COMP3702 Artificial Intelligence (Semester 2, 2022)
Assignment 3: HEXBOT Reinforcement Learning

Name: Alex Viller

Student ID: 45375325

Student email: a.viller@uqconnect.edu.au

Note: Please edit the name, student ID number and student email to reflect your identity and **do not modify the design or the layout in the assignment template**, including changing the paging.

Question 1 (Complete your full answer to Question 1 on the remainder of page 1)

a)

i. Q-learning and value iteration both work by having a table of values which contains the best action to do based on the current state.

ii. Q-learning and value iteration both use a discount factor ɣ which tell the agent how much to consider the previous states reward.

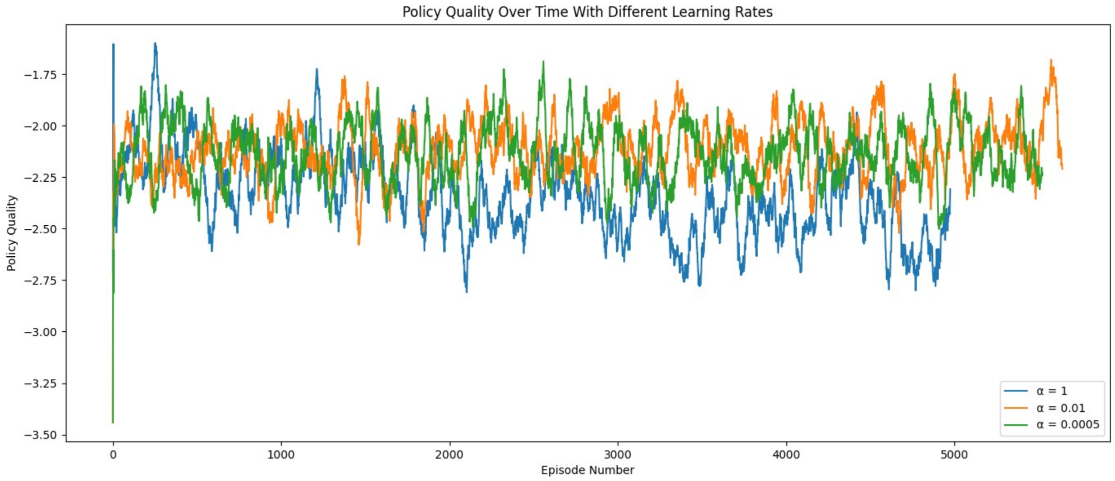
b) q-learning does not know the rewards or probabilities ahead of time. It has to learn these throughout the training process to build up the q-table.

Question 2 (Complete your full answer to Question 2 on page 2)

a) Off policy means we are learning the value of the optimal policy no matter what it does. Leading to dangerous paths being taken which could result in a much lower overall reward. On policy however learns the value of the policy being followed, optimising the exploration as well. This difference can be found in the code in lines 205 and 206 which show selecting a next action to take before the frame is finished. Allowing for an optimised exploration path.

b) Q-learning does not solve for my case but I would expect it to go along the dangerous lower path as that is the most optimal policy. SARSA in fact follows this dangerous path which is unexpected due to the theory of SARSA I would expect the agent to take the safer top path. However there might be some difference in time taken to learn not allowing for the agent to see this path as valid. We can also see that the hazard penalty is less than obstacle penalty. This actually causes what appears to be the dangerous path to be the safe path.

Question 3 (Complete your full answer to Question 3 on page 3)



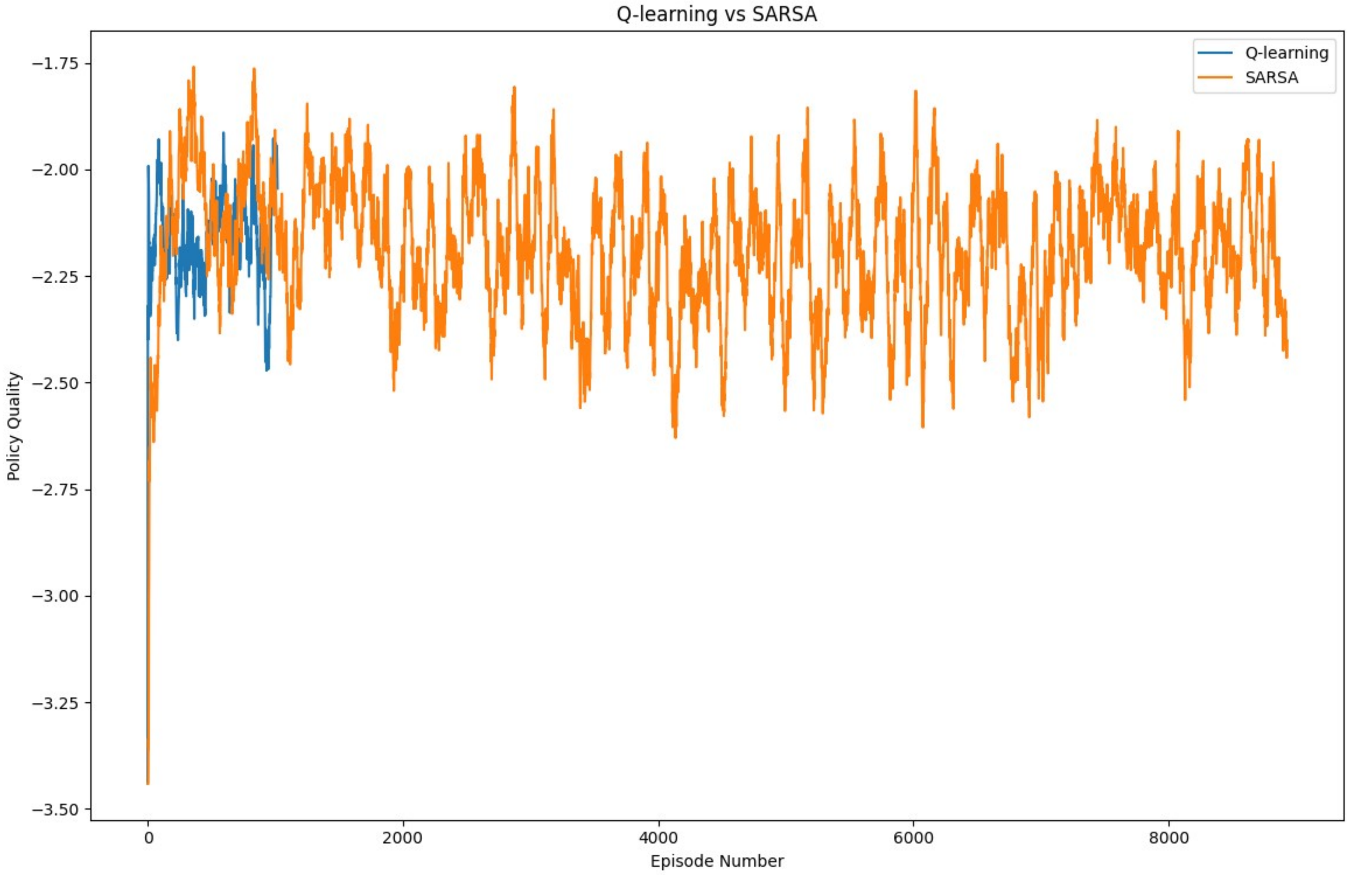
a) see above

b) Varying the learning rate changes how big of a step we take in the direction of minimal loss. If this step is too small we may never reach any kind of minima or get stuck in a local minima. If this step is too large it may be impossible to ever reach a minima of any kind as we will always over shoot. Thus with the perfect medium we can find the global minimum and this the best possible policy. To clarify by minimum I am meaning closest to 0 in our case.

We can see that with the learning rate of 1 we have the lowest quality of policy and the function actually stopped early after having too high of a cost for too long. Next best we see the very small learning rate of 0.0005 but we see that we have gotten stuck in a local minima which clearly isn’t the highest quality policy. Then finally we see the learning rate of 0.01 between these two extremes lasts the longest before being cancelled by having too high of a cost and produces the highest quality of result.

These values can be further trained to produce better and better results and they can also be made to change value as you go to avoid getting stuck in local minima and then to avoid overshooting the global minima by accident.

Question 4 (Complete your full answer to Question 4 on page 4)



a) see above

b) We can see that q-learning doesn’t train for as many episodes, despite having the same number of max frames. This is due to sarsa finding exit conditions much faster and more often. This allows for sarsa to further optimise the path it takes to get a better policy. Q-learning however has very little time to find it’s optimised policy and so instead just ends up taking the first policy it finds that does decently well. Also important to note is that q-learning is not able to solve the environment in this case.