基於多智能體生成對抗模仿學習開發強化式造市交易策略

摘要

選擇權造市策略是近年財金領域研究的重要議題，聚焦於如何做出最佳的掛單行為，在取得價格優勢的同時最大化價差獲利。不同於財務工程中設計精準定價模型的做法，本計畫提出一套不需要繁瑣設定，透過模仿他人行為獲取具獲利性的交易策略。本計畫主要採用多智能體模仿學習，模仿真實世界中不同造市商的造市策略，再利用生成式學習探索更高獲利的行為，進而建構出更具優勢的造市交易策略。

Abstract

Options market-making strategy has emerged as a significant research topic in the field of finance in recent years, focusing on how to execute optimal order placement to gain price advantages while maximizing profit from bid-ask spreads. Unlike the approach of designing precise pricing models in financial engineering, this project proposes a method that does not require intricate configurations, aiming to acquire profitable trading strategies by imitating the behavior of other market makers.

This project primarily employs multi-agent imitation learning, imitating the market-making strategies of various real-world market makers. It then utilizes generative learning to explore behaviors that yield higher profits, thus constructing a more advantageous market-making trading strategy.

目錄

圖目錄

表目錄

壹、 緒論

1.1 研究背景

選擇權是一種常見的衍生性金融商品，其定義為買方支付權利金後，便有在特定時間用指定價格買進或賣出標的資產的權利。作為一種衍生性金融商品，更為常見的選擇權交易是在次級市場中，針對選擇權合約本身進行的交易。選擇權合約的交易方法為申報買進價(bid price)或賣出價(ask price)及欲交易口數，再透過集中交易所根據「價格優先、時間優先」的撮合原則決定成交，即較高買進價或較低賣出價者優先撮合，若價格相同則依下單順序進行撮合。

在選擇權市場中，除了一般的交易人之外，還有一群特殊的市場參與者，我們稱之為造市商(Market Maker)。造市商負責提供買賣雙向報價，建立市場基本流動性，進而活絡市場交易。為確保造市者充分發揮其功能，台灣期貨交易所規範選擇權造市商需在交易人詢價後的 20 秒內提供買賣報價；而為增加造市者誘因，國內期貨市場提供予造市者之優惠為期交所手續費折減。較少的手續費使造市商具有交易摩擦較小的優勢，若再透過高頻交易，即可透過賺取微小的買賣價差來累積巨大的獲利。

1.2 研究動機

造市商藉由買賣報價賺取價差利潤，但要成為新進造市商卻會碰到許多難題。首先，商品報價的金額多由定價模型來決定，而複雜的衍生性金融商品定價模型牽涉到多種參數設定及估計。其次，由於選擇權市場存在多個造市商，只有具競爭優勢的造市單才能順利撮合，而作為新進造市商，在面臨造市經驗不足，或定價模型不具有足夠的優勢之下，報價出的造市單很可能無法與他人競爭而順利成交。

為了避免從頭建置定價模型帶來的困難，我們提出了一種解法：向現有的造市商學習造市策略。在我們取得的台指選擇權造市資料中，占比最大的三家造市商分別為澳帝華期貨、法銀巴黎證券以及元大證券。其中前兩家皆為外資造市商，進行多國的造市策略並持續獲利，可驗證其交易策略足夠穩定，能夠在不同國家的商品都獲得競價優勢。然而交易策略屬於商業機密，他人無法直接取得相關交易細節並學習，因此我們希望能透過模仿三家造市商過去的造市行為，從中反推出交易策略，讓下單策略能夠在取得價格優勢的同時最大化價差獲利。

1.3 研究目的

本研究所探討的問題為從其他造市商的交易策略中找尋具有優勢的掛單點位。我們採用強化學習中的模仿學習(Imitation Learning)，並採取多智能體模型，

讓三個智能體與由台指選擇權五檔報價歷史資料所建構出的模擬交易環境互動，並以最佳化智能體做出的委託單與其代表的實際造市商委託單之相似程度為目標，模擬出貼近個別真實造市商的下單行為。再利用生成式學習產生多種樣本，使智能體探索並提升策略，獲得一套報酬率較高的投資策略。在這篇計畫中我們有以下貢獻：

1. 提出一套方法，在不需要設計複雜的定價模型時，也可以得到能成功獲利的交易策略。
2. 引入生成式學習，探索過往交易資料中未有的行為，使策略取得更高的獲利性。

貳、 文獻探討

本研究的核心技術為深度強化學習(Deep Reinforcement Learning, DRL)之中的模仿學習。以下將針對人工智慧應用於交易的歷史與模型進行簡述。

2.1 強化學習

強化學習(Reinforce Learning)是智能體(agent)和環境(environment)互動並學習最佳策略(policy)的一種演算法，通常應用在順序決策(sequential decision making)的任務。在強化學習的框架中，智能體會觀察環境的狀態(state)並做出對應的動作(action)，環境會依現在的狀態和動作給予對應的獎勵(reward)，並依狀態轉移(state transition)機率轉移到新的狀態，而智能體的目標便是最大化得到的獎勵[1]。 和一般的監督式學習(Supervised Learning)不同，強化學習需要依序和環境互動(即順序決策)，每次動作不僅會影響獲得的獎勵，動態的環境同時也會有不同的轉變。

深度強化學習為強化學習的一種改進。DRL 將深度類神經網路(Deep Neural Network)應用在強化學習上，不僅增強模型在函數逼近(function approximation)上的精確度，更能應用在高維度的資料集上，增加模型的複雜度及準確率。如近期十分熱門的 AlphaGO、研究上廣泛應用的深度 Q 網路(Deep Q-Network, DQN)等。

目前強化學習被廣泛應用在各類的問題上，如電玩[2]、棋類遊戲[3]、機器人[4]等。而在財金領域，強化學習的應用也是近期研究的重點之一，相關研究有注重在整合技術指標和價量資訊，改進交易策略以及產生交易訊號等[5][6]、預測價格或趨勢[7]、投資組合管理和最佳化[8][9]、造市等。

2.2 模仿學習

模仿學習是強化學習的一個分支。一般的強化學習會以已知的獎勵函式給予智能體獎勵，目標是最佳化智能體的策略。模仿學習則相反，給予一組專家數據，但獎勵函式未知，期望智能體能做出接近專家數據的行為[10]。

2.2.1 行為複製

行為複製是在真實世界蒐集多組環境狀態跟專家行為作為資料集，將狀態輸入至類神經網路後，比對輸出值與該狀態對應的專家行為，透過反覆調整類神經元間連接的權重，習得一套用來制定軌跡(trajectories)的策略。

行為複製的框架和監督式學習類似，對智能體輸入當下的環境，目標是盡可能模仿出類似專家行為的解答，其優勢是智能體可以快速且有效的模仿出專家行為，且智能體不必和環境進行互動；但若要達到一般性的高正確率，行為複製需要蒐集大量又多元的資料集，才能確保專家數據以及學習到的策略會涵蓋大部分的狀況。而在強化學習的應用上，行為複製更有複合誤差(compounding error)的問題。由於強化學習強調順序決策，一旦在訓練過程出現和專家數據的偏差，便可能出現專家數據並未涵蓋的數據，導致行為複製失效或偏離。因此目前行為複製通常做為快速訓練模型起點的一種手段，如 AlphaGO 先模仿人類棋手的下棋邏輯，得到一套不錯的策略為基底後，再進行其餘的強化學習模型的訓練。

2.2.2 生成對抗模仿學習

為了避免逆向強化學習的高計算量，史丹佛大學的一個團隊在 2016 年提出了生成對抗模仿學習的框架[13]，迴避了逆向強化學習的繁雜步驟，同時確保在智能體學習的過程中，能直接從專家軌跡學習如何做出決策。生成對抗模仿學習的目標在於模仿專家策略的佔用度量(occupancy measure)，即希望智能體產生的軌跡之分布和專家策略的軌跡分布越接近越好。生成對抗模仿學習採用了生成對抗網路(Generative Adversarial Network, GAN)的框架，訓練一個生成器(generator)和判別器(discriminator)，前者用來產生策略，後者用來判定輸入的數據和專家軌跡的相像程度。此時判別器和獎勵函數的功能類似，而生成器也能直接依據和專家軌跡的相似程度來學習策略。生成對抗模仿學習不僅解決了行為複製並無和環境互動的缺陷，同時也改進了逆向強化學習的缺點，藉由生成器和判別器的對抗，使策略的佔用度量更接近專家策略的佔用度量，進而達成模仿學習的目的。

多智能體生成對抗模仿學習(Multi-Agent GAIL)[14]是生成對抗模仿學習的一種應用，但從單一智能體的框架轉成多智能體決策之問題。由於智能體間獎勵函數可能不一致，而有衝突或合作等不同情況，因此會產生多個均衡解，即單一智能體的最佳策略會仰賴其他的智能體決策。在多智能體生成對抗模仿學習的框架下，生成器會以分散式的方式控制所有智能體的決策，而判別器會判

定各智能體軌跡以及其對應的專家軌跡的相像程度。多智能體生成對抗模仿學習本質上仍是模仿學習，只是相較於先前所述的方法，更加擴展了問題應用的範圍以及複雜度，且對於存在複數策略的環境能有更好的應用。

2.3 強化學習應用於造市及其缺陷

造市商是指提供市場流動性的交易商。和一般的交易人不同，造市商並無投資方向性，亦即不靠做多或做空而獲利，而是透過控制手中買賣單的存貨，賺取買賣單之價差來獲利。在強化學習的框架中，智能體透過和市場的互動，觀察歷史價格、存貨、成交量等，決定報價的價格、數量等，提供報價並依規定進行搓合，在評估各式不同限制，如存貨、風險之下，嘗試獲取最大的報酬。由於造市過程有其時間順序性的限制，及大量且複雜的限價委託簿(limit order book, LOB)，使得應用深度強化學習於造市委託的研究逐漸受到重視。[15]

目前將強化學習應用在造市策略的研究，相較於其他金融領域，仍有極大的發展空間[16]。而在已知的研究文獻中，多數方法仍將模型建立在模擬的資料上，而非真實世界的資料集。如[17]採用了模擬資料與只以簡單規則交易的市場參與者，[18]採用了模擬資料，以及使用加入時間限制的造市策略。且大多數模型仍侷限在單一智能體的造市學習，與現實市場上的情況仍有歧異。[19]使用了真實世界的 10 種證券進行造市，但並未考慮其他造市商的競爭。[20]使用了兩個智能體進行造市，但兩者為共同產生最後決策，而非多智能體間的競爭。本研究採用了模仿學習的方式，並以多智能體的模仿學習，嘗試模擬真實世界中三大造市商的行為，我們期望透過模擬三家台指選造市商的掛單行為，嘗試結合各造市商的優勢策略並最佳化掛單邏輯及點位，期望獲得最大之報酬，再以模仿學習的策略為出發點，利用生成式學習來提升策略獲利。

參、 研究方法及步驟

3.1 研究架構

本研究主要分為兩個階段。第一階段為基於多智能體之行為複製造市策略，其目的是利用行為複製演算法，使三個智能體分別模仿對應的造市商，進而獲得該造市商的交易策略。此階段的流程如圖 1 所示，我們先針對資料進行前處理及篩選，並利用造市名單產生標籤資料。產生資料集後，先以預訓練集產生三個智能體的雛形，再透過多智能體行為複製演算法訓練智能體參與造市競爭。

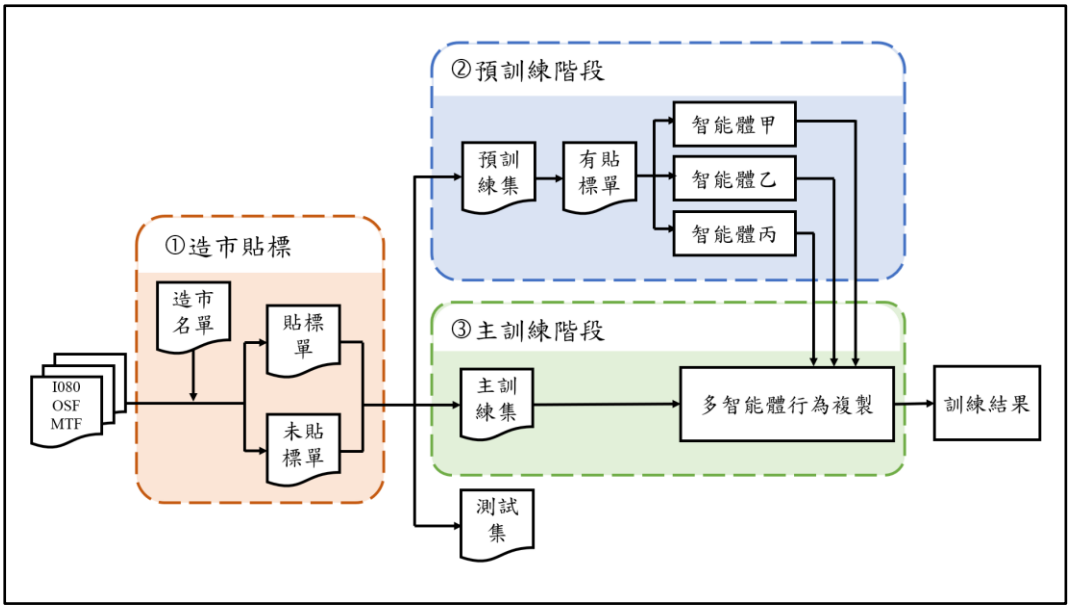

圖 1、基於多智能體之行為複製造市策略流程圖

第二階段為生成式造市策略與模擬交易，此部分旨在透過生成式學習產生多種樣本，探索更具獲利性的行為，進而提升實驗一習得之交易策略的成效。圖 2 為第二階段的流程，先針對造市策略加入擾動產生多種不同的軌跡，接著篩選獲利較原先軌跡更優勢的資料後，讓智能體再次學習，並進行模擬交易評估該策略的獲利性。

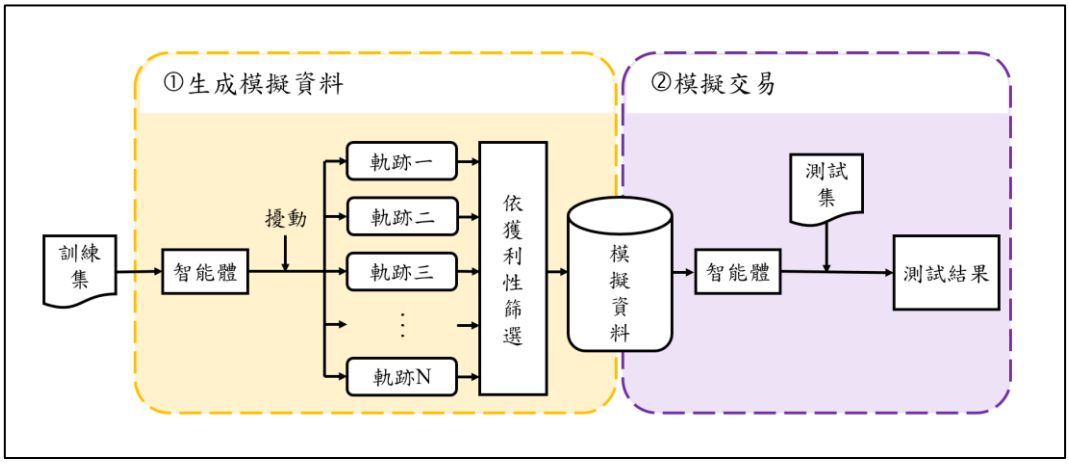

圖 2、生成式造市策略與模擬交易流程圖

3.2 資料蒐集與前處理

本研究以台灣期貨交易所提供 2017 年三月至六月的選擇權委託簿揭示訊息(I080)、選擇權委託檔(OSF)以及選擇權成交檔(MTF)來建置台指選擇權造市交易模型。除了上述三種資料外，本研究還利用台灣期交所的每日造市撮合名單，作為造市交易貼標的依據。

本研究針對每月近一月之台指選擇權之造市單，其履約價距離當日標的中價最近的各三檔月買賣選擇權 (每天挑選 6 種)，透過選擇權委託檔的委託時間欄位與選擇權委託簿揭示檔的時間欄位，找到每筆造市委託單最接近的揭示訊息，取得共 711229 筆造市委託行為資料。再依照期交所的台指選擇權權利金報價單位規定，轉換成以點(tick)為單位的買入價-賣出價配對(bid-ask pair)。

我們依時間順序來切分資料，前 3/10(2017/03/01~2017/04/12)為預訓練集，中間 3/10(2017/04/13~2017/05/22)為訓練集，後 4/10(2017/05/23~2017/06/30)為測試集。

3.3 馬可夫決策過程及特徵工程

3.3.1 馬可夫決策過程

在強化學習的過程中，智能體和環境互動的過程被定義成馬可夫決策過程(Markov Decision Process, MDP)，其假設智能體的動作只會受到當下的狀態所影響。此次實驗可以定義成透過找出最佳的策略，來解決一個馬可夫決策過程的問題。對於一個學習策略的智能體，MDP 被定義成一元組(tuple)： (S, A, P, γ, R)

- S為一組有限的狀態空間(State Space)
- $A = \{a_1, a_2, \cdots, a_k\}$為一系列動作的集合
- $P(s', s) = Pr(s'|s, a)$表示在s的情況下執行動作a，狀態轉移到s'的機率
- $\gamma \in [0,1)$為折扣因子(discount factor)
- $R(s, a)$為獎勵函式，表示在s下執行動作a時，智能體會獲得的獎勵

在時間為t時，智能體會觀察到環境的狀態s_t，依其資訊做出動作a_t，環境依獎勵函式$R(s_t, a_t)$給予獎勵r_t，同時環境依狀態轉移機率轉移至下一環境s_{t+1}，重複迭代後便會得到序列$(s_0, a_0, r_0, \cdots, s_t, a_t, r_t, \cdots)$[18]。

3.3.2 狀態空間

1. Order Imbalance(OI)：OI衡量的是在給定視窗大小 ,最佳五檔委買力度與最佳五檔委賣力度的差異。

$$OI = \frac{\sum_{i=1}^{5} p_i^{\text{buy}} q_i^{\text{buy}} - \sum_{i=1}^{5} p_i^{\text{sell}} q_i^{\text{sell}}}{\sum_{i=1}^{5} p_i^{\text{buy}} q_i^{\text{buy}} + \sum_{i=1}^{5} p_i^{\text{sell}} q_i^{\text{sell}}} \tag{1}$$

2. Time Frequency：距離上次掛單間隔的時間。

$$\Delta T_{t,n} = \text{time}_{t-n} - \text{time}_{t-n-1}, \text{where } n = 0,1, \ldots, w-1 \tag{2}$$

3. Mid price: 定義中價為最佳委買價與最佳委賣價的平均值。

$$P^{mid} = \frac{P^{b_1} + P^{a_1}}{2} \tag{3}$$

4. Mid Price Return: 定義 t 的價格變化為與 t-1 的增減幅。

$$return^{mid} = \frac{p_t^{\text{mid}} - p_{t-1}^{mid}}{p_{t-1}^{\text{mid}}} \tag{4}$$

5. Implied Volatility(σ): 利用 Black-Scholes Formula 推出該檔選擇權的隱含波動度
6. Greeks(Δ、Γ、Θ、ρ、V): 選擇權隱含波動度的延伸指標
7. Spread: 最佳賣價與最佳買價的差距

$$spread = P^{a_1} - P^{b_1} \tag{5}$$

8. Volume Imbalance(VI): VI衡量最佳五檔委買量與最佳五檔委賣量的差異。

$$VI = \frac{\sum_{i=1}^{5} q_i^{\text{buy}} - \sum_{i=1}^{5} q_i^{\text{sell}}}{\sum_{i=1}^{5} q_i^{\text{buy}} + \sum_{i=1}^{5} q_i^{\text{sell}}} \tag{6}$$

9. Weighted Average Price(WAP): 以對手委託量加權得出的價格，為一種判斷標的真實價格的方式

$$WAP = \frac{P^{b_1} * q^{a_1} + P^{a_1} * q^{b_1}}{q^{a_1} + q^{b_1}} \tag{7}$$

10. Put or Call(PC): 判斷該選擇權是 put or call
11. TTM: 選擇權與到期日的距離

Lookback window: 針對上述的 feature(PC、TTM、time frequency 以外)，分別計算 rolling window∈[10,20,30,40,50]的平均值以及標準差，列入 state，得到 state space 共 145 種。

3.3.3 智能體動作空間

本實驗設定買價差與賣價差各有 0(市價)與 1~5(分別表示最佳 5 檔)，共六種可能的離散行為，其中買價差為市場中價減去買價報價的值，而賣價差為賣價報價扣除市場中價之值。我們將一組買賣價差視為一個動作，智能體每次同時決定買價與賣價的點位。

Action ID	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Ask	0	1	2	3	4	5	0	1	2	3	4	5	0
Bid	0	0	0	0	0	0	1	1	1	1	1	1	2
Action ID	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**
Ask	1	2	3	4	5	0	1	2	3	4	5	0	1
Bid	2	2	2	2	2	3	3	3	3	3	3	4	4
Action ID	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**			
Ask	2	3	4	5	0	1	2	3	4	5			
Bid	4	4	4	4	5	5	5	5	5	5			

表 1、智能體動作空間

3.3.4 獎勵函式

給定環境狀態s以及專家數據$a \in [0,1,2,3,4,5]$，智能體的策略π及其參數θ，獎勵函式(損失函數)為

$$R_t = \frac{\sum_{i=1}^{n}[(a_i^{bid} - \pi_\theta(s_i)^{bid})^2 + (a_i^{ask} - \pi_\theta(s_i)^{ask})^2]}{n} \quad (8)$$

3.4 基於多智能體之行為複製造市策略

3.4.1 造市單貼標

我們擁有一份於 2017 年 3 月至 6 月，各造市商於各項商品的成交搓合紀錄。然而這份數據的時間粒度為日，因此無法精確對應至我們用於訓練的逐筆委託單資料。

我們首先將造市名單與成交檔合併，在每日的每種商品中，若有一個成交價只屬於某造市商，即假定此價格的成交單皆為該造市商的行為，根據此貼標邏輯，我們可獲的 0.1%的有標資料。除了價格行為外，我們也發現某些造市商具有特殊的委託量偏好，於是使用流水單號將成交檔與委託檔合併，藉由委託量去擴充貼標。根據此法有 4.6%的委託單可以區分出造市商。

3.4.2 訓練階段

由於智能體在訓練階段初期的策略非常仰賴初始的參數設定，為避免隨機初始化造成的不穩定性，我們先以只有標籤資料的預訓練集，分別針對三個智能體(對應真實世界的三大造市商)進行監督式學習，其損失函數定義於 3.3.4 節，目的為使智能體有初步的造市策略。由於預訓練集資料筆數較少，我們固定訓練輪數為 50000 輪，目標是使智能體過擬合預訓練資料，並以該結果做為主訓練階段的初始模型。我們以智能體(1)擬合澳帝華期貨、智能體(2)擬合法銀巴黎證券、智能體(3)擬合元大證券。

主訓練階段的核心邏輯為模仿學習。2.2.1 節介紹到行為複製是在真實世界大量蒐集環境狀態-專家行為配對，透過比對類神經網路在某狀態輸出的行為與該狀態對應的專家行為，反覆調整類神經網路，以習得策略。本實驗從當日所有造市委託單取得 3.3.2 節的狀態資料及 3.3.3 節的造市商真實下單行為，作為環境狀態-專家行為配對資料集。我們在每一回合的訓練中，使智能體比對其輸出值與專家行為來學習策略。以下為行為複製演算法[21]。

演算法 1 Behavior Cloning

Collect a set of trajectories demonstrated by the expert D
Select a policy representation π_θ
Select an objective function L
Optimize L w.r.t. the policy parameter θ using D
return optimized policy parameters θ

在多智能體的框架下，三個智能體皆會在同一狀態下給出對應的動作，然而我們認為一個智能體的動作在真實世界中只會是某單一造市商的行為，因此我們設計了一個評估行為相似性的方法，使得每次只有與專家行為最相近的智能體會受到更新。由於智能體皆已預訓練完成，每個智能體皆會有不同的策略，以判定機制決定學習的資料可使每個智能體學習和自己的策略相近的造市單，來達到區分造市商的目的。

偏向性

bid \ ask	0	1	2	3	4	5
0	0	1	2	3	4	5
1	-1	0	1	2	3	4
2	-2	-1	0	1	2	3
3	-3	-2	-1	0	1	2
4	-4	-3	-2	-1	0	1
5	-5	-4	-3	-2	-1	0

當專家行為 (3,1)的(偏向差，點位差)

bid \ ask	0	1	2	3	4	5
0	(-2,4)	(-1,3)	(0,2)	(1,1)	(2,2)	(3,3)
1	(-3,3)	(-2,2)	(-1,1)	(0,0)	(1,1)	(2,2)
2	(-4,4)	(-3,3)	(-2,2)	(-1,1)	(0,2)	(1,3)
3	(-5,5)	(-4,4)	(-3,3)	(-2,2)	(-1,3)	(0,4)
4	(-6,6)	(-5,5)	(-4,4)	(-3,3)	(-2,4)	(-1,5)
5	(-7,7)	(-6,6)	(-5,5)	(-4,4)	(-3,5)	(-2,6)

相似性排序

bid \ ask	0	1	2	3	4	5
0	12	9	2	4	6	7
1	13	11	8	1	4	6
2	15	13	11	8	2	5
3	16	15	13	11	9	3
4	17	16	15	13	12	10
5	18	17	16	15	14	13

圖 3、專家行為是(3,1)時，不同動作的相似度排序

圖 3 為專家行為是(3,1)的排序。首先，每一個買賣點位組合皆有偏向性，計算方式為買價點位減去賣價點位，針對每種組合皆可計算其與專家行為偏向性的差距。此外還需計算每種組合與專家行為的點位差距，算法為買價點位差之絕對值與賣價點位差之絕對值的總和。由於偏向性表示掛單者的買賣傾向，比點位差更為重要，因此我們先以偏向性排序相似度，相同偏向性時再依點位差排序，並只更新相似度排序最高的智能體。若有複數相似度排序最高的智能體，則全數進行更新。

3.5 生成式造市策略與模擬交易

3.5.1 生成軌跡

生成式造市策略的核心邏輯為產生多種軌跡並挑選較佳的軌跡，再以產生之軌跡產生新資料集，再重新進行訓練。產生軌跡的方式如下：我們在智能體產生報價時加入擾動，擾動的方式分為兩種：第一種是採用 ϵ-greedy 策略，使智能體在 ϵ=0.1 的機率下採取隨機的雙邊報價，此種策略可使不同軌跡有較大的差異，達到更強的探索效果；第二種是在既有軌跡上引入雜訊。由於智能體在產生決策時，是透過 Sigmoid 函數給出 0~1 的值，放大五倍後再四捨五入取整，進而給出 0~5 的整數值，以此對應市價及一到五檔的行為，我們在 Sigmoid 函數的輸出值上加入一個隨機變數$X \sim N(\mu = 0, \sigma = 0.1)$，使輸出值經過放大和取整後，有機會作出高於或低於原先報價一個檔位的動作。此種策略可使產出的軌跡接近於原始軌跡，在保有原本的邏輯下進行一定程度的探索。

3.5.2 模擬交易

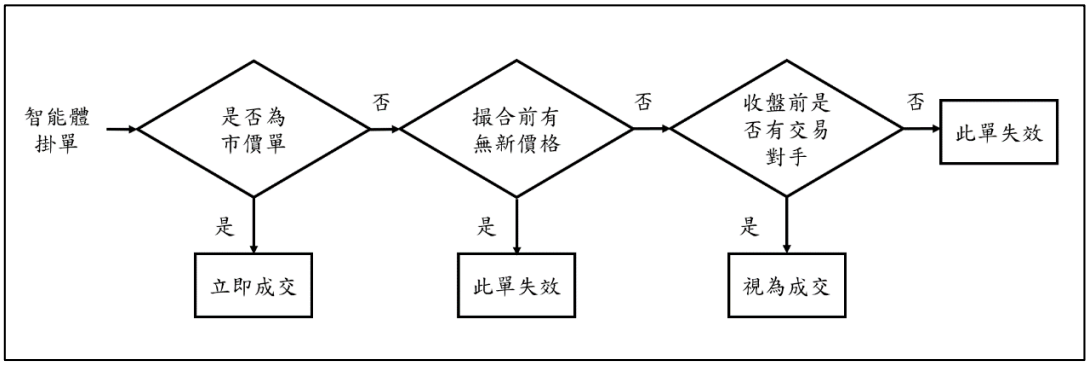

圖 4、模擬交易搓合邏輯

在模擬交易中，我們設定交易時段為上午八點五十分至下午一點二十五分，以避免開盤與收盤時價格波動劇烈導致的不穩定性。智能體在市場上存在造市單時，參考先前最近的揭示檔選擇委託價格進行掛單，行為包含市價成交或是選用一到五檔的價格作為報價，共有六種可能行為。撮合規則如圖 4 所示，假設智能體是採取市價單，此單立即成交。如果是限價單，為符合選擇權交易人詢價與造市者報價機制之規定，當同商品出現新價格時，舊的掛單視為無效。有效的限價單如果在當天收盤前有與之相配之對手價，例如揭示檔最佳買價高於或等於某掛單的賣價，則此單視為成交。為了方便判定以及增加成交機會，我們不考慮逐筆搓合的排隊機制，也不限定成交時，市價必須穿越委託價。倘若在收盤前此單始終無交易對手，則原有掛單視為無效掛單。若有成交，其成交價皆採用掛單時所報的價格，成交量限定在造市者報價之最低口數(20 口)，以消除不同成交量的影響。由於造市商極度注重流動性，因此當日收盤前如果智能體仍有存貨，將以當日收盤價進行出清。

3.5.3 實驗流程

此實驗的流程如下：我們以模仿學習後的模型參數來生成軌跡，每輪固定生成 100 個軌跡，然後以模擬交易環境檢驗軌跡的獲利性，針對每天的每種商品，挑出所有軌跡中有前 10%獲利的軌跡，拼湊成新的訓練集並進行訓練，再以模擬交易環境評估新訓練的模型獲利能力。如此進行三輪，並每次評估模型進步成效以及獲利能力。

肆、 實驗結果

4.1 基於多智能體之行為複製造市策略實驗結果

	澳帝華期貨	法銀巴黎證券	元大證券	標籤資料筆數
真實資料分布	37.36%	4.35%	58.28%	4066
智能體(1)	**40.69%**	3.44%	55.87%	810
智能體(2)	37.49%	**5.02%**	57.49%	1793
智能體(3)	35.24%	4.05%	**60.71%**	1463
提升	**+8.91%**	**+15%**	**+4.17%**	

表 2、行為複製實驗結果

表 2 為每個智能體學習到所有貼標的造市單中，三大造市商標籤資料的比例分布。真實資料分布是在貼標處理過後，三大造市商標籤資料在訓練集的比例分布。我們以智能體(1)擬合澳帝華期貨、智能體(2)擬合法銀巴黎證券、智能體(3)擬合元大證券，並用黑字標示對應的比例。由於進行模仿學習造市策略時，會藉由相似度判斷機制來決定智能體的訓練資料，使得智能體不一定會學習到所有訓練資料。各智能體的標籤資料筆數標示於右方。

由於貼標的造市單數量過少(只有訓練集的約 2%)，只依貼標造市單訓練並不能學習到準確且穩定的造市策略，因此我們使用所有的造市單訓練智能體，並以貼標的造市單來判定智能體可能代表的造市商。我們定義總體準確率為三個智能體在指定為一家造市商後，所學習到指定貼標造市單的比例。

在假設一個智能體代表一家造市商的前提下，我們希望每個智能體能只學習到一家造市商的造市單，在訓練智能體時，會使三個智能體同時掛單，由於正確的標籤單只會屬於一家造市商，在理想狀況下，只會有一個智能體學習到該標籤

單，依此便可使每個智能體都只會傾向擬合某家造市商。我們分別計算該智能體在學習到的貼標造市單中，三大造市商所佔的比例，如智能體(1)在其學習到的貼標造市單中，有 40.69%是屬於澳帝華期貨。如此便可計算三個智能體分別代表三個不同造市商時的準確率。

在各智能體學習到的樣本數量上，由於受到原始標籤的稀少問題，各智能體學習到的造市商標籤分布接近於原始標籤數據。若以現有的標籤資料解釋，三者習得指定的造市商標籤的比例皆有提升，在智能體間的比較中，以預訓練造市商標籤訓練的智能體，可以在訓練時競爭到該造市商最大的%數，顯示預訓練集可建立基礎的造市策略，解決多智能體初始網路參數的問題，並能藉由訓練集的相似度判斷，進一步強化智能體的策略。而若綜合三種策略，我們可以取得 35.48%的總體準確率，此數字由表格中粗體的數字，亦即我們預先指定的造市商進行平均而得。我們依照模型習得的造市單分布，來推測在市場上交易時，各模型皆可以依此分布獲得造市單。依此邏輯，比起盲目猜測僅能取得 33%的準確率，即三個智能體皆依市場分布進行猜測的綜合模型，也有略為的提升。

至此我們取得了三種不同的造市策略，分別來自於模仿不同的造市策略，我們以預訓練集的方式取得基礎造市策略，並以三者競爭的方式強化各智能體的策略，實驗結果顯示我們能提升區別各造市商的能力，同時三個智能體也以現有的造市商交易邏輯建立了一套交易策略。此實驗的假設建立在正確的擴充貼標資料，以及標籤資料的分布和真實市場一致的情況之下，礙於資料限制，我們並無法驗證擴充的貼標資料之正確性，也無法驗證其餘非標籤資料是否依照標籤資料的分布，和真實情況或許存在差距。

4.2 生成式造市策略與模擬交易實驗結果

	智能體(1)	智能體(2)	智能體(3)
行為複製階段	-20,089,466 點	-6,513,546 點	-4,896,396 點
	405,419 筆	424,179 筆	261,632 筆
第一輪生成	-3,225,984 點	-9,832,194 點	-3,763,798 點
	262,498 筆	333,980 筆	184,534 筆
第二輪生成	**1,344,204 點**	-7,930,066 點	-3,760,752 點
	199,604 筆	304,309 筆	178,508 筆
第三輪生成	116,490 點	**-3,435,768 點**	**-3,323,818 點**

	180,039 筆	284,620 筆	177,322 筆

表 3、ϵ-greedy 生成器之實驗結果(單位:點，1 點=50 元)

	智能體(1)	智能體(2)	智能體(3)
行為複製階段	-20,089,466 點	**-6,513,546 點**	-4,896,396 點
	405,419 筆	424,179 筆	261,632 筆
第一輪生成	**-436,780 點**	-12,145,030 點	-5,627,618 點
	313,172 筆	323,152 筆	205,816 筆
第二輪生成	-1,704,744 點	-10,205,970 點	-4,120,174 點
	269,476 筆	297,079 筆	189,468 筆
第三輪生成	-1,321,742 點	-14,659,878 點	**-3,658,976 點**
	237,115 筆	279,447 筆	179,914 筆

表 4、雜訊生成器之實驗結果(單位:點，1 點=50 元)

上表為不同代理人以兩種擾動方式所產生的軌跡後，每輪的模擬交易結果。由於各選擇權價格基準不同，會產生不同跳動的幅度，我們依照台灣期貨交易所的規範，統一轉換成點數的形式，此外，我們以測試集的所有造市單進行委託和模擬交易，取得的結果是-3,631,094 點，我們以此為此實驗的基準，來判斷經模型篩選後的交易軌跡是否有更佳的獲利。由於交易筆數也是考慮獲利時非常重要的依據，我們將交易筆數列在獲利下方，同時以黑字標示各智能體在此實驗的最佳獲利。依照 3.5.3 節的流程，我們將實驗一的模型和三輪生成軌跡進行獲利能力比較。

行為複製階段中所習得的造市策略中，由於不同造市策略可能產生掛單點為的重疊，智能體可能擬合到不同造市商的造市策略，使得過程可能學習到許多雜訊。智能體 1 和智能體 2 中，兩者作出類似的交易筆數，但獲利點數卻相差甚大，顯示何時進場會顯著的影響獲利。另外，行為複製階段中，三個智能體皆無法贏過基準，顯示原先策略確實具有很大的改善基礎。

在 ϵ-greedy 生成器的結果下，模型的表現皆有進一步的提升，相較於基準，也有略為的提升。掛單數量減少的情形可能是因為亂數決定動作時，會使模型做出更不易成交的決策，若學習到避開大量虧損的交易決策，如此便可更進一步提升獲利。ϵ-greedy 生成器具有更強的探索能力，有助於學習到更加的進場時間，

同時智能體(1)也成功地取得了正獲利。而在雜訊生成器的結果下，只有智能體(1)和智能體(3)有較行為複製的軌跡更好的表現。由於雜訊生成器相較於 ϵ-greedy 生成器更仰賴初始軌跡的好壞，推測是在行為複製時的結果尚未達到一個較好的造市策略，導致雜訊生成器表現較差。同時雜訊生成器所產生的軌跡相較 ϵ-greedy 生成器，其減少的交易次數也較少，但獲利性卻並未贏過 ϵ-greedy 生成器產生的軌跡，代表選擇正確的交易時間，避開可能會虧損的交易，是造市商和造市策略非常重要的一環，而 ϵ-greedy 生成器能輔助智能體學習此點。

隨著訓練輪數增加，交易次數從第一輪減少 100,000 筆左右，到最後只減少約莫 10,000 筆，顯示可獲利的軌跡可能會收斂到一定程度，智能體學習到的交易筆數也會逐步減少到一定次數。在本實驗中，每輪獲利的變動幅度較大，推測是因為給予的擾動較於劇烈。由於初始軌跡的獲利能力較差，我們期望透過大幅度的擾動，使模型探索可能的獲利策略，故設置較大的擾動，導致各軌跡的變化性較大，來增加新資料的多樣性。同時也可能因為擾動導致模型成交的價格不同，或是成交的方向不同，導致累積過多存貨，進而因為不留倉的設置導致虧損。

至此，我們採用不同生成器產生多種交易軌跡，使智能體學習好的交易軌跡，藉此提升智能體的獲利能力。我們獲得下列結論：ϵ-greedy 生成器的效果比 noise 生成器好，原因可能在探索程度的強弱之別；ϵ-greedy 生成器可以有效提升基礎造市策略，在回測框架下，也有智能體可以達到正獲利，相較於實驗一的模型，得到了巨大的提升。

伍、 結論與未來展望

本研究以多智能體生成對抗式演算法為出發點，嘗試利用多智能體模仿不同造市商的造市策略，並在模仿的基礎下建立新的造市策略。我們主要有以下貢獻：

- 我們先從擴充造市單貼標開始，並依此基礎上進行了實驗一的模仿學習，嘗試擬合不同的造市策略。我們提升了三個模型對於其指定造市商的表現，相較於原始資料分布，提升了辨別市場上造市商的能力

- 我們在實驗二中，在模仿學習的基礎上加入了生成式學習的手法，以資料增強的方式提升原先造市策略的獲利能力。我們藉由生成器產生軌跡，使智能體學習好的軌跡，相較於實驗一的智能體，我們能有效提升智能體的獲利能力

由於造市資料的缺乏，我們沒有辦法驗證貼標邏輯的正確性，這使得基於貼標的後續研究皆無法確保正確的區別造市商。而對於模擬交易環境的擬真性，也是考量獲利能力的一大因素，惟本研究以簡易的模擬環境進行，並不保證實盤策略的可行程度，且造市環境及因素複雜，也無法真正模擬專業造市商的交易決策。

另外，還可嘗試更多不同的生成器參數、增加輪數或是每輪的生成軌跡數量，使產生的軌跡更多樣化。透過本次研究，我們嘗試應用強化學習及深度學習的技術於高頻的造市策略，除了前述所提及的問題之外，也期望本研究能提供一般投資者對於造市策略的新觀點，不再仰賴極為複雜的定價策略，而是直接向市場學習。

陸、 參考資料

[1] K. Arulkumaran, M. P. Deisenroth, M. Brundage and A. A. Bharath, "Deep Reinforcement Learning: A Brief Survey," in *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 26-38, Nov. 2017, doi: 10.1109/MSP.2017.2743240.

[2] Mnih, Volodymyr, et al. "Playing atari with deep reinforcement learning." *arXiv preprint arXiv:1312.5602* (2013).

[3] Silver, David, et al. "A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play." *Science* 362.6419 (2018): 1140-1144.

[4] Matarić, Maja J. "Reinforcement learning in the multi-robot domain." *Robot colonies* (1997): 73-83.

[5] Pricope, Tidor-Vlad. "Deep reinforcement learning in quantitative algorithmic trading: A review." *arXiv preprint arXiv:2106.00123* (2021).

[6] Fischer, Thomas G. *Reinforcement learning in financial markets-a survey*. No. 12/2018. FAU Discussion Papers in Economics, 2018.

[7] Meng, Terry Lingze, and Matloob Khushi. "Reinforcement learning in financial markets." *Data* 4.3 (2019): 110.

[8] Sato, Yoshiharu. "Model-free reinforcement learning for financial portfolios: a brief survey." *arXiv preprint arXiv:1904.04973* (2019).

[9] Hambly, Ben, Renyuan Xu, and Huining Yang. "Recent advances in reinforcement learning in finance." *arXiv preprint arXiv:2112.04553* (2021).

[10] Torabi, Faraz, Garrett Warnell, and Peter Stone. "Recent advances in imitation learning from observation." *arXiv preprint arXiv:1905.13566* (2019).

[11] Zhifei, Shao, and Er Meng Joo. "A survey of inverse reinforcement learning techniques." *International Journal of Intelligent Computing and Cybernetics* 5.3 (2012): 293-311.

[12] Arora, Saurabh, and Prashant Doshi. "A survey of inverse reinforcement learning: Challenges, methods and progress." *Artificial Intelligence* 297 (2021): 103500.

[13] Ho, Jonathan, and Stefano Ermon. "Generative adversarial imitation learning." *Advances in neural information processing systems* 29 (2016).

[14] Song, Jiaming, et al. "Multi-agent generative adversarial imitation learning." *Advances in neural information processing systems* 31 (2018).

[15] Spooner, Thomas, et al. "Market Making via Reinforcement Learning." *arXiv*

preprint arXiv:1804.04216 (2018).

[16] Gašperov, Bruno, et al. “Reinforcement learning approaches to optimal market making.” *Mathematics* 9.21 (2021): 2689.

[17] Ganesh, Sumitra, et al. “Reinforcement learning for market making in a multi-agent dealer market.” *arXiv preprint arXiv:1911.05892* (2019).

[18] Lim, Ye-Sheen, and Denise Gorse. "Reinforcement learning for high-frequency market making." *ESANN 2018-Proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Esann,* 2018.

[19] LSpooner, Thomas, et al. "Market Making via Reinforcement Learning." *arXiv preprint arXiv:1804.04216* (2018)

[20] Patel, Yagna. "Optimizing market making using multi-agent reinforcement learning." *arXiv preprint arXiv:1812.10252*

[21] Osa, Takayuki, et al. “An algorithmic perspective on imitation learning.” *Foundations and Trends® in Robotics* 7.1-2 (2018): 1-179.