Exercise 1

Marco Di Francesco - 100632815
ELEC-E8125 - Reinforcement Learning

October 3, 2022

Task 1.1

Look at 1 and 2

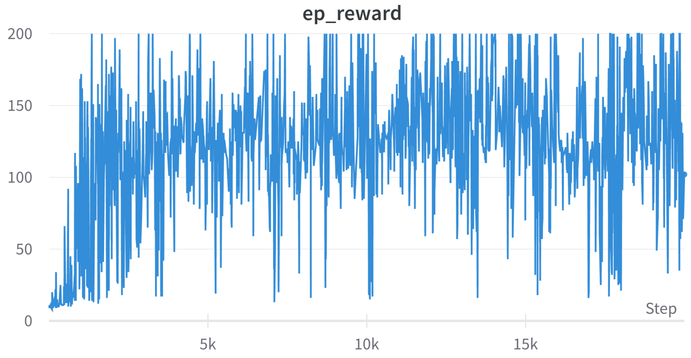

Figure 1: Reward plot for $\epsilon = 0.1$

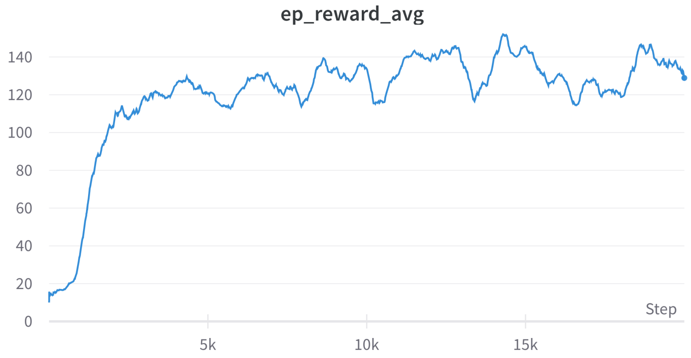

Figure 2: Average reward plot for $\epsilon = 0.1$

Look at 3 and 4, with $b = 2222$

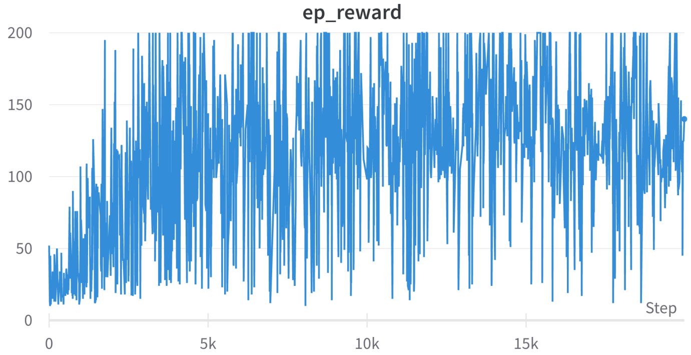

Figure 3: Reward plot using GLIE for $b = 2222$

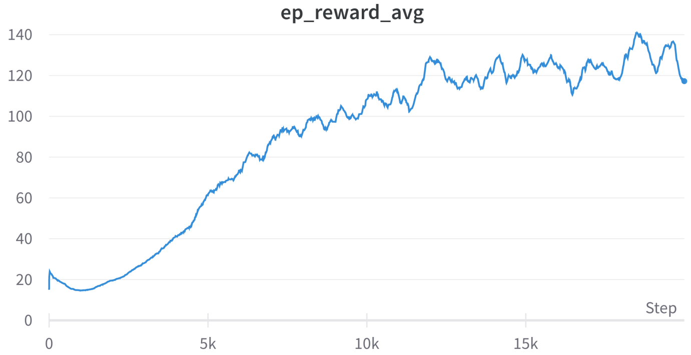

Figure 4: Average reward plot using GLIE for $b = 2222$

Task 1.2

Look at 5

Question 1

a) Before training all values will be 0s, thus black map.

b) After a single episode the value map will have very very small values the central values ($x \simeq 0, \theta \simeq 0$) and all other values are 0. This values will start from the central point (initial point) and there will be one line of values in one direction decreasing starting from the central value.

c) Halfway through the training the heatmap will be colored similarly, because at 10k iterations the values were already pretty good and it just improved marginally.

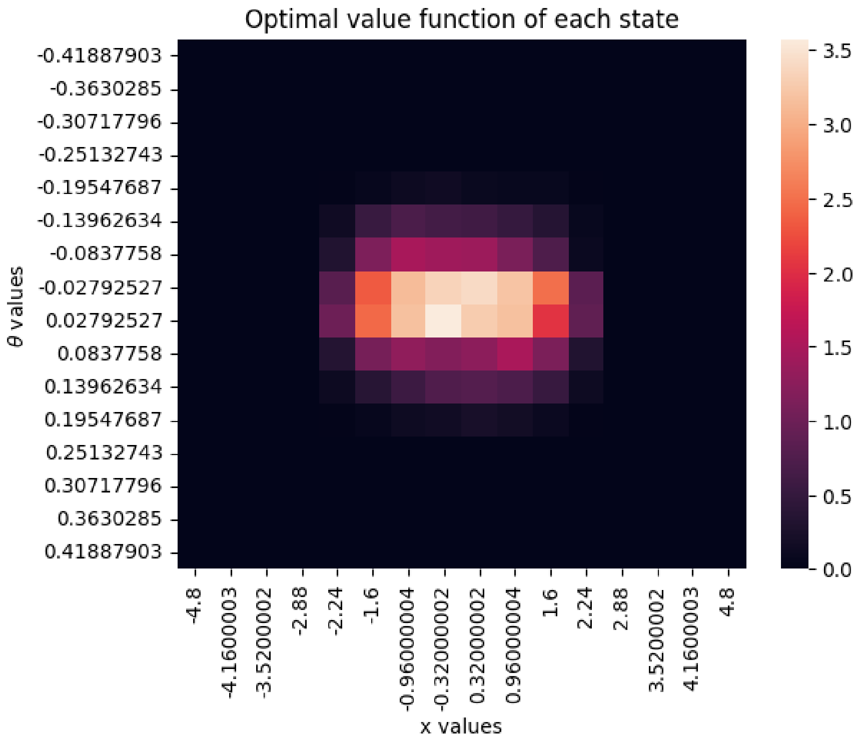

Figure 5: Heatmap of the optimal value function of each state

Question 1.3

Look at 6 and 7.

Question 2.1

The model with $\epsilon = 0$ and $initial\_q = 50$ performs better.

Question 2.2

The values of the Q function in the second case were initialized with a high value having 2 pros:

- optimistical initial value: the value may be closer to the target value
- incentive in exploration: in our case with $\epsilon = 0$ and we always chose the greedy action with $initial\_q = 0$, but with $initial\_q = 50$ we were incentived in explorating state-action paris that have were not explored

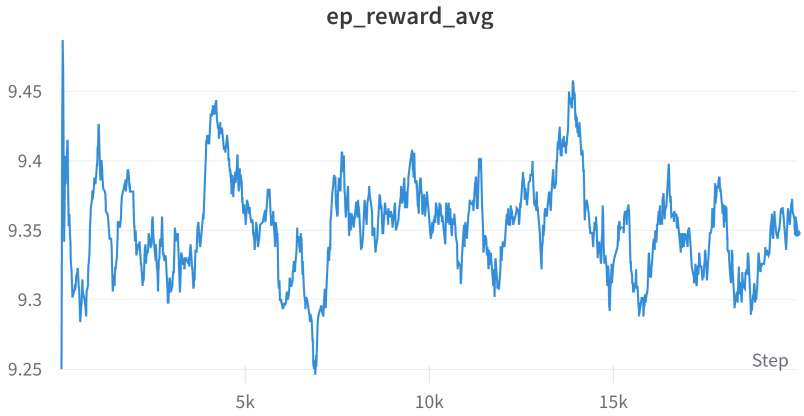

Figure 6: Average reward plot for $\epsilon = 0$ and $initial\_q = 0$

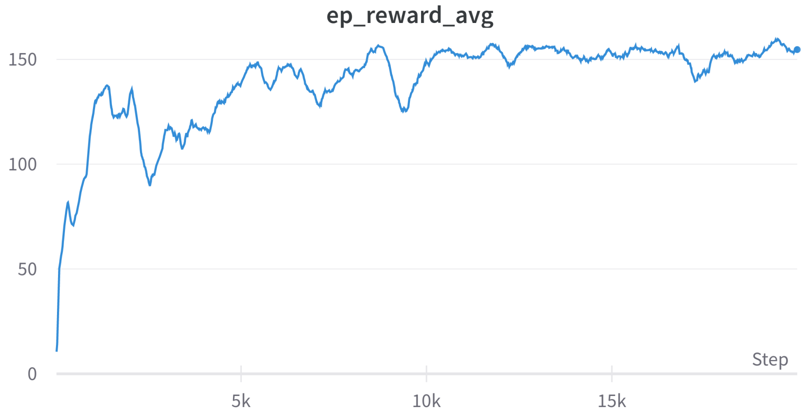

Figure 7: Average reward plot for $\epsilon = 0$ and $initial\_q = 50$

Task 2

Look at 8

Question 3

The lander does not learn how to lend between the flag poles. There may be multiple reasons for this to happen:

- The algorithm does not explore all states because ϵ is too high: having a greedy policy it may not explore all states and finds a suboptimal solution
- There are too many states to explore in the case of high states dimensionality: 20.000 iterations are not enough to expore all of them. Q-learning does not update the values it has not seen during the exploration process, for this reason if the algorithm does not explore a state it is not able to give a correct answer for this state action pair.

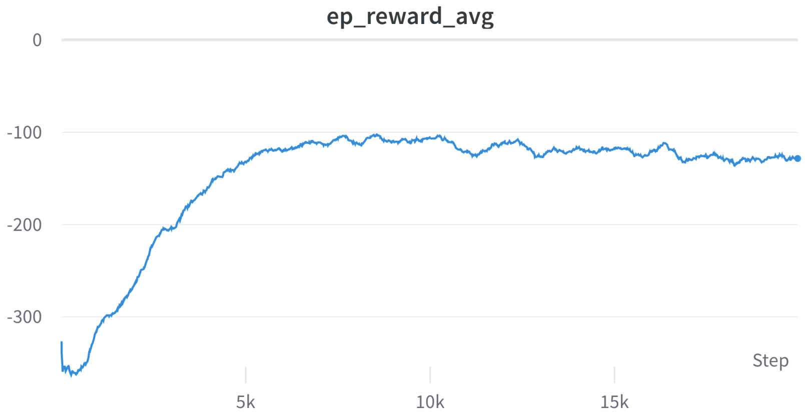

Figure 8: Average reward plot with $glie\_b = 2222$

- Continuous space systems: Q-learning does not work on continuous states systems thus requiring discretization to assign an action to a discrete value, because Q-learning is only considering that exacy state-action pair.