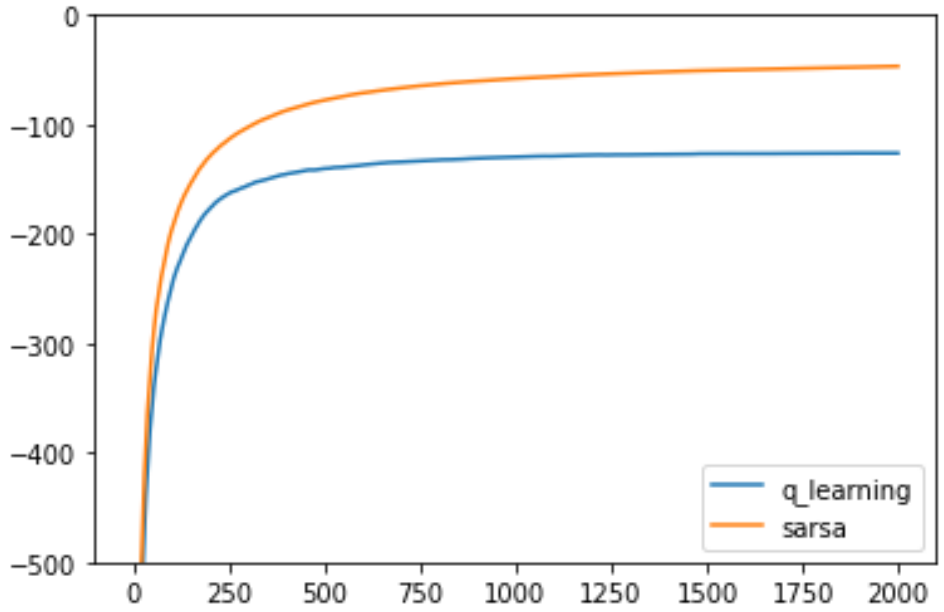
108061217 鍾永桓

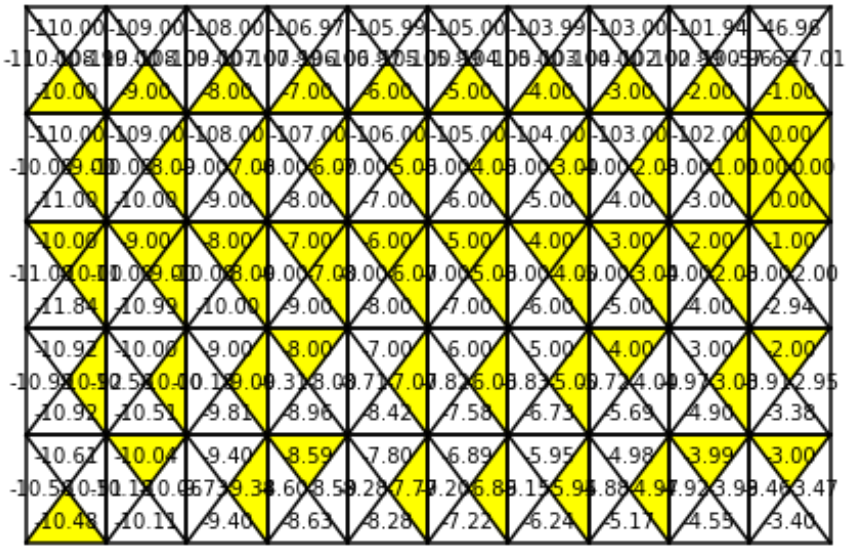
1. Plot the average rewards of Sarsa and Q-learning, and explain your result.(20%)



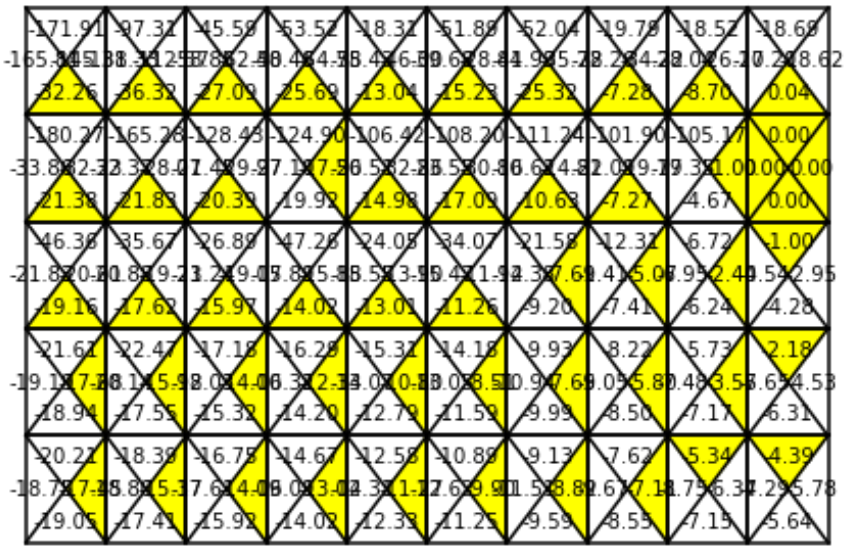
從曲線圖的結果來看，平均而言 Sarsa 在多個 episode 所收斂到的平均 reward 會比採用 Q-learning 更小，主要的原因應該是 Sarsa 所採用的策略相較 Q-learning 更為保守，Sarsa 會盡可能避免踩進 swamp，而 Q-learning 則可能為了選擇較短的路徑而冒較大的風險，這導致其在隨機動作選取時有更大的可能掉進 swamp，因此才會平均下來的 reward 較 Sarsa 低。

2. Plot the Q-values of Sarsa and Q-learning, and explain your result.(10%)

(1)Q-learning

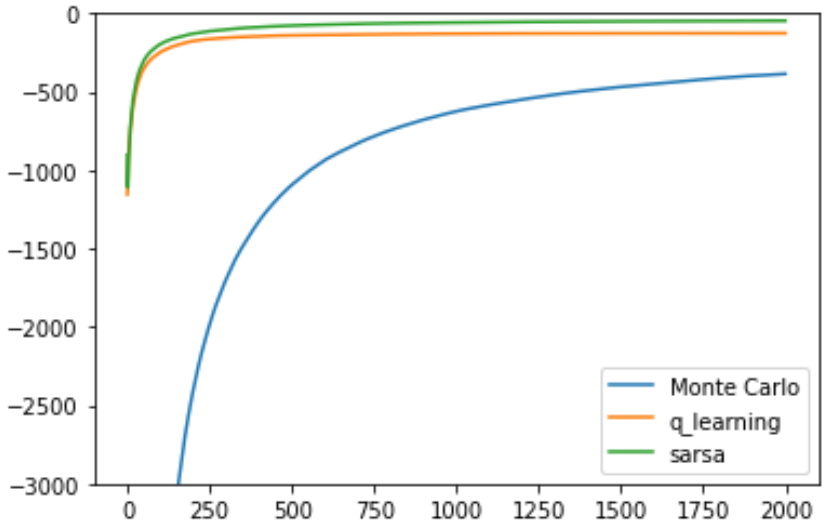


(2)Sarsa



圖中所表現出的 Q-value 也大致符合預期，由 Q-learning 所產生的策略比較大膽，所以從起點到終點前會偏好不斷向右直走一路到達終點，不過這條直走的路徑很接近 swamp，因此風險也相對很高，所以 Sarsa 就採取了比較保守的作法，由圖可看出 Sarsa 偏好先從起點往下走約兩步後才開始向右，如此的好處是因隨機行動而踩到 swamp 的機率便很低，比較安全，但是卻會導致步數增加。

3. Complete Monte Carlo, and compare average rewards.(10%)



由圖主要可以看出 Monte Carlo 收斂的速度很慢，訓練時間也更長，甚至跑到超過 2000 個 episode 都還沒能完全收斂，主要是由於 Monte Carlo 是回合制的更新，而且本題的圖又比之前更大，導致其相比於使用時間差法更難收斂。