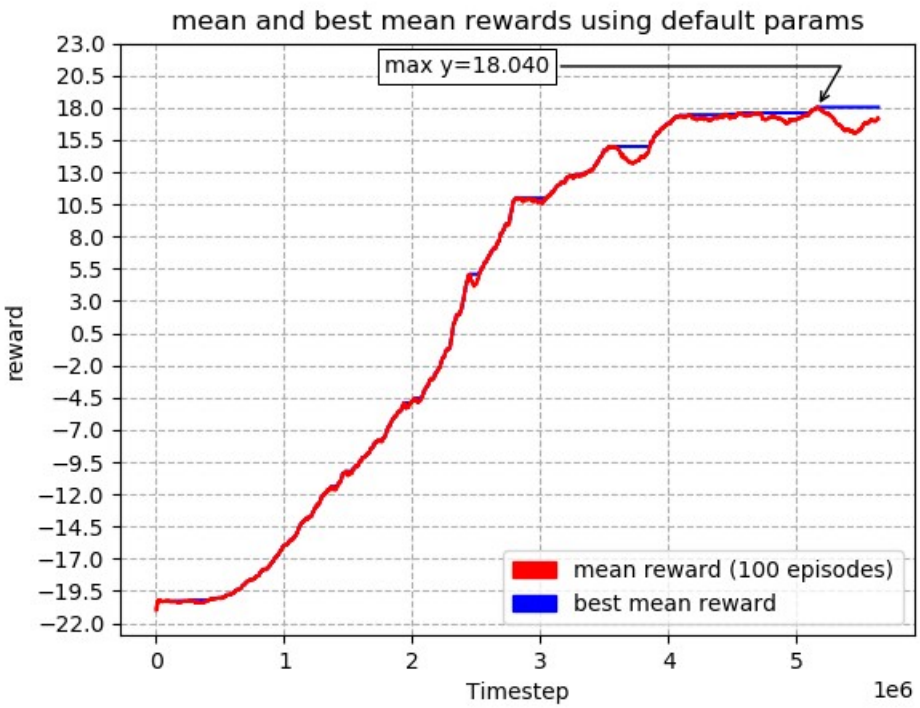
RL Project Report

057931354

Question 1:

See the plot below. Reached score (best mean reward) of 18.04 with default hyperparameters. Number of steps > 4 million.



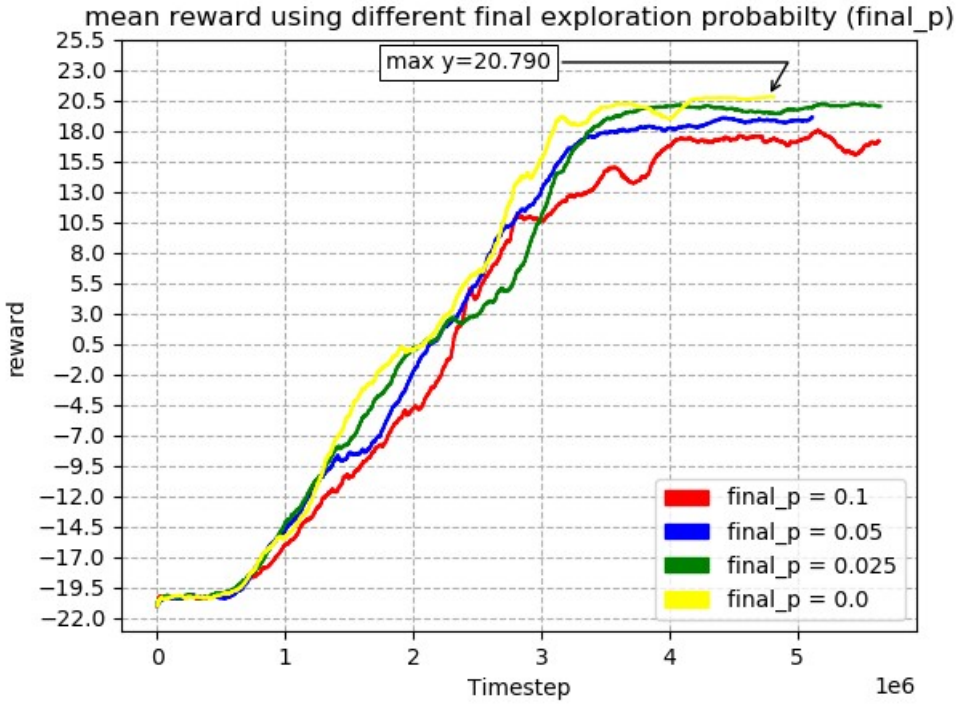
Question 2:

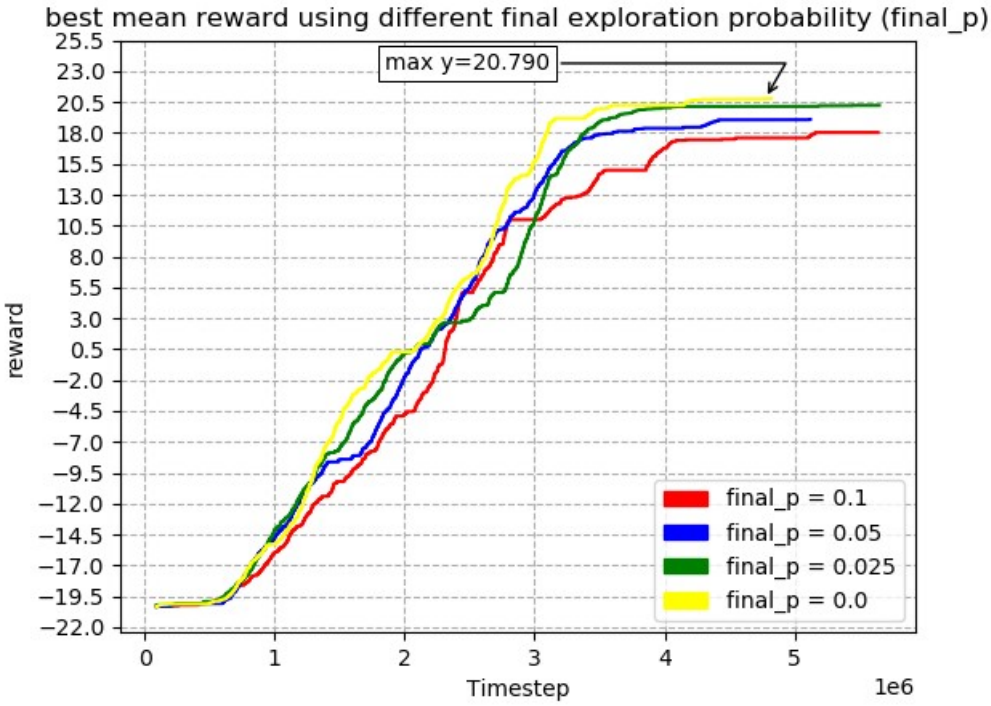
See the two plots below where the first one shows the mean reward of four experiments (three new ones plus the one used in Question 1) and the second shows the best mean reward of the same four experiments.

I have experimented with the exploration schedule scheme, in fact, with exactly a single parameter – the final exploration probability (*final_p*) in the linear interpolation scheme. Just to recall epsilon (*eps*), the probability

to select a random action *a*, is equal to *fraction * final_p + (1 – fraction) * initilal_p* where *fraction = t / schedule_timepoints*. In the default setting *schedule_timepoints = 1,000,000, final_p = 0.1* and *initial_p = 1.0*. This means that we start with *eps = 1.0*, namely, fully randomized exploration, and as we are making more and more steps we are decreasing gradually the epsilon such that at million steps it reaches 0.1, and from this point it stays fixed to 0.1. In terms of expolarion and exploitation – as we play more and more games, the time we spend on exploration decreases and the time we spend on exploitation increases. It is a good stategy since as the agent is experiencing with more and more games, as it is approaching to the goal, the estimated Q function is becoming more and more accurate, and therefore we can feel more and more confident on it. Now, if we go back to graph of Question 1, one can easily observe that at around 4 million steps the implementation reached its highest score (which is less than the maximum possible score) and from this point it stays fixed to that score. Since it is evident from the graph (of Question 1) that gradually decreasing epsilon improves the reward, one can hope that by decreasing the epsilon, even to a lower value than 0.1, the extra improvement to the reward would be gained, and, indeed, it this is what happened.

As the two graphs below show, results are improved when *final_p* is decreased. The graphs plot three additional experiments with different values of *final_p*: 0.05, 0.025 and 0. Highest score of 20.79 achieved when *final_ p = 0*. To summarize, best setting is gained when we start with fully exploration, gradually changing relative proportion between exploration and exploitation until exploitation becomes *fully dominant* after million steps.


mean reward using different final exploration probabilty (final_p)
max y=20.790
reward
Timestep
1e6
final_p = 0.1
final_p = 0.05
final_p = 0.025
final_p = 0.0
25.5
23.0
20.5
18.0
15.5
13.0
10.5
8.0
5.5
3.0
0.5
−2.0
−4.5
−7.0
−9.5
−12.0
−14.5
−17.0
−19.5
−22.0
0
1
2
3
4
5


best mean reward using different final exploration probability (final_p)
max y=20.790
reward
Timestep
1e6
final_p = 0.1
final_p = 0.05
final_p = 0.025
final_p = 0.0
25.5
23.0
20.5
18.0
15.5
13.0
10.5
8.0
5.5
3.0
0.5
−2.0
−4.5
−7.0
−9.5
−12.0
−14.5
−17.0
−19.5
−22.0
0
1
2
3
4
5