Artificial Intelligence (521495A), Spring 2022
Exercise 6 : Reinforcement learning
Deadline for reports: Wed 2.3.2021 23:59 (+1h)

This handout contains three problems related to reinforcement learning part of the course. Problems 1 and 2 are pre-exercises to support learning and solutions to them are provided in **solutions6.pdf**. For Problem 3, the answer is not given and you should return a report for your solution. **Include in the report only the solution to Problem 3**. The solution gives max. 1 point, which is taken into account in course grading.

Problems 1-3: Background material is provided in Lectures 13 (Track 2). See also the course book Chapter 21.

Problem 1. Similar to Problem 4 in Exercise 5, consider a game where you repeatedly draw a card (with replacement) that is 2, 3, or 4. This time you don't know how many of 2, 3, or 4 cards are in the deck (i.e., the distribution of cards and due to that the transition function of MDP is unknown). You can either Draw or Stop if the total score of the cards you have drawn is less than 7. Otherwise, you must Stop. When you Stop, your reward is equal to your total cumulative score (up to 6), or zero if you exceed 6 (i.e., 7 or higher). When you Draw, you receive no reward. There is no discount ($\gamma = 1$). Apply model-based reinforcement learning to estimate the transition and reward functions based on training episodes below. Each line in an Episode is a tuple containing (s, a, s', r).

Episode 1	Episode 2	Episode 3
0, Draw, 3, 0	0, Draw, 4, 0	0, Draw, 2, 0
3, Draw, 6, 0	4, Draw, 7, 0	2, Draw, 4, 0
6, Stop, Done, +6	7, Stop, Done, 0	4, Stop, Done, +4

Episode 4	Episode 5	Episode 6
0, Draw, 2, 0	0, Draw, 2, 0	0, Draw, 3, 0
2, Draw, 4, 0	2, Draw, 5, 0	3, Draw, 5, 0
4, Draw, 6, 0	5, Draw, 8, 0	5, Stop, Done, +5
6, Stop, Done, +6	8, Stop, Done, 0	

Estimate the transition $T(s, a, s')$ and reward $R(s, a, s')$ function values for

(a) $T(0, Draw, 2)$

(b) $T(0, Draw, 3)$

(c) $T(4, Stop, Done)$

(d) $R(4, Draw, 6)$

(e) $R(6, Stop, Done)$

(f) $R(4, Stop, Done)$

Problem 2. Consider the same setting as in Problem 1, but now apply model-free reinforcement learning. The learning rate is $\alpha = 0.5$ and there is no discount ($\gamma = 1$). Use fixed reward function (similar to Problem 4 in Exercise 5) as follows

$$\mathrm{R}(\mathrm{s},\ \mathrm{Stop}\ ,\mathrm{s}') = \mathrm{s} \text{ if } \mathrm{s} \leq 6$$
$$\mathrm{R}(\mathrm{s}, \mathrm{a}, \mathrm{s}') = 0 \text{ otherwise.}$$

(a) Use direct evaluation and training episodes from Problem 1 to update the state policy values.

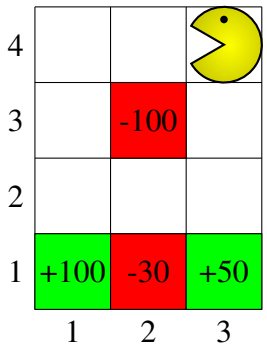
(b) Formulate Temporal Difference learning update equation and calculate state policy values using the observations from training episodes 1-6. V values are initialized to zeros.

(c) Use direct evaluation to calculate following Q-values (i.e., averaging the discounted reward).

1. $Q(0, Draw)$

2. $Q(2, Draw)$

Problem 3* [1p]. Consider a 3 x 4 grid world given below where an agent is exploring the environment and trying to learn the optimal policy. Rewards for taking the *Exit* action from one of the shaded states are shown. Taking this action moves the agent to the Done state, and the MDP terminates. Other actions are, N: *move north*, E: *move east*, S: *move south*, W: *move west*. Assume $\gamma = 1$ and $\alpha = 0.5$. For Q-learning, training episodes are given below. Each line in an Episode is a tuple containing (s, a, s', r).



Episode 1	Episode 2
(3,4), S, (3,3), 0	(3,4), S, (3,3), 0
(3,3), S, (3,2), 0	(3,3), S, (3,2), 0
(3,2), S, (3,1), 0	(3,2), W, (2,2), 0
(3,1), Exit, Done, +50	(2,2), S, (2,1), 0
	(2,1), Exit, Done, -30

Episode 3	Episode 4
(3,4), W, (2,4), 0	(3,4), W, (2,4), 0
(2,4), W, (1,4), 0	(2,4), W, (1,4), 0
(1,4), S, (1,3), 0	(1,4), S, (1,3), 0
(1,3), E, (2,3), 0	(1,3), S, (1,2), 0
(2,3), Exit, Done, -100	(1,2), S, (1,1), 0
	(1,1), Exit, Done, +100

(a) Calculate the following Q-values using direct evaluation (i.e., averaging the discounted reward) from the training episodes.

1. $Q((2,4),W)$

2. $Q((3,3),S)$

(b) Formulate the Q-learning update equation and run the iteration to update Q-values based on training episodes. All Q-values are initilized to zero. Iterate through episodes 1, 2, 3 and 4 in that particular order **two** times. How many non-zero Q-values there are after **two** runs through the episodes?