1. Title

基於MDP的迷你黑白棋AI

1. Background and Motivation

2.1 Motivation

基於搜索估值和黑白棋AI，在使用同形表、歷史記錄、節點排序等減少重複搜索以及增加剪枝的頻率的技巧後，要再進一步提升棋力，就只能改善棋局估值函數或使用更多更快的處理器來加深搜索深度了。然而，要能盡可能準確地描述棋局的優劣，需要一定的下棋能力，有些甚至是棋手憑經驗和說不清道不明的直覺做的決策，這樣就好難提升程式的棋力了。

2.2 Background

基於MDP的黑白棋AI可以在經過足夠量的訓練後得出合法步的好與壞。

2.3 Problem Definition

在每一步的落子決策中選擇對自己優勢最大的那一步。

1. Methodology

3.1 Method chosen

QLearning

3.2 Reason

容易實現，且QLearning對下一state的value的定義是該state中能帶來最大好處的動作所帶來的value，這和下棋時每一步都想要提高自己的勝率的目標相符。

3.3 Justification

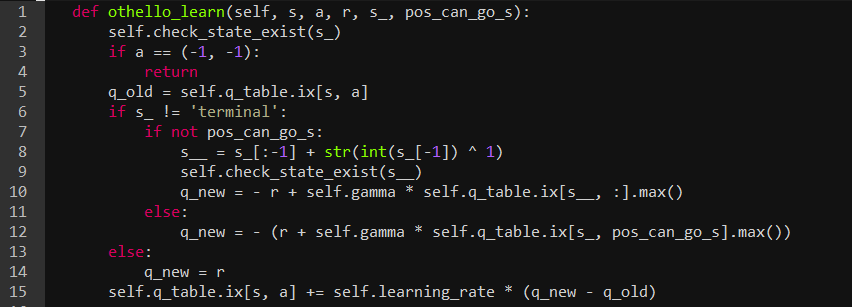
已知3 × 3 的黑白棋是先手必勝，而4 × 4的黑白棋的先手必敗。本專題的程式在3×3黑白棋中通過QLearning學習100局後，可以得出先手有很大優勢的結論，而在4 × 4的黑白棋中，經過超過3千次的學習後，得出了先手有很大劣勢的結論。這些均與事實相符。

1. Data Collection and Analysis Result

4.1 Data Collection

通過自己與自己對下來創造數據。

4.2 Model Formulation



新Q值主要在第12行得出，即對方選最佳合法步後的value與reward的負值（負值是對方優勢相對於己方而言），並在第15行更新。

但黑白棋會出現對方沒合法步的狀況，這時會跳過對方回合，也就是第8至10行。（沒這麼做會導致這state往下的值永遠傳不上來）跳過對方回合後又輪到己方下棋，因此第10行從自己的合法步中選擇最佳步，這時不用負值。（因為這裡的值是己方的值）

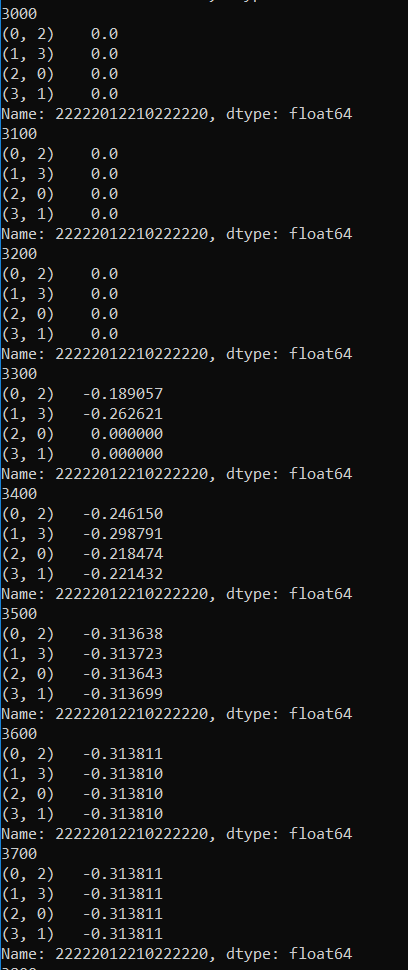
而若自己沒合法步（第3行判斷）便直接跳過更新。（因為這state沒必要更新，而在這一state的前一個state時，已經跳過這一state而直接用下一state的值更新到上一state的表格裡面了）

4.3 Analysis

下圖為3×3黑白棋經過100次訓練的結果，可見先手方（黑方）的合法步均為 正值，與必勝結果相符。



下圖為4×4黑白棋經過3000次訓練以後的效果，可見先手方（黑方）在經過3700次訓練後，最開始的那四個合法步的值均為負值，這亦與4×4先手必敗的結果相符。



1. Conclusion

5.1 Summarize

結果還算成功，但效果有待加強。4×4用暴力搜不到一秒便可得出結果，而訓練3千多次會花費數分鐘。但與搜索相比，訓練後只需查表便能得出下一步的位置。

5.2 Future research

a. 旋轉+反射減少state數

有些看似不同的棋面通過旋轉和反射會形成相同的棋面，可將這些棋面視為同一個，那麼一次更新就能比目前的效率高數倍。

b. Sarsa lambda加快訓練速度

QLearning是單步更新，效率較不佳，若時候sarsa lambda則能多步更新，縮短訓練時間

c. DQN (Deep-Q-Network)

在棋面是8×8甚至是10×10時，傳統QLearning或Sarsa或許就在state的儲存上就有困難了，這時可能需要神經網路來取代表格做state與action 生成value的過程。

1. Member

陳聲發 F74057049

王皓玄 F74047052