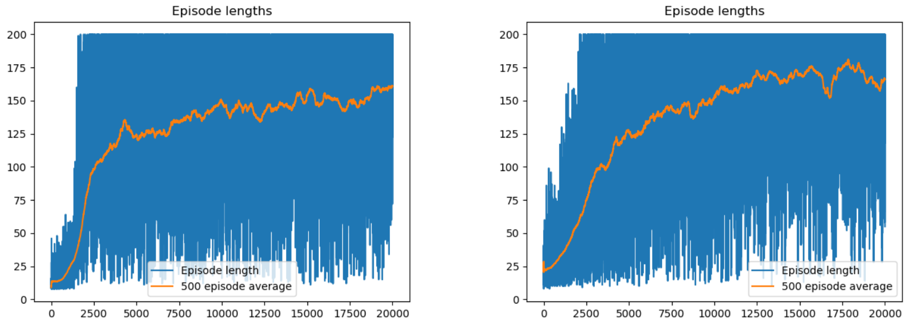
RL Homework 3

Ananth Mahadevan

October 27, 2019

Question 1

The heatmap training times for $\epsilon = 0.2$ and GLIE is seen in Figure 1



(a) $\epsilon = 0.2$

(b) GLIE

Figure 1: Training Performance of constant ϵ and GLIE

The heatmap for Task 2 is shown below

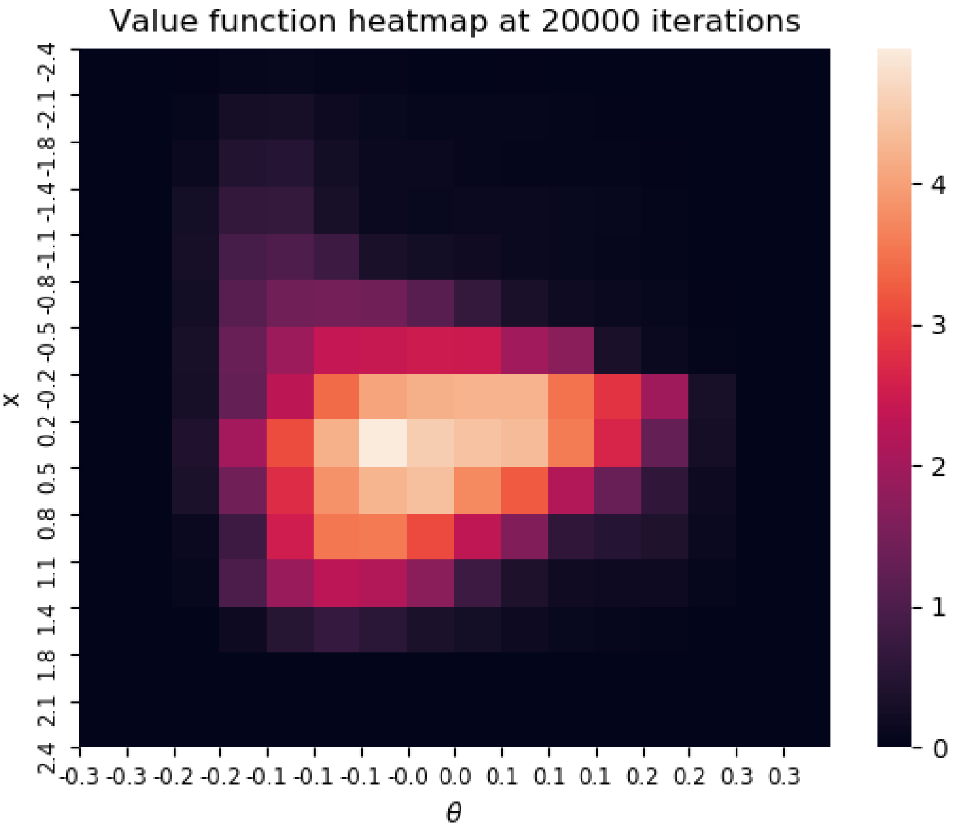


Figure 2: Heatmap of value function

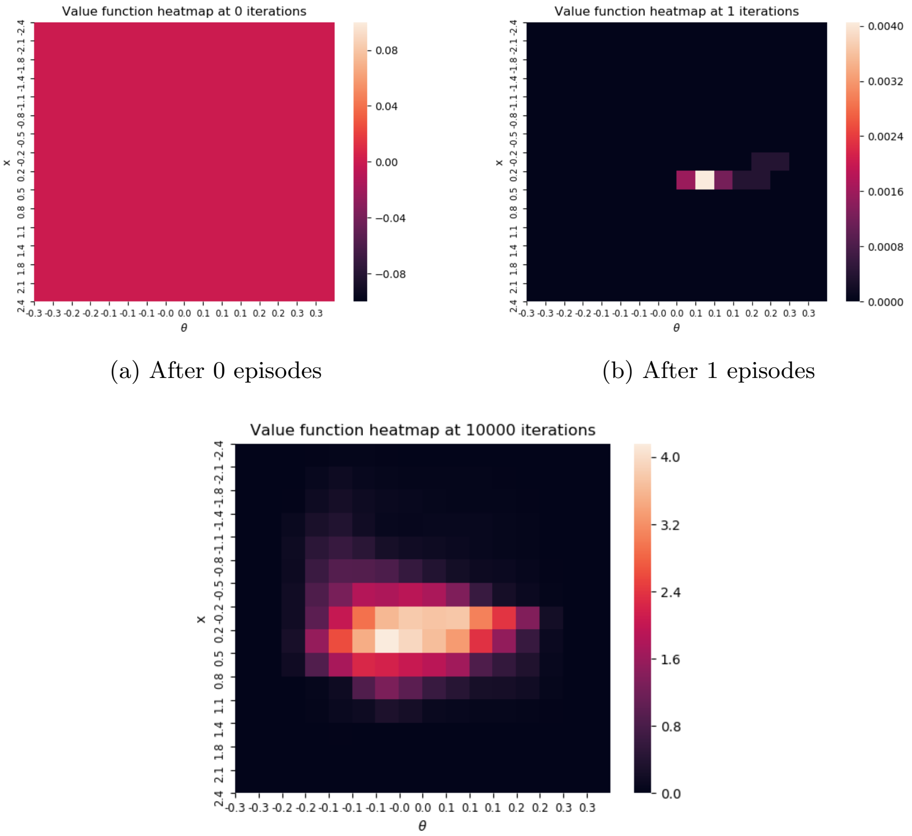
My hypothesis for what the heatmap of the value function would have looked like in the various cases are

- **before training**
  Before training the heatmap would have be all uniformly zero as the value of the Q learning gird would be initialized to zero by before training and no value has been learnt yet.

- **after a single episode**
After a single episode I guess it would update states very close to the center ($x = 0$) and small angles. This change in the value will also be very small. This is because the first episode is most likely to fail and hence only the handful of states in which it managed to keep the pole upright would get a reward to have changed the Q learning grid's initial values of 0. And as the rewards would likely be low the change from 0 is also small.

- **Halfway through training**
By this point I would expect the heatmap to more or less converge to roughly how the final heatmap would look. By this I mean that the exploration has been carried out until halfway through, would cover most of the states around the middle of the environment ($x = 0$ and $\theta = 0$). This is because it has explored most of the state space and as the value of ϵ is decreasing further the values will not change too much after the halfway point the amount of exploration will decrease and the values will just converge to a better level till the end of the episodes. This is also supported by the Figure 1b the average episode length is similar

All of these claims can be verified by the plots of the heatmap as seen in



(a) After 0 episodes

(b) After 1 episodes

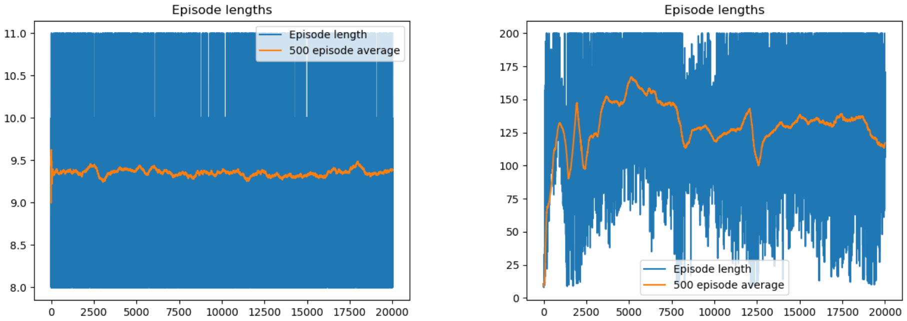
(c) After 10,00 episodes (halfway)

Figure 3: Heatmap of Value function after different number of episodes

Question 2.1

The training graphs for the two cases are seen Figure 4

Based on these plots we can see that the case when `q_init` = 50, the model is able to stay alive longer and also has much better rewards than the model when `q_init`=0.



(a) Initial Q value = 0

(b) Initial Q value = 50

Figure 4: Training Performance of $\epsilon = 0$ and different initial Q values

Question 2.2

When the initial value of the Q learning grid is 0 and the ϵ is also 0, this means that the policy will just choose to be greedy. But because the state space hasn't been explored at all the exploitation of the rewards is also the little it gets. In the first couple iterations it would stay alive for a very short period of time as it would always take the first action as all actions have 0 value. This in turn only provides very small reward for those state actions it managed to stay alive in. As now these are the non-zero actions, it will choose to exploit this constantly for the rest of the episodes. Hence it can only gain the rough same amount of reward as it refuses to explore other states that still 0 but might be better strategies for the overall environment. It is clearly seen in Figure 4. This issue of greedy exploitation is to a degree mitigated when the initial Q learning value is optimistic. This is shown in the next question.

Question 2.3

The reason that the case when the `q_init`=50 there is has **optimistic initial values**. This is explained in S&B book Chapter 2.6, where they mention that most methods have a *bias* on their initial values. This is not an issue with a fixed learning rate α which corrects any issue. Initial bias is not a problem as it can also make the model more optimistic in the start and explore much better. This works because when the initial values are highly optimistic it chooses some action which has a reward of much less than the initial value of 50, hence the update reduces the value of that state-action, now it can choose the other actions that are still 50 till it finds out that those are also not optimal. This is very nicely described in the book as the learner becoming disappointed with the optimistic states and tries many other actions before they converge. This is then in a way inducing the exploration that would have happened in an ϵ-greedy policy. This clearly only happens when the initial Q learning values are ridiculously optimistic which would lead it to be disappointed and then explore to converge. This is the reason why the exploration is nearly non-existant when initial Q learning value is 0, the model performs poorly.

Question 3.1

The performance for the Lunar Lander environment is much better measured as the cumulative reward at the end of the episode rather than length of the episode as it was done in the Cartpole environment. This is seen in Figure 5 This clearly shows that the model is unable to learn anything as it is unable to get above 0 in average. It seems to land very rarely while testing. It seems to be able to navigate in the environment properly but is unable to achieve the goal of landing.

Question 3.2

The lunar lander environment is more complex than the cartpole especially due to a couple of reasons.The sheer size of the discretized state space might mean that more exploration might be necessary for the model to learn a better policy that would help it land. Hence the discrete state space that ranges in

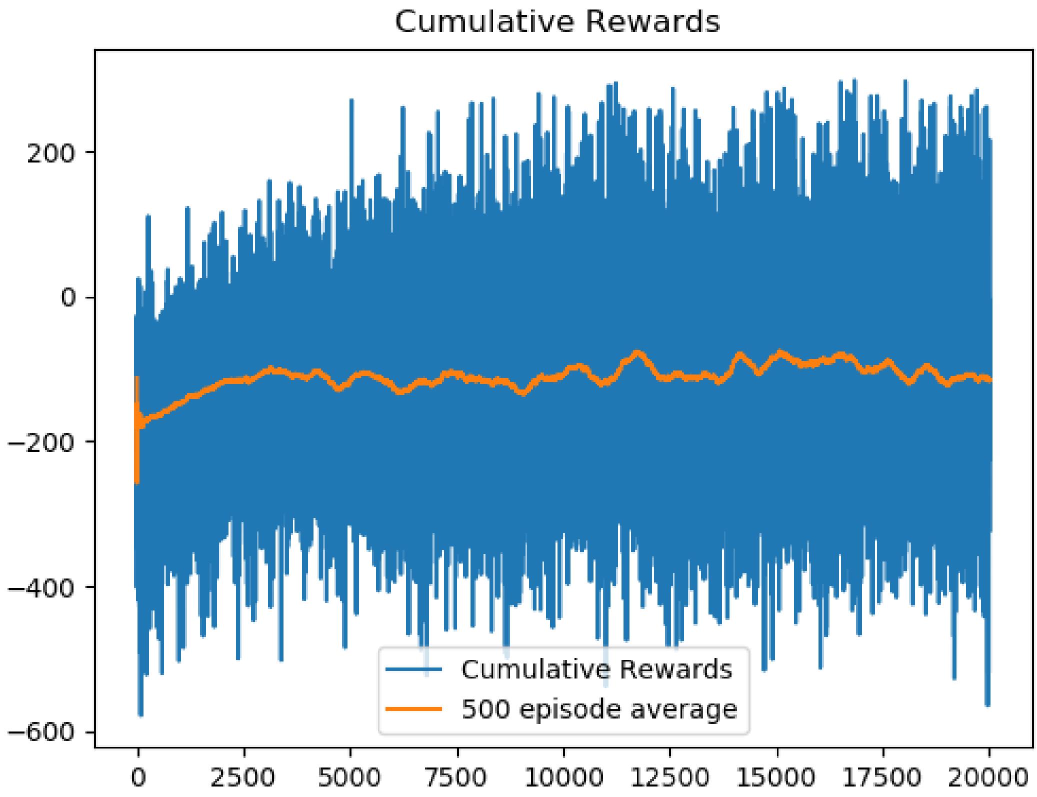


Figure 5: Training performance of cumulative reward for Lunar Lander

the millions might be unable to converge effectively in any time frame. This might be strategies like not doing anything when it's above the landing site and then firing the thrusters really fast to land softly. The rewards structure for the lunar lander is also more delayed as landing is the last time step in an episode and that garners the most positive reward and also most of the time, firing the main thrusters the model accrues more negative rewards, hence the action space is filled with most actions that lead to negative rewards. This might just mean that the environment might need to train for much longer for the rewards from successfully landing to also update the state action values which are more in the beginning of the episode.