RBE595 - Week 7 Assignment

Keith Chester

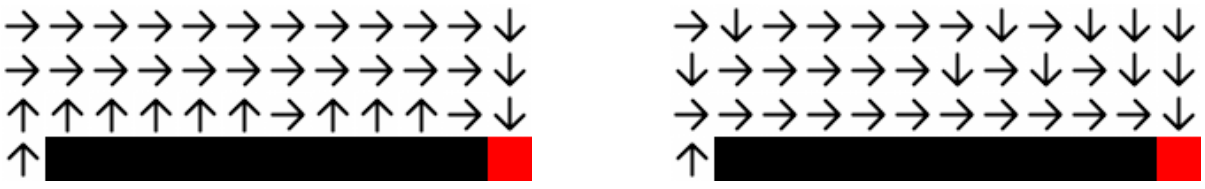
Due date: February 26, 2023

In this assignment, we aimed to create a SARSA and a Q-Learning agent for a 2-dimensional grid world. In this world, there exists terminal states - a goal position, and "cliffs". Falling off of a cliff returns a reward of -100, and every transition brings in a reward of -1.

Attached to this document in Figure 1 is the SARSA and Q-Learning outcomes with $\alpha = 0.1, \epsilon = 0.05$ and $10,000$ episodes. Here we see that the SARSA prefers either the middle or top route, whereas Q-Learning chooses the optimal route along the cliff. This is because SARSA utilizes epsilon-greedy action selection when calculating possible next-step rewards, which results in the next-step value occasionally being the large -100 reward. This means that the perceived value of cells near the cliffs are lower than what they should be. Q-Learning utilizes a straight Q-max selection which avoids this. Figure 2 presents a plot of the rewards after each episode. This plot shows a general convergence, with epsilon greedy creating occasional spikes of -100 rewards to cause the seemingly "wide" drops throughout the chart.

Why is Q-Learning converging to lower rewards despite finding the optimal route quicker? This is because Q-Learning tends to overestimate the values of future actions, whereas SARSA provides more accurate estimates via taking into account the actual actions taken for the next state. This overestimation leads to a higher probability of Q-Learning choosing suboptimal actions, even though Q-values were high. This is what Doulbe Q-Learning attempts to fix with dual value tracking of Q values for a given state/action pair. Q-Learning finds the optimal route faster as it's off-policy versus the on-policy approach SARSA does.

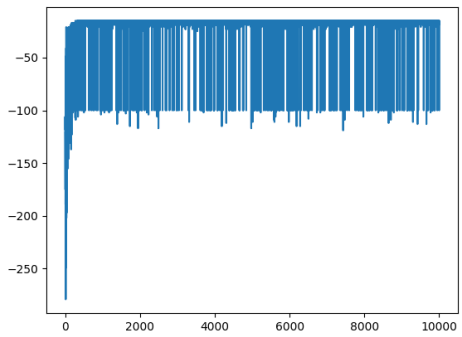
Finally our Figure 3 demonstrates an experiment where we re-run both models, slowly lowering ϵ by increments of 0.01. As ϵ approaches to 0 and our exploration moves closer towards a pure $\max Q(S, a)$ approach, SARSA also finds the optimal path instead of staying wary of cliffs.



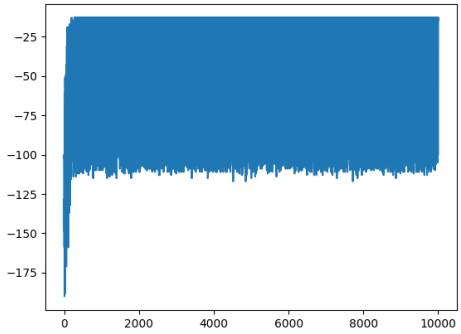
(a) SARSA Agent

(b) Q-Learning Agent

Figure 1: Agent paths

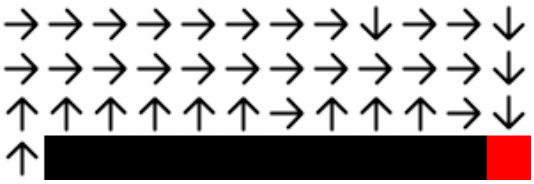


(a) SARSA Agent Rewards

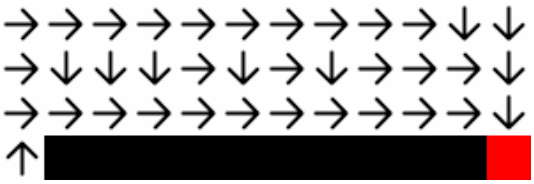


(b) Q-Learning Agent Rewards

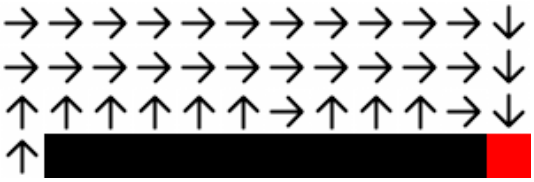
Figure 2: Agent Rewards



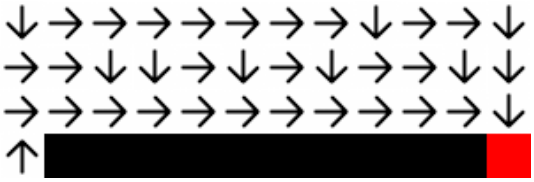
(a) SARSA $\epsilon = 0.05$



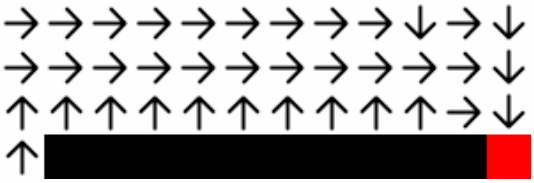
(b) Q-Learning $\epsilon = 0.05$



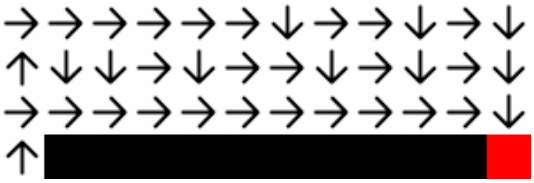
(c) SARSA $\epsilon = 0.04$



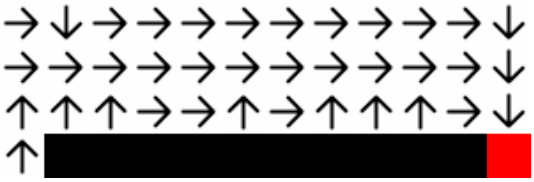
(d) Q-Learning $\epsilon = 0.04$



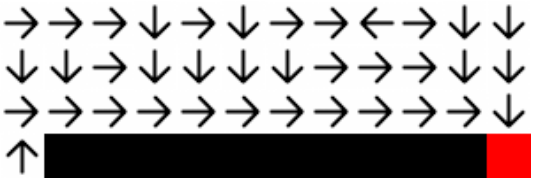
(e) SARSA $\epsilon = 0.03$



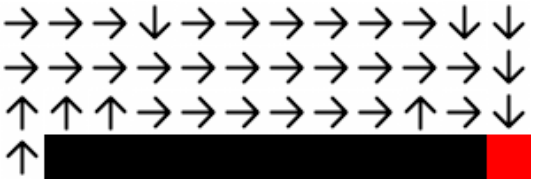
(f) Q-Learning $\epsilon = 0.03$



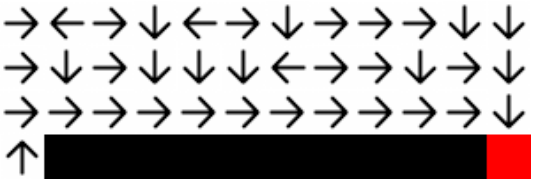
(g) SARSA $\epsilon = 0.02$



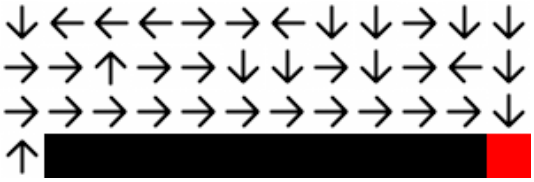
(h) Q-Learning $\epsilon = 0.02$



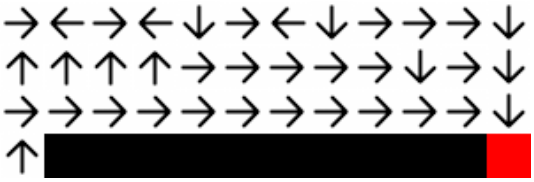
(i) SARSA $\epsilon = 0.01$



(j) Q-Learning $\epsilon = 0.01$



(k) SARSA $\epsilon = 0.00$



(l) Q-Learning $\epsilon = 0.00$

Figure 3: Lowering *epsilon* by 0.01 steps