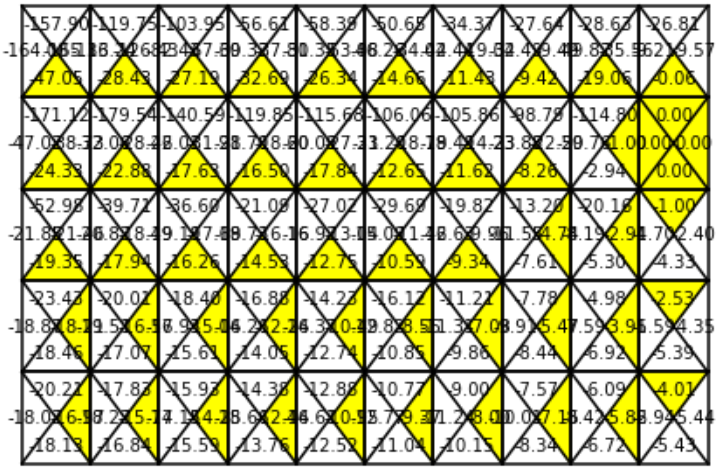
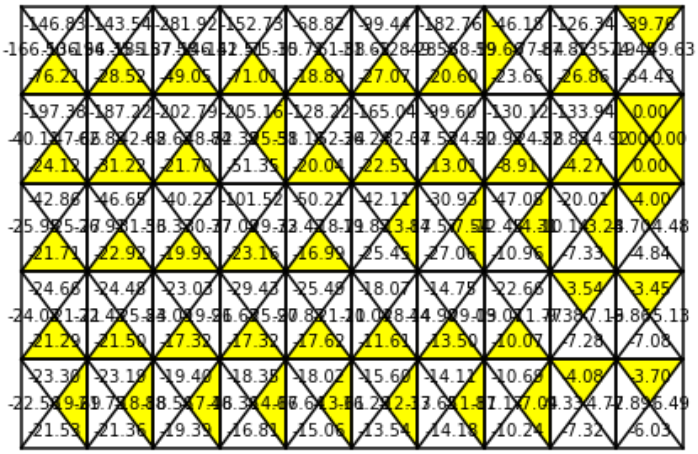
1. Plot the Q-values of Sarsa and 5-steps Sarsa, and explain your result.(15%)

   由 Q-values table 來看兩者皆不會直接走最短路徑到達終點，主要應是要避免落入 swamp 所造成的 reward 減少，不過雖然接繞遠路，兩者的 state-action value 分布還是有所差別，Sarsa 主要都是是連續直走後轉彎在連續直走，5-steps Sarsa 則相對有更多的轉彎，也就是 Sarsa 會一直所採取的行動變化較少，就為不靈活，但相對而言 5-step Sarsa 的行動更為複雜，例如轉彎次數較多。

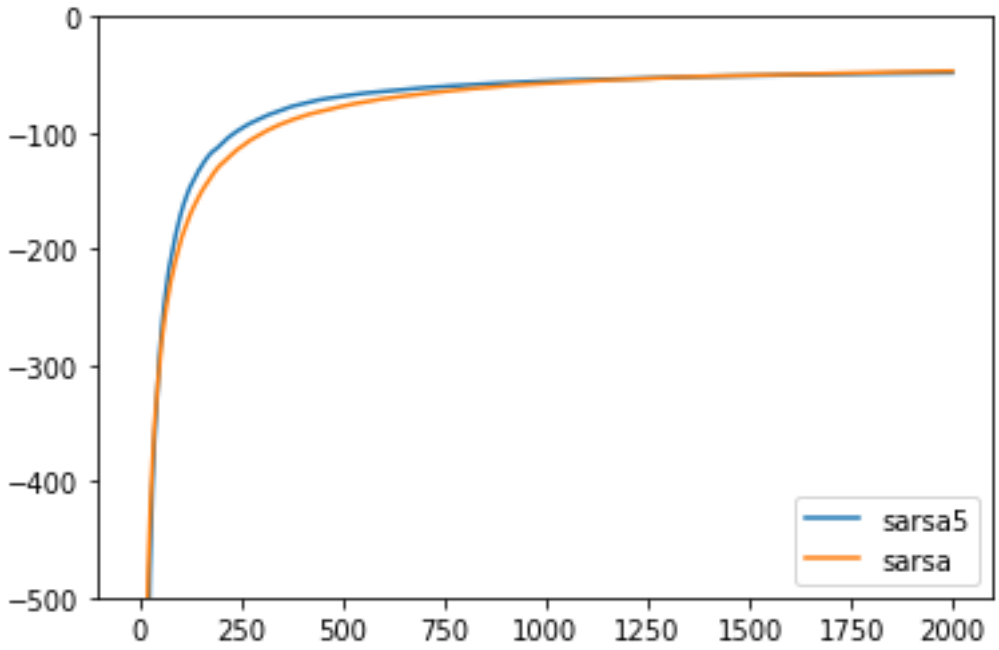




2. Plot the average returns of Sarsa and 5-steps Sarsa, and explain your result(15%)

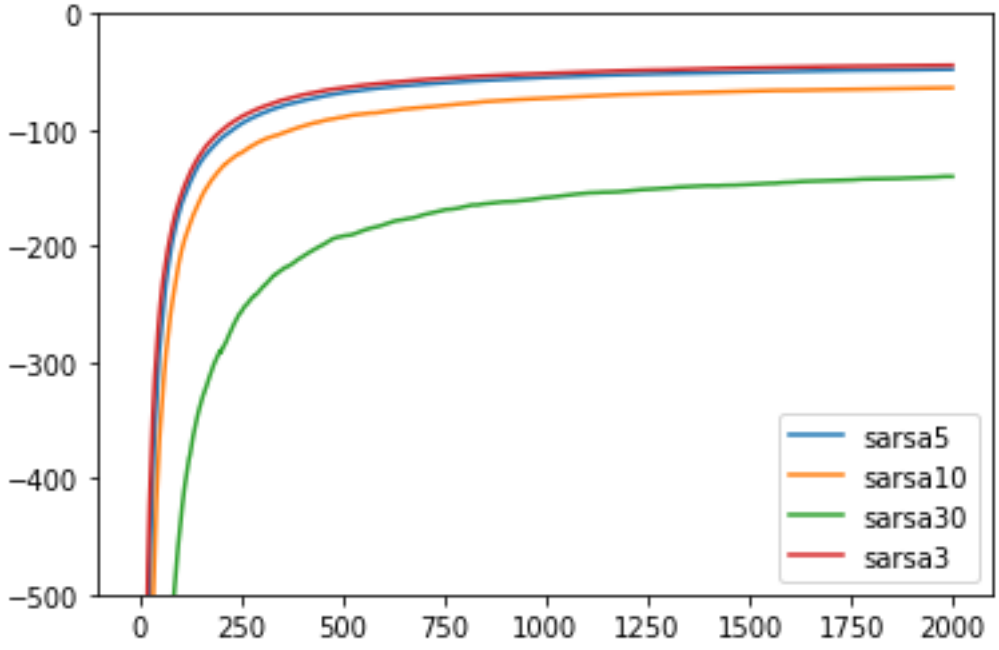
根據圖上的結果可以看出雖然兩個方法最終收斂到的結果非常接近，但是 5-step Sarsa 收斂比較快，它一次可以累積 5 步的資訊以此更新，但是 Sarsa 一

次只能用 1 步的資訊來更新，5 步的資訊中有更高機率包含有用的資訊，因此可以使 5-step Sarsa 學習的更加快速。



3. Varying n-steps and get average returns, then compare by overlap the plot(10%)

根據圖中的結果顯示未必 step 的數量越大越好，本題在 step3~5 時表現較為優異，收斂速度快且所收斂到的值也較小，但到 step10 收斂速度明顯下滑，30 又下滑更嚴重，而且收斂到的值也更小，有這樣的現象主要原因是雖然 step 變多時每次 state-action value 改變時會用到最後 n 步的資訊，步數越多可以累積更多資訊，但也相對減少了 state-action value 改變的次數，在學習過程 state-action value 改變較慢，相較之下 step 少時可以在學習中有更多次的 state-action value 改變，而因為在學習中 state-action value 不斷改變，因此選擇的行動也會越來越好，return 也不斷變小，因此要選擇適當的 step 來確保累積足夠的資訊，但又不要導致 state-action value 改變次數過少。


0
−100
−200
−300
−400
−500
0
250
500
750
1000
1250
1500
1750
2000
sarsa5
sarsa10
sarsa30
sarsa3