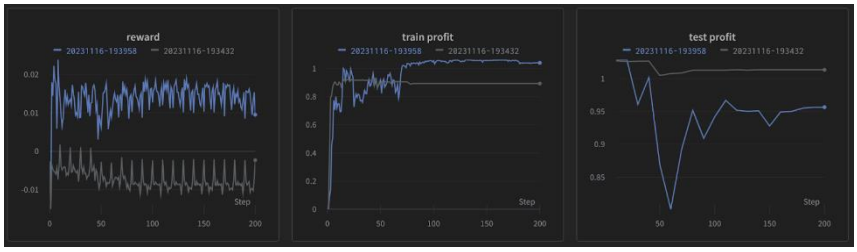
# CE6023 Homework2 Report

學號：111502531　　姓名：趙啟翔

1. **(10%) Policy Gradient 方法**

1.1 請閱讀及跑過範例程式，並試著改進 Reward 計算的方式。

- 灰色線是 Baseline，藍色線是改進後 Reward 的結果




1.2 請說明你如何改進 Reward 的算法，而不同的算法又如何影響訓練結果？

在使用 Policy Gradient 時，目標是直接優化一步的 Reward，使得一步的獎勵最大化。而我嘗試加入獎勵衰減權重 gamma=0.75 最大化多步獎勵。

1.3 補充說明

我有嘗試使用 test.csv 訓練過，但是發現好像訓練不太起來。在 train.csv 中，2013/5~2017/6 之間的資料進行訓練，是最接近 test.csv 的。


2. **(15%) 試著修改與比較至少三項超參數（神經網路大小、一個 batch 中的回合數等），並說明你觀察到什麼。**

2.1 更改超參數 1 – Network Hidden Size

- 藍色線是 Policy Gradient(有做獎勵衰減)並將 Network Hidden Size 設為[256,256](原始值) 。
- 橘色線是 Policy Gradient(有做獎勵衰減)並將 Network Hidden Size 設為[32,32] 。


- 觀察

  Reward Chart：兩條線都在零附近震盪，但橘色線較穩定。

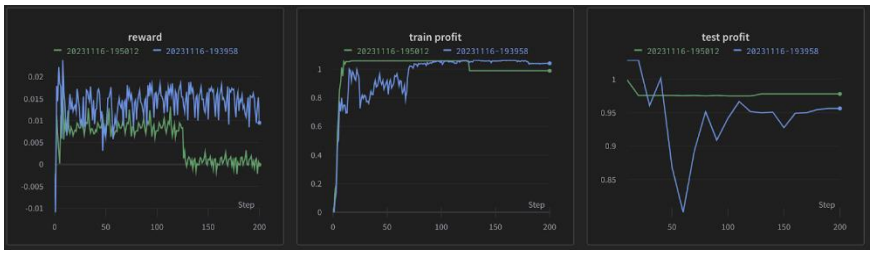
  Train Profit Chart：藍色線比橘色線學習得更好，利潤更高。

  Test Profit Chart：藍色線一剛始表現得不錯但後來下降，可能是 overfitting；橘色線最初表現不錯，但後來急劇下降，可能因為學習能力有限。

  總結：隱藏層越大的學得更多，但可能 overfitting；隱藏層教小的學習相較穩定，但能力有限。

2.2 更改超參數 2 – Learning Rate

- 藍色線是 Policy Gradient(有做獎勵衰減)並將 Learning Rate 設為 1e-3(原始值) 。
- 綠色線是 Policy Gradient(有做獎勵衰減)並將 Network hidden size 設為 1e-2。




- 觀察

  Reward Chart：藍色線相較綠色線有較小的波動。

  Train Profit Chart： 藍色線穩定上升，而綠色線則顯示快速提升後穩定。

  Test Profit Chart：兩條線皆在測試階段時都相對穩定，但藍色線在高點波動，綠色線則呈輕微上升趨勢。

  總結：較低的學習率在訓練和測試中表現出更穩定的學習過程。

  較高的學習率在訓練利潤上升快，但後期增長放緩。

2.3 更改超參數 3 – Batchsize

- 藍色線是 Policy Gradient(有做獎勵衰減)並將 Batchsize 設為 512(原始值) 。
- 綠色線是 Policy Gradient(有做獎勵衰減)並將 Batchsize 設為 64。


- 觀察

  Reward Chart：藍色線和綠色線波動大致相同，沒有顯著差異。

  Train Profit Chart： 藍色線的利潤增長穩定，而綠色線則顯示利潤有較快的上升，後期趨於平穩。

  Test Profit Chart：藍色線在測試階段表現較為穩定，而綠色線則在初期下降，後期回升。

  總結：Batchsize 較大在訓練和測試中表現出較為穩定的趨勢。而 Batchsize 較小在訓練利潤上升速度較快，但在測試階段的表現較為波動。

3. **(15%) 請同學們從 Q Learning、Actor-Critic、PPO、DDPG、TD3 等眾多 RL 方法中擇一實作， 並說明你的實作細節。**

在眾多 Reinforcement Learning 的方法中，我選擇實作的是 Proximal Policy Optimization（PPO）。我採用了以下的參數設定：

- Gamma 設為 0.9，以表示未來獎勵相比於當前獎勵的重要程度較低
- Lambda 設為 0.9，在計算 GAE 時使用，有助於減少方差並保留一定程度的偏差
- Batch size 設為 512
- 對觀察值進行了標準化處理
- 對 advantage function 進行了標準化處理，以平衡不同狀態下的獎勵差異
- 採用 one step TD-return 進行優勢估計，以增加學習的效率
- PPO clip threshold 設為 0.2，用以限制策略更新的範圍，從而提高學習的穩定性
- Learning rate 設為 1e-3，這是一個相對較小的值，有助於防止在學習過程中出現過大的參數更新
- Actor network 和 Critic network 分別設定了[256, 256]的網絡結構，這表示每個網絡都有兩個隱藏層，每層包含 256 個神經元

**4. (10%) 請具體比較（數據、作圖等）你實作的方法與 Policy Gradient 方法有何差異，並說明其各自的優缺點為何?**

3.1 實作方法差異

- Baseline 實作方法是用 Policy Gradient
- 我這次作業使用的方法是 Proximal Policy Optimization

3.2 Policy Gradient

- 優點

  適合處理連續動作空間的問題，因為它可以直接輸出一個動作，而不是動作的概率分佈。

- 缺點

  策略梯度方法往往有較高的方差，需要更多的樣本才能穩定學習，這使得訓練過程變慢。

3.3 Proximal Policy Optimization

- 優點

  通過限制策略更新幅度，提供了一個更穩定且高效的學習過程。

- 缺點

  PPO 的效果在很大程度上依賴於超參數（如剪裁範圍）的選擇，這可能需要大量的試錯來找到最佳設定。