Reinforcement Learning - Homework 6
Vishal I B, 917809310

The grid world is similar to the one from the previous problem but without the wind and an added cliff. When the next state is on the cliff, the reward is -100 and the state changes to the start state.

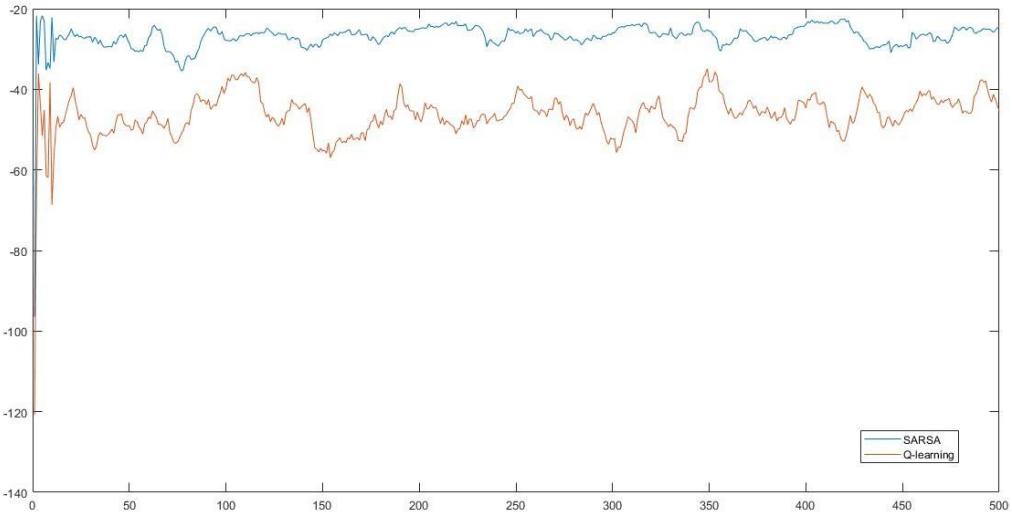
Sarsa is an on-policy TD control algorithm that uses quintuple of events (St, At, Rt+1, St+1, At+1).
The Q(s, a) values are initialized to zero for all values of s, a. Looping for each episode, we initialize the state, and for each step of the episode, we take action and observe the reward and the next state. The next action is chosen by being epsilon-greedy with respect to Q. Q is updated. The state and action are also updated. This continues until the state reaches the final state. Q is initialized as a cell array of vectors for each state, action pair.

Q-Learning is an off-policy TD algorithm with the target policy being greedy and the behavioral policy being e-greedy. So, the Q value is updated using the target policy(greedy).

Expected SARSA is a learning algorithm that is just like Q-learning except that instead of the maximum over next state–action pairs it uses the expected value, taking into account how likely each action is under the current policy. The policy was designed as an e-greedy policy.

The following are the performance graphs that were generated:
We can see that the expected SARSA performs better than SARSA and Q-learning.

Part 1: SARSA vs Q-Learning



Part 2: SARSA vs Q-Learning vs Expected SARSA