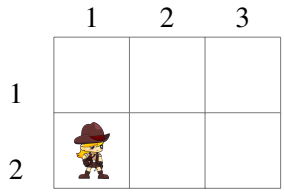
Exercise 1 :

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.

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.

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.
