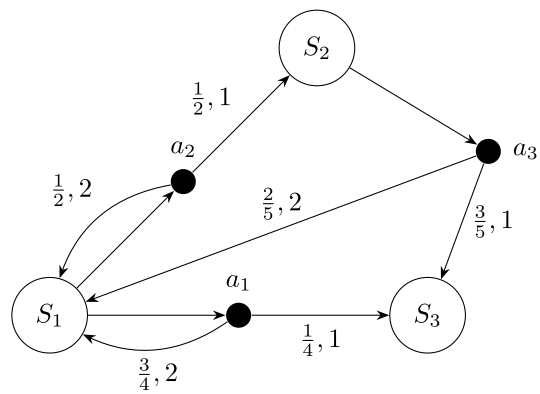
# ECE421: Introduction to Machine Learning — Fall 2024

## Worksheet 7 Solution: 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.]

**Answer.** We use $Q^\star(s, a)$ to denote the expected cost of reaching a goal state if one starts in state $s$, executes action $a$ and then acts according to the optimal policy. We must calculate $Q^\star(S_1, a_1)$ and $Q^\star(S_1, a_2)$ to find $\pi^\star(S_1)$. Observe that

$$Q^\star(S_1, a_1) = \frac{1}{4}\left(1 + C^\star(S_3)\right) + \frac{3}{4}\left(2 + C^\star(S_1)\right) = 7$$
$$Q^\star(S_1, a_2) = \frac{1}{2}\left(1 + C^\star(S_2)\right) + \frac{1}{2}\left(2 + C^\star(S_1)\right) = 7.1$$

Since $Q^\star(S_1, a_1) < Q^\star(S_1, a_2)$, $\pi^\star(S_1) = a_1$.

On a separate note, observe that since $S_2$ has only one available action, we have

$$Q^\star(S_2, a_3) = C^\star(S_2) = 4.2.$$

**1.b** Suppose that we follow policy $\pi$, 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 $C_\pi(S_2)$.

**Answer.** Since the given policy chooses $a_2$ at $S_1$, we simply ignore $a_1$ during our computation. We first generate the following set of equations.

$$C_\pi(S_1) = Q_\pi(S_1, a_2) = \frac{1}{2}(1 + C_\pi(S_2)) + \frac{1}{2}(2 + C_\pi(S_1))$$
$$C_\pi(S_2) = Q_\pi(S_2, a_3) = \frac{3}{5}(1 + C_\pi(S_3)) + \frac{2}{5}(2 + C_\pi(S_1))$$
$$C_\pi(S_3) = 0.$$

Solving this system of equations would result in

$$C_\pi(S_1) = 2.2/0.3 = 7.33,$$
$$C_\pi(S_2) = 1.4 + 0.4C_\pi(S_1) = 1.4 + 0.4(7.33) = 4.333.$$

**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]

**Answer.** **2.a** is False and **2.b** is False. Note that in stochastic MDPs, $\alpha = 1$ would prevent convergence to the optimal Q-values because the updates would overwrite previous information without averaging over the stochastic outcomes. However, this is not a concern in deterministic environments.

**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 $\alpha$ 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 $\epsilon$-greedy policy.

(d) A greedy policy.

**Answer.** (b) and (c). For Q-learning to converge, every state-action pair $(s, a)$ must occur infinitely often. A uniform random policy and an $\epsilon$-greedy policy will achieve this. A fixed optimal policy will not take any suboptimal actions and so will not explore enough. Similarly a greedy policy will stop taking actions the current Q-values suggest are suboptimal, and so will never update the Q-values for supposedly suboptimal actions. (This is problematic if, for example, an action most of the time yields no reward but occasionally yields very high reward. After observing no reward a few times, Q-learning with a greedy policy would stop taking that action, never obtaining the high reward needed to update it to its true value.)

## 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.]

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

**Answer.** In lecture, we had used direct evaluation to estimate $V_\pi$. Here, we are asked to use direct evaluation to estimate $Q(s, a)$. We should average the sum of discounted rewards for transitions like $s, a, \ldots$.

$Q((1, 2), N) = 0$ since we had never observed this (state, action) pair.
$Q((2, 2), E) = \frac{-100/4+100/4+100/8}{3}$

**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.

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

**Answer.**
$Q((1, 3), S)$ : Never
$Q((2, 2), E)$ : Episode 5
$Q((3, 2), S)$ : Episode 3

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

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

**Answer.**
$V^*((1,3)) = \frac{100}{16}$
$V^*((2,2)) = \frac{100}{4}$
$V^*((3,2)) = \frac{100}{2}$

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

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

**Answer.**
On Episode 2: $Q((3,2),N) = (1-\alpha)0 + \alpha[0 + \gamma(-100)] = -25$
On Episode 1: $Q((1,2),S) = (1-\alpha)0 + \alpha[0 + \gamma(-100)] = -25$
On Episode 5: $Q((2,2),E) = (1-\alpha)0 + \alpha[0 + \gamma \max_{a'} Q((3,2),a)] = \frac{25}{4}$

**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:

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

**Answer.**
$w_1 = 0 + \frac{1}{2}[(0 + \frac{1}{2}(-100)) - 0] \cdot f_1((1,2)) = -25$
$w_2 = 0 + \frac{1}{2}[(0 + \frac{1}{2}(-100)) - 0] \cdot f_2((1,2)) = -50$
$w_3 = 0 + \frac{1}{2}[(0 + \frac{1}{2}(-100)) - 0] \cdot f_3(S) = -75$

**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 $(2,2)$?

**Answer.** E