AI 3000 / CS 5500 : Reinforcement Learning

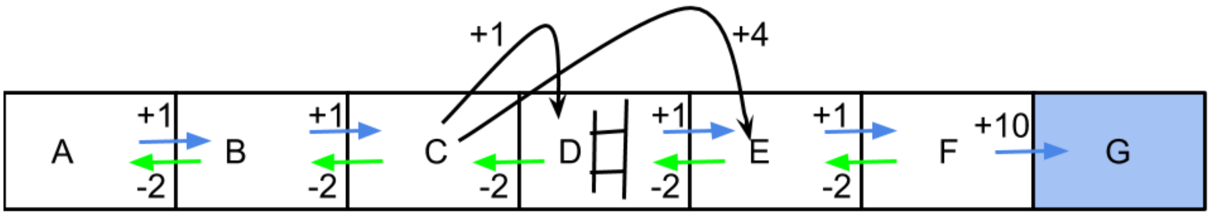
Assignment № 3

Due Date : 06/10/2023

Couse Instructor : Easwar Subramanian 21/09/2023

Problem 1 : Model Free Methods

Consider the MDP shown below with states $\{A, B, C, D, E, F, G\}$. Normally, an agent can either move *left* or *right* in each state. However, in state C, the agent has the choice to either move *left* or *jump* forward as the state D of the MDP has an hurdle. There is no *right* action from state C. The *jump* action from state C will place the agent either in square D or in square E with probability 0.5 each. The rewards for each action at each state s is depicted in the figure below alongside the arrow. The terminal state is G and has a reward of zero. Assume a discount factor of $\gamma = 1$.



Consider the following samples of Markov chain trajectories with rewards to answer the questions below

- $A \xrightarrow{+1} B \xrightarrow{+1} C \xrightarrow{-2} B \xrightarrow{+1} C \xrightarrow{+1} D \xrightarrow{+1} E \xrightarrow{+1} F \xrightarrow{+10} G$
- $A \xrightarrow{+1} B \xrightarrow{+1} C \xrightarrow{+1} D \xrightarrow{+1} E \xrightarrow{+1} F \xrightarrow{+10} G$
- $A \xrightarrow{+1} B \xrightarrow{+1} C \xrightarrow{+4} E \xrightarrow{+1} F \xrightarrow{+10} G$
- $A \xrightarrow{+1} B \xrightarrow{+1} C \xrightarrow{+4} E \xrightarrow{-2} D \xrightarrow{+1} E \xrightarrow{+1} F \xrightarrow{+10} G$
- $A \xrightarrow{+1} B \xrightarrow{+1} C \xrightarrow{+4} E \xrightarrow{-2} D \xrightarrow{+1} E \xrightarrow{+1} F \xrightarrow{-2} E \xrightarrow{+1} F \xrightarrow{+10} G$

(a) Evaluate $V(s)$ using first visit Monte-Carlo method for all states s of the MDP. (2 Points)

(i) $V(A) = (14 + 15 + 17 + 16 + 15)/5 = 77/5 = 15.4$

(ii) $V(B) = (13 + 14 + 16 + 15 + 14)/5 = 72/5 = 14.4$

(iii) $V(C) = (12 + 13 + 15 + 14 + 13)/5 = 67/5 = 13.4$

(iv) $V(D) = (12 + 12 + 12 + 11)/4 = 47/4 = 11.75$

(v) $V(E) = (11 + 11 + 11 + 10 + 9)/5 = 47/4 = 10.4$

(vi) $V(F) = (10 + 10 + 10 + 10 + 9)/5 = 9.8$ and $V(G) = 0$

(b) Which states are likely to have different value estimates if evaluated using every visit MC as compared to first visit MC ? Why ? (2 Points)

States $\{B, C, E, F\}$ are likely to have different value estimates when evaluated using every visit MC as these are visited more than once in a single rollout.

(c) Fill in the blank cells of the table below with the Q-values that result from applying the Q-learning update for the 4 transitions specified by the episode below. You may leave Q-values that are unaffected by the current update blank. Use learning rate $\alpha = 0.7$. Assume all Q-values are initialized to -10. (2 Points)

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

	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						
Transition 2						
Transition 3						
Transition 4						

Q-Evaluations are provided in the table. A state-action is only updated when a transition is made from it. Q(C; left), Q(E; left), and Q(F; right) state-actions are never experienced and so these values are never updated. The Q-learning update rule is given by,

$$Q(s, a) \leftarrow Q(s, a) + \alpha[r + \gamma \max_{a'} Q(s', a') - Q(s, a)]$$

Using the above update rule, the four updates are given by,

$$\begin{aligned} -7.2 &= -10 + 0.7(4 + 0) \\ -9.3 &= -10 + 0.7(1 + 0) \\ -10.91 &= -10 + 0.7(-2 + 0.7) \\ -9.09 &= -9.3 + 0.7(1 - 0.7) \end{aligned}$$

On transition 2, the Qs for F are still both 0, so the update increases the value by the reward +1 times the learning rate. On transition 3, the reward of -2 and $Q(E; right) = 0:5$ are included in the update. On transition 4, $Q(F; left)$ is now -0:75 but $Q(F; right)$ is still 0 so the next update to $Q(E; right)$ uses 0 in the max over the next state's action

	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		-7.2				
Transition 2				-9.3		
Transition 3					- -10.91	
Transition 4				-9.09		

(d) After running the Q-learning algorithm using the four transitions given above, construct a greedy policy using the current values of the Q-table in states C, E and F. (1 Point)

The greedy policy in states C, E and F is given by,

$$\pi(s) = \begin{Bmatrix} \text{jump}, & \text{for } s = C \\ \text{right}, & \text{for } s = E \\ \text{right}, & \text{for } s = F \end{Bmatrix}$$

(e) For the Q-Learning algorithm to converge to the optimal Q function, a necessary condition is that the learning rate, α_t, which is the learning rate at the t-th time step would need to satisfy the Robinns-Monroe condtion. In here, the time step t refers to the t-th time we are updating the value of the Q value of the state-action pair (s, a). Would the following values for learning rate α_t obey Robbins Monroe conditions ? (3 Points)

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

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

The series $\sum_{i=1}^{\infty} \frac{1}{t}$ is harmonic series and it does not converge. In fact, one can rewrite the series in the following way (by re-grouping terms)

$$\begin{aligned} \sum_{i=1}^{\infty} \frac{1}{t} &= 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \cdots \\ &> 1 + \frac{1}{2} + \frac{1}{2} + \frac{1}{4} + \frac{1}{4} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} \cdots \\ &= 1 + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \cdots = \infty \end{aligned}$$

A generalization of the Harmonic series is the p-series (Hyperharmonic series) defined as $\sum_{i=1}^{\infty} \frac{1}{t^p}$ for any +ve real number p. The p-series converges for all $p > 1$ (overharmonic series) and diverges for all $p \leq 1$. So, one can now use the above property to get the following results.

α_t	$\sum \alpha_t$	$\sum \alpha_t^2$	Algo converges
$\frac{1}{t}$	∞	$< \infty$	Yes
$\frac{1}{t^2}$	$< \infty$	$< \infty$	No

(f) A RL agent collects experiences of the form $(s_t, a_t, r_{t+1}, s_{t+1}, a_{t+1})$ to update Q values. At each time step, to choose an action, the agent follows a fixed policy π with probablity 0.5 or chooses an action in uniform random fashion. Assume the updates are applied infinitely often, state-action pairs are visited infintely often, the discount factor $\gamma < 1$ and the learning rate scheduling is appropriate.

(i) The Q learning agent performs following update

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

Will this update converge to the optimal Q function ? Why or Why not ? If not, will it converge to anything at all ? (2.5 Points)

Yes. Q-learning is an off-policy control algorithm and the target is based on Bellman optimality condition. Provided other conditions as stated in the question are true, this update will converge to Q^*.

(ii) Another reinforcemnt learning called SARSA agent, performs the following update

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

Will this update converge to the optimal Q function ? Why or Why not ? If not, will it converge to anything at all ? (2.5 Points)

No, it will not converge to optimal Q function. Rather it will be converge $Q^{\pi'}$ where π' is a policy that, at each time step, chooses an action based on policy π with probablity 0.5 or chooses an action in uniform random fashion. Recall that SARSA update is on-policy.