CS 188
Spring 2020

Exam Prep 5

Q1. RL

Pacman is in an unknown MDP where there are three states [A, B, C] and two actions [Stop, Go]. We are given the following samples generated from taking actions in the unknown MDP. For the following problems, assume $\gamma = 1$ and $\alpha = 0.5$.

(a) We run Q-learning on the following samples:

s	a	s'	r
A	Go	B	2
C	Stop	A	0
B	Stop	A	-2
B	Go	C	-6
C	Go	A	2
A	Go	A	-2

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

(i) $Q(C, Stop) =$ ________________________

(ii) $Q(C, Go) =$ ________________________

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

- $f_1(s, a) = 1$
- $f_2(s, a) = \begin{cases} 1 & a = \text{Go} \\ -1 & a = \text{Stop} \end{cases}$

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

s	a	s'	r
A	Go	B	4
B	Stop	A	0

What are the weights after the first update? (using the first sample)

(i) $w_1 =$ ________________________

(ii) $w_2 =$ ________________________

What are the weights after the second update? (using the second sample)

(iii) $w_1 =$ ________________________

(iv) $w_2 =$ ________________________

Q2. Reinforcement Learning

(a) Each True/False question is worth 1 points. Leaving a question blank is worth 0 points. **Answering incorrectly is worth -1 points.**

(i) [*true* or *false*] Temporal difference learning is an online learning method.

(ii) [*true* or *false*] Q-learning: Using an optimal exploration function leads to no regret while learning the optimal policy.

(iii) [*true* or *false*] In a deterministic MDP (i.e. one in which each state / action leads to a single deterministic next state), the Q-learning update with a learning rate of $\alpha = 1$ will correctly learn the optimal q-values (assume that all state/action pairs are visited sufficiently often).

(iv) [*true* or *false*] A small discount (close to 0) encourages greedy behavior.

(v) [*true* or *false*] A large, negative living reward ($\ll 0$) encourages greedy behavior.

(vi) [*true* or *false*] A negative living reward can always be expressed using a discount < 1.

(vii) [*true* or *false*] A discount < 1 can always be expressed as a negative living reward.

(b) Given the following table of Q-values for the state A and the set of actions $\{Forward, Reverse, Stop\}$, what is the probability that we will take each action on our next move when we following an ϵ-greedy exploration policy (assuming any random movements are chosen uniformly from all actions)?

$Q(A, Forward) = 0.75$
$Q(A, Reverse) = 0.25$
$Q(A, Stop) = 0.5$

Action	Probability (in terms of ϵ)
Forward	
Reverse	
Stop	