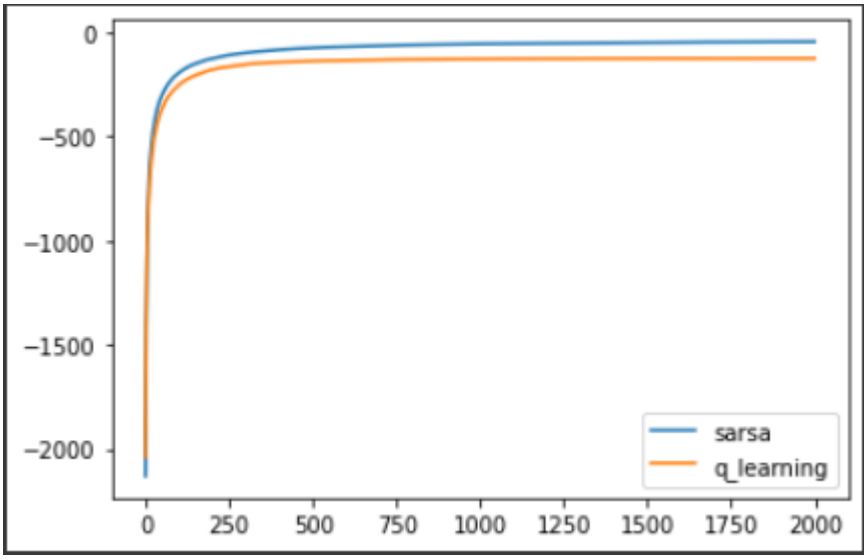
Experiments and Analysis (40%)

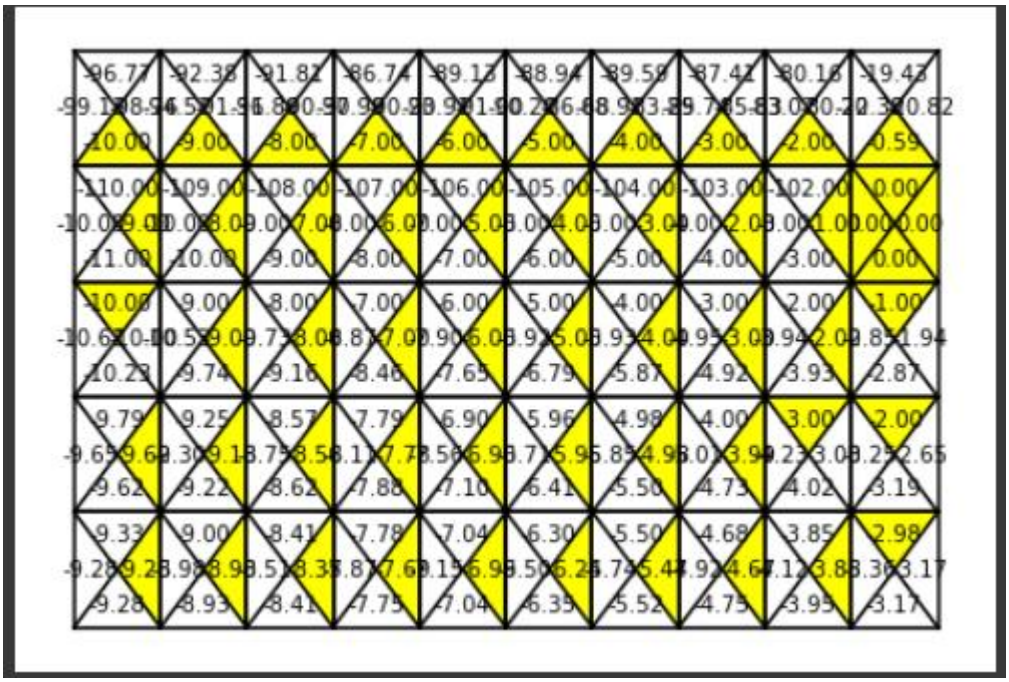
1. Plot the average rewards of Sarsa and Q-learning, and explain your result



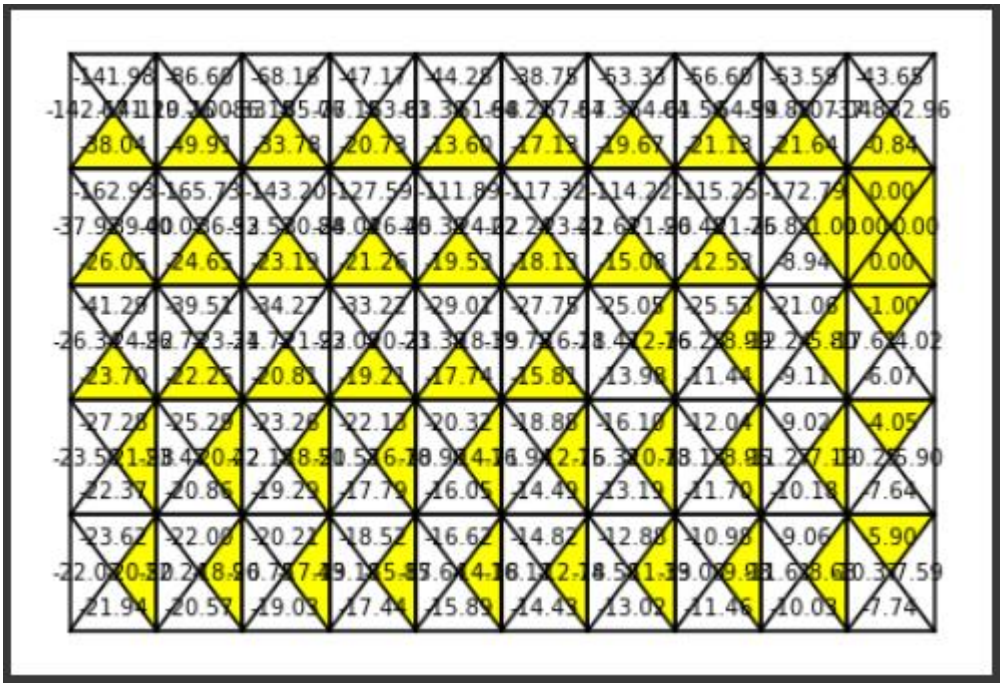
可以看出 Sarsa 為 on policy learning，因為其最後的平均的 reward 會比較高，因為他會走比較安全的路線，而 Q-learning 為異策略，所以其會學習最佳策略，但這也導致增加走到 swamp 的機率，所以其最後的 reward 會比較小。

2. Plot the Q-values of Sarsa and Q-learning, and explain your result.

Q table of Q-learning

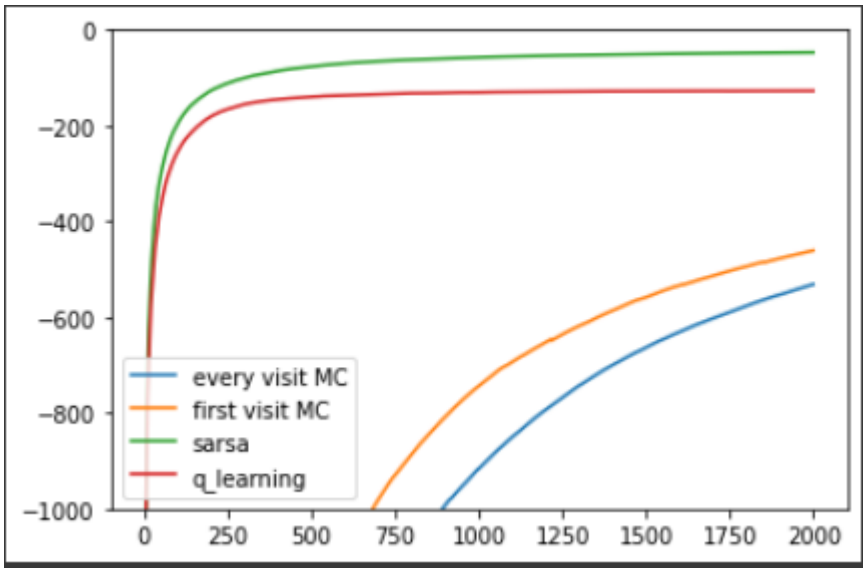


Q table of Sarsa



Q-learning 會學習最佳路徑，因為異策略，所以決定動作時不會考慮離 swamp 太近的問題，所以才可以學習出最佳路徑，而 Sarsa 為同策略，會考慮到離 swamp 太近導致 reward 變得更小，所以會走較遠的路徑。

3. Complete Monte Carlo, and compare average rewards



可以看出，sarsa 有最高的平均獎勵，因為同策略而且每一步都會更新，所以學習很快，而為何蒙地卡羅法會有較低的平均獎勵，是因為他在學習時必須走到最後才能更新 Q table，這會造成她一開始走很多不是最佳路徑的方法，所以導致整體平均會很低，而 Q-learning 為第二高的平均獎勵，但卻沒有 Sarsa 來的高，就跟上

述說的一樣，因為異策略導致其平均獎勵會較低一點，但其擁有最佳路徑。