Developing A Multi-Agent and Self-Adaptive Framework with Deep Reinforcement Learning for Dynamic Portfolio Risk Management

Zhenglong Li
The University of Hong Kong
Hong Kong, China
lzlong@hku.hk

Vincent Tam
The University of Hong Kong
Hong Kong, China
vtam@eee.hku.hk

Kwan L. Yeung
The University of Hong Kong
Hong Kong, China
kyeung@eee.hku.hk

ABSTRACT

Deep or reinforcement learning (RL) approaches have been adapted as reactive agents to quickly learn and respond with new investment strategies for portfolio management under the highly turbulent financial market environments in recent years. In many cases, due to the very complex correlations among various financial sectors, and the fluctuating trends in different financial markets, a deep or reinforcement learning based agent can be biased in maximising the total returns of the newly formulated investment portfolio while neglecting its potential risks under the turmoil of various market conditions in the global or regional sectors. Accordingly, a multi-agent and self-adaptive framework namely the MASA is proposed in which a sophisticated multi-agent reinforcement learning (RL) approach is adopted through two cooperating and reactive agents to carefully and dynamically balance the trade-off between the overall portfolio returns and their potential risks. Besides, a very flexible and proactive agent as the market observer is integrated into the MASA framework to provide some additional information on the estimated market trends as valuable feedbacks for multi-agent RL approach to quickly adapt to the ever-changing market conditions. The obtained empirical results clearly reveal the potential strengths of our proposed MASA framework based on the multi-agent RL approach against many well-known RL-based approaches on the challenging data sets of the CSI 300, Dow Jones Industrial Average and S&P 500 indexes over the past 10 years. More importantly, our proposed MASA framework shed lights on many possible directions for future investigation.

KEYWORDS

Deep Reinforcement Learning; Multi-Agent; Risk Management; Self-Adaptive; Portfolio Optimisation

ACM Reference Format:
Zhenglong Li, Vincent Tam, and Kwan L. Yeung. 2024. Developing A Multi-Agent and Self-Adaptive Framework with Deep Reinforcement Learning for Dynamic Portfolio Risk Management. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 9 pages.

This work is licensed under a Creative Commons Attribution International 4.0 License.

Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand. © 2024 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

1 INTRODUCTION

Computational Finance (CF) [12, 28, 37] is a very active research area involving the studies of computational approaches to tackle many different challenging and practical problems in Finance. Conventionally, algorithmic methods had been employed to simulate various investment strategies and their plausible results in the financial markets. Recently, many researchers have tried to explore the potential uses of machine learning approaches [31] including the support vector machines [35], deep learning (DL) or reinforcement (RL) learning approaches [17, 39, 42, 43] in a diversity of real-world applications [9, 25, 27, 46] in CF. Among these applications, DL or RL approaches such as the Twin Delayed DDPG (TD3) algorithm [10] for dynamic environments with continuous action spaces have been adapted as reactive agents [6] to quickly learn and respond with new investment strategies for portfolio management under the highly turbulent financial market environments in recent years. In many cases, due to the very complex correlations among various financial sectors, and the fluctuating trends in different financial markets, a deep or reinforcement learning based agent can be mainly focused on maximising the total returns of the newly formulated investment portfolio while ignoring the potential risks of the new investment portfolio under the turmoil of various market conditions such as the unpredictable and sudden changes of the market trends frequently occurring in the global or regional sectors of financial markets, especially due to the COVID-19 pandemic, natural disasters brought by extreme weathers, and local conflicts across different regions, etc.

To overcome the above pitfall, a **m**ulti-**a**gent and **s**elf-**a**daptive framework namely the *MASA* is proposed in this work in which two cooperating and reactive agents are utilised to implement a radically new multi-agent RL scheme so as to carefully and dynamically balance the trade-off between the overall returns of the newly revised portfolio and their potential risks especially when the concerned financial markets are highly turbulent. The first cooperating agent is based on the TD3 algorithm targeted to optimise the overall returns of the current investment portfolio while the second intelligent agent is based on a complete constraint solver, or possibly any efficient local optimisers such as the evolutionary algorithms [36] or particle swarm optimisation (PSO) methods [19], trying to adjust the current investment portfolio in order to minimise its potential risks after considering the estimated market trend as provided by another adaptive agent as the market observer in the proposed MASA framework. Clearly, the multi-agent RL scheme of the proposed MASA framework may help to produce more balanced investment portfolios in terms of both portfolio returns and potential

risks with the clear division of works between the two cooperating agents to continuously learn and adapt from the underlying financial market environment. It is worth noting that multi-agent RL-based frameworks have been actually considered in some previous research studies. For instance, a TD3-based multi-agent deep reinforcement learning (DRL) approach [45] was investigated in a previous work to improve the function approximation error and complex mission adaptability through applications to the mixed cooperation-competition environment in a general perspective. Yet instead of relying on the complex and dual-centered Q-network to reduce the bias of function estimation as in the previous work, our proposal has uniquely focused on using the TD3-based agent to firstly optimise on the overall returns of the newly revised portfolio with some possibly under-estimated bias/error in its potential risks to be quickly rectified by the second solver-based agent using a loosely-coupled and pipelining computational model to tackle this specific and challenging problem of dynamic portfolio risk management in the real-world applications of CF. It should be noted that by adopting the loosely-coupled and pipelining computational model, the proposed MASA framework will become more resilient and reliable since the overall framework will continue to work even when any particular agent fails. Moreover, to make the proposed MASA framework more adaptive to the extremely volatile environments of financial markets, the market observer as a very flexible and proactive agent to continuously provide the estimated market trends as valuable feedbacks for the other two cooperating agents to quickly adapt to the ever-changing market conditions. Undoubtedly, this simply highlights another key difference of our proposal on the multi-agent RL scheme when compared to those multi-agent RL-based frameworks examined in the previous studies. Furthermore, when the market observer agent is implemented as a deep neural network such as the multi-layer perceptron (MLP) [40] model, the resulting MASA framework can be extended as a DRL approach for dynamic portfolio management in CF.

To demonstrate the effectiveness of our proposal, a prototype of the proposed MASA framework is implemented in Python and tested on a GPU server installed with the Nvidia RTX 3090 GPU card. The attained empirical results demonstrate the potential strengths of our proposed MASA framework based on the multi-agent RL approach against many well-known RL-based approaches on the challenging data set of the CSI 300, Dow Jones Industrial Average (DJIA) and S&P 500 indexes over the past 10 years. More importantly, our proposed MASA framework shed lights on many possible directions including the exploration of utilising different meta-heuristic based optimisers such as the PSO for the solver-based agent, various machine learning approaches for the market observer agent, or the potential applications of the proposed MASA framework for various resource allocation, planning or disaster recovery problems in which the risk management is very critical for our future investigation.

2 THE PRELIMINARIES

2.1 Reinforcement Learning

As one of the active research areas in machine learning [11], RL [20] is mainly focused on how intelligent agents make rational decisions on actions based on specific observations in a possibly unexplored environment in order to maximise the cumulative rewards of the performed actions with respect to the underlying environment. The key focus of RL approaches is to strive for a *balance* between the *exploration* of the unknown environment and the *exploitation* of the current knowledge gained through the iterative learning process. The underlying environment is usually stated in the form of a partially observable Markov decision process (POMDP) [32] in which dynamic programming techniques [29] can be employed to solve the involved POMDP. Yet the RL approaches are targeted to handle large POMDPs where exact methods like the dynamic programming techniques may fail since the RL approaches do not need to assume any prior knowledge of the involved POMDP to represent the underlying environment. Clearly, the RL approaches are very suitable to explore the uncharted and also unpredictable environments of various financial markets when solving a diversity of real-world problems in CF.

In recent years, the RL approaches have attained remarkable successes for portfolio optimisation in which RL-based investment strategies have demonstrated adaptive and fast learning abilities to adjust the portfolios for maximising the overall returns after a targeted trading period. Among the numerous RL approaches, a successful example is the TD3 algorithm [10] as a model-free, online, off-policy reinforcement learning method. Generally speaking, a TD3 agent is an actor-critic reinforcement learning agent that is aimed to look for an optimal policy to maximise the expected cumulative long-term reward. For portfolio optimisation, the expected cumulative long-term reward of the TD3 agent can be straightforwardly formulated as the expected overall returns of the concerned portfolio after a specific trading period. Yet with the highly volatile financial market conditions, it can be difficult for most RL approaches to strive for a good balance between the intrinsically conflicting objectives of maximising returns and also minimising the risks of portfolios over a specific trading period. In most cases with turbulent market conditions, increasing the portfolio returns will likely increase the potential risks that may possibly lead to great and sudden losses of the investment portfolio in an extremely short period of time due to some unexpected crises.

2.2 Multi-Agent Systems

Multi-agent systems (MAS) [18] is a core and very active research area of artificial intelligence [47] in which many different perspectives and methodologies including the neural networks [13] or evolutionary algorithms [8] have been adopted and contributed to the latest development of MAS. In many real-world applications such as various challenging problems in CF, multiple intelligent agents may try to optimise their own returns and/or other objective(s), that may unavoidably collide with the interests of other investors with the same objective(s).

Besides, there are other research studies [3, 4] describing how MAS may facilitate the simulations in research studies of CF and Computational Economics. To more precisely model from the perspective of MAS, each agent in the multi-agent market simulation environment may find it difficult to learn a static investment strategy due to the fluctuating market dynamics. Thus, the involved agents may need to deploy intelligent algorithms capable of learning

to compete well with adaptive mechanisms in the adversarial market environments. In addition, studying intelligent trading through such simulations from a multi-agent perspective can lead to many exciting research directions with possible findings of relevance to policy makers and investors. An example is the market simulator (MAXE) [3] to examine different types of agent behaviour, market rules and anomalies on market dynamics through the simulation of large-scale MAS.

It is worth noting that there are some previous research studies investigating the potential uses of RL-based algorithms in MAS for many applications. For instance, a TD3-based multi-agent DRL approach [45] was examined in a previous work to improve the function approximation error and complex mission adaptability through applications to the mixed cooperation-competition environment from a general perspective. Essentially, the TD3-based DRL approach makes use of the complex and dual-centered Q-network to reduce the bias of function estimation. On the other hand, our proposal of the MASA framework has focused on using the TD3-based agent to firstly optimise on the overall returns of the newly revised portfolio while leaving the potential risks as the possibly under-estimated error to be effectively handled by the second solver-based agent with a loosely-coupled and pipelining mechanism for dynamic portfolio risk management in CF. Through adopting the loosely-coupled and pipelining computational model, the proposed MASA framework will become a dependable MAS with high availability and reliability since the resulting MAS will continue to work even when any specific agent fails.

2.3 Portfolio Optimisation in Computational Finance

Portfolio optimisation is a very challenging multi-objective optimisation problem in CF where the uncharted and highly volatile financial market environments can be difficult for many intelligent algorithms or well-known mathematical programming [34] approaches to tackle. Conventionally, many investors and researchers utilised specific financial indicators such as the moving averages [33] or the relative strength index [1], together with the heuristic or machine learning approaches including the follow-the-winner, follow-the-loser, pattern-matching or meta-learning algorithms [22] to try to capture the momentum of price changes. Recently, there have been many interesting research studies trying to apply DL or RL techniques to explore the turbulent and uncharted financial market environments. For instance, [25, 44] consider the news data as an additional information for portfolio management while [17, 42] utilise specific modules as intelligent agents to carefully deal with the assets information and then capture the correlations among the involved assets.

To facilitate our subsequence discussion, some essential concepts including the portfolio value, both the short-term and long-term risks of a portfolio, etc. related to portfolio management in CF are given as below.

Definition 2.1. (Portfolio Value) The total value of a portfolio at time t is

$$C_t = \sum_{i=1}^{N} a_{t,i} \times p^c_{t,i}, \tag{1}$$

where N is the number of assets in a portfolio, $a_{t,i}$ is the weight of ith asset, and $p^c_{t,i}$ is the close price of ith asset at time t.

Accordingly, each investment portfolio is constrained as below.

$$\forall a_{t,i} \in \mathbf{A}_t : \quad a_{t,i} \geq 0, \sum_{i=1}^{N} a_{t,i} = 1, \tag{2}$$

where $\mathbf{A}_t \in \mathbf{A}$ is the weight vector $\mathbf{A}$ at time t. Clearly, the summation of all the allocation weights $a_{t,i}$ for the total N assets of a complete portfolio should be 1.

Based on the well-known Markowitz model [29], both the short-term and long-term risks of a portfolio over a specific trading period can be defined in terms of the corresponding covariance-weighted risk and the volatility of strategies as follows.

Definition 2.2. (Short-term Portfolio Risk) The short-term portfolio risk $\sigma_{p,t}$ at time t is defined as below.

$$\begin{aligned} \sigma_{p,t} &= \sigma_\beta + \sigma_{\alpha,t} \\ \sigma_{\alpha,t} &= \sqrt{\mathbf{A}_t^T \Sigma_k \mathbf{A}_t} = \|\Sigma_k \mathbf{A}_t\|_2, \end{aligned} \tag{3}$$

where $\sigma_{\alpha,t}$ is the trading strategy risk, σ_β is the market risk and $\mathbf{A}_t \in \mathcal{R}^{N\times 1}$ is the matrix of weights. The covariance matrix $\Sigma_k \in \mathcal{R}^{N\times N}$ between any two assets can be calculated by the rate of daily returns of assets in the past k days.

Definition 2.3. (Long-term Portfolio Risk) The long-term portfolio risk Vol_p is defined as the strategy volatility that is the sampled variance of the daily return rates $r_{p,t}$ of a trading strategy over the whole trading period. $\bar{r}_{p,t}$ is the average daily return rate.

$$Vol_p = \sqrt{\frac{252}{T-1} \sum_{t=1}^{T} \left(r_{p,t} - \bar{r}_{p,t}\right)^2}. \tag{4}$$

Besides, the following gives a formal definition of the Sharpe ratio as one of the most widely adopted performance measures on the risk-adjusted relative returns of a portfolio.

Definition 2.4. (Sharpe Ratio) The Sharpe Ratio (SR) is a performance indicator for evaluating a portfolio in terms of the total annualized returns R_p, risk-free rate r_f and annualized long-term portfolio risk Vol_p.

$$\text{SR} = \frac{R_p - r_f}{Vol_p}. \tag{5}$$

More importantly, it should be noted that the portfolio optimisation problem in CF is used in this work to demonstrate the feasibility of our proposed MASA framework for risk management under the highly volatile and unknown environments. In the future investigation, it would be interesting to explore how the multi-agent RL-based approach of the proposed MASA framework can be adapted to various planning or resource allocation problems under certain hostile and unknown environments such as those for disaster recovery or emergency management.

3 THE PROPOSED MULTI-AGENT AND SELF-ADAPTIVE FRAMEWORK

To overcome the pitfall of the RL-based approaches to bias on optimising the investment returns, a **m**ulti-**a**gent and **s**elf-**a**daptive

framework namely the *MASA* is proposed in this work in which two cooperating and reactive agents, namely the RL-based and solver-based agents, are utilised to implement a radically new multi-agent RL scheme in order to dynamically balance the trade-off between the overall returns of the newly revised portfolio and potential risks especially when the financial markets are highly turbulent.

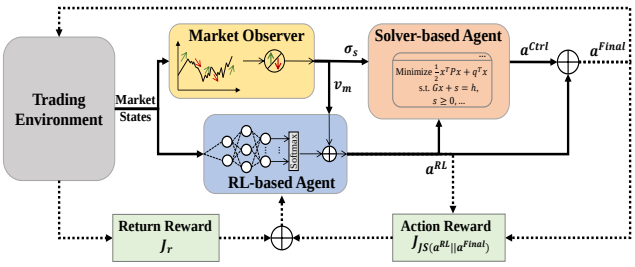

Figure 1: The System Architecture of the Proposed MASA Framework

Figure 1 reviews the overall system architecture of the proposed MASA framework in which the RL-based agent is based on the TD3 algorithm to optimise the overall returns of the current investment portfolio while the solver-based agent is based on a complete constraint solver, or possibly any efficient local optimisers such as the evolutionary algorithms [36] or PSO methods [19], that works to further adapt the investment portfolio returned by the RL-based agent so as to minimise its potential risks after considering the estimated market trend as provided by the market observer of the proposed MASA framework. In essence, through the clear division of works between both RL-based and solver-based agents to continuously learn and adapt from the underlying financial market environment with the support by market observer agent, the multi-agent RL scheme of the proposed MASA framework may help to attain more balanced investment portfolios in terms of both portfolio returns and potential risks when compared to those portfolios returned by the RL-based approaches. It is worth noting that the proposed MASA framework adopts a loosely-coupled and pipelining computational model among the three cooperating and intelligent agents, thus making the overall multi-agent RL-based approach more resilient and reliable since the overall framework will continue to work in the worst case of any individual agent being failed.

In addition, to make the proposed MASA framework more adaptive to the extremely volatile environments of financial markets, the market observer agent will continuously provide the estimated market trends as valuable feedbacks for both RL-based and solver-based agents to quickly adapt to the ever-changing market conditions. Furthermore, when the market observer agent is implemented as a deep neural network such as the MLP [40] model, the resulting MASA framework can be extended as a multi-agent DRL approach for dynamic portfolio management in CF. The empirical evaluation results of the market observer agent implemented as an algorithmic approach [38], the MLP and another deep learning models are carefully analysed in Section 4.

The pseudo-code of the training procedure of the proposed MASA framework is shown in Algorithm 1 to illustrate how the

Algorithm 1 The Training Procedure of the MASA Framework

Input: T as the total number of trading days, $MaxEpisode$ as the maximum number of episodes, the settings of RL-based agent, and the selected market observer agent.
Output: The revised RL policy π^* and (possibly) updated market observer agent.
Initialise the RL policy π_0, the market observer agent and memory tuple $\hat{D}$ and $\hat{M}$.
for $k = 1$ to $MaxEpisode$ **do**
  Reset the trading environment and set the initial action a_0^{Final} and a_0^{RL}.
  for $t = 1$ to T **do**
    Observe the current market state o_t
    Calculate the reward r_{t-1} by a_{t-1}^{Final}
    Store tuple $(o_{t-1}, a_{t-1}^{Final}, a_{t-1}^{RL}, o_t, r_{t-1})$ in $\hat{D}$
    Store tuple $(o_{t-1}, o_t, \sigma_{s,t-1}, v_{m,t-1})$ in $\hat{M}$
    Invoke the market observer agent to compute the suggested risk boundary $\sigma_{s,t}$ and market vector $v_{m,t}$ as the additional feedback for updating both the RL-based and solver-based agents
    Invoke the RL-based agent π_t to generate the current action a_t^{RL} as portfolio weights
    Invoke the solver-based agent to generate the adjusted action a_t^{Ctrl}
    Adjust the current portfolio by $a_t^{Final} = a_t^{RL} + a_t^{Ctrl}$
    Execute the portfolio order with a_t^{Final}
    if the RL policy update condition is triggered **then**
      Update the RL policy π by learning the historical trading data from $\hat{D}$
    end if
    if the predefined update condition of the market observer agent is triggered **then**
      Update the market observer agent by learning the historical profile $\hat{M}$
    end if
  end for
end for
return the best RL policy π^* and the possibly updated market observer agent

3 cooperating agents are working with each other to adaptively achieve the conflicting objectives of optimising returns and minimising risks in response to the possibly highly turbulent financial market conditions. Firstly, before the iterative training process is started, all the relevant information including the RL policy, the market state information stored in the market observer agent, etc. are initialised. During the training process, the current market state information o_t such as the most recent downward or upward trend of the underlying financial market over the past few trading days will be collected as the basic information for the subsequent computation of the market observer agent. Besides, the reward of the previously executed action a_{t-1}^{Final} will be computed as the feedback for the RL-based algorithm to revise its RL policy. The market observer agent will then be invoked to compute the suggested risk boundary $\sigma_{s,t}$ and market vector $v_{m,t}$ as some additional feedback

for updating both the RL-based and solver-based agents on the latest market conditions. As aforementioned, to maintain the flexibility and self-adaptivity of the proposed MASA framework, there can be various approaches including the algorithmic approach such as the directional changes [2, 38], deep neural networks such as the MLP or other DL approaches [15, 41] that can be considered in more detail in Section 4. More importantly, it should be noted that both the RL-based and solver-based agents are already secured with the current market information as the most valuable feedback obtained from the existing trading environment as shown in Figure 1. The provision of the suggested market condition information by the market observer agent are used solely as an additional information to quickly adapt and enhance the performance of both the RL-based and solver-based agents especially when the latest market conditions are highly volatile. In the worst cases when the suggested market condition produced by the market observer agent can be incorrect as 'noises' to mislead the search of both RL-based and solver-based agents for possibly biased actions in specific trading days, the *self-adaptive nature* of the reward mechanism of the RL-based agent to adapt from the underlying trading environment in the subsequent iterations of training, and also the *auto-corrective learning capability* of the intelligent market observer algorithm will help to ensure that such misleading noises can be effectively and quickly fixed over a longer period of trading to gain more valuable domain knowledge and insights about the underlying market conditions through updating the learning history profile $\hat{M}$ of the market observer agent. Interestingly, as observed from the empirical evaluation results obtained in Section 4, there can be fairly impressive enhancements in the ultimate performance of both RL-based and solver-based agents even when some relatively simple algorithmic approach based on the directional changes is used to implement the market observer agent in the proposed MASA framework on the challenging data sets of the CSI 300, DJIA and S&P 500 indexes over the past 10 years. Clearly, for a deeper understanding of the ultimate impacts of the suggested information by the market observer agent to the other two intelligent agents, the proposed MASA framework should be applied to more challenging data sets in CF or other application domains for more in-depth and thorough analyses in the future research studies. After the market observer agent is invoked, the RL-based agent will be triggered to generate the current action a_t^{RL} as portfolio weights that can be further revised by the subsequent solver-based agent after considering its own risk management strategy and the suggested market condition provided by the market observer agent. All in all, through adopting this loosely-coupled and pipelining computational model, the resulting MASA framework will continue to work as a dependable MAS even when any individual agent fails.

Figure 2 demonstrates the strengths of the reward-based guiding mechanism adopted by the proposed MASA framework to gradually enhance the various policies constructed by the single-agent RL-based approach, the proposed MASA framework without the reward-based guiding mechanism, and the proposed MASA framework utilising the reward-based guiding mechanism. The single-agent RL-based approach can update the policy π_{t+1} into the relatively more optimal set (i.e., the blue shaded area) by maximising the total returns of portfolios yet it may possibly neglect the

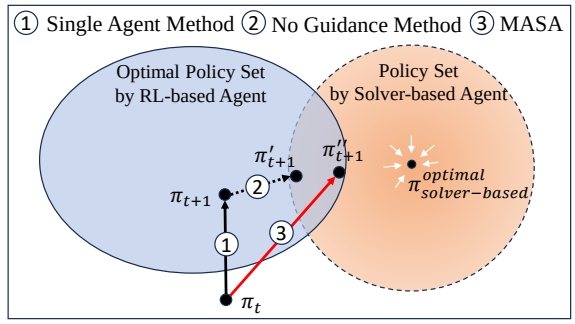

Figure 2: An Illustration of the Guiding Mechanism of the MASA Framework to Gradually Enhance the Constructed Policies of the RL-Based Agent

potential risks. On the other hand, the red shaded area of Figure 2 represents the policy set as recommended by the solver-based agent to minimise the potential risks for portfolio management. When working independently, each of the agents cannot combine the best advantages to achieve a more optimal portfolio for both objectives on the overall returns and potential risks. Besides, as shown in Figure 2, when all the 3 proposed agents working together without any intelligent guiding mechanism such as the reward-based guidance, the resulting framework can be easily stuck in specific local minima. Accordingly, through the reward-based guiding mechanism as adopted by the MASA framework to carefully respond to the ever-changing environment, both the RL-based and solver-based agents can iteratively enhance the current investment portfolio with respect to both objectives of the overall returns and potential risks after considering the valuable feedback from the third market observer agent. At the same time, the reward-based guiding mechanism of the MASA framework utilises an entropy-based divergence measure such as the Jensen–Shannon divergence (JSD) [26, 30] for promoting the diversity of the generated action sets as an intelligent and self-adaptive strategy to cater for the highly volatile environments of various financial markets.

The augmented reward function for the RL-based agent is depicted as follows.

$$J(\theta) = \lambda_1 J_r(\theta) + \lambda_2 J_{JS}(\theta), \tag{6}$$

where λ_1 and λ_2 are the learning rates of the return reward $J_r(\theta)$ and the action reward $J_{JS}(\theta)$. To maximise the overall returns of the current investment portfolio, the $J_r(\theta)$ can be computed as the sum of the logarithm of returns as stated in Equation (7).

$$\begin{aligned} J_r(\theta) &= \frac{1}{T} \log C_0 \prod_{t=1}^{T} r_t \\ &= \frac{1}{T} \left(\log C_0 + \sum_{t=1}^{T} \log r_t \right), \end{aligned} \tag{7}$$

where C_0 is the initial portfolio value, T is the number of trading days, and $r_t = \frac{C_t}{C_{t-1}}$ is the growth rate of portfolio at t.

The action-guided reward $J_{JS}(\theta)$ to promote the diversity of the action sets as generated by the proposed MASA framework is

defined as below.

$$J_{JS}(\theta) = -\frac{1}{T}\sum_{t=1}^{T} D_{JS}\left(\mathbf{a}_t^{RL} \,||\, \mathbf{a}_t^{Final}\right), \tag{8}$$

where $\mathbf{a}_t^{RL}$ is the action generated by the RL-based agent at t, $\mathbf{a}_t^{Final}$ is the adjusted action after considering the actions as recommended by both the RL-based and solver-based agents, and D_{JS} is the JSD to measure the similarity between $\mathbf{a}_t^{RL}$ and $\mathbf{a}_t^{Final}$ as two probability distributions of the actions generated by the MASA framework.

4 AN EMPIRICAL EVALUATION

Datasets: To demonstrate the effectiveness of the proposed MASA framework in tackling the real-world portfolio risk management with conflicting objectives under the mostly uncharted and highly volatile financial market environments, a preliminary prototype of the proposed MASA framework is implemented in Python, and evaluated on a GPU server machine installed with the AMD Ryzen 9 3900X 12-Core processor running at 3.8 GHz and two Nvidia RTX 3090 GPU cards. Furthermore, the MASA framework is compared with other methods on three challenging yet representative data sets of CSI 300, Dow Jones Industrial Average (DJIA) and S&P 500 indexes from September 2013 to August 2023 in which the first five-year data is used to train the model, followed by the subsequent data set of two years to validate the trained model. Lastly, all the validated models of various approaches are evaluated on the data set of the latest three years. The top 10 stocks of each index are selected to construct the investment portfolio in terms of the company capital. In addition, all the involved data sets contain both upward and downward trends of stock prices, and also various patterns of fluctuation for different market conditions so as to avoid any possible bias toward a specific approach under the evaluation.

Comparative Methods: Ten representative methods based on algorithmic or RL approaches are carefully selected to compare against the MASA framework. The Constant Rebalanced Portfolio (CRP) [5] is the vanilla strategy of equal weighting. The Exponential Gradient (EG) method [14] is based on the follow-the-winner approach while the Online Moving Average Reversion (OLMAR) [21], Passive Aggressive Mean Reversion (PAMR) [24], and Robust Median Reversion (RMR) [16] approaches follow the loser assets during trading. The Correlation-driven Nonparametric Learning Strategy (CORN) [23] is a heuristic strategy to match historical investment patterns. Moreover, the four latest RL-based portfolio optimisation approaches are considered. Ensemble of Identical Independent Evaluators (EIIE) [17] is based on a convolution-based neural network to extract the features of assets while Portfolio Policy Network (PPN) [46] consists of a recurrent-based and a convolution-based neural networks to capture the sequential information and correlations between assets. Besides, Relation-Aware Transformer (RAT) [42] is a transformer-based model to learn the patterns from price series. Lastly, the TD3 with a profit maximisation strategy (TD3-Profit) [10] as the classical RL approach is included for the comparison.

Besides, to evaluate the profitability and risk management of the concerned approaches, four commonly adopted performance metrics including the Annual Return (AR), Maximum Drawdown (MDD), Sharpe Ratio (SR), and short-term portfolio risk (Risk) are considered. Specifically, the SR is a comprehensive metric to indicate the balance between the portfolio returns and risks as attained by each approach. All the reported results are averaged over 10 runs.

Performance Analysis: Table 1 reviews the performance of various well-known RL-based approaches against that of the proposed MASA framework using different market observers, with the symbol ↑ to denote the preference of a larger value in the metrics of AR and SR while the symbol ↓ denoting the favour of a smaller value in MDD and Risk. From the results of the CSI 300 data set, the AR of the MASA frameworks is at least 1.5% larger than those of other methods while maintaining the portfolio risks at a relatively low level. In particular, the MASA framework integrated with an MLP-based market observer achieves the highest AR at 8.87% and the highest SR at 0.27, thus demonstrating the higher capability of all the proposed agents in the MASA framework to balance the trade-offs among different objectives.

For the attained results on the DJIA index, the MASA-DC approach utilising the directional changes (DC) method [38] as the market observer agent to estimate the market trends significantly outperforms other baseline models in all metrics. Specially, the MASA-DC approach attains the AR about 4% higher than that of the single-agent TD3-Profit approach while reducing the maximum possible losses by 3% when compared to the TD3-Profit in terms of MDD. For a clear presentation of the overall results, Figure 3 shows the changes of portfolio values of each approach under evaluation. In addition, Figure 4 shows an interesting example of upward trends in the DJIA market where the MASA-DC approach achieves competitive returns while maintaining the potential risks at a relatively lower level as compared to those of other approaches. On the contrary, when the financial market stays for long periods of downward trends, the portfolio values of those baseline models dramatically decreases as shown in Figure 3. Yet the MASA-DC approach can manage to effectively minimise the losses during such adverse market conditions when compared to the other approaches. Figure 4 reveals such a challenging example of downward trends in which the short-term risk of the involved portfolio can be managed well by the MASA-DC approach with less fluctuation even if the market index drops over 10%, thus confirming the effectiveness of the DC-based market observer agent to timely capture the environment changes as the valuable feedback for the solver-based agent to adjust the actions for balancing multiple objectives. Furthermore, a similar performance is attained by the MASA approach on the other two indexes for which the corresponding graphs of portfolio values and risk comparison can be found in the Appendix for a more detailed investigation.

Table 1 reviews the performance of various approaches on the S&P 500 data set in which the OLMAR obtains a relatively high AR due to its loser tracking strategy that may typically invest almost the whole capital in a single asset. Yet such strategy may not be able to balance the risk in a portfolio and possibly fail in financial markets of downward trends. Thus, the OLMAR gets a relatively higher MDD of 68.52% where it may suffer from huge potential risks. Similar to the results obtained on the CSI 300 and DJIA data sets, both the MASA-MLP and MASA-LSTM approaches obtain the best performance on balancing the returns and potential risks, achieving a SR of around 0.9 and a MDD of 26%.

Moreover, the Wilcoxon rank-sum test [7] is used to compare the statistical significance of the MASA framework against the

Table 1: The Performance of Various Well-Known RL-Based Approaches Against the Proposed MASA Framework on Different Challenging Data Sets of Financial Indexes

Market	CSI 300				DJIA				S&P 500			
Models	**AR**(%) ↑	**MDD**(%) ↓	**SR**↑	**Risk**↓	**AR**(%) ↑	**MDD**(%) ↓	**SR**↑	**Risk**↓	**AR**(%) ↑	**MDD**(%) ↓	**SR**↑	**Risk**↓
CRP	7.19	33.96	0.19	0.0131	11.44	19.66	0.58	0.0095	18.09	37.12	0.65	0.0143
EG	7.19	33.94	0.19	0.0131	11.39	19.66	0.57	0.0095	18.03	36.88	0.65	0.0142
OLMAR	-3.50	55.67	-0.17	0.0217	-13.89	59.89	-0.54	0.0179	**29.34**	68.52	0.55	0.0275
PAMR	-14.24	49.32	-0.47	0.0223	-37.72	81.72	-1.35	0.0174	3.63	59.08	0.05	0.0263
CORN	-2.06	59.28	-0.13	0.0219	1.62	41.76	0.00	0.0122	-6.90	62.97	-0.21	0.0200
RMR	5.77	41.96	0.07	0.0215	-12.98	61.23	-0.51	0.0180	-4.12	85.48	-0.09	0.0280
EIIE	6.84	31.77	0.18	0.0122	10.81	18.24	0.58	0.0089	16.50	35.80	0.63	0.0135
PPN	6.74	**31.19**	0.18	0.0119	10.50	17.95	0.57	0.0087	16.65	34.17	0.65	0.0130
RAT	6.78	31.48	0.18	0.0121	10.62	18.19	0.57	0.0088	17.03	34.92	0.65	0.0133
TD3-Profit	7.18	33.97	0.19	0.0128	11.45	19.65	0.58	0.0095	18.09	37.12	0.65	0.0143
MASA-MLP	**8.87**	31.78	**0.27**	**0.0119**	13.17	19.89	0.69	0.0088	22.49	26.50	**0.92**	0.0116
MASA-LSTM	8.72	31.83	0.26	0.0121	13.50	19.58	0.71	0.0087	22.12	26.61	0.90	0.0117
MASA-DC	8.70	31.77	0.25	0.0120	**15.52**	**16.21**	**0.80**	**0.0086**	14.88	**24.29**	0.60	**0.0112**

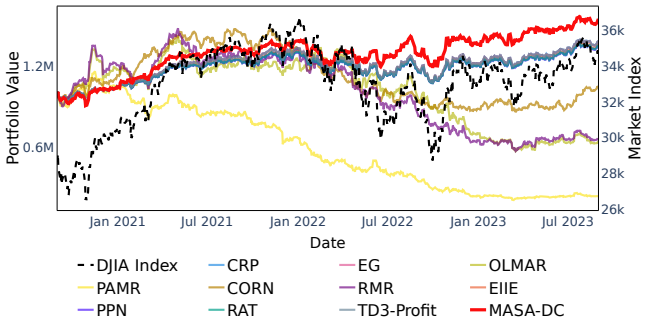

Figure 3: A Comparison of the Portfolio Values of Different Approaches on the DJIA Index

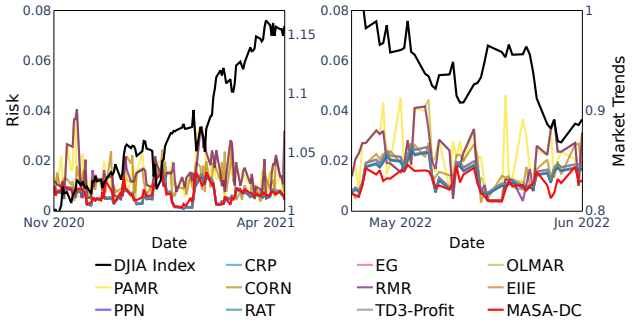

Figure 4: The Risk Comparison on the Uptrend and Downtrend Cases of the DJIA Index

other approaches with a significance level of 0.05. Clearly, the performance of the MASA is statistically significant against that of other compared approaches on all three data sets except for the specific result attained by OLMAR approach on the S&P 500 index. Generally speaking, the MASA approach beats all other approaches on the three challenging data sets through the 3 cooperating agents to carefully optimise the possibly conflicting objectives under the highly uncharted environments of various financial markets.

Table 2: The Ablation Study of the Proposed MASA Framework on the CSI 300 Index

Models		**AR**(%) ↑	**MDD**(%) ↓	**SR**↑	**Risk**↓
Single-Agent	TD3-Profit	7.18	33.97	0.19	0.0128
	TD3-PR	7.21	33.96	0.19	0.0128
	TD3-SR	7.18	33.98	0.19	0.0128
Dual-Agent	MASA-w/oMktObs	8.17	32.71	0.23	0.0121
Triple-Agent	MASA-MLP	8.46	32.26	0.25	**0.0119**
	MASA-LSTM	8.27	32.42	0.24	0.0121
(w/o Action Reward)	MASA-DC	7.93	32.34	0.21	**0.0119**
Triple-Agent	MASA-MLP	**8.87**	31.78	**0.27**	**0.0119**
	MASA-LSTM	8.72	31.83	0.26	0.0121
(with Action Reward)	MASA-DC	8.70	**31.77**	0.25	0.0120

Ablation Study: Table 2 shows the results of the ablation study of the proposed MASA framework on the CSI 300 index in which three variants of the TD3-based models are used to compare with the MASA framework utilising the TD3 approach to implement the RL-based agent. Specifically, the TD3-Profit model is targeted to maximise total profits while the TD3-PR model combines both profit maximisation and short-term risk minimisation. Besides, the TD3-SR approach uses the SR as the reward function. For the dual-agent model, the MASA-w/oMktObs approach combines both the RL-based and solver-based agents to balance the trade-offs of the portfolio optimisation yet there is no market observer agent to provide any additional market information. For the proposed triple-agent MASA framework, the market observer agent is implemented by the MLP, LSTM and DC method respectively. The single-agent approach obtains around 7.20% of returns per year yet with potential losses of 33% during the trading period. Moreover, the MASA-w/oMktObs efficiently reduces the investment risks to avoid great losses even when no extra information about market changes is provided. Thus, the total AR of the dual-agent model is increased by 1% against that of single-agent model. Meanwhile, a relatively higher SR is achieved by the MASA-w/oMktObs due to a better trade-off between returns and risks. Furthermore, the MASA can better estimate the potential risks while pursuing higher returns

after considering more latest market information from the market observer agent. The resulting risks and MDD are dropped to 0.0119 and 31.7% respectively, with some further improvement on both the total returns and SR. To demonstrate the effectiveness of the solver-based agent on the risk management, Figure 5 shows the relationship between market state changes and the sum of weights adjusted by the solver-based agent. Clearly, the solver-based agent makes larger weighting adjustments to manage risks at a relatively low level after considering the market information provided by the market observer, especially when the potential risks are increased sharply over successive episodes. Obviously, the ablation studies confirm the contributions of both the solver-based and market observer agents in the proposed MASA framework that can effectively tackle the trade-offs between different objectives under the highly volatile environments of financial markets.

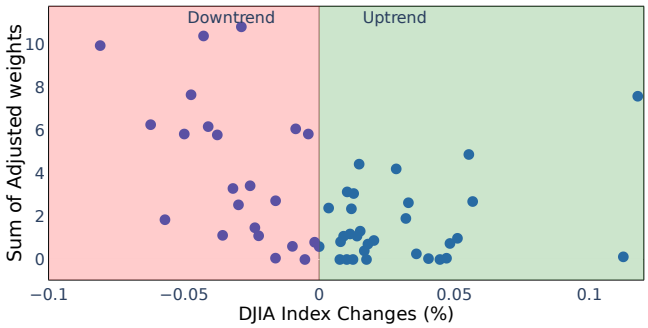

Figure 5: The Contribution of Solver-based Agents on Different Market States

Table 2 shows the effectiveness of the reward of the action generated by the RL-based agent in the proposed MASA framework. The MASA variant without considering the reward of the action by the RL-based agent still performs better than the single-agent or dual-agent framework in balancing the profits and risks especially when the MLP or LSTM model is used for the market observer agent. When considering the rewards of generated actions, the risk-aware information provided by the solver-based agent can guide the policy of the RL-based agent toward higher profits and less potential risks for which the MASA framework can enhance the AR by 0.5% and the MDD by 1%.

Furthermore, the top 20 and 30 stocks of each index are selected to study the scalability of the MASA framework on large-scale portfolios, except for the CSI 300 index due to the limited data sources. When constructing a portfolio of 20 assets in the DJIA market, the MASA-MLP achieves the highest SR of 0.80 and the highest AR of 14% while the best baseline approach obtains a SR of 0.61 and a AR of 11% only. After increasing the portfolio size to 30 assets, the MASA framework still has a significant improvement against those of other approaches on all metrics. Similar results are obtained by the MASA framework in the S&P 500 market, that can be found in the Appendix. Undoubtedly, all the obtained results validate that the MASA framework can achieve a better performance in balancing different goals when the problem size increases.

5 CONCLUDING REMARKS

In recent years, deep or reinforcement learning (RL) approaches have been adapted as intelligent and reactive agents to quickly learn and respond with newly revised investment strategies for portfolio management under the highly volatile financial market environments particularly under the threats of global or regional conflicts, pandemics, natural disasters, etc. Yet due to the very complex correlations among different financial sectors, and the fluctuating trends in various financial markets, a deep or reinforcement learning based agent can be biased in maximising the total returns of the newly formulated investment portfolio while neglecting its potential risks under the turmoil of various market conditions in the global or regional sectors. Accordingly, a multi-agent and self-adaptive framework namely the MASA is proposed in which a sophisticated multi-agent reinforcement learning approach is adopted through both of the RL-based and solver-based agents working to carefully and dynamically balance the trade-off between the overall portfolio returns and their potential risks. In addition, a very flexible and proactive agent as the market observer is integrated into the proposed MASA framework to provide the estimated market conditions and trends as additional information for multi-agent RL approach to carefully consider so as to quickly adapt to the ever-changing market conditions.

To demonstrated the potential advantages of our proposal, a prototype of the proposed MASA framework is evaluated against various well-known RL-based approaches on the challenging data sets of the CSI 300, Dow Jones Industrial Average and S&P 500 indexes over the past 10 years. The obtained empirical results clearly reveal the remarkable performance of our proposed MASA framework based on the multi-agent RL approach when compared against those of other well-known RL-based approaches on the 3 data sets of widely recognised financial indexes in China and the United States. More importantly, our proposed MASA framework shed lights on many possible directions for future investigation. First, the thorough investigation on using different meta-heuristic based optimisers such as the evolutionary algorithms or the PSO for the solver-based agent should be interesting. Besides, experimenting various intelligent approaches for the market observer agent is worth exploring. Last but not least, the potential applications of the proposed MASA model for various resource allocation, planning or disaster recovery in which the risk management is critical and timely should be very valuable for our future studies.

ACKNOWLEDGMENTS

The authors wish to express our deepest gratitude to Professor Edward Tsang for his fruitful discussion on this work, and also the anonymous reviewers for their valuable feedback.

REFERENCES

[1] Rodrigo Alfaro and Andres Sagner. 2010. Financial Forecast for the Relative Strength Index. (2010).
[2] Amer Bakhach, Edward PK Tsang, and Hamid Jalalian. 2016. Forecasting directional changes in the fx markets. In *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 1–8.
[3] Peter Belcak, Jan-Peter Calliess, and Stefan Zohren. 2022. Fast Agent-Based Simulation Framework with Applications to Reinforcement Learning and the Study of Trading Latency Effects. In *Multi-Agent-Based Simulation XXII*, Koen H. Van Dam and Nicolas Verstaevel (Eds.). Springer International Publishing, Cham, 42–56.

[4] Jan-Peter Calliess and Stefan Zohren. 2021. Agent-Based Models in Finance and Market Simulations. (2021). https://oxford-man.ox.ac.uk/projects/agent-based-models-in-finance-and-market-simulations/

[5] Thomas M Cover. 1991. Universal Portfolios. *Mathematical finance* (1991).

[6] Nigel Cuschieri, Vince Vella, and Josef Bajada. 2021. TD3-Based Ensemble Reinforcement Learning for Financial Portfolio Optimisation. *FinPlan 2021* (2021), 6.

[7] Joaquín Derrac, Salvador García, Daniel Molina, and Francisco Herrera. 2011. A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms. *Swarm and Evolutionary Computation* 1, 1 (2011), 3–18.

[8] Rafał Dreżewski and Krzysztof Doroz. 2017. An agent-based co-evolutionary multi-objective algorithm for portfolio optimization. *Symmetry* 9, 9 (2017), 168.

[9] Yitong Duan, Lei Wang, Qizhong Zhang, and Jian Li. 2022. FactorVAE: A Probabilistic Dynamic Factor Model Based on Variational Autoencoder for Predicting Cross-sectional Stock Returns. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[10] Scott Fujimoto, Herke Hoof, and David Meger. 2018. Addressing Function Approximation Error in Actor-critic Methods. In *Proceedings of the International Conference on Machine Learning*. PMLR.

[11] John W Goodell, Satish Kumar, Weng Marc Lim, and Debidutta Pattnaik. 2021. Artificial intelligence and machine learning in finance: Identifying foundations, themes, and research clusters from bibliometric analysis. *Journal of Behavioral and Experimental Finance* 32 (2021), 100577.

[12] Abhishek Gunjan and Siddhartha Bhattacharyya. 2023. A brief review of portfolio optimization techniques. *Artificial Intelligence Review* 56, 5 (2023), 3847–3886.

[13] Reza Hafezi, Jamal Shahrabi, and Esmaeil Hadavandi. 2015. A bat-neural network multi-agent system (BNNMAS) for stock price prediction: Case study of DAX stock price. *Applied Soft Computing* 29 (2015), 196–210.

[14] David P Helmbold, Robert E Schapire, Yoram Singer, and Manfred K Warmuth. 1998. On-line Portfolio Selection Using Multiplicative Updates. *Mathematical Finance* (1998).

[15] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[16] Dingjiang Huang, Junlong Zhou, Bin Li, Steven CH Hoi, and Shuigeng Zhou. 2016. Robust median reversion strategy for online portfolio selection. *IEEE Transactions on Knowledge and Data Engineering* 28, 9 (2016), 2480–2493.

[17] Zhengyao Jiang, Dixing Xu, and Jinjun Liang. 2017. A Deep Reinforcement Learning Framework for the Financial Portfolio Management Problem. *arXiv preprint arXiv:1706.10059* (2017).

[18] Michael Kampouridis, Panagiotis Kanellopoulos, Maria Kyropoulou, Themistoklis Melissourgos, and Alexandros A. Voudouris. 2022. Multi-agent systems for computational economics and finance. *AI Communications* 35, 4 (sep 2022), 369–380. https://doi.org/10.3233/aic-220117

[19] James Kennedy and Russell Eberhart. 1995. Particle swarm optimization. In *Proceedings of ICNN'95-international conference on neural networks*, Vol. 4. IEEE, 1942–1948.

[20] Petter N Kolm and Gordon Ritter. 2020. Modern perspectives on reinforcement learning in finance. *Modern Perspectives on Reinforcement Learning in Finance (September 6, 2019). The Journal of Machine Learning in Finance* 1, 1 (2020).

[21] Bin Li and Steven CH Hoi. 2012. On-Line Portfolio Selection with Moving Average Reversion. In *Proceedings of the International Conference on Machine Learning*. PMLR.

[22] Bin Li and Steven CH Hoi. 2014. Online Portfolio Selection: A Survey. *ACM Computing Surveys (CSUR)* (2014).

[23] Bin Li, Steven CH Hoi, and Vivekanand Gopalkrishnan. 2011. Corn: Correlation-driven Nonparametric Learning Approach for Portfolio Selection. *ACM Transactions on Intelligent Systems and Technology* (2011).

[24] Bin Li, Peilin Zhao, Steven CH Hoi, and Vivekanand Gopalkrishnan. 2012. PAMR: Passive Aggressive Mean Reversion Strategy for Portfolio Selection. *Machine learning* (2012).

[25] Qianqiao Liang, Mengying Zhu, Xiaolin Zheng, and Yan Wang. 2021. An Adaptive News-Driven Method for CVaR-sensitive Online Portfolio Selection in Non-Stationary Financial Markets.. In *Proceedings of the IJCAI Conference on Artificial Intelligence*.

[26] J. Lin. 1991. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory* 37, 1 (1991), 145–151. https://doi.org/10.1109/18.61115

[27] Yang Liu, Qi Liu, Hongke Zhao, Zhen Pan, and Chuanren Liu. 2020. Adaptive Quantitative Trading: An Imitative Deep Reinforcement Learning Approach. In *Proceedings of the AAAI conference on artificial intelligence*.

[28] Costis Maglaras, Ciamac C Moallemi, and Muye Wang. 2022. A deep learning approach to estimating fill probabilities in a limit order book. *Quantitative Finance* 22, 11 (2022), 1989–2003.

[29] Harry M Markowitz and G Peter Todd. 2000. *Mean-variance Analysis in Portfolio Choice and Capital Markets*. John Wiley & Sons.

[30] M.L. Menéndez, J.A. Pardo, L. Pardo, and M.C. Pardo. 1997. The Jensen-Shannon divergence. *Journal of the Franklin Institute* 334, 2 (1997), 307–318. https://doi.org/10.1016/S0016-0032(96)00063-4

[31] Mojtaba Nabipour, Pooyan Nayyeri, Hamed Jabani, S Shahab, and Amir Mosavi. 2020. Predicting stock market trends using machine learning and deep learning algorithms via continuous and binary data; a comparative analysis. *IEEE Access* 8 (2020), 150199–150212.

[32] Martin L Puterman. 1990. Markov decision processes. *Handbooks in operations research and management science* 2 (1990), 331–434.

[33] Aistis Raudys, Vaidotas Lenčiauskas, and Edmundas Malčius. 2013. Moving averages for financial data smoothing. In *Information and Software Technologies: 19th International Conference, ICIST 2013, Kaunas, Lithuania, October 2013. Proceedings 19*. Springer, 34–45.

[34] Bartosz Sawik. 2012. Bi-criteria portfolio optimization models with percentile and symmetric risk measures by mathematical programming. *Przeglad Elektrotechniczny* 88, 10B (2012), 176–180.

[35] M Sivaram, E Laxmi Lydia, Irina V Pustokhina, Denis Alexandrovich Pustokhin, Mohamed Elhoseny, Gyanendra Prasad Joshi, and K Shankar. 2020. An optimal least square support vector machine based earnings prediction of blockchain financial products. *IEEE Access* 8 (2020), 120321–120330.

[36] Rainer Storn and Kenneth Price. 1997. Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. *Journal of global optimization* 11, 4 (1997), 341–359.

[37] Edward PK Tsang. 2023. *AI for Finance*. CRC Press.

[38] Edward PK Tsang, Ran Tao, Antoaneta Serguieva, and Shuai Ma. 2017. Profiling high-frequency equity price movements in directional changes. *Quantitative finance* 17, 2 (2017), 217–225.

[39] Heyuan Wang, Shun Li, Tengjiao Wang, and Jiayi Zheng. 2021. Hierarchical Adaptive Temporal-Relational Modeling for Stock Trend Prediction.. In *Proceedings of the IJCAI Conference on Artificial Intelligence*.

[40] Ruoxi Wang, Rakesh Shivanna, Derek Cheng, Sagar Jain, Dong Lin, Lichan Hong, and Ed Chi. 2021. Dcn v2: Improved deep & cross network and practical lessons for web-scale learning to rank systems. In *Proceedings of the web conference 2021*. 1785–1797.

[41] Zhicheng Wang, Biwei Huang, Shikui Tu, Kun Zhang, and Lei Xu. 2021. DeepTrader: a deep reinforcement learning approach for risk-return balanced portfolio management with market conditions Embedding. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 643–650.

[42] Ke Xu, Yifan Zhang, Deheng Ye, Peilin Zhao, and Mingkui Tan. 2021. Relation-aware Transformer for Portfolio Policy Learning. In *Proceedings of the IJCAI Conference on Artificial Intelligence*.

[43] Mengyuan Yang, Xiaolin Zheng, Qianqiao Liang, Bing Han, and Mengying Zhu. 2022. A Smart Trader for Portfolio Management based on Normalizing Flows. In *Proceedings of the IJCAI Conference on Artificial Intelligence*.

[44] Yunan Ye, Hengzhi Pei, Boxin Wang, Pin-Yu Chen, Yada Zhu, Ju Xiao, and Bo Li. 2020. Reinforcement-learning Based Portfolio Management with Augmented Asset Movement Prediction States. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[45] Fengjiao Zhang, Jie Li, and Zhi Li. 2020. A TD3-based multi-agent deep reinforcement learning method in mixed cooperation-competition environment. *Neurocomputing* 411 (2020), 206–215. https://doi.org/10.1016/j.neucom.2020.05.097

[46] Yifan Zhang, Peilin Zhao, Qingyao Wu, Bin Li, Junzhou Huang, and Mingkui Tan. 2020. Cost-sensitive Portfolio Selection via Deep Reinforcement Learning. *IEEE Transactions on Knowledge and Data Engineering* (2020).

[47] Xiao-lin Zheng, Meng-ying Zhu, Qi-bing Li, Chao-chao Chen, and Yan-chao Tan. 2019. FinBrain: when finance meets AI 2.0. *Frontiers of Information Technology & Electronic Engineering* 20, 7 (2019), 914–924.