RL Homework 3

Ananth Mahadevan

October 23, 2019

Contents

1 Question 1

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

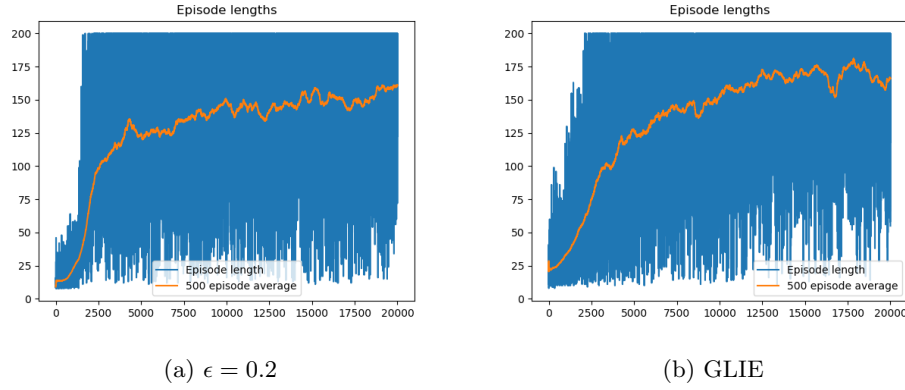(a) $\epsilon = 0.2$ (b) GLIE

Figure 1: Training Performance of constant ϵ and GLIE
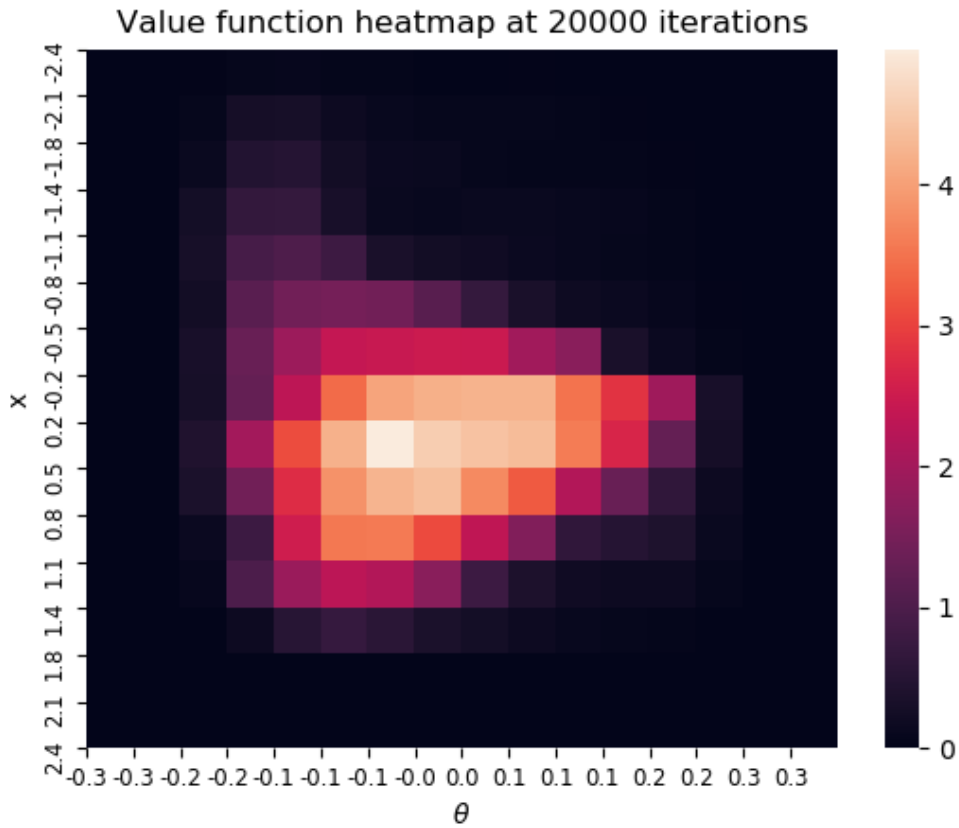
The heatmap for Task 2 is shown below

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 been randomly close to zero as the value of the Q learning gird would be initialized to zero by a random function.

3

- **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.

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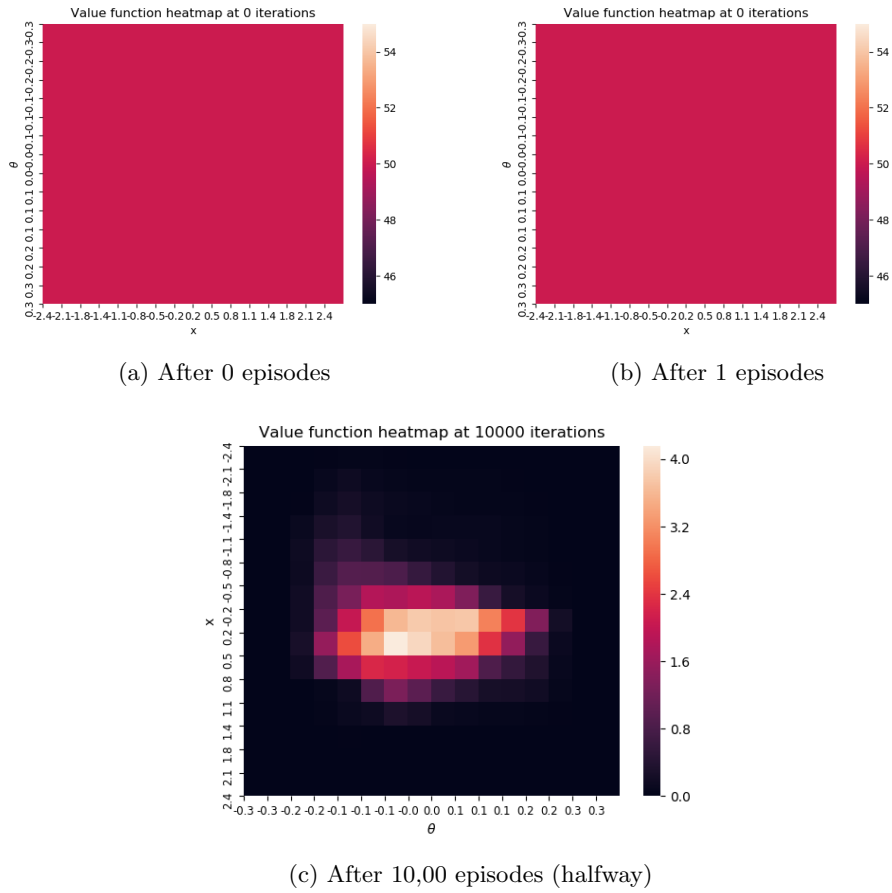

(a) After 0 episodes

(b) After 1 episodes

(c) After 10,00 episodes (halfway)

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