Exercise 2

Marco Di Francesco - 100632815
ELEC-E8125 - Reinforcement Learning

September 26, 2022

1 Task 1

1.1 Question 1.1

The agent is the boat, the environment is the grid that the boat can explore.

1.2 Question 1.2

The value is 0 because does not get updated from the initial value. The algorithm looks at the next states and tries to take the best one, because there is no next state (the episode ends) the value stays 0.

1.3 Question 1.3

With reward -2 for hitting the rock, it made the dangerous path.

No, it does not choose the same path. With reward -10 for hitting the rock it made the safe path.

2 Task 2

No, considering $\epsilon = 10^{-4}$ the algorithm does not converge in 30 iterations.

Generally the policy is converging before the value function, this anyway depends on the meaning we have of convergence and it's related to ϵ . In our example with $\epsilon = 10^{-4}$ it converges after 38 iterations and the policy does not have changes after 35 iterations, but if we would have taken $\epsilon = 10^{-2}$ then the value function would have converged before the policy.

3 Task 3

As said in the previous task, it took 38 iterations.

4 Task 4

- Mean: 0.6242525662068833
- Standard deviation: 1.3741594932705201

4.1 Question 4.1

They are both discounted sums of the rewards (in our case of the last reward), but the discounted return is a random variable that considers only one episode, while the value function is the expectation of all the possible episodes. Moreover, the value function can start from every step, while the discounted return starts from a fixed step.

4.2 Question 4.2

In the value iteration algorithm we are making the assumption that the observation corresponds to the state, this means in our example that the boat knew the ending state and computed the best path starting from there. On the other hand, in an unknown environment we must give rewards while exploring our surroundings. For this reason our approach cannot directly be applied to this problem.

As a side note, we are making the assumption that the environment is not changing in time, that may of may not be a realistic assumption to do.