Exercise 5

[Avik Banerjee (3374885), Soumyadeep Bhattacharjee (3375428)]

Text in italics are notes taken during the exercise

1 Random Walk

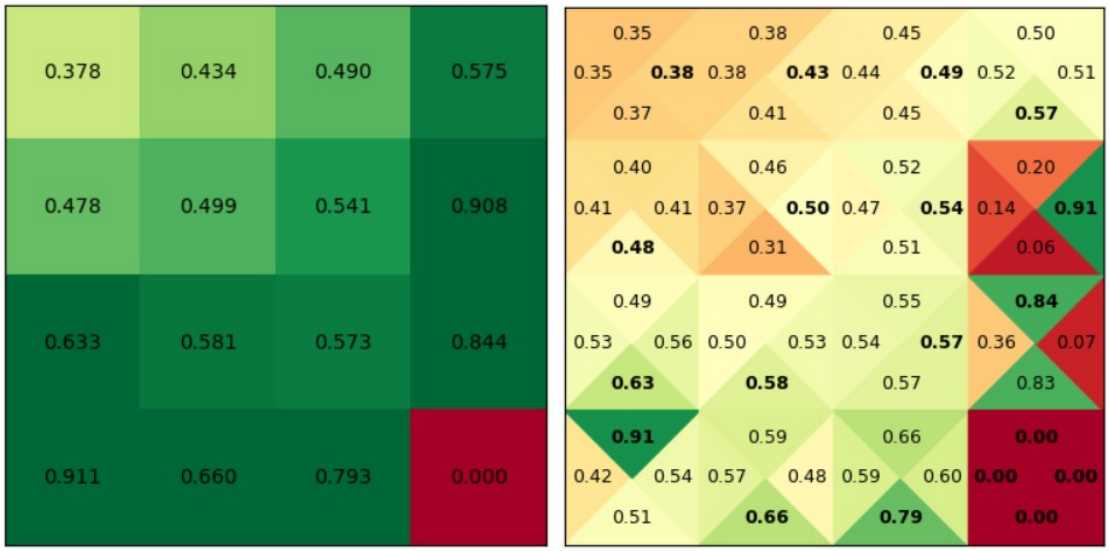
The value of each state is initialized to 0.5 before the first episode starts, while the final states are set to 0. In the first epoch, as we proceed from $C \rightarrow A$, $V(S_t)$ results in 0 for states C and B, as both the previous and current states have $V(S) = 0.5$. However, when transitioning from A to the leftmost state, the value functions for the consecutive states are not the same.

$$V(S_t) \leftarrow V(S_t) + \alpha[R_{t+1} + \gamma V(S_{t+1}) - V(S_t)]$$
$$\text{where } V(S_t) = 0.5, V(S_{t+1}) = 0$$

Plugging in all the remaining values, we get $V(S_t) = 0.45$, whereby the value changes by 0.05.

2 Sarsa and Q-learning on the FrozenLake

a) The state-value function, action-value function for Sarsa are as follows:

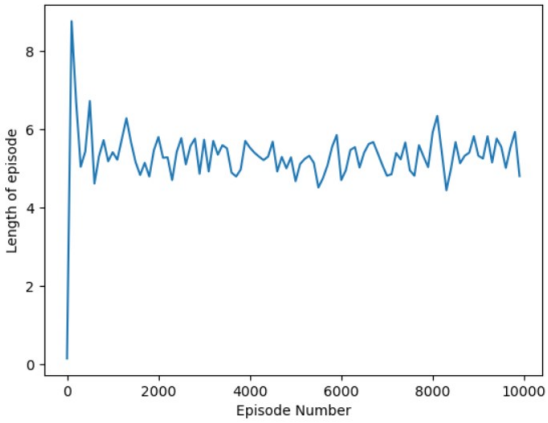


(a) State-value function

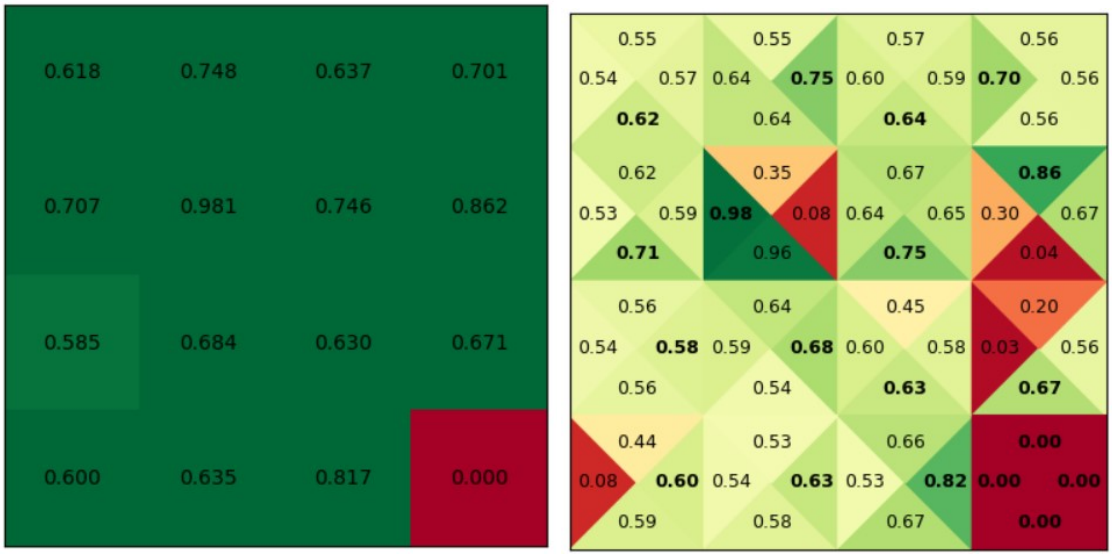
(b) Action-value function

The policy:

The average episode length over every 100 episodes:



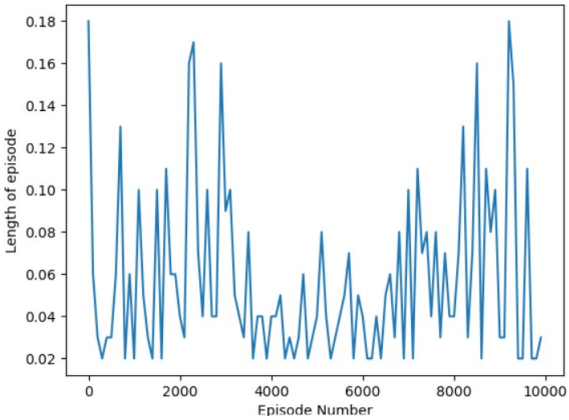
b) The state-value function, action-value function for Q-learning are as follows:



(a) State-value function (b) Action-value function

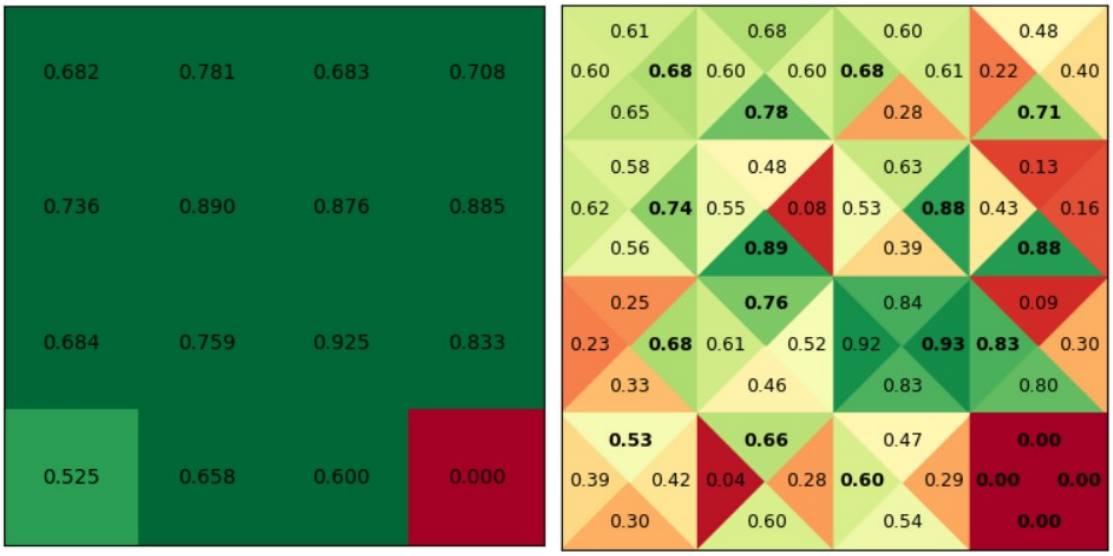
The policy:

The average episode length over every 100 episodes:



As is evident from the episode-length graphs, the optimal policy obtained at the end of 10000 iterations may not have the lowest episode length, there are multiple minima that are reached before all episodes are covered. *Sarsa actually converges to an optimal version of the epsilon-greedy policy since it is on-policy, whereas Q-learning, being on-policy converges to the actual optimal policy.*

c) The non-slippery version makes the episode lengths converge faster, which means the optimal policy is derived faster, which is prominently visible in the episode-length plot of Q-learning.
The state-value function, action-value function for Sarsa are as follows:

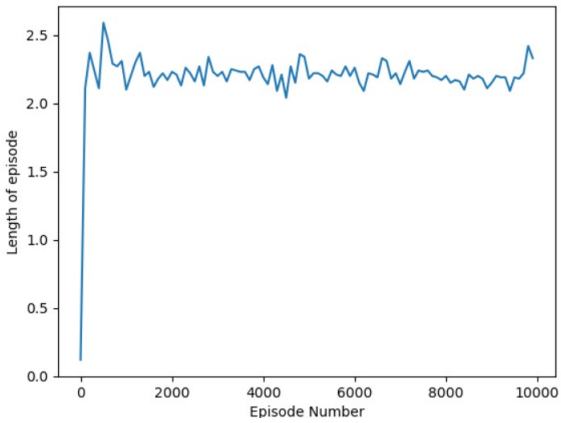


(a) State-value function

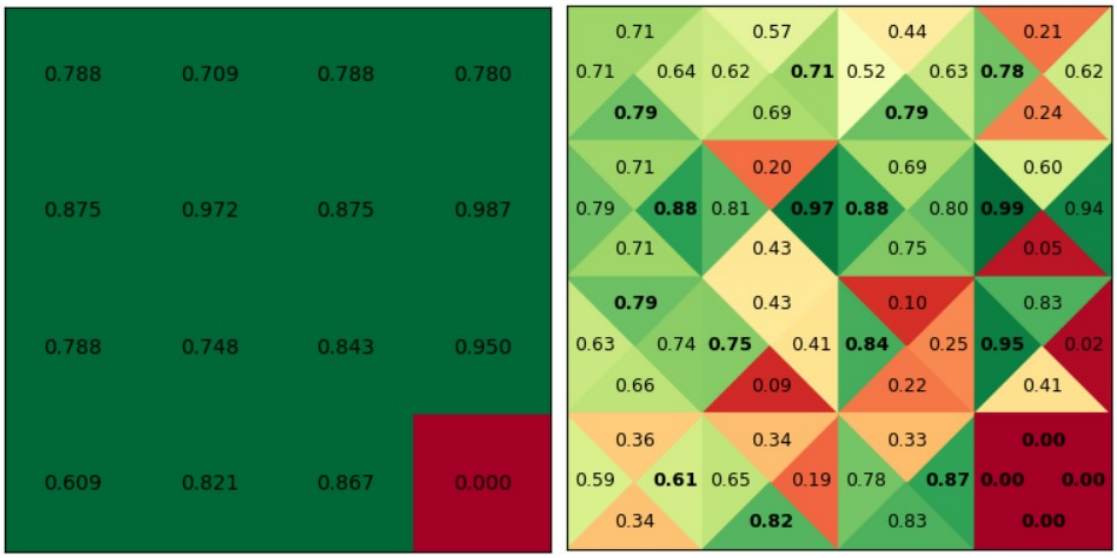
(b) Action-value function

The policy for Sarsa:

The average episode length over every 100 episodes:



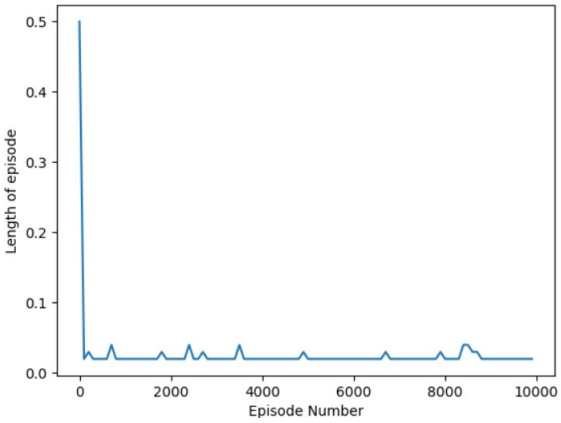
The state-value function, action-value function for Q-learning are as follows:



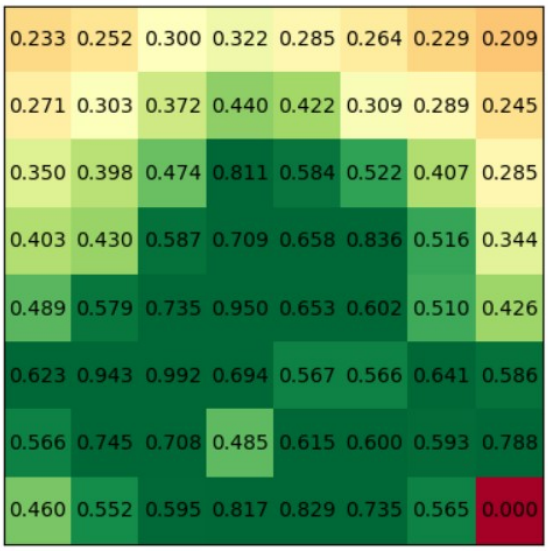
(a) State-value function (b) Action-value function

The policy:

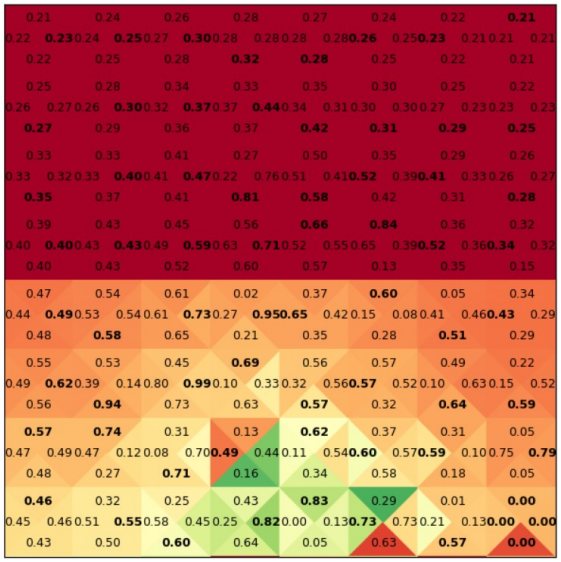
The average episode length over every 100 episodes:



d) The state-value function, action-value function for Sarsa are as follows:

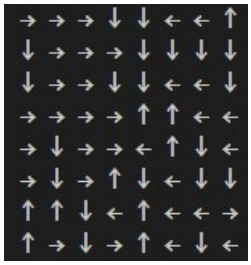


(a) State-value function

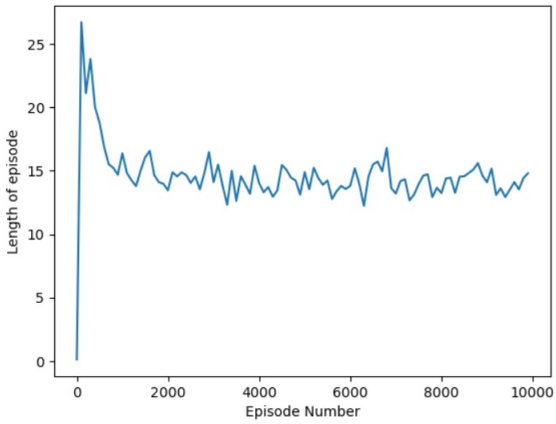


(b) Action-value function

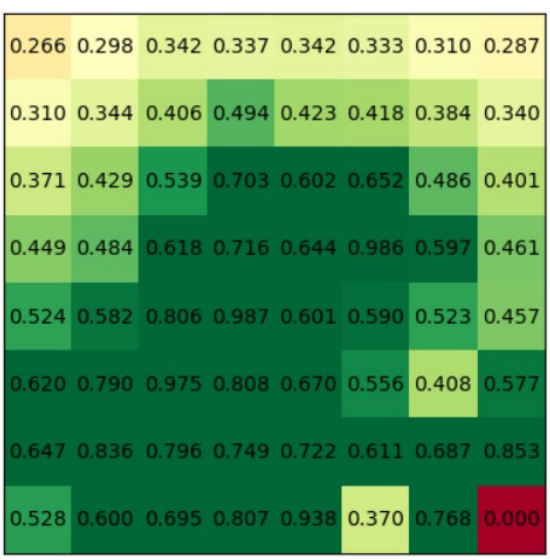
The policy for Sarsa:



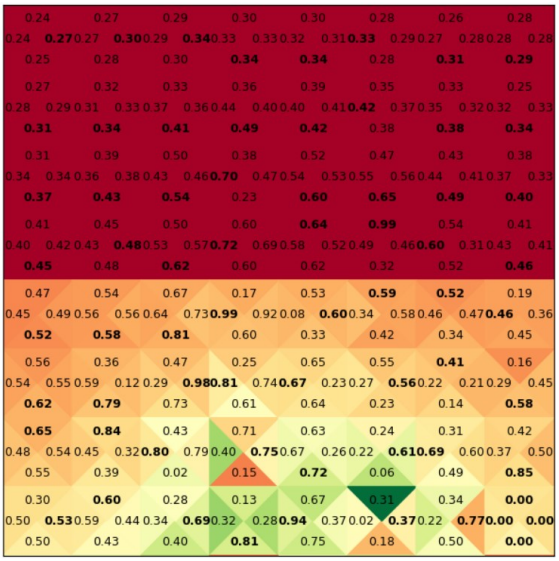
The average episode length over every 100 episodes:



The state-value function, action-value function for Q-learning are as follows:

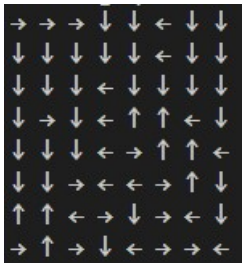


(a) State-value function



(b) Action-value function

The policy:



The average episode length over every 100 episodes: