16-662 Robot Autonomy Homework 4: Austin Windham

1. Questions

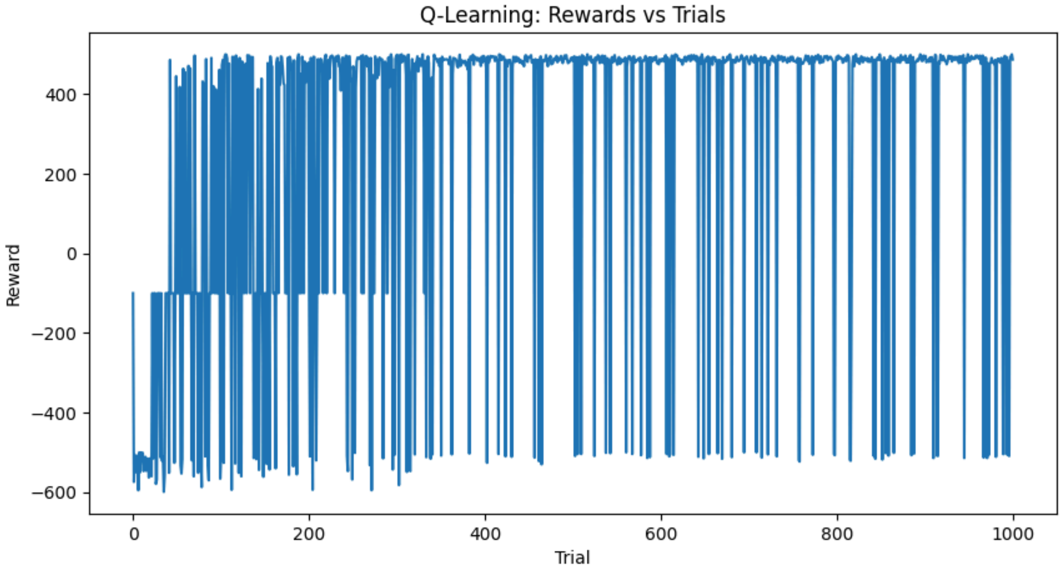
A. The rewards randomly dip to a low number during training because the Q-learning algorithm explores different actions, including suboptimal ones in order to learn the optimal policy. This exploration can lead the agent to take actions that result in lower rewards temporarily before learning the optimal strategy. The rewards also randomly dip sometimes because of the rho value that affects the probability that the agent is successfully taking the action it chooses. Sometimes the agent is trying to take the action with the highest reward, but because we have a rho value, the agent sometimes moves in a different direction.

B. Yes, the arrows plotted in q_values.png make sense. They represent the best action to take in each state according to the learned Q-values. Each arrow points in the direction of the action with the highest Q-value in that state, which gives a visualization of the optimal policy learned by the agent.

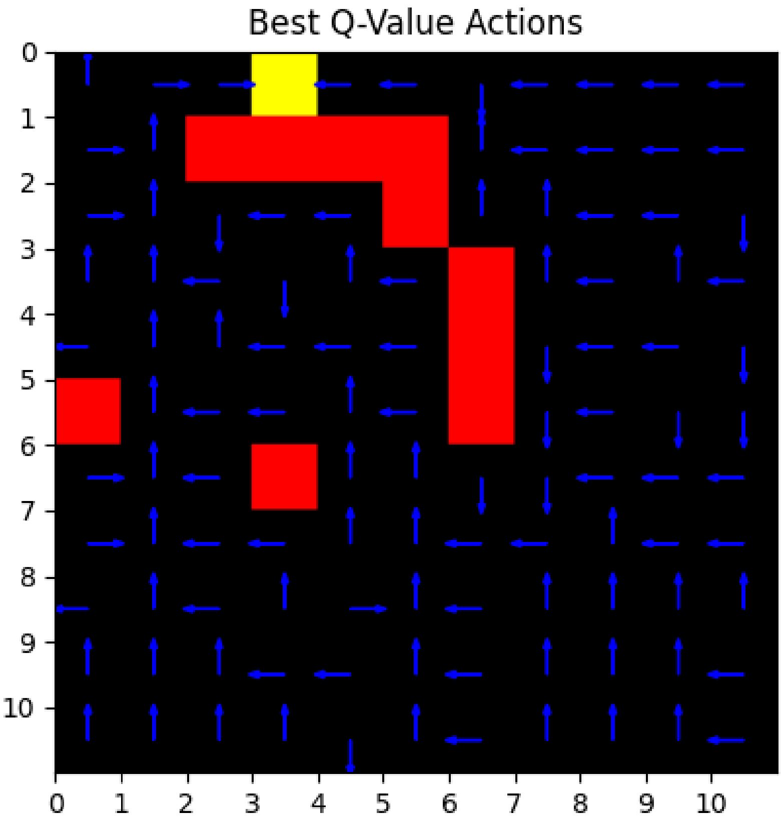
C.

No, the Q Agent does not always follow the optimal path with respect to Q values in the generated videos. Sometimes, it might deviate from the optimal path due to the exploration-exploitation trade-off. The agent may choose to explore suboptimal actions with a certain probability based on the epsilon value instead of always following the action with the highest Q-value. Additionally, stochasticity in the environment or imperfect learning can also lead to deviations from the optimal path. Additionally because of the rho value, sometimes the agent is not performing the action that it is trying to. The algorithm may have also not trained long enough yet to find the best policy.

2. The q_learning_training.png generated by the script is below.



3. The q_values.png generated by the script is below.



4. The average reward for the random agent is -432.392, and the average reward for the Q agent is 481.206, so the Q learning agent was much more effective.

5. Role of hyperparameters.

A. Q_ALPHA controls the learning rate of the algorithm. Increasing the value increases the speed of learning and makes the agent more sensitive to new information, but it also makes it more prone to fluctuations in the learning process.

B. Q_GAMMA is the discount factor in the algorithm. Increasing this value increases the importance of future rewards, which makes the algorithm place a priority on long-term rewards. Conversely, decreasing the value does the opposite and places more importance on immediate rewards.

C. Q-EPSILON is the value that sets the probability for the exploration-exploitation process. Increasing this value makes the algorithm more likely to explore new actions. This can help find a better path, but it can also take longer time for the algorithm to train. Conversely, decreasing this value will make the algorithm less likely to explore new actions, causing it to possibly miss the optimal path, but it also speeds up the training times.

D. ENV_RHO is the probability of performing a random action instead of the action with the highest Q-value. Increasing this value makes the agent more likely to perform a random action, which can cause it to not follow the optimal path. Increasing this value will also probably increase the learning time. Decreasing the value will do the opposite; it will make the agent more likely to perform the action with the highest Q-value. This will decrease inaccurate movements and cause the model to converge faster. The algorithm also does not have any feedback, so when it moves incorrectly, it returns the reward for the action that it wanted to perform, and not the one it performed. This can cause errors when trying to develop the optimal policy because the agent is not learning the corresponding reward to the correct action.