Assignment-3

AI21BTECH11021

1)

(c)

s	a	r	s	a	r	s	a	r	s	a	r	s
C	jump	4	E	right	1	F	left	-2	C	right	1	F

→ learning rate $\alpha = 0.7$

Assumming all Q-values initially set to -10

Transition 1 $(s, a, r, s') = (C, \text{jump}, 4, E)$

$$Q[s,a] = Q(s,a) + \alpha \left[R(s,a,s') + \gamma \max_{a'} Q(s',a') - Q(s,a) \right]$$

$$Q(c, \text{jump}) = \underbrace{Q(c, \text{jump})}_{-10} + 0.7\left(4 + 1.0\underbrace{(-10)}_{\text{initiall all} -10} - (-10)\right)$$

$$= -10 + 0.7(4) = \boxed{-7.2}$$ {remaining unchanged}

Transition 2 $(s, a, r, s') = (E, \text{right}, 1, F)$

$$Q[E, \text{right}] = \underbrace{Q[E, \text{right}]}_{-10} + 0.7\left(1 + (1.0)(-10) - (-10)\right)$$

$$= -10 + 0.7(1) = \boxed{-9.3}$$

Transition 3 $(s, a, r, s') = (F, \text{left}, -2, E)$

$$Q[F, \text{left}] = Q[F, \text{left}] + 0.7\left(-2 + (1.0)\overbrace{(-9.3)}^{\max(-10, -9.3)} - (-10)\right)$$

$$= \boxed{-10.91}$$

Transition 4 $(s, a, r, s') = (E, right, 1, F)$

$$Q[E, right] = \underline{Q[E, right]} + 0.7\left(1 + (1.0)\overbrace{(-10)}^{max(-10,-10.91)} - (-9.3)\right)$$

$$= -9.3 + 0.7\ \cancel{(+0.3)}\ +0.99$$

$$= -9.09$$

	Q(C, left)	Q(C, jump)	Q(E, left)	Q[E, right]	Q(F, left)	Q(F, right)
Initial	−10	−10	−10	−10	−10	−10
Transition 1	−10	−7.2	−10	−10	−10	−10
Transition 2	−10	−7.2	−10	−9.3	−10	−10
Transition 3	−10	−7.2	−10	−9.3	−10.91	−10
Transition 4	−10	−7.2	−10	−9.09	−10.91	−10

d) Constructing greedy policy using above table

$$\pi(C) = \underset{a}{\mathrm{argmax}}\ Q(C, a) = \text{jump} \quad \begin{bmatrix} \text{jump} \rightarrow -7.2 \\ \text{left} \rightarrow -10 \end{bmatrix}$$

$$\pi(E) = \underset{a}{\mathrm{argmax}}\ Q(E, a) = \text{right} \quad \begin{cases} \text{left} \rightarrow -10 \\ \text{right} \rightarrow -9.09 \end{cases}$$

$$\pi(F) = \underset{a}{\mathrm{argmax}}\ Q(F, a) = \text{right} \quad \begin{cases} \text{right} \rightarrow -10 \\ \text{left} \rightarrow -10.91 \end{cases}$$

e) Robinns - Monroe condition:

In order for the Q-learning Algorithm to converge we need

$$\sum \alpha_t = \infty$$
$$\sum \alpha_t^2 < \infty$$

i) $\boxed{\alpha_t = \frac{1}{t}}$ $\quad \sum \alpha_t = \sum_{t=1}^{\infty} \frac{1}{t} = \left(1 + \frac{1}{2} + \frac{1}{3} + \cdots \infty\right)$
$= \infty$

(Proof by contradiction)

Assume $1 + \frac{1}{2} + \frac{1}{3} + \cdots = H$ (fixed bounded)

Then $1 + \frac{1}{2} + \underbrace{\frac{1}{3} + \frac{1}{4}} + \underbrace{\frac{1}{5} + \frac{1}{6}} + \underbrace{\frac{1}{7} + \frac{1}{8}} + \cdots = H$

$1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{4} + \frac{1}{6} + \frac{1}{6} + \frac{1}{8} + \frac{1}{8} + \cdots \leq H$

$1 + \frac{1}{2} + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \cdots \leq H \Rightarrow H + \frac{1}{2} \leq H$

But $H = H$ so it is false

Only make sence if $H \to \infty$

So $\therefore H \to \infty$

So Harmonic series is divergent $\boxed{\sum \alpha_t = \infty}$

ii) $\sum \alpha_t^2 = \sum_{t=1}^{\infty} \frac{1}{t^2}$. We know $0 \leq \frac{1}{t^2} \leq \frac{1}{t^2 - t}$ ($\forall t \geq 2$)

$$\sum_{t=2}^{\infty} \frac{1}{t^2 - t} = \sum_{t=2}^{\infty} \left(\frac{1}{t-1} - \frac{1}{t}\right) = \left(1 - \cancel{\frac{1}{2}}\right) + \left(\cancel{\frac{1}{2}} - \cancel{\frac{1}{3}}\right) + \left(\cancel{\frac{1}{3}} - \frac{1}{4}\right) + \left(\cancel{\frac{1}{t-1}} - \frac{1}{t}\right)$$

$$\lim_{t \to \infty} 1 - \frac{1}{t} = 1$$

$$\sum_{t=1}^{\infty} \frac{1}{t^2} = 1 + \sum_{t=2}^{\infty} \frac{1}{t^2} < 1 + \sum_{t=2}^{\infty} \frac{1}{t^2 - t} = 2$$

$$\Rightarrow \boxed{\sum_{t=1}^{\infty} \frac{1}{t^2} < 2} \quad \text{Hence} \quad \boxed{\sum \alpha_t^2 < \infty}$$

Hence this satisfy Robbins Monroe conditions.

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

But we want $\boxed{\sum \alpha_t = \infty}$.

So this does not obeys Robbins Monroe conditions

Also $\sum \alpha_t^2 = \sum \frac{1}{t^4} \le \sum \frac{1}{t^2} < 2 < \infty$ but 1st condition failed.

f.) We know that

$$\max_a Q^*(s,a) = E\left[R(s,a,s') + \gamma \max_a Q^*(s',a)\right]$$

Now In Q learning → It is an off policy Algorithm and we take one step at a time TD(0) so it is written in the form.

$$Q(s_t,a_t) = Q(s_t,a_t) + \alpha_t\left[R(s_t,a_t,s_{t+1}) + \gamma \max_{a'} Q(s_{t+1},a') - Q(s_t,a_t)\right]$$

This converges if it follows the given

a) State & action spaces are finite

b) All state-action pairs are visited infinetly often

c) Robbin - Monroe Conditions Must be satisfied

→ Even though the agent follows fixed policy π → with 0.5 prob & 0.5 → random fashion As we explore the states with high reward the Q-value functions get updated according (ie according to our π if we visit a state more no of times doesn't gurantee that Q^* will change → It is unknown ~~but~~ fixed irrespective of our exploration policy so visiting All states - infinetly often will eventually gets us to Q^* → Hence the Algorithm will converge to optimal Q-function. As we select $\max_a Q(s,a)$ At the End.

ii) We know SARSA is an on-policy algorithm So if the Agent follows a fixed policy π with probability 0.5 and with 0.5 → chooses randomly

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha_t \left[r_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t) \right]$$

Here the states which are visited often by our fixed policy → Has some problems. The conditions for convergence in SARSA

1) Robbins - Monro Condition
2) Every state-action pair visited often
3) The policy is greedy with respect to policy derived from Q in the limit
4) Controlled Markov chain is communicating: every state follows markovian Assumption
5) $Var(R(s,a)) < \infty$ R → reward function

But if we observe the 3rd condition our policy is fixed so if π^* is not same as our fixed policy π then 3rd wont hold

This can be seen of some states are visited more often then if even though it can have small +ve reward but over time it gets accumulated more & more unlike Q-learning there is no $\max_a Q(s,a)$ to stop this

so this is like an off-policy situation

$\therefore$ SARSA may not converge to optimal Q-function $\left\{ \text{It may converge if given fixed policy itself is the optimal policy.} \right\}$