AI3000/CS5500: Reinforcement Learning
Assignment 3

Taha Adeel Mohammed
CS20BTECH11052

Problem 1: Model Free Methods

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

Using the first visit Monte-Carlo method, we have:

$$V(A) = \frac{14+15+17+16+15}{5} = 15.4$$
$$V(B) = \frac{13+14+16+15+14}{5} = 14.4$$
$$V(C) = \frac{12+13+15+14+13}{5} = 13.4$$
$$V(D) = \frac{12+12+12+11}{4} = 11.75$$
$$V(E) = \frac{11+11+11+10+9}{5} = 10.2$$
$$V(F) = \frac{10+10+10+10+9}{5} = 9.8$$
$$V(G) = 0$$

(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)

The States C, E, and F are likely to have different value estimates if evalluated using every visit MC as compared to first visit MC. This is because the first visit MC method only considers the first visit to a state, and ignores all subsequent visits. The every visit MC method, on the other hand, considers all visits to a state. Hence, since C, E, and F are visited multiple times in some episodes, they might have different final estimated values.

(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)

In Q-learning, the update rule is given by:

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

Hence applying these updates for the given episode, we have:

	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 points)

After running the Q-learning algorithm, the Q-table for C, E, and F is as follows:

	C	E	F
left	-10	-10	-10.91
right	-	-9.09	-10
jump	-7.2	-	-

In the greedy policy, we choose the action with the highest Q-value for each state. Hence, the greedy policy is as follows:

$$\pi(C) = \text{jump}$$
$$\pi(E) = \text{right}$$
$$\pi(F) = \text{right}$$

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

The Robbins-Monroe condition is given by:

$$\sum_{t=1}^{\infty} \alpha_t = \infty \quad \text{and} \quad \sum_{t=1}^{\infty} \alpha_t^2 < \infty$$

For $\alpha_t = \frac{1}{t}$, we have:

$$\sum_{t=1}^{\infty} \alpha_t = \sum_{t=1}^{\infty} \frac{1}{t} = \infty$$
$$\sum_{t=1}^{\infty} \alpha_t^2 = \sum_{t=1}^{\infty} \frac{1}{t^2} = \frac{\pi^2}{6} < \infty$$

Hence it satisfies the Robbins-Monroe condition.

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

For $\alpha_t = \frac{1}{t^2}$, we have:

$$\sum_{t=1}^{\infty} \alpha_t = \sum_{t=1}^{\infty} \frac{1}{t^2} = \frac{\pi^2}{6} < \infty$$

Hence it does not satisfy the Robbins-Monroe condition.

(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.

Given that the agent follows a fixed policy π with probablity 0.5, and chooses an action in uniform random fashion with probablity 0.5. Let π_R denote the random policy. Then the agent can be said to be following a policy π', where:

$$\pi'(a|s) = \frac{1}{2}\pi(a|s) + \frac{1}{2}\pi_R(a|s)$$

(i) The Q learning agent performs following update

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

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

As we have shown above, the agent can be viewed as following a fixed behavioural policy π', i.e $a_t \sim \pi'(a_t|s_t)$. Then the update rule is give by:

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

This is basically Q-learning with a fixed behavioural policy. Hence, it will converge to the optimal Q function $Q*$, as long it visits all state-action pairs infinitely often, and the learning rate scheduling is appropriate. The target policy $\pi*$ is the greedy policy with respect to $Q*$, and the behavioural policy π' is the fixed policy π with probablity 0.5, and the random policy π_R with probablity 0.5. Hence, the conditions for convergence of Q-learning with a fixed behavioural policy are satisfied, and the update will converge to the optimal Q function $Q*$, as it has been shown in class.

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

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

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

We can notice that the update rule for above SARSA is the same as the update rule used in Temporal Difference (TD) on policy learning algorithm with the fixed behavioural policy π', i.e $a_t \sim \pi'(a_t|s_t)\ \forall\ t$. Hence, Q will converge to $Q^{\pi'}$ and not the optimal Q function $Q*$. Since our policy is not being updated every time step, the policy will not converge to the optimal policy $\pi*$, and hence Q will not converge to the optimal Q function $Q*$. However, it will converge to $Q^{\pi'}$.

Problem 2: Game of Tic-Tac-Toe

Implemented in `CS20BTECH11052-tictactoe.ipynb`.