CS5500 : Reinforcement Learning

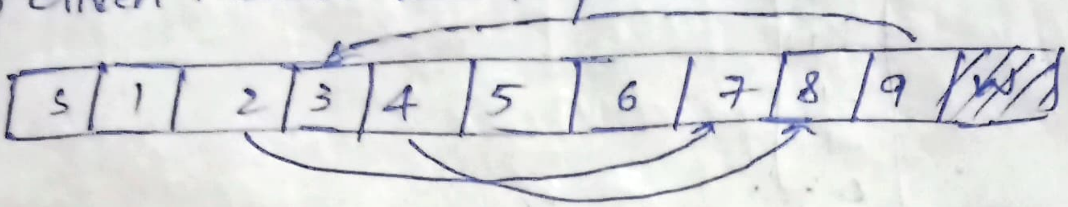
Exam No 1

- Abburi Venkata Sai Mahesh

- CS18BTECH11001

Problem 1:

a) Given Markov reward process is



States:

As 2, 4, 9 are states having either ladder and snake any die roll leading to these places will redirect to othe states. So we can ignore these state (The ans will be same even if we consider these states)

$\therefore S = \{S, 1, 3, 5, 6, 7, 8, W\}$

where W is terminal state

others are non terminal states.

Transition probability matrix

	S	1	3	5	6	7	8	W
S	0	0.25	1/4	0	0	1/4	1/4	0
1	0	0	1/4	1/4	0	1/4	1/4	0
3	0	0	0	1/4	1/4	1/4	1/4	0
5	0	0	1/4	0	1/4	1/4	1/4	0
6	0	0	1/4	0	0	1/4	1/4	1/4
7	0	0	1/4	0	0	1/4	1/4	1/4
8	0	0	1/4	0	0	0	1/2	1/4
W	0	0	0	0	0	0	0	1

b) Reward function

The reward is -1 for non terminal state
0 for terminal state

$$\therefore R = \begin{bmatrix} -1 \\ -1 \\ -1 \\ -1 \\ -1 \\ -1 \\ -1 \\ 0 \end{bmatrix}$$

discount factor : 1

Using Bellman equation

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

$$= \begin{bmatrix} -7.08333 \\ -7 \\ -6.6666 \\ -6.6666 \\ -5.3333 \\ -5.3333 \\ -5.3333 \end{bmatrix}$$ [∵ Considering only non-terminal states]

∴ The expected number of die throws $= 7.08$
~ 7

2 a) Given $M = \langle S, A, P, R, \gamma \rangle$

$$R(s,a) = R_1(s,a) + R_2(s,a)$$

Given value action function for policy π

$Q_1^{\pi}(s,a)$ with reward function $R_1(s,a)$

value action function $Q_2^{\pi}(s,a)$ with

reward function $R_2(s,a)$

So $$Q_1^{\pi}(s,a) = E_{\pi}\left(\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \middle| s_t = s, a_t = a\right) \quad r \sim R_1(s,a)$$

$$Q_2^{\pi}(s,a) = E_{\pi}\left(\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \middle| s_t = s, a_t = a\right) \quad r \sim R_2(s,a)$$

we can say that $$Q^{\pi}(s,a) = Q_1^{\pi}(s,a) + Q_2^{\pi}(s,a)$$

But we cannot exactly say that

$$Q^*(s,a) = Q_1^*(s,a) + Q_2^*(s,a)$$

Because the variation of $Q_1^*(s,a)$ might not be propotional to $Q_2^*(s,a)$ so the optimal value will not be for same action

So it is not possible to combine the action value functions in a simpler manner.

b) Given

$M = \langle S, A, P, R, \gamma \rangle$

$f, g : S \times A \to \mathbb{R}$.

$(L_f)(s,a) = R(s,a) + \gamma P(s,a), V_f(s)$

where. $V_f(s) = \max_a f(\cdot, a)$.

Consider

$$\|L_f - L_g\|_\infty = \| \cancel{R(s,a)} + \gamma P(s,a) V_f(s) - \cancel{R(s,a)} + \gamma P(s,a) V_g(s) \|_\infty$$

$$= \| \gamma P(s,a) [V_f(s) - V_g(s)] \|_\infty$$

$$\leq \gamma \| V_f(s) - V_g(s) \|_\infty \quad [\because P(s,a) \leq 1]$$

Now consider

$$\|V_f - V_g\| = | \max_a f(\cdot, a) - \max_a g(\cdot, a) |$$

$$\leq \max_a \| f(\cdot, a) - g(\cdot, a) |$$

$$\leq \max_a \max_s \| f(s,a) - g(s,a) \|$$

$$\leq \|f - g\|_\infty$$

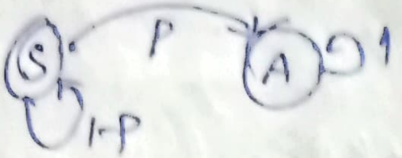
$$\Rightarrow \|L_f - L_g\|_\infty \leq \gamma \| V_f(s) - V_g(s) \|_\infty$$

$$\leq \gamma \|f - g\|_\infty$$

Hence proved.

3. Given Markov process - transition probabilities.

	S	A
S	1-P	P
A	0	1



a) This shows that A is a terminal state
if we start from state s

then $S \to A$ $\quad SA$

$S \to S \to A$ $\quad S^2A$

⋮

$S \to S \to S - S \to A$ $\quad S^K A$

i.e once we hit A we exit

∴ we can write general form of trajectory as $S^K A$ where $K \geq 1$ [∵ S repeated for K times and exited with A]

b) We calculate MC for first K trajectories

$S \to A$ $\quad r = 1$

$S \to S \to A$ $\quad r = 2$

⋮

$S \to S \to S \cdots S \to A$ $\quad r = K$

$$V(S) = \frac{\sum_{K=1}^{K}(K)}{K} = \frac{\frac{K(K+1)}{2}}{K} = \frac{K+1}{2}$$

$$\boxed{\therefore V(S) = \frac{K+1}{2}}$$

c) We consider every visit MC for first k trajectories

$S \to A \quad r = 1$

$S \to S \to A \quad r = 2+1$

$S \to S \to S \to A \quad r = 3+2+1 \cdots$

$\vdots$

$S \to S \to S \to \cdots S \to A \quad r = k + (k-1) + \cdots 1$

$$\therefore V(S) = \frac{\sum_{k=1}^{k} \frac{k(k+1)}{2}}{\sum_{k=1}^{k} k} = \frac{\frac{1}{2}\left[\frac{k(k+1)(2k+1)}{6} + \frac{k(k+1)}{2}\right]}{\frac{k(k+1)}{2}}$$

$$= \frac{1}{2}\left[\frac{2k+1}{3} + 1\right]$$

$$\boxed{\therefore V(S) = \frac{k+2}{3}}$$

d) True value of $V(S) = (I-P)^{-1} R$

$P = \begin{bmatrix} 1-p & p \\ 0 & 1 \end{bmatrix} \quad R = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ ($\because$ given)

So by removing terminal state

$P = [1-p] \quad R = [1]$

$V(S) = \left[[1] - [1-p]\right]^{-1} [1]$

$= [p]^{-1} [1]$

$$\boxed{\therefore V(S) = 1/p}$$

c) Consider the expectation of every visit MC

$$E[V(s)] = E\left[\frac{K+2}{3}\right]$$

$$= \sum_{k=1}^{\infty} \left(\frac{k+2}{3}\right) Pr(s \text{ repeats } k \text{ times})$$

$$= \sum_{k=1}^{\infty} \left(\frac{k+2}{3}\right)(1-p)^{k-1} \cdot p$$

$$= \frac{p}{3} \underbrace{\sum_{k=1}^{\infty} (k+2)(1-p)^{k-1}}_{\text{sum of } \infty \text{ AGP}}$$

$$= \frac{p}{3}\left[\frac{3}{1-(1-p)} + \frac{1(1-p)}{(1-(1-p))^2}\right]$$

$$= \frac{p}{3}\left[\frac{3}{p} + \frac{1-p}{p^2}\right]$$

$$= \frac{\cancel{p}}{3}\left[\frac{1+2p}{p^{\cancel{2}}}\right]$$

$$\boxed{\therefore E[V(s)] = \frac{1+2p}{3p}} \neq V(s)$$

$\therefore$ The MC estimate is biased.

f) The first visit MC has low bias & high variance but the Every visit MC has high bias & low variance.

Both MC converges to the unique $\hat{V}$ as the no. of trajectories goes to ∞.

$\therefore$ By law of large numbers both converges.

5) a) for the TD(λ)

$$G_t^\lambda = (1-\lambda)\sum_{n=1}^{T-t-1} \lambda^{n-1} G_{t:t+n} + \lambda^{T-t-1} G_t$$

when $\lambda = 1$

$$G_t^\lambda = 0 + \lambda^{T-t-1} G_t$$

$$= (1)^{T-t-1} G_t$$

$$G_t^1 = G_t$$

Therefore it is Monte carlo method as state, action process goes all the way to the end.

b) TD(0) is low variance and high bias

~~TD(n) is~~

TD(n) is high variance & low bias

we consider TD(λ) as a trade off between variance and bias.

c) If all rewards are scaled with a positive constant, the expected reward is scaled with that constant. Therefore the best policy is not affected.

d) If the Behaviour policy is deterministic the chances for the exploration would decrease and might not perform well for a ~~the~~ stachoistic target policy.

∴ It may not be beneficial.

e) The convergence take place under following conditions

1. state and action spaces are finite
2. All state-action pairs are visited infinitely often.
3. Robbins-Monroe Condition

$$\sum_t a_t = \infty, \quad \sum_t a_t^2 < \infty$$

f) The MC method for policy evaluation is the sample mean for the distribution of rewards. As sample mean is a random variable and is an estimator of population mean, the expected value of sample mean is same as population mean. There fore policy evaluation MC method is unbiased estimator.

h) The value iteration update for state s

is
$$V_{t+1}(s) = \max_{a \in A} R(s,a) + \gamma \sum_{s' \in S} P(s' | s,a) V_t(s')$$

As the $V_{t+1}(s)$ depends only on values in V_t, $V_t(s)$ and not on any other entries of V_{t+1}, it is possible to parallelize the calculations.