Exercise 2

Stefano Baggetto - 879118
ELEC-E8125 - Reinforcement Learning

September 27, 2020

Question 1

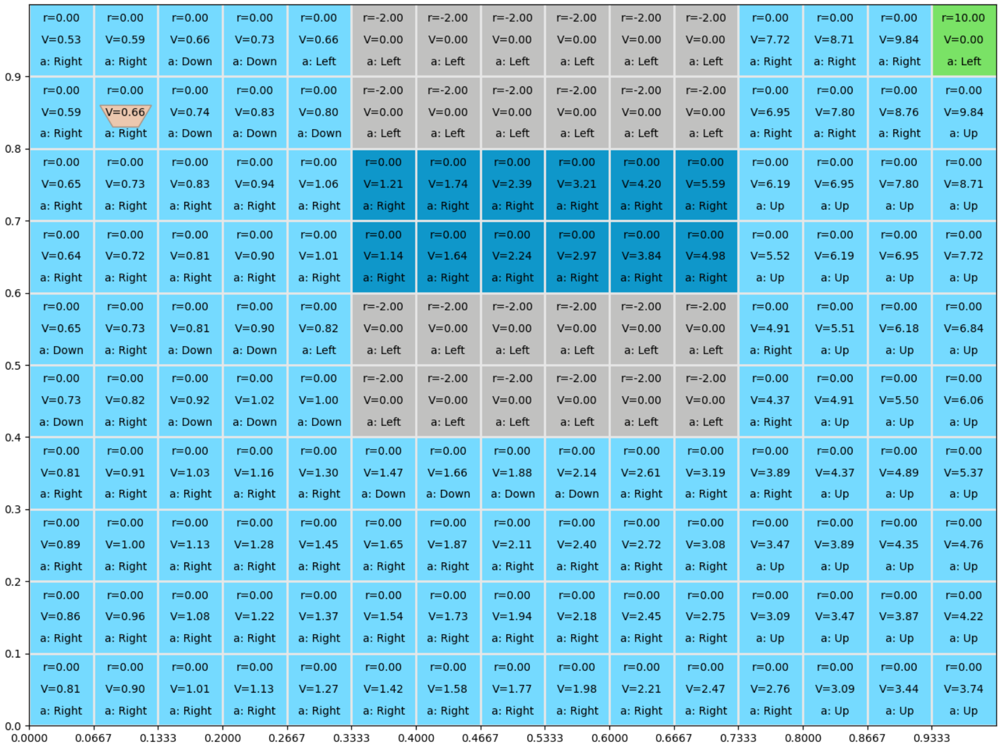
The agent is the sailor on a boat. The sailor and the boats are situated in a grid environment representing a different state. The states are the sea with the wind, the rocks and the harbour. The sea is where the sailor can move taking into account that there is the wind which can randomly change the wanted behaviour of the boat. There is a part of the sea which is more dangerous because it's surrounded by rocks. The rocks represent some failing states. If the boat hits a rock it fails to complete the task and it's game over. Lastly, the harbour is where the sailor wants to go in order to fulfil the task successfully.

Task 1

After implementing the value iteration each block changed its value accordingly. For the update values see Figure 1.

Question 2

The value of those states is 0 because if the boat reaches one of those states, the task is finished, they are terminal states, there aren't any further possible states. If the boat reaches the harbour the process finishes with success. Instead, if the boat hits a rock in its path to the harbour the process finishes with a fail. The value function $v_\pi(s)$ is the expected return when starting in s and following the policy π but the result will be zero if there are no future states.



r=0.00 V=0.53 a: Right	r=0.00 V=0.59 a: Right	r=0.00 V=0.66 a: Down	r=0.00 V=0.73 a: Down	r=0.00 V=0.66 a: Left	r=-2.00 V=0.00 a: Left	r=-2.00 V=0.00 a: Left	r=-2.00 V=0.00 a: Left	r=-2.00 V=0.00 a: Left	r=-2.00 V=0.00 a: Left	r=-2.00 V=0.00 a: Left	r=0.00 V=7.72 a: Right	r=0.00 V=8.71 a: Right	r=0.00 V=9.84 a: Right	r=10.00 V=0.00 a: Left
r=0.00 V=0.59 a: Right	r=0.00 V=0.66 a: Right	r=0.00 V=0.74 a: Down	r=0.00 V=0.83 a: Down	r=0.00 V=0.80 a: Down	r=-2.00 V=0.00 a: Left	r=-2.00 V=0.00 a: Left	r=-2.00 V=0.00 a: Left	r=-2.00 V=0.00 a: Left	r=-2.00 V=0.00 a: Left	r=-2.00 V=0.00 a: Left	r=0.00 V=6.95 a: Right	r=0.00 V=7.80 a: Right	r=0.00 V=8.76 a: Right	r=0.00 V=9.84 a: Up
r=0.00 V=0.65 a: Right	r=0.00 V=0.73 a: Right	r=0.00 V=0.83 a: Right	r=0.00 V=0.94 a: Right	r=0.00 V=1.06 a: Right	r=0.00 V=1.21 a: Right	r=0.00 V=1.74 a: Right	r=0.00 V=2.39 a: Right	r=0.00 V=3.21 a: Right	r=0.00 V=4.20 a: Right	r=0.00 V=5.59 a: Right	r=0.00 V=6.19 a: Up	r=0.00 V=6.95 a: Up	r=0.00 V=7.80 a: Up	r=0.00 V=8.71 a: Up
r=0.00 V=0.64 a: Right	r=0.00 V=0.72 a: Right	r=0.00 V=0.81 a: Right	r=0.00 V=0.90 a: Right	r=0.00 V=1.01 a: Right	r=0.00 V=1.14 a: Right	r=0.00 V=1.64 a: Right	r=0.00 V=2.24 a: Right	r=0.00 V=2.97 a: Right	r=0.00 V=3.84 a: Right	r=0.00 V=4.98 a: Right	r=0.00 V=5.52 a: Up	r=0.00 V=6.19 a: Up	r=0.00 V=6.95 a: Up	r=0.00 V=7.72 a: Up
r=0.00 V=0.65 a: Down	r=0.00 V=0.73 a: Right	r=0.00 V=0.81 a: Down	r=0.00 V=0.90 a: Down	r=0.00 V=0.82 a: Left	r=-2.00 V=0.00 a: Left	r=-2.00 V=0.00 a: Left	r=-2.00 V=0.00 a: Left	r=-2.00 V=0.00 a: Left	r=-2.00 V=0.00 a: Left	r=-2.00 V=0.00 a: Left	r=0.00 V=4.91 a: Right	r=0.00 V=5.51 a: Up	r=0.00 V=6.18 a: Up	r=0.00 V=6.84 a: Up
r=0.00 V=0.73 a: Down	r=0.00 V=0.82 a: Right	r=0.00 V=0.92 a: Down	r=0.00 V=1.02 a: Down	r=0.00 V=1.00 a: Down	r=-2.00 V=0.00 a: Left	r=-2.00 V=0.00 a: Left	r=-2.00 V=0.00 a: Left	r=-2.00 V=0.00 a: Left	r=-2.00 V=0.00 a: Left	r=-2.00 V=0.00 a: Left	r=0.00 V=4.37 a: Right	r=0.00 V=4.91 a: Up	r=0.00 V=5.50 a: Up	r=0.00 V=6.06 a: Up
r=0.00 V=0.81 a: Right	r=0.00 V=0.91 a: Right	r=0.00 V=1.03 a: Right	r=0.00 V=1.16 a: Right	r=0.00 V=1.30 a: Right	r=0.00 V=1.47 a: Down	r=0.00 V=1.66 a: Down	r=0.00 V=1.88 a: Down	r=0.00 V=2.14 a: Down	r=0.00 V=2.61 a: Right	r=0.00 V=3.19 a: Right	r=0.00 V=3.89 a: Right	r=0.00 V=4.37 a: Up	r=0.00 V=4.89 a: Up	r=0.00 V=5.37 a: Up
r=0.00 V=0.89 a: Right	r=0.00 V=1.00 a: Right	r=0.00 V=1.13 a: Right	r=0.00 V=1.28 a: Right	r=0.00 V=1.45 a: Right	r=0.00 V=1.65 a: Right	r=0.00 V=1.87 a: Right	r=0.00 V=2.11 a: Right	r=0.00 V=2.40 a: Right	r=0.00 V=2.72 a: Right	r=0.00 V=3.08 a: Right	r=0.00 V=3.47 a: Up	r=0.00 V=3.89 a: Up	r=0.00 V=4.35 a: Up	r=0.00 V=4.76 a: Up
r=0.00 V=0.86 a: Right	r=0.00 V=0.96 a: Right	r=0.00 V=1.08 a: Right	r=0.00 V=1.22 a: Right	r=0.00 V=1.37 a: Right	r=0.00 V=1.54 a: Right	r=0.00 V=1.73 a: Right	r=0.00 V=1.94 a: Right	r=0.00 V=2.18 a: Right	r=0.00 V=2.45 a: Right	r=0.00 V=2.75 a: Right	r=0.00 V=3.09 a: Up	r=0.00 V=3.47 a: Up	r=0.00 V=3.87 a: Up	r=0.00 V=4.22 a: Up
r=0.00 V=0.81 a: Right	r=0.00 V=0.90 a: Right	r=0.00 V=1.01 a: Right	r=0.00 V=1.13 a: Right	r=0.00 V=1.27 a: Right	r=0.00 V=1.42 a: Right	r=0.00 V=1.58 a: Right	r=0.00 V=1.77 a: Right	r=0.00 V=1.98 a: Right	r=0.00 V=2.21 a: Right	r=0.00 V=2.47 a: Right	r=0.00 V=2.76 a: Right	r=0.00 V=3.09 a: Up	r=0.00 V=3.44 a: Up	r=0.00 V=3.74 a: Up

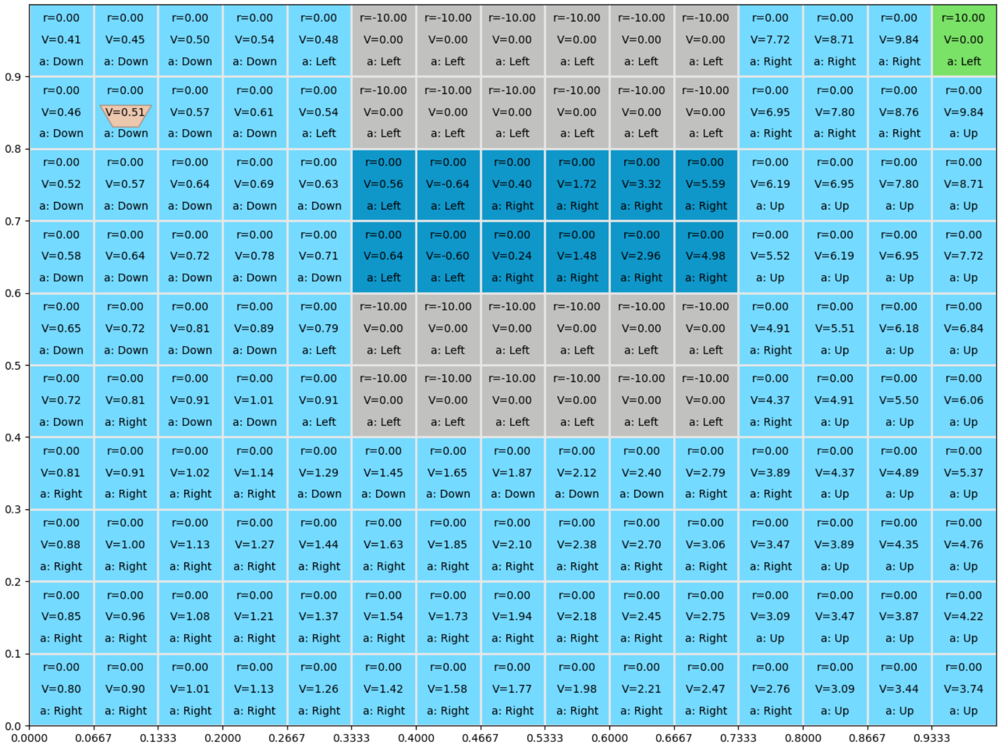
Figure 1: π and v_π for each state ($reward_{\text{rocks}} = -2$)

Task 2

The optimal policy is computed according to the value iteration adopting a greedy approach. The selected next step is the one with the highest value.

Question 3

The sailor chose the dangerous path between the rocks, successfully reaching the harbour. It will hit a rock less than 50% of the times. After changing the rock penalty to -10, the sailor decides to keep a fair distance from the rocks and followed a safe path to the harbour



r=0.00 V=0.41 a: Down	r=0.00 V=0.45 a: Down	r=0.00 V=0.50 a: Down	r=0.00 V=0.54 a: Down	r=0.00 V=0.48 a: Left	r=-10.00 V=0.00 a: Left	r=-10.00 V=0.00 a: Left	r=-10.00 V=0.00 a: Left	r=-10.00 V=0.00 a: Left	r=-10.00 V=0.00 a: Left	r=-10.00 V=0.00 a: Left	r=0.00 V=7.72 a: Right	r=0.00 V=8.71 a: Right	r=0.00 V=9.84 a: Right	r=10.00 V=0.00 a: Left
r=0.00 V=0.46 a: Down	r=0.00 V=0.51 a: Down	r=0.00 V=0.57 a: Down	r=0.00 V=0.61 a: Down	r=0.00 V=0.54 a: Left	r=-10.00 V=0.00 a: Left	r=-10.00 V=0.00 a: Left	r=-10.00 V=0.00 a: Left	r=-10.00 V=0.00 a: Left	r=-10.00 V=0.00 a: Left	r=-10.00 V=0.00 a: Left	r=0.00 V=6.95 a: Right	r=0.00 V=7.80 a: Right	r=0.00 V=8.76 a: Right	r=0.00 V=9.84 a: Up
r=0.00 V=0.52 a: Down	r=0.00 V=0.57 a: Down	r=0.00 V=0.64 a: Down	r=0.00 V=0.69 a: Down	r=0.00 V=0.63 a: Down	r=0.00 V=0.56 a: Left	r=0.00 V=-0.64 a: Left	r=0.00 V=0.40 a: Right	r=0.00 V=1.72 a: Right	r=0.00 V=3.32 a: Right	r=0.00 V=5.59 a: Right	r=0.00 V=6.19 a: Up	r=0.00 V=6.95 a: Up	r=0.00 V=7.80 a: Up	r=0.00 V=8.71 a: Up
r=0.00 V=0.58 a: Down	r=0.00 V=0.64 a: Down	r=0.00 V=0.72 a: Down	r=0.00 V=0.78 a: Down	r=0.00 V=0.71 a: Down	r=0.00 V=0.64 a: Left	r=0.00 V=-0.60 a: Left	r=0.00 V=0.24 a: Right	r=0.00 V=1.48 a: Right	r=0.00 V=2.96 a: Right	r=0.00 V=4.98 a: Right	r=0.00 V=5.52 a: Up	r=0.00 V=6.19 a: Up	r=0.00 V=6.95 a: Up	r=0.00 V=7.72 a: Up
r=0.00 V=0.65 a: Down	r=0.00 V=0.72 a: Down	r=0.00 V=0.81 a: Down	r=0.00 V=0.89 a: Down	r=0.00 V=0.79 a: Left	r=-10.00 V=0.00 a: Left	r=-10.00 V=0.00 a: Left	r=-10.00 V=0.00 a: Left	r=-10.00 V=0.00 a: Left	r=-10.00 V=0.00 a: Left	r=-10.00 V=0.00 a: Left	r=0.00 V=4.91 a: Right	r=0.00 V=5.51 a: Up	r=0.00 V=6.18 a: Up	r=0.00 V=6.84 a: Up
r=0.00 V=0.72 a: Down	r=0.00 V=0.81 a: Right	r=0.00 V=0.91 a: Down	r=0.00 V=1.01 a: Down	r=0.00 V=0.91 a: Left	r=-10.00 V=0.00 a: Left	r=-10.00 V=0.00 a: Left	r=-10.00 V=0.00 a: Left	r=-10.00 V=0.00 a: Left	r=-10.00 V=0.00 a: Left	r=-10.00 V=0.00 a: Left	r=0.00 V=4.37 a: Right	r=0.00 V=4.91 a: Up	r=0.00 V=5.50 a: Up	r=0.00 V=6.06 a: Up
r=0.00 V=0.81 a: Right	r=0.00 V=0.91 a: Right	r=0.00 V=1.02 a: Right	r=0.00 V=1.14 a: Right	r=0.00 V=1.29 a: Down	r=0.00 V=1.45 a: Down	r=0.00 V=1.65 a: Down	r=0.00 V=1.87 a: Down	r=0.00 V=2.12 a: Down	r=0.00 V=2.40 a: Down	r=0.00 V=2.79 a: Right	r=0.00 V=3.89 a: Right	r=0.00 V=4.37 a: Up	r=0.00 V=4.89 a: Up	r=0.00 V=5.37 a: Up
r=0.00 V=0.88 a: Right	r=0.00 V=1.00 a: Right	r=0.00 V=1.13 a: Right	r=0.00 V=1.27 a: Right	r=0.00 V=1.44 a: Right	r=0.00 V=1.63 a: Right	r=0.00 V=1.85 a: Right	r=0.00 V=2.10 a: Right	r=0.00 V=2.38 a: Right	r=0.00 V=2.70 a: Right	r=0.00 V=3.06 a: Right	r=0.00 V=3.47 a: Right	r=0.00 V=3.89 a: Up	r=0.00 V=4.35 a: Up	r=0.00 V=4.76 a: Up
r=0.00 V=0.85 a: Right	r=0.00 V=0.96 a: Right	r=0.00 V=1.08 a: Right	r=0.00 V=1.21 a: Right	r=0.00 V=1.37 a: Right	r=0.00 V=1.54 a: Right	r=0.00 V=1.73 a: Right	r=0.00 V=1.94 a: Right	r=0.00 V=2.18 a: Right	r=0.00 V=2.45 a: Right	r=0.00 V=2.75 a: Right	r=0.00 V=3.09 a: Up	r=0.00 V=3.47 a: Up	r=0.00 V=3.87 a: Up	r=0.00 V=4.22 a: Up
r=0.00 V=0.80 a: Right	r=0.00 V=0.90 a: Right	r=0.00 V=1.01 a: Right	r=0.00 V=1.13 a: Right	r=0.00 V=1.26 a: Right	r=0.00 V=1.42 a: Right	r=0.00 V=1.58 a: Right	r=0.00 V=1.77 a: Right	r=0.00 V=1.98 a: Right	r=0.00 V=2.21 a: Right	r=0.00 V=2.47 a: Right	r=0.00 V=2.76 a: Right	r=0.00 V=3.09 a: Up	r=0.00 V=3.44 a: Up	r=0.00 V=3.74 a: Up

Figure 2: π and v_π for each state ($reward_{\text{rocks}} = -10$)

Question 4

For a small amount of iteration, the policy and the value function don't converge. Although the policy is faster to converge than the value function The fact that the policy is faster because in order to find the optimal policy we need that the two functions are correctly related and they don't need to converge

Task 3

At 36 Iterations the value function converges with a threshold of 0.0001. After some experiments, I could lower the iterations down to 15 to make the sail still to reach the harbour.

1 Task 4

Episodes ended reaching the harbour: 598
Episodes ended hitting a rock: 402
Mean of the discounted returns: 0.6649
Standard deviation of the discounted returns: 1.3528

Question 5

the relationship between the discounted return and the value iteration function can be seen as follow. The value function can be describe as:

$$v_\pi(s) \doteq \mathbb{E}_\pi \left[G_t \mid S_t = s \right]$$

The expected discounted return G_t for a state s is, from the definition, equal to the value function v_π on the state s.

Question 6

No, in a reinforcement learning problem involving a robot exploring an unknown environment, value iteration would could not be applied directly as in this exercise. The value iteration approach is based on the Bellman equation, which requires complete knowledge of the dynamics of the environment. Therefore, without knowing the environment, we cannot use this technique.