CS 188
Summer 2023

Discussion 5D

Q1. Reinforcement Learning

Imagine an unknown environments with four states (A, B, C, and X), two actions ($\leftarrow$ and $\rightarrow$). An agent acting in this environment has recorded the following episode:

s	a	s'	r	Q-learning iteration numbers (for part b)
A	$\rightarrow$	B	0	1, 10, 19, ...
B	$\rightarrow$	C	0	2, 11, 20, ...
C	$\leftarrow$	B	0	3, 12, 21, ...
B	$\leftarrow$	A	0	4, 13, 22, ...
A	$\rightarrow$	B	0	5, 14, 23, ...
B	$\rightarrow$	A	0	6, 15, 24, ...
A	$\rightarrow$	B	0	7, 16, 25, ...
B	$\rightarrow$	C	0	8, 17, 26, ...
C	$\rightarrow$	X	1	9, 18, 27, ...

(a) Consider running model-based reinforcement learning based on the episode above. Calculate the following quantities:

$\hat{T}(B, \rightarrow, C) =$ _____

$\hat{R}(C, \rightarrow, X) =$ _____

(b) Now consider running Q-learning, repeating the above series of transitions in an infinite sequence. Each transition is seen at multiple iterations of Q-learning, with iteration numbers shown in the table above.

After which iteration of Q-learning do the following quantities first become nonzero? (If they always remain zero, write *never*).

$Q(A, \rightarrow)$? _____

$Q(B, \leftarrow)$? _____

(c) True/False: For each question, you will get positive points for correct answers, zero for blanks, and negative points for incorrect answers. Circle your answer **clearly**, or it will be considered incorrect.

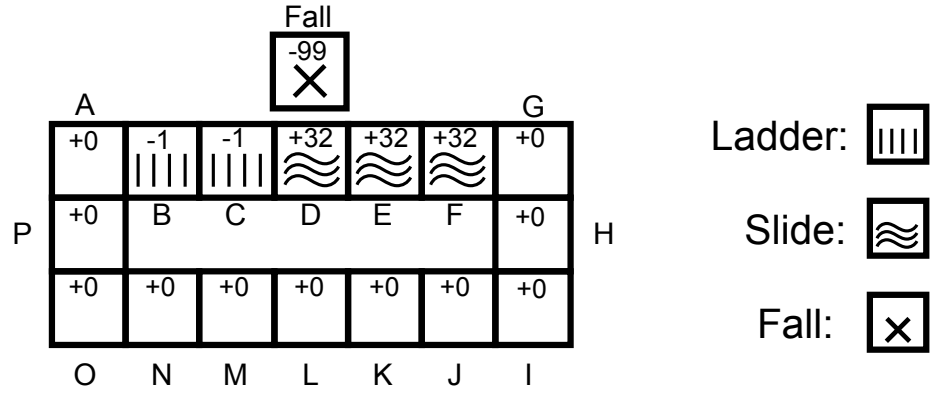
(i) [*true* or *false*] In Q-learning, you do not learn the model.

(ii) [*true* or *false*] For TD Learning, if I multiply all the rewards in my update by some nonzero scalar p, the algorithm is still guaranteed to find the optimal policy.

(iii) [*true* or *false*] In Direct Evaluation, you recalculate state values after each transition you experience.

(iv) [*true* or *false*] Q-learning requires that all samples must be from the optimal policy to find optimal q-values.

Q2. RL: Dangerous Water Slide



Suppose now that several years have passed and the water park has not received adequate maintenance. It has become a dangerous water park! Now, each time you choose to move to (or remain at) one of the ladder or slide states (states B-F) there is a chance that, instead of ending up where you intended, you fall off the slide and hurt yourself. The cost of falling is -99 and results in you getting removed from the water park via ambulance. The new MDP is depicted above.

Unfortunately, you don't know how likely you are to fall if you choose to use the slide, and therefore you're not sure whether the fun of the ride outweighs the potential harm. You use reinforcement learning to figure it out!

For the rest of this problem assume $\gamma = 1.0$ (i.e. no future reward discounting). You will use the following two trajectories through the state space to perform your updates. Each trajectory is a sequence of samples, each with the following form: (s, a, s', r).

Trajectory 1: (A, East, B, -1), (B, East, C, -1), (C, East, D, +32)
Trajectory 2: (A, East, B, -1), (B, East, Fall, -99)

(a) What are the values of states A, B, and C after performing temporal difference learning with a learning rate of $\alpha = 0.5$ using only Trajectory 1?

$V(A) =$ $V(B) =$ $V(C) =$

(b) What are the values of states A, B, and C after performing temporal difference learning with a learning rate of $\alpha = 0.5$ using both Trajectory 1 and Trajectory 2?

$V(A) =$ $V(B) =$ $V(C) =$

(c) What are the values of states/action pairs (A, South), (A, East), and (B, East) after performing Q-learning with a learning rate of $\alpha = 0.5$ using both Trajectory 1 and Trajectory 2?

$Q(A, \text{South}) =$ $Q(A, \text{East}) =$ $Q(B, \text{East}) =$