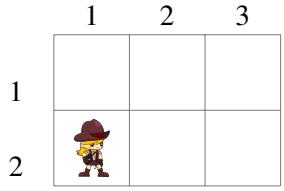
Exercise 1 :

Consider the following variation of the example from the lecture, where we do not know neither the probabilistic transition function or the reward function:



Apply Q-learning on this example, with a learning rate $\alpha = 0.9$, and a discount factor $\gamma = 0.8$. Specifically:

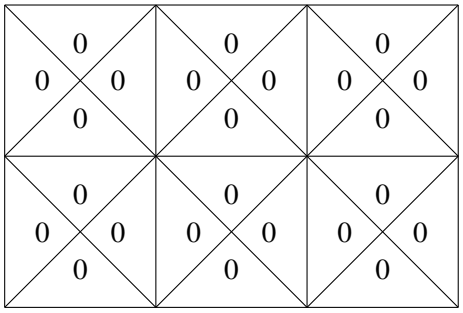
(i) Define what is the table that the agent uses for learning. Initialize it to 0.

(ii) Consider that the agent performs the following sequence of actions. How the table is updated? For each of the steps indicate the new table and how the values are updated.

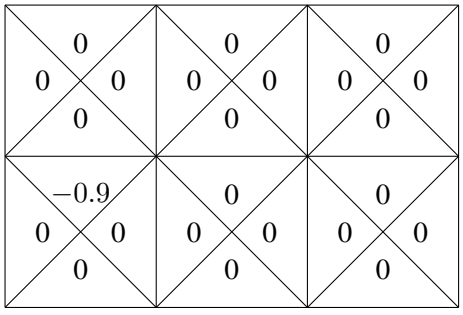
- north, go to (1, 1), receive reward of -1
- east go to (2, 1), receive reward of $+5$
- east go to (2, 2), receive reward of -2
- west go to (1, 2), receive reward of $+2$

(iii) If the agent decides to act greedily, what action would the agent perform next? Justify your answer.

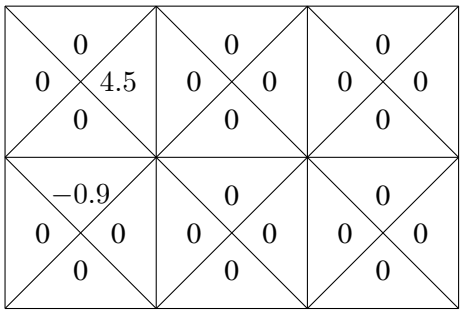
Solution:



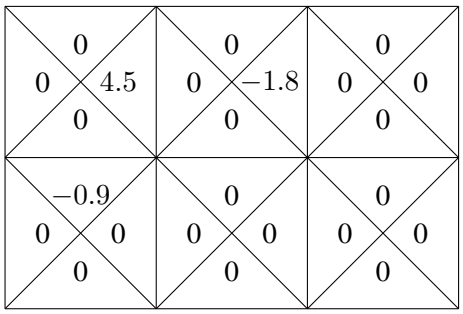
- north, go to (1, 1), receive reward of -1



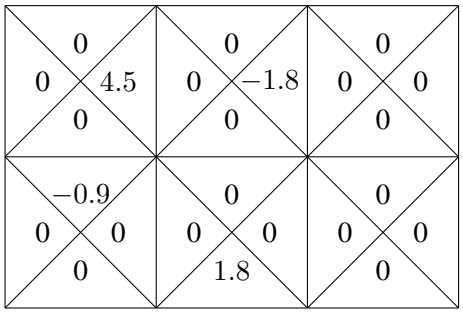
- east go to (2, 1), receive reward of $+5$



- east go to (2, 2), receive reward of -2



- west go to (1, 2), receive reward of $+2$



- The agent will apply any of the actions that maximize $Q(s, a)$, so it will choose to move in any direction except north.

Exercise 2 :

Consider an environment with two states $S = \{s_1, s_2\}$ and two actions $A = \{a_1, a_2\}$. The initial state is s_1, and there are no terminal states.

When we apply action a_1, we move to state s_1. In that case, we obtain a reward of $+1$ if we came from s_1, or $+7$ if we come from s_2.

When we apply action a_2 we move to s_2. In that case, we obtain a reward of -2 if we come from s_1, or $+2$ if we come from s_2.

We assume that the discount factor is $\gamma = 0.95$ and the learning rate is $\alpha = 0.9$.

(i) Specify the probabilistic transition function.

(ii) Specify the reward function. Note: in this example the reward depends on the action and the state you are coming from, so it should be a function of the form $R(s, a)$.

(iii) Trace through the first few steps of the Q-learning algorithm and with all Q values initially set to zero. Sample actions according to the explotation principle, choosing always the best action according to the current Q-values, and if both actions tie prefer a_1 by default. Explain why it is necessary to force exploration through probabilistic choice of actions, in order to ensure convergence to the true Q values.

(iv) Trace through the first few steps of the Q-learning algorithm with all Q values initially set to zero. Sample actions according to ϵ-greedy extrategy, using the following "random" selection for each of the steps: random (a_2 is chosen), random (a_1 is chosen), greedy, random (a_2 is chosen).

(v) Determine what is the optimal value function V^* ($\gamma = 0.95$).
Hint: consider that V^* is the solution to the Bellmann Equation.
Hint2: When you have an expression that maximizes over the expected reward of multiple actions you can guess what the optimal policy is. If you do, then you should at the end confirm that the values you got satisfy the Bellmann Equation.

(vi) Determine what is the optimal Q-value function Q^* ($\gamma = 0.95$). Hint: Explain how the optimal value and Q-value function are related, and use the values in the previous part.

Solution:

(i) Probabilistic transition function is: $P(s_1 \mid s_1, a_1) = 1$. $P(s_2 \mid s_1, a_2) = 1$ $P(s_1 \mid s_2, a_1) = 1$ $P(s_2 \mid s_2, a_2) = 1$

In this case the transition function is deterministic.

(ii) The reward function is:

$R(s_1, a_1) = 1$

$R(s_1, a_2) = -2$

$R(s_2, a_1) = +7$

$R(s_2, a_2) = +2$

(iii) Starting at s_1 with the following Q-values

	s_1	s_2
a_1	0	0
a_2	0	0

We pick action a_1. We reach s_1, and update the Q-value:

	s_1	s_2
a_1	0.9	0
a_2	0	0

As, $Q(s_1, a_1) > Q(s_1, a_2)$, we will keep always selecting a_1 and never explore what happens if we execute a_2, which could lead to better rewards.

	s_1	s_2
a_1	0.9	0
a_2	0	0

(iv) Now, we run Q=learning according to the ϵ-greedy strategy with the randomness provided in the exercise:

	s_1	s_2
a_1	0	0
a_2	0	0

We pick action a_2. We reach s_2, and update the Q-value: $expectedval = R(s_1, a_2) + \gamma \max_a Q(s_2, a) = -2 + 0 = -2$

$Q'(s_1, a_2) = 0.1 \cdot Q(s_1, a_2) + 0.9 \cdot expectedval = -1.8$

	s_1	s_2
a_1	0	0
a_2	-1.8	0

We pick action a_1. We reach s_1, and update the Q-value: $expectedval = R(s_2, a_1) + \gamma \max_a Q(s_1, a) = +7 + 0 = +7$

$Q'(s_2, a_1) = 0.1 \cdot Q(s_2, a_1) + 0.9 \cdot expectedval = 6.3$

	s_1	s_2
a_1	0	6.3
a_2	-1.8	0

We greedily pick action a_1. We reach s_1, and update the Q-value: $expectedval = R(s_1, a_1) + \gamma \max_a Q(s_1, a) = +1 + 0$

$Q'(s_1, a_1) = 0.1 \cdot Q(s_1, a_1) + 0.9 \cdot expectedval = 0.9$

	s_1	s_2
a_1	0.9	6.3
a_2	-1.8	0

We randomly pick action a_2. We reach s_2, and update the Q-value: $expectedval = R(s_1, a_2) + \gamma \max_a Q(s_2, a) = +1 + 0.95 \cdot 6.3 = 5.985$

$Q'(s_1, a_2) = 0.1 \cdot Q(s_1, a_2) + 0.9 \cdot expectedval = -0.18 + 5.3865 \approx 5.2$

	s_1	s_2
a_1	0.9	6.3
a_2	5.2	0

The algorithm is already starting to converge towards the optimal policy. However, if it wasn't due to randomly picking a_2, when the agent assigned it a reward of -1.8, compared to the reward of $a_1(0.9)$, we would have stuck in a local minima again.

(v) V^* is the optimal value function, which satisfies the Bellman equation:

$$\forall s \in S, V^*(s) = \max_a \sum_{s'} P(s' \mid s, a)(R(s, a, s') + \gamma V^*(s')$$

Replacing for our tiny example, with simplified reward function, we get

$$V^*(s_1) = \max(R(s_1, a_1) + \gamma V^*(s_1), R(s_1, a_2) + \gamma V^*(s_2))$$
$$V^*(s_2) = \max(R(s_2, a_1) + \gamma V^*(s_1), R(s_2, a_2) + \gamma V^*(s_2))$$

Replacing the known values we get:

$$V^*(s_1) = \max(1 + 0.95V^*(s_1), -2 + 0.95V^*(s_2))$$
$$V^*(s_2) = \max(7 + 0.95V^*(s_1), 2 + 0.95V^*(s_2))$$

Intuitively, the cycle that provides maximum reward each 2 steps is $s_1 \xrightarrow{a_2} s_2 \xrightarrow{a_1} s_1$, with a total reward of 5 each $2steps$. By fixing these actions we get: (note: to make this formal, one could compute the value for each of the following equations).

$$V^*(s_1) = -2 + 0.95V^*(s_2)$$
$$V^*(s_2) = 7 + 0.95V^*(s_1)$$

Solving via substitution:

$$V^*(s_1) = -2 + 0.95(7 + 0.95V^*(s_1)) \approx 4.65 + 0.9V^*(s_1)$$

$$V^*(s_1) = 46.5$$
$$V^*(s_2) = 51.175$$

We can confirm that these values satisfy the Bellmann equation:

$$46.5 = \max(1 + 0.95 \cdot 46.5, -2 + 0.95 \cdot 51.175) = \max(45.175, 46.6) \approx 46.5$$

(we get an approximation due to rounding in some operations)

$$V^*(s_2) = \max(7 + 0.95 \cdot 46.5, 2 + 0.95 \cdot 51.157) = \max(51.175, 50.6) = 51.175$$

(vi) $Q^*(s, a)$ is the optimal value starting from s and choosing a. Therefore, in this example:

$Q^*(s_1, a_1) = R(s_1, a_1) + \gamma \cdot V^*(s_1) = 1 + 0.95 \cdot 46.5 = 45.175$

$Q^*(s_1, a_2) = R(s_1, a_2) + \gamma \cdot V^*(s_2) = -2 + 0.95 \cdot 51.175 = 46.5$

$Q^*(s_2, a_1) = R(s_2, a_1) + \gamma \cdot V^*(s_1) = 7 + 0.95 \cdot 46.5 = 51.175$

$Q^*(s_2, a_2) = R(s_2, a_2) + \gamma \cdot V^*(s_2) = 2 + 0.95 \cdot 51.175 = 50.6$

Q^*	s_1	s_2
a_1	45.175	46.5
a_2	51.175	50.6

Note also that these are exactly the values that we computed when we were confirming that V^* satisfies the Bellmann Equation. Basically, we were confirming if our choice of policy was indeed the correct one or choosing some other action could yield better expected rewards.

Exercise 3 :

Please decide for each of the following statements whether it is true or false and justify your answer (1-3 sentences per statement).

1. It is possible to learn the optimal policy for an MDP by interacting with the environment, even if we do not know a priori the transition probabilities or the reward function.
2. Q-learning is guaranteed to converge to the optimal Q-value function independently of the strategy to select actions at each step.

Solution:

1. True, by applying Q-learning for example (one needs to be careful of having a good choice of action selection and update the learning rate).
2. False. For example, if we always select the same action, we cannot possibly learn anything about what will be the reward or states that could be reached by applying different actions.