Reinforcement - Learning

Assignment - 2

Question - 1)

a) we terminate the algorithm when

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

Lets say we didn't stop it at this condition, instead we let the value function converge.

The iteration scheme is

$$v^{(k+1)}(s) = \sum_a \pi(a|s) \sum_{s'} P_{ss'}^a \left(R_{ss'}^a + \gamma v^{(k)}(s)\right) \quad -(1)$$

$$v^{(k+2)}(s) = \sum_a \pi(a|s) \sum_{s'} P_{ss'}^a \left(R_{ss'}^a + \gamma v^{(k+1)}(s)\right) \quad -(2)$$

(2) - (1)

$$v^{(k+2)}(s) - v^{(k+1)}(s) = \sum_a \pi(a|s) \sum_{s'} P_{ss'}^a \gamma \left(v^{(k+1)}(s) - v^{(k)}(s)\right) \quad -(3)$$

we know that $\| v^{(k+1)} - v^{(k)} \|_\infty \leq \varepsilon$

$$\Rightarrow \left| v^{(k+1)}(s) - v^{(k)}(s) \right| \leq \varepsilon \quad \forall s \in S$$

$$\Rightarrow -\varepsilon \leq v^{(k+1)}(s) - v^{(k)}(s) \leq \varepsilon \quad -(4)$$

Using (3) and (4)

$$\left| v^{(k+2)}(s) - v^{(k+1)}(s) \right| \leq \sum_a \pi(a|s) \gamma \varepsilon \quad \forall s$$

$$\boxed{-\gamma \varepsilon \leq v^{(k+2)}(s) - v^{(k+1)}(s) \leq \gamma \varepsilon}$$

similarly we can say

$$-\gamma^2\varepsilon \leq v^{(k+3)} - v^{(k+2)}(s) \leq \gamma^2\epsilon$$

Now

$$-\gamma\varepsilon \leq v^{(k+2)}(s) - v^{(k+1)}(s) \leq \gamma\varepsilon$$

$$-\gamma^2\varepsilon \leq v^{(k+3)}(s) - v^{(k+2)}(s) \leq \gamma^2\varepsilon$$

$$\vdots$$

Adding these equations

$$-(\gamma\varepsilon + \gamma^2\varepsilon + \ldots) \leq v^{\pi}(s) - v^{(k+1)}(s) \leq \gamma\varepsilon + \gamma^2\varepsilon + \ldots$$

$$|v^{\pi}(s) - v^{(k+1)}(s)| \leq \gamma\varepsilon + \gamma^2\varepsilon + \ldots \leq \frac{\gamma\varepsilon}{1-\gamma}$$

$$\Rightarrow \boxed{\|v^{\pi} - v^{(k+1)}\|_{\infty} \leq \frac{\gamma\varepsilon}{1-\gamma}}$$

Hence Proved

b) We ~~kn~~ know that L (Bellman optimality operator) is a contraction map.

Since V^π is the optimal value,

$$L(V^\pi) = V^\pi \Leftrightarrow V^\pi \text{ is fixed for}$$

from the properties of contraction map we know that

$$\| L(V^{(k)}) - L(V^\pi)\|_\infty \leq \gamma \| V^{(k)} - V^\pi \|_\infty \quad (1)$$

from our algorithm

$$L(V^{(k)}) = V^{(k+1)}$$

Substituting in (1) & using $L(V^\pi) = V^\pi$

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

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

$$\vdots$$

$$\| V^{(2)} - V^\pi \|_\infty \leq \gamma \| V^{(1)} - V^\pi \|_\infty$$

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

c)

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

P^a is probability ~~vector~~ matrix which has all positive entries.

~~we know that $P^a x \geq 0$ if $x \geq 0$~~

Claim : $P^a x \geq 0$ if $x \geq 0$

Proof all the entries of P are positive, and also all the entries of x are positive, we ~~me~~ take inner product in rowwise manner all the terms would again be non negative.

$$\Rightarrow \quad P^a(v-u) \geq 0$$

$$\Rightarrow \quad P^a v \geq P^a u$$

adding R^a on both sides

$$P^a v + R^a \geq P^a u + R^a$$

This holds for all a

$$\therefore \quad \max_a (P^a v + R^a) \geq \max_a (P^a u + R^a)$$

$$\boxed{\Rightarrow \quad L(v) \geq L(u)}$$

Hence Proved

Problem-2

Part-a

For P

$$\|P(v) - P(u)\| \leq \gamma_1 \|v - u\|$$

and for Q

$$\|Q(v) - Q(u)\| \leq \gamma_2 \|v - u\|$$

In the first equation substitute v by $Q(v)$ and u by $Q(u)$. We get,

$$\|P(Q(v)) - P(Q(u))\| \leq \gamma_1 \|Q(v) - Q(u)\|$$

From the second equation

$$\|P(Q(v)) - P(Q(u))\| \leq \gamma_1 \gamma_2 \|v - u\|$$

$\therefore$ $P \circ Q$ is a contraction mapping

Now, there's nothing unique about P or Q, we could just swap them and we would still get the same result. Or we would follow the same procedure and get

$$\|Q(P(v)) - Q(P(u))\| \leq \gamma_1 \gamma_2 \|v - u\|$$

Part-b

From the solution to above subpart, we can see that the contraction coefficient for both the composite functions is $\gamma_1 \gamma_2$ where γ_1, γ_2 are the contraction coefficients for P and Q respectively.

Part-c

The value iteration scheme under the operator would converge to a unqiue solution if it follows

$$\|B(v) - B(u)\| \leq \|v - u\|$$

Writing B as a composite function, we can arrive at a general solution with a specific behaviour

$$\|F(L(v)) - F(L(u))\| \leq \|v - u\|$$

Assume $\|F(v) - F(u)\| \leq \eta \|v - u\|$ Now

$$\|F(L(v)) - F(L(u))\| \leq \eta \|L(v) - L(u)\| \leq \eta\gamma \|v - u\|$$

We get that $\eta < \frac{1}{\gamma}$

Problem - 3)

a) A generic form of the trajectory is

$$\underbrace{S \dots}\, A$$

S can occur ≥ 1 times

b) First visit MC makes sense only when we are given some experience

However, we can find the value vector by calculating expectation of G_t

For a trajectory ~~with m~~ of size m, the state S would have been visited $m-1$ times

$G_t^{(m)} = m - 1$ since state S is visited $m-1$ times

Note that the trajectory starts with S

$$E[G_t] = \sum_{m=2}^{\infty} p(1-p)^{m-2}(m-1)$$

$$= p + 2p(1-p) + 3p(1-p)^2 + \dots$$

$$V(S) = p + 2p(1-p) + 3p(1-p)^2 + \dots$$

$$(1-p)V(S) = p(1-p) + 2p(1-p)^2 + \dots$$

$$pV(S) = p + p(1-p) + p(1-p)^2 + \dots$$

$$pV(s) = \frac{p}{p}$$

$$V(s) = 1$$

for the value at A. All the trajectories end with A and that's the first visit to A in that trajectory

Since, the reward for being in state A is 0

$$V(A) = 0$$

$$\therefore V = \begin{pmatrix} \frac{1}{p} \\ 0 \end{pmatrix}$$

c) Again consider a trajectory

$$G_t^{(m)} = \frac{1+2+\dots+(m-1)}{m-1} = \frac{\cancel{(m-1)}\, m}{2\cancel{(m-1)}}$$

$$= \frac{m}{2}$$

Again we are considering the trajectories with state S for every visit MC since $V(A)=0$ using the same argument is the previous subpart.

$$E[G_t] = V(S) = \sum_{m=2}^{\infty} p(1-p)^{m-2}\,\frac{m}{2}$$

$$v(s) = p + p(1-p)\frac{3}{2} + p(1-p)^2 2 + \dots$$

$$v(s)(1-p) = p(1-p) + (1-p)^2\frac{3}{2} + \dots$$

$$v(s)p = p + \frac{1}{2}(p(1-p) + p(1-p)^2 + \dots)$$

$$= p + \frac{1}{2}\frac{\cancel{p}(1-p)}{\cancel{p}}$$

$$= p + \frac{1}{2} - \frac{p}{2}$$

$$v(s) = \frac{p+1}{2p} = \frac{1}{2} + \frac{1}{2p}$$

d) Using the exact method

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

$$= \left(\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} 1-p & p \\ 0 & 0 \end{bmatrix}\right)^{-1} \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$= \begin{bmatrix} p & -p \\ 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$= \frac{1}{p}\begin{bmatrix} 1 & p \\ 0 & p \end{bmatrix}\begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{p} \\ 0 \end{bmatrix}$$

e) Yes, every visit MC estimate is biased.

The true value function evaluated at state S is $\frac{1}{p}$

while every visit MC gives us $\frac{1}{2} + \frac{1}{2p}$

we know that $p < 1$

$$\frac{1}{2p} > \frac{1}{2}$$

$$\frac{1}{p} - \frac{1}{2p} > \frac{1}{2}$$

$$\Rightarrow \quad \frac{1}{p} > \frac{1}{2} + \frac{1}{2p}$$

$\therefore$ there is bias

f) Every visit MC will converge faster because the number of samples would be far more in case of every visit MC compared to first visit MC.

Problem - 4

a) $$E_\pi(\delta_t \mid S_t = s) = E_\pi[r_{t+1} + \gamma V^\pi(S_{t+1}) - V^\pi(S_t) \mid S_t = s]$$

Expectation is linear

$$E_\pi[r_{t+1} + \gamma V^\pi(S_{t+1}) \mid S_t = s] - E[V^\pi(S_t) \mid S_t = s]$$

$$\Rightarrow \quad V^\pi(s) - V^\pi(s)$$

$$\boxed{= 0}$$

b) $$E_\pi[\delta_t \mid S_t = s, A_t = a] = E_\pi[r_{t+1} + \gamma V^\pi(S_{t+1}) - V^\pi(S_t) \mid S_t = s, A_t = a]$$

$$= \underbrace{E_\pi[r_{t+1} + \gamma V^\pi(S_{t+1}) \mid S_t = s, a_t = a]}_{Q^\pi(s,a)} - \underbrace{E[V^\pi(S_t) \mid S_t = s, a_t = a]}_{V^\pi(s)}$$

$$= Q^\pi(s,a) - V^\pi(s)$$

c) $$G_t^\lambda = (1-\lambda) \sum_{n=1}^{\infty} \lambda^{n-1} G_t^{(n)}$$

The first coefficient is

$(1-\lambda)$

~~The coeffici~~ The subsequent coefficients are

$(1-\lambda)\lambda, (1-\lambda)\lambda^2, \ldots$

by definition

$$\cancel{(1-\lambda)}\, \lambda^{\eta(\lambda)} \le \frac{\cancel{(1-\lambda)}}{2}$$

$$\eta(\lambda) \log \lambda \le -\log 2$$

$$\eta(\lambda) \le \frac{-\log 2}{\log \lambda}$$

we can write

$$\eta(\lambda) \leq \log_{1/\lambda} 2$$

given $\eta(\lambda) = 3$

$$3 \leq \log_{1/\lambda} 2$$

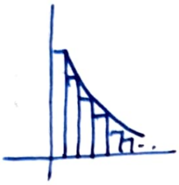
$$\lambda^{-3} \geq 2$$

$$\boxed{\lambda \geq 2^{-1/3}}$$

Problem 5

$\alpha_t = \frac{1}{t^p}$

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



$$\int_1^{\infty} \frac{1}{t^p} dt \leq \sum_{t=1}^{\infty} \frac{1}{t^p} \leq 1 + \int_1^{\infty} \frac{1}{t^p} dt$$

when $p > 1$

$$\frac{1}{p-1} \leq \sum_{t=1}^{\infty} \frac{1}{t^p} \leq \frac{p}{p-1}$$

when $p < 1$, the summation diverges

when $p = 1$

$$\log t \Big|_1^{\infty} \leq \sum_{t=1}^{\infty} \frac{1}{t} \leq 1 + \log t \Big|_1^{\infty}$$

the summation diverges

Summary

$$\sum \frac{1}{t^p} \begin{cases} \text{converges} & p > 1 \\ \text{diverges} & p \leq 1 \end{cases}$$

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

$\sum \alpha_t = \infty$ $\}$ $p = 1$ diverges

$\sum \alpha_t^2 < \infty$ $\}$ $p = 2$ converges

$\therefore$ it would converge

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

$\sum \alpha_t = \sum \frac{1}{t^2} < \infty$ $\}$ $p = 2$ converges

$\sum \alpha_t^2 = \sum \frac{1}{t^4} < \infty$ $\}$ $p = 4$ converges

$\therefore$ it wouldn't converge

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

$\sum \alpha_t = \sum \frac{1}{t^{2/3}} = \infty$ $\}$ $p = 2/3$ diverges

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

$\therefore$ it would converge

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

$\sum \alpha_t = \sum \frac{1}{t^{1/2}} = \infty$ $\}$ $p = \frac{1}{2}$ diverges

$\sum \alpha_t^2 = \sum \frac{1}{t} = \infty$ $\}$ $p = 1$ diverges

$\therefore$ it won't converge