**Question 1.**

Q-learning is closely related to the Value Iteration algorithm for Markov decision processes.

**a)** (5 marks) Describe two key similarities between Q-learning and Value Iteration.

In both Q-learning and Value iteration there is a table containing a value for all states, that shows the agent how preferably it is to be in that state.

Another similarity between the two algorithms, is that they both try to compute the values for each state. When the values for all states are computed and converged, the agent knows the optimal action from every state.

**b)** (5 marks) Give one key difference between Q-learning and Value Iteration.

One key difference between Q-learning and Value Iteration is that in Value Iteration the agent knows the transition probabilities and rewards. In Q-learning the agent needs to do exploration to learn these rewards and probabilities.

For Questions 2, 3 and 4, consider test case ex3.txt for Q-learning and the equivalent SARSA test case, ex4.txt.

**Question 2.**

**a)** (5 marks) Explain the difference between off-policy and on-policy reinforcement learning algorithms. Explain where this difference is represented in the Q-learning and SARSA algorithms.

An off-policy reinforcement algorithm learns the value of all policies, not matter what this policy is. On-policy reinforcement learning learns the value of the policy that has the best value at that moment. The off-policy algorithms try to learn the value through exploration, even though the policy being explored can end up in a unpreparable state.

In Q-learning, when updating the Q-tables, a next state is chosen by an exploration strategy. However, in SARSA after an action is selected with an exploration strategy, this action is saved and will be used in the next exploration phase.

**b)** (5 marks) How does the difference between off-policy and on-policy algorithms affect the way in which Q-learning and SARSA solve test cases ex3.txt and ex4.txt? If you were unable to solve these test cases, it is sufficient to answer with reference to what you think would happen, based on the set up described in the test case files.

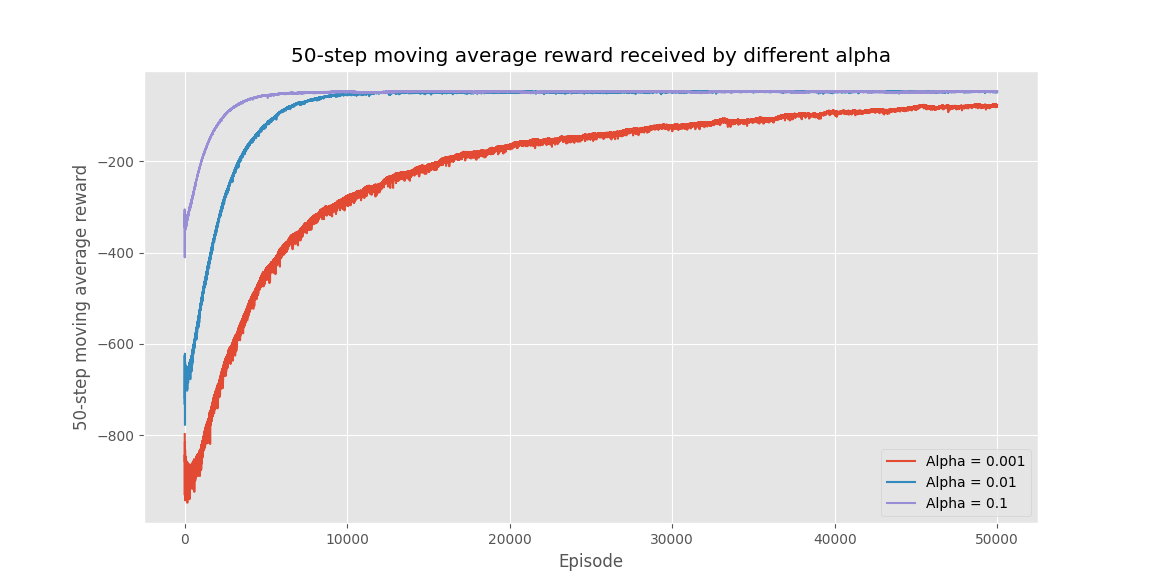
Both Q-learning and SARSA choose the shorter, but riskier path close to the hazards. I think it make sense that Q-learning choses the riskier path, because Q-learning only choses one action at a time. However, I would think that SARSA had chosen the less risky, but longer way around the obstacles. This is because SARSA is on-policy learning, and I would think that the first action would be to go the less risky path around the obstacles.

For questions 3 and 4, you are asked to plot the solution quality at each episode, as given by the 50-step moving average reward received by your learning agent. At time step t, the 50-step moving average reward is the average reward earned by your learning agent in the episodes [t − 50, t], including episode restarts. If the Q-values imply a poor-quality policy, this value will be low. If the Q-values correspond to a high-value policy, the 50-step moving average reward will be higher. We are using a moving average here because the reward is received only occasionally and there are sources of randomness in the transitions and the exploration strategy.

**Question 3.**

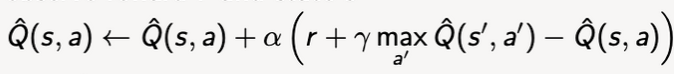
**a)**

(5 marks) Plot (all on a single plot) the quality of the policy learned by Q-learning in testcase ex3.txt against episode number for three different fixed values of the learning\_rate (which is called α in the lecture notes and in many texts and online tutorials), as given by the 50-step moving average reward (i.e. for this question, do not adjust α over time, rather keep it the same value throughout the learning process). Your plot should display the solution quality up to an episode count where the performance stabilises. This may take a significant number of episodes (e.g. >50,000) depending on the learning rates used. Note that the policy quality may still be noisy, but the algorithm’s performance will stop increasing and its average quality will level out. Your plot should include axis labels and a legend.

****

**b)** (5 marks) Discuss the effect of varying the learning\_rate. You should make reference to the Q- learning algorithm to support your discussion. If you were able to generate a plot in Q3a, you may also make reference to this in your discussion.

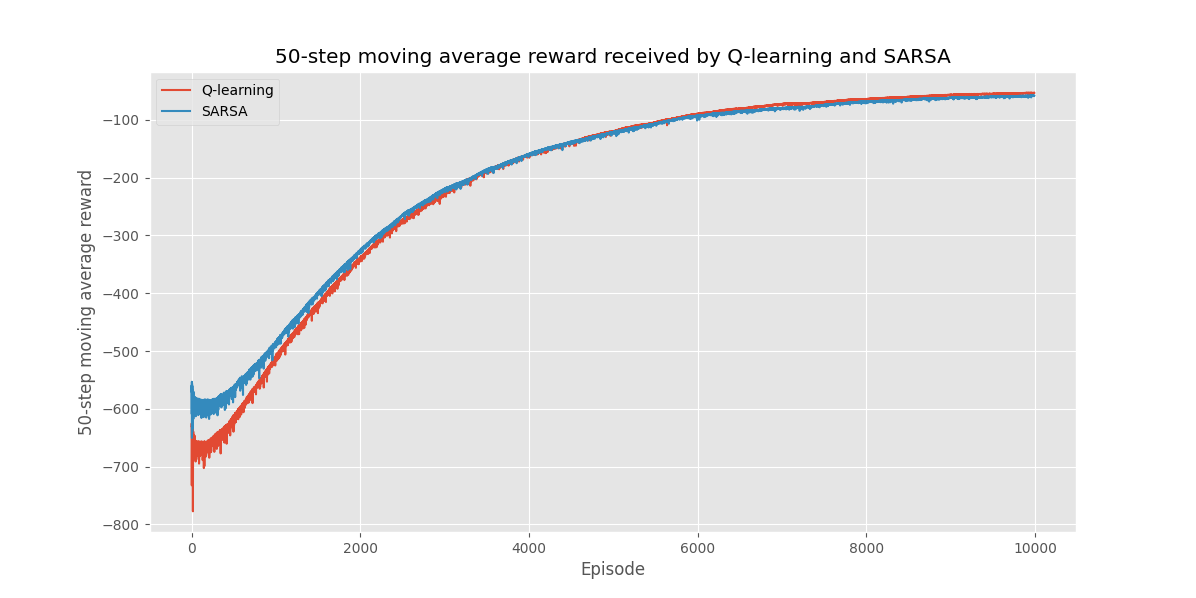
The difference in learning rate can be seen in the three different plots in Q3a. The Q-learning algorithm updates the Q-values by some learning factor alpha. If this learning factor is high, the Q-values will be updated faster by the new values seen in the parenthesis under. When alpha is low, the updated should be much slower. As you can see from the three plots over, when the learning rate is 0.1 the episode rewards converges after around 500 0 episodes. However when the learning rate is 0.01 and 0.001 respectivly, the episode rewards converges after 9000 and 60 000 episodes.



**Question 4.**

**a)**

(5 marks) Plot (on a single plot) the quality of the learned policy against episode number under Q- learning and SARSA in test cases ex3.txt and ex4.txt respectively, as given by the 50-step moving average reward. Your plot should display the solution quality up to an episode count where the performance of both algorithms stabilise. Your plot should include axis labels and a legend.



**b)**

(5 marks) With reference to your plot, compare the learning trajectory of the two algorithms, and their final solution quality. Discuss the way the solution quality of Q-learning and SARSA change as they learn to solve the testcase, both as they learn and once they have stabilised.

As we can see, in the first 3000 episodes approximately SARSA has a better reward than Q-learning. This is expected, since SARSA learn the values of the current policy that has the best value. However, after around 3000 episodes, Q-learning starts to get the same rewards as SARSA. This is due to the fact SARSA keeps learning the value of the best policy, while Q-learning also learns the value of other policies. This is decided by either UCB or epsilon-greedy. In the plot over, epsilon greedy is used. From around 6000 episodes can be see that Q-learning has passed SARSA in average reward, and from around 8000 episodes, they have both stabilized. Q-learning continues to have a slightly better reward than SARSA. This can be explained by the same reason why Q-learning after around 3000 episodes almost has the same reward as SARSA.