HW 4 - CSE 473 Spring 2021

Due Date: Friday, May 21st, 2021 at *11:59 pm PDT*
Total Points: 20 points

- This homework can be done individually or in groups of two.
    - Groups of two should submit their assignment as a group on gradescope.

1.Short Answer Questions[3 Points]

1. (1 pt) For Q-learning to converge to the optimal policy, is it necessary for the agent to eventually start acting according to the optimal policy?
2. (1 pt) Briefly describe the conditions that are required for Q-learning to converge to the optimal policy.
3. (1 pt) Will approximate Q-learning with feature vectors always converge to the optimal policy, if the agent visits all of the states a sufficient number of times and the learning rate is decreased at an appropriate rate?

2.Reinforcement Learning[8 Points]

You are playing a peculiar card game, but unfortunately you were not paying attention when the rules were described. You did manage to pick up that for each round you will be holding one of three possible cards [Ace, King, Jack] ([A, K, J], for short) and you can either Bet or Pass, in which case the dealer will reward you points and possibly switch out your card. You decide to use Q-Learning to learn to play this game, in particular you model this game as an MDP with states [A, K, J], actions [Bet, Pass] and discount $\gamma = 1$. To learn the game you use $\alpha = 0.25$.

A) Say you observe the following rounds of play (in order):

s	a	s'	r
A	Bet	K	4
J	Pass	A	0
K	Pass	A	-4
K	Bet	J	-12
J	Bet	A	4
A	Bet	A	-4

(4 pts) What are the estimates for the following Q-values as obtained by Q-learning? All Q-values are initialized to 0.

i) Q(J, Pass) = _____

ii) Q(J, Bet) = _____

B) For this next part, we will switch to a feature based representation. We will use two features:

$$f_1(s, a) = 1$$

$$f_2(s, a) = \begin{cases} 1 & a = \text{Bet} \\ 0 & a = \text{Pass} \end{cases}$$

Starting from initial weights of 0, compute the updated weights after observing the following samples:

s	a	s'	r
A	Bet	K	8
K	Pass	A	0

(2 pts) What are the weights after the first update, in other words, after using the first sample?

i) w1 = _____

ii) w2 = _____

(2 pts) What are the weights after the second update, in other words, after using the second sample?

iii) w1 = _____

iv) w2 = _____

3.Reinforcement Learning [3 pts]

Given the following list of Q-values for state s and the set of actions {*Left*, *Right*, *Fire*}:

Q(s, *Left*) = 0.15

Q(s, *Right*) = 0.95

Q(s, *Fire*) = 0.5

What is the probability that we will take each action on our next move when following an ε-greedy exploration policy (assuming all random movements are chosen uniformly from all actions)?

Action	Probability, in terms of ε
Left	
Right	
Fire	

4.Probability and Uncertainty[6 points]

A. (2 pts) Suppose Boolean random variables A and B are **independent** of each other (They can only have two values of True and False). Determine the missing entries x and y in the joint distribution of P(A, B) shown below.$P(A = T,\ B = T) = 0.4$

$$P(A = T,\ B = F) = 0.1$$
$$P(A = F,\ B = T) = x$$
$$P(A = F,\ B = F) = y$$

B. For these problems, assume we have three random variables A, B, C with possible instantiations a, b, c, respectively.
   1. (2 pts) The conditionalized version of the general product rule is P(a,b|c) = P(a|b,c)P(b|c). Show how to derive this rule using the definition of conditional probability.
   2. (2 pts) If P(a,b,c) = 0.01, P(a|b,c) = 0.2, and P(b|c) = 0.1. What is P(c)?