人工智慧HW2

Reinforce-learning: Model free approach

機械4B 109303552 周奕安

2024 / 04 / 07

目錄

[**一、** **目的:** 2](#_Toc163386414)

[**二、** **方法** 2](#_Toc163386415)

[**三、** **模擬結果** 3](#_Toc163386416)

[**四、** **討論** 5](#_Toc163386417)

[**五、** **結論** 7](#_Toc163386418)

1. **目的:**

本次目標為透過Monte Carlo method 進行model free approach，目標為以最少步數抵達最大值，目標函數繪製的3D mesh如Fig01，並從運算時間與步數探討參數導致的差異。

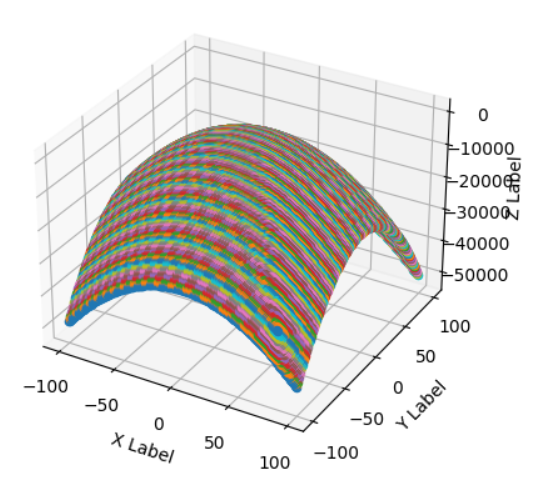


Fig01:目標函數

1. **方法**

本實驗以Monte Carlo method搭配ε greedy method模擬，前提為已知終點位置(確保episode終止)，將狀態設為座標，獎勵為z軸上升距離，並包含9種動作，共模擬10000個episode，範圍為x、y在-5~5，每個episode透過更新random seed隨機在不同起點。而為了確保探索階段不會收斂過快，將ε greedy method的ε分為兩個階段(Fig02): 探索階段與收斂階段，在探索階段設定較高的ε下限避免過快收斂，而收斂階段則是有較低的ε下限，ε不為0是為了避免在特定位置卡住導致無法收斂。

評價方式包含運算總時間及最後10%的(最佳步數/實際步數)取平均(1代表最佳、數字越小成果越差)。而實驗項目包含reward的衰減率γ、ε、ε的衰減率、探索階段佔episode的比例，透過python argparse 套件及batch file測試個參數的影響並優化參數。透過(1/評價)可推算實際步數與最佳解的倍數(如評分0.5代表所需步數為最佳解的2倍)。

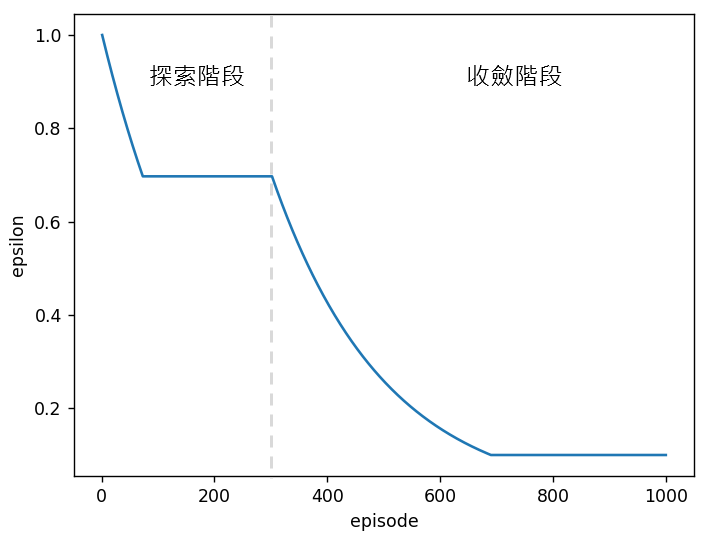
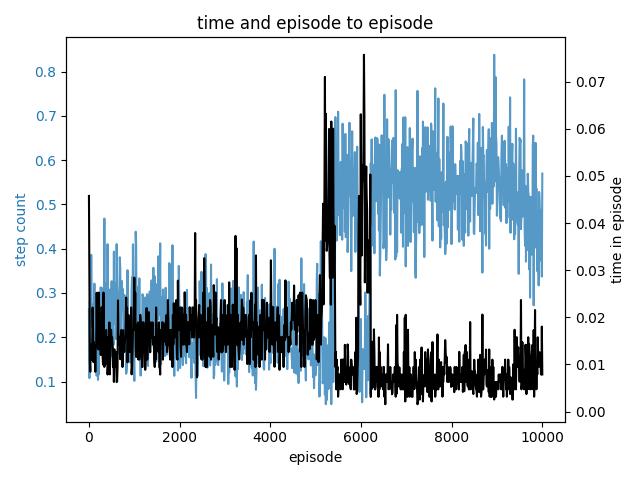
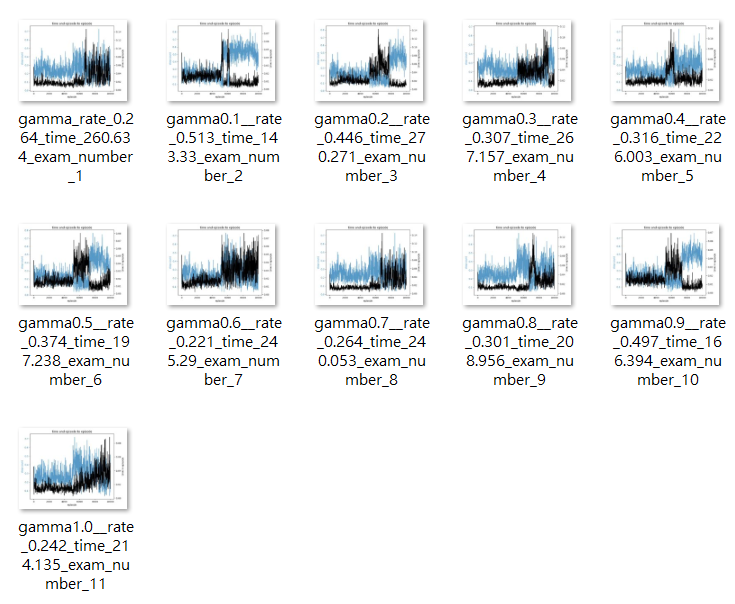


Fig02:ε收斂方式(範例)

1. **模擬結果**

在每個實驗都會輸出圖形、模型的評分及總時間(Fig03)，由於圖表太多，僅列出一部份。而各參數的影響見Fig04~08，下個章節會根據模擬結果進行探討。

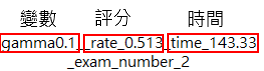


Fig03:圖形(左)、實驗資料夾(右)、檔名資訊(下)

Fig04:γ對模型影響

Fig05:ε對模型影響(<=0.5無法收斂)

Fig06:ε discount對模型影響

Fig07: discover 階段比例對模型影響

Fig08: discover 階段 ε threshold 對模型影響

1. **討論**

經過多次模擬後，各參數的影響如上圖，而最終根據實驗與討論結果，嘗試組合出最佳的參數。

γ對模型影響在這個問題及獎勵機制中，根據線性回歸，在γ較小時有較佳的成果，不過決定係數不高，再加上實驗與機率有關，因此這項參數影響相對不明確，在這組實驗中，平均評價為0.305。

ε雖然數據量較少，但對模型影響相對明顯，在ε<=0.5無法在1,000,000步內抵達終點，而ε越大，除了評價越高，所需的時間越短，Fig09列出ε=0.6、0.9的圖形，我認為時間評價較佳的原因是多次探索，使value table較完整，且較不容易在初期卡在相同位置，時間的降低的原因除了所需步數少，在初期所花的時間也差很多，在ε較小時，初期的運算時間差異明顯很大。這組數據平均評價為0.258

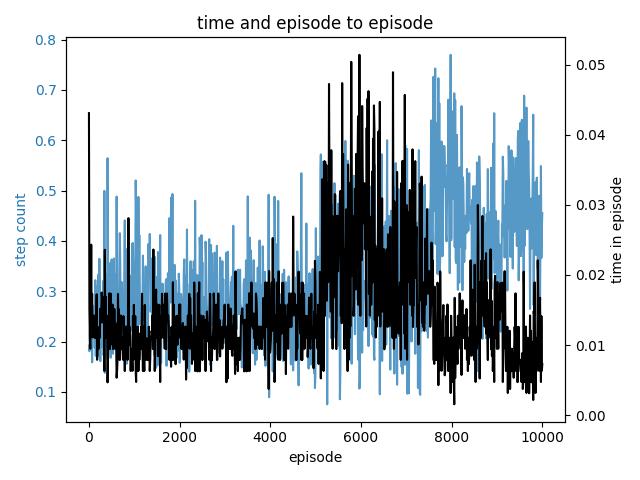
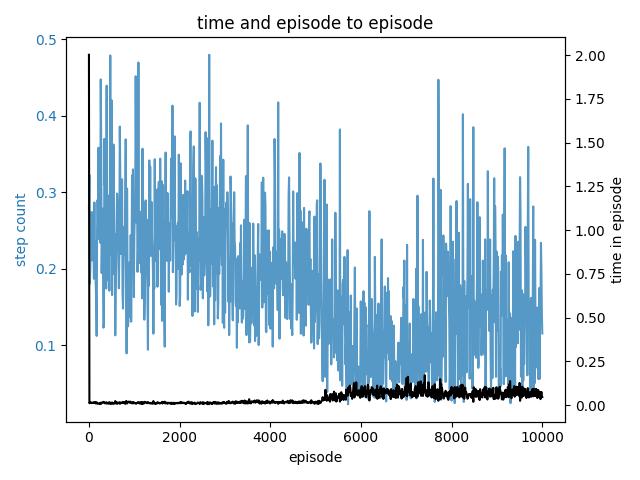


Fig09: ε=0.6(左)，ε=0.9(右)

接著是ε衰減率，由於episode很多，所以將它設為0.98~1的數值，在這組數據中並未發現明顯的關聯性，比較特別的是ε衰減率為1時，所需的時間與評價表現都明顯不佳，因為全程的決策都是隨機決定的，這組數據的平均評價為0.287。

而探索階段的比例有較高的決定係數，隨著探索階段的占比提升，模型評價也越高，所需的時間也越短。我認為是因為在探索接短有較多收集資資訊的效果，使每個state-action的value都很齊全，最終收斂效果較好。這組數據的平均評價為0.276。

最後是探索階段ε閾值，這項參數的影響也不明顯，值得注意的是在探索階段ε閾值=1時有最佳效果與最長時間，我認為是因為在這個階段幾乎一直在隨機移動，需要更多時間來抵達終點，但資訊也收集的特別完整，這個數值在設定時需考慮ε衰減率與探索階段的比例，這三者都會影響決策的機率。這組數據的平均評價為0.297。

最後透過上述觀察組合參數，但參數間會互相影響，最終最佳結果為0.502，不如過程中發現的最大值0.513，但能相對穩定訓練出不錯的模型(Fig10)，平均值為0.429，遠高於其他實驗平均值。另外發現episode在超過一定數值後沒有明顯影響，而初始設定的範圍越小結果越好，在-3~3的區間評價可達0.796(Fig11)。

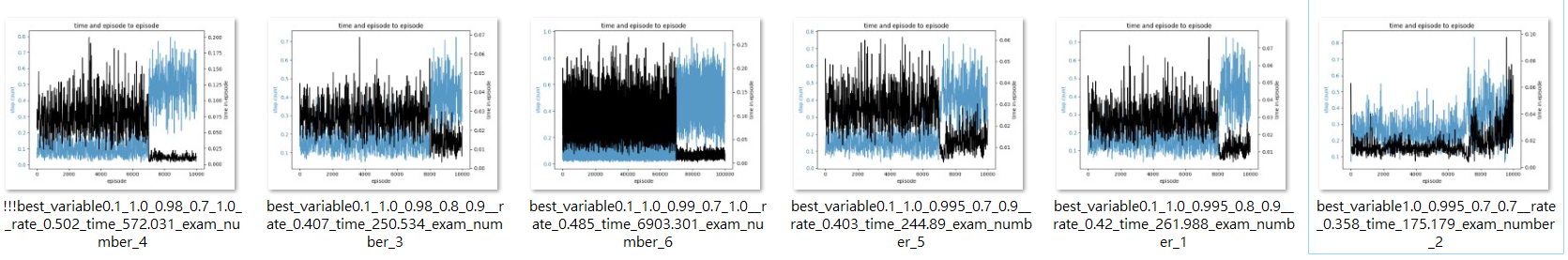


Fig10: 優化參數實驗

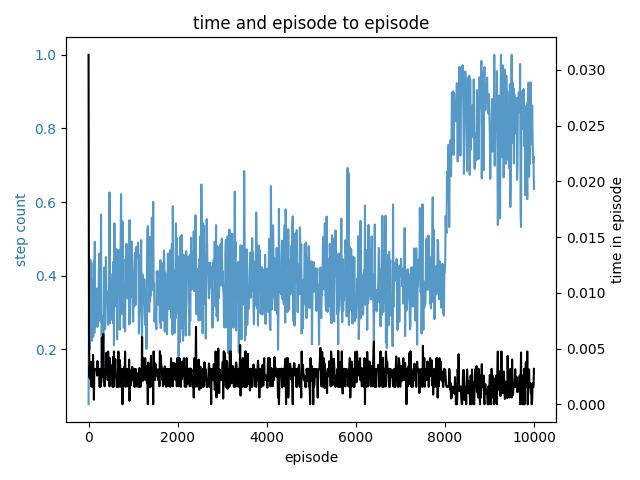


Fig11: 小區間(-3~3)成果

1. **結論**

透過實驗，觀察了各種參數對模型的影響並發現ε和探索階段的佔比影響最明顯，而嘗試根據結果優化參數，在-5~5的區間中模擬取得最高評價0.502(需要最佳步數的1.99倍的步數)，且平均評價為0.429(需要最佳步數的2.33倍的步數)，而在-3~3的區間中模擬更取得0.796的評價(需要最佳步數的1.26倍的步數)。雖然並未取得完美的成果，但比最初隨機設參數容易成功。