* "Assignment-2" *

Nelakuditi Rahul Naga

AI20BTECH11029

1) a) We know that the algorithm terminates when the following condition is satisfied:

$$\| V_{K+1} - V_K \|_\infty \leq \epsilon \quad \rightarrow ①$$

→ Now consider the norm given by $\| V_K - V^\pi \|_\infty$. Using triangle inequality we have:

$$\| V_K - V^\pi \|_\infty \leq \| V_K - V_{K+1} \|_\infty + \| V_{K+1} - V^\pi \|_\infty \quad \rightarrow ②$$

→ Let L^π be the Bellman evaluation operator. We know that $L^\pi(V_K) = V_{K+1}$. Also, L^π is a γ-contraction map. Hence we have:

$$\| V_{K+1} - V^\pi \|_\infty = \| L^\pi(V_K) - L^\pi(V^\pi) \|_\infty \quad (\because L^\pi(V^\pi) = V^\pi)$$

$$\leq \gamma . \| V_K - V^\pi \|_\infty \quad \rightarrow ③$$

From ①, ②, ③ we have:

$$\| V_K - V^\pi \|_\infty \leq \epsilon + \gamma . \| V_K - V^\pi \|_\infty$$

$$\| V_K - V^\pi \|_\infty \leq \frac{\epsilon}{1-\gamma} \quad \rightarrow ④$$

From ③, ④ we can conclude that:

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

b) From equation ③ in the previous part of this question we have:

$$\| V_{K+1} - V^\pi \|_\infty \leq \gamma . \| V_K - V^\pi \|_\infty$$

→ Similarly we have, $\| V_K - V^\pi \|_\infty \leq \gamma . \| V_{K-1} - V^\pi \|_\infty \Rightarrow \| V_{K+1} - V^\pi \|_\infty \leq \gamma^2 . \| V_{K-1} - V^\pi \|_\infty$

Hence, applying the above inequality iteratively we obtain the following ineq-

-uality :-

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

c) Consider a state $s \in S$. There exists actions a and a' that satisfy the following relations :-

$$L(u(s)) = R(s,a) + \gamma . \sum_{s'} P(s'|s,a).u(s') \rightarrow (1)$$

$$L(v(s)) = R(s,a') + \gamma . \sum_{s'} P(s'|s,a').v(s') \rightarrow (2)$$

It is easy to infer that :-

$$L(v(s)) \geq R(s,a) + \gamma . \sum_{s'} P(s'|s,a). v(s') \rightarrow (3)$$

The operation (1) - (3) gives :-

$$L(u(s)) - L(v(s)) \leq \gamma . \sum_{s'} P(s'|s,a).(u(s') - v(s')) \rightarrow (4)$$

→ Since $u \leq v$ we have $u(s') - v(s') \leq 0$. Hence from equation (4) we can infer that:-

$$L(u(s)) - L(v(s)) \leq 0 \quad \text{for an arbitrary } s \in S$$

→ As state 's' is chosen arbitrarily, the above equation holds for all $s \in S$ which implies that $L(u) \leq L(v)$. Hence, Bellman Optimality operator is monotonic.

2) a) It is given that P, Q are two contractions defined on a normed vector space $\langle V, \|.\| \rangle$. By definition we have:-

$$\| P(u) - P(v) \| \leq \gamma_P \| u - v \| \quad \forall u,v \in V \rightarrow (1)$$

$$\| Q(u) - Q(v) \| \leq \gamma_Q \| u - v \| \quad \forall u,v \in V \rightarrow (2)$$

Here, γ_P & $\gamma_Q \in [0,1)$. Now consider $P \circ Q$ map :

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

$$\leq \gamma_P . \gamma_Q . \| u - v \| \quad \forall u,v \in V$$

Since $\gamma_P \cdot \gamma_Q \in [0,1)$, we can conclude that $P \circ Q$ is a contraction map on normed vector space $\langle V, \|.\| \rangle$. Now consider $Q \circ P$ map :-

$$\| Q \circ P(u) - Q \circ P(v) \| = \| Q(P(u)) - Q(P(v)) \|$$
$$\leq \gamma_Q \cdot \| P(u) - P(v) \|$$
$$\leq \underbrace{\gamma_Q \cdot \gamma_P}_{\in [0,1)} \| u - v \| \quad \forall u, v \in V$$

Therefore, $Q \circ P$ is also a contraction map on normed vector space $\langle V, \|.\| \rangle$.

b) It is easy to see from part (a) of this question that the contraction coefficient for $P \circ Q$ and $Q \circ P$ is given by $\gamma_P \cdot \gamma_Q$. Since γ_P & $\gamma_Q \in [0,1)$ their product i.e; $\gamma_P \cdot \gamma_Q$ also belongs to $[0,1)$.

c) It is required that the value iteration algorithm converge to a unique solution. for this to happen, the operator $B = F \circ L$ must be a contraction map i.e; both F & L should be contraction maps under ∞-norm. We have :-

$$\| B(u) - B(v) \|_\infty \leq \gamma \cdot \| u - v \|_\infty \quad \exists \gamma \in [0,1) \; ; \forall u, v \in V$$

→ Also, let V_* be the optimal value function. If the map B converges to V_*, then V_* must be a fixed point of B i.e;

$$B(V_*) = F(L(V_*)) = V_*$$

→ Under the above conditions, the value iteration algorithm converges to the unique solution V_* if we use operator B instead of L.

3) a) It is easy to see that state A is an absorbing state. Therefore the general form for a trajectory starting from state 'S' is given by SSSS... (K times) A where K: 1, 2, ...

b) We will consider a trajectory starting from state 'S' as follows to estimate v(S) using first visit MC:

Trajectory: SS... (K times) A

Using the above single trajectory, the estimate for v(S) using first visit MC is given by:-

$$v(S) = \frac{1+1+1+\ldots(K \text{ times})}{1} \quad (\because \text{Reward for being in state } S = 1)$$

$$v(S) = K$$

c) We will consider a trajectory starting from state 'S' as follows to estimate v(S) using every visit MC:

Trajectory: SS.... (K times) A

Using the above single trajectory, the estimate for v(S) using every visit MC is given by:-

$$v(S) = \frac{K + (K-1) + (K-2) + \ldots 1}{K}$$

$$v(S) = \frac{\frac{\cancel{K}\cdot(K+1)}{2}}{\cancel{K}}$$

$$v(S) = \frac{K+1}{2}$$

d) We know that inorder to calculate the true value function, we have to use the Bellman Evaluation equation which is given by:-

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

It is given that $r = 1$. Also, P and R are given by :-

$$P = \begin{array}{c} \\ S \\ A \end{array}\!\!\begin{array}{cc} S & A \\ \left[\begin{array}{c} 1-p \\ 0 \end{array}\right. & \left.\begin{array}{c} p \\ 1 \end{array}\right] \end{array}$$

$$R = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \begin{array}{c} S \\ A \end{array}$$

→ As we desire to calculate $v(S)$, we will exclude the terminal/absorbing state 'A' from P and R. Then we have:-

$$V(S) = (1 - (1-p))^{-1} . (1)$$

$$V(S) = (p)^{-1}$$

$$V(S) = \frac{1}{p}$$

→ We have obtained the above matrix equation to find $v(S)$ since $P = [1-p]$ and $R = [1]$ if we exclude the absorbing state 'A'.

e) Consider $E[v(S)]$ for every visit MC. It is given by :-

$$E[v(S)] = E\left[\frac{K+1}{2}\right]$$

$$= \sum_{K=1}^{\infty} \left(\frac{K+1}{2}\right) . (1-p)^{K-1} . p$$

$$= \frac{p}{2} . \left(\sum_{K=1}^{\infty} (1-p)^{K-1} + \sum_{K=1}^{\infty} K . (1-p)^{K-1} \right)$$

Upon evaluting the summations we have:-

$$\sum_{K=1}^{\infty} (1-p)^{K-1} = \frac{1}{p}$$

$$\sum_{K=1}^{\infty} K . (1-p)^{K-1} = \frac{1}{p^2}$$

$$\therefore \quad E[v(S)] = \frac{p}{2}\left(\frac{1}{p} + \frac{1}{p^2}\right) = \frac{1 + \frac{1}{p}}{2}.$$

→ We have derived in part (d) of this question that $v(s) = \frac{1}{P}$.

$\therefore$ for every visit MC, $E[v(s)] \neq v(s) \Rightarrow$ Every visit MC estimate is biased.

f) The first visit MC method converges to $v(s)$ by the law of large numbers as the number of trajectories/ experiences increase.

→ Every visit MC method also converges to $v(s)$ as the number of trajec-tories increase (by the law of large numbers). But the convergence of every visit MC is not as straight-forward as convergence of first visit MC, since there might be multiple samples from the same trajectory that contribute to the mean calculation in every visit MC unlike first visit MC.

4) a) We have $\delta_t = r_{t+1} + \gamma . v^{\pi}(s_{t+1}) - v^{\pi}(s_t)$. We have to calculate $E_{\pi}(\delta_t \mid S_t = s)$ when δ_t uses the true state value function v^{π}.

$$E_{\pi}(\delta_t \mid S_t = s) = E_{\pi}(r_{t+1} + \gamma . v^{\pi}(s_{t+1}) - v^{\pi}(s_t) \mid S_t = s)$$

$$= E_{\pi}(r_{t+1} + \gamma . v^{\pi}(s_{t+1}) \mid S_t = s) - E_{\pi}(v^{\pi}(s_t) \mid s_t = s)$$

$$= v^{\pi}(s) - v^{\pi}(s)$$

$$= 0.$$

→ While evaluating $E_{\pi}(\delta_t \mid S_t = s)$ above, we made use of the fact that $E_{\pi}(r_{t+1} + \gamma . v^{\pi}(s_{t+1}) \mid s_t = s) = v^{\pi}(s)$ for true state value function v^{π}.

b) We have $\delta_t = r_{t+1} + \gamma . v^{\pi}(s_{t+1}) - v^{\pi}(s_t)$. We have to calculate $E_{\pi}(\delta_t \mid S_t = s, A_t = a)$ when δ_t uses the true state value function v^{π}.

$$E_{\pi}(\delta_t \mid S_t = s, A_t = a) = E_{\pi}(r_{t+1} + \gamma . v^{\pi}(s_{t+1}) - v^{\pi}(s_t) \mid s_t = s, A_t = a)$$

$$= E_{\pi}(r_{t+1} + \gamma . v^{\pi}(s_{t+1}) \mid S_t = s, A_t = a) - E_{\pi}(v^{\pi}(s_t) \mid S_t = s, A_t = a)$$

Using the fact that $E_\pi(\delta_{t+1} + \gamma . v^\pi(S_{t+1}) \mid S_t = s, A_t = a) = Q^\pi(s,a)$ we have :-

$$E_\pi(\delta_t \mid S_t = s, A_t = a) = Q^\pi(s,a) - v^\pi(s)$$

c) The weight corresponding to n-step return i.e; $G_t^{(n)}$ in the expression for G_t^λ is given by $(1-\lambda).\lambda^{n-1}$. Clearly, w_1 = Weight corresponding to $G_t^{(1)} = 1-\lambda$. It is given that after $\eta(\lambda)$ time-steps, the weight would have fallen to half of the initial value i.e; w_1.

$$\therefore \quad w_{\eta(\lambda)} = \cancel{(1-\lambda)}.\lambda^{\eta(\lambda)-1} = \frac{w_1}{2} = \frac{\cancel{1-\lambda}}{2}$$

$$\Rightarrow (\eta(\lambda)-1). \ln(\lambda) = \ln(\tfrac{1}{2}) \quad \text{(taking logarithm on both sides)}$$

$$\Rightarrow \eta(\lambda) = \frac{\ln(\frac{1}{2})}{\ln(\lambda)} + 1$$

→ Now it is given that $\eta(\lambda) = 3$. Therefore we have :-

$$3 = \frac{\ln(\frac{1}{2})}{\ln(\lambda)} + 1$$

$$\Rightarrow \ln(\lambda) = \frac{\ln(\frac{1}{2})}{2} = \ln\left(\frac{1}{\sqrt{2}}\right)$$

$$\therefore \lambda = \frac{1}{\sqrt{2}}$$

5) In order to check the divergence and convergence of $\sum_{t=1}^{\infty} \alpha_t$ and $\sum_{t=1}^{\infty} \alpha_t^2$ respectively, we will use the "Integral-Test". It states that :-

→ If $f(x)$ is a decreasing positive function defined on $[1,\infty)$, then the series $\sum_{n=1}^{\infty} f(n)$ converges if and only if the integral $\int_1^\infty f(x)dx$ converges.

→ It is easy to see that all α_t given in the problem are decreasing positive functions defined on $[1,\infty)$. This holds true even for α_t^2.

i) $\alpha_t = \frac{1}{t}$

→ To check the convergence / divergence of $\sum_{t=1}^{\infty} \frac{1}{t}$, consider $\int_1^{\infty} \frac{1}{t} dt$

$= [\ln(t)]_1^{\infty} = \infty \Rightarrow \sum_{t=1}^{\infty} \frac{1}{t} = \infty$.

→ $\sum_{t=1}^{\infty} \alpha_t^2 = \sum_{t=1}^{\infty} \frac{1}{t^2}$. Consider $\int_1^{\infty} \frac{1}{t^2} dt = \left[\frac{-1}{t}\right]_1^{\infty} = 1 \Rightarrow \sum_{t=1}^{\infty} \frac{1}{t^2} < \infty$.

$\therefore \alpha_t = \frac{1}{t}$ satisfies Robbins-Monroe condition ⇒ TD

Algorithm converges to true $v(s)$.

ii) $\alpha_t = \frac{1}{t^2}$

→ $\sum_{t=1}^{\infty} \alpha_t = \sum_{t=1}^{\infty} \frac{1}{t^2}$. Consider $\int_1^{\infty} \frac{1}{t^2} dt = 1 \Rightarrow \sum_{t=1}^{\infty} \frac{1}{t^2} < \infty$.

→ $\sum_{t=1}^{\infty} \alpha_t^2 = \sum_{t=1}^{\infty} \frac{1}{t^4}$. Consider $\int_1^{\infty} \frac{1}{t^4} dt = \frac{1}{3} \Rightarrow \sum_{t=1}^{\infty} \frac{1}{t^4} < \infty$.

$\therefore \alpha_t = \frac{1}{t^2}$ does not satisfy Robbins-Monroe condition ⇒ TD

Algorithm doesn't converge to true $v(s)$.

iii) $\alpha_t = \frac{1}{t^{2/3}}$

→ $\sum_{t=1}^{\infty} \alpha_t = \sum_{t=1}^{\infty} \frac{1}{t^{2/3}}$. Consider $\int_1^{\infty} \frac{1}{t^{2/3}} dt = \infty \Rightarrow \sum_{t=1}^{\infty} \frac{1}{t^{2/3}} = \infty$.

→ $\sum_{t=1}^{\infty} \alpha_t^2 = \sum_{t=1}^{\infty} \frac{1}{t^{4/3}}$. Consider $\int_1^{\infty} \frac{1}{t^{4/3}} dt = 3 \Rightarrow \sum_{t=1}^{\infty} \frac{1}{t^{4/3}} < \infty$.

$\therefore \alpha_t = \frac{1}{t^{2/3}}$ satisfies Robbins-Monroe condition ⇒ TD

Algorithm converges to true $v(s)$.

iv.) $\alpha_t = \frac{1}{t^{1/2}}$

→ $\sum_{t=1}^{\infty} \alpha_t = \sum_{t=1}^{\infty} \frac{1}{t^{1/2}}$. Consider $\int_1^{\infty} \frac{1}{t^{1/2}} dt = \infty \Rightarrow \sum_{t=1}^{\infty} \frac{1}{t^{1/2}} = \infty$.

$\rightarrow \quad \sum_{t=1}^{\infty} \alpha_t^2 = \sum_{t=1}^{\infty} \frac{1}{t}$. consider $\int_1^{\infty} \frac{1}{t} dt = \infty \Rightarrow \sum_{t=1}^{\infty} \frac{1}{t} = \infty$

$\therefore \alpha_t = \frac{1}{t^{1/2}}$ does not satisfy Robbins-Monroe condition which implies that TD algorithm doesn't converge to true $v(s)$.

* Now we have $\alpha_t = \frac{1}{t^p}$ where $p > 0$. for the Robbins-Monroe condition to be satisfied we should have the following:-

$$\rightarrow \sum_{t=1}^{\infty} \alpha_t = \sum_{t=1}^{\infty} \frac{1}{t^p} = \infty$$

$$\rightarrow \sum_{t=1}^{\infty} \alpha_t^2 = \sum_{t=1}^{\infty} \frac{1}{t^{2p}} < \infty$$

$\rightarrow$ It is easy to see that if $p \leq 1$ we have:-

$$\sum_{t=1}^{\infty} \frac{1}{t^p} \geq \sum_{t=1}^{\infty} \frac{1}{t} = \infty \Rightarrow \sum_{t=1}^{\infty} \frac{1}{t^p} = \infty$$

$\therefore \sum_{t=1}^{\infty} \alpha_t = \infty$ if $p \leq 1 \rightarrow$ (1)

$\rightarrow$ Now for $\sum_{t=1}^{\infty} \frac{1}{t^{2p}}$ we need the integral $\int_1^{\infty} \frac{1}{t^{2p}} dt = \left[\frac{-1}{(2p-1).t^{2p-1}} \right]_1^{\infty}$ to converge. It is easy to see that the integral converges if and only if $2p-1 > 0 \Rightarrow p > \frac{1}{2}$.

$\therefore \sum_{t=1}^{\infty} \alpha_t^2 < \infty$ if $p > \frac{1}{2} \rightarrow$ (2)

$\therefore$ From (1) & (2) we can conclude that TD algorithm converges to true $v(s)$ only if $\frac{1}{2} < p \leq 1$.