CSCD84: Artificial Intelligence
Worksheet: RL

Q1: True or False

- Temporal difference learning is a model-based learning method. [True/False]
- In a deterministic MDP, Q-learning with a learning rate of $\alpha = 1$ cannot learn the optimal q-values. [True/False]

Q2: Properties of reinforcement learning algorithms

Assuming we run for infinitely many steps, for which exploration policies is Q-learning guaranteed to converge to the optimal Q-values for all state-action pairs. Assume we chose reasonable values for α and all states of the MDP are connected via some path. (Select all that apply)

□ A fixed optimal policy.

□ A fixed policy taking actions uniformly at random.

□ An ϵ-greedy policy.

□ A greedy policy.

Q3: Grid-World Reinforcement

Consider the grid-world given below and Pacman who is trying to learn the optimal policy. All shaded states are terminal states, *i.e.*, the MDP will take the exit action and collect the corresponding reward once it arrives in a shaded state. The other states have the *North*, *East*, *South*, *West* actions available, which deterministically move Pacman to the corresponding neighboring state (or have Pacman stay in place if the action tries to move out of the grid). Assume the discount factor $\gamma = 0.5$ and the Q-learning rate $\alpha = 0.5$ for all calculations. Pacman starts in state (1, 3). For this question, Pacman does not have to learn the values for the terminal (shaded) states, these are given to him and remain fixed.

(a) The agent starts from the top left corner and you are given the following episodes from runs of the agent through this grid-world. Each line in an Episode is a tuple containing (s, a, s', r). Fill in the following Q-values obtained from direct evaluation from the samples (round to three decimal places):

$Q((4,2),N) =$ $\qquad$ $Q((1,2),N) =$ $\qquad$ $Q((2,2),E) =$

3	ᗧ	+10	-100	+100
2				
1	-100	-100	+30	-100
	1	2	3	4

Episode 1	Episode 2	Episode 3	Episode 4	Episode 5
(1,3), S, (1,2), 0	(1,3), S, (1,2), 0	(1,3), S, (1,2), 0	(1,3), S, (1,2), 0	(1,3), S, (1,2), 0
(1,2), S, (1,1), 0	(1,2), E, (2,2), 0	(1,2), E, (2,2), 0	(1,2), E, (2,2), 0	(1,2), E, (2,2), 0
(1,1), Exit, D, -100	(2,2), E, (3,2), 0	(2,2), E, (3,2), 0	(2,2), N, (2,3), 0	(2,2), E, (3,2), 0
	(3,2), E, (4,2), 0	(3,2), S, (3,1), 0	(2,3), Exit, D, +10	(3,2), E, (4,2), 0
	(4,2), N, (4,3), 0	(3,1), Exit, D, +30		(4,2), N, (4,3), 0
	(4,3), Exit, D, +100			(4,3), Exit, D, +100

(b) As we studied in the lectire, Q-learning is an online algorithm to learn optimal Q-values in an MDP with unknown rewards and transition function. Given the episodes in part (a), fill in the episode at which the following Q values first become non-zero. If the specified Q value never becomes non-zero, write never.

$$Q((1,3),S) = \qquad Q((2,2),E) = \qquad Q((3,2),E) =$$

(c) What is the value of the optimal value function V^* at the following states: (Unrelated to answers from previous parts)

$$V^*((1,3)) = \qquad V^*((2,2)) = \qquad V^*((3,2)) =$$

(d) Using Q-Learning updates, what are the following Q-values after the above five episodes:

$$Q((3,2),N) = \qquad Q((1,2),S) = \qquad Q((2,2),E) =$$

(e) Consider a feature based representation of the Q-value function: $Q_f(s,a) = w_1 f_1(s) + w_2 f_2(s) + w_3 f_3(a)$, where $f_1(s)$ and $f_2(s)$ are the x coordinate of the state and the y coordinate of the state, respectively. Furthermore, $f_3(\text{N}) = 1$, $f_3(\text{W}) = 2$, $f_3(\text{S}) = 3$, $f_3(\text{E}) = 4$, $f_3(\text{Exit}) = 1$.

i) Given that all w_i are initially 0, what are their values after the first episode:

$$w_1 = \qquad w_2 = \qquad w_3 =$$

ii) Assume the weight vector $\mathbf{w}$ is equal to $(1,1,1)$. What is the action prescribed by the Q-function in state $(2,2)$?