Exercise 3

Jari Mattila - 35260T
ELEC-E8125 - Reinforcement Learning

October 18, 2021

Task 1.1

The training performance plots for Task 1.1 are presented below using a constant value of $\epsilon = 0.2$ in Figure 1 and reducing the value of ϵ over time (i.e. using the GLIE formula from the lecture) in Figure 2.

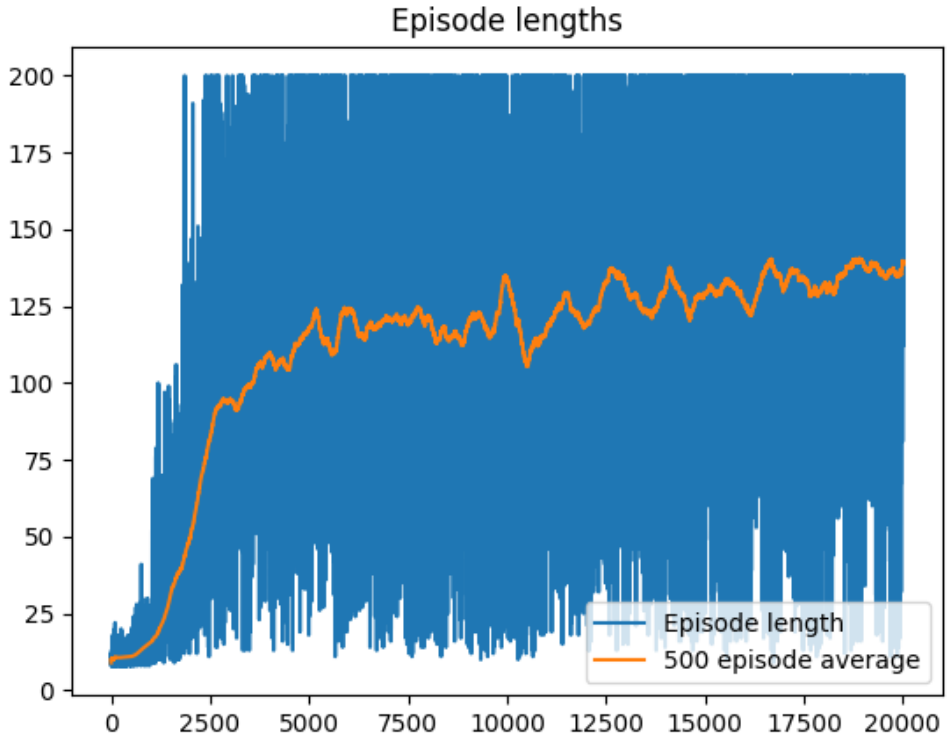

Figure 1: Training performance using a constant value of $\epsilon = 0.2$.

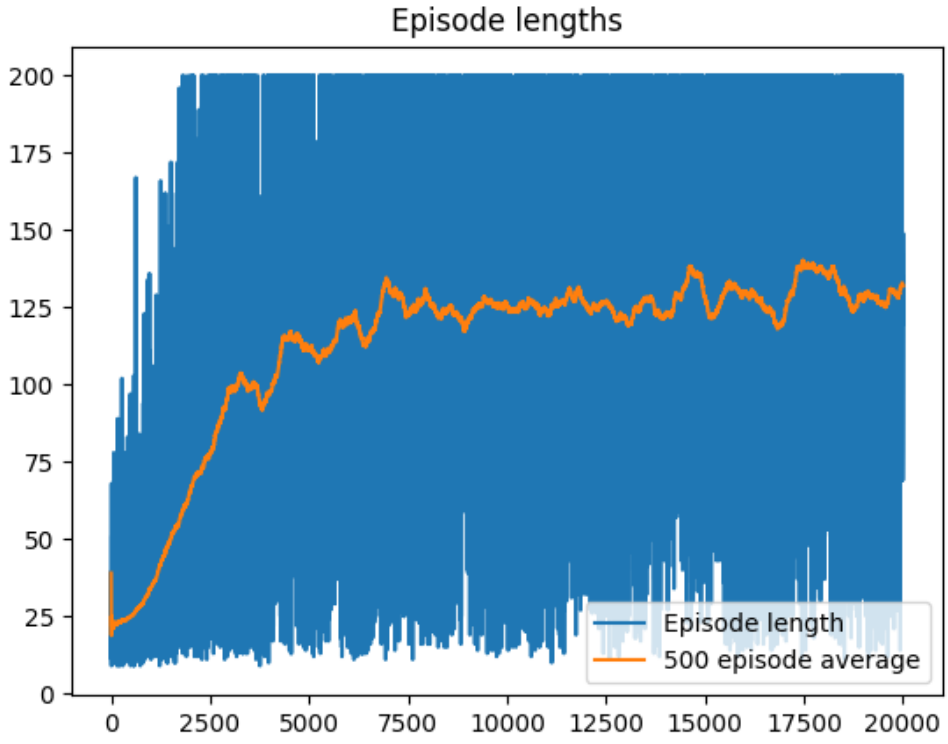

Figure 2: Training performance using the GLIE formula for ϵ.

Source files: qlearning.py, q_values_task1_1.npy, value_func_task1_1.npy

Task 1.2

The heatmap from the end of the training is presented in Figure 3. From Exercise 1 the states of the cartpole are: $\{x,\ \dot{x},\ \theta,\ \dot{\theta}\}$. For plotting, the values of the value function are averaged over $\dot{x}$ and $\dot{\theta}$.

Question 1

What do you think the heatmap would have looked like:

(a) before the training?

Before the training Q-values and the value function values are initialized to zero thus the heatmap would show a zero value for each state.

(b) after a single episode?

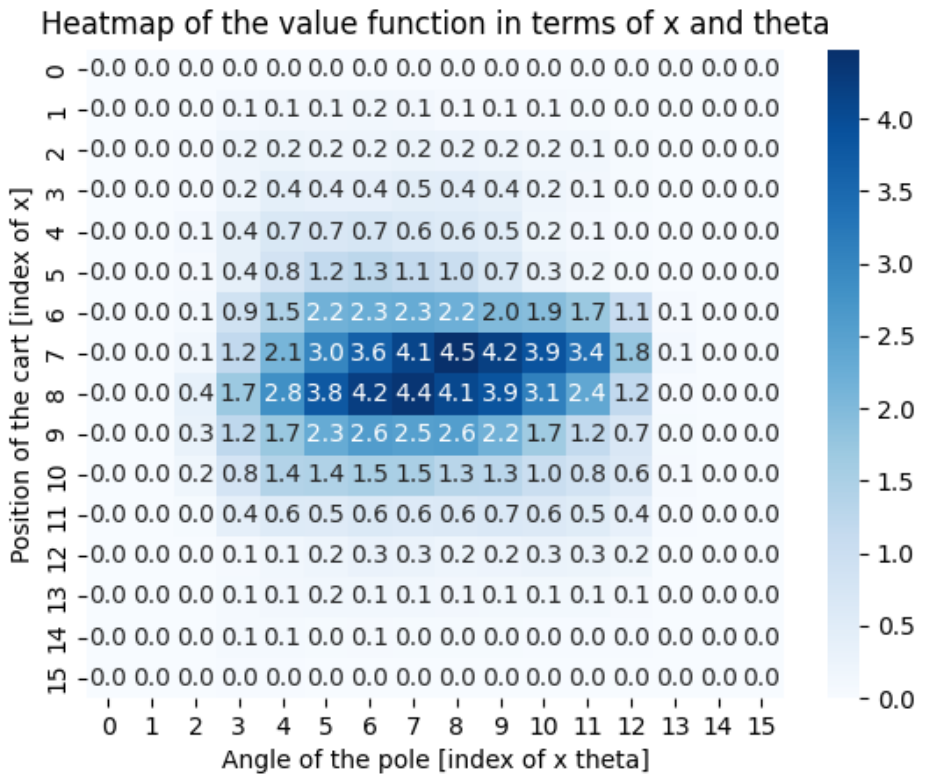

Figure 3: Heatmap of the value function in terms of x and θ.

The positions in the middle row of the heat map would show very small values (around $1..5e^{-3}$) because the initial state indices are as follows: (8, 8, 8, 7) and only the values around the initial state are affected after a single episode.

(c) halfway through the training?

The values of the value function have not yet converged to the final values, otherwise the heatmap looks very similar to the final heatmap.

Task 1.3

The training performance plots with the greedy policy ($\epsilon = 0$) are presented in Figure 4 where the Q function estimates are initialized to 0 and Figure 5 where the initial Q function estimates are set to 50 for all states and actions.

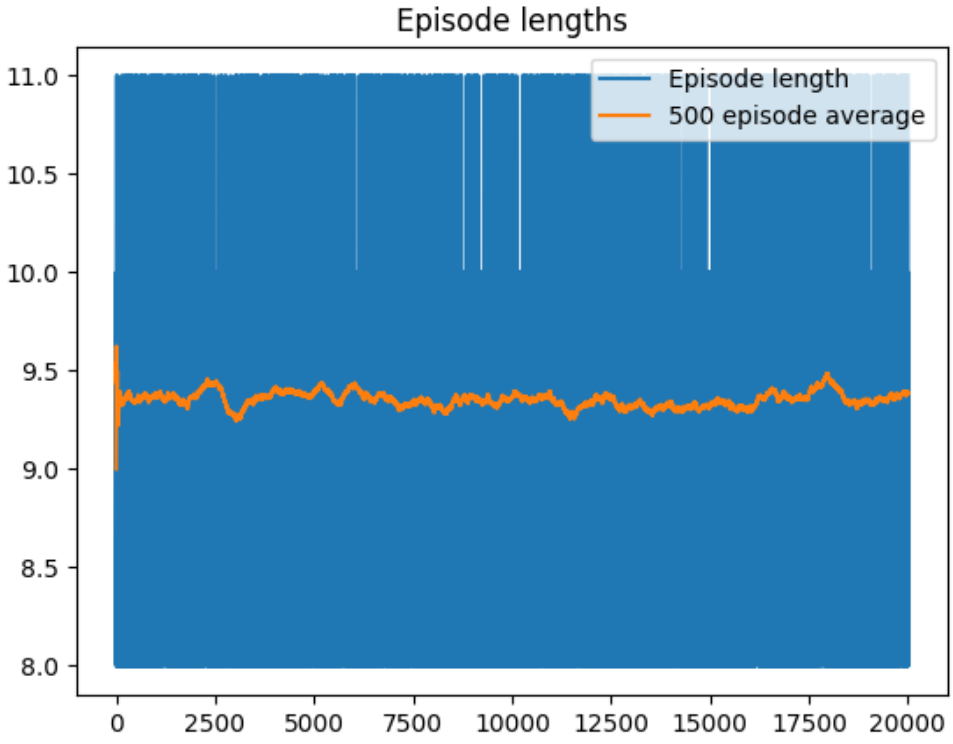

Figure 4: Training performance using ϵ with zero value (i.e. greedy policy) and keeping the intial estimates of the Q function at 0.

Question 2

Based on the results you observed in Task 1.3, answer the following questions:

Question 2.1

In which case does the model perform better?

The model performs clearly better in the case of Figure 5 where the initial estimates of the Q function are set to 50 for all states and actions.

Question 2.2

Why is this the case, and how does the initialization of Q values affect exploration?

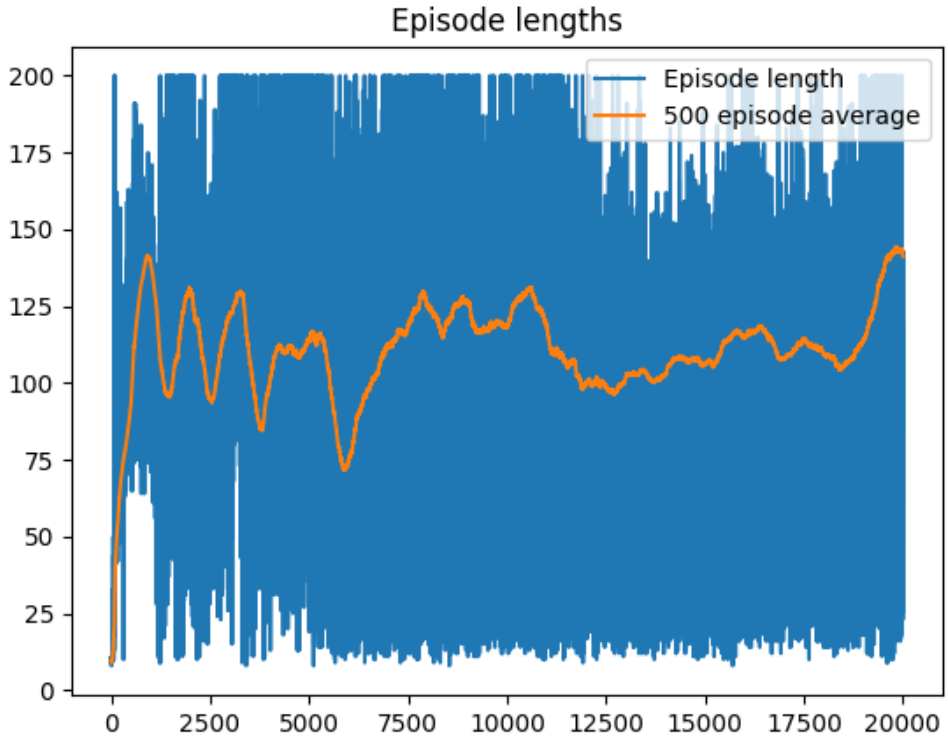

Figure 5: Training performance using ϵ with zero value (i.e. greedy policy) and setting the initial estimates of the Q function to 50 for all states and actions.

If the Q values are initialized to zero along with the greedy policy ($\epsilon = 0$), the action of zero is always selected according to the optimal policy and the policy has no way of exploring other alternatives. Setting the initial Q function estimates to 50 for all states and actions will resolve this bottleneck.

Task 2

The training performance of the lunar lander for Task 2 is presented in Figure 6.

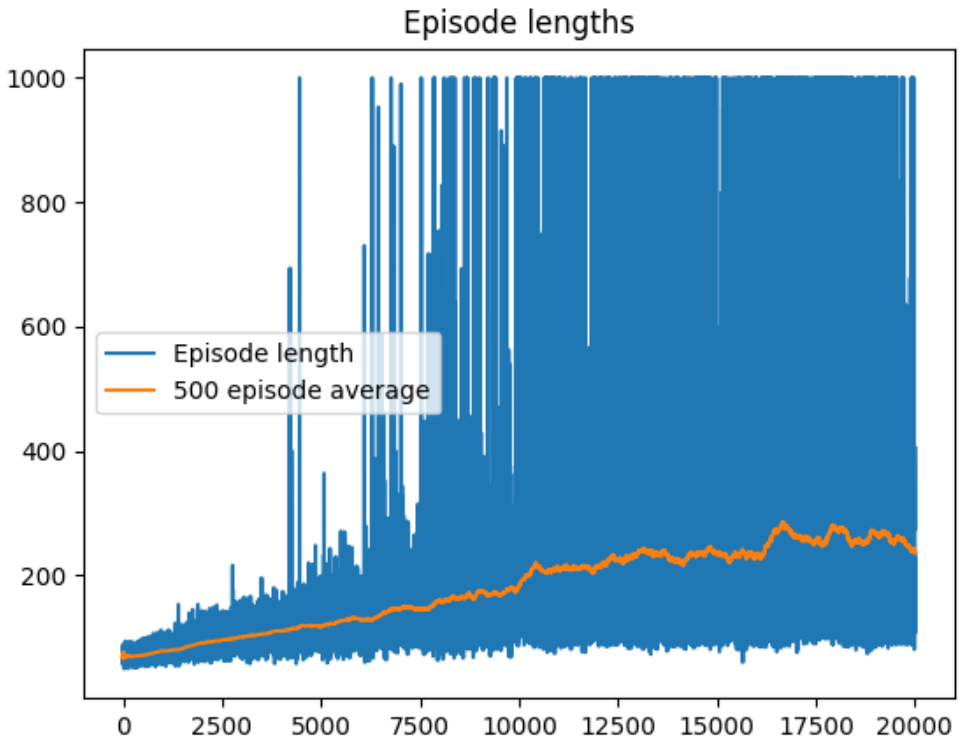

Figure 6: Training performance of the lunar lander using ϵ with GLIE.

Question 3

Is the lander able to learn any useful behaviour? Why/why not?

The lander managed to land successfully a few times so something must have been learned. The episode lengths of Figure 6 can show that the time without crashing has increased along the number of episodes during training.