{1}{4} \sum V^{N-1}(s))$

$= \max(3s^2+5, \frac{1}{4}(140))$

$= \max(3s^2+5, 35)$

$= \{35, 35, 35, 53\}$

$V^{N-3}(s) = \max(3s^2+5, \frac{1}{4} \sum V^{N-2}(s))$

$= \max(3s^2+5, 39.5)$

$= \{39.5, 39.5, 39.5, 53\}$

$V^{N-4}(s) = \max\left(3s^2+5, \underbrace{\frac{1}{4}(39.5 \times 3 + 53)}_{<53}\right)$

$= \{x, x, x, 53\}$ where $x > 32$ & $x < 53$

So we can observe that when $n \leq N-2$

$V^{n}(s) = \begin{cases} 53 & \text{if } s = 4 \Rightarrow \text{action = 'Quit'} \\ x & \text{otherwise} \Rightarrow \text{action = 'Continue'} \end{cases}$ where $x < 53$

Hence the optimal policy is

for $n = N$ $\pi(s) = \{ \text{drop} \ \forall s \in \{1,2,3,4\}$ [as it is last time step]

for $n = N-1$ $\pi(s) = \begin{cases} \text{continue} & \forall s \in \{1,2\} \\ \text{Quit} & \text{otherwise} \end{cases}$

for $n \leq N-2$ $\pi(s) = \begin{cases} \text{continue} & \forall s \in \{1,2,3\} \\ \text{Quit} & \text{if } s = 4 \end{cases}$

g) As we have seen in the above that the optimal policy for $n = N$, $n = N-1$, $n \leq N-2$ are not equal and so we can say that the optimal policy is not stationary. Even though the policy is same for $n \leq N-2$, it differed from $n = N-1$ & $n = N$. So it is a <u>non-stationary</u> optimal policy.

3. Given that

$M = < S, A, P, R, \gamma >$, $|S| < \infty$, $|A| < \infty$, $\gamma \in [0, 1)$

$\hat{M} = < S, A, P, \hat{R}, \gamma >$

$$|R(s, a, s') - \hat{R}(s, a, s')| = \epsilon$$

policy π, V^{π} and $\hat{V}^{\pi}$ are value functions for M & $\hat{M}$

a) we know that

$$V^{\pi}(s) = E_{\pi}\left[\sum_{k=0}^{\infty} \gamma^{k} r_{t+k+1} \mid s_t = s\right]$$

$$\hat{V}^{\pi}(s) = E_{\pi}\left[\sum_{k=0}^{\infty} \gamma^{k} \hat{r}_{t+k+1} \mid s_t = s\right]$$

$$\Rightarrow V^{\pi}(s) - \hat{V}^{\pi}(s) = E_{\pi}\left[\sum_{k=0}^{\infty} \gamma^{k} (r_{t+k+1} - \hat{r}_{t+k+1}) \mid s_t = s\right]$$

$$\Rightarrow |V^{\pi}(s) - \hat{V}^{\pi}(s)| = \left|E_{\pi}\left[\sum_{k=0}^{\infty} \gamma^{k} (r_{t+k+1} - \hat{r}_{t+k+1}) \mid s_t = s\right]\right|$$

$$\leq \left[\sum_{k=0}^{\infty} \gamma^{k} \cdot \epsilon\right]$$

$$\leq \frac{1}{1-\gamma} \cdot \epsilon$$

$$\boxed{\therefore |V^{\pi}(s) - \hat{V}^{\pi}(s)| \leq \frac{\epsilon}{1-\gamma}}$$

b) Let us consider that π_1 be the optimal policy for MDP 'M'

i.e $V_* = V^{\pi_1}$

Let us consider that π_2 be the optimal policy for MDP '$\hat{M}$'

i.e $\hat{V}_* = \hat{V}^{\pi_2}$

from the before subquestion

$$|V^{\pi_1}(s) - \hat{V}^{\pi_1}(s)| \leq \frac{\epsilon}{1-\gamma} \quad - (1)$$

$$|V^{\pi_2}(s) - \hat{V}^{\pi_2}(s)| \leq \frac{\epsilon}{1-\gamma} \quad - (2)$$

From the definition of optimal policy

$$V^{\pi_1}(s) \geq V^{\pi_2}(s) \quad -③ \quad [\because \pi_1 \text{ is optimal for } V$$

$$\hat{V}^{\pi_1}(s) \leq \hat{V}^{\pi_2}(s) \quad -④ \quad \pi_2 \text{ is optimal for } \hat{V}]$$

Now consider ②

$$-\frac{\epsilon}{1-\gamma} \leq V^{\pi_2}(s) - \hat{V}^{\pi_2}(s) \leq \frac{\epsilon}{1-\gamma}$$

$$V^{\pi_2}(s) \geq \hat{V}^{\pi_2}(s) - \frac{\epsilon}{1-\gamma}$$

$$\Rightarrow V^{\pi_1}(s) \geq \hat{V}^{\pi_2}(s) - \frac{\epsilon}{1-\gamma} \quad [\because \text{from } ③]$$

$$\Rightarrow \boxed{\hat{V}^{\pi_2}(s) - V^{\pi_1}(s) \leq \frac{\epsilon}{1-\gamma}} \quad -⑤$$

Now consider ①

$$-\frac{\epsilon}{1-\gamma} \leq V^{\pi_1}(s) - \hat{V}^{\pi_1}(s) \leq \frac{\epsilon}{1-\gamma}$$

$$\hat{V}^{\pi_1}(s) \geq V^{\pi_1}(s) - \frac{\epsilon}{1-\gamma}$$

$$\Rightarrow \hat{V}^{\pi_2}(s) \geq V^{\pi_1}(s) - \frac{\epsilon}{1-\gamma}$$

$$\Rightarrow \boxed{\hat{V}^{\pi_2}(s) - V^{\pi_1(s)} \geq -\frac{\epsilon}{1-\gamma}} \quad -⑥$$

from ⑤ and ⑥ we can conclude that

$$-\frac{\epsilon}{1-\gamma} \leq \left(\hat{V}^{\pi_2}(s) - V^{\pi_1(s)}\right) \leq \frac{\epsilon}{1-\gamma}$$

$$\Rightarrow \left|\hat{V}^{\pi_2}(s) - V^{\pi_1}(s)\right| \leq \frac{\epsilon}{1-\gamma}$$

$$\Rightarrow \left|\hat{V}_* - V_*\right| \leq \frac{\epsilon}{1-\gamma} \quad [\because \text{from the beginning}]$$

$$\boxed{\therefore \left|\hat{V}_* - V_*\right| \leq \frac{\epsilon}{1-\gamma}}$$

c) Given that

$M < S, A, P, R, \gamma >$

$\hat{M} < S, A, P, \hat{R}, \gamma >$

$|R(s,a,s') - \hat{R}(s,a,s')| = \epsilon$

Let us consider a system with ~~two~~ three states and define the Reward functions as follows

$R(s_1, a, s_2) = \epsilon$ $\quad \hat{R}(s_1, a, s_2) = 0$ $\quad S = \{s_1, s_2, s_3\}$

$R(s_1, a, s_3) = 0$ $\quad \hat{R}(s_1, a, s_3) = \epsilon$

Here the $|R(s,a,s') - \hat{R}(s,a,s')| = \epsilon$ is maintained along with the same states, Action, probability table, discount factor.

But we can able to see that the optimal policy for M is is from s_1 to s_2. where as the optimal policy for $\hat{M}$ is from s_1 to s_3.

So the optimal policy Need not be same.

$\boxed{\therefore \text{M and } \hat{M} \text{ can have different optimal policy}}$

4. a) discount factor (γ):

The discount factor depends on the distance to the terminal states. It means that, if the distant factor (γ) is small, it is preffered to select the closest exit point. else if the distant factor (γ) is high, it is preffered to select the distant unit.

Noise factor (η):

When there is more risk in moving to the exit point, the success probability should be high and therefore choosing low noise factor. Where as if the risk is low the success probability may be low and we can choose a high noise factor.

- close exit but risk the cliff

  γ is low = 0.1

  η is ~~high~~ low = 0

- distant exit but risk the cliff

  γ is high = 0.9

  η is low = 0

- close exit by avoiding the cliff

  γ is low = 0.1

  η is high = 0.5

- distant exit by avoiding the cliff

  γ is high = 0.9

  η is high = 0.5

5. a) Given that

$$\|V_{k+1} - V_k\|_\infty \leq \epsilon \ , \quad \epsilon > 0$$

$$\gamma \|V_{k+1} - V_k\|_\infty \leq \gamma\epsilon$$

$$\|\gamma V_{k+1} - \gamma V_k\|_\infty \leq \gamma\epsilon$$

$$\|\gamma V_{k+1} - \gamma V^\pi + \gamma V^\pi - \gamma V_k\|_\infty \leq \gamma\epsilon$$

$$\|\gamma(V_{k+1} - V^\pi) - \gamma(V_k - V^\pi)\|_\infty \leq \gamma\epsilon \quad - (1)$$

we know that

$$V_{k+1} = \max_{a \in A} [R^a + \gamma P^a V_k]$$

Bellman optimality equation is a contraction mapping

$$L: V \to V$$

$$L(V) = \max_{a \in A} [R^a + \gamma P^a V]$$

$$L(V_k) = V_{k+1}$$

For a contraction mapping, we know that

$$\|L(V) - L(\omega)\|_\infty \leq \gamma \|V - \omega\|$$

$$\|L(V_k) - L(V^\pi)\|_\infty \leq \gamma \|V_k - V^\pi\|_\infty \quad - (2)$$

Now consider (1)

From triangle inequality

$$\gamma \|V_k - V^\pi\|_\infty - \gamma \|V_{k+1} - V^\pi\|_\infty \leq \gamma\epsilon$$

$$\|L(V_k) - L(V^\pi)\|_\infty - \gamma \|V_{k+1} - V^\pi\|_\infty \leq \gamma\epsilon$$

$$\|V_{k+1} - V^\pi\|_\infty - \gamma \|V_{k+1} - V^\pi\|_\infty \leq \gamma\epsilon$$

$$\Rightarrow (1-\gamma) \|V_{k+1} - V^\pi\|_\infty \leq \gamma\epsilon$$

$$\boxed{\therefore \|V_{k+1} - V^{\pi}\|_\infty \leq \frac{\gamma \epsilon}{1-\gamma}}$$

Hence proved.

b) Consider equation ② from subquestion a

$$\|L(V_k) - L(V^{\pi})\|_\infty \leq \gamma \|V_k - V^{\pi}\|_\infty$$

$$\Rightarrow \|V_{k+1} - V^{\pi}\|_\infty \leq \gamma \|V_k - V^{\pi}\|_\infty$$

$$\Rightarrow \|V_{k+1} - V^{\pi}\|_\infty \leq \gamma \left[\gamma \|V_{k-1} - V^{\pi}\|_\infty\right]$$

$$\Rightarrow \|V_{k+1} - V^{\pi}\|_\infty \leq \gamma^2 \|V_{k-1} - V^{\pi}\|_\infty$$

Similarly after k iterations

$$\|V_{k+1} - V^{\pi}\|_\infty \leq \gamma^k \|V_1 - V^{\pi}\|_\infty$$

$$\boxed{\therefore \|V_{k+1} - V^{\pi}\|_\infty \leq \gamma^k \|V_1 - V^{\pi}\|_\infty}$$

Hence proved.

c) Given that $L(v) = \max\limits_{a \in A} [R^a + \gamma P^a v]$

and $u \leq v$

$$\Rightarrow P^a u \leq P^a v \quad \forall a \in A$$

$$\Rightarrow \gamma P^a u \leq \gamma P^a v \quad \forall a \in A$$

$$\Rightarrow R^a + \gamma P^a u \leq R^a + \gamma P^a v \quad \forall a \in A$$

$$\Rightarrow \max_{a \in A} [R^a + \gamma P^a u] \leq \max_{a \in A} [R^a + \gamma P^a v]$$

$$\Rightarrow L(u) \leq L(v) \Rightarrow L \text{ is monotonic}$$

Hence proved.

6 a) Given P and Q are contractions defined on a normal vector space $\langle V, \|\cdot\| \rangle$.

$\Rightarrow \|P(u) - P(v)\| \le \gamma_P \|u - v\| \quad \forall\, u, v \in V$ — ①

$\Rightarrow \|Q(u) - Q(v)\| \le \gamma_Q \|u - v\| \quad \forall\, u, v \in V$ — ②

Now consider

$$\|P \circ Q(u) - P \circ Q(v)\| = \|P(Q(u)) - P(Q(v))\|$$

$$\le \gamma_P \|Q(u) - Q(v)\| \quad [\text{from ①}]$$

$$\le \gamma_P \cdot \gamma_Q \|u - v\| \quad [\text{from ②}]$$

$$\forall\, u, v \in V$$

Hence $P \circ Q$ is also a contraction on V.

Now consider

$$\|Q \circ P(u) - Q \circ P(v)\| = \|Q(P(u)) - Q(P(v))\|$$

$$= \gamma_Q \|P(u) - P(v)\| \quad [\text{from ②}]$$

$$= \gamma_Q \cdot \gamma_P \|u - v\| \quad [\text{from ①}] \quad \forall\, u, v \in V$$

Hence $Q \circ P$ is also a contraction on V

b) from (a), we can observe that the contraction coefficient for $P \circ Q$ and $Q \circ P$ are same and equal to $\gamma_P \cdot \gamma_Q$ where $\gamma_P \in [0,1)$ and $\gamma_Q \in [0,1)$

$\Rightarrow \gamma_P \cdot \gamma_Q \in [0,1)$ is a valid Lipshitz coefficient.

c) Given, $B = F \circ L$ where L is the Bellman optimality operator.

if we replace L with B in the value iteration algorithm the algorithm must converge to a unique solution. This implies that $F \circ L$ should be a contraction.

i.e $\|F \circ L(u) - F \circ L(v)\|_\infty \leq \gamma \|u - v\|_\infty \quad \exists \gamma \in [0,1)$
$\forall u, v \in V$

For the iterative algorithm to converge to an optimal fixed point V_*, the function value $F \circ L$ at V_* must be V_* as well

$$\Rightarrow F \circ L(V_*) = V_*$$

These two conditions are necessary for the value iteration algorithm to converge to a unique solution.