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	E	right	1	F

→ learning rate $\alpha = 0.7$

Assumming all Q-values initially set to -10

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

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

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

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

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

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

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

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

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

$$= \boxed{-10.91}$$

Transition4 $(s, a, r, s') = (E, \text{right}, 1, F)$

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

$$= -9.3 + 0.7\,\cancel{(+0.3)} + 0.21$$

$$= -9.09$$

	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
Transition1	−10	−7.2	−10	−10	−10	−10
Transtion2	−10	−7.2	−10	−9.3	−10	−10
Transtion3	−10	−7.2	−10	−9.3	−10.91	−10
Transition4	−10	−7.2	−10	−9.09	−10.91	−10

d) Constructing greedy policy using above table

$\pi(C) = \underset{a}{\text{argmax}}\, Q(C,a) = \text{jump}$ $\left[\begin{array}{l} \text{jump} \to -7.2 \\ \text{left} \to -10 \end{array}\right]$

$\pi(E) = \underset{a}{\text{argmax}}\, Q(E,a) = \text{right}$ $\left\{\begin{array}{l} \text{left} \to -10 \\ \text{right} \to -9.09 \end{array}\right.$

$\pi(F) = \underset{a}{\text{argmax}}\, Q(F,a) = \text{right}$ $\left[\begin{array}{l} \text{right} \to -10 \\ \text{left} \to -10.91 \end{array}\right.$

---

e) In order for Q-learning Alg to converge to optimal Q-function $\alpha_t \to$ learning rate must satsify Robinns-Monroe condition

i) $\alpha_t = 1/t$

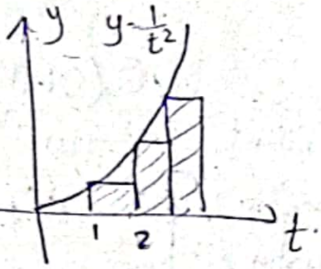
ii) $\alpha_t = 1/t^2$

Sol Robinns-Monroe condition if $\alpha_t \rightarrow$ learning rate

$$\sum \alpha_t = \infty$$
$$\sum \alpha_t^2 < \infty \text{ (ie bounded)}$$

i) $\boxed{\alpha_t = 1/t}$ $\quad \sum \alpha_t = \sum_{t=0}^{\infty} \frac{1}{t} \rightarrow \infty$ {Harmonic series diverge}

$$\sum \alpha_t^2 = \sum_{t=1}^{\infty} \frac{1}{t^2} \leq \int_1^{\infty} f(t)\, dt$$



We know by Integral test as $m \rightarrow \infty$ Both are equal

$$\int_1^{\infty} f(t)\, dt = \lim_{m \to \infty} \int_1^{m} f(t)\, dt \geq \sum_{t=1}^{\infty} \frac{1}{t^2}$$

$$\Rightarrow \lim_{k \to \infty} \int_1^{m} \frac{1}{t^2}\, dt = \left[\frac{-1}{t}\right]_1^m = \lim_{m \to \infty} \left[1 - \frac{1}{m}\right] \rightarrow 1 \text{ (converges to 1)}$$

So $\sum \frac{1}{t^2}$ also converges.

---

ii) $\boxed{\alpha_t = \frac{1}{t^2}}$ $\quad \sum \alpha_t \longleftrightarrow 1$ (Proved above)

But we want this to diverge

So this Robinns-Monroe condition failed.

---

f)

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

$\rightarrow$ This Equation is derived from bellman Equation & we replaced.

value function with Q - function.

$V(s_{t+1}) = \max_{a'} Q(s_{t+1}, a')$

→ Given it is sampled infinetly often in such cases

$E[r_{t+1} + \gamma V(s_{t+1})] \xrightarrow[\text{to}]{\text{converges}} V(s_t)$

As sampling tends to ∞

→ Now Above is just the incremental form of this Equation so next we have ε-greedy policy with $\varepsilon = 0.5$

→ Now eventhough we choose 0.5 probability some random action for exploration this is for finding more optimal paths eventually after exploring all (∞ sampling often) We can be sure no more to explore so our optimal policy is achieved (ie even though we explore randomly the update equation's actions is unchanged ↳ action to be taken after certain point of time.)

→ Furthermore from incremental form we need α_t to obey Robinns-Monroe condition. For mathematically convergence to work.

→ NOTE: As Q-learning is off policy π can be any policy need not be optimal.

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

$E[r_{t+1} + \gamma Q(s_{t+1}, a_{t+1})] \longrightarrow V(s_{t+1})$

This holds if we sample infinetly often But we need to do it in on-policy fashion. Then this will converge

So Simple argument to previous we can say ~~Even~~ though we use ϵ-greedy over the time this will converge to optimal policy.

→ Here also we need α_t to have theoretical gurantees in incremental form to Converge. Also $\gamma < 1$ which limits/bounds the value inside expectation.

Scanned with CamScanner