International Review of Economics and Finance 98 (2025) 103996

Contents lists available at ScienceDirect

International Review of Economics and Finance

journal homepage: www.elsevier.com/locate/iref

High-dimensional multi-period portfolio allocation using deep reinforcement learning

Yifu Jiang [a], Jose Olmo [a,b,*], Majed Atwi [a]

[a] Department of Economic Analysis, University of Zaragoza, Gran Via 2, 50005, Zaragoza, Spain
[b] Department of Economics, University of Southampton, Highfield Campus, SO17 1BJ, Southampton, United Kingdom

ARTICLE INFO

JEL classification:
C45
C61
G11
G17

Keywords:
Multi-period portfolio selection
High-dimensional portfolios
Risk aversion
Portfolio constraints
Deep reinforcement learning

ABSTRACT

This paper proposes a novel investment strategy based on deep reinforcement learning (DRL) for long-term portfolio allocation in the presence of transaction costs and risk aversion. We design an advanced portfolio policy framework to model the price dynamic patterns using convolutional neural networks (CNN), capture group-wise asset dependence using WaveNet, and solve the optimal asset allocation problem using DRL. These methods are embedded within a multi-period Bellman equation framework. An additional appealing feature of our investment strategy is its ability to optimize dynamically over a large set of potentially correlated risky assets. The performance of this portfolio is tested empirically over different holding periods, risk aversion levels, transaction cost rates, and financial indices. The results demonstrate the effectiveness and superiority of the proposed long-term portfolio allocation strategy compared to several competitors based on machine learning methods and traditional optimization techniques.

1. Introduction

Portfolio management supports investors in making decisions on how to allocate resources and funds across a set of assets and over time. Traditional portfolio selection methods typically consider single-period returns. Markowitz (1952) pioneered the mean-variance optimization model, which remains foundational in portfolio theory. This framework seeks to maximize the expected return for a given level of risk by considering the variance of asset returns. Despite its simplicity and wide application, Markowitz's paradigm has limitations, particularly in addressing long-term investment horizons and dynamic market conditions.

Long-term portfolio allocation focuses on how investors can optimally allocate investment assets over extended periods of time to maximize returns while managing risks (Escobar et al., 2016; Fan et al., 2024; Lucey & Muckley, 2011). One of the first contributions in this area was Merton (1969, 1971), who extended the portfolio selection framework to a continuous-time setting, incorporating intertemporal choice and dynamic strategies for long-term investors. His work introduced the concept of dynamic asset allocation, emphasizing the importance of adjusting portfolio weights over time in response to changes in market conditions and investor preferences. It is widely understood, at least since the work of this author, as seen by Samuelson (1969), that the solution to a multi-period portfolio choice problem can be very different from the solution to a static portfolio choice problem. Unfortunately, intertemporal asset allocation models are hard to solve in closed form unless strong assumptions on the investor's objective function or the statistical distribution of asset returns are imposed.

* Corresponding author. Department of Economic Analysis, University of Zaragoza, Gran Via 2, 50005, Zaragoza, Spain.
E-mail addresses: 849201@unizar.es (Y. Jiang), joseolmo@unizar.es (J. Olmo), matwi@unizar.es (M. Atwi).

https://doi.org/10.1016/j.iref.2025.103996
Received 2 October 2024; Received in revised form 25 February 2025; Accepted 26 February 2025
Available online 27 February 2025

Traditionally, the extension from single-period to multi-period portfolio optimization has been addressed using stochastic dynamic programming. Samuelson (1969) and Bellman (1957) developed methods for solving dynamic optimization problems, allowing the consideration of future states and decisions in portfolio management. Nevertheless, the lack of closed-form solutions for optimal portfolios in multi-period settings has limited the applicability of Merton's model and has not displaced Markowitz's paradigm. This situation began to change due to several developments in numerical methods and continuous time finance models. More specifically, some authors such as Barberis (2000) and Brennan et al. (1997, 1999), among a few others, provide discrete-state numerical algorithms to approximate the solution of the portfolio problem over infinite horizons. Other articles obtain closed-form solutions to the Merton model in a continuous time framework with a constant risk-free interest rate and a single risky asset if long-lived investors have power utility defined over terminal wealth (Kim & Omberg, 1996) or if investors have power utility defined over consumption (Watcher, 2002), or if the investor has Epstein and Zin (1989, 1991) utility with intertemporal elasticity of substitution equal to one (Campbell & Viceira, 1999; Schroder & Skiadas, 1999). Approximate analytical solutions to the Merton model have been developed by Campbell et al. (2003) and Campbell and Viceira (1999, 2001, 2002) for models exhibiting an intertemporal elasticity of substitution not too far from one.

An alternative to solve the investor's optimal portfolio choice problem is proposed by Ait-Sahalia and Brandt (1999, 2001), and Brandt and Clara (2006). These authors show how to select and combine variables to best predict the optimal portfolio weights, both in single-period and multi-period contexts. Moreover, Laborda and Olmo (2017) focus directly on the dependence of the portfolio weights on the predictor variables through a linear parametric portfolio policy rule. This characterization allows them to apply the generalized method of moments to sample analogues of the multi-period Euler equations that characterize the optimal portfolio choice.

Standard methods based on solving Bellman equations struggle with large datasets. Hambly et al. (2023) noted that portfolio optimization often involves high dimensionality, complex non-linear relationships, and constraints, making it difficult for traditional algorithms to adapt to changing market environments and large-scale data. These limitations lead to suboptimal portfolio strategies in dynamic situations. Recent advances in machine learning and artificial intelligence have significantly impacted portfolio management. Deep reinforcement learning (DRL), as explored by Jiang et al. (2017), Wang and Zhou (2020), and Cong et al. (2022), offers robust frameworks for developing adaptive and dynamic portfolio strategies. These methods leverage vast amounts of data and sophisticated algorithms to optimize asset allocation in real time. Compared to standard portfolio allocation methods such as Markowitz's paradigm, DRL aims to search for optimal sequences of actions and then obtain a multi-step task, where the objective is to achieve the maximum cumulative reward (Sutton & Barto, 2018). This allows DRL to adapt to complex, high-dimensional, and dynamic environments, making it an attractive method for improving traditional portfolio allocation.

The literature applying these techniques for multi-period portfolio allocation is rapidly growing. Aboussalah et al. (2021) applied the DRL approach based on convolutional neural networks (CNN) to construct optimal portfolios in a multi-period investment scenario. Corsaro et al. (2022) investigated the application of L1-regularization with machine learning and neural networks-based automatic selection to perform multi-period portfolio selection. Wei et al. (2021) and Chen and Ge (2021) showed the benefits of stochastic neural network algorithms to incorporate asymmetric investor sentiment and construct an investment portfolio that balances the return and risk over multiple holding periods. In a recent study, Cui et al. (2024) applied DRL to multi-period portfolio selection considering different risk aversion levels.

Other challenges to constructing optimal investment strategies over multi-period investment horizons are the presence of transaction costs and other market frictions. Research by Constantinides (1986) and Liu and Loewenstein (2002) incorporated these factors into dynamic models to obtain more realistic strategies that account for the costs associated with rebalancing portfolios. Building on this, Eom and Park (2017) investigated the impact of common factors by implementing a comparative correlation matrix on stocks. They emphasized that portfolios constructed by considering market factors significantly outperform other approaches in terms of diversification. However, with the increasing availability of high-dimensional and high-frequency financial data, more sophisticated models need to be developed. The works by Bernardi and Catania (2018) and Zhao et al. (2023) utilized copula models and Monte Carlo methods to capture dependencies and optimize portfolios in high-dimensional settings. Recent studies use the DRL framework to extract cross-asset dependence features in financial investments (Xu et al., 2020; Zhang et al., 2022). Notably, Marzban et al. (2023) introduced a WaveNet structure in a DRL framework for capturing cross-asset dependence and improve long-term portfolio optimization.

In this paper, we design an advanced portfolio policy framework to construct a multi-period portfolio selection model. To do this, we model the dynamics of asset prices and their cross-sectional dependence using machine learning methods (dynamic patterns of asset prices are obtained using convolutional neural networks, and group-wise asset dependence is modelled via WaveNet). The optimal long-term portfolio allocation is obtained by solving a multi-period Bellman equation combined with DRL methods. More specifically, we integrate the multi-period Bellman equation with DRL into a Markov Decision Process (MDP) framework. This approach accommodates different levels of risk aversion and is developed under a set of portfolio constraints. Empirical results demonstrate the effectiveness and superiority of our proposed portfolio strategy in various real-world settings. To show this, we conduct an in-depth comparative analysis under various investment holding periods, revealing the impact of the investment horizon, level of risk aversion, and portfolio constraints on portfolio management.

The rest of the paper is organized as follows: Section 2 presents the multi-period portfolio selection model and portfolio constraints. In Section 3, we propose an advanced portfolio policy framework based on DRL. The empirical results and analysis are provided in Section 4 and Section 5 concludes.

2. Multi-period portfolio optimization

This section introduces the theoretical framework for constructing multi-period optimal investment portfolios. According to Barberis (2000), one widely adopted strategy is the buy-and-hold approach, where investors choose a set of assets and hold onto them for an extended period before adjusting portfolio allocation. In contrast, the re-balancing strategy consists of periodically adjusting the portfolio to adapt to market fluctuations. In this paper we focus on the latter approach. Thus, we consider a dynamic multi-period portfolio optimization problem with a rebalancing strategy which optimizes the allocation of capital among financial assets over time.

We allocate funds into an investment portfolio at the beginning of each period over a planning horizon that extends h periods: t, t+1, t+2, …, $t + h$. We assume that a financial market has N risky assets, and the closing prices of these assets comprise a price vector $\boldsymbol{x}_t \in \mathbb{R}_+^N$ at time period t, where $\boldsymbol{x}_t = (x_{1,t}, \cdots, x_{N,t})$ with $x_{i,t}$ being the price of asset i. A portfolio is managed with a vector of these asset weights $\boldsymbol{\omega}_t = (\omega_{1,t}, \cdots, \omega_{N,t})^{\mathrm{T}} \in \mathbb{R}^N$, where $\omega_{i,t}$ denotes the proportion of portfolio value invested on the i-th asset. At the end of each period, investors could actively adjust the value of their portfolios in accordance with the realized returns and the most recent data available from financial markets. The vector of asset returns is expressed as:

$$\boldsymbol{r}_t = (r_{1,t}, r_{2,t}, \cdots, r_{N,t})^{\mathrm{T}} = \left(\frac{x_{1,t} - x_{1,t-1}}{x_{1,t-1}}, \frac{x_{2,t} - x_{2,t-1}}{x_{2,t-1}}, \cdots, \frac{x_{N,t} - x_{N,t-1}}{x_{N,t-1}}\right)^{\mathrm{T}}, \tag{1}$$

where $r_{i,t}$ represents the i-th asset return at time t. The portfolio value at period t is denoted by p_t and given, under the assumption that the change in portfolio weighs is small compared to the portfolio value, by the following expression:

$$p_t = p_{t-1}\left(1 + \boldsymbol{\omega}_{t-1}^{\mathrm{T}} \boldsymbol{r}_t\right) = p_{t-1}\left(1 + \sum_{i=1}^{N} \omega_{i,t-1} r_{i,t}\right). \tag{2}$$

The logarithmic rate of portfolio return at time t is defined as:

$$\widehat{r}_t = ln\left(\frac{p_t}{p_{t-1}}\right) = ln\left(1 + \boldsymbol{\omega}_{t-1}^{\mathrm{T}} \boldsymbol{r}_t\right) = ln\left(1 + \sum_{i=1}^{N} \omega_{i,t-1} r_{i,t}\right). \tag{3}$$

Sales and purchases of assets typically incur transaction costs, such as exchange fees and execution fees. Thus, it is necessary to consider the transaction cost in our portfolio trading selection. Here, let ξ denote the proportional cost level of each trading, and we set transaction cost rates for purchases and sales equal to ξ_t at time t. Furthermore, let $\psi_t \in [0, 1]$ denote the total transaction cost, which is defined as:

$$\psi_t = \xi_t \sum_{i=1}^{N} \left|\omega_{i,t} - \omega_{i,t-1}\right|. \tag{4}$$

The updated logarithmic rate of return at the end of period t is given by:

$$\widehat{r}_t = ln\left(\frac{p_t'}{p_{t-1}'}\right) = ln\left(\frac{(1 - \psi_t)p_t}{p_{t-1}'}\right) = ln\left((1 - \psi_t)\left(1 + \boldsymbol{\omega}_{t-1}^{\mathrm{T}} \boldsymbol{r}_t\right)\right). \tag{5}$$

Therefore, after considering transaction costs, the terminal portfolio value at time $t + h$ is expressed as (Moody et al., 1998; Zhang et al., 2022):

$$p_{t+h}' = p_t\left((1 - \psi_{t+1})\left(1 + \boldsymbol{\omega}_t^{\mathrm{T}} \boldsymbol{r}_{t+1}\right)\right)\left((1 - \psi_{t+2})\left(1 + \boldsymbol{\omega}_{t+1}^{\mathrm{T}} \boldsymbol{r}_{t+2}\right)\right) \cdots \left((1 - \psi_{t+h})\left(1 + \boldsymbol{\omega}_{t+h-1}^{\mathrm{T}} \boldsymbol{r}_{t+h}\right)\right) = p_t \prod_{k=1}^{h} \left((1 - \psi_{t+k})\left(1 + \boldsymbol{\omega}_{t+k-1}^{\mathrm{T}} \boldsymbol{r}_{t+k}\right)\right). \tag{6}$$

According to Eq. (6), the total portfolio log return over a planning horizon of h holding periods is expressed as:

$$\begin{aligned} R_h &= \sum_{k=1}^{h} \widehat{r}_{t+k} = \sum_{k=1}^{h} ln\left((1 - \psi_{t+k})\left(1 + \boldsymbol{\omega}_{t+k-1}^{\mathrm{T}} \boldsymbol{r}_{t+k}\right)\right) \\ &= \sum_{k=1}^{h} ln(1 - \psi_{t+k}) + \sum_{k=1}^{h} ln\left(1 + \boldsymbol{\omega}_{t+k-1}^{\mathrm{T}} \boldsymbol{r}_{t+k}\right). \end{aligned} \tag{7}$$

We consider the mean-variance function to model the one-period utility function representing investor's short-term preferences:

$$u_t = \widehat{r}_t - \lambda \sigma_t^2, \tag{8}$$

where λ is a risk-aversion parameter that balances the quantity placed on the maximization of portfolio return rate $\widehat{r}_t$ and the minimization of portfolio risk σ_t^2. The long-term portfolio allocation problem is characterized by the maximization of the investor's multi-period utility function computed over h periods and denoted by $U_{t,h}$. Investor's impatience is modelled by introducing a time

preference parameter γ that discounts future returns (Jaisson, 2022; Olschewski et al., 2021) such that the multi-period objective utility over h periods is defined as:

$$U_{t,h} = u_{t+1} + \gamma^1 u_{t+2} + \cdots + \gamma^{h-1} u_{t+h} = \sum_{k=1}^{h} \gamma^{k-1} u_{t+k}, \gamma \in [0,1], \tag{9}$$

when $\gamma = 0$, the investor's focus is solely on immediate return. For $0 < \gamma < 1$, the reward sequence converges given that the individual reward is finite.

Investors aim to construct a strategy that maximizes its expected long-term utility $U_{t,h}$ over the h-period investment horizon. The optimal portfolio is characterized by a weight matrix $\boldsymbol{\omega} = (\boldsymbol{\omega}_{t+1}, \boldsymbol{\omega}_{t+2}, \cdots, \boldsymbol{\omega}_{t+h})$ obtained over h periods and can be achieved by solving a multi-period optimization problem that incorporates a set of portfolio constraints, namely, a budget constraint, a turnover constraint, and a box constraint. The budget constraint is given by the following condition:

$$\sum_{i=1}^{N} \omega_{i,t} = 1, \forall t. \tag{10}$$

Similarly, the turnover constraint reduces the effect of the transaction costs on portfolio returns. Most studies use the average turnover when evaluating the influence of transaction costs, as it estimates the portfolio weight updates. The portfolio turnover (TO) constraint at time t can be expressed as:

$$TO_{i,t} = \left|\omega_{i,t} - \omega_{i,t-1}\right| \leq TO_t^{max}, \forall i, \forall t, \tag{11}$$

where TO_t^{max} is the maximum turnover rate of each asset at time t, $0 \leq TO_t^{max} \leq 1$. Finally, we also include a box constraint that avoids extreme investment positions and fosters the presence of diversification. To do this, we set an upper and lower bound for the maximum and minimum weights in the portfolio. Thus, the box constraint is defined as:

$$0 \leq \omega_{i,t} \leq \omega_i^{max}, \forall i, \forall t. \tag{12}$$

For simplicity, we only allow for long positions on the assets, and the objective function characterizing the optimal multi-period portfolio is expressed as:

$$\begin{cases} \max_{\omega} E\left[U_{t,h}\right] = \max_{\omega} \sum_{k=1}^{h} \gamma^{k-1} E[u_{t+k}] \\ = \max_{\omega} \left(E\left[\widehat{r}_{t+1} - \lambda\sigma_{t+1}^2\right] + \cdots + \gamma^{h-1} E\left[\widehat{r}_{t+h} - \lambda\sigma_{t+h}^2\right] \right). \end{cases} \tag{13}$$

Traditional optimization techniques (Bertsimas & Sim, 2004; Kamali et al., 2019) face difficulties in achieving a solution to the above optimal investment strategy of Eq. (13). For instance, convex optimization and Lagrange optimization methods may not fully leverage historical data and converge, instead, to a local rather than global optimal solution for the long-term portfolio strategy. Fortunately, DRL is a model-free dynamic programming strategy that can be adopted to tackle the above decision-making problem by learning the optimal policy in dynamic markets (Jaisson, 2022; Olschewski et al., 2021). Thus, the following section proposes a DRL-based portfolio approach to address an optimization problem in the same spirit of Eq. (13) but also including the above constraints as the components of the optimization function in the learning process.

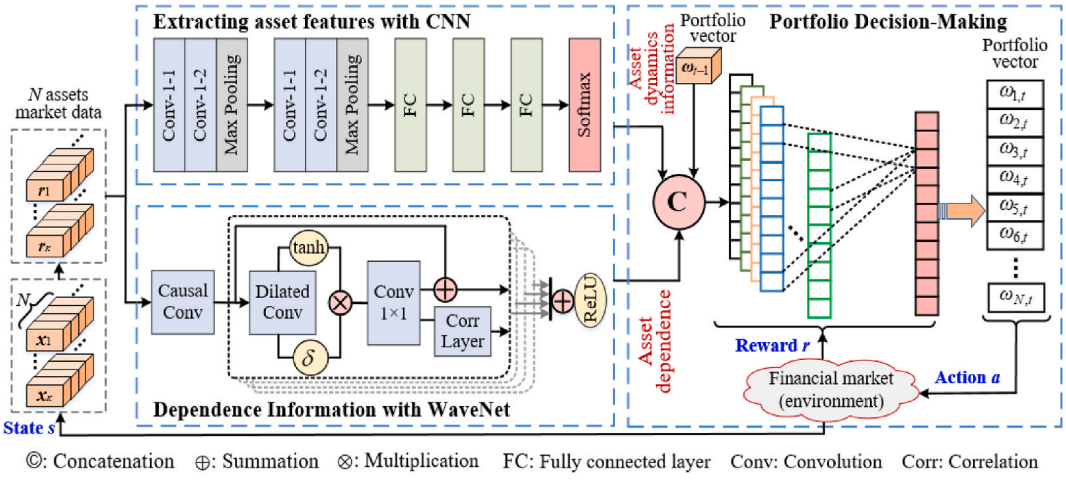


Fig. 1. Portfolio framework based on DRL with CNN and WaveNet.

3. Investor's long-term optimization problem

As mentioned above, the multi-period optimization problem defined in Section 2 will be addressed by proposing a portfolio framework based on DRL with CNN and WaveNet, as shown in Fig. 1.

The designed portfolio policy framework mainly includes three components. To begin with, the sequential information based on CNN is adopted to capture the dynamic patterns in each asset price. Secondly, the cross-dependence between the assets in the portfolio is modelled using WaveNet. This is particularly important in high-dimensional environments. Finally, the decision-making module is used to perform optimal portfolio allocation across assets. The following three subsections provide an in-depth exploration of each part.

3.1. Extraction of dynamic price sequence information based on CNN

Gu et al. (2018) stated that CNN can combine features and achieve a higher precision accuracy on large-scale datasets than other deep learning methods. In this case, we develop a sequential information module based on CNN to extract the changes in each asset price. This module aims to estimate the dynamic characteristics of high-dimensional vectors of asset prices and effectively model time series data. Moreover, by applying CNN models to analyse historical price data, we can predict the conditional volatility of asset returns in future periods.

As illustrated in Fig. 1, CNN extracts the nonlinear dynamic features of each asset separately on its multidimensional input data, thereby greatly improving the accuracy of the price mapping features. Additionally, the structure of CNN for price dynamic sequential information extraction comprises not only input and output layers but also convolutional layers, pooling layers, fully connected layers, and so on. The analysis focuses on the following key layers:

Input layer: This layer pre-processes the original asset returns, including normalizing the amplitude into the same range [0, 1], which reduces the interference caused by differences in the value range of data in various dimensions.

Convolutional layer: The convolutional kernels are used as effective methods for price feature extraction, and the result obtained from price data after convolution is called a Feature Map. The specific process can be expressed as:

$$y_{j,n} = \delta\left(b + \sum_{l=0}^{Z}\sum_{m=0}^{M} \varpi_{l,m} x'_{j+l,n+m}\right), \tag{14}$$

where δ represents the activation function, b is the shared bias parameter, Z and M are the length and width of the local receptive field, respectively, and $\varpi_{l,m}$ denotes the shared weight parameters between neurons, $x'_{j+l,n+m}$ is the data corresponding to the input matrix received by the convolutional layer. Widely used activation functions include the sigmoid, tanh, and ReLU functions, among a few others. The first two specifications are usually observed in fully connected layers, while the latter ReLU function applied commonly in convolutional layers is given by:

$$\delta(x') = ReLU(x') = \begin{cases} x', x' \geq 0 \\ 0, x' < 0 \end{cases}. \tag{15}$$

The fully connected layer is prone to overfitting due to its large number of parameters and the relationship between all elements of the output and input. Therefore, ReLU functions are added between each layer in the model as non-linear activation units to prevent overfitting and increase non-linear expression ability.

Pooling layer: The main objective of this layer is to remove unimportant samples from the Feature Map and thus reduce the number of parameters. Max pooling preserves the maximum value within each small block, which is equivalent to preserving the best matching result for that block.

Fully connected layer: Each node of the fully connected layer is connected to all the nodes of the previous layer and is used to synthesize the features extracted from the previous side. All neurons between the two layers are connected with weights, and the fully connected layer is usually at the tail of the convolutional neural network.

Softmax layer: This layer provides non-linear modeling capability by mapping the output results of the convolutional layer into nonlinear maps, which can effectively capture the asset price dynamics.

3.2. Cross-asset dependence information extraction based on WaveNet

WaveNet can effectively capture the cross-asset dependence information in portfolio management (Marzban et al., 2023). In this subsection, we apply the WaveNet framework to estimate the time-varying dependence across assets in the portfolio, denoted as $\boldsymbol{Q}_t$, where the dependence dynamics between two assets i and j, denoted as $q_{i,j,t}$, is adjusted over time based on a neural network function φ. More specifically, we use the WaveCorr layer from Marzban et al. (2023) as our convolution layer for capturing asset dependence in the WaveNet. This is associated with the following correlation layer (Corr-layer) function set:

$Q = \left\{q_{i,j,t} \in \mathbb{R}, a \in \mathbb{R}\right\}$, (16) where the vector of asset dependence $q_{i,t}(r_t) = \left[q_{i,1,t}, q_{i,2,t}, \cdots, q_{i,N,t}\right]$ between asset i and the remaining assets depends on a neural network function φ, expressed as:

$$\boldsymbol{q}_{i,t}(\boldsymbol{r}_t) = \left(\varphi(r_{i,t}) \odot (\mathbf{1}\boldsymbol{\omega}_0^{\mathrm{T}}) + \sum_{j=1}^{N} \varphi(r_{j,t}) \odot \left(\mathbf{1}\boldsymbol{\omega}_j^{\mathrm{T}}\right) \right) \mathbf{1} + a, \tag{17}$$

where a is the bias for accelerating neural network fitting. A general model operating directly on the assets returns is provided, where the joint probability of an input stream $r_t = [r_{1,t}, \cdots, r_{N,t}]$ is modelled as the product of the probabilities conditional on the realization of past returns (Marzban et al., 2023; Van Den Oord et al., 2016), i.e.,

$$\eta(\boldsymbol{r}_t) = \prod_{i=1}^{N} \eta\left(r_{i,t} | r_{1,t-1}, \cdots, r_{N,t-1}\right). \tag{18}$$

Each sample $r_{i,t}$ of the i-th asset is conditioned on the samples at all previous time steps. The causal convolution operation extracts the price dynamics, but it may need large kernel sizes and layers. Thus, apart from adopting causal convolution, the dilated operation is also applied to meet the exponentially large receptive fields with only a few layers while maintaining the network and computational efficiency. In the WaveNet structure, a softmax distribution is adopted to model the conditional distribution $\eta(r_t)$, even if the asset returns are implicitly continuous. In addition, residual and parameterized skip connections are adopted to enhance the training convergence. With the help of WaveNet, the dependence information between assets can be constructed in a multi-block framework as illustrated in Fig. 1.

3.3. Multi-period portfolio decision-making based on DRL

The dynamic asset price features and dependence information obtained from CNN and WaveNet are combined in 'C', as shown in Fig. 1, as the input of the deterministic policy gradient (DPG) model for portfolio decision-making.

3.3.1. MDP with multi-period Bellman equation

DRL typically combines the MDP framework to address the challenges posed by the multi-period Bellman equation. MDP provides a mathematical foundation for describing the interaction between an agent and an environment, incorporating elements such as states, actions, rewards, and state transitions.

The portfolio management problem (Eq. (13)) is defined as a MDP with a tuple $(\mathscr{S}, \mathscr{A}, \mathscr{P}, u)$. Specifically, the learning agent (i.e., an investor) observes one state $s_t \in \mathscr{S}$ (i.e. assets' daily prices, latest asset returns, cross dependencies) from the market and then chooses an action $a_t \in \mathscr{A}$ (i.e., portfolio weight vector $\boldsymbol{\omega}_t$). Afterwards, the agent will achieve an instantaneous reward u_t and observe the next state s_{t+1}. Here, let $\pi(a_t, s_t)$ denote a portfolio policy, mapping from observed states with a Markovian transition probability $\mathscr{P}(s_{t+1}|s_t = s, a_t = a)$ over available actions that the agent selects. Note that the objective function at the t-th period u_t in Eq. (8) is the immediate reward function, and the multi-period utility U_h over h periods in Eq. (9) is the long-term reward function.

According to the discussions of the multi-period optimization constraints and objective function in Section 2, the investor aims to trade off risk and return in the portfolio under several constraints. In applying the DRL framework to optimization problems, similar to the studies of Zhang et al. (2022) and Marzban et al. (2023), we also incorporate the turnover and box constraints in Eq. (11) and Eq. (12) as penalties within the objective function. This approach enables the model to incorporate penalty functions on the utility function to penalise portfolio weights not satisfying the constraints. In this context, the risk-averse and constraint-awareness reward function at the t-th single period is designed as follows:

$$u_t = \widehat{r}_t - \lambda\sigma_t^2 - c_1 \sum_{i=1}^{N} max\left(0, TO_{i,t} - TO_t^{max}\right) - c_2 \sum_{i=1}^{N} max\left(0, \omega_{i,t} - \omega_i^{max}\right), \tag{19}$$

where c_1 and c_2 are the parameters for the unsatisfied maximum turnover constraint (Eq. (11)) and maximum weight constraint (Eq. (12)). The period reward function above is penalized if the turnover $TO_{i,t}$ or portfolio weight $\omega_{i,t}$ goes beyond TO_t^{max} and ω_i^{max}, respectively. Note that the portfolio weight of each asset i is non-negative and the sum of portfolio weights over all asset is equal to 1.

The parameters c_1 and c_2 are set by balancing the trade-off between the portfolio return $\widehat{r}_t$ and the two punishment values and need to be carefully selected. If parameters are too large, the reward function will sacrifice most of the portfolio return and risk performance in order to meet the turnover and box constraints. In contrast, the impact of the parameters on the reward function is limited if c_1 and c_2 are too small. We calibrate the selection of c_1 and c_2 and choose values given by $c_1 = c_2 = 0.5$. These values are selected using the empirical insights obtained in Zhang et al. (2022) and Van Den Oord et al. (2016).

The state-value function for policy in MDP over a planning horizon of h holding periods can be described as follows:

$$V_\pi(s) = E_\pi\left(U_{t,h} \middle| s_t = s\right) = E_\pi\left(\sum_{k=1}^{h} \gamma^{k-1} u_{t+k} \middle| s_t = s \right), \tag{20}$$

where $V_\pi(s)$ is the expected reward under the policy π and the state s. The expectation is computed based on the agent's policy mapping π. Similarly, we define the action-value function for the policy π by using Q_π as follows:

$$Q_\pi(s,a)=E_\pi\left(U_{t,h}\middle|s_t=s,a_t=a\right)=E_\pi\left(\sum_{k=1}^{h}\gamma^{k-1}u_{t+k}\middle|s_t=s,a_t=a\right), \tag{21}$$

where $Q_\pi(s,a)$ denotes the expected reward function at state s when performing action a and following policy π. For simplicity, the transition probability is denoted by $\mathscr{P}^a_{ss'}=\mathscr{P}(s_{t+1}|s_t=s,a_t=a)$. Additionally, the expected reward for transitioning from the current state s to the next state s' in the next period (s_{t+1}) by taking action a is denoted by $u^a_{ss'}=E(u_{t+1}|s_t=s,a_t=a,s_{t+1}=s')$.

The self-consistency of the value function indicates that certain recursive relationships are required to be met. The multi-period Bellman equation $V_\pi(s)$ can be expressed as:

$$\begin{aligned}
V_\pi(s)&=E_\pi\left(U_{t,h}\middle|s_t=s\right)\\
&=E_\pi\left(u_{t+1}+\gamma u_{t+2}+\gamma^2u_{t+3}+\cdots\gamma^{h-1}u_{t+h}\middle|s_t=s\right)\\
&=E_\pi\left(u_{t+1}+\sum_{k=2}^{h}\gamma^{k-1}u_{t+k}\middle|s_t=s\right)\\
&=\sum_a\pi(s,a)\sum_{s'}\mathscr{P}^a_{ss'}\left(u^a_{ss'}+\gamma E_\pi\left(\sum_{k=2}^{h}\gamma^{k-1}u_{t+k}\middle|s_{t+1}=s'\right)\right)\\
&=\sum_a\pi(s,a)\sum_{s'}\mathscr{P}^a_{ss'}\left(u^a_{ss'}+\gamma V_\pi(s')\right)\\
&=\sum_a\pi(s,a)Q_\pi(s,a).
\end{aligned} \tag{22}$$

The solution to the Bellman equation is the value function.

3.3.2. Multi-period portfolio based on DRL

The learning agent (investor) aims to achieve the multi-period portfolio reward U_h over h periods. Here, we adopt a DRL-based deterministic policy gradient to obtain the optimal portfolio policy. As shown in Fig. 2, the multi-period investment has h state-action pairs i.e., (s_{t+1},a_{t+1}), (s_{t+2},a_{t+2}), …, (s_{t+h},a_{t+h}), and the agent aims to maximize both the state-value function $V_\pi(s)$ and the action-value $Q_\pi(s,a)$. This is achieved by selecting the optimal action vector $a=(a_{t+1},a_{t+2},\cdots,a_{t+h})$ over h horizons based on the observed state vector $s=(s_{t+1},s_{t+2},\cdots,s_{t+h})$.

Deep learning with a set of neural network parameters $\boldsymbol{\theta}$ is used to specify the policy in the DRL framework, i.e., $\pi_{\boldsymbol{\theta}}(s,a)$. The objective of DRL is to maximize $U_{t,h}$ over the time interval $[t+1,t+h]$ generated by $\pi_{\boldsymbol{\theta}}(s,a)$ as expressed:

$$\max_{\boldsymbol{\theta}}J(\pi_\theta)=\mathbb{E}_{\pi_\theta(s,a)}\left(U_{t,h}(u_{t+1}(\boldsymbol{\omega}_{t+1}),u_{t+2}(\boldsymbol{\omega}_{t+2}),\cdots,u_{t+h}(\boldsymbol{\omega}_{t+h}))\right). \tag{23}$$

DPG learning methods enable the agents to learn portfolio strategies through real-time interaction with financial markets. The agent continuously observes market information and learns adaptive strategies during their interaction. This method is usually applicable to real-time decision-making for financial markets based on the current observed state of the environment. We use the state distribution $\rho^\pi(s)$ such that the objective function in Eq. (23) is given by:

$$\begin{aligned}
J(\pi_{\boldsymbol{\theta}})&=\int_{\mathscr{S}}\rho^\pi(s)\int_{\mathscr{A}}\pi_{\boldsymbol{\theta}}(s,a)Q_\pi(s,a)dads\\
&=\mathbb{E}_{s\sim\rho^\pi,a\sim\pi_{\boldsymbol{\theta}}}[Q_\pi(s,a)],
\end{aligned} \tag{24}$$

and the gradient of the objective function (see Sutton and Barto, 2018) is expressed as:

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}}J(\pi_{\boldsymbol{\theta}})&=\int_{\mathscr{S}}\rho^\pi(s)\int_{\mathscr{A}}\nabla_{\boldsymbol{\theta}}\pi_{\boldsymbol{\theta}}(s,a)Q_\pi(s,a)dads\\
&=\mathbb{E}_{s\sim\rho^\pi,a\sim\pi_{\boldsymbol{\theta}}}[\nabla_{\boldsymbol{\theta}}\,log\,\pi_{\boldsymbol{\theta}}(s,a)Q_\pi(s,a)].
\end{aligned} \tag{25}$$

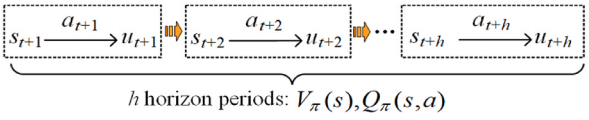


Fig. 2. The multi-period portfolio trajectory based on DRL.

Note that the action-value function $Q_\pi(s,a)$ in Eq. (24) and Eq. (25) is computed or updated by Eq. (21) using a neural network (value network). DPG adjusts the parameters $\boldsymbol{\theta}$ of the strategy towards the gradient direction of the objective function to maximize the objective function. The mathematical expression for parameter update is as follows:

$$\boldsymbol{\theta}' \leftarrow \boldsymbol{\theta} + \alpha \nabla_{\boldsymbol{\theta}} J(\pi_{\boldsymbol{\theta}}), \tag{26}$$

where α denotes the learning rate and $\nabla(\cdot)$ is the first order partial derivative. It is necessary to sample the state and actions under the corresponding distribution function.

The decision-making process is to evaluate the potential growth of assets in the near future and consider the portfolio weight vector of the investment based on the previous action a_t to obtain a new portfolio weight $\boldsymbol{\omega}_t$. In this case, the investor can achieve the multi-period portfolio weight matrix $\boldsymbol{\omega} = (\boldsymbol{\omega}_{t+1}, \boldsymbol{\omega}_{t+2} \cdots, \boldsymbol{\omega}_{t+h})$ over h investment periods; $\boldsymbol{\omega}$ captures the investment behavior of optimizing agents, ultimately guiding the asset portfolio selection action $\boldsymbol{a} = (a_{t+1}, \cdots, a_{t+h})$.

4. Empirical application

This section illustrates the performance of the proposed portfolio approach under different scenarios and datasets. We operate in a high-dimensional setting, interpreted as a portfolio with more than 50 assets (Ding et al., 2021) except when analysing the Dow Jones financial index that is comprised by 30 assets.

4.1. Datasets and competing portfolio construction methods

We consider three important financial indices reflecting the performance of stock markets and the overall economy for the US and Canada. For the US, we consider the S&P100 index and the DJIA. These indices capture similar dynamics of the US stock market, however, whereas the S&P100 index is value-weighted the DJIA is price-weighted and comprised by a much smaller number of stocks. The Canadian stock market is represented by the S&P/TSX Composite Index. Our data span from 04/01/2010 until 12/07/2023 and are divided into a training set and a test set, with the training period spanning from 04/01/2010 to 31/12/2018 and the testing period from 02/01/2019 to 12/07/2023 used for out-of-sample evaluation of the different models and methods.

The main purpose of our empirical study is to see the performance of our proposed approach for portfolio allocation under different choices of the investment horizon. To assess the robustness of the results and the influence of important factors such as risk aversion, constraints on the portfolio weights, or the presence of transaction costs, we carry out the analysis for different values of these parameters. As a second empirical contribution, we compare the performance of our proposed long-term investment strategy with existing methods in the literature that act as benchmarks. Most of these investment strategies used for model comparison are state-of-the-art techniques taking advantage of machine learning methods but we also consider the equally-weighted (EW) portfolio to capture a more traditional and naïve portfolio allocation strategy.

More formally, we consider the following methods to construct optimal investment portfolios: 1) Our proposed advanced multi-period DRL-based portfolio method combined with the WaveNet-enabled dependence information and CNN-enabled sequential information. This method is denoted as MP-Adv-DRL-Cor; 2) The multi-period cost-sensitive portfolio selection method using CNN to extract the dynamic asset return features, temporal correlational convolution block (TCCB) to perform asset correlation and portfolio policy network (PPN) to obtain the portfolio selection, respectively. This method is denoted as MP-CS-PPN-Cor (Zhang et al., 2022); 3) The multi-period DPG-based portfolio method with Ensemble of Identical Independent Evaluators (EIIE) algorithm (Jiang et al., 2017), denoted by MP-DPG; 4) The equally-weighted portfolio method, denoted by EW, 5) The single-period DRL-based portfolio method combined with WaveNet and CNN, denoted by SP-Adv-DRL-Cor. Note that the MP-Adv-DRL-Cor, MP-CS-PPN-Cor, and MP-DPG methods optimize the objective reward function provided in Eq. (19) that includes transaction costs in the optimal portfolio problem. The inclusion of Strategy 5 is to assess the differences in portfolio performance of our proposed methodology between long- and short-term investment horizons.

All of the strategies except the EW portfolio require some prior definition of a set of hyperparameters. In order to improve the learning efficiency of machine learning-based portfolio algorithms, it is crucial to choose appropriate hyperparameters. The values characterizing the architecture of the neural network models, DRL method, and portfolio constraints are shown in Table 1. These values are standard in the related literature and adopted by Sutton and Barto (2018), Zhang et al. (2022), Marzban et al. (2023), and Hambly et al. (2023), among others. For example, we choose a moderate learning rate of 0.0002 as in Sutton and Barto (2018) to

Table 1
Hyperparameter values.

Hyperparameter	Value	Hyperparameter	Value
Learning rate α	2×10^{-4}	Decay rate	0.9999
Optimizer γ	Adam	Planning horizon h	36
Discount factor	0.98	Look back window size	36
Mini-batch size	32	Number of epochs	1000
Hidden layers of CNN	2	Parameter c_1, c_2	0.5
Hidden layers of WaveNet	7	Maximum weight ω_i^{max}	0.7
Hidden layer size	256	Maximum turnover TO_t^{max}	0.5

ensure a balance between convergence speed and model stability. Instead, if the learning rate is very small (e.g., $a = 0.00001$), the convergence speed is slow and the training time increases significantly. In contrast, if the learning rate is very large (e.g., $a = 0.01$), it may lead to dramatic fluctuations and an unstable training process. The number of hidden layers and their size are also taken from the related literature. If too many hidden layers are set the model may become overly complex, leading to excessive computational complexity. Conversely, an insufficient number of hidden layers may limit the model's learning ability, which in turn affects model performance. It is also worth discussing the choice of portfolio constraints and the discount rate. Following the work of Marzban et al. (2023), the maximum weight is set to be 0.7 to avoid excessive leverage on specific assets, and the discount rate is set to 0.98, following the suggestions in Sutton and Barto (2018), Zhang et al. (2022), and Hambly et al. (2023).

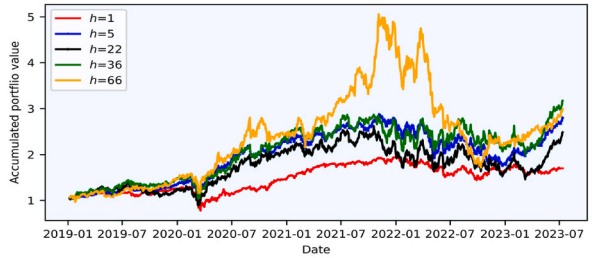
4.2. Performance measures

All of the following empirical results are evaluated using out-of-sample data ("test data"). Different metrics are adopted to measure portfolio performance. Firstly, as an indicator of the return on investment, the accumulated portfolio value (APV) is used to evaluate the increase in portfolio value over time. This performance measure considers the effect of transaction costs and is expressed as:

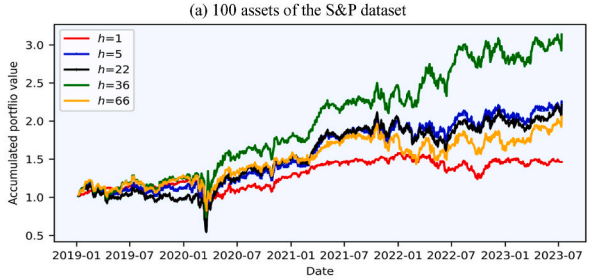
$$APV = p_0 \prod_{t=1}^{T} \left((1 - \psi_{t+k})\left(1 + \boldsymbol{\omega}_{t+k-1}^{T} \boldsymbol{r}_{t+k}\right)\right), \tag{27}$$

where p_0 is the initial value of the portfolio and ψ_t represents the percentage of transaction Costs. APV typically focuses on total value without considering the underlying risk in the Portfolio. To control for the underlying risk, we follow the investment literature and employ The Sharpe ratio (SR) as a second indicator of performance:

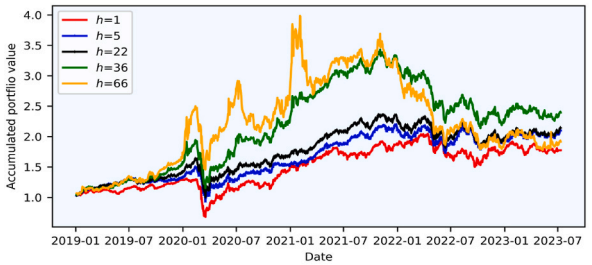
$$SR = \frac{E_t\left(\widehat{r}_{t+1} - r_{t,f}\right)}{\sigma_t}, \tag{28}$$



(a) 100 assets of the S&P dataset



(b) 30 assets of the DJIA dataset



(c) 50 assets of the S&P/TSX dataset

Fig. 3. APV for MP-Adv-DRL-Cor under different investment horizons h for the three financial indices over the out-of-sample evaluation period 02/01/2019 to 12/07/2023. $\xi = 0.01\%$ for transaction costs and $\lambda = 0.01$ for the risk aversion coefficient.

where $\widehat{r}_t$ is the rate of return as defined in Eq. (5) at time t and $r_{t,f}$ is the risk-free rate. The Sharpe ratio uses the standard deviation as a measure of risk without differentiating between upward and downward volatility, which means it may overly focus on short-term adverse fluctuations and overlook positive ones. Consequently, to better identify and evaluate a portfolio's downside risk, we turn to the Maximum drawdown (MDD). MDD measures actual losses and reflects the maximum potential loss a portfolio may face. It is defined as:

$$\mathrm{MDD} = \max_{t:j>t} \frac{\left(p_t' - p_j'\right)}{p_t'}, \tag{29}$$

where $j > t$; p_t is the total value of the portfolio at time t as expressed in Eq. (2), and p_j is the aggregate value of the portfolio at time j. In general, a lower value of the MDD metric reflects a more stable and lower-risk investment.

4.3. Effects of investment horizon on portfolio performance

We evaluate empirically the influence of the holding period h on our proposed MP-Adv-DRL-Cor method. This investment strategy corresponds to Method 1 in the above description of the portfolio strategies. For comparison purposes, we consider as a separate investment strategy, called Method 5, the myopic version of Method 1 that is constructed for an investment horizon of $h = 1$. The baseline parameters considered for this exercise are $\xi = 0.01\%$ for the transaction costs and $\lambda = 0.01$ for the risk aversion coefficient. These parameters represent very low levels of transaction costs and risk aversion, respectively. The purpose of using these values is to fully capture the potential of our investment strategy over a multi-period horizon by minimizing the effect of risk aversion and transaction cost penalties on the objective function. The results of this exercise for the three financial indices under investigation are shown in Fig. 3 and Table 2.

The results from Fig. 3 and Table 2 show that extending the holding period ($h = 1, 5, 22, 36$) generally leads to increased portfolio gains under the proposed MP-Adv-DRL-Cor method but also results in higher annual realized volatilities. This is because when considering extended investment periods, investors can maximize the total utility over the long term, which typically avoids making myopic short-term investment decisions. The trade-off between mean return and volatility is captured by the Sharpe ratio. The performance of this indicator is not monotonic over the investment horizon. The results for the S&P100 and DJIA financial indices show an increase in Sharpe ratio up to $h = 36$ trading days and then a drop in value due to a decrease in portfolio return accompanied by an increase in volatility. The results for the Canadian index are similar in the sense that the performance metrics report an optimal investment horizon beyond which portfolio performance decreases. However, in contrast to the US stock indices, portfolio performance measured by the Sharpe ratio reaches a peak at $h = 22$ days and then drops. The reason for this drop in profitability adjusted for risk is a substantial increase in volatility not followed by a similar rise in mean return. In fact, the mean return falls from $h = 36$ to 66 investment horizon. Therefore, continually extending the investment horizon does not always yield better portfolio performance. Thus, the horizon period h needs to be carefully selected. These results are broadly consistent with the literature on long-term portfolio allocation. The profitability of the portfolio increases with the investment horizon, while the annual volatility is also expected to rise.

Table 2
Portfolio performance metrics under different holding periods h on three datasets.

Holding period	Annual return (%)	Annual volatility (%)	Sharpe Ratio	Maximum drawdown (%)	Turnover
S&P 100 Index					
$h = 1$	12.48	24.57	0.508	39.80	0.007
$h = 5$	25.72	28.89	0.890	39.19	0.087
$h = 22$	22.37	37.08	0.603	44.13	0.099
$h = 36$	29.21	36.14	0.808	32.09	0.170
$h = 66$	27.51	39.43	0.698	66.10	0.107
DJIA Index					
$h = 1$	8.808	21.43	0.411	35.34	0.007
$h = 5$	19.79	33.51	0.590	51.72	0.080
$h = 22$	19.23	32.49	0.592	52.63	0.103
$h = 36$	28.88	34.32	0.841	44.24	0.109
$h = 66$	17.40	31.32	0.555	36.22	0.092
S&P/TSX Composite Index					
$h = 1$	13.58	25.35	0.536	48.01	0.008
$h = 5$	17.91	21.58	0.830	37.94	0.009
$h = 22$	18.42	22.10	0.833	38.65	0.020
$h = 36$	21.38	27.08	0.790	40.83	0.074
$h = 66$	15.50	37.11	0.418	55.38	0.197

4.4. Portfolio performance under different risk aversion levels

The next exercise shows the effect of the risk aversion coefficient λ for different holding periods. We consider $h = 5$ and $h = 36$ for which Fig. 3 shows that the relationship between portfolio performance and investment horizon is monotonically increasing for small values of risk aversion. The aim of this exercise is to increase the level of risk aversion and see the effect on portfolio performance for each investment horizon. For space considerations, we fix the transaction cost rate at $\xi = 0.01\%$ and consider only the S&P100 index.

Fig. 4 shows that the cumulative portfolio value tends to decrease as λ increases, and its trajectories also become less volatile. Table 3 provides further clarity on the rationale for these dynamics. Both annual return and volatility decline as the level of risk aversion increases, however, the drop is more acute in mean return than in volatility. This downward trend is especially noticeable when λ rises from 0.1 to 1. An increase in λ implies that investors are more inclined to select conservative strategies to mitigate portfolio risks. This preference leads to a decline in trading frequency and investment activity, as shown in the Turnover column of Table 3. Specifically, when the risk aversion coefficient is very high, i.e. $\lambda = 1$, portfolio volatility is significantly reduced. Consequently, the potential for substantial annual returns and a high Sharpe ratio is limited. For instance, under $h = 36$, the annual return is only 9.54% with $\lambda = 1$ compared to 29.21% when $\lambda = 0.01$. The figures in Table 3 also provide some insights into the fruitfulness of considering higher investment horizons for a given level of risk aversion. Thus, the comparison of rows with same levels of risk aversion across panels suggests that both the annual return and the volatility are higher as h increases. However, the ratio between both quantities given by the Sharpe ratio is slightly more favourable to the longer investment horizon reinforcing the idea that longer investment horizons may be more profitable for a given degree of risk aversion.

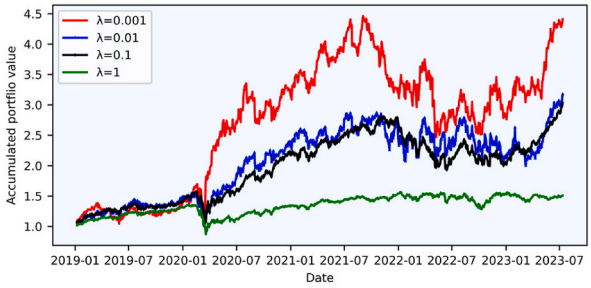
A related question is how to choose realistic values of the risk aversion coefficient for the reward functions introduced above. Zhang et al. (2022), in a similar context, explored different values of this coefficient and found a better balance between risk and return when λ is around 0.01. Similar results are obtained in our empirical exercise when the risk aversion coefficient rises from 0.001 to 0.1 for $h = 5$, however, it is worth noting the excellent performance of the method for very low levels of risk aversion under longer investment horizons.

4.5. Portfolio performance under different transaction costs

This subsection evaluates the role of transaction costs on the performance of multi-period investment portfolios. As in the previous exercises, the optimal portfolios are constructed using the MP-Adv-DRL-Cor method. Risk aversion is fixed at $\lambda = 0.01$ and transaction costs vary between a low value given by $\xi = 0.05\%$ and a high value given by 0.5%. For comparison purposes, we only consider optimal portfolios comprised by the assets in the S&P100 index. Fig. 5 and Table 4 show that the annual returns are higher when the transaction cost rate ξ is low at 0.05%, compared to the annual returns for the higher rate of 0.5%. This result holds across investment horizons. In contrast, portfolio annual volatility is hardly affected by the presence of transaction costs, entailing a decrease in Sharpe ratios as ξ increases across values of h.

The presence of transaction costs mainly affects the profitability of the portfolios and does not increase risk. It also has a major

(a) Portfolio performance under different risk aversion levels λ when h =5

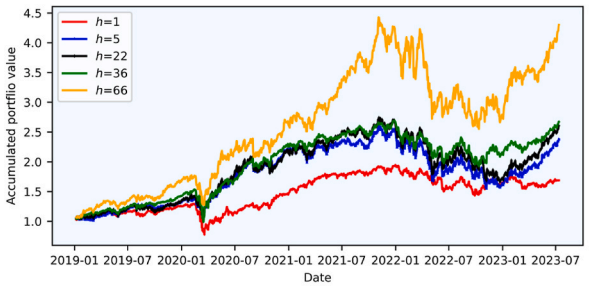


(b) Portfolio performance under different risk aversion levels λ when h =36

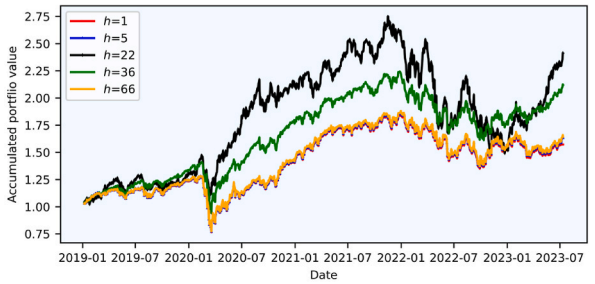
Fig. 4. APV for MP-Adv-DRL-Cor under different risk aversion levels for the S&P100 index over the out-of-sample evaluation period 02/01/2019 to 12/07/2023. $\xi = 0.01\%$ for transaction costs and $h = 5$ (top panel) and $h = 36$ (bottom panel).

Table 3
Portfolio performance metrics under different risk aversion coefficients λ.

λ	Annual return (%)	Annual volatility (%)	Sharpe ratio	Maximum drawdown (%)	Turnover
$h = 5$					
$\lambda = 0.001$	26.11	29.96	0.872	39.55	0.082
$\lambda = 0.01$	25.71	28.89	0.890	39.19	0.087
$\lambda = 0.1$	21.25	23.55	0.902	31.17	0.021
$\lambda = 1$	9.037	19.70	0.459	36.08	0.017
$h = 36$					
$\lambda = 0.001$	39.99	38.44	1.014	44.66	0.333
$\lambda = 0.01$	29.21	36.14	0.808	32.09	0.170
$\lambda = 0.1$	27.91	25.99	1.074	31.64	0.064
$\lambda = 1$	9.539	19.22	0.496	36.05	0.032



(a) Portfolio performance under different holding periods h when ξ=0.05%



(b) Portfolio performance under different holding periods h when ξ=0.5%

Fig. 5. APV for MP-Adv-DRL-Cor under different investment horizons h for the S&P100 index over the out-of-sample evaluation period 02/01/2019 to 12/07/2023 for $\xi = 0.05\%$ (top panel) and $\xi = 0.5\%$ (bottom panel).

Table 4
Portfolio performance metrics for different periods h when $\xi = 0.05\%$ and 0.5%.

Holding period	Annual return (%)	Annual volatility (%)	Sharpe ratio	Maximum drawdown (%)	Turnover
Transaction cost rate $\xi = 0.05\%$					
$h = 1$	12.32	24.57	0.502	39.82	0.007
$h = 5$	21.20	27.45	0.772	40.35	0.019
$h = 22$	23.63	28.52	0.829	39.32	0.033
$h = 36$	24.38	24.83	0.982	35.54	0.030
$h = 66$	38.21	33.02	1.157	42.34	0.063
Transaction cost rate $\xi = 0.5\%$					
$h = 1$	10.56	24.56	0.430	40.00	0.007
$h = 5$	11.48	24.58	0.467	39.97	0.006
$h = 22$	21.56	30.38	0.710	45.73	0.005
$h = 36$	18.14	23.19	0.782	33.25	0.006
$h = 66$	11.64	24.52	0.475	39.80	0.006

impact on the turnover of the portfolios. Under the presence of transaction costs, investors have fewer incentives to rebalance their portfolios under changes in investment opportunities over time. The effect on portfolio turnover is particularly visible for longer investment horizons, we observe a decrease of an order of magnitude in the performance measure for high values of the transaction

costs. Importantly, the results for $\xi = 0.05\%$ are very similar to the results reported in Table 2 for $\xi = 0.01\%$ and suggest that portfolios' risk-adjusted profitability is monotonically increasing on the investment horizon. In contrast, for high transaction costs, portfolio profitability is reduced but not the underlying volatility, implying that optimal portfolios constructed for intermediate investment horizons report higher Sharpe ratios.

The results in Fig. 5 and Table 4 are qualitatively similar to the analysis of risk aversion. Transaction costs reduce portfolio profitability. Interestingly, this result is stronger as the investment horizon rises entailing sharper declines for $h = 66$.

4.6. Portfolio performance comparisons

The previous subsections have shown the ability of our proposed procedure to construct optimal portfolios for different investment horizons and under different choices of transaction costs and risk aversion. This subsection complements this analysis by comparing the performance of our procedure against existing competitors, most of them drawn from the machine learning literature on portfolio allocation. The aim of this exercise is to show that the MP-Adv-DRL-Cor procedure outperforms these methods under most scenarios given by different levels of risk aversion and transaction costs. Figs. 6–8 present empirical results for the S&P100 index, DJIA, and the S&P/TSX index for investment horizons $h = 1$, 22, 66 with $\xi = 0.05\%$ and $\lambda = 0.1$. Additional results for other combinations of transaction costs and risk aversion ($\xi = 0.5\%$ and $\lambda = 0.1$, $\xi = 0.05\%$ and $\lambda = 1$) are available from the authors upon request.

The APV associated to the EW method is the same across investment horizons and serves as a helpful benchmark for comparing the performance of the remaining competitors. For $h = 1$, the APVs are surprisingly similar across investment strategies suggesting that for $\xi = 0.05\%$ and $\lambda = 0.1$, the use of the proposed techniques based on machine learning methods are not necessarily superior to naïve investment methods such as the EW portfolio. Sophisticated techniques are superior in settings characterized by investment over multiple periods. The MP-Adv-DRL-Cor method exhibits superior performance than the other four approaches for investment horizons greater than one period. It is also worth noting the good performance of MP-CS-PPN-Cor, however, there are cases such as $h = 22$ for

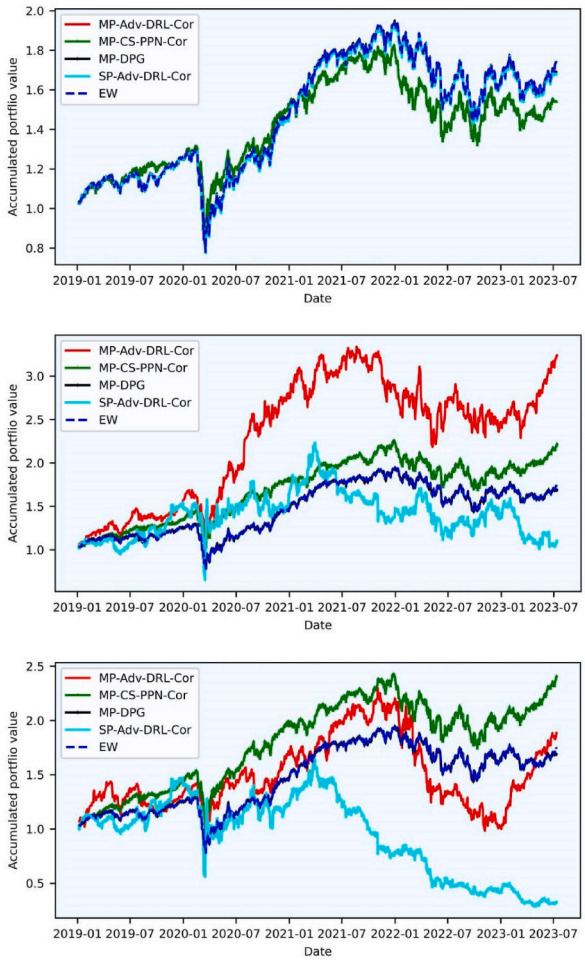


Fig. 6. APV for the five investment portfolios evaluated for S&P100 Index data over the period 02/01/2019 to 12/07/2023 for $h = 1$ (top panel), $h = 22$ (middle panel) and $h = 66$ (bottom panel) when $\xi = 0.05\%$ and $\lambda = 0.1$.

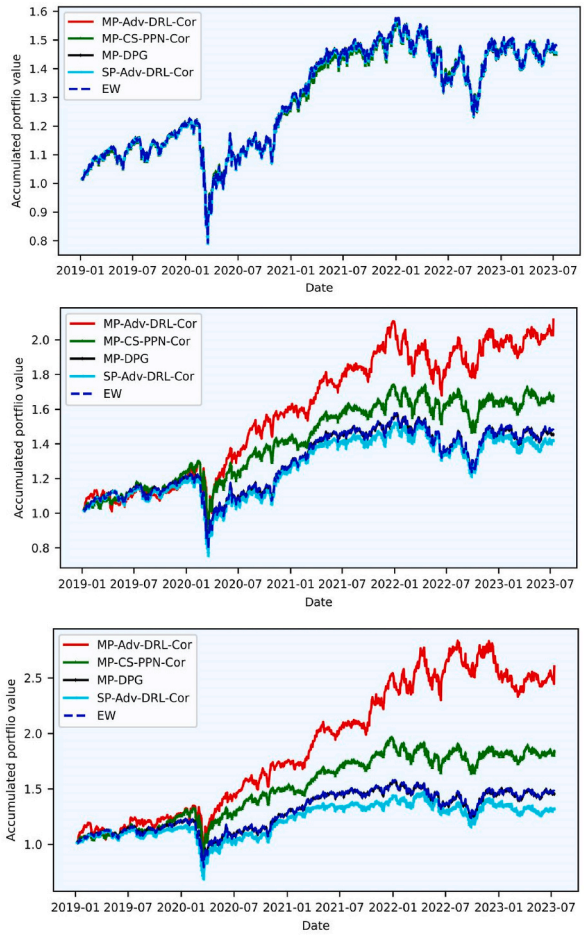


Fig. 7. APV for the five investment portfolios evaluated for DJIA Index data over the period 02/01/2019 to 12/07/2023 for $h = 1$ (top panel), $h = 22$ (middle panel) and $h = 66$ (bottom panel) when$\xi = 0.05\%$ and $\lambda = 0.1$.

the DJIA index in which the method reports very poor results. Table 5 formalizes the results in Fig. 6 for the S&P100 index. For $h = 1$, we observe similar results for the five performance measures under investigation across methods. The results for $h = 22$ in the middle panel reveal the strong performance of MP-Adv-DRL-Cor in terms of annual return and Sharpe ratio and the poor performance of MP-DPG.

The MP-Adv-DRL-Cor and MP-CS-PPN-Cor investment strategies are the best performers for the three data sets. Both methods adopt correlation networks, however, MP-CS-PPN-Cor adopts the TCCB framework. This method is not invariant to the ordering of the assets in the portfolio implying that portfolio performance can greatly vary when the ordering of the assets changes. Unlike TCCB, the WaveNet approach contains a simpler permutation invariant structure that can efficiently capture asset correlation, thus achieving higher portfolio performance under different choices of transaction costs and risk aversion. This result is because the multi-period strategic investment employs a dynamic asset allocation strategy to adapt the portfolio weights under time-varying financial conditions. Also, the multi-period strategy is more effective in managing portfolio's overall risk by considering long-term risk and return. In contrast, a single-period (tactical investing) focuses more on short-term market volatility and takes advantage of market opportunities to maximize wealth in the short term while ignoring long-term investment opportunities.

4.7. Dimensionality effects on portfolio performance

An interesting feature of our proposed procedure is the ability to work with high-dimensional portfolios given by a large number of investment assets. This is illustrated in the following empirical exercise in which we also compare the performance of MP-Adv-DRL-Cor against the four competitors discussed above. To do this, we consider the universe of assets in the S&P500 index and take random subsets of 50, 75, 100 and 200 assets. The results are provided in Fig. 9 and Table 6. Each panel of Fig. 9 represents the APV of the five competing portfolios for different number of assets. The empirical results provide overwhelming evidence on the outperformance of our proposed approach across different choices of the number of assets. The differences in APV are more important as the number of assets comprising the portfolios increases. Interestingly, the dynamics and magnitudes of the cumulative portfolio returns do not vary

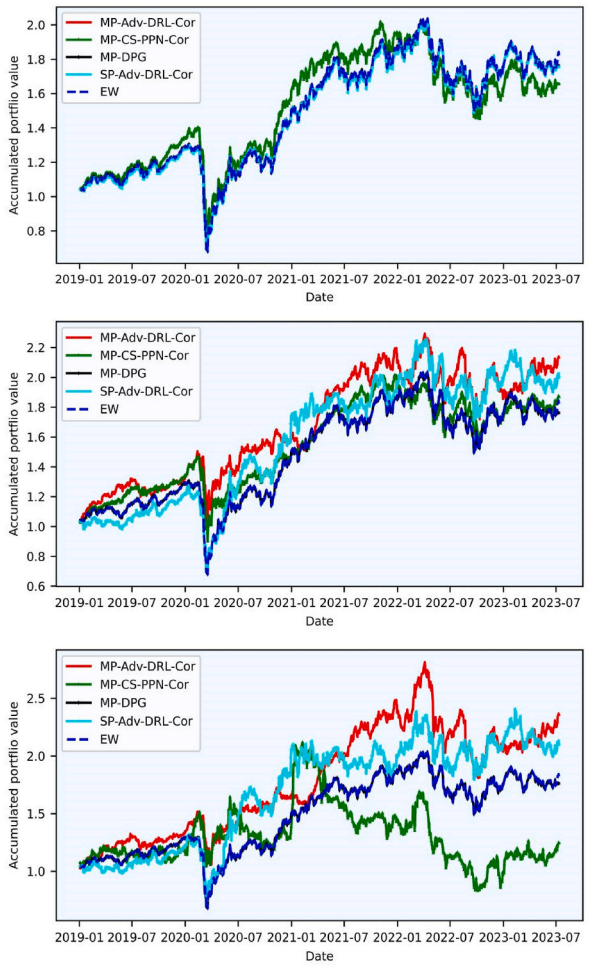
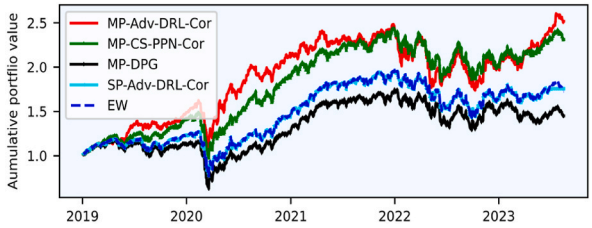


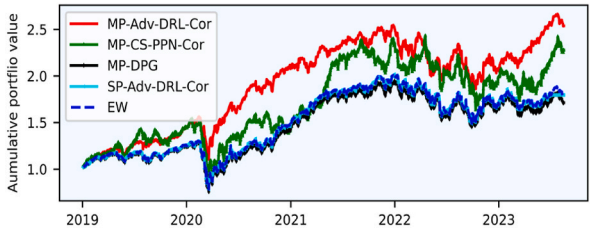
Fig. 8. APV for the five investment portfolios evaluated for S&P/TSX Index data over the period 02/01/2019 to 12/07/2023 for $h = 1$ (top panel), $h = 22$ (middle panel) and $h = 66$ (bottom panel) when $\xi = 0.05\%$ and $\lambda = 0.1$.

Table 5
Portfolio performance metrics for five portfolios when $h = 1$, 22, and 66 under S&P100 index.

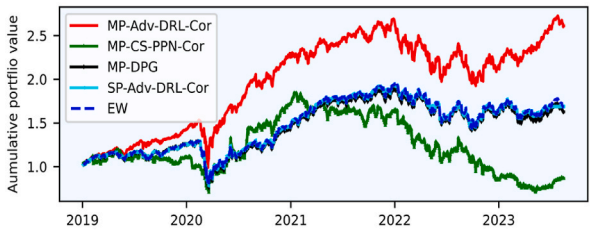
Method	Annual return (%)	Annual volatility (%)	Sharpe ratio	Maximum drawdown (%)	Turnover
$h = 1$					
MP-Adv-DRL-Cor	12.32	24.57	0.502	39.82	0.007
MP-CS-PPN-Cor	10.05	23.75	0.423	36.62	0.006
MP-DPG	12.19	24.69	0.494	40.10	0.011
SP-Adv-DRL-Cor	12.32	24.57	0.502	39.82	0.007
EW	13.03	24.59	0.530	39.80	0.006
$h = 22$					
MP-Adv-DRL-Cor	29.80	31.65	0.942	34.53	0.192
MP-CS-PPN-Cor	19.32	21.74	0.888	28.84	0.005
MP-DPG	2.110	49.05	0.043	57.62	0.730
SP-Adv-DRL-Cor	12.32	24.57	0.502	39.82	0.007
EW	13.03	24.59	0.530	39.80	0.006
$h = 66$					
MP-Adv-DRL-Cor	15.11	35.99	0.420	57.15	0.116
MP-CS-PPN-Cor	21.52	23.33	0.922	32.56	0.006
MP-DPG	−22.20	53.24	−0.417	82.54	0.809
SP-Adv-DRL-Cor	12.32	24.57	0.502	39.82	0.007
EW	13.03	24.59	0.530	39.80	0.006



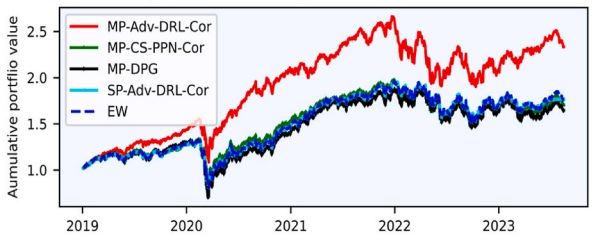
(a) Performance comparisons with 50 assets



(b) Performance comparisons with 75 assets



(c) Performance comparisons with 100 assets



(d) Performance comparisons with 200 assets

Fig. 9. APV across investment portfolios for random subsets of the S&P500 Index over the period 02/01/2019 to 12/07/2023 for different numbers of assets.

much with the number of assets, suggesting that the weight allocated to each asset is quite small. Increasing the number of assets does not contribute massively to improve annual return but helps to mitigate risk through improved diversification of the portfolio. This is achieved for all methods but our MP-Adv-DRL-Cor approach shows superior performance for larger portfolios.

Table 6 confirms these results and provides further insights obtained from alternative performance measures considering risk exposure along with portfolio annual return. Increasing the number of assets in the portfolio rises annual return keeping portfolio variance mostly constant. In particular, the results show that the MP-Adv-DRL-Cor strategy is superior to the MP-CS-PPN-Cor approach, and the performance gap widens as the number of assets increases. Such superiority is mainly attributed to the ability of the WaveNet approach to model asset mutual dependence. In conclusion, the above results confirm the superiority of the two-stream learning frameworks and the importance of applying suitable machine learning models for modeling asset dependencies. The results also confirm the learning ability of our method in addressing high-dimensional portfolio problems.

5. Conclusion

This paper proposes an advanced multi-period portfolio selection method that employs DRL for decision-making, convolutional neural networks to extract the dynamics of asset prices and WaveNet to identify cross-dependencies among the set of investment assets. The proposed approach is capable of solving multi-period investment portfolio problems in high-dimensional settings characterized by an investment pool of many stocks. An extensive empirical application to different datasets, levels of risk aversion and transaction costs shows the good performance of the proposed portfolio allocation procedure for different investment horizons. We find a monotonic

Table 6
Portfolio performance metrics for five investment strategies constructed from different numbers of stocks from the S&P500 index.

Method	Annual return (%)	Annual volatility (%)	Sharpe ratio	Maximum drawdown (%)	Turnover
50 assets					
MP-Adv-DRL-Cor	22.86	23.24	0.984	32.72	0.018
MP-CS-PPN-Cor	17.00	21.22	0.801	31.14	0.005
MP-DPG	12.13	25.62	0.473	41.18	0.092
SP-Adv-DRL-Cor	13.29	25.09	0.530	39.55	0.007
EW	13.93	25.10	0.555	39.54	0.006
75 assets					
MP-Adv-DRL-Cor	24.01	23.04	1.042	29.23	0.045
MP-CS-PPN-Cor	19.29	31.43	0.614	43.53	0.401
MP-DPG	13.67	25.63	0.533	41.32	0.033
SP-Adv-DRL-Cor	13.87	24.85	0.558	39.83	0.007
EW	14.59	24.86	0.587	39.82	0.006
100 assets					
MP-Adv-DRL-Cor	24.38	24.83	0.952	35.54	0.030
MP-CS-PPN-Cor	15.40	23.78	0.680	37.41	0.009
MP-DPG	12.28	25.78	0.477	41.40	0.034
SP-Adv-DRL-Cor	12.32	24.57	0.502	39.82	0.007
EW	13.02	24.59	0.530	39.80	0.006
200 assets					
MP-Adv-DRL-Cor	22.34	23.47	0.952	30.17	0.027
MP-CS-PPN-Cor	13.48	23.18	0.581	37.41	0.013
MP-DPG	12.51	27.28	0.458	46.83	0.080
SP-Adv-DRL-Cor	13.43	23.64	0.568	39.01	0.007
EW	14.03	23.66	0.593	39.01	0.006

relationship between risk-adjusted profitability and the investment horizon for low levels of risk aversion and transaction costs. Increasing levels of risk aversion affect the performance of long-term investment portfolios by reducing incentives to invest over longer horizons. The presence of transaction costs also affects the performance of long-term portfolios by reducing the net annual return above and beyond the reduction observed for short-term portfolios while keeping portfolio volatility roughly constant. Our results also show the outperformance of our proposed procedure against competing methods for constructing optimal portfolios drawn from the machine learning literature in asset allocation. These results are robust to different factors such as the number of assets, risk aversion levels, and the presence of transaction costs.

Auhtor contribution

Yifu Jiang: Conceptualization, Methodology, Software, Empirical Application, Data curation, Write up, Jose Olmo: Literature review, investigation, supervision, reviewing, Majed Atwi: Literature review, investigation, supervision.

Acknowledgements

Jose Olmo acknowledges financial support from Agencia Aragonesa para la Investigación y el Desarrollo (ARAID) and from Grant PID2023-147798NB-I00 funded by Ministerio de Ciencia, Innovación y Universidades.

References

Aboussalah, A. M., Xu, Z., & Lee, C. G. (2021). What is the value of the cross-sectional approach to deep reinforcement learning? *Quantitative Finance, 22*(6), 1091–1111.
Ait-Sahalia, Y., & Brandt, M. (2001). Variable selection for portfolio choice. *The Journal of Finance, 56*, 1297–1351.
Barberis, N. (2000). Investing for the long run when returns are predictable. *The Journal of Finance, 55*, 225–264.
Bellman, R. (1957). A Markovian decision process. *Journal of Mathematics and Mechanics*, 679–684.
Bernardi, M., & Catania, L. (2018). Portfolio optimisation under flexible dynamic dependence modelling. *Journal of Empirical Finance, 48*, 1–18.
Bertsimas, D., & Sim, M. (2004). The price of robustness. *Operations Research, 52*(1), 35–53.
Brandt, M., & Clara, P. S. (2006). Dynamic portfolio selection by augmenting the asset space. *The Journal of Finance, 61*, 2187–2217.
Brennan, M. J., Schwartz, E. S., & Lagnado, R. (1997). Strategic asset allocation. *Journal of Economic Dynamics and Control, 21*(8–9), 1377–1403.
Brennan, M. J., Schwartz, E. S., & Lagnado, R. (1999). The use of treasury bill futures in strategic asset allocation programs. In W. T. Ziemba, & J. Mulvey (Eds.), *World wide asset and liability modeling*. Cambridge University Press.
Campbell, J., Chan, Y., & Viceira, L. (2003). A multivariate model of strategic asset allocation. *Journal of Financial Economics, 67*(1), 41–80.
Campbell, J., & Viceira, L. (1999). Consumption and portfolio decisions when expected returns are time varying. *Quarterly Journal of Economics, 114*(2), 433–495.
Campbell, J., & Viceira, L. (2001). Who should buy long-term bonds? *The American Economic Review, 91*, 99–127.
Campbell, J., & Viceira, L. (2002). *Strategic asset allocation: Portfolio choice for long-term investors*. New York, *NY*: Oxford University Press.
Chen, S., & Ge, L. (2021). A learning-based strategy for portfolio selection. *International Review of Economics & Finance, 71*, 936–942.
Cong, L. W., Tang, K., Wang, J., & Zhang, Y. (2022). AlphaPortfolio: Direct construction through deep reinforcement learning and interpretable AI. *Available at SSRN, 3554486*.
Constantinides, G. M. (1986). Capital market equilibrium with transaction costs. *Journal of Political Economy, 94*(4), 842–862.

Corsaro, S., De Simone, V., Marino, Z., & Scognamiglio, S. (2022). l1-Regularization in portfolio selection with machine learning. *Mathematics, 10*(4), 540.
Cui, T. X., Du, N. J., Yang, X. Y., & Ding, S. S. (2024). Multi-period portfolio optimization using a deep reinforcement learning hyper-heuristic approach. *Technological Forecasting and Social Change, 198*, Article 122944.
Ding, Y., Li, Y., & Zheng, X. (2021). High dimensional minimum variance portfolio estimation under statistical factor models. *Journal of Econometrics, 222*(1), 502–515.
Eom, C., & Park, J. W. (2017). Effects of common factors on stock correlation networks and portfolio diversification. *International Review of Financial Analysis, 49*, 1–11.
Epstein, L., & Zin, S. (1989). Substitution, risk aversion, and the temporal behavior of consumption and asset returns: A theoretical framework. *Econometrica, 57*, 937–969.
Epstein, L., & Zin, S. (1991). Substitution, risk aversion, and the temporal behavior of consumption and asset returns: An empirical investigation. *Journal of Political Economy, 99*, 263–286.
Escobar, M., Ferrando, S., & Rubtsov, A. (2016). Portfolio choice with stochastic interest rates and learning about stock return predictability. *International Review of Economics & Finance, 41*, 347–370.
Fan, Q., Wu, R., Yang, Y., & Zhong, W. (2024). Time-varying minimum variance portfolio. *Journal of Econometrics, 239*(2), Article 105339.
Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., & Chen, T. (2018). Recent advances in convolutional neural networks. *Pattern Recognition, 77*, 354–377.
Hambly, B., Xu, R., & Yang, H. (2023). Recent advances in reinforcement learning in finance. *Mathematical Finance, 33*(3), 437–503.
Jaisson, T. (2022). Deep differentiable reinforcement learning and optimal trading. *Quantitative Finance, 22*(8), 1429–1443.
Jiang, Z. Y., Xu, D. X., & Liang, J. J. (2017). A deep reinforcement learning framework for the financial portfolio management problem. *arXiv preprint arXiv: 1706.10059*.
Kamali, R., Mahmoodi, S., & Jahandideh, M. T. (2019). Optimization of multi-period portfolio model after fitting best distribution. *Finance Research Letters, 30*, 44–50.
Kim, T. S., & Omberg, E. (1996). Dynamic nonmyopic portfolio behavior. *Review of Financial Studies, 9*, 141–161.
Laborda, R., & Olmo, J. (2017). Optimal asset allocation for strategic investors. *International Journal of Forecasting, 33*(4), 970–987.
Liu, H., & Loewenstein, M. (2002). Optimal portfolio selection with transaction costs and finite horizons. *Review of Financial Studies, 15*(3), 805–835.
Lucey, B. M., & Muckley, C. (2011). Robust global stock market interdependencies. *International Review of Financial Analysis, 20*(4), 215–224.
Markowitz, H. M. (1952). Portfolio selection. *The Journal of Finance, 7*(1), 77.
Marzban, S., Delage, E., Li, J. Y. M., Desgagne-Bouchard, J., & Dussault, C. (2023). WaveCorr: Deep reinforcement learning with permutation invariant convolutional policy networks for portfolio management. *Operations Research Letters, 51*(6), 680–686.
Merton, R. (1969). Lifetime portfolio selection under uncertainty: The continuous time case. *The Review of Economics and Statistics, 51*, 247–257.
Merton, R. (1971). Optimum consumption and portfolio rules in a continuous time model. *Journal of Economic Theory, 3*, 373–413.
Moody, J., Wu, L., Liao, Y., & Saffell, M. (1998). Performance functions and reinforcement learning for trading systems and portfolios. *Journal of Forecasting, 17*(5-6), 441–470.
Olschewski, S., Diao, L., & Rieskamp, J. (2021). Reinforcement learning about asset variability and correlation in repeated portfolio decisions. *Journal of Behavioral and Experimental Finance, 32*, Article 100559.
Samuelson, P. (1969). Lifetime portfolio selection by dynamic stochastic programming. *The Review of Economics and Statistics, 51*, 239–246.
Schroder, M., & Skiadas, C. (1999). Optimal consumption and portfolio selection with stochastic differential utility. *Journal of Economic Theory, 21*, 68–126.
Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press, 9780262039246 http://www.scholarpedia.org/article/Reinforcement_learning.
Van Den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., & Kavukcuoglu, K. (2016). WaveNet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499, 12*.
Wang, H., & Zhou, X. Y. (2020). Continuous time mean-variance portfolio selection: A reinforcement learning framework. *Mathematical Finance, 30*(4), 1273–1308.
Watcher, J. (2002). Portfolio and consumption decisions under mean-reverting returns: An exact solution for complete markets. *Journal of Financial and Quantitative Analysis, 37*, 63–91.
Wei, J., Yang, Y. X., Jiang, M., & Liu, J. G. (2021). Dynamic multi-period sparse portfolio selection model with asymmetric investors' sentiments. *Expert Systems with Applications, 177*, Article 114945.
Xu, K., Zhang, Y. F., Ye, D. H., Zhao, P. L., & Tan, M. K. (2020). Relation-aware transformer for portfolio policy learning. *International Joint Conference on Artificial Intelligence*. https://doi.org/10.24963/ijcai.2020/641
Zhang, Y., Zhao, P., Wu, Q., Li, B., Huang, J., & Tan, M. (2022). Cost-sensitive portfolio selection via deep reinforcement learning. *IEEE Transactions on Knowledge and Data Engineering, 34*(1), 236–248.
Zhao, T., Ma, X., Li, X., & Zhang, C. (2023). Asset correlation based deep reinforcement learning for the portfolio selection. *Expert Systems with Applications, 221*, Article 119707.