108061217 鍾永桓

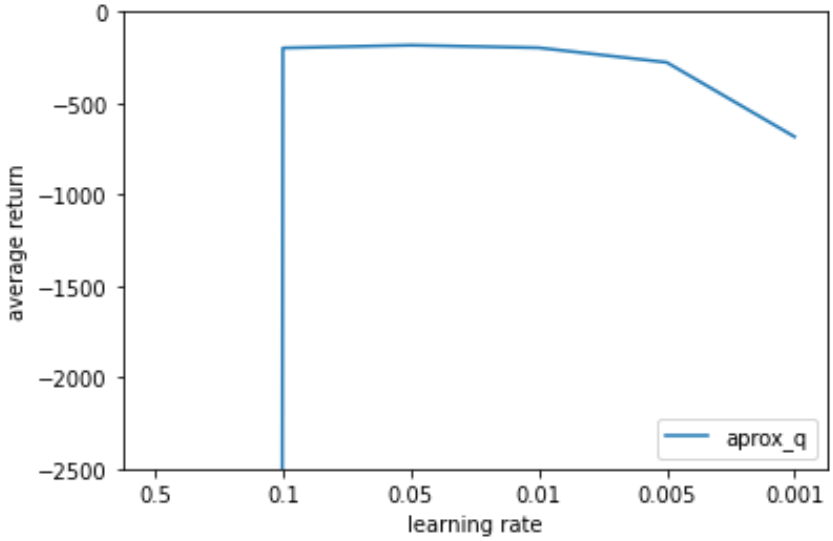
Question 1

11 Line 的行為是參數更新，它相當於 Q-table method 當中的動作價值函數更新，因為無論是動作價值函數或是參數都是要用來估計 state-action value。無法用 Q-table method 的原因在使用 Q-table 時狀態不可以過多，必須要確保代理人可以拜訪所有的狀態足夠次數，在狀態非常多時若採用 Q-table method，可能無法保證每個狀態被拜訪到足夠次數，如此會使 state-action value 無法適當收斂，但若是使用參數表達的函數來估計 state-action value，則可以用相對小很多的維度來達成，因此在狀態很多時無法藉由直接更新 Q-table 來使 state-action value 的估計收斂，故將 state-action value 的表示成用某些參數表達的函數，藉由更新這些參數來改變 state-action value 的估計。

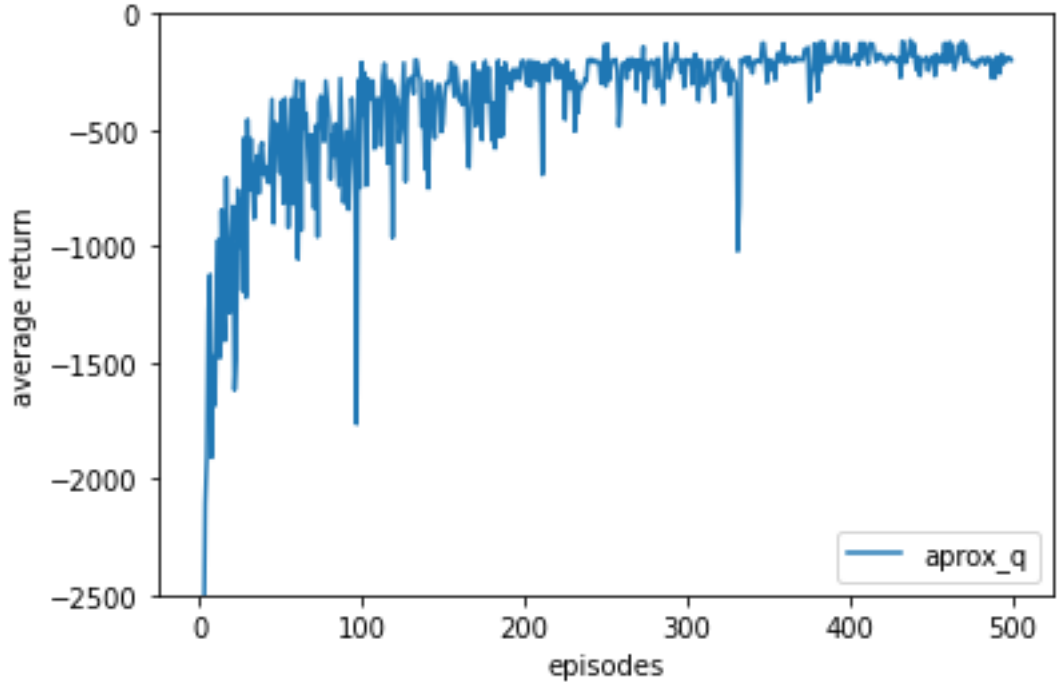
Experiment 1

根據所得到的結果，當 learning rate 太大時參數將會無法正常收斂，導致 average return 極小，而在 learning rate 約為 0.05 可以最良好的收斂，有最高的 average return，不過若 learning rate 太小，也沒辦法順利收斂。

Learning rate	Average return
0.5	-500000.0
0.1	-198.2
0.05	-182.89
0.01	-196.92
0.005	-276.78
0.001	-683.88



Experiment 2



最後 100 回合的 average returns: `-191.83`