EE551000-System-Theory-Hw2

謝昉澂

109061589

Implementation:

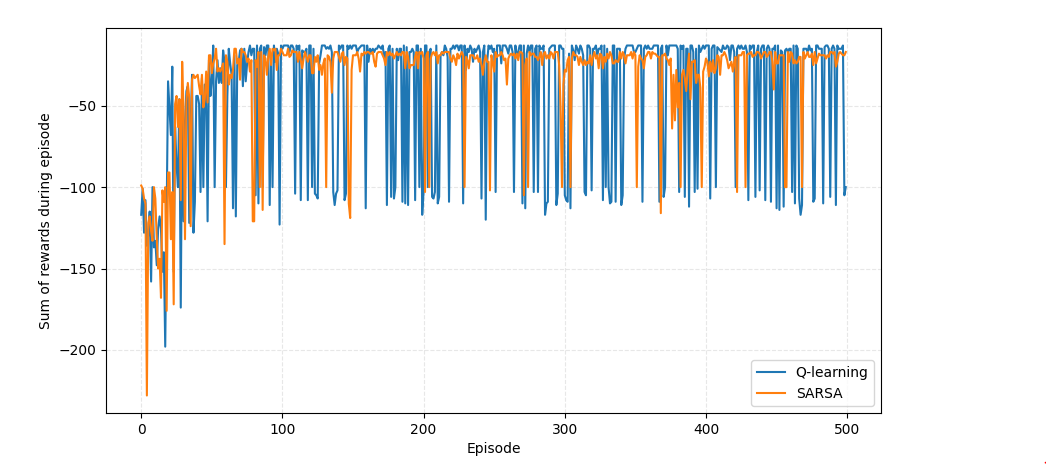
更改algo.py 裡的演算法

q-learning以epsilon-greedy當作behavior policy，以greedy當作target policy。

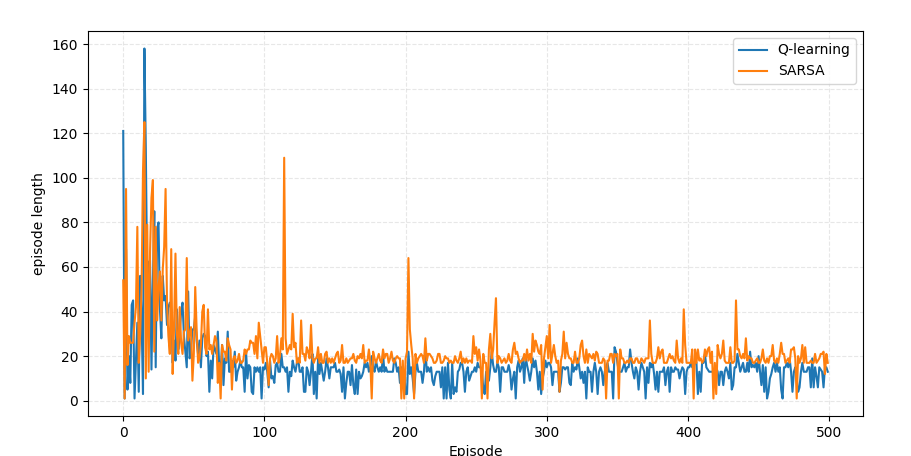
sarsa以epsilon-greedy當作behavior policy，和target policy。

Experiments and analysis:

1.Plot curves of different methods into a figure. (As example above)



2. Plot the episode length (time steps taken per episode) v.s. episode. What do you observe?

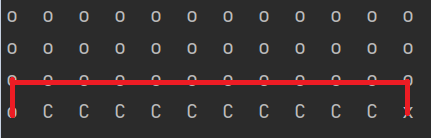


Sarsa回合的長度通常比q-learning長。

Render and show the trajectory of each method. What do you observe?

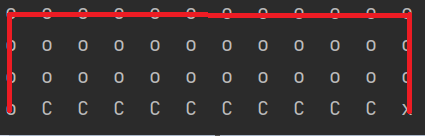
q-learning:

會選擇最佳的路徑



Sarsa:

會選擇最安全的路徑



Observe the reward curve of each algorithm. We can observe that the reward curve of SRASA is more stable than Q-learning (less severe drop to -100). Please explain.

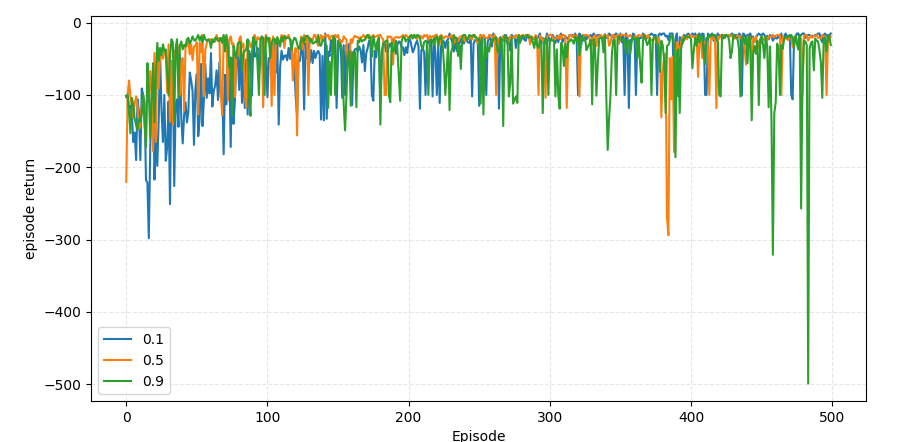
因為sarsa會選擇遠離懸崖的路徑走，所以偶爾使用隨機策略的時候較不容易直接掉進懸崖。

Why is Q-learning considered an off-policy control method? How about SARSA?

q-learning以epsilon-greedy當作behavior policy，以greedy當作target policy，兩policy不同所以是off-policy。

sarsa以epsilon-greedy當作behavior policy，和target policy，兩policy相同，所以是on-policy。

Vary the TD learning rate 𝛼, what happens?



以sarsa觀察，當alpha降低會導致收斂速度下降，因為新資料的權重降低。