國立政治大學金融學系

碩士學位論文

Black-Litterman 模型結合強化學習之投資組合配置

Black-Litterman Portfolios with Reinforcement Learning Derived View

指導教授：廖四郎 博士

研究生：李雍群 撰

中 華 民 國 一一一 年 六 月

DOI:10.6814/NCCU202200681

摘 要

本研究嘗試將強化學習 (Reinforcement Learning) 應用於預測金融資產價格走勢，並結合 Black-Litterman 模型建構全球性多元投資組合。本研究使用近端策略優化演算法 (Proximal Policy Optimization, PPO)，以資產價量資料預測資產價格漲跌及漲跌幅度，並將預測結果作為 Black-Litterman 模型中的投資者觀點進行資產配置，比較投資組合在不同獎勵設定及不同更新次數下的績效表現。本研究以美國五檔不同資產類別 ETF 作為基礎資產，研究結果顯示強化學習在一定更新次數上具有預測力，本研究建立之投資組合績效在更新次數 600000 皆能贏過其餘基準模型。另外，對於強化學習而言，以不同獎勵設定訓練模型比起增加更新次數對績效有著較大的影響。

關鍵詞：投資組合、強化學習、Black-Litterman 模型、近端策略優化

DOI:10.6814/NCCU202200681

Abstract

In this thesis, we try to apply reinforcement learning to forecasting price trends of finance assets. We combine the forecasts with the Black-litterman model and construct various globally diversified portfolios. In our paper, we use the Proximal Policy Optimization algorithm to forecast assets' price trends by historical price and volume. The prediction of results are used as the investor's views in the Black-Litterman model for asset allocation. This study compares the performance of the portfolios under different reward setting and number of updates. The empirical results show that reinforcement learning has predictive power at a certain number of updates. The portfolios performance in this study outperform the benchmark portfolios at 600,000 updates. In addition, for the reinforcement learning, training the model with different reward setting has a greater impact on performance than increasing the number of updates.

Keywords: Portfolio, Reinforcement Learning, Black-Litterman Model, Proximal Policy Optimization

DOI:10.6814/NCCU202200681

目 次

DOI:10.6814/NCCU202200681

表次

DOI:10.6814/NCCU202200681

圖次

DOI:10.6814/NCCU202200681

第一章　緒論

第一節　研究背景與動機

近年來，隨著資訊科技的高速發展，電腦擁有更為強大的運算能力，對於人工智慧的發展提供良好的條件，在各領域都有著蓬勃的發展這股浪潮也對金融領域帶來許多創新，開始有許多新技術新議題與金融領域結合，透過資訊領域中的各種演算法，像是機器學習、深度學習和強化學習等，透過歷史的資訊防範違約或者預測股價，藉由這些技術為金融領域帶來更多的可能性，也解決過去單純人力無法達成的問題。

資產配置這個議題，在財務領域中一直是一個值得探討的問題，該如何在瞬息萬變的金融市場決定出資產權重，是一件非常困難的事，過去也有非常多文獻探討此問題，得益於科技的快速發展，資產配置也利用相關技術，透過機器找出歷史資料與未來趨勢的關係決定各資產的權重，像有文獻使用強化學習的演算法來配置權重，但這種方法往往只會有輸入資料跟輸出結果，我們無法解釋其中的運作原理，故本文結合傳統資產配置模型與強化學習，透過強化學習決定漲跌幅，取代傳統理論中難以決定的主觀觀點，並解釋其具有的經濟意涵。

Markowitz (1952) 透過平均數—變異數模型(Mean-Variance Model) 提出現代投資組合理論(Modern Portfolio Theory)，投資人可透過此理論建構出效率前緣(Efficient Frontier)，再根據不同的風險偏好決定資產權重，奠定了當代其他投資組合理論的基礎，後續許多經濟學家所提出的理論皆以此為延伸。Black and Litterman (1992)在貝式(Bayes) 框架下嘗試解決 Markowitz 模型中的問題，提出了加入投資人對資產主觀看法的 Black-Litterman 模型，較能符合實際情形。但投資人的看法往往不容易決定，根據市場表現每個人都有不同的觀點，而在此理論中這個觀點對於權重有著明顯影響，進一步改變投資組合績效。

DOI:10.6814/NCCU202200681

本研究利用強化學習，依據各資產的市場變化調整預期漲跌幅，形成一組完全透過機器決定的觀點，強化學習透過獎勵的設計，讓模型學習在不同環境下，應採用何種決策，此決策不會影響到下一期的環境，可視為單期決策過程，不同的獎勵設定就會有不同的決策結果。先前的文獻中，強化學習有不少成功的案例，Moody and Saffell (2001) 利用強化學習決定標普 500 指數該買進或賣出，顯示出由演算法決定的策略比起買進並持有具有更好的績效。因此，本研究利用強化學習，決定下一期的預測值，進而決定出投資觀點帶入 Black-Litterman 模型中，相較於傳統上由投資人決定的方法，更具有一致性，期望建構出的投資組合能夠擁有更良好的績效表現。

第二節　研究目的

本文研究目的主要探討透過各資產的市場價格變化，利用強化學習動態決定不同時期的預測漲跌幅，再依據此預測形成投資觀點，進而建構投資組合。在選擇資產方面，為了有效分散風險，故選擇五檔不同資產類別的美國 ETF 作為基礎資產，採用資產的價量資料作為輸入變數，預測資產價格未來走勢。本研究採用 Black-Litterman 模型建構投資組合，將強化學習設定不同的獎勵值、不同訓練次數所得到的預測結果作為模型中的投資者觀點，並根據不同的風險趨避程度建立投資組合，期望建構之投資組合績效能夠優於市值加權投資組合、等值加權投資組合及風險中立投資組合。

本文章節安排如下：第二章為文獻回顧，敘述過往的投資組合理論以及強化學習在投資領域的應用；第三章為研究方法，介紹 Black-Litterman 模型與強化學習，以及本文的研究方法與策略；第四章為實證分析，說明如何應用強化學習以及介紹所使用商品及參數選擇，並分析實證結果；第五章為結論與建議，總結本研究結果及未來改進方向。

DOI:10.6814/NCCU202200681

第二章　文獻回顧

第一節　投資組合理論

Markowitz (1952) 提出現代投資組合理論，利用平均數—變異數模型進行資產配置，說明投資組合的報酬率和風險之間具有抵換關係，投資人無法在追求高報酬的同時要求投資組合存在低的波動度，此理論奠定投資理論的基礎。Treynor (1961)、Sharpe (1964)、Lintner (1965) 及 Black (1972) 在投資組合理論的基礎上發展出資本資產定價模型(Capital Asset Pricing Model, CAPM)，說明資產的風險及預期報酬率具有線性關係，認為能夠藉由增加投資組合內的標的降低非系統性風險。Black and Litterman (1992) 結合 Markowitz 模型及 CAPM 在貝式的架構下發展出 Black-Litterman 模型，使投資者能夠加入主觀的觀點，並且獲得相對應的投資權重。Donthireddy (2018) 透過隨機森林演算法預測資產報酬率，並將結果結合 Black-Litterman 模型，得到投資組合權重。

第二節　強化學習之投資領域應用

近年來，在科技快速發展下使得機器運算能力大幅提升，帶動強化學習在財務金融領域應用有著顯著發展，最早 Neumeier (1998) 使用強化學習中 Q 學習(Q-Learning) 演算法，由機器判斷德國股價指數 DAX 買進時機，研究結果發現此策略績效優於買進並持有的策略。Moody and Saffell (2001) 認為獎勵除了考慮投資組合報酬外，也要考慮到投資組合的風險，分別使用強化學習兩大分支的不同演算法，基於價值(Value-Based)中的 Q 學習及基於策略(Policy-Based)中的直接強化學習 (Direct Reinforcement)，決定標普 500 買進時機，發現兩種方法績效皆優於買進並持有的策略，除此之外，還發現基於策略的直接強化學習優於基於價值的 Q 學習，原因在於基於策略的演算法較能即時調整模型，資產價格走勢較雜亂，使用基於價值的演算法容易受到雜訊影響。

DOI:10.6814/NCCU202200681

除了以上單純由機器判斷買進或賣出資產，強化學習還能直接決定出投資組合中各資產權重，Jiand and Liang (2017) 使用基於策略中的確定性策略梯度 (Deterministic Policy Gradient, DPG)，透過投資組合中 12 檔加密貨幣資產開高低收價決定出權重。Zhang et al. (2020) 使用基於策的策略梯度 (Direct Policy Gradient)，提出將卷積神經網路 (Convolutional Neural Networks, CNN)和長短期記憶神經網路(Long Short-Term memory, LSTM) 結合而成的投資組合策略網路 (Portfolio Policy Network, PPN)，透過兩種不同的深度學習演算法找出資產價格之特徵，此外，將獎勵函數設為成本敏感獎勵 (Cost-sensitive Reward)，除了投資組合報酬率以外還考慮風險及周轉率，藉此降低轉換投資組合權重造成的手續費，故此模型擁有不錯的績效。

總結以上文獻，Black-Litterman 模型讓投資人能夠加入對資產主觀觀點，但觀點的形成往往不太容易獲得，因此 Donthireddy (2018) 以機器學習所預測的結果作為投資者觀點，進而建立 Black-Litterman 投資組合。近年來強化學習在資產配置方面有許多應用，Moody and Saffell (2001) 研究中發現基於策略的方法較適用於資產配置。因此本研究嘗試將以上這些研究方法做結合，利用基於策略的演算法預測資產報酬率作為投資者觀點，並建構 Black-Litterman 投資組合，觀察由此種方法建立的投資組合是否能夠獲得優於以往得績效表現。本文使用 Schulman et al.(2017) 提出的近端策略優化 (Proximal Policy Optimization, PPO)，傳統基於策略方法在收集決策過程中需要花費過多的時間，PPO 在保有策略梯度的精神下，能夠更有效率，花費更少時間，故本文採用 PPO 演算法。

DOI:10.6814/NCCU202200681

第三章　研究方法

第一節　Black-Litterman 模型

Black-Litterman 模型在 Markowitz 模型的基礎下進行修正，利用貝式機率的觀念，將市場均衡報酬作為事前分配(prior distribution)，結合投資人對資產的預期觀點後，形成事後分配(posterior distribution)，再根據事後分配進行最佳化配置，獲得投資組合中各資產的權重，此模型最大的優點在於能夠加入投資者的主觀觀點進行評估。

3.1.1 事前分配

假定市場資產其超額報酬（資產報酬率減去無風險利率）皆服從常態分配，若此時有 N 項資產，則服從以下多元常態分配：

$$R \sim N(\mu, \Sigma) \tag{3.1}$$

其中，R：超額報酬，為 $N \times 1$ 的向量；μ：期望超額報酬，為 $N \times 1$ 的向量；Σ：共變異數矩陣，為 $N \times N$ 的矩陣

Black-Litterman 模型以市場均衡超額報酬為出發點，若市場是有效率的，那最佳投資組合就會是透過市值加權的投資組合，此時我們就可以利用反向最適化(Reverse Optimization) 透過共變異數矩陣及市值權重計算隱含市場均衡超額報酬，並獲得期望超額報酬的機率分配如下：

$$\Pi = \delta \Sigma w_{mkt} \tag{3.2}$$

$$\mu \sim N(\Pi, \tau\Sigma) \tag{3.3}$$

DOI:10.6814/NCCU202200681

其中，Π：市場均衡超額報酬，為 N × 1 的向量；δ：風險趨避係數，係數越大則代表對風險的厭惡程度越高；w_{mkt}：市值加權權重，為 N × 1 的向量；τ：調整因子，代表估計值的不確定性，若不存在誤差則$\tau = 0$。本文參考 Meucci (2010) 中設定 $\tau = \frac{1}{T}$，T 為樣本數。

3.1.2 投資者觀點

假設投資者對於 N 項資產具有 K 個相互獨立的主觀看法（K ≤ N），則投資者根據其主觀看法建立一個 K × N 的 P 矩陣代表觀點中的資產權重，另外搭配一個 K × 1 的 Q 向量代表觀點值：

$$P = \begin{bmatrix} V_{11} & \cdots & V_{1N} \\ \vdots & \ddots & \vdots \\ V_{K1} & \cdots & V_{KN} \end{bmatrix}, \quad Q = \begin{bmatrix} q_1 \\ \vdots \\ q_K \end{bmatrix}$$

其中，V_{kn}：投資者對於第 n 個資產第 k 個主觀觀點，$n = 1,2,\cdots,N$，$k = 1,2,\cdots,K$；q_k：第 k 個觀點值，$k = 1,2,\cdots,K$

但在現實中投資者對於觀點並不是完全確定，因此會加入一個誤差項 v，假設 v 服從常態分配，用來代表觀點的不確定性，v 為 K × 1 的向量，這些觀點將會滿足下列式子：

$$P\mu = Q + v, \quad v \sim N(0, \Omega) \tag{3.4}$$

$$where \quad \Omega = \begin{bmatrix} \omega_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \omega_{KK} \end{bmatrix}$$

其中，Ω：觀點的不確定程度，為 K × K 對角矩陣；ω_{kk}：第 k 個觀點的不確定程度，$k = 1,2,\cdots,K$。

舉例來說：若市場中存在四種資產，投資者有以下兩個主觀看法：

DOI:10.6814/NCCU202200681

1. 看法一是認為資產三將有 10%的超額報酬；

2. 看法二是認為資產一的超額報酬較資產二多出 2%，則：

$$P = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & -1 \end{bmatrix}, \quad Q = \begin{bmatrix} 10\% \\ 2\% \end{bmatrix}$$

透過以上例子可以看出，投資者得觀點並不僅限於絕對觀點，資產間的相對觀點也是可以被接受的。本研究採用質性觀點，參考 Meucci (2010) 將投資者的主觀看法透過各資產的波動度設定，並參考 He and Litterman (1999) 中將觀點不確定性假設與市場波動幅度相同，滿足以下式子：

$$Q(k) = (P\Pi)(k) + \eta_k \sqrt{(P\Sigma P')(k,k)}, \quad k = 1,2,\cdots,k \tag{3.5}$$

$$where \quad \Omega = diag(\tau P\Sigma P')$$

其中，P 為 N × N 的單位矩陣；$\eta_k \in \{-2,-1,1,2\}$，分別代表預期資產會大跌、小跌、小漲及大漲。

3.1.3 事後分配

將事前分配加上投資者的主觀看法，透過貝式方法即可求出期望超額報酬的事後分配為：

$$\mu = N(\mu_{BL}, \Sigma_{BL}^{\mu}) \tag{3.6}$$

$$where \quad \mu_{BL} = \Pi + \tau\Sigma P'(\tau P\Sigma \mathrm{P}' + \Omega)^{-1}(Q - P\Pi)$$

$$\Sigma_{BL}^{\mu} = ((\tau\Sigma)^{-1} + P'\Omega^{-1}P)^{-1}$$

透過上式可以看出 μ_{BL} 可以分為兩項，第一項 Π 為市場均衡超額報酬，第二項為 Q − PΠ 的倍數，當投資人的觀點與市場觀點相同時Q − PΠ 會等於零。可以將資產超額報酬分配寫作：

DOI:10.6814/NCCU202200681

$$R = \mu + \varepsilon, \quad \varepsilon \sim N(0, \Sigma) \tag{3.7}$$

假設 μ 跟 ε 兩項互相獨立，則可以將超額報酬的事後分配改寫成：

$$R \sim N(\mu_{BL}, \Sigma_{BL}), \quad \Sigma_{BL} = \Sigma + \Sigma_{BL}^{\mu} \tag{3.8}$$

當有了 μ_{BL} 與 Σ_{BL} 以後，即可通過最適化模型來獲得最後的 Black-Litterman 投資組合權重：

$$w_{BL}^{*} = (\delta \Sigma_{BL})^{-1} \mu_{BL} \tag{3.9}$$

第二節　強化學習

本節中將會介紹強化學習的基本架構，說明強化學習的運作方式，接著介紹基於策略方法中的傳統策略梯度，最後介紹本文所使用之 PPO 演算法，詳細說明此演算法的原理，並介紹相較於傳統方法的改善。

3.2.1. 基本架構

如圖 3.1 所示，要描述一個強化學習的過程，必須要有一個代表做決策的主體也就是代理人 (Agent)，那與代理人互動的就是環境 (Environment)，當第 t 期時，環境傳給代理人當下做決策所需的資訊或觀察值，稱為狀態 s_t，代理人收到以後，會根據設定好的策略 (Policy) 做出相對應的動作 a_t，這個策略可以是任意函數 $\pi_\theta(a_t|s_t)$，可以看出策略函數是以 θ 為參數之函數，在做出動作以後，環境會給出一個相對應的獎勵 r_t，透過不斷循環上述的過程，直到決策結束，通常決策結束代表達到所設定之次數或者達到設定目標。我們可以將整個決策過程(Trajectory) 表示為 $\tau^i = \{s_1^i, a_1^i, s_2^i, a_2^i, \cdots, s_T^i, a_T^i\}$，其中 T 代表共

DOI:10.6814/NCCU202200681

有 T 期決策，而上標則為第 i 次決策過程。

圖 3.1：強化學習基本架構

當執行完決策過程 τ^i 後，會得到一連串的獎勵值，將所有值相加即為總獎勵 $R(\tau^i) = \sum_{t=1}^{T} r_t^i$，而基於策略的強化學習目的就是透過調整策略函數 π_θ 中的參數 θ ，達到總獎勵越大越好，找到最佳策略的過程必須透過演算法，下一小節中會介紹基於策略方法中的策略梯度(Policy Gradient) 演算法。

3.2.2. 策略梯度

透過強化學習的基本架構，我們知道基於策略之強化學習需藉由調整策略函數 π_θ 中的參數 θ，使得執行完決策過程 τ^i 後，所得到總獎勵$R(\tau^i) = \sum_{t=1}^{T} r_t^i$ 能夠越大越好，假設共有 N 次決策過程，每次決策過程中有 T 期，因此目標函數為極大化期望總獎勵，可表示為下式：

$$Max\ E_{\tau \sim p_\theta(\tau)}[R(\tau)] = \sum_{i=1}^{N} R(\tau^i) p_\theta(\tau^i) \qquad (3.10)$$

其中

DOI:10.6814/NCCU202200681

$$p_\theta(\tau^i) = p(s_1^i)p_\theta(a_1^i|s_1^i)p(s_2^i|s_1^i,a_1^i)p_\theta(a_2^i|s_2^i)p(s_3^i|s_2^i,a_2^i)\cdots$$
$$= p(s_1^i)\prod_{t=1}^{T} p_\theta(a_t^i|s_t^i)p(s_{t+1}^i|s_t^i,a_t^i) \quad (3.11)$$

如 (3. 10) 式所示，期望總獎勵為經過每一決策過程 τ^i 所得到總獎勵 $R(\tau^i)$ 乘上其發生的機率 $p_\theta(\tau^i)$ 相加以後就可以獲得。前一小節中我們知道決策過程可表示成 $\tau^i = \{s_1^i, a_1^i, s_2^i, a_2^i, \cdots, s_T^i, a_T^i\}$，在這過程中前面的動作會影響到後面的環境，因此可以將 τ^i 發生的機率 $p_\theta(\tau^i)$ 看成兩個部分，其中 $p_\theta(a_t^i|s_t^i)$ 為受到策略參數 θ 控制，在 t 時點環境的狀態為 s_t^i 時，代理人選擇執行動作 a_t^i 的機率，而另一部份 $p(s_{t+1}^i|s_t^i,a_t^i)$，為受環境影響不可藉由改變參數 θ 來控制，也就是在給定 t 時點環境狀態 s_t^i 且執行動作 a_t^i 下，$t+1$ 發生狀態 s_{t+1}^i 的機率。將這些從決策開始到結束時發生的全部機率相乘，即可獲得 $p_\theta(\tau^i)$，如 (3. 11) 所示。

前面說到要極大化期望總獎勵，和傳統機器學習極小化變異數的方法剛好相反，需透過梯度上升法(Gradient Ascent) 來更新參數 θ，設定學習率為 η 時，更新參數的方式如下：

$$\theta \leftarrow \theta + \eta \nabla E_{\tau\sim p_\theta(\tau)}[R(\tau)] \quad (3.12)$$

透過 (3. 12) 的方式不斷更新 θ 找出最優策略，因此須先計算目標函數之梯度，計算結果如下：

$$\nabla E_{\tau\sim p_\theta(\tau)}[R(\tau)] \approx \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T_i} R(\tau^i)\nabla log\,\{p_\theta(a_t^i|s_t^i)\} \quad (3.13)$$

根據 (3. 12) 及 (3. 13) 式可知，更新的方向由總獎勵 $R(\tau^i)$ 決定，當總獎勵 $R(\tau^i)$ 為正時，代表此決策過程 τ^i 比較好，往 $\nabla log\,\{p_\theta(a_t^i|s_t^i)\}$ 的方向更新參數，也就是增加在狀態 s_t^i 下執行 a_t^i 的機率；當總獎勵 $R(\tau^i)$ 為負時，代表此決

DOI:10.6814/NCCU202200681

策過程 τ^i 比較差，就會朝著 $\nabla log\{p_\theta(a_t^i|s_t^i)\}$ 的反方向更新，也就是減少在狀態 s_t^i 下執行 a_t^i 的機率。但透過總獎勵 $R(\tau^i)$ 來更新參數會產生問題，當總獎勵是正的，並不代表這個決策過程中每一個動作皆是好的；當總獎勵是負的，也並不代表這個決策過程中每一個動作皆是差的，因此需要透過不同的方式，來衡量決策過程中個別動作的貢獻。

使用總獎勵 $R(\tau^i)$ 來更新參數有上述問題，因此在強化學習會採用優勢函數(Advantage Function) 來取代總獎勵，而優勢函數 $A_t^\theta(s_t, a_t)$ 能夠更有效的衡量個別動作的貢獻，假設 γ 為介於 0 到 1 的折現因子(Discount Factor)，優勢函數如下：

$$A_t^\theta(s_t^i, a_t^i) = \sum_{t'=t}^{T_i} \gamma^{t'-t} r_{t'}^i - V(s_t) \tag{3.14}$$

$$where \quad V(s_t) = E[\sum_{t'=t}^{T_i} \gamma^{t'-t} r_{t'}^i \,|s = s_t^i]$$

根據 (3. 14) 式可以看出，優勢函數由兩個部分所構成，分別是 $\sum_{t'=t}^{T_i} \gamma^{t'-t} r_{t'}^i$ 及 $V(s_t) = E\left[\sum_{t'=t}^{T_i} \gamma^{t'-t} r_{t'}^i \,\middle|s = s_t^i\right]$，第一部分的 $\sum_{t'=t}^{T_i} \gamma^{t'-t} r_{t'}^i$ 與總獎勵不同的地方在於只計算執行完動作 a_t^i 之後獲得的獎勵，因為過去的獎勵與 t 時點才執行的動作無關，另一個不同的地方則是新增折現因子 γ，距離現在執行動作 a_t^i 越遠的獎勵照理來說受到此動作的影響越小，因此透過介於 0 到 1 的折現因子即可將未來的獎勵折現到現在。優勢函數的第二部分 $V(s_t)$，則為某一個狀態 s_t^i 下，所預期未來獲得的獎勵折現，這項主要用來判斷當執行動作 a_t^i 得到的未來獎勵是否優於平均。

使用優勢函數取代前面的總獎勵 $R(\tau^i)$，即可解決上述說到使用總獎勵產生的問題，因此將期望總獎勵梯度 (3. 13) 改寫如下：

DOI:10.6814/NCCU202200681

$$\nabla E_{\tau\sim p_\theta(\tau)}[R(\tau)] \approx \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T_i} A_t^\theta(s_t^i, a_t^i)\nabla log\,\{p_\theta(a_t^i|s_t^i)\} \quad (3.15)$$

與 (3.13) 式類似，改進後的梯度函數 (3.15) 由優勢函數 $A_t^\theta(s_t^i, a_t^i)$ 決定更新的方向，當 $A_t^\theta(s_t^i, a_t^i)$ 為正時，代表在狀態 s_t^i 下執行動作 a_t^i 所獲得的未來獎勵折現後大於平均值，故往 $\nabla log\,\{p_\theta(a_t^i|s_t^i)\}$ 的方向更新參數，也就是增加 $p_\theta(a_t^i|s_t^i)$ 的機率；相反地，當 $A_t^\theta(s_t^i, a_t^i)$ 為負時，則往 $\nabla log\,\{p_\theta(a_t^i|s_t^i)\}$ 的反方向更新參數，也就是減少 $p_\theta(a_t^i|s_t^i)$ 的機率。結合 (3.12) 與 (3.15)式，梯度上升法更新參數 θ 可以用下列式子表示，其中學習率為 η ；

$$\theta \leftarrow \theta + \eta \times \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T_i} A_t^\theta(s_t^i, a_t^i)\nabla log\,\{p_\theta(a_t^i|s_t^i)\} \quad (3.16)$$

3.2.3. 近端策略優化(PPO)

上一小節中介紹關於傳統的策略梯度演算法，在給定參數 θ 下，透過策略 π_θ 與環境互動以後獲得決策過程，接著使用梯度上升法來更新參數，此種方法每更新一次參數就必須要重新與環境互動，導致花費大量的時間收集決策過程，因此 PPO 演算法目的為了改善傳統策略梯度的問題，使收集到的決策過程能夠重複使用，增加更新參數的效率，並降低演算法運行所需的時間。

為解決上述問題，PPO 與傳統策略梯度不同的地方在於採用 Off-Policy 的方式來更新參數，意思就是更新參數 θ 時，PPO 並不是使用策略 π_θ 所收集的決策過程，而是改用另外一個策略 $\pi_{\theta'}$ 收集的決策過程來更新參數 θ ，故可以解決傳統策略梯度效率不足的問題。但因收集決策過程改變， θ 與 θ' 還是存在著一定程度的差異，這之間的轉換需要應用重要性採樣(Importance Sampling)的概念，因此本小節中介紹 PPO 運作方法前，先介紹如何透過重要性採樣將收集資料的過程改變。

DOI:10.6814/NCCU202200681

假設存在兩個不同的分配 p 及 q，且有一個函數 $f(x)$，若無法透過 p 分配採樣只能透過 q 分配的話，要計算 $f(x)$ 在 p 分配下的期望值，則滿足下式：

$$E_{x\sim p}[f(x)] = \int f(x)p(x)dx = \int f(x)\frac{p(x)}{q(x)}q(x)dx = E_{x\sim q}\left[f(x)\frac{p(x)}{q(x)}\right] \quad (3.17)$$

因為是透過不同的分配來採樣，故需要乘上一個修正項 $\frac{p(x)}{q(x)}$ 以彌補兩個分配之間的差異，以上這個過程即為重要性採樣，透過 (3.17) 式改寫 (3.15) 為：

$$\nabla E_{\tau\sim p_\theta(\tau)}[R(\tau)] = E_{\tau\sim p_{\theta'}(\tau)}\left[\frac{p_\theta(a_t|s_t)}{p_{\theta'}(a_t|s_t)}A_t^{\theta'}(s_t,a_t)\nabla log p_\theta(\tau)\right] \quad (3.18)$$

即可透過 (3.18) 反推出透過 $\pi_{\theta'}$ 更新參數 θ 之目標函數 $J^{\theta'}(\theta)$，如下式：

$$J^{\theta'}(\theta) = E_{\tau\sim p_{\theta'}(\tau)}\left[\frac{p_\theta(a_t|s_t)}{p_{\theta'}(a_t|s_t)}A_t^{\theta'}(s_t,a_t)\right] \quad (3.19)$$

根據 (3.17) 式來看，理論上使用重要性採樣可以將 $p_\theta(\tau)$ 換成任意的 $p_{\theta'}(\tau)$，但在實作上兩者還是不能相差太多，雖然兩者的期望值相等，但當差距變大時，兩者的變異數也會隨之變大，故在 (3.19) 目標函數中加入限制式來避免，如下式：

$$\begin{aligned} J^{\theta'}(\theta) = E_{\tau\sim p_{\theta'}(\tau)}[\min(&\frac{p_\theta(a_t|s_t)}{p_{\theta'}(a_t|s_t)}A_t^{\theta'}(s_t,a_t), \\ &clip(\frac{p_\theta(a_t|s_t)}{p_{\theta'}(a_t|s_t)}, 1+\varepsilon, 1-\varepsilon)A_t^{\theta'}(s_t,a_t)] \end{aligned} \quad (3.20)$$

其中加入 $clip\left(\frac{p_\theta(a_t|s_t)}{p_{\theta'}(a_t|s_t)}, 1+\varepsilon, 1-\varepsilon\right)$ 即為避免兩者差距太大，$clip$ 函數意思為當 $\frac{p_\theta(a_t|s_t)}{p_{\theta'}(a_t|s_t)}$ 介於 $1+\varepsilon$ 和 $1-\varepsilon$ 之間時，則為 $\frac{p_\theta(a_t|s_t)}{p_{\theta'}(a_t|s_t)}$，而當大於 $1+\varepsilon$ 為 $1+\varepsilon$，小於 $1-\varepsilon$ 為 $1-\varepsilon$。可以將 (3.20) 式看成兩個部分，當 $A_t^{\theta'}(s_t,a_t)$ 為正時

DOI:10.6814/NCCU202200681

代表狀態 s_t 下執行 a_t 的表現是好的，因此增加其機率 $p_\theta(a_t|s_t)$，但因為加上限制條件，若 $\frac{p_\theta(a_t|s_t)}{p_{\theta'}(a_t|s_t)}$ 增加到超過 $1+\varepsilon$，則限制為 $1+\varepsilon$；反之，若 $A_t^{\theta'}(s_t,a_t)$ 為負時代表狀態 s_t 下執行 a_t 的表現是差的，因此降低其機率 $p_\theta(a_t|s_t)$，若 $\frac{p_\theta(a_t|s_t)}{p_{\theta'}(a_t|s_t)}$ 減少到低於 $1-\varepsilon$，則限制為 $1-\varepsilon$，透過在目標函數加上限制即可讓參數更新時差距不會更進一步擴大，如圖 3.2 所示。

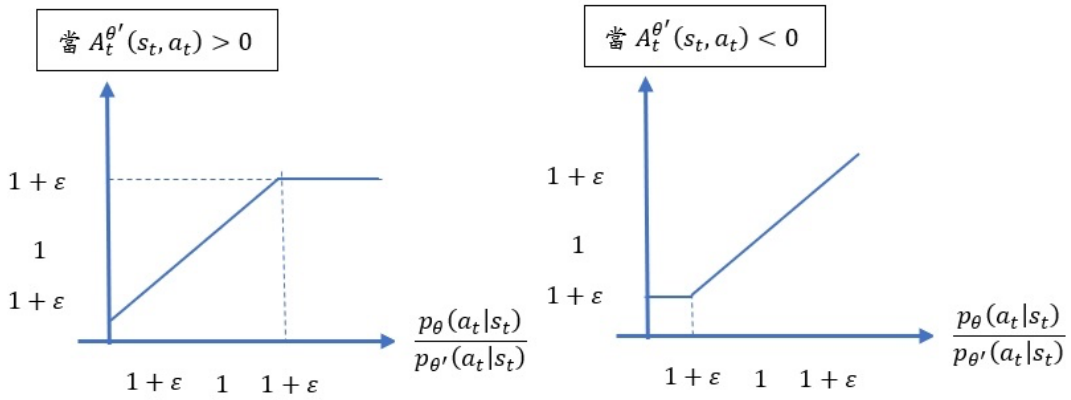

圖 3.2：目標函數圖解

而 PPO 演算法中額外新增兩條目標式增進演算法的效果，如下式：

$$\max J_{PPO}^{\theta'}(\theta) = E_{\tau\sim p_{\theta'}(\tau)}[J_{CLIP}^{\theta'}(\theta) - c_1 J_{VF}^{\theta'}(\theta) + c_2 H^{\theta'}(s_t)] \qquad (3.21)$$

其中：

$$J_{VF}^{\theta'}(\theta) = \left[\sum_{t'=t}^{\mathrm{T}} \gamma^{t'-t} r_{t'}^{i} + \gamma^{\mathrm{T}-\mathrm{t}}\mathrm{V}(\mathrm{s_T}) - \mathrm{V}(\mathrm{s_T})\right]^2 \qquad (3.22)$$

$$\mathrm{H}^{\theta'}(s_t) = -\sum p_{\theta'}(a_t|s_t)\log p_{\theta'}(a_t|s_t) \qquad (3.23)$$

目標式 (3.22) 為均方誤差(Mean Square Error)，目的是使估計出的 $\mathrm{V}(\mathrm{s_t})$ 能夠越來越準，誤差越小越好；而目標式 (3.23) 為執行不同動作的熵(Entropy)，若是熵越大，則代表越能夠選擇到不同的動作，增加此限制式能幫助最初選擇到不

DOI:10.6814/NCCU202200681

同的動作。且設置 c_1 和 c_2 兩個超參數調整權重，因此 PPO 之目標式則為 (3.21) 所示。

第三節　投資策略

根據以下幾種不同的條件限制建立相應的投資組合：

4.3.1. 極大化夏普比率

極大化投資者額外承擔每一單位風險所獲得的額外收益：

$$max \frac{w'_{BL,t}\mu_{BL,t}}{\sqrt{\delta w'_{BL,t}\Sigma w_{BL,t}}}$$
$$s.t. \quad w'_{BL,t}\mathbf{1} = 1, \quad 0 \leq w_{BL,t} \leq 1$$

4.3.2. 基準投資組合

本研究以市值加權投資組合、等值加權投資組合以及風險平價投資組合作為比較基準。市值加權投資組合即依照基礎資產的市值作為分配權中的依據；等值加權投資組合則是將每種基礎資產的權重都設為相等；而風險平價投資組合則根據資產波動度建立投資組合，使各資產承受相同風險。

DOI:10.6814/NCCU202200681

第四章　實證分析

本研究參考 Donthireddy (2018)，以美國市場中流動性較高且不同資產類別的五檔 ETF 做為建構投資組合的基礎資產，根據價量資料透過強化學習由不同獎勵設定及運行次數，分別判斷各資產未來 30 天的漲跌幅，將決定好的結果代入 Black-Litterman 模型中，得到各成分股權重，即為本研究之投資組合。

本章節的架構上，第一節先介紹資料來源與前處理，第二節介紹本文之研究方法，說明基本架構、模型設定及研究流程，最後第三節則是分析研究結果。

第一節　資料來源與前處理

本研究參考 Donthireddy (2018)，其商品代號、名稱及代表資產類別如表 4.1 所示：

表 4.1 基礎資產商品代號、名稱及代表資產類別

商品代號	名稱	代表資產類別
EEM	iShares MSCI Emerging Markets ETF	新興市場
EFA	iShares MSCI EAFE ETF	已開發市場
GLD	SPDR Gold Shares ETF	黃金
IYR	iShares US Real Rstate ETF	美國房地產
TLT	iShares 20+ Year Treasury Bond ETF	美國長期公債

本研究採用資料以日為單位，採用期間為 2006 年 1 月 3 號至 2021 年 11 月 26，所有資料皆取自 Yahoo Finance。

表 4.2 基礎資產相關係數矩陣

	EEM	EFA	GLD	IYR	TLT
EEM	1.00	0.62	0.54	0.42	0.27
EFA	0.62	1.00	-0.03	0.80	0.31
GLD	0.54	-0.03	1.00	0.23	0.62
IYR	0.42	0.80	0.23	1.00	0.72
TLT	0.27	0.31	0.62	0.72	1.00

DOI:10.6814/NCCU202200681

由表 4.2 可以發現，黃金類型 GLD 與已開發市場 EFA 及美國房地產 IYR 有較低的相關性，甚至對於 EFA 為負相關；而長期公債類的 TLT 與股票市場的 EEM 及 EFA 也有較低的相關性，因此我們可以預期利用這五檔 ETF 建構出的投資組合能夠具有風險分散的效果

第二節　強化學習之應用

本節主要說明如何使用強化學習，將市場的價量資料與主觀觀點做連結，在第一小節，會先定義本研究中強化學習的基本架構，如狀態、動作和獎勵，第二小節中介紹本研究所使用的四種不同獎勵設定，第三小節介紹模型中神經網路如何建構和參數選擇，最後第三小節說明研究流程及應用方式。

4.2.1. 基本架構

強化學習主要由三個基本元素構成，為狀態、動作和獎勵，本研究中，狀態 $s_t = [X_{t-30}, X_{t-29}, \cdots, X_{t-1}]$，$s_t$ 為資產過去 30 天的價量資料，其中 $X_t = \left[P_t^{(C)}, Y_t, O_t, H_t, L_t, S_t, V_t\right]$ 包含七種不同的價量資料，令 $P_t^{(O)}, P_t^{(H)}, P_t^{(L)}, P_t^{(C)}$ 分別為第 t 天資產的開高低收價，而 v_t 為第 t 天的成交量，透過價量資料定義以上變數，$Y_t = \frac{P_t^{(C)} - P_{t-1}^{(C)}}{P_t^{(C)}}$ 為第 t 天之報酬率，$O_t = \frac{P_t^{(O)}}{P_t^{(C)}}$，$H_t = \frac{P_t^{(H)}}{P_t^{(C)}}$，$L_t = \frac{P_t^{(L)}}{P_t^{(C)}}$ 為第 t 天的開盤價、最高價及最低價佔當天收盤價的比例，透過此一設定，呈現出整天的價個波動，S_t 為資產過去五天之標準差，$V_t = \frac{v_t - v_{t-1}}{v_t}$ 為第 t 天成交量的變化，藉由資產過去 30 天的價量變數，試圖找出未來 30 天資產漲跌幅。

當代理人接收到觀察值也就是上述市場價量資料後決定出對應的動作，根據 (3.5) 式的方法，機器需判斷出未來 30 天資產為大跌、小跌、小漲或大漲，因此將動作的數量設為 4，而動作空間則對應到 (3.5) 式中的 η_k，故動作空間 $a_t = [-2, -1, 1, 2]$，當決定好動作 a_t 後，依據所設定的獎勵方式給與對應

DOI:10.6814/NCCU202200681

的獎勵，本研究中設置四種不同獎勵方式，並比較獎勵對於績效之影響，四種獎勵設定將於下一小節中介紹。

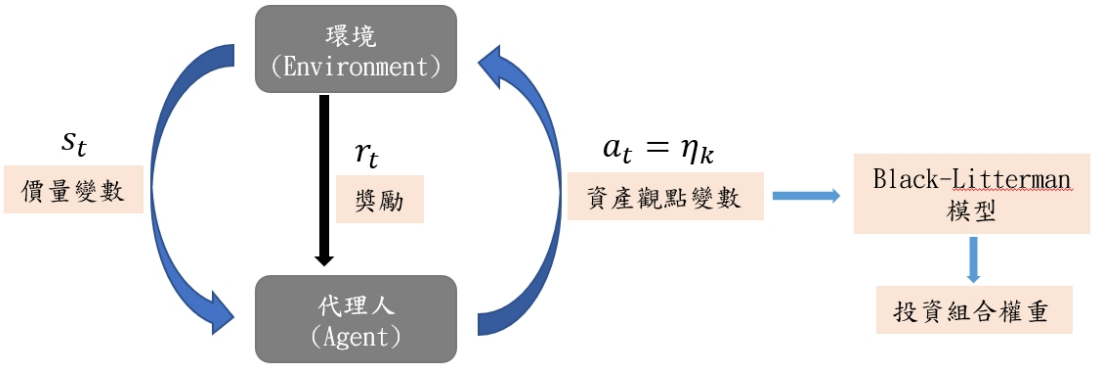

圖 4.1：強化學習應用基本架構

圖 4.1 為本研究的強化學習應應基本架構圖，將對價量資料和資產未來觀點經由強化學習作結合，代理人接收狀態 s_t，判斷未來 30 天資產的漲跌幅，代入觀點的函數以後，使用 Black-Litterman 模型得到投資組合權重，其中根據不同的獎勵設定，觀察獎勵與績效的相關性。

4.2.2. 獎勵設定

強化學習中，獎勵的設定對於代理人來講是很重要的，不管演算法的種類為何，基本上強化學習的訓練方式為極大化所獲得的獎勵，因此在本文中設定四種不同的獎勵，觀察獎勵對於實證結果的影響。本文的強化學習可以看作一個分類問題，透過代理人觀察過去 30 天的狀態預測資產未來 30 天的漲跌幅，將結果分為四類，故獎勵設定大致上是以比較預測值與實際值的差異。詳細獎勵設定如下，其中 Act 代表實際值，Pre 代表預測值：

1. Type I：完全預測正確給予獎勵

第一個獎勵設定最為簡單，設定當預測值符合實際值時，代理人獲得 1 分的獎勵，其餘預測錯誤的部分皆不獲得獎勵，獎勵如下表所示：

DOI:10.6814/NCCU202200681

Pre / Act	-2	-1	1	2
-2	1	0	0	0
-1	0	1	0	0
1	0	0	1	0
2	0	0	0	1

2. Type II：根據兩者差距給予獎勵

第二種獎勵設定方式則為根據兩者的差距來給予代理人獎勵，當差距越小獲得較高的獎勵，而差距越大則獎勵越低，獎勵和預測值及實際值的關係，可以透過下列數學式表示：

$$reward = -(Pre - Act)^2 \tag{4. 1}$$

將 (4. 1) 式轉換為表格型式，即為下表所示：

Pre / Act	-2	-1	1	2
-2	0	-1	-9	-16
-1	-1	0	-4	-9
1	-9	-4	0	-1
2	-16	-9	-1	0

透過以上表格可以看出，當預測值與實際值差距越大時，給予越大的懲罰，也又是給予負的獎勵值。

3. Type III：自行獎勵設定

第三種獎勵設定方式則是依據金融上的知識給予不同獎勵，此獎勵設計理念為完全正確預測給予 5 分，預測方向正確但漲跌幅度不同給予 3 分，預

DOI:10.6814/NCCU202200681

測方向不同是較難以接受的，因此按照錯誤程度給予懲罰性的獎勵 −3 及 −10 分，如下表所示：

Pre / Act	-2	-1	1	2
-2	5	3	-10	-10
-1	3	5	-3	-10
1	-10	-3	5	3
2	-10	-10	3	5

會這樣設計獎勵值是因為預測未來漲跌並不是一件容易的事，如果能夠完全預測需要給一個正的獎勵值，而如果是預測方向正確但漲跌幅度不同，也對後面資產配置有一定的幫助，因此給予小於精準預測的獎勵值，最不能接受的則是實際值與預測值差距過大，為了避免這情形發生因此給予代理人負的獎勵。

4. Type IV：自訂獎勵（不均衡分類）

一般來說，分類資產未來漲跌趨勢，並不會平均分配，而是小漲及小跌這兩類數目較多，故參考 Lin et al. (2020) 論文中對於不均衡分類(Imbalance Classification) 的設定來決定獎勵函數，如下表所示：

Pre / Act	-2	-1	1	2
-2	1	$-\lambda$	$-\lambda$	-1
-1	-1	λ	$-\lambda$	-1
1	-1	$-\lambda$	λ	-1
2	-1	$-\lambda$	$-\lambda$	1

會這樣設定主要是因為 Lin et al. (2020) 認為當總數比較少的類別被預測正確或錯誤應該給予較大的獎勵或懲罰，而總數比較多的類別被預測正確或

DOI:10.6814/NCCU202200681

錯誤則給予較小的獎勵或懲罰，故設定一個小於 1 的參數 λ，本文參考論文中的設定 $\lambda = \frac{N(\text{大漲U大跌})}{N(\text{小漲U小跌})}$，即為兩種類個數之比例。

4.2.3. 模型參數與設定

強化學習中，代理人進行決策的方式是依據策略函數 $\pi_\theta(a_t|s_t)$，能夠依照不同的需求設定策略函數，而本文使用神經網路，神經網路模型如圖 4.2 所示，其中括號內為神經元個數。神經網路模型輸入值為 s_t，也就是從環境接收到的觀察值，而輸出分為兩個部分，上半部輸出代理人決定的預測值，下半部則輸出優勢函數中的 $V(s_t)$，即為先前所說狀態 s_t 下，執行動作得到的平均獎勵。會選擇神經網路 LSTM 來處理原因在於觀察到的狀態為資產過去 30 天的價量資料，這些資料為時間序列資料，故採用 LSTM 處理，而後續皆採用一般的神經網路 NN，接著使用演算法 PPO，不斷更新神經網路模型中的參數，藉此找出最適模型。

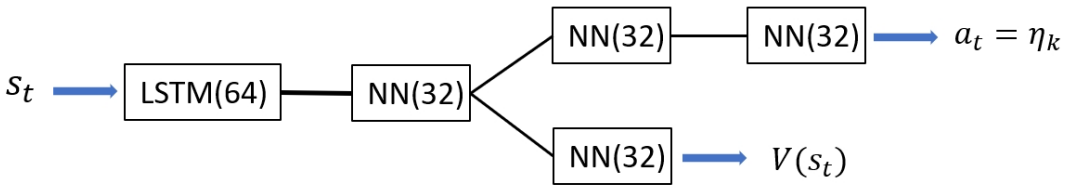

圖 4.2：神經網路模型

模型決定後，設定強化學習所需使用到之超參數，學習率為 0.0005，本文除了研究不同獎勵設定對於績效的影響外，還比較不同更新次數是否會對結果造成很大的影響，因次設定三種不同更新次數，分別是 200000、600000 及 1000000 次。

4.2.4. 研究流程

DOI:10.6814/NCCU202200681

本研究每 30 日重新配置投資組合權重，透過收盤價對收盤價計算報酬率，將資料分為訓練期與測試期，模型先透過訓練期進行訓練，挑選出訓練期中最適模型當作最終模型，再將測試期資料丟進最終模型中決定出預測值，本研究訓練期 2006 年 1 月 3 日至 2016 年 12 月 30 日，而測試期為 2017 年 1 月 3 日至 2021 年 11 月 26 日。第 t 天的決策過程會根據訓練期或測試期有所不同，如圖 4. 3 所示。

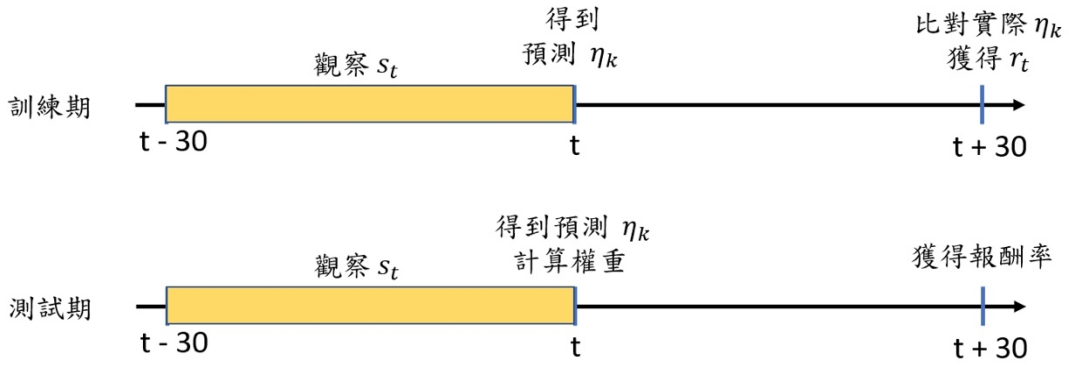

圖 4. 3：第 t 天決策過程

在訓練期時，第 t 天的決策過程為觀察出前 30 天的價量資料作為觀察值 s_t，接著由模型得到預測值 η_k，當$t+30$時比對實際值與預測值的差異得到獎勵 r_t；而測試期時，與訓練期不同的地方在於少了比對個過程，得到各資產預測值後即代入 Black-Litterman 模型計算投資組合權重，而在$t+30$時獲得投資組合報酬率並更新權重。

為使訓練資料量足夠多，切分資料在訓練期時採用移動窗格的方式，將 30 天的價量資料當作一組觀察值，故訓練期共有 2735 組資料；而測試期的目的在於決定投資組合權重，因此不使用移動窗格，而是每 30 天為一組觀察值，故總共只有 40 組資料。

DOI:10.6814/NCCU202200681

第三節　實證結果

本節中將會介紹本研究衡量績效的指標，以及分別觀察基準投資組合、極大化夏普比率投資組合與極小化變異數投資組合的績效表現，最後則是特別觀察更新次數對於績效的影響。

4.3.1. 績效評估

在衡量投資組合績方面，本文使用數個指標以檢測投資組合的績效表現，例如平均年化報酬率、年化標準差、夏普比率（Sharpe Ratio）、賺賠比（Omega Ratio or Gain Ratio）、最大策略虧損（Maximum Drawdown, MDD）、卡瑪比率（Calmar Ratio）。績效指標計算方式如表 4.3 所示。

表 4.3 投資組合績效指標

投資組合績效指標	公式
平均年化報酬率	$\bar{r} = \frac{1}{T}\sum_{t=1}^{T} r_t$
年化標準差	$\sigma = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(r_t - \bar{r})^2}$
夏普比率	$Sharpe = \frac{\bar{r} - r_f}{\sigma}$
賺賠比	$Omega = \frac{\frac{1}{T}\sum_{t=1}^{T}\max(0, r_t)}{\frac{1}{T}\sum_{t=1}^{T}\max(0, -r_t)}$
最大策略虧損	$MDD = \max_{\tau\in(0,T)}(0, \max_{t\in(0,\tau)}(0, -r_{\{t,T\}}))$
卡瑪比率	$Calmar = \frac{\bar{r} - r_f}{MDD}$

後續將透過以上指標評估不同投資策略的績效表現，包括基準投資組合、極大化夏普比率投資組合及極小化變異數投資組合。

DOI:10.6814/NCCU202200681

4.3.2. 基準投資組合

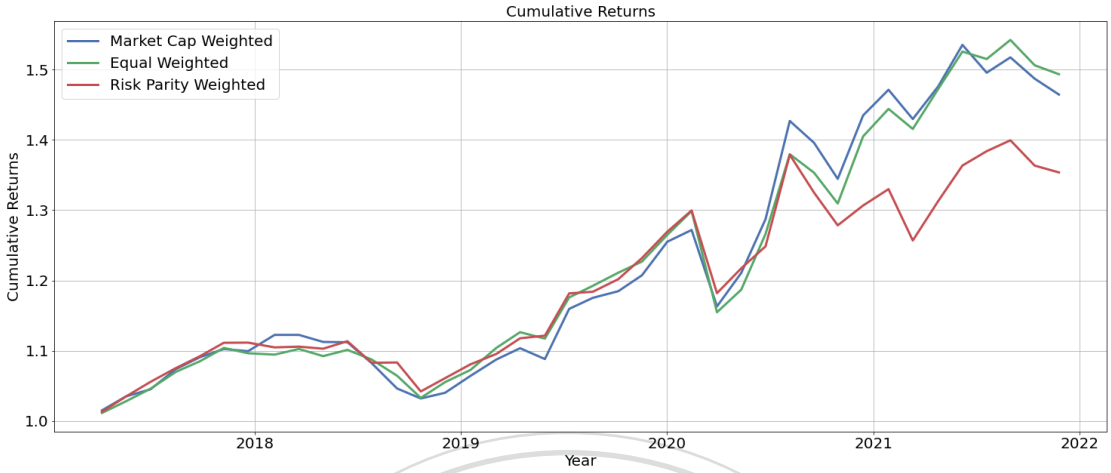

圖 4.4：基準投資組合累積報酬率

基準投資組合在測試期間的累積報酬率如圖 4.4 所示。透過累積報酬率的圖形可以發現三種不同的基準投資組合 2020 年時受到疫情的衝擊下，皆出現的大幅度的回檔，而其中市值加權投資組合與等值加權投資組合擁有差不多的累積報酬率，但風險平價投資組合透過使每個資產的相同波動度的方式分配權重，相較其餘兩種表現較差。基準投資組合績效表現如下表所示：

表 4.4 基準投資組合績效表現

	Mean p.a(%)	SD p.a(%)	Sharpe	Omega	MDD(%)	Calmar
Market Cap-Weighted	12.12	11.78	1.03	1.33	8.55	1.42
Equal Weighted	12.78	11.56	1.11	1.28	11.08	1.15
Risk Parity Weighted	9.50	11.09	0.86	0.66	9.06	1.04

DOI:10.6814/NCCU202200681

4.3.3. 極大化夏普比率投資組合

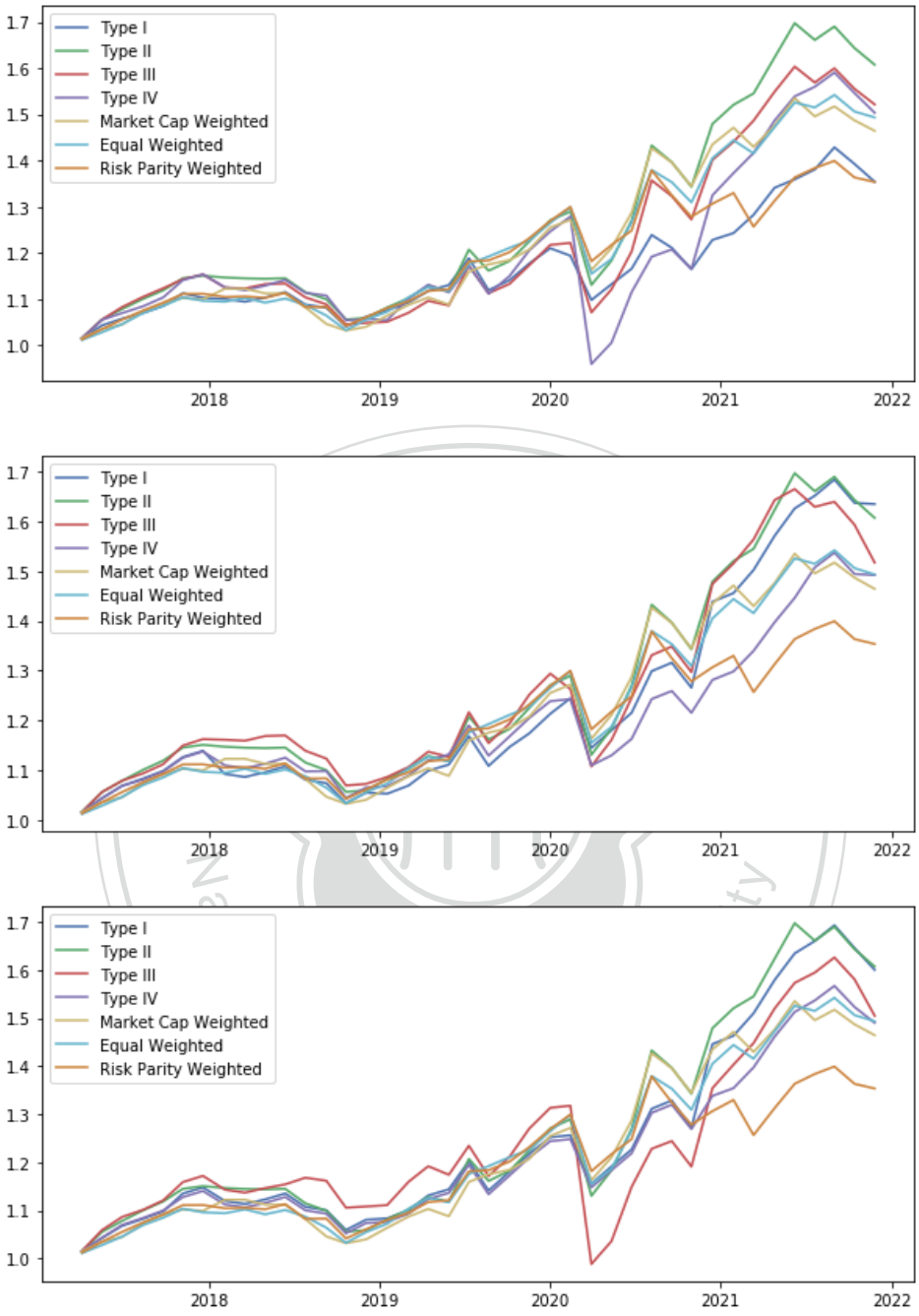

圖 4.5：不同獎勵設定與基準投資組合比較圖

圖 4.5 由上到下分別為更新次數 200000、600000 及 1000000，可以看出不管哪種獎勵的設定方式，大致上平均年化報酬率都能超過基準投資組合，透過下表詳細績效表現可以更容易看出彼此的關係：

DOI:10.6814/NCCU202200681

表 4.5 極大化夏普比率投資組合績效表現

	更新次數	Mean p.a(%)	SD p.a(%)	Sharpe	Omega	MDD (%)	Calmar
Type I	0.2m	9.52	10.41	0.91	0.92	9.27	1.03
	0.6m	15.89	12.36	1.29	1.00	9.24	1.72
	1m	15.14	12.42	1.22	0.84	8.04	1.88
Type II	0.2m	15.303	14.794	1.034	1.25	12.32	1.242
	0.6m	15.296	14.796	1.033	1.25	12.32	1.241
	1m	15.296	14.796	1.033	1.25	12.32	1.241
Type III	0.2m	13.417	14.97	0.90	1.00	12.32	1.09
	0.6m	13.320	15.06	0.88	0.99	14.43	0.92
	1m	13.045	19.73	0.66	0.64	24.97	0.52
Type IV	0.2m	13.01	19.59	0.66	0.83	24.97	0.52
	0.6m	12.76	11.24	1.13	0.67	10.72	1.19
	1m	12.71	10.40	1.22	0.77	8.03	1.58

根據表 4.5 來看四種不同獎勵設定下，投資組合在測試期間之績效表現，基本上除了 Type I 獎勵設定在更新次數 200000 時比起基準投資組合的平均年化報酬率要低許多，推測原因為模型並未完全學習；而 Type III 獎勵設定下，隨著跟新次數的增加，報酬率雖大致上相同但卻使得標準差增加，推測其原因為模型往錯誤的方向預測。其餘的獎勵設定及更新次數大致上皆有不弱於基準投資組合的績效，但在獲得高報酬率的同時普遍伴隨著更高的波動度，因此單除比較夏普比率時發現只剩下少數仍優於基準投資組合中最好的等值加權投資組合。

DOI:10.6814/NCCU202200681

4.3.4. 極小化變異數投資組合

除了極大化夏普比率外，本文還對強化學習產生的結果做極小化變異數投資組合，研究結果發現不管哪一種獎勵設定方式與更新次數，在選擇極小化變異數投資組合時產生結果皆為一致的。結果如下：

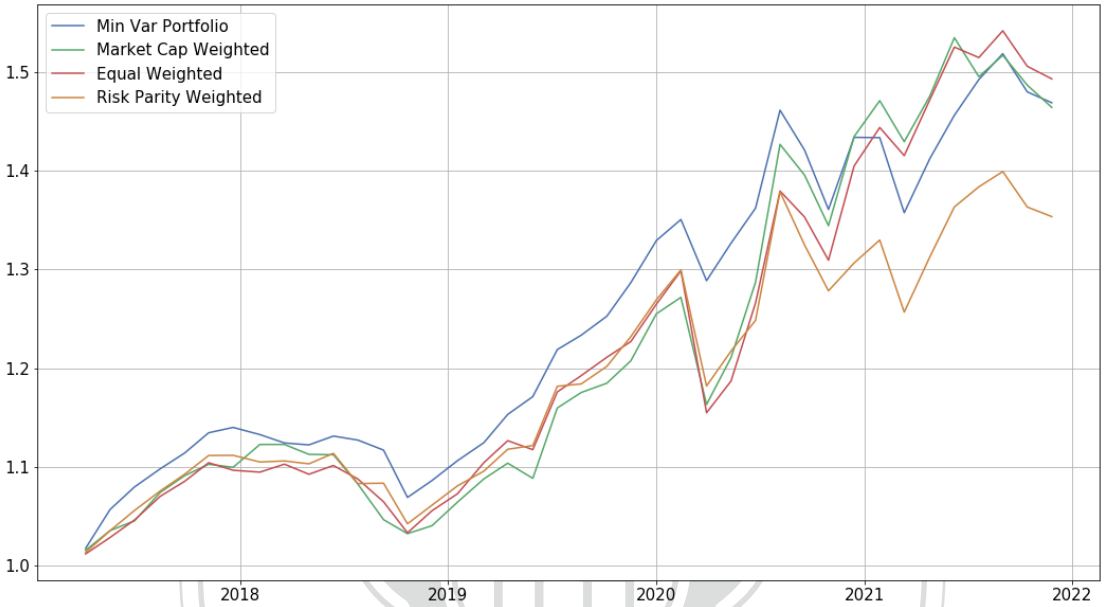

圖 4.6：極小化變異數投資組合績效比較圖

表 4.6 極小化變異數投資組合績效比較表

	Mean p.a(%)	SD p.a(%)	Sharpe	Omega	MDD(%)	Calmar
Market Cap-Weighted	12.12	11.78	1.03	1.33	8.55	1.42
Equal Weighted	12.78	11.56	1.11	1.28	11.08	1.15
Risk Parity Weighted	9.50	11.09	0.86	0.66	9.06	1.04
Min Variance Portfolio	12.23	9.39	1.32	1.19	7.11	1.72

根據以上圖表可以看出，極小化變異數投資組合在平均年畫報酬率上並沒有特別出眾的績效表現，略低於等值加權投資組合，但在波動度方面事全部最小的，也具有較小的回落，因此具有較大的夏普比率及卡瑪比率。

DOI:10.6814/NCCU202200681

4.3.5. 更新次數對績效之影響

本研究中對每個資產每種獎勵設定皆做三種不同更新次數的設定，分別為200000、600000 及 1000000，探討更新次數對於投資組合的績效之影響，故在本小節著重於次數與績效的相關性，而不再與基準資產做比較，因此將不同獎勵設定分開討論，根據表 4.5 可以看出，除了 Type I 平均年化報酬率有顯著提升以外，其餘的獎勵設定隨著更新次數的增加對於報酬率基本為持平，甚至還更下降。

為了方便觀察，因此將相同獎勵設定，不同更新次數的投資組合績效擺在同一張圖進行比較，藉此觀察出更新次數對於績效之間的關係，如下圖。

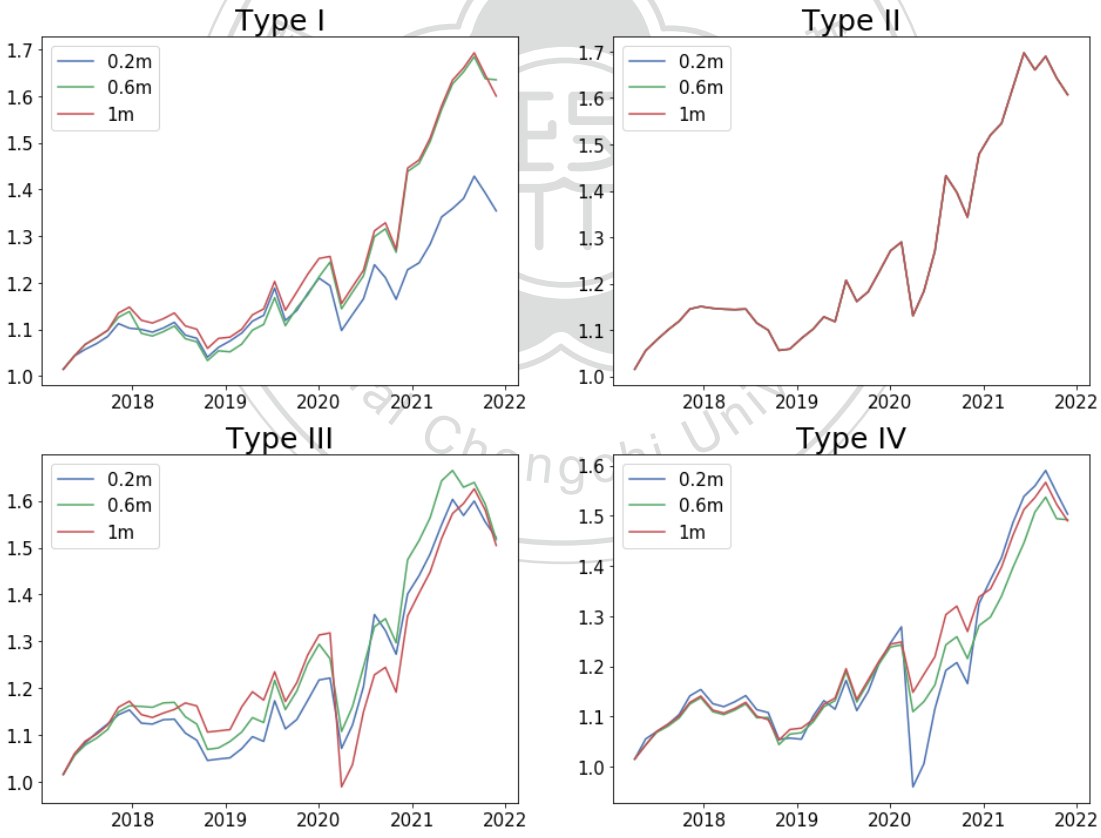

圖 4.7：同獎勵、不同更新次數之績效比較

DOI:10.6814/NCCU202200681

第五章　結論與建議

第一節　結論

本研究以 2006 年 1 月至 2021 年 11 月美國五檔不同資產類別的 ETF 商品，藉由強化學習，每 30 天預測標的資產價格的漲跌及漲跌幅度，將預測結果代入 Black-Litterman 模型的投資者主觀觀點，並進行資產配置，建構出風險分散的投資組合，並比較不同獎勵設定及更新次數不同對投資組合績效之影響。

實證發現：透過不同的獎勵設置，對結果影響十分巨大，因為強化學習中獎勵函數是訓練的一大重點，反而更新次數的影響相對來說就比較小，找到一個好的獎勵函數對於訓練模型絕對會起到加分的作用。在策略績效方面，本研究所建構之投資組合普遍表現出超越基準投資組合的績效。其中，獎勵函數設置為完全預測正確獲得獎勵這種方法所建構出的投資組合，擁有較高的平均年化報酬與夏普比率。

第二節　未來展望

本研究嘗試將 Black-Litterman 模型與強化學習中的 PPO 演算法做結合，此研究方法仍有許多可改進之處。對於未來研究方向，有以下幾點建議：

1. 強化學習中設定獎勵函數是一件困難的事，而本研究所使用的獎勵為比較基礎的設定。未來可嘗試使用不同獎勵，也許可獲得更好的績效表現。
2. 本研究因為選擇資產為 ETF 因此對於投資組合權重限制只能做多，未來可嘗試其他不同總類的資產，例如期貨，在交易上的限制相對較少，也許能夠獲得更好的表現。
3. 本研究觀察值採用過去 30 天之價量資量，未來的研究可嘗試使用其他觀察值，或許會獲得更好的預測準確率。

DOI:10.6814/NCCU202200681

參考文獻

[1] Black, F., & Litterman, R. (1991), "Asset Allocation: Combining Investor Views with Market Equilibrium." *The Journal of Fixed Income*, *1*(2), 7-18.

[2] Black, F., & Litterman, R. (1992), "Global Portfolio Optimization." *Financial Analysts Journal*, *48*(5), 28-43.

[3] Black, F., Jensen, M. C., & Scholes, M. (1972), "The Capital Asset Pricing Model: Some Empirical Tests." In M. Jenson, ed., *Studies in the Theory of Capital Markets* (Praeger, New York, NY).

[4] Donthireddy, P. (2018), "Black-Litterman Portfolios with Machine Learning derived Views." ResearchGate. Retired April 12, 2022, from https://www.researchgate.net/publication/326489143_Black-Litterman_Portfolios_with_Machine_Learning_derived_Views

[5] He, G., & Litterman, R. (2002), "The intuition behind Black-Litterman model portfolios." *Available at SSRN 334304*.

[6] Jiang, Z., & Liang, J. (2017, September), "Cryptocurrency Portfolio Management with Deep Reinforcement Learning." In 2017 *Intelligent Systems Conference (IntelliSys)*, 905-913, IEEE.

[7] Lin, E., Chen, Q., & Qi, X. (2020), "Deep reinforcement learning for imbalanced classification." *Applied Intelligence*, *50*(8), 2488-2502.

[8] Lintner, J. (1965), "Security Prices, Risk, and Maximal Gains from Diversification." *The Journal of Finance*, *20*(4), 587-615.

[9] Markowitz, H.(1952), "Portfolio selection." *The Journal of Finance*,7(1),77-91

[10] Meucci, A. (2010), "The Black-Litterman Approach: Original Model and Extensions." *Shorter version in, The Encyclopedia Of Quantitative Finance, Wiley*.

DOI:10.6814/NCCU202200681

[11] Moody, J., & Saffell, M. (2001), "Learning to Trade Via Direct Reinforcement." *IEEE transactions on neural Networks*, *12*(4), 875-889.

[12] Neuneier, R. (1997), "Enhancing Q-Learning For Optimal Asset Allocation." *Advances In Neural Information Processing Systems*, *10*. 936-942

[13] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017), "Proximal Policy Optimization Algorithms." arXiv:1707.06347

[14] Sharpe, W. F. (1964), "Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk." *The journal of finance*, *19*(3), 425-442.

[15] Treynor, J. L. (1961), "Market Value, Time, and Risk." *Available at SSRN 2600356.*

[16] Zhang, Y., Zhao, P., Li, B., Wu, Q., Huang, J., & Tan, M. (2020), "Cost-Sensitive Portfolio Selection Via Deep Reinforcement Learning." *IEEE Transactions on Knowledge and Data Engineering*

DOI:10.6814/NCCU202200681