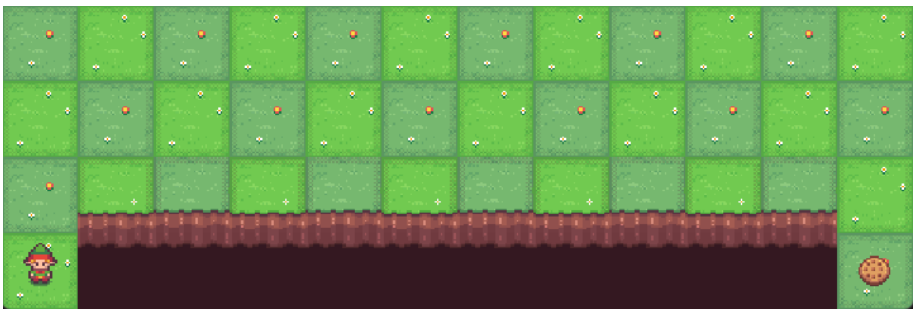
Assignment 6

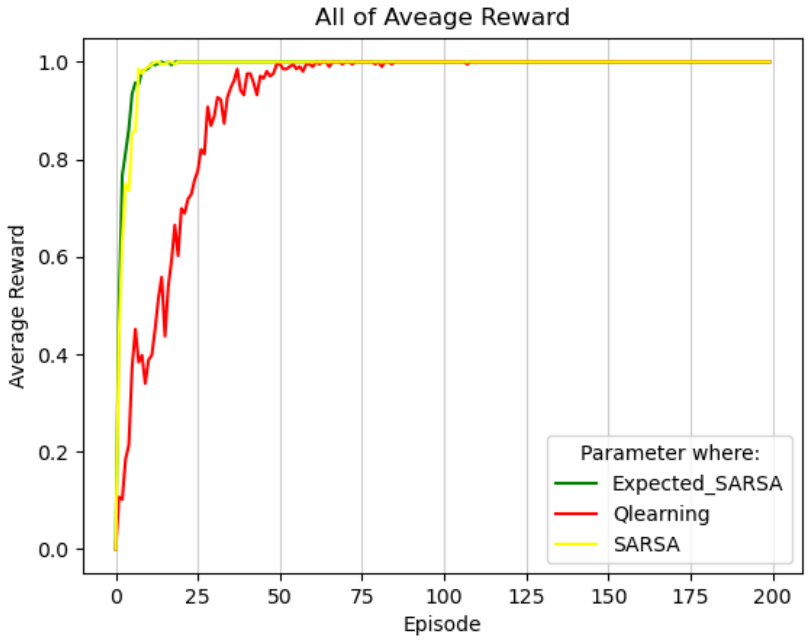
61075029H Jun Yu SHEN 沈峻宇

1. Using OpenAI gym compare the convergence of Q-Learning, SARSA, and ESARSA based on one of the following desired environments.

   - BlackJack
   - Cliff Walking
   - Frozen Lake

   Testing with Cliff Walkin





On-Policy learning algorithms are the algorithms that evaluate and improve the same policy which is being used to select actions. That

means we will try to evaluate and improve the same policy that the agent is already using for action selection. In short , [Target Policy = Behavior Policy]. SARSA is using On-policy.
Off-Policy learning algorithms evaluate and improve a policy that is different from Policy that is used for action selection. In short, [Target Policy != Behavior Policy]. Q learning is using Off-policy.
From the results, the on policy approach is more suitable for the Cliff Walking.

Expected Sarsa is like Q-learning but instead of taking the maximum over next state-action pairs, we use the expected value, taking into account how likely each action is under the current policy.

$$\mathrm{Q}(S_t, A_t) \leftarrow \mathrm{Q}(S_t, A_t) + \alpha \left[R_{t+1} + \gamma \sum_a \pi(a|S_{t+1}) Q(S_{t+1}, a) - Q(S_t, A_t) \right]$$

Except for this change to the update rule, the algorithm otherwise follows the scheme of Q-learning. It is more computationally expensive than Sarsa but it eliminates the variance due to the random selection of A_{t+1}.
So in the result, Expected Sarsa is better than Sarsa.