國立政治大學金融學系

碩士學位論文

強化學習下動態調整風險偏好之投資組合配置: 以台灣 50 指數為例

Portfolio Allocation with Dynamic Risk Aversion via Reinforcement Learning : Evidence from Taiwan 50 Index

指導教授：林士貴 博士

研究生：陳昱成 撰

中 華 民 國 一 一 〇 年 六 月

DOI:10.6814/NCCU202100814

摘要

Markowitz (1952) 提出現代投資組合理論，透過均數-變異數模型(Mean-Variance Model) 為投資人進行資產配置，並設風險趨避參數調整報酬率和風險之間的比例，但在實務中，此風險趨避參數難以動態調整。本研究使用強化學習(Reinforcement Learning) 中的近端策略優化 (Proximal Policy Optimization, PPO)，依據不同市場變化，動態調整每一天風險趨避參數，當市場情況好時，投資人偏好承擔較高風險，獲得更高報酬，當市場情況壞時，投資人風險偏好趨於保守。本研究以台灣 50 指數當作整體市場走勢，比較強化學習輸入過去不同時間週期資訊之結果，研究結果顯示，不論輸入時間週期長短，強化學習績效皆能贏過固定風險趨避參數下均數-變異數模型，說明利用強化學習，能解決實務上風險趨避參數難以動態調整之問題。

關鍵詞：均數-變異數模型、風險趨避、強化學習、近端策略優化

DOI:10.6814/NCCU202100814

Abstract

Markowitz (1952) proposed Modern Portfolio Theory, which used the Mean-Variance Model to allocate assets for investors, and set the risk aversion parameter to adjust the ratio between return and the risk. But in practice, this risk aversion parameter is difficult to adjust dynamically. In our paper, we use Proximal Policy Optimization in reinforcement learning to dynamically adjust daily risk aversion parameters according to different market changes. In a bull market, investors prefer to take higher risks and get higher returns. On the other hand, in a bear market, investors' risk appetite tends to be conservative. This study uses the Taiwan 50 Index as the overall market trend, and compares the results of inputting different time periods of information into the model. The results show that regardless of the length of the input time period, the performance of the model can outperform the mean-variance model under fixed risk aversion parameters. Explain that the use of reinforcement learning can solve the problem of difficulty in dynamic adjustment of risk aversion parameters in practice.

Keywords：Mean-Variance model, Risk Aversion, Reinforcement Learning, Proximal Policy Optimization

DOI:10.6814/NCCU202100814

目 次

DOI:10.6814/NCCU202100814

表 次

DOI:10.6814/NCCU202100814

圖 次

DOI:10.6814/NCCU202100814

第一章 緒論

近年來，由於電腦運算能力大幅提升，人工智慧與大數據蓬勃發展，使許多應用得到飛躍性的突破，如影像辨識和自然語言處理，這波浪潮也為金融領域帶來創新，越來越多新技術與金融應用作結合。如使用機器學習預測客戶違約率，藉由大量過去客戶資訊，自動篩選出高違約率客戶，使銀行更容易偵測和防範違約事件。或者運用深度學習預測股價，演算法由過去資產價格走勢或相關總體指標，找出彼此間關係並預測股價。藉由人工智慧與大數據等技術，的確為金融領域帶來更多可能性，並解決許多問題。

在財務領域中，資產配置一直是值得研究的議題，金融環境複雜而多變，如何依據此變化配置不同資產權重非常困難，過往也有非常多文獻探討此問題。近年來金融科技的快速發展，資產配置也可運用相關新技術，像有文獻使用強化學習，因應各資產價格走勢，配置各資產權重(Jiang and Liang 2017; Zhang et al. 2020)，但此作法無法解釋其中運作原理。故本文將強化學習與傳統資產配置模型結合，藉由強化學習動態調整傳統模型中的參數，解決傳統模型之問題並解釋其經濟意涵。

第一節 研究動機

Markowitz (1952) 提出現代投資組合理論 (Modern Portfolio Theory)，透過均數-變異數模型配置不同資產間的權重，模型說明建構投資組合時，報酬率和風險之間有抵換關係，當投資人追求高報酬時，必然伴隨高風險，當投資人期望降低風險，報酬也會相對地下降，依不同投資人之風險偏好，模型設置風險趨避參數調整報酬與風險的比例，基於模型下，給定不同風險趨避參數，投資人可建構出效率前緣，效率前緣上每一點皆為最適投資組合。但在模型中，風險趨避參數往往難以決定，Basak and Chabakauri (2010) 求解多期均數-變異數模型時假設風

DOI:10.6814/NCCU202100814

險趨避參數固定不變以簡化模型，但投資人在不同時期下，應有不同的風險偏好，Bjork et al. (2014) 提出風險趨避參數與投資人現有財富應呈負相關。而本文認為，風險趨避參數除了與現有財富有關外，投資人也會依據市場表現，在不同時期有不同風險偏好。舉例來說，當市場表現好時，投資人會想承擔較高風險，以獲取更高報酬，風險趨避參數會變小，反之，當市場表現差時，投資人會趨於保守，以避免損失，風險趨避參數會變大，而在過往文獻中也有類似結果，Rosenberg and Engle (2001) 實證風險趨避程度與景氣循環呈負關係。因此，依據外在市場變化，投資人會有不同風險偏好。

本研究利用強化學習，依據市場變化動態調整風險趨避參數。機器學習分為三種方法，監督式學習 (Supervised Learning)、非監督式學習 (Un-supervised Learning)和強化學習，其中強化學習藉由獎勵設計，讓模型學習在不同環境下，應執行何種決策，而此決策會改變下一期環境，並藉此影響後續的決策，可視為多期決策過程。而一般應用於投資領域的機器學習方法為監督式學習，多為預測資產價格或股票漲跌，但皆為固定樣本下求解，不考慮序列關係，故強化學習相較於監督式學習更適合應用在本研究。且在過往文獻中，強化學習在資產配置有不少成功的應用，Moody and Saffell (2001) 利用強化學習，決定不同時期下標普 500 指數應買進或賣出，此策略比買進持有績效還好，Zhang et al. (2020) 提出新的強化學習模型，決定不同時期加密貨幣資產權重，此模型績效勝過以往強化學習模型及傳統策略。因此，本研究利用強化學習之特性，處理每段時期市場變化，決定下一期投資人風險偏好建構均數-變異數模型，為投資人在每一期配置最適投資組合。

第二節 研究目的

本文研究目的主要探討如何根據市場環境變化，利用強化學習動態調整不同時期的風險趨避參數，再依據不同時期的參數建構均數-變異數模型，以求得最

DOI:10.6814/NCCU202100814

適的投資組合。研究分為模型與實證兩部分，模型部分，介紹如何利用強化學習，將市場環境變化和動態調整風險趨避參數做連結；實證部分，使用台灣 50 指數價量資料代表市場環境變化，並依據此變化動態調整風險趨避參數，接著，利用求得之參數建構最適馬可維茲投資組合。

本文後續章節架構如下：第二章回顧投資組合理論、投資人風險偏好與強化學習應用在投資領域之相關文獻；第三章介紹均數-變異數模型與強化學習方法；第四章以台灣 50 指數和其成分股實證，說明強化學習如何應用並分析實證結果；第五章為本文研究結論和未來改進方向。

DOI:10.6814/NCCU202100814

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，透過均數-變異數模型進行資產配置，模型說明報酬率和風險之間具有抵換關係，並依投資人自身風險偏好，設置風險趨避參數調整報酬率和風險之比率，依據不同風險趨避參數，投資人可建構出一條效率前緣，效率前緣上每一點為不同風險偏好下之最適投資組合，在特定預期報酬率下極小化風險，或在特定風險下極大化報酬率，依據自身風險偏好，可在效率前緣選一點建構投資組合。

但均數-變異數模型只適用於單期決策，投資人進行資產配置時，只考慮極大化下期報酬率及極小化下期風險，無法應用在多期決策，顯然與實務不符合。故 Basak and Chabakauri (2010) 應用動態規劃 (Dynamic Programming) 對多期均數-變異數模型求出解析解。雖 Basak and Chabakauri (2010) 考慮多期決策，但在求解時，未探討不同時期下風險趨避參數的變化，一律將風險趨避參數設置為固定常數，此假設非常不合理，投資人在不同時期，其風險偏好應不同。因此本論文將考慮不同時期下風險趨避參數的變化並求出資產配置之權重。

第二節 投資人風險偏好

投資人風險偏好會隨著時間變動，過往已有多篇文獻探討，Rosenberg and Engle (2001) 使用標普 500 指數選擇權估計定價核 (Pricing Kernel)之參數和風險趨避參數，再將風險趨避參數與過往文獻指出和景氣循環相關之變數，如長短公債利差、消費成長率等等，做多元迴歸，實證出風險趨避參數與景氣呈負相關。Diaz and Esparcia (2019) 整理過往經濟和財務上，如何決定風險偏好之文獻，認為過去文獻中，常假設固定風險偏好以簡化模型，但此假設無法反映投資人真實

DOI:10.6814/NCCU202100814

的狀況，而近年來，越來越多文獻認為風險偏好應隨時間變動，且與景氣呈負相關，並將此關係代入模型求解，由此可知，動態調整風險偏好較固定不變更能反映投資人狀態。

故在均數-變異數模型中，假設風險趨避參數為常數非常不合理，所以 Bjork et al. (2011) 移除 Basak and Chabakauri (2010) 風險趨避參數為常數的假設，認為風險趨避參數與投資人現有財富呈負相關，當投資人財富較多時，風險趨避參數應較小，較能承受較高風險，但當投資人財富較少時，風險趨避參數應較大，無法承受較高的損失，Bjork et al. (2011) 將此財富與風險趨避關係代入模型，求得新的解析解，相較於 Basak and Chabakauri (2010)，更能反映投資人實際風險偏好。

Li and Li (2013) 將 Bjork et al. (2011) 之結果應用至保險公司決策，運用多期均數-變異數模型決定有再保險下，保險公司如何依模型進行資產配置，並比較風險趨避參數固定和動態調整之結果，由結果可知，當風險趨避參數與現有財富呈負相關時，所求得之結果較合理。

Zhang et al. (2017) 將 Bjork et al. (2011) 之結果延伸至多風險資產，並考慮投資人之負債問題，使結果更符合實務，並得出考慮多風險資產較單一風險資產更具經濟意涵。以上文獻顯示風險趨避參數應隨時間變動，且和投資人現有財富和景氣呈負相關，並非固定不變。

第三節 強化學習之投資領域應用

近年來，強化學習在財務領域應用蓬勃發展，最早 Neuneier (1998) 使用基於價值 (Value-Based) 方法中的 Q 學習 (Q-learning)，決定何時該買進德國股價指數 DAX，此策略績效較買進持有還好。Moody and Saffell (2001) 在強化學習

DOI:10.6814/NCCU202100814

架構中提出兩種新的獎勵函數，分別是可微分夏普比率 (Differential Sharpe Ratio) 和可微分下檔風險比率 (Differential Downside Deviation Ratio)，認為獎勵除了考慮投資組合報酬外，也將風險考慮進去，接著，使用強化學習中兩種不同演算法，基於策略 (Policy-Based) 方法中的直接強化學習 (Direct Reinforcement) 和基於價值方法中的 Q 學習，決定何時該買進、不動作和賣出標普 500 指數，由結果可知，此兩種演算法之績效皆比買進持有好，說明經過強化學習決定進出場時機，比單純買進持有績效更佳。另外，比較兩種不同演算法之績效，發現直接強化學習又比 Q 學習好，原因為基於策略方法能依據每一期的獎勵，直接調整模型，在財務領域中，每一期所做之投資決策，都能在下一期藉由股價漲跌得到反饋。而基於價值方法目的為估計獎勵函數，但資產價格走勢為非定態且受許多雜訊影響，導致無法準確估計獎勵函數，故相較於基於價值方法，在財務領域中，較適用於能即時調整模型的基於策略方法。

近年來，由於電腦運算能力大幅提升，相較於單純決定買進或賣出資產，強化學習能使用更複雜的模型，處理大量資料並直接決定各資產權重，Jiang and Liang (2017) 使用基於策略方法中的確定性策略梯度 (Deterministic Policy Gradient)，利用不同資產之開高低收價，決定投資組合中 12 種加密貨幣資產權重，雖然績效並沒有特別突出，但卻提出一個新架構，為強化學習應用在資產配置提供更多可能性。Zhang et al. (2020) 使用基於策略方法中的策略梯度 (Direct Policy Gradient)，一樣也利用不同資產間的開高低收價，決定投資組合中不同加密貨幣的權重，但提出新的投資組合策略網路 (Portfolio Policy Network，簡稱 PPN) 和獎勵函數，成本敏感獎勵 (Cost-sensitive Reward)，PPN 將卷積神經網路 (Convolutional Neural Networks，簡稱 CNN) 及長短期記憶神經網路 (Long Short-Term Memory，簡稱 LSTM) 結合，分別處理不同資產價格間的關係和同一資產價格走勢，較能萃取出資產的特徵，加上新設計的獎勵函數除了考慮投資組合風險外，也考慮投資組合周轉率，藉此降低手續費，故此模型績效較過往強化學習

DOI:10.6814/NCCU202100814

應用至資產配置之文獻都還好。

綜合以上文獻，近年強化學習在資產配置有不少成功應用，且都認為在該領域中，基於策略方法較基於價值方法還要適合，因為基於價值方法主要目的是估計獎勵，面對複雜的價格走勢，無法準確估計，而在資產配置中，這一期作出投資決策，下一期便能得知報酬率，並依此調整模型，故較適合基於策略方法。而相較傳統基於策略方法中的策略梯度，本文使用之演算法為 Schulman et al. (2017) 提出近端策略優化 (Proximal Policy Optimization，簡稱 PPO)，PPO 改善傳統基於策略方法收集決策過程花費時間過多的問題，讓 PPO 保有策略梯度的精神，但卻更有效率，花費時間更少，故本文採用 PPO。

DOI:10.6814/NCCU202100814

第三章 研究方法

第一節 馬可維茲投資組合理論

均數-變異數模型之目標式如 (3.1.1) 式，說明投資人在建構投資組合時，目的為極小化投資組合標準差和極大化投資組合報酬，而不同投資人會依據各自的風險趨避參數來分配報酬率和風險之比例，其中，風險趨避參數介於 0 和 1，當風險趨避參數為 0 時，代表投資人只考慮報酬，當風險趨避參數為 1 時，代表投資人只考慮風險，藉由此參數調整，可了解投資人當下的風險偏好，而此參數在本研究中會依據市場環境變化，隨時間變動。除了上述目標式外，加了限制式 (3.1.2) 和限制式 (3.1.3)，限制不同資產權重總合為 1，且資產權重介於 0 到 1，不允許放空。

$$\min \quad \lambda_t \sum_{i=1}^{N}\sum_{j=1}^{N} w_{i,t}w_{j,t}\sigma_{ij,t} - (1-\lambda_t)\sum_{i=1}^{N} w_{i,t}u_{i,t} \tag{3.1.1}$$

$$\text{subject to} \ \sum_{i=1}^{N} w_{i,t} = 1 \tag{3.1.2}$$

$$0 \le w_{i,t} \le 1 \tag{3.1.3}$$

其中 N 為不同資產數量，$w_{i,t}$ 為投資組合中資產 i 在第 t 期權重，$u_{i,t}$ 為資產 i 在第 t 期預期報酬，$\sigma_{ij,t}$ 為資產 i 與資產 j 在第 t 期報酬共變異數，λ_t 為第 t 期風險趨避參數。

第二節 強化學習

3.2.1 小節會先介紹強化學習基本架構，說明強化學習如何運作，3.2.2 小節會介紹傳統基於策略方法中的策略梯度，3.2.3 小節會介紹本研究使用之演算法

DOI:10.6814/NCCU202100814

PPO，說明演算法原理和流程，和如何改善傳統方法，最後，在 3.2.4 小節會介紹如何將強化學習應用至動態調整風險趨避參數。

3.2.1 基本架構

強化學習基本架構如圖 3-1 所示，其中，代理人 (Agent) 代表作決策之主體，環境 (Environment) 代表與代理人互動的場景，在 t 期時，環境會給定代理人決

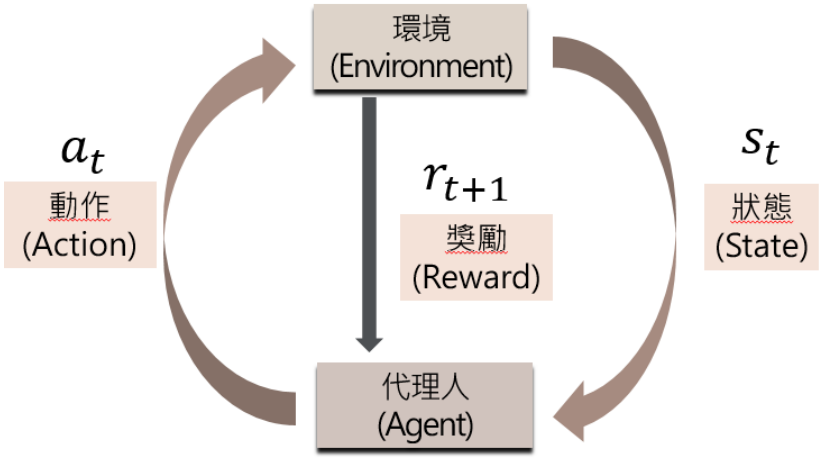


圖 3.1：強化學習基本架構

策所需資訊，稱為狀態 s_t，當代理人接收到狀態，會依據策略 (Policy) 做出對應動作 a_t，策略可表達為任意函數 $\pi_\theta(a_t|s_t)$，θ 為函數之參數，接著依據代理人執行的不同動作，會得到相應獎勵 r_{t+1}，此過程會不斷循環，直到決策結束。而從決策開始到結束，為一決策過程 $\tau^i = \{s_1^i, a_1^i, r_2^i, s_2^i, a_2^i, r_3^i, \dots, s_{T-1}^i, a_{T-1}^i, r_T^i\}$，其中 $t = 1, \dots, T$ 期決策，$i = 1, \dots, N$ 次決策過程。

強化學習中的基於策略方法目的是希望執行完決策過程 τ^i 後，所得到總獎勵 $R(\tau^i) = \sum_{t=1}^{T-1} r_{t+1}^i$ 越大越好，為此，演算法需調整策略 π_θ 中之參數 θ 以達到目的，接著，介紹基於策略方法中的策略梯度，說明演算法如何找到最適策略。

3.2.2 策略梯度

DOI:10.6814/NCCU202100814

假設第 i 次決策過程 τ^i 後得到總獎勵為 $R(\tau^i)=\sum_{t=1}^{T-1} r_{t+1}^i$，策略梯度演算法藉由調整策略 π_θ 中之參數 θ，使執行完決策過程 τ^i 後，所得到總獎勵 $R(\tau^i)=\sum_{t=1}^{T-1} r_{t+1}^i$ 越大越好。因此目標式如 (3.2) 式，目標為極大化期望總獎勵。期望總獎勵為每一次決策過程 τ^i 所得到總獎勵 $R(\tau^i)$，乘上每一次決策過程發生機率 $p_\theta(\tau^i)$。$p_\theta(\tau^i)$ 分為兩部分，其中 $p(s_{t+1}^i|s_t^i,a_t^i)$ 受環境影響，不可控制，為給定在 t 時間狀態 s_t^i 與動作 a_t^i 下其 $t+1$ 時間發生狀態 s_{t+1}^i 機率，而 $p_\theta(a_t^i|s_t^i)$ 則受策略參數 θ 控制，給定 t 時間狀態 s_t^i 下其 t 時間執行動作 a_t^i 之機率。再將決策開始至結束所發生機率全部相乘，即可得到 $p_\theta(\tau^i)$，如 (3.3) 式。

$$Max \quad E_{\tau\sim p_\theta(\tau)}(R(\tau))=\sum_{i=1}^{N} R(\tau^i)p_\theta(\tau^i) \tag{3.2}$$

此處

$$p_\theta(\tau^i)=p(s_1^i)p_\theta(a_1^i|s_1^i)p(s_2^i|s_1^i,a_1^i)p_\theta(a_2^i|s_2^i)p(s_3^i|s_2^i,a_2^i)\ldots$$

$$=p(s_1^i)\prod_{t=1}^{T} p_\theta(a_t^i|s_t^i)p(s_{t+1}^i|s_t^i,a_t^i) \tag{3.3}$$

決定完目標式後，需使用梯度上升法 (Gradient Ascent) 不斷更新參數 θ，以找出最適策略參數 θ 極大化未來期望總獎勵 (3.2) 式，因此需先計算目標式 (3.2) 式之梯度，推導如下式

$$\nabla E_{\tau\sim p_\theta(\tau)}[R(\tau)]=\sum_\tau R(\tau)\nabla p_\theta(\tau) \;=\; \sum_\tau R(\tau)\,p_\theta(\tau)\frac{\nabla p_\theta(\tau)}{p_\theta(\tau)}$$

$$=\;\sum_\tau R(\tau)\,p_\theta(\tau)\nabla log p_\theta(\tau)$$

$$=E_{\tau\sim p_\theta(\tau)}[R(\tau)\nabla log p_\theta(\tau)]$$

DOI:10.6814/NCCU202100814

$$\approx \frac{1}{N}\sum_{i=1}^{N} R(\tau^i)\nabla log p_\theta(\tau^i)$$

$$= \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T_n-1} R(\tau^i)\nabla log p_\theta(a_t^i|s_t^i) \quad (3.4)$$

由梯度 (3.4) 式可知，總獎勵 $R(\tau^i)$ 決定梯度 $\nabla\log\{p_\theta(a_t^i|s_t^i)\}$ 更新之方向，舉例總獎勵 $R(\tau^i)$ 為正時，代表此決策過程 τ^i 較好，更新方向往梯度 $\nabla\log\{p_\theta(a_t^i|s_t^i)\}$ 方向更新，即增加在狀態 s_t^i 執行動作 a_t^i 之機率 $p_\theta(a_t^i|s_t^i)$，當總獎勵 $R(\tau^i)$ 為負時，代表此決策過程 τ^i 較差，更新方向往梯度 $\nabla\log\{p_\theta(a_t^i|s_t^i)\}$ 反方向更新，即減少在狀態 s_t^i 執行動作 a_t^i 之機率 $p_\theta(a_t^i|s_t^i)$。但每一次執行動作 a_t^i 得到的獎勵 r_{t+1}^i 佔總獎勵 $R(\tau^i)$ 權重不同，有些動作貢獻總獎勵較多，有些動作貢獻較少，故以總獎勵 $R(\tau^i)$ 決定更新方向不合理，需有另一種方法，衡量個別動作的貢獻。

在強化學習中，會採用優勢函數 (Advantage Function) $A_t^\theta(s_t^i, a_t^i)$ 來取代總獎勵 $R(\tau^i)$，優勢函數能衡量在某一個狀態 s_t^i 下，執行動作 a_t^i 所得到之貢獻，數學式如下，其中 γ 為折現因子

$$A_t^\theta(s_t^i, a_t^i) = \sum_{t'=t}^{T_i-1}\gamma^{t'-t} r_{t'+1}^i - V(s_t)，其中\ V(s_t) = E[\sum_{t'=t}^{T_i-1}\gamma^{t'-t} r_{t'+1}^i | s = s_t^i]$$

優勢函數可分為兩部分，第一部分為 $\sum_{t'=t}^{T_i-1}\gamma^{t'-t} r_{t'+1}^i$，與總獎勵 $R(\tau^i)$ 不同之處有兩點，第一點為因為未來的獎勵會受現在執行動作 a_t^i 影響越來越小，故新增折現因子 γ 將未來得到的獎勵折現到現在。第二點為因為過去的獎勵與現在執行的動作 a_t^i 無關，故只衡量從第 t 期開始，到結束的獎勵總和，而非從第一期到結束之獎勵總和。第二部分為 $V(s_t) = E[\sum_{t'=t}^{T_i-1}\gamma^{t'-t} r_{t'+1}^i | s = s_t^i]$，為衡量在某一個狀態 s_t^i 下，還沒執行動作 a_t^i，能得到的平均值，用來判斷執行動作 a_t^i

DOI:10.6814/NCCU202100814

後所得到的 $\sum_{t'=t}^{T_i-1}\gamma^{t'-t}r_{t'+1}^i$，是否比平均好。

決定優勢函數後，將其取代總獎勵 $R(\tau^i)$，將梯度 (3.4) 式改寫為 (3.5) 式，

$$\nabla E_{\tau\sim p_\theta(\tau)}[R(\tau)] \approx \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T_n-1} A_t^\theta(s_t^i,a_t^i)\nabla log p_\theta(a_t^i|s_t^i) \tag{3.5}$$

其意義為優勢函數 $A_t^\theta(s_t^i,a_t^i)$ 為正時，也就是在狀態 s_t^i 下，執行動作 a_t^i 所得到的 $\sum_{t'=t}^{T_i-1}\gamma^{t'-t}r_{t'+1}^i$ 大於平均值時，增加其機率 $p_\theta(a_t^i|s_t^i)$，反之，總獎勵 $A_t^\theta(s_t^i,a_t^i)$ 為負時，也就是當 $\sum_{t'=t}^{T_i-1}\gamma^{t'-t}r_{t'+1}^i$ 小於平均值時，減少其機率 $p_\theta(a_t^i|s_t^i)$。算出梯度 (3.4) 式後，使用梯度上升法更新參數 θ，藉由不斷更新，找到最適策略中的參數 θ，數學式如下，其中學習率為 η：

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

3.2.3 近端策略優化(PPO)

給定參數 θ 下，傳統策略梯度演算法流程為先用策略 π_θ 與環境互動，收集 $i=1,2,\dots,N$ 筆決策過程 τ^i 後，使用梯度上升法更新參數 θ。但每更新一次參數 θ 就必須重新收集決策過程，導致演算法將時間都花費在此，過程非常耗時且沒效率，PPO 目的為改善此問題，讓收集的決策過程能夠重複利用，使更新參數更有效率及省時。

當更新參數 θ 時，PPO 不使用策略 π_θ 收集決策過程，改用另一個策略 $\pi_{\theta'}$ 收集並更新策略 π_θ 中的參數 θ，其中 θ' 固定，可解決重新收集決策過程的問題，因收集過程改變，便需改寫梯度 (3.5) 式至 (3.6) 式，其中應用到重要性採樣 (Importance Sampling)，接著由梯度 (3.6) 式反推回目標式 (3.7) 式。

DOI:10.6814/NCCU202100814

$$\nabla E_{\tau\sim p_\theta(\tau)}[R(\tau)] = E_{\tau\sim p_{\theta'}(\tau)}\left[\frac{p_\theta(a_t|s_t)}{p_{\theta'}(a_t|s_t)}A_t^{\theta'}(s_t,a_t)\nabla log p_\theta(\tau)\right] \tag{3.6}$$

$$J^{\theta'}(\theta) = E_{\tau\sim p_{\theta'}(\tau)}\left[\frac{p_\theta(a_t|s_t)}{p_{\theta'}(a_t|s_t)}A_t^{\theta'}(s_t,a_t)\right] \tag{3.7}$$

但重要性採樣有一個問題，雖然期望值相等，但 $p_\theta(\tau)$ 和 $p_{\theta'}(\tau)$ 這兩個分配差距越大，變異數也會隨之放大，所以 $p_\theta(\tau)$ 和 $p_{\theta'}(\tau)$ 這兩個分配差距不能過大，故在目標式 (3.7) 式加入限制，改寫為 (3.8) 式。

$$J_{CLIP}^{\theta'}(\theta) = E_{\tau\sim p_{\theta'}(\tau)}[\min(\frac{p_\theta(a_t^i|s_t^i)}{p_{\theta'}(a_t^i|s_t^i)}A_t^{\theta'}(s_t,a_t),$$

$$clip(\frac{p_\theta(a_t^i|s_t^i)}{p_{\theta'}(a_t^i|s_t^i)},1+\varepsilon,1-\varepsilon)A_t^{\theta'}(s_t,a_t))] \tag{3.8}$$

其中 $clip(\frac{p_\theta(a_t^i|s_t^i)}{p_{\theta'}(a_t^i|s_t^i)},1+\varepsilon,1-\varepsilon)$ 代表當 $\frac{p_\theta(a_t^i|s_t^i)}{p_{\theta'}(a_t^i|s_t^i)}$ 介於 $1+\varepsilon$ 和 $1-\varepsilon$ 時，為 $\frac{p_\theta(a_t^i|s_t^i)}{p_{\theta'}(a_t^i|s_t^i)}$，若大於 $1+\varepsilon$ 為 $1+\varepsilon$，若小於 $1-\varepsilon$ 為 $1-\varepsilon$。

為了解決 $p_\theta(\tau)$ 和 $p_{\theta'}(\tau)$ 分配差距過大的問題，目標式 (3.8) 可分為兩部分，如圖 3.2，當 $A_t^{\theta'}(s_t,a_t)$ 為正時，一樣增加在狀態 s_t，執行動作 a_t 之機率 $p_\theta(a_t|s_t)$，但當 $\frac{p_\theta(a_t^i|s_t^i)}{p_{\theta'}(a_t^i|s_t^i)}$ 增加到大於 $1+\varepsilon$ 時，限制其為 $1+\varepsilon$，反之，當 $A_t^{\theta'}(s_t,a_t)$ 為負時，減少在狀態 s_t，執行動作 a_t 之機率 $p_\theta(a_t|s_t)$，但當 $\frac{p_\theta(a_t^i|s_t^i)}{p_{\theta'}(a_t^i|s_t^i)}$ 減少到小於 $1-\varepsilon$ 時，限制其為 $1-\varepsilon$，透過以上限制，讓 $p_\theta(\tau)$ 和 $p_{\theta'}(\tau)$ 的分配在更新過程中差距不會擴大。

DOI:10.6814/NCCU202100814

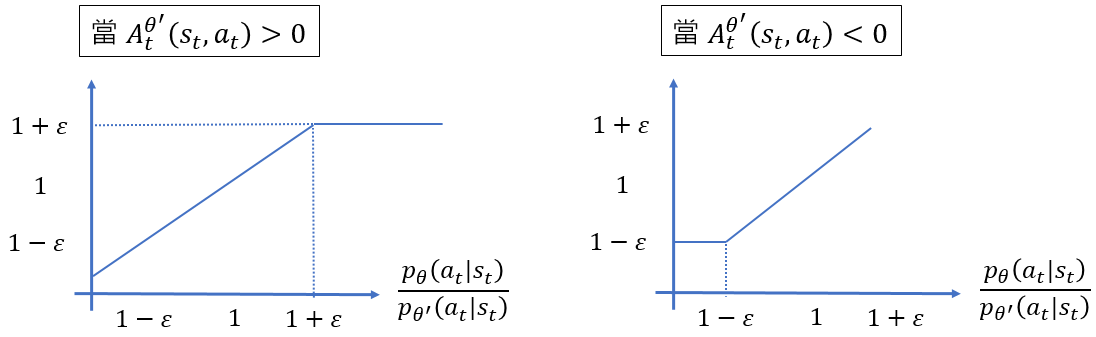


圖 3.2 目標式圖解

除了目標式 (3.8) 外，PPO 還分別加了目標式 (3.9) 和目標式 (3.10)，其中

$$J_{VF}^{\theta'}(\theta)=[\sum_{t'=t}^{T-1}\gamma^{t'-t}r_{t'+1}^{i}+\gamma^{T-t}V(s_T)-V(s_t)]^2 \tag{3.9}$$

$$H^{\theta'}(s_t)=-\sum p_{\theta'}(a_t|s_t)\,log\,p_{\theta'}(a_t|s_t) \tag{3.10}$$

目標式 (3.9) 為均方誤差 (Mean Square Error)，目的要讓估計 $V(s_t)$ 越來越準，目標式 (3.10) 為執行不同動作的熵 (Entropy)，當熵越大，代表越能平均選到不同的動作，有助於一開始更新時探索 (Exploration) 不同動作。最後，PPO 將目標式修改為極大化 (3.11) 式，並設超參數 c_1 和 c_2 調整其權重。

$$\max \quad J_{PPO}^{\theta'}(\theta)=E_{\tau\sim p_{\theta'}(\tau)}[J_{CLIP}^{\theta'}(\theta)-c_1J_{VF}^{\theta'}(\theta)+c_2H^{\theta'}(s_t)] \tag{3.11}$$

3.2.4 強化學習動態風險趨避之應用

本小節說明如何利用強化學習，將市場環境變化和動態決定風險趨避參數連結在一起，首先定義本研究中強化學習的三個基本元素，分別為狀態 s_t、動作 a_t 和獎勵 r_{t+1}。

本研究中，狀態 $s_t=[X_{t-n},\,X_{t-n+1},\dots,\,X_{t-1}]$，$s_t$ 為過去 n 天台灣 50 指數價量資料，其中 $X_t=[Y_t,\,O_t,\,H_t,\,L_t,\,S_t,\,V_t]$ 包含六種不同價量變數，令 $P_t^{(O)},P_t^{(H)},P_t^{(L)},P_t^{(C)}$ 分別為第 t 天台灣 50 指數的開高低收價，v_t 為第 t 天的

DOI:10.6814/NCCU202100814

成交量，定義六種價量變數 $Y_t = \frac{P_t^{(C)} - P_{t-1}^{(C)}}{P_{t-1}^{(C)}}$ 為第 t 天報酬率，$O_t = \frac{P_t^{(O)}}{P_t^{(C)}}$, $H_t = \frac{P_t^{(H)}}{P_t^{(C)}}$, $L_t = \frac{P_t^{(L)}}{P_t^{(C)}}$ 為第 t 天開高低價除以收盤價，藉由此調整，呈現每一天價格波動，S_t 為第 t 天過去五天報酬率標準差，$V_t = \frac{v_t - v_{t-1}}{v_{t-1}}$ 為第 t 天成交量變化。藉由過去 n 天六種台灣 50 指數的價量變數，間接反映台灣整體金融市場情況。

當接收到市場變化後，會動態調整每一期風險趨避參數，也就是每一期的動作 $a_t = \lambda_t$，其中 $\lambda_t = \{0, 1\}$，市場大漲時，期望選擇極大化報酬率投資組合 (λ_t =0)，使均數-變異數模型在只考慮報酬下，所建構出的投資組合能獲得比大盤更高的報酬，反之，市場大跌時，期望選擇最小變異數投資組合 (λ_t =1)，使均數-變異數模型在只考慮風險下，建構出的投資組合下跌幅度能較大盤少。

當決定風險趨避參數 λ_t 後，需要獎勵 r_{t+1} 衡量此次決策，當 $t+1$ 期時可算出極大化報酬率投資組合 (λ_t = 0) 和最小變異數投資組合 (λ_t = 1)之報酬率，分別為 $Y_{t+1}^{\lambda_t=0}$ 和 $Y_{t+1}^{\lambda_t=1}$，其中 $Y_{t+1}^{\lambda_t}$ 代表在風險趨避參數 λ_t 下，$t+1$ 期均數-變異數模型求得之投資組合報酬率。而獎勵 $r_{t+1} = Y_{t+1}^{\lambda_t} - \frac{Y_{t+1}^{\lambda=0} + Y_{t+1}^{\lambda=1}}{2}$，為風險趨避參數 λ_t 所對應投資組合報酬率減掉 $Y_{t+1}^{\lambda=0}$ 和 $Y_{t+1}^{\lambda=1}$ 之平均，當獎勵 r_{t+1} 為正時，代表 t 期選擇的 λ_t 對應之投資組合報酬率較大，反之，當獎勵 r_{t+1} 為負時，代表在 t 期選擇的 λ_t 對應之投資組合報酬率較小，藉由上述設計，讓模型學到在特定大盤走勢下，選擇何種風險趨避參數 λ_t 才能使下一期投資組合報酬率較大。

最後，強化學習動態風險趨避之應用基本架構如圖 3.3，首先，代理人藉由接收狀態 s_t，觀察台灣 50 指數價量變化，判斷風險趨避參數 λ_t 為 0 或 1，決定風險趨避參數 λ_t 後，將之代入均數-變異數模型，求出本研究投資組合權重，最後，再藉由獎勵 r_{t+1} 去評量決策的好壞，並依此不斷更新自己的決策，

DOI:10.6814/NCCU202100814

藉由不斷重複上述步驟，達到依據市場變化，動態調整風險趨避參數。

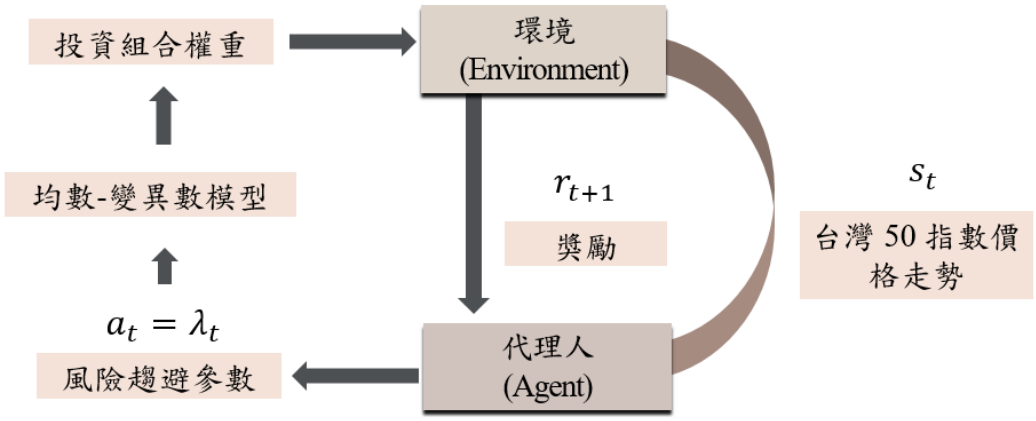


圖 3.3：強化學習應用基本架構

第三節 績效衡量指標

本研究總共有四種績效衡量指標，透過以下四種指標，比較不同投資組合績效及強化學習預測表現，而在衡量績效時，本研究不考慮流動性風險和手續費。第一種指標為累加報酬，本研究投資組合皆為每日調整資產權重，故我們將每日報酬累加起來，觀察不同時期的漲跌情形。第二種指標為夏普比率 (Sharpe Ratio)，式子為 $\frac{Y_p}{\sigma_p}$，衡量承擔每一單位風險，能獲得多少報酬，其中 Y_p 為投資組合報酬率，σ_p 為投資組合標準差。第三種指標為索提諾比率 (Sortino Ratio)，式子為 $\frac{Y_p}{\sigma_{p-negative}}$，衡量承擔每一單位下行風險，能獲得多少報酬，其中 $\sigma_{p-negative}$ 為投資組合負報酬標準差，因正報酬波動對績效衡量有正面影響，所以只衡量投資組合負報酬波動。最後一種指標為預測準確率，因為獎勵函數的設計，我們希望強化學習選到正獎勵越多越好，故設預測準確率為衡量強化學習選到正的獎勵之天數比率，數學式為 $\frac{\#(r_{t+1} > 0)}{T}$。

DOI:10.6814/NCCU202100814

第四章 實證分析

本研究使用台灣 50 指數價量資料代表台灣金融市場變化，並依據此環境變化，用強化學習決定合適的風險趨避參數，接著，將決定好的風險趨避參數和台灣 50 指數成分股過去報酬代入均數-變異數模型，得到各成分股權重，為本研究之投資組合。

本章節的架構上，第一節會先介紹資料來源與前處理，第二節會介紹強化學習模型設定和研究流程，最後第三節會分析實證結果。

第一節 資料來源與前處理

因台灣 50 指數為季調整成分股，故描述資料範圍皆以季為單位，本研究採用期間為 2004 年第一季到 2020 年第四季，日期為 2004 年 3 月 22 日至 2021 年 3 月 22 日，其中各季成分股名單資料來源為公開資訊觀測站，台灣 50 指數和其成分股開高低收價及成交量資料從 TEJ 取得，頻率為日資料。

台灣 50 指數成分股部分不符合本研究模型設定，故將歷史資料作如下調整及簡化以符合設定，第一，受限於強化學習模型，每季成分股皆需保持 50 家，台灣 50 指數在 2006 年第三季及 2011 年第三季成分股為 51 家，故分別在這兩季刪除成分股廣輝及裕隆，因廣輝下一季會被剔除成分股並下市，而裕隆為 2011 年第三季新增之成分股，剔除此兩家公司對台灣 50 指數造成影響較小，故刪除以符合模型設定。第二，部分成分股因新上市、合併等等因素，無過去歷史報酬率代入均數-變異數模型求解，缺失部分以當天其餘成分股平均報酬率代替，藉由此調整，避免後續求解均數-變異數模型受此缺失值影響，造成模型結果不符合實際情況。

DOI:10.6814/NCCU202100814

第二節 模型設定

在強化學習架構中，面對不同狀態 s_t，代理人依據策略函數 $\pi_\theta(a_t|s_t)$ 進行決策，而策略函數為神經網路，神經網路模型設定如圖 4.1，其中括弧內為神經

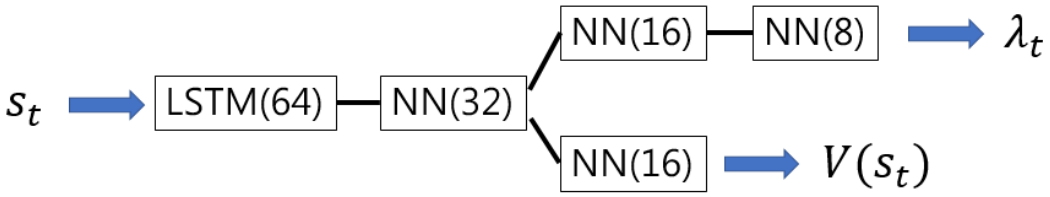


圖 4.1：神經網路模型

元個數。神經網路模型輸入狀態 s_t，輸出分為兩部分，右上部分輸出動作，也就是風險趨避參數 λ_t，右下部分輸出優勢函數中的 $V(s_t)$，為衡量在狀態 s_t 下，能得到的平均獎勵，由於狀態包含過去 n 天台灣 50 指數價量變數，具有時間性，故採用擅長處理序列之神經網路 LSTM 處理，後續都採用一般神經網路 NN 連接，接著使用演算法 PPO，不斷更新神經網路模型中的參數 θ，找出最適的模型。

決定完模型後，設定學習率為 0.0001，更新次數為 4000000 次，因強化學習模型訓練時，並無像一般監督式學習有樣本答案，而是藉由與環境多次互動所收集的決策過程進行參數更新，故一般學習率會設置較小且更新次數較監督式學習任務多。除了模型超參數設定，歷史狀態包含過去 n 天台灣 50 指數價量變數，設歷史狀態天數 n 為 5、20 和 60 天，分別代表一星期、一個月及一季的價量資訊，觀察不同時間長度對動態調整風險趨避參數是否有影響。另外，求解均數-變異數模型時，需過去 m 天各成分股報酬率計算歷史平均報酬率及共變異數，歷史報酬波動天數 m 一樣設定為 5、20 和 60 天，觀察不同均數-變異數模型對結果的影響。

DOI:10.6814/NCCU202100814

本研究為每日重新配置投資組合權重，計算報酬率方式為收盤價對收盤價，第 t 天的決策流程如圖 4.2。第 t 天時，算出狀態 s_t，包含過去 n 天台灣 50

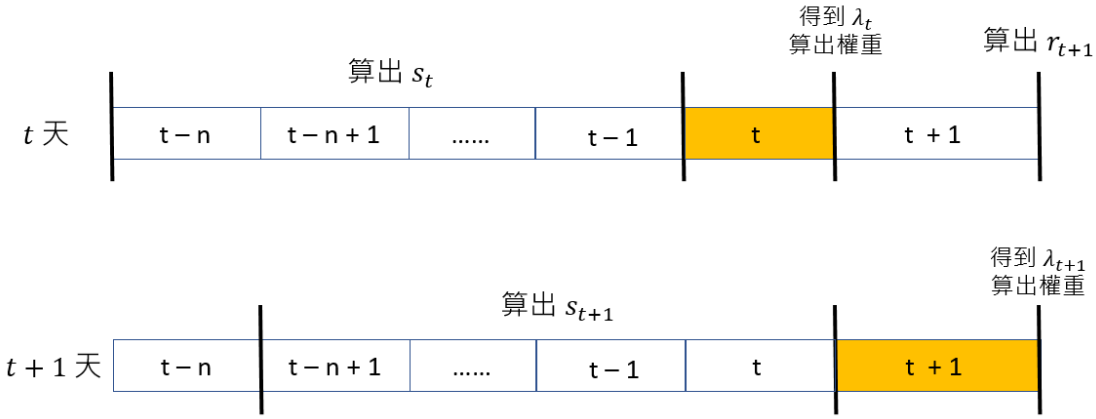


圖 4.2：第 t 天決策流程

指數價量變數，並由模型得到風險趨避參數 λ_t，接著，將參數 λ_t 及過去報酬變異 m 天報酬率與標準差代入均數-變異數模型算出成分股權重，到了第 t＋1 天時，可由第 t 天和第 t＋1 天收盤價變化算出各成分股報酬率，進而得到投資組合報酬率，並算出獎勵 r_{t+1}，同時，重複上述步驟算出風險趨避參數 λ_{t+1} 及對應的成分股權重，接著不斷進行決策，直到決策結束。

本研究將資料切分為訓練集和測試集，首先，模型會先以訓練集資料訓練神經網路 $\pi_\theta(a_t|s_t)$ 中的參數 θ，並挑選訓練結果最好的模型當作最終模型，再將測試集資料輸入最終模型進行動態調整風險趨避參數。本研究切分訓練集和測試集方式如圖 4.3，白色部分為訓練集，深色部分為測試集，因台灣 50 指數成分股為季調整，故切分訓練集和測試集皆以季為單位，首先，先以 2004 年第一季

DOI:10.6814/NCCU202100814

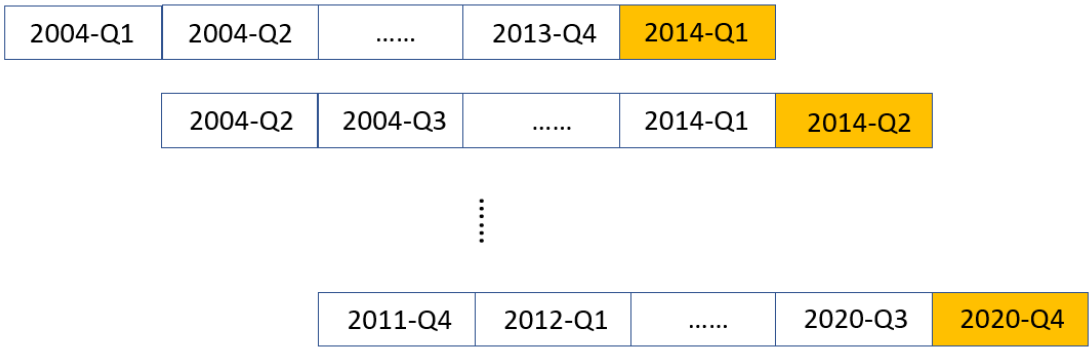


圖 4.3：切分訓練集和測試集

到 2013 年第四季，總共 40 季，相當 10 年的資料當作訓練集，訓練模型如何動態調整每一天風險趨避參數，接著使用訓練完的模型調整下一季，2014 年第一季每一天風險趨避參數，接著採移動窗口，往前遞移一季，用 2004 年第二季到 2014 年第一季當作訓練集，2014 年第二季當作測試集，不斷遞移下去，直到 2020 年第四季。藉由移動窗口的設計，讓模型能夠隨時間調整，更能準確判斷風險趨避參數。

第三節 實證結果

本節共分為三個小節，在 4.3.1 小節會分析與實證結果比較之基準投資組合，接著，在 4.3.2 小節會介紹強化學習投資組合的績效比較，並比較不同參數對結果的影響，最後，在 4.3.3 小節，探討利用強化學習求得之風險趨避參數與股價走勢之關係。

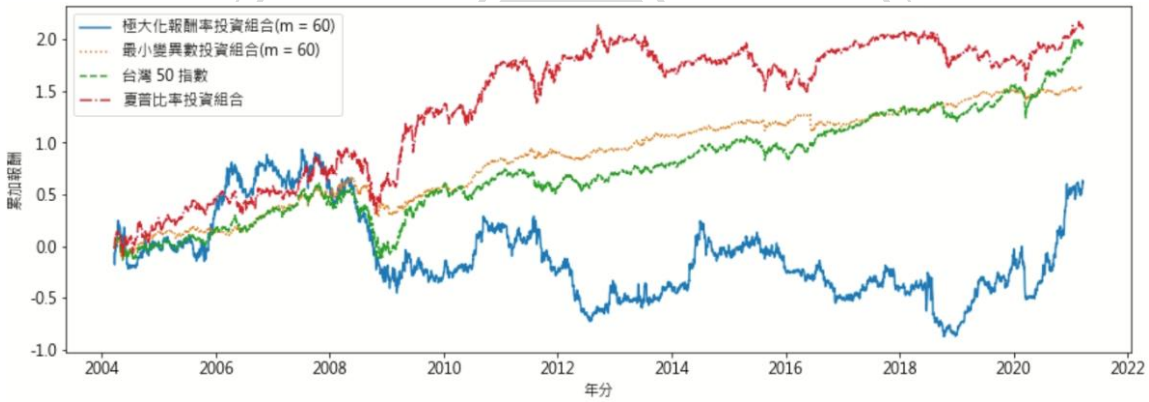
4.3.1 基準投資組合分析

本研究分四種基準投資組合與實證結果比較，分別為台灣 50 指數、極大化報酬率投資組合 ($\lambda_t = 0$) 和最小變異數投資組合 ($\lambda_t = 1$) 及均數-變異數模型目標式改為極大化夏普比率的夏普比率投資組合。因本文目的為動態調整風險趨避參數，期望在市場大漲選擇極大化報酬率投資組合 ($\lambda_t = 0$)，承受較高風險獲

DOI:10.6814/NCCU202100814

取更高報酬，在市場大跌選擇最小變異數投資組合 (λ_t =1)，下跌幅度較小，所以應用強化學習後，績效應比台灣 50 指數和固定風險趨避參數還要好，另外，在風險趨避參數未知下，一般求解均數-變異數模型常將目標式改為極大化夏普比率求解，夏普比率投資組合雖也可視為動態調整風險趨避參數，但其目的為整體投資組合夏普比率最大，無法因應市場漲跌調整對應的風險趨避參數，故本研究之投資組合績效應勝過夏普比率投資組合。

在利用強化學習動態調整風險趨避參數之前，可先觀察資料期間，2004 年到 2021 年不同基準投資組合累加報酬走勢，如圖 4.4，其中均數-變異數模型皆以過去報酬波動天數 m = 60 為例。雖然整體來說，夏普比率投資組合累加報酬較



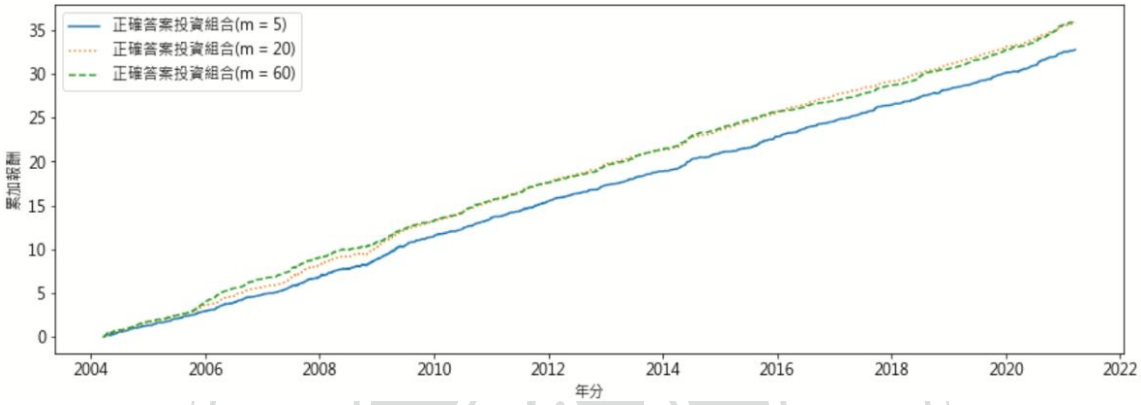
Note：圖為歷史報酬波動天數 m 為 60 下，台灣 50 指數、極大化報酬率投資組合 (λ_t =0)、最小變異數投資組合 (λ_t =1) 和夏普比率投資組合 2004 年到 2021 年累加報酬。

圖 4.4：不同基準投資組合累加報酬圖

其餘基準投資組合還好，但觀察不同時期，發現市場情況較好時，如 2020 年到 2021 年台股市場頻頻創新高時，選擇極大化報酬率投資組合 (λ_t =0) 上漲幅度較其餘基準投資組合大，而當市場情況較差時，如 2008 年到 2009 年金融海嘯時，選擇最小變異數投資組合 (λ_t =1)，下跌幅度較其餘基準投資組合小，由此可知，依據市場變化調整風險趨避參數求得之投資組合績效表現應能勝過基準投資組合。

DOI:10.6814/NCCU202100814

進行實證結果前，可先觀察強化學習依據市場變化正確調整每一天的風險趨避參數之績效表現如何。因我們希望強化學習得到正獎勵越多越好，所以本研究定義正確答案投資組合，為在第 t 天選擇第 t＋1 天報酬率較高之投資組合所對應風險趨避參數 λ_t，圖 4.5 為不同歷史報酬波動天數 m 下的正確答案投資組合累加報酬圖，由圖可知，不管歷史報酬波動天數 m 為多少，累加報酬率皆為穩定向上，代表假如強化學習能正確地調整每一天風險趨避參數，能獲得相當



Note：圖為不同歷史報酬波動天數 m 下，正確答案投資組合在 2004 年到 2021 年累加報酬。

圖 4.5：不同歷史報酬波動天數之正確答案投資組合累加報酬圖

可觀的報酬，另外從相關係數來看，台灣 50 指數報酬率和正確答案投資組合中風險趨避參數之相關係數皆為負，其中與正確答案投資組合 (m＝5) 相關係數為 -0.104，與正確答案投資組合 (m＝20) 相關係數為 -0.198，與正確答案投資組合 (m＝60) 相關係數為 -0.278，代表當台灣 50 指數報酬越大，風險趨避參數傾向越小，而當台灣 50 指數報酬越小時，風險趨避參數傾向越大。由以上結果可知，依據市場變化動態調整每一天風險趨避參數 λ_t 有其必要，接下來，分析本研究建置的強化學習投資組合是否能學到此關係，準確調整每一天風險趨避參數。

4.3.2 強化學習投資組合績效分析

在 4.3.2 小節，會先介紹強化學習之訓練結果和測試結果，比較強化學習投資組合和基準投資組合之績效，其中設歷史狀態天數 n 和歷史報酬波動天數 m

DOI:10.6814/NCCU202100814

一致，因為當均數-變異數模型使用過去 m 天計算歷史報酬率平均數和變異數，我們預期強化學習需輸入對應歷史狀態天數 n 才能動態調整風險趨避參數。最後觀察不同參數組合 n 和 m 之結果，與上述想法是否一致或不同。

本研究共訓練 84 種模型，訓練 10 年後，分別測試 2014 年第一季至 2020 年第四季共 28 季，乘上不同歷史狀態天數 n 為 5、20 和 60 天，因訓練模型總共有 84 種，無法列出全部結果，故只列出 84 種模型平均期末累加報酬和預測準確率，如表 4.1，由表可知，當輸入模型資訊越多，也就是歷史狀態天數 n 越大，期末平均累加報酬和預測準確率越高，訓練結果越好。雖然訓練結果最高準確率只有 0.7362，最低為 0.5730，但平均期末累加報酬都遠超過其餘四種基準投資組合，說明模型有學到如何動態調整風險趨避參數，另外，因為正確答案為每一期選擇對應報酬率較高之風險趨避參數，故獎勵皆為正，預測準確率為 1。

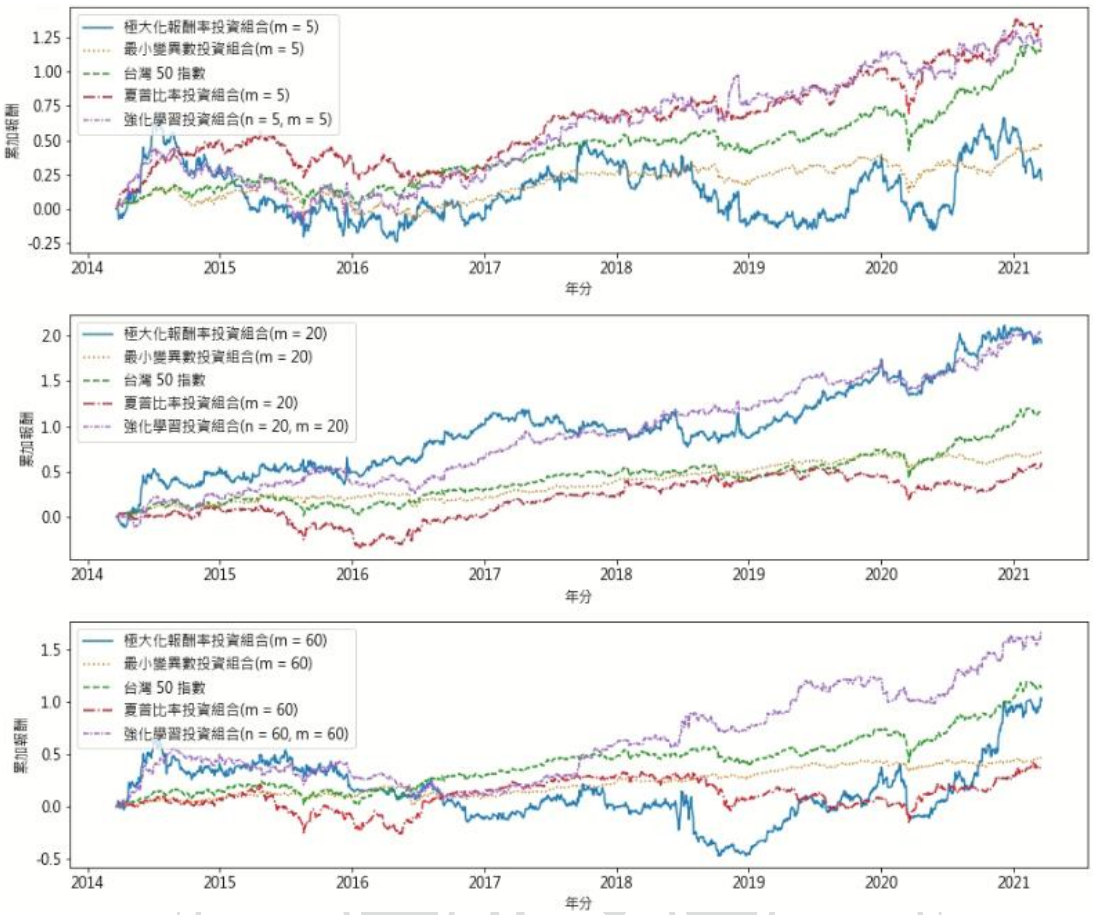
表 4.1：不同歷史狀態天數之平均訓練結果

	平均期末累加報酬	預測準確率
正確答案投資組合 (m = 5)	19.3708	1
正確答案投資組合 (m = 20)	21.1076	1
正確答案投資組合 (m = 60)	20.4598	1
強化學習投資組合(n = 5, m = 5)	2.8832	0.573
強化學習投資組合(n = 20, m = 20)	6.6691	0.6144
強化學習投資組合(n = 60, m = 60)	12.583	0.7359

Note：表為不同歷史狀態天數 n 和歷史報酬波動天數 m 下，強化學習投資組合 84 種訓練平均結果和正確答案投資組合比較，結果包含期末累加報酬和預測準確率。

本研究測試不同歷史狀態天數 n 為 5、20 和 60 天，2014 年第一季至 2020 年第四季共 28 季，圖 4.6 為不同歷史狀態天數之強化學習投資組合與其餘四種基準投資組合累加報酬比較圖，首先，由圖可知，不論歷史狀態天數

DOI:10.6814/NCCU202100814



Note：圖為比較在固定歷史報酬波動天數 m 下，不同歷史狀態天數 n 下強化學習投資組合測試集結果和四種基準投資組合累加報酬比較。

圖 4.6：不同投資組合累加報酬比較圖

n 為多少，強化學習投資組合績效皆比固定風險趨避參數之投資組合還要好，但只有當 $n = 20$ 時，強化學習投資組合跟極大化報酬率投資組合 $(\lambda_t = 0)$ 累加報酬相當，本文推測原因如下，在測試集期間，歷史報酬波動天數 $m = 20$ 下極大化報酬率投資組合 $(\lambda_t = 0)$ 之累加報酬並不如 $m = 5$ 和 $m = 60$ 時上下波動，而是呈現向上的趨勢，說明測試集間風險趨避參數與景氣並不如預期呈現負相關，導致訓練好的模型無法在這段期間動態調整風險趨避參數，而當歷史狀態天數 $n = 5$ 和 $n = 60$ 時可以看出，在某些時期，如 2020 年初時，強化學習投資組合傾向預測風險趨避參數為 1，避開由疫情導致的大跌。在 2014 年初和 2019 年，台灣 50 指數呈上升趨勢，強化學習投資組合傾向預測風險趨避參數為 0，而非選擇較保守的情況，所以有較高的累加報酬。除了與固定風險趨避參

DOI:10.6814/NCCU202100814

數比較外，當歷史狀態天數 n = 20 和 n = 60 時，強化學習投資組合績效贏過夏普比率投資組合，當 n = 5 時，測試結果與之相當，代表本研究選擇極大化報酬或極小化風險之極端策略勝過同時考慮報酬及風險之夏普比率投資組合，說明模型有依據市場變化，動態調整風險趨避參數，在大漲時報酬率較夏普比率投資組合高，在大跌時損失較少。

除了衡量累加報酬外，表 4.2 為不同投資組合績效比較表，整體來說，因最

表 4.2：不同投資組合績效比較表

		年化報酬率	年化標準差	夏普比率	索提諾比率
	台灣 50 指數	0.1679	0.1603	1.0479	1.4106
n = 5, m = 5	極大化報酬率投資組合	0.0309	0.3865	0.0798	0.1172
	最小變異數投資組合	0.0689	0.1349	0.5109	0.6641
	夏普比率投資組合	0.1970	0.2015	0.9778	1.3806
	強化學習投資組合	0.1715	0.2709	0.6331	0.8921
n = 20, m = 20	極大化報酬率投資組合	0.2820	0.3616	0.7798	1.1654
	最小變異數投資組合	0.1056	0.1042	1.0129	1.1737
	夏普比率投資組合	0.0882	0.1932	0.4562	0.6341
	強化學習投資組合	0.3008	0.2563	1.1734	1.5601
n = 60, m = 60	極大化報酬率投資組合	0.1490	0.3800	0.3922	0.5575
	最小變異數投資組合	0.0680	0.0948	0.7176	0.7761
	夏普比率投資組合	0.0529	0.2000	0.2647	0.3619
	強化學習投資組合	0.2437	0.2622	0.9295	1.2756

Note：表為比較在固定歷史報酬波動天數 m 下，不同歷史狀態天數 n 下強化學習投資組合測試集結果和四種基準績效比較，績效衡量指標包括年化報酬率、年化標準差、夏普比率和索提諾比率。

小變異數投資組合目標式為極小化風險，故年化標準差最小。而強化學習投資組合因為每一天動態調整極大化報酬率投資組合和最小變異數投資組合，所以年化標準差介於兩者之間。另外，除了參數 n, m = 20 下強化學習投資組合，台灣 50 指數在夏普比率和索提諾比率都贏過其餘投資組合，因在測試集 2014 年到 2021 年中，台灣 50 指數除了 2020 年因疫情大跌外無大幅下跌，導致波動較

DOI:10.6814/NCCU202100814

小，故在夏普比率和索提諾比率都贏過強化學習投資組合，而在參數 n, m = 20 下，極大化報酬率投資組合績效表現異常的好，導致強化學習投資組合績效贏過台灣 50 指數。綜觀以上結果，雖然本研究強化學習投資組合並無明顯贏過台灣 50 指數，但績效卻都遠遠贏過固定風險趨避參數之投資組合，說明強化學習有學到動態調整風險趨避參數。

除了上述績效比較外，也可利用預測準確率，觀察強化學習投資組合是否選到正確的風險趨避參數，表 4.3 為不同歷史狀態天數之強化學習投資組合預測準確率，由表可知，預測準確率都落在 0.5 附近，說明與正確答案投資組合相比，

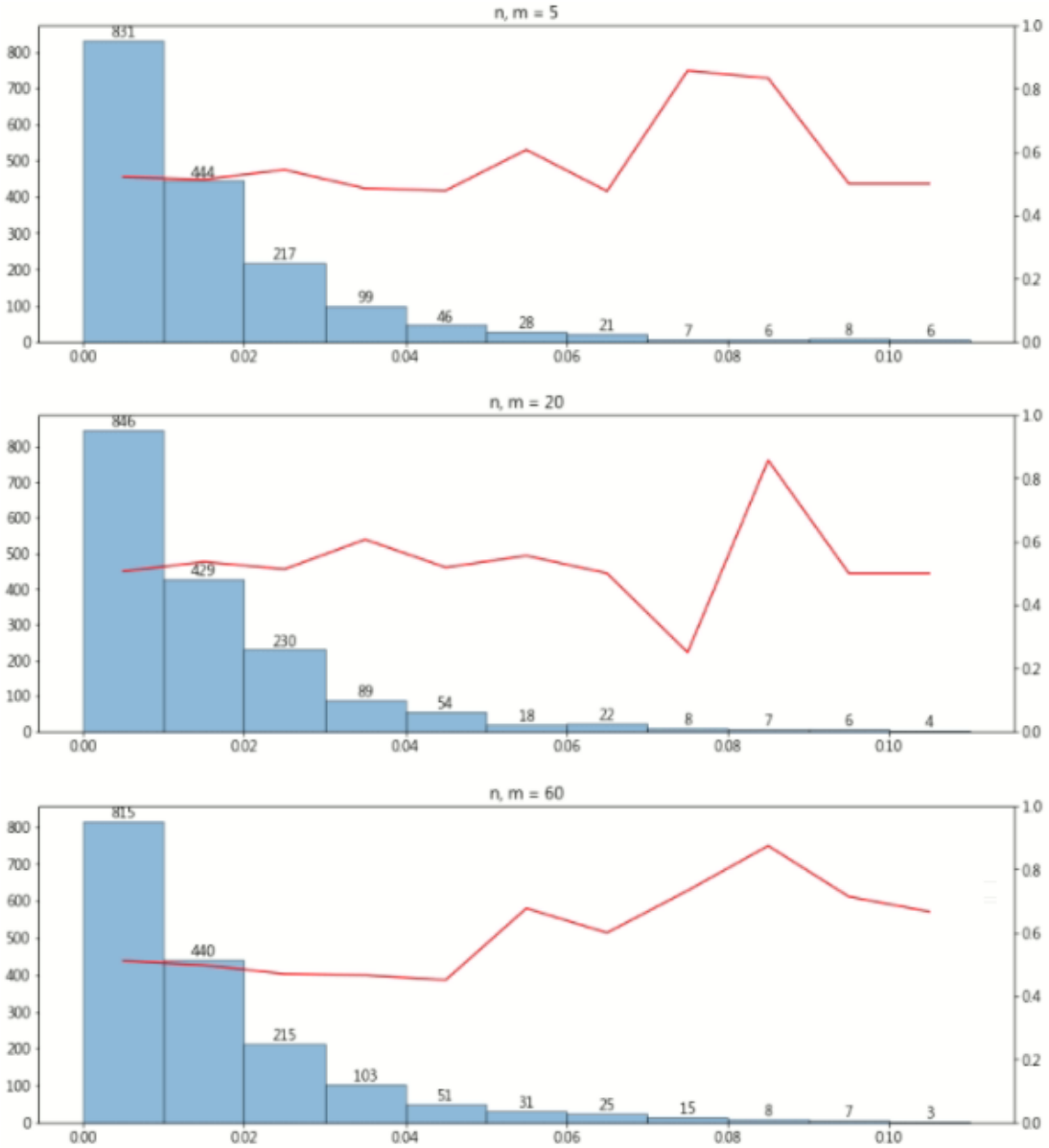
表 4.3：不同歷史狀態天數之強化學習投資組合預測準確率

	預測準確率
強化學習投資組合(n = 5, m = 5)	0.5219
強化學習投資組合(n = 20, m = 20)	0.5207
強化學習投資組合(n = 60, m = 60)	0.5073

強化學習投資組合只選對大約一半風險趨避參數，雖預測準確率結果很差，但強化學習投資組合累加報酬還是贏過固定風險趨避參數之投資組合，由圖 4.7 可解釋原因，橫軸為極大化報酬率投資組合 ($\lambda_t = 0$) 和最小變異數投資組合 ($\lambda_t = 1$) 報酬率差距並取絕對值，縱軸左邊為個數，右邊代表準確率，直方圖代表在此測試集中報酬率差距的個數，折線圖代表不同報酬率差距對應之預測準確率，由圖可知，模型面對大部分的情況下，也就是報酬率差距為 5% 以下，預測準確率較低，原因為本研究強化學習獎勵函數之設計，當極大化報酬率投資組合 ($\lambda_t = 0$) 和最小變異數投資組合 ($\lambda_t = 1$) 報酬率差距不大時，對應獎勵較小，代表選擇風險趨避參數 0 和 1 結果類似，故強化學習預測準確率大約 0.5，但當報酬率差距擴大時，對應獎勵較大，預測準確率大幅升高，使強化學習較能選到對應的風險趨避參數，雖然大漲大跌容易預測正確，但其對應天數較少，所以整體來說，強化學習預測準確率只有 0.5，但累加報酬還是贏過固定風險趨避參數之

DOI:10.6814/NCCU202100814

投資組合，說明強化學習在市場變化波動較大時能動態調整風險趨避參數。而當歷史狀態天數為 $n = 5$ 和 $n = 60$ 時，強化學習投資組合與固定風險趨避參數之投資組合累加報酬相差較大，此現象更為明顯。



Note：圖為不同參數 n, m 下強化學習投資組合之結果，直方圖代表極大化報酬率投資組合 ($\lambda_t = 0$) 和最小變異數投資組合 ($\lambda_t = 1$) 報酬率間差距的個數，折線圖代表對應的預測準確率。

圖 4.7：歷史狀態天數與對應歷史報酬波動天數之固定參數投資組合間報酬差距累積圖及相應預測準確率

本研究強化學習架構中，影響結果重要的參數有兩部分，分別為歷史狀態天

DOI:10.6814/NCCU202100814

數 n 及歷史報酬波動天數 m 如何決定。當歷史狀態天數 n 不同，代表模型接收到不同時間周期的資訊，會影響模型如何學習及判斷，而歷史報酬波動天數 m 會影響均數-變異數模型中報酬率平均數和標準差的計算，當歷史報酬波動天數 m 越大，代表考慮的天期越長，均數-變異數模型較不受短期波動影響。在實證結果都設歷史狀態天數 n 和歷史報酬波動天數 m 一致，接下來，我們探討歷史狀態天數 n 及歷史報酬波動天數 m 不同組合下與實證結果是否一致或有不同現象。

表 4.4 為不同參數間測試集強化學習投資組合之累加報酬比較，由表可知，在固定歷史報酬波動天數 m 下，當歷史狀態天數與歷史報酬波動天數一致時，強化學習投資組合期末累加報酬都較高，強化學習較能準確調整風險趨避參數。

表 4.4：不同歷史狀態天數與歷史報酬波動天數之期末累加報酬

歷史報酬波動天數	歷史狀態天數	期末累加報酬	夏普比率
m = 5	n = 5	1.1663	0.6331
	n = 20	-0.2162	-0.1246
	n = 60	0.9594	0.5387
m = 20	n = 5	1.2999	0.7724
	n = 20	2.0458	1.1734
	n = 60	1.4018	0.8139
m = 60	n = 5	0.0835	0.0476
	n = 20	0.5064	0.2889
	n = 60	1.6578	0.9295

Note：表為固定歷史報酬波動天數 m 下，不同歷史狀態天數 n 之強化學習投資組合測試結果，結果包含期末累加報酬和夏普比率。

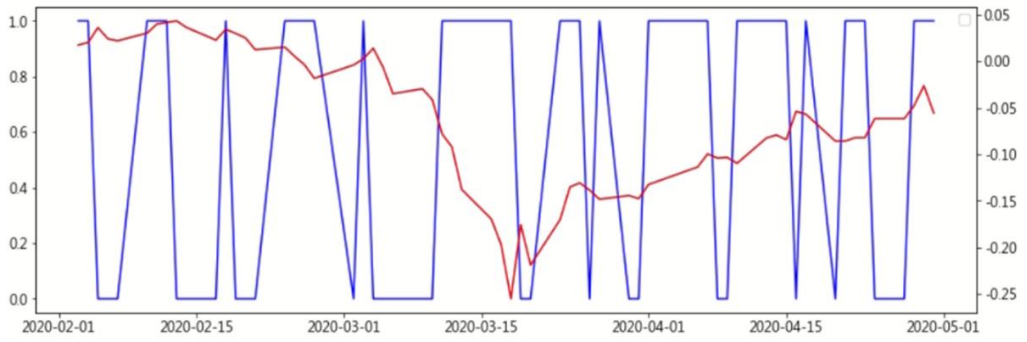
由表 4.4 可得知歷史狀態天數 n 和歷史報酬波動天數 m 不同組合，會影響強化學習是否能準確動態調整風險趨避參數。當歷史狀態天數 n 及歷史報酬波動天數 m 一致時，強化學習才能有效調整風險趨避參數，本文認為因均數-變異數模型考慮天期不同，所使用之報酬率平均數及標準差波動也會不同，以 m =

DOI:10.6814/NCCU202100814

60 之均數-變異數模型為例，因輸入過去天期較長，所使用報酬率平均數及標準差波動也會較小，而當強化學習輸入較短期資訊時，如歷史狀態天數 $n = 5$ 或 $n = 20$，無法有效捕捉到長期的趨勢，故本文認為在選擇參數上，歷史狀態天數及歷史報酬波動天數所考慮天數應一致，強化學習才能準確地調整風險趨避參數。

4.3.3 風險趨避程度與股價景氣之關係

由 Rosenberg & Engle (2001) 實證結果可知，風險趨避程度與景氣呈負相關，而由上述實證結果顯示，強化學習在市場變化較大時能動態調整風險趨避參數。故觀察利用強化學習求得之風險趨避參數，與景氣是否如文獻結果呈負相關，圖 4.8 為參數 $n, m = 60$ 下，2020 年 2 月至 2020 年 4 月，強化學習求得之風險趨避參數與台灣 50 指數累加報酬比較圖，由圖可知，在 2020 年 3 月台股因



Note：圖為 2020 年 2 月至 2020 年 4 月，歷史狀態天數 n 和歷史報酬波動天數 m 為 60 下，強化學習投資組合之風險趨避參數與台灣 50 指數累加報酬比較。

圖 4.8：風險趨避參數與台灣 50 指數累加報酬比較圖

為疫情大跌時，風險趨避參數傾向為 1，代表當市場大幅下跌時，強化學習傾向選擇最小變異數投資組合 ($\lambda_t = 1$)，風險趨避參數較大。由相關係數可知，在此期間風險趨避參數和台灣 50 指數累加報酬相關係數為 -0.23，說明強化學習求得之風險趨避參數與景氣呈負相關，與 Rosenberg & Engle (2001) 實證結果一致，當市場表現好時，投資人會想承擔多一點風險，來獲取較高的報酬，反之，當市場表現差時，投資人會想承擔少一點風險，來避免資產的損失。

DOI:10.6814/NCCU202100814

第五章 結論與未來展望

過往均數-變異數模型中的風險趨避參數難以在實務中衡量，而由 Rosenberg & Engle (2001) 可知風險趨避程度與景氣呈負相關，故本文利用強化學習，依據台股市場變化動態調整風險趨避參數。實證結果顯示，強化學習投資組合績效皆贏過固定風險趨避參數之投資組合，表示強化學習在不同時期下能有效調整投資人風險偏好，配置最適的投資組合。另外，由於本研究獎勵函數的設計，導致當極大化報酬率投資組合 ($\lambda_t = 0$) 和最小變異數投資組合 ($\lambda_t = 1$) 報酬率差距擴大時，得到的獎勵較大，強化學習預測準確率大幅提高，說明在市場大漲大跌時，投資人風險偏好較明顯，強化學習能有效地動態調整風險趨避參數。

由後續不同參數測試發現，影響強化學習表現的關鍵因素為不同歷史狀態天數 n 和歷史報酬波動天數 m 之組合有關，建構投資組合時，不同歷史報酬波動天數之均數-變異數模型代表考慮不同時間週期的價格走勢，面對不同時間週期的均數-變異數模型，強化學習需輸入對應歷史狀態天數，才能捕捉到此價格走勢，準確調整風險趨避參數。

最後從台股因疫情下跌期間可以發現，風險趨避程度與景氣表現呈負向關係，如同 Rosenberg & Engle (2001) 實證結果，當市場表現好時，投資人會想承擔多一點風險，來獲取較高的報酬，反之，當市場表現差時，投資人會想承擔少一點風險，來避免資產的損失。

未來研究方向可分為兩點，第一點為將投資人現有財富加入強化學習架構，由 Bjork et al. (2011) 可知，投資人現有財富與風險趨避參數呈負相關，故可在狀態中增加投資人現有財富此一變數或設置新的獎勵函數考慮此關係，讓模型能學到風險趨避參數除了與景氣有關外，也與投資人自身財富有關連，模型更能準確地調整風險趨避參數。

DOI:10.6814/NCCU202100814

第二點為強化學習除了選擇極大化報酬率投資組合 ($\lambda_t = 0$) 和最小變異數投資組合 ($\lambda_t = 1$) 外，也可加入夏普比率投資組合，期望在市場大漲時選擇極大化報酬率投資組合 ($\lambda_t = 0$)，在市場大跌時選擇最小變異數投資組合 ($\lambda_t = 1$)，在市場變動不大下選擇夏普比率投資組合，藉此改善強化學習在市場波動不大時預測準確率過低的問題。

除了上述未來發展方向外，模型未來也可做改善，主要可分為兩部分，第一部分為均數-變異數模型，未來應使用新增限制式之均數-變異數模型，限制模型只能投資固定公司家數和考慮手續費，使本研究更符合實務應用。第二部分為強化學習模型，在參數設置上，本研究只探討不同歷史狀態天數 n 和不同歷史報酬波動天數 m 對結果之影響，對於其餘模型參數都沒做過多的調整，如神經網路架構和神經元個數、PPO 中的超參數，未來應著重模型參數設定之測試，以找到最適模型動態調整風險趨避參數。

DOI:10.6814/NCCU202100814

參考文獻

中文文獻

[1] 劉上瑋 (2017)。深度增強學習在動態資產配置上之應用 ：以美國 ETF 為例。國立政治大學金融研究所碩士論文。

英文文獻

[1] Basak, S., & Chabakauri, G. (2010). Dynamic mean-variance asset allocation. *The Review of Financial Studies*, *23*(8), 2970-3016.

[2] Björk, T., Murgoci, A., & Zhou, X. Y. (2014). Mean–variance portfolio optimization with state-dependent risk aversion. *Mathematical Finance: An International Journal of Mathematics, Statistics and Financial Economics*, *24*(1), 1-24.

[3] Díaz, A., & Esparcia, C. (2019). Assessing risk aversion from the investor's point of view. *Frontiers in psychology*, *10*, 1490.

[4] Gold, C. (2003, March). FX trading via recurrent reinforcement learning. In *2003 IEEE International Conference on Computational Intelligence for Financial Engineering, 2003. Proceedings.* (pp. 363-370). IEEE.

[5] Jiang, Z., & Liang, J. (2017, September). Cryptocurrency portfolio management with deep reinforcement learning. In *2017 Intelligent Systems Conference (IntelliSys)* (pp. 905-913). IEEE.

[6] Li, Y., & Li, Z. (2013). Optimal time-consistent investment and reinsurance strategies for mean–variance insurers with state dependent risk aversion. *Insurance: Mathematics and Economics*, *53*(1), 86-97.

[7] Markowitz, H. (1959). Portfolio selection. *Journal of Finance*, 7, 77–98.

[8] Moody, J., & Saffell, M. (2001). Learning to trade via direct reinforcement. *IEEE Transactions on Neural Networks*, *12*(4), 875-889.

[9] Neuneier, R. (1998). Enhancing Q-learning for optimal asset allocation. *Advances in neural information processing systems* (pp. 936-942).

[10] Rosenberg, J. V., & Engle, R. F. (2002). Empirical pricing kernels. *Journal of Financial Economics*, *64*(3), 341-372.

DOI:10.6814/NCCU202100814

[11] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

[12] Zhang, Y., Wu, Y., Li, S., & Wiwatanapataphee, B. (2017). Mean-variance asset liability management with state-dependent risk aversion. *North American Actuarial Journal*, *21*(1), 87-106.

[13] 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/NCCU202100814