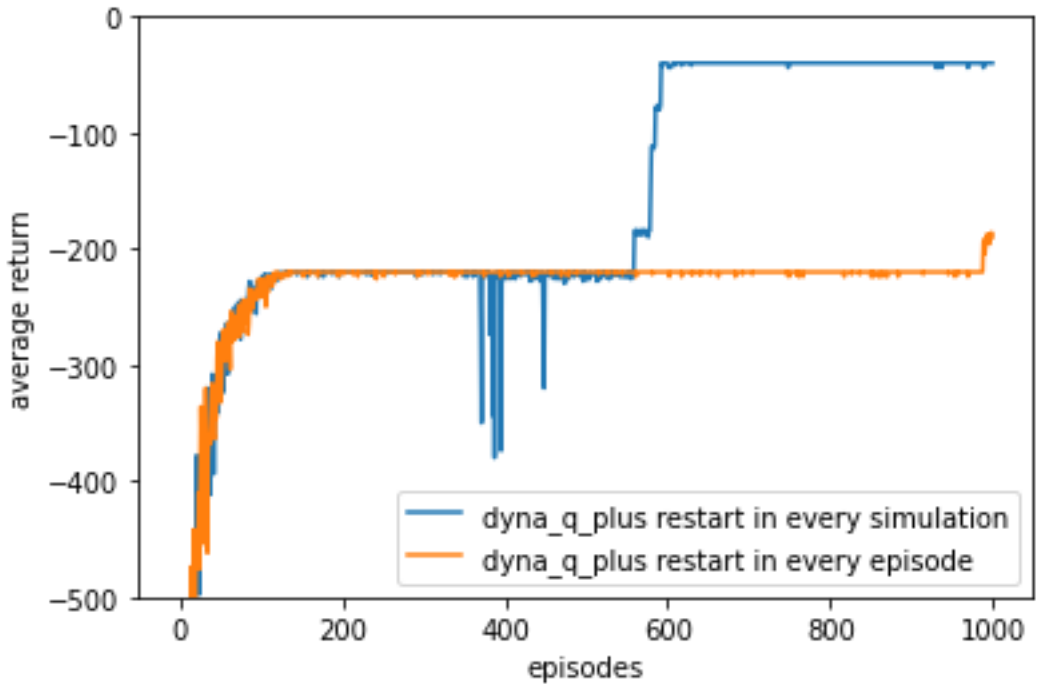
108061217 鍾永桓 HW7

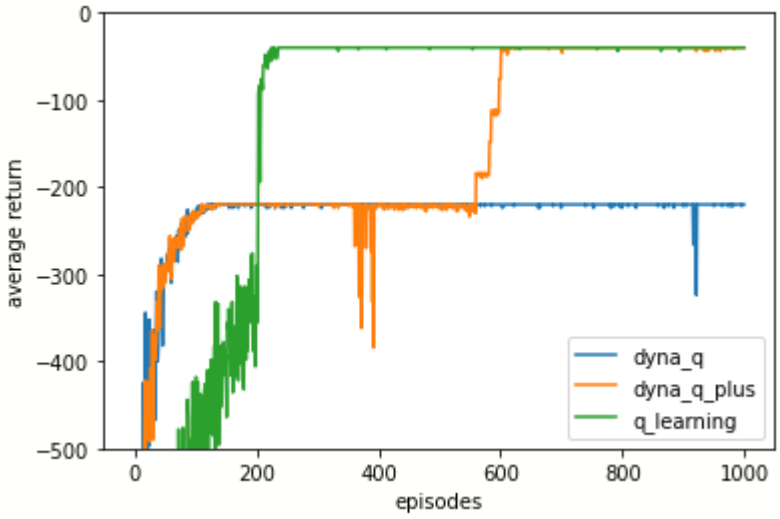
Experiment 1 (20%)

In dyna-Q+ algorithm line20 ,τ need to restart in every episode or restart in every simulation?please do the experiment and explain your answer.



根據實驗的結果 restart in every simulation 應為正確的選擇，restart in every simulation 才能使 reward 更高，主要原因應該是若在每個 episode 結束就從新設為 0 並沒有辦法很有效的累積內在獎勵，在環境改變後內在獎勵也是被從新由 0 開始計算，無法得知環境改變前哪些狀態動作組合很久沒有採用過，故而無法鼓勵針對這些很久沒使用過的狀態動作組合的探索。

Result(60%)



Question 1

Why Q-learning can react instantly when environment change?

Q-learning 是一種異策略時間差分控制法，普通的 Q-learning 之所以可以針對環境及時做出反應是因為它採用不同的目標策略和行為策略，它在更新 Q-value 時都會採用 Greedy 的動作選擇，因為更新時的動作選擇為 Greedy 使它相較 Sarsa 等方法更具有探索能力，更新會更加偏向找出最佳解法，故環境改變後它也會很快就開始從新的環境中找最佳解法，而它跟 Dyna-Q 和 Dyna-Q plus 相比，Dyna-Q 雖然也採用 Greedy 的動作選擇，但是它所採用的間接強化學習選用的是過去已經採取過的狀態行動組合，但若環境變化，可能會有原本過去不可能採用的狀態行動組合，模型會無法很快的學習到來自這些過去沒有的狀態或行動的經驗，故而依然依靠模型由過去環境累積的經驗來更新 Q-value，這使 Dyna-Q 無法很快對環境做出反應，而 Dyna-Q plus 因為可以在規劃過程產生現實沒出現過的狀態動作配對，而且還藉由內在獎勵來鼓勵針對很久沒拜訪過的狀態動作配對的探索，使模型可以在非穩定的環境下更快被修正。