EE551000-System-Theory-Hw2

謝昉澂

109061589

Implementation:

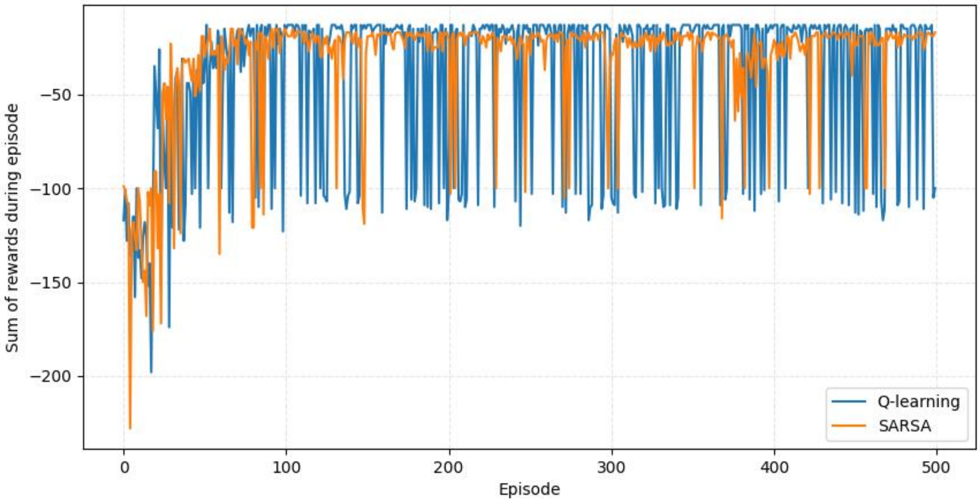
更改 algo.py 裡的演算法

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

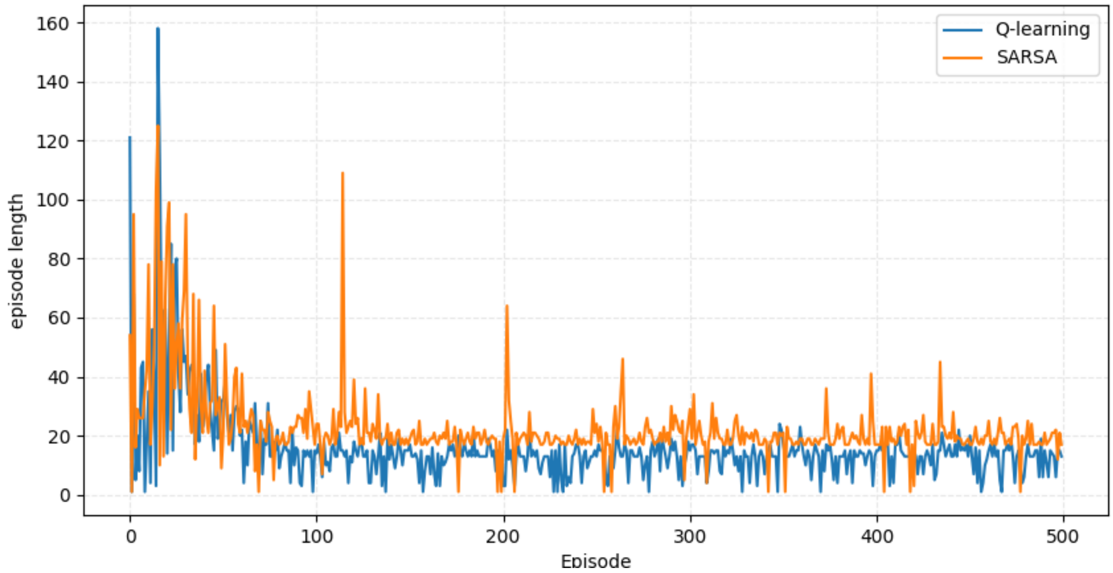
sarsa 以 epsilon-greedy 當作 behavior policy，和 target policy。

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?

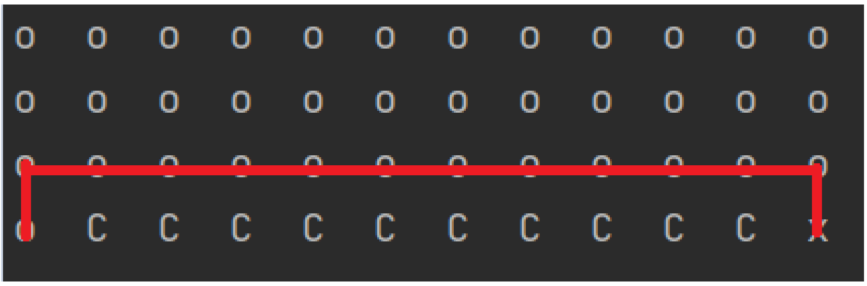


Sarsa 回合的長度通常比 q-learning 長。

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

q-learning:

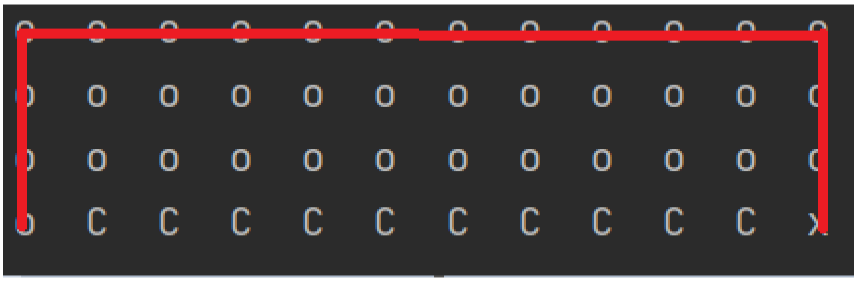
會選擇最佳的路徑



```
O   O   O   O   O   O   O   O   O   O   O   O
O   O   O   O   O   O   O   O   O   O   O   O
O   O   O   O   O   O   O   O   O   O   O   O
O   C   C   C   C   C   C   C   C   C   C   x
```

Sarsa:

會選擇最安全的路徑



```
O   O   O   O   O   O   O   O   O   O   O   O
O   O   O   O   O   O   O   O   O   O   O   O
O   O   O   O   O   O   O   O   O   O   O   O
O   C   C   C   C   C   C   C   C   C   C   x
```

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.

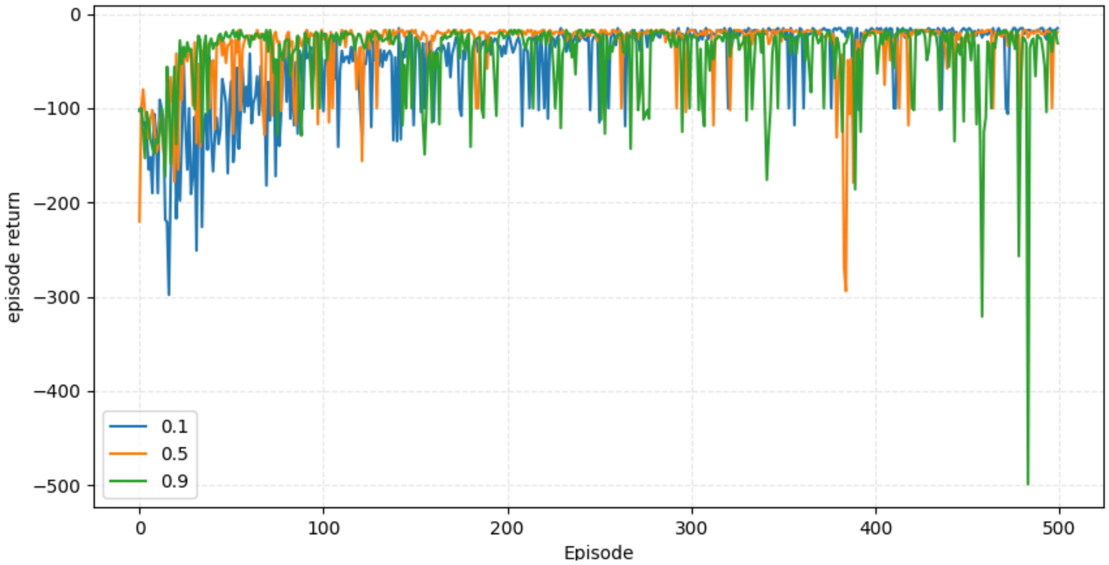
因為 sarsa 會選擇遠離懸崖的路徑走，所以偶爾使用隨機策略的時候較不容易直接掉進懸崖。

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。