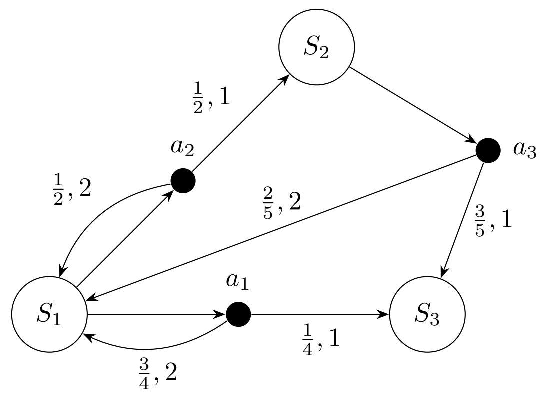
ECE421: Introduction to Machine Learning — Fall 2024

Worksheet 7: MDP and RL

Q1 An MDP with a single goal state S_3 is given below. The probability and distance of each transition is denoted next to the corresponding edge. For instance, the pair $(\frac{1}{2}, 2)$ next to the edge connecting a_2 to S_1 denotes the probability and distance of this transition, respectively.



1.a Given the optimal expected cost $C^\star(S_1) = 7, C^\star(S_2) = 4.2$, and $C^\star(S_3) = 0$, calculate the optimal policy for state S_1, *i.e.*, $\pi^\star(S_1)$.
[**NOTE:** In this question, we used distance/cost instead of reward/return, and $C^\star(\cdot)$ denotes the optimal expected distance. In such MDP, the optimal policy minimizes the expected distance.]

1.b Suppose that we follow policy π, where we pick action a_2 in state S_1 and action a_3 in state S_2. Calculate the expected cost of S_1 and S_2 for this policy, *i.e.*, $C_\pi(S_1)$ and $C_\pi(S_2)$.

Q2 **[True or False]**

2.a Temporal difference learning is a model-based learning method. [True/False]

2.b In a *deterministic MDP*, Q-learning with a learning rate of $\alpha = 1$ cannot learn the optimal q-values. [True/False]

Q3 **[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) A fixed optimal policy.

(b) A fixed policy taking actions uniformly at random.

(c) An ϵ-greedy policy.

(d) A greedy policy.

Q4 **[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** (N), **East** (E), **South** (S), **West** (W) 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**.

3		+10	-100
2			
1	-100	-100	+100
	1	2	3

Table 1: Pacman grid-world. Assume that the discount factor $\gamma = 0.5$ and the Q-learning rate $\alpha = 0.5$.

4.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

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

Q-values obtained from direct evaluation from the samples.
[**NOTE:** You do not need to simplify your answer and can leave it as summation of fractions.]

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

4.b As we studied in the lecture, 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.

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

4.c What is the value of the optimal value function $V^\star$ at the following states: (Unrelated to answers from previous parts)

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

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

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

4.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, $f_3(\text{N}) = 1$, $f_3(\text{W}) = 2$, $f_3(\text{S}) = 3$, $f_3(\text{E}) = 4$, $f_3(\text{Exit}) = 1$.

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

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

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