MASTER'S THESIS 2024-2025 – WORKING PAPER 2

Reinforcement learning for dynamic assets allocation to fulfill futures Liabilities: T3D algorithm with Digital Portfolio Theory framework

Etienne, LARCHET
Finance, Risk & Compliance
2001775

Supervisor: Bushra GHUFRAN

Introduction

Asset analysis targets the ideal portfolio allocation by minimizing risk while maximizing returns. One of the eminent figures in portfolio strategy is Harry Markowitz, whose 1952 paper introduced Modern Portfolio Theory (MPT), widely recognized as the foundation of portfolio optimization. MPT provides a framework to construct investment portfolios that maximize expected returns for a given level of risk. It emphasizes diversification and reduces risk by combining assets with small correlations. It leads to an efficient frontier of optimal portfolios, that is drawn using Sharpe ratio.

For banks and especially life insurance companies, effective portfolio optimization is crucial, as it must balance both asset returns, and future liabilities given by the client contracts.

To address this, the Asset Liability Management (ALM) department is primary focused on optimizing asset allocation while managing liabilities. As defined by Wekwete et al. (2023: 1), ALM aims to "*[…] derive an optimal investment asset allocation strategy for reducing interest rate risk exposure by considering both current and future liabilities*". [1]

A classic ALM approach is Redington's Immunization model, introduced in 1952 (Redington, 1952, cited by Wekwete et al., 2023 [1], and Shiu, 1990 [2]). This model focuses on matching the duration of assets and liabilities so that changes in interest rates affect both and therefore reduce interest rate risk. However, the traditional ALM framework has notable limitations. One significant drawback is the heavy reliance on human judgment for investment decisions and reallocations [1]. This reliance makes traditional approaches vulnerable to behavioral biases, such as *"[…] confirmation bias, overconfidence, recency bias, availability bias, and many other biases.*" (Syed & Bansal, 2018, Rabbani et al., 2021, Chiu et al., 2022, Bondt et al., 2013, cited by Wekwete et al., 2023 : 2) [1]. These biases can distort decision-making and affect the overall performance of ALM strategies.

One way to balance these issues is to reduce human involvement in the decision-making process. Many researchers have demonstrated that the use of machine learning offers superior results in complex decision environments like ALM (Wekwete et al., 2023, Jang & Seong, 2023, Fontoura et al., 2019, Lim et al., 2022, Ruyu et als., 2024). [1] [3] [4] [5] [6]

Machine learning (ML), a branch of artificial intelligence, focuses on developing systems that can learn from data and improve their performance over time without explicit programming. To our knowledge, three types of learning in ML coexist: supervised learning, unsupervised learning, and reinforcement learning.

This paper will not elaborate on the first two, as they are not well suited for portfolio optimization due to their lack of learning abilities once the model has been trained. Reinforcement learning, on the other hand, is highly suited for ALM and portfolio optimization tasks. RL is a type of machine learning where an agent interacts with its environment and learns by receiving feedback based on its actions. The learning process is iterative, where the agent refines its strategy (that is called policy) to maximize the cumulative reward over time. Several strategies have been implemented, the main one being the Bellman equation that handles the times series and discount factor. This equation is detailed in Ashin and Tikhon 2022 paper [7]. The underlying structure that governs RL is the Markov Decision Process (MDP), which formalizes the decision-making framework.

The MDP is defined as a "*class of stochastic sequential decision processes in which the cost and transition functions depend only on the current state of the system and the current action"* (Puterman, 1990: 1) [8].

A sequential decision process being a model where the decision maker (called agent) observes the state (S) of the system and performs an action (A) from a set of available actions. Puterman explains the consequences of choosing an action as *"[...] twofold; the decision maker receives an immediate reward, and specifies a probability distribution on the subsequent system state. If the probability distribution is degenerate, the problem is deterministic. The decision maker's objective is to choose a sequence of actions called a policy, that will optimize the performance of the system over the decision making horizon. Since the action selected at present affects the future evolution of the system, the decision maker cannot choose his action without taking into account future consequences.*" (Puterman, 1990: 1) [8].

In other words, the agent's goal is to learn the optimal policy that maximizes the expected cumulative reward over time. In financial markets, this translates into continuously adjusting a portfolio to maximize returns while minimizing risk, all in response to a changing market environment.

In the context of ALM, Deep Deterministic Policy Gradient (DDPG), a variant of reinforcement learning, has been widely used within the Modern Portfolio Theory framework to optimize asset allocations. DDPG is an actor-critic algorithm specifically designed to handle continuous action spaces, which makes it ideal for tasks such as portfolio rebalancing. The algorithm learns a deterministic policy (the actor) and evaluates it using a value function (the critic), improving its decisions over time based on feedback from the environment. This type of deterministic algorithm is privileged over machine learning stochastic policy gradient as they require less computing power in high dimensional environment (Silver et al., 2014) [9].

Gap in literature

Despite the advances brought by DDPG in ALM and portfolio optimization, existing research has primarily focused on its application within the MPT framework. However, the gap in literature lies in the fact that MPT is inherently myopic and not suited for long-horizon risk management, particularly when incorporating multi-period objectives or mean-reverting risks. This has been studied by C. Kenneth Jones that compared the MPT with Intertemporal Portfolio Theory (IPT) and Digital Portfolio Theory (DPT) (Kenneth Jones, 2017) [10]. The paper put in lights DPT as a hybrid framework, that gives a single-period non-myotic solution and estimates mean-reversion risk levels.

Current research has yet to explore the potential of advanced reinforcement learning algorithms, such as Twin-Delayed Deep Deterministic Policy Gradient (TD3) that is an extension of DDPG that addresses some of its shortcomings and introduces improvements like using two critics to mitigate overestimation bias. A consensus of researchers considers TD3 as an extend and improvement of *"the DDPG algorithm by introducing new features to make it increasingly stable during training and to improve convergence speeds"*(Jiang et al., 2024: 4) [11]. Its usage in finance is still not widely implemented, with few papers using it, however as of my knowledge, no one has used it with modern framework like Digital Portfolio Theory which is designed for long-term portfolio optimization.

This thesis addresses this gap by investigating how TD3 can be integrated into ALM within the context of DPT, providing a more dynamic and flexible approach to asset allocation. Unlike MPT, DPT accommodates long-horizon risks and mean-reversion, making it a more suitable framework for institutional investors aiming to balance returns and liabilities over time.

Through this exploration, the research aims to answer the following question:

Can the TD3 reinforcement learning algorithm, within the DPT framework, improve asset allocation outcomes in ALM by optimizing portfolio performance and fulfilling its long-term liabilities?

Research Objectives and Approach

Therefore, the purpose of this thesis is twofold. The primary goal is to build and train a model using the TD3 algorithm, which is particularly well-suited for high-dimensional and continuous action spaces, such as those found in financial portfolios. The model will be trained using the DPT framework, which emphasizes the long-horizon, mean-reversion nature of financial returns, as opposed to the short-term, risk-return trade-offs inherent in MPT. The model will follow two guiding principles for the design of the policy and reward-punishment function:

- Short-term goal: maximize returns while simultaneously reducing systemic risk. The model will have to optimize asset allocations dynamically, adjusting based on market signals to capitalize on opportunities while keeping systemic risk under control.
- Long-term goal: address long-term liabilities and progressively allocate assets to less volatile instruments as the liabilities approach maturity. This will require the model to balance short-term portfolio growth with the need to manage long-term obligations and ensure that sufficient assets are available when liabilities are due.

The second purpose of the thesis will be to benchmark the performance of TD3 against DDPG, using either MPT framework or DPT framework. By comparing TD3 and DDPG across these two frameworks. This research will provide insights into the advantages of adopting a long-horizon, mean-reversion approach and the potential for reinforcement learning for asset allocation strategies in ALM.

Data Requirements

To train and evaluate the models, a diverse set of financial and liability-related data will be required. Most will be historical financial market data such as price and volume, as well as technical indicators, including MACD, RSI, MA, standard deviation, Calmar, Sharpe, Treynor, and Sortino ratio, as well as Maximum Draw Down (MDD). In case of missing values in financial data, the gaps are filled using the backfill method.

Future liabilities as well as interest rates projections will be given by HSBC France Life Insurance. The model will also handle assets allocation with fixed transactions costs for buy and sell transactions.

Academic literature

Modern Portfolio Theory

Modern Portfolio Theory (MPT), developed by Harry Markowitz in his works of 1952 and 1959, fundamentally changed the way investors approach portfolio construction. Before MPT, investment decisions were largely based on individual asset characteristics, such as returns and risks, without a structured framework for considering how assets interact within a portfolio. Markowitz introduced a mathematical and systematic approach to investing, emphasizing that asset selection must account for the interrelationships between assets, particularly their covariances, to optimize the risk-return tradeoff.

Markowitz’s mean-variance framework established two fundamental principles:

- For a given level of risk, investors should maximize expected return.
- For a given level of expected return, investors should minimize risk (variance).

As summed by Elton and Gruber, these principles led to the concept of the efficient frontier and the mean-variance, which represent a method to determine optimal portfolios offering the highest expected return for a given level of risk (Elton & Gruber, 1997) [12]. The efficient frontier became the cornerstone of MPT, illustrating that diversification reduces risk not just by holding more assets but by selecting assets with low or negative correlations.

At the heart of MPT lies the recognition that asset returns do not move independently, and therefore in the construction of a portfolio, an investor “*had to consider how each security co-moved with all other securities*” (Elton & Gruber, 1997:2) [12]. This interdependence means that well-diversified portfolios can achieve a lower total risk than the sum of individual risks. By quantifying both expected returns and risks, MPT provides a framework for constructing portfolios that balance an investor’s risk tolerance with their return objectives.

MPT operates under the assumption of normally distributed returns and a single-period investment horizon, simplifying mathematical modeling but limiting its applicability in more complex or multi-period scenarios.

Limitations

While MPT remains a foundational framework, several extensions and modifications have addressed its limitations:

1. **Mean-variance simplification**: MPT assumes that the investor's sole objective is to maximize returns for a given level of risk, measured by variance. The simplification of focusing only on mean and variance in the portfolio optimization problem was criticized by Tobin in 1958 (cited by Elton & Gruber, 1997) [12]. Other factors, such as skewness and kurtosis, are ignored, which led to the emergence of other portfolio theories including such indicators (Lee, 1977, Kraus & Litzenberger, 1976, cited by Elton & Gruber, 1997) [12].

2. **Liabilities**: The theory primarily focuses on asset returns without adequately considering liabilities. In their 1992 paper, Elton and Gruber emphasized the importance of incorporating liabilities into portfolio optimization (Elton & Gruber, 1992, cited in their 1997 paper) [12]. This approach recognizes that for institutional investors, such as pension funds or insurance companies, the asset allocation process must account for future cash flow obligations, considering the uncertainty and systematic risks inherent to these liabilities.

3. **Stationarity of Inputs**: MPT assumes that returns are independent and identically distributed (IID), with no predictive time variation, and focuses solely on the trade-off between risk and return in the immediate future (Kenneth Jones, 2009) [13]. This myopic approach fails to account for long-term risks, such as mean-reversion effects or horizon-dependent risks, making it less suitable for investors with extended time horizons

4. **Quadratic computation**: In the portfolio optimization problem, MPT's reliance on quadratic programming often results in unstable and extreme portfolio weights, which are sensitive to estimation errors in means and covariances. The quadratic computation is inherent to the MPT, that relies on a covariance matrix for the mean-variation calculation. Furthermore, optimizers of the mean-variance problem are estimation error amplifiers (Michaud, 1989, cited by Kenneth Jones, 2009) [13] .

5. **Single period problem:** The estimation of mean return and mean variance for each asset is over a single period. To solve the multi period problem, several researchers processed with a sequence of single period problems (Fama, 1970, Hakansson, 1970 and 1974, cited by Elton & Gruber, 1997) [12]. However, this method follows the assumption that returns are independent between periods, which were proven later to be dependents (Fama and French, 1989, Cambell and Shiller, 1988, cited by Elton & Gruber, 1997) (Kenneth Jones, 2017) [12] [10].

Digital Portfolio Theory

To address these gaps in the MDP, C. Kenneth Jones introduced a theorical portfolio optimization framework named Digital Portfolio Theory (DPT) (Kenneth Jones, 2001) [14]. This framework extends the foundational concepts of Modern Portfolio Theory (MPT) by incorporating advanced techniques from digital signal processing to address the limitations of MPT, particularly for long-term investment horizons. While MPT optimizes portfolios by balancing expected returns and risk in a single period, DPT addresses the complexities of long-term investment by incorporating mean-reversion risks, holding periods, and the dynamic nature of financial markets.

The Innovation of Digital Portfolio Theory

DPT incorporates insights from digital signal processing, a field traditionally associated with engineering and communications. At its core, DPT decomposes portfolio variance into systematic and unsystematic components while also accounting for periodic risks, such as calendar anomalies and mean-reversion effects. Anomalies returns on months (Linn & Lockwood, 11988, Hensel & Ziemba, 1996, Penman, 1987), seasons (Wachtel, 1942) and presidential elections periods (Booth & Booth,1999), have been cited by Kenneth Jones in his 2001 [14] and 2009 [13] papers on the DPT. Mean reversion refers to the tendency of asset returns to revert to their long-term average over time (Poterba & Summers, 1988, Fama & French, 1998, cited by Kenneth Jones, 2001) [14]. This decomposition enables investors to manage risks across different time horizons, aligning portfolio decisions with long-term goals and expectations.

A fundamental technical innovation in DPT is its use of the Fourier transform, which allows for the decomposition of financial time series into frequency components. By treating asset returns as signals, the Fourier transform translates these time-domain data into the frequency domain, enabling the identification of periodic risk components such as short-term volatility and long-term mean-reversion trends. Low frequencies represent stable, long-term risks, while high frequencies capture short-term fluctuations. This granular view of risk allows investors to manage portfolio exposures to specific frequencies, tailoring their portfolios to align with anticipated cycles, such as quarterly earnings effects or multi-year mean-reversion patterns.

DPT's integration of Fourier analysis facilitates a refined decomposition of risk into systematic and unsystematic components. Systematic risk, which reflects market-wide influences, and unsystematic risk, which pertains to individual securities or sectors, are analyzed in the context of multiple time horizons. By using autocovariance data derived from the Fourier-transformed signals, DPT enables precise risk tuning. Investors can explicitly control the proportion of systematic and unsystematic risks they wish to bear, ensuring that their portfolios align with their risk tolerance and investment objectives. This decomposition also helps achieve efficient diversification by mitigating unsystematic risk without over-diversifying into excessively large portfolios.

Another notable advancement of DPT is its ability to impose constraints on portfolio size, addressing practical challenges associated with over-diversification and the management of large portfolios. Unlike MPT, which often leads to portfolios with many negligible allocations, DPT employs mixed integer programming (MIP) to set a specific limit on the number of assets in the portfolio. This constraint is implemented using a zero-one integer variable. By allowing investors to specify their preferred portfolio size, DPT accommodates diverse investment styles, whether emphasizing concentrated active management or broader passive diversification. This is crucial because no consensus has been found by research on the ideal portfolio number assets to achieve optimal diversification (Kenneth Jones, 2009) [13], but also because the DPT can be used by individuals investors, which compose small portfolios, and by institution investors that diversify in hundreds of different assets.

Another critical technical feature of DPT is its prevention of extreme portfolio weights, a common drawback of MPT (Green & Hollifield, 1992, cited by Kenneth Jones, 2017) [10]. MPT's quadratic optimization approach amplifies estimation errors in expected returns and covariances, often resulting in highly unstable and concentrated allocations. DPT addresses this by utilizing linear programming, which avoids quadratic terms and is less sensitive to input errors. Furthermore, DPT incorporates shrinkage techniques to stabilize estimates of means and covariances, pulling them toward more robust, generalized values. These enhancements reduce the likelihood of extreme allocations and ensure that portfolio weights remain stable and realistic. Additional constraints, such as upper and lower bounds on individual asset weights, allow investors to further control their allocations, preventing any single security from dominating the portfolio or being allocated an insignificant share.

Applications in Portfolio Management

DPT is particularly well-suited for institutional investors, such as pension funds, that manage long-term liabilities. By incorporating mean-reversion effects and horizon-dependent risks, DPT helps these investors construct portfolios that align with their extended holding periods and risk tolerances. Furthermore, DPT's ability to manage periodic risks makes it an effective tool for tactical adjustments, such as positioning portfolios to benefit from anticipated seasonal trends or macroeconomic cycles.

Limits of the DPT

Digital Portfolio Theory (DPT), while innovative, has several limitations that may hinder its practical application. One significant challenge is its reliance on large datasets. Empirical tests of DPT often require extensive historical data, requiring 16 years or more data prices to accurately capture long-term mean-reversion patterns and periodic risks. This requirement can be restrictive, particularly for emerging markets or new asset classes with limited historical records.

Also, since its introduction in 2001, there appears to be no significant evidence of its adoption in practical financial applications or academic benchmarks. Unlike the MPT, which remains the dominant framework for portfolio optimization, DPT has yet to be tested extensively in real-world scenarios or compared rigorously against the established methodologies. As a result, while DPT offers intriguing potential, further empirical testing and validation are essential to evaluate its practical effectiveness and determine whether it can truly outperform or complement MPT in portfolio optimization.

A Comparison with MPT

When compared to MPT and intertemporal portfolio choice models, DPT emerges as a superior framework for long-term portfolio optimization. While MPT is limited to single-period decisions and intertemporal models are often computationally complex and reliant on restrictive assumptions, DPT strikes a balance between practicality and sophistication. Its linear programming approach allows for efficient optimization, while its integration of mean-reversion risks ensures relevance for long-term investment strategies.

	Digital Portfolio Theory	**Modern Portfolio Theory**
Time Horizon	Long-term focus; incorporates mean-reversion and periodic risks.	Short-term focus; assumes a single-period optimization.
Risk	Separates risk into systematic, unsystematic, and periodic components.	Low control on systematic and unsystematic risks
Data Requirements	Requires extensive historical data (16+ years) to analyze long-term trends and autocorrelations.	Requires shorter historical datasets to estimate means and covariances for a single period.
Portfolio Control	Control on the number and weights of assets	No direct control over portfolio size and assets weights requirements, with tendencies to extreme weights
Mean-Reversion Handling	Explicitly incorporates mean-reversion risk, allowing for hedging and speculative strategies across different time horizons.	Assumes IID returns, ignoring mean-reversion and long-term trends.
Periodic Risk Management	Integrates periodic risk (e.g., calendar anomalies) using frequency-domain analysis.	Does not account for periodic risks or time-dependent autocorrelation.
Computational Efficiency	More scalable for large universes due to linear programming, even with constraints like portfolio size.	Computationally expensive for large universes due to the inversion of large covariance matrices and quadratic programming.
Practical Applications	Not widely adopted; requires further empirical testing to validate its practical effectiveness.	Well-established and widely used by both researchers and companies.

Reinforcement learning

In recent years, many researchers have proven the efficiently of machine learning in portfolio weights optimization, most of them using the mean-variance theory of the MPT as base to the reward function. This section will recompose the advancement over time of empirical studies on the resolution of the portfolio optimization problem with the leverage of machine learnings algorithms.

Machine learning with the MPT

In 2022, Pinelis and Ruppert propose a machine learning algorithm that implements the MPT (Pinelis & Rubbert, 2022) [15]. They created a utility-maximizing framework that optimizes portfolio weights between a market index and a risk-free asset using two Random Forest models to capture expected returns and volatility. The first Random Forest model predicts the expected monthly excess returns using macroeconomic and financial variables such as payout yields, term spreads, and inflation rates, while the second estimates prevailing volatility using lagged realized volatility and similar predictors. Empirical results validate the efficacy of this implementation, demonstrating a 28% improvement in Sharpe ratios over traditional buy-and-hold strategies, with a heavy annualized alpha of 3.4%.

Type of reallocation strategy

The same year, in 2022, some researchers begin the implementation of reinforcement learning to solve the portfolio optimization problem (Lim et al, 2022) [5]. The focus is on the assets allocation strategies with the RL algorithm testing strategic and tactical asset allocation (SAA and TAA). As explained by Bouyé, SAA is a medium-to-long term strategy that bonds each asset with minimal and maximal allocation, while TAA is a short term strategy that shifts the assets weights depending on the market conditions (Bouyé, 2018) [16]. The challenges of investors, being the conciliation of the two strategies (Aglietta et al., 2007, cited by Bouyé, 2018) [16].

The selected reinforcement learning is a Q-learning type of algorithm, which unlike most similar studies is a value-based algorithm that uses neurons layers (in opposition with policy based and actor critic algorithms). The research uses inputs derived from technical indicators, including EMAs and MACD values. The implementation demonstrated that a gradual rebalancing method incorporating Long Short-Term Memory (LSTM) models for price prediction yielded the best results, improving NAV returns by 27.9% to 93.4% over a full rebalancing strategy without predictions. The explanation was the gradual adjustments reduced transaction costs and penalties associated with abrupt changes, leading to superior risk management and enhanced overall returns compared to individual assets within the portfolios.

DRL with the MPT using Tucker decomposition

Then, in 2023, Jang and Seong successfully implemented the MPT in a Deep Reinforcement Learning (DRL) algorithm with risk-adjusted returns and diversification principles (Jang & Seong, 2023) [3]. For the first time, a reinforcement algorithm utilizes the Tucker decomposition to bridge the gap between historical price correlations and technical indicators, such as moving averages, RSI, and MACD. This multimodal approach processes high-dimensional tensors (multi-dimensional vectors) representing the covariance of returns and technical features. These tensors are then subjected to 3D convolutional neural networks for feature extraction and dimensionality reduction, followed by a deep deterministic policy gradient (DDPG) reinforcement learning framework. The reward function was designed to maximize the portfolio's terminal value while accounting for transaction costs and risk, reflected through metrics like the Sharpe ratio and maximum drawdown.

Deep Reinforcement Learning

More recently, another study was conducted by Yan et al. in 2024, that also integrates the MPT for portfolio optimization. The article especially fills the gaps in the literature about the lack of consideration for transaction costs and risk volatility in existing RL-based portfolio optimization models. The proposed framework uses the Deterministic Policy Gradient (DPG) algorithm, which is well-suited for continuous decision-making and robust against the inherent nonstationarity of financial markets. Inputs to the model include normalized price vectors of assets, comprising historical data for opening, closing, highest, and lowest prices across multiple timeframes. Additionally, convolutional neural networks (CNNs) and long short-term memory networks (LSTMs) are employed to extract respectively asset correlations and temporal patterns. Unlike the precedent study, authors choose not to provide any technical indicators, to give more control for the CNNs and LSTM algorithms with raw prices data. The MPT was implemented in the reward function with a risk-cost reward function to optimize the trade-off between portfolio returns and risk. The model incorporates transaction costs by modeling them using the one-norm of portfolio transaction vectors. Specifically, the function is designed to penalize frequent trading and estimation errors while rewarding stable accumulative portfolio returns.

Implementation of the TD3 algorithm

Jiang et al. achieve another breakthrough in the portfolio optimization problem by utilizing the Twin Delayed Deep Deterministic Policy Gradient (TD3) algorithm, providing a comparison with the well-established DDPG approach (Jiang et al., 2024) [11]. This algorithm is employed to address the complexities of portfolio optimization in high-dimensional and dynamic financial markets by improving upon the DDPG with the introduction of three key enhancements: target policy smoothing, double Q-learning, and delayed policy updates. These updates reduce overestimation bias and improve training stability (Fujimoto et al, 2018) [17]. The TD3 framework utilizes actor-critic architecture, where the actor network outputs deterministic actions, and two critic networks estimate the Q-values (expected cumulative reward function) to avoid overestimation. Innovations include target policy smoothing, which applies clipped Gaussian noise to the action output to improve generalization, and a delayed update mechanism, where the actor is updated less frequently than the critics to enhance convergence stability.

The method also models the portfolio optimization problem as MDP, with states defined by historical asset prices and portfolio weights, actions representing asset reallocation decisions and rewards incorporating mean-variance framework, as well as transaction costs, and risk aversion. The proposed framework successfully balances risk and return, achieving reduced maximum drawdowns and higher cumulative returns compared to other deep reinforcement learning methods.

Multiagent DRL

Lastly, Cheng and Sun proposed an advance algorithm that also relies on TD3 algorithm, but implementing a multiagent structure (Cheng & Sun, 2024) [18]. They underscore many gaps in the literature, including the inability of single-agent frameworks to handle chaotic, multi-asset environments and the tendency of traditional models to overfit or produce unstable returns. The framework allows multiple agents to independently explore distinct assets while sharing learned parameters with a global network. Each agent focuses on specific stock features, with two primary modules: the Trading Action Module (TAM) and the Trading Portfolio Module (TPM). The TAM integrates CNNs and LSTMs to analyze candlestick patterns (open, close, high low of daily prices) generating asset-specific trading actions (long, short, hold). TPM evaluates these actions, assigns asset scores based on technical indicators, and determines portfolio weights dynamically. This modular design avoids the pitfalls of single-agent models by enabling robust exploration of diverse market conditions. The reward function uses the Sortino ratio, which prioritizes downside risk management by focusing on negative deviations from expected returns, thereby enhancing stability in volatile markets. The multi-agent frame introduces asynchronous training via Asynchronous Advantage Actor-Critic (A3C), wherein each agent optimizes its environment-specific policy before aggregating results globally. This decentralized approach accelerates training and reduces inter-agent correlation issues.

Application in Assets Liabilities Management

This last section synthesizes insights from two seminal works on Wekwete et al. (2023) [1] and Fontoura et al. (2019) [4], which apply DRL frameworks to revolutionize ALM practices. As of our knowledge, these 2 studies constitute the overall work done on the subject, highlighting the important gap in literature, and hindrance compared to the explored dynamic assets allocation models.

Traditional ALM approaches, particularly duration matching through Redington Immunization (Shiu, 1990) [2], aim to align the timing of asset cash flows with liability outflows. However, as explained by Wekwete, this method requires frequent rebalancing, which is both labor-intensive and prone to human biases, such as overconfidence and recency effects (Wekwete et al., 2023). Therefore, the need for automated, adaptive, and robust solutions, is paving the way for DRL-based approaches.

In Wekwete et al. (2023) [1], the state space includes liability durations and asset maturities derived from Monte Carlo simulations, and actions represent asset allocation decisions, such as the weights assigned to short- and long-term bonds. The reward function is designed to minimize the absolute mismatch between asset and liability durations, thus ensuring effective duration matching.

Similarly, Fontoura et al. (2019) leverage the DDPG algorithm. Unlike Wekwete et al. approach that discretize state spaces into scenario trees, DDPG enables the use of continuous state and action spaces, which better reflect the stochastic and dynamic nature of ALM. The algorithm comprises an actor-critic architecture: the actor maps the actions (buy, short, hold asset allocations), while the critic evaluates the quality of these actions based on Q-values. By incorporating a discount factor in the reward function, the model emphasizes immediate gains while maintaining a forward-looking perspective on liability management.

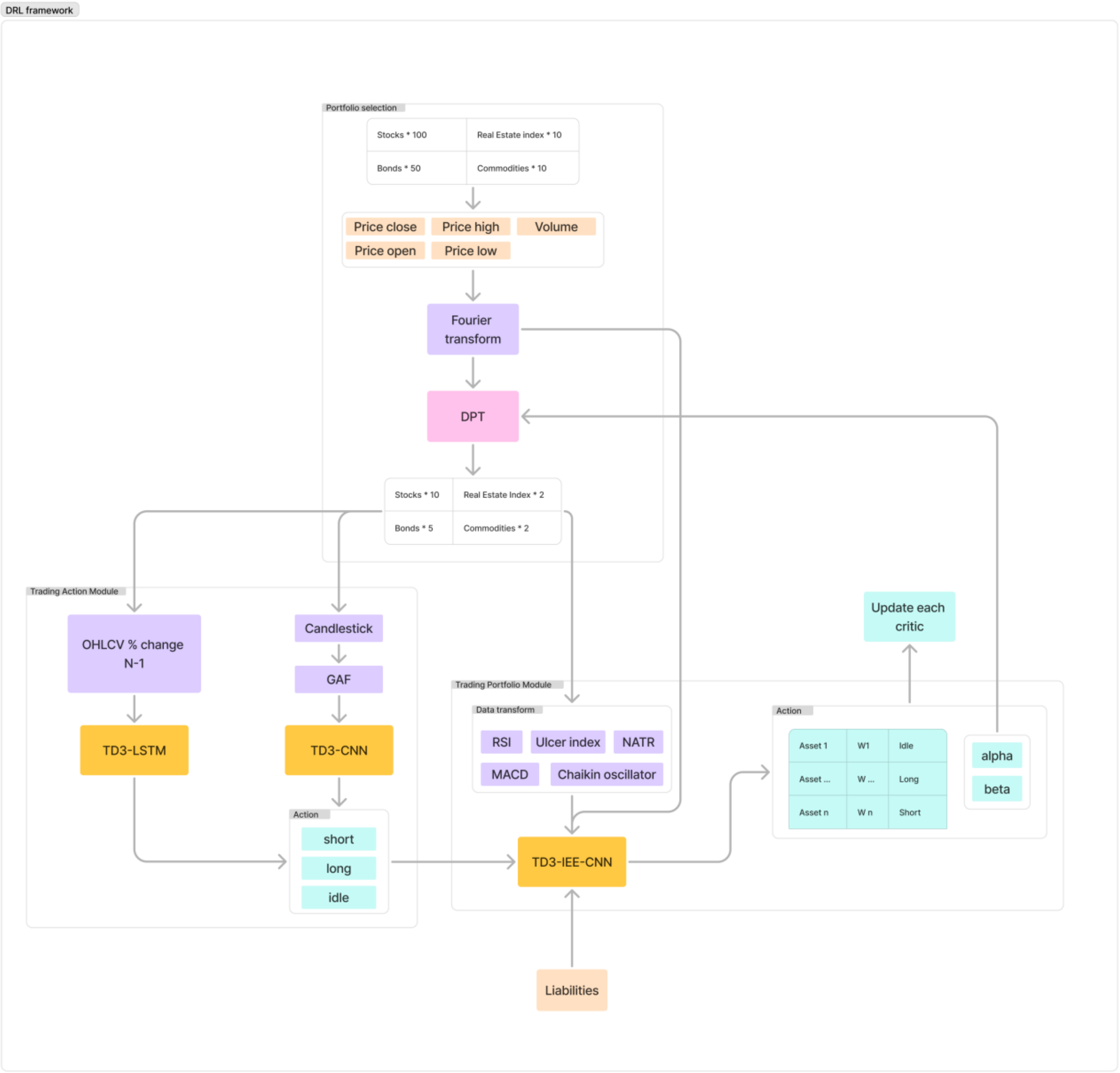
Academic conclusion and model introduction

This thesis lays the groundwork for a novel approach to Asset–Liability Management by integrating Digital Portfolio Theory with a multiagent Deep Reinforcement Learning framework. While Modern Portfolio Theory has traditionally shaped portfolio optimization, it struggles with long-term and mean-reversion dynamics that are critical for institutional investors. DPT addresses this gap by incorporating Fourier transforms, frequency-domain analysis, and advanced constraints, making it more suitable for horizon-dependent risks and large-scale asset universes.

On the machine learning side, Deep Reinforcement Learning has demonstrated robust performance in dynamic asset allocation, although existing studies predominantly employ MPT-based reward functions. Moreover, single-agent RL models may struggle with the complexity of multi-asset environments. As such, the multiagent extension of RL, supported by techniques like Twin Delayed Deep Deterministic Policy Gradient offers a framework that can better handle multiple asset dynamics, reduce overfitting, and address scalability challenges.

Against this backdrop, the goal of our research is to develop a multiagent DRL model that leverages the DPT framework to adaptively manage asset weights and positions. By doing so, the system aims to maximize returns while considering long-term liabilities, thereby ensuring that portfolios remain well-positioned to meet future obligations.

Model description (in progress)



Results (in progress)

Conclusion (in progress)

References

[1] T. A. Wekwete, R. Kufakunesu and G. van Zyl, "Application of deep reinforcement learning in asset liabiltiy management," *Intelligent Systems with Applications 20,* 2023.

[2] E. S. Shiu, "On Redington's theory of immunization," in *Insurance: Mathematics and Economics*, vol. 9, 1990, pp. 171-175.

[3] J. Jang and N. Seong, "Deep reinforcement learning for stock portfolio optimization by connection with modern portfolio theory," *Expert Systems with Applications,* vol. 218, 2023.

[4] A. Fontoura, E. Bezerra and D. Haddad, "A Deep Reinforcement Learning Approach to Asset-Liabilities Management," 2019.

[5] Q. Y. E. Lim, Q. Cao and C. Quek, "Dynamic portfolio rebalancing though reinforcement learning," *Neural Computing and Applications,* vol. 34, pp. 7125-7135, 2022.

[6] Y. Ruyu, J. Jiafei and H. Kun, "Reinforcement learning for deep portfolio optimization," *Electronic Research Archive,* 2024.

[7] R. Ashwin and J. Tikhon, Foundations of Reinforcement Learning with Applications in Finance, 2022.

[8] M. L. Puterman, "Markov Decision Processes," in *Handbooks in Operations Research and Management Science*, vol. 2, 1990, pp. 331-434.

[9] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra and M. Riedmiller, "Deterministic Policy Gradient Algorithms," *Proceedings of Machine Learning Research,* vol. 32, pp. 387-395, 2014.

[10] C. Kenneth Jones, "Modern Portfolio Theory, Digital Portfolio Theory and Intertemporal Portfolio Choice," *American Journal of Industrial and Business Management,* vol. 7, pp. 833-854, 2017.

[11] Y. Jiang, J. Olmo and M. Atwi, "Deep reinforcement learning for portfolio selection," *Global Finance Journal,* vol. 62, 2024.

[12] E. J. Elton and M. J. Gruber, "Modern portfolio theory, 1950 to date," *Journal of BANKING & FINANCE,* pp. 1743-1759, 1997.

[13] C. Kenneth Jones, "Digital Portfolio Theory: Portfolio Size versus Alpha, Beta, and Horizon Risk," 2009.

[14] C. Kenneth Jones, "Digital Portfolio Theory," *Computational Economics,* vol. 18, pp. 287-316, 2001.

[15] M. Pinelis and D. Ruppert, "Machine learning portfolio allocation," *The Journal of Finance and Data Science,* vol. 8, pp. 35-54, 2022.

[16] É. BOUYÉ, "Allocation stratégique des actifs et gestion de l'investissement à long terme par les investisseurs institutionnels," *Risque et économie,* vol. 69, pp. 505-531, 2018.

[17] S. Fujimoto, H. Hoof and D. Meger, "Addressing Function Approximation Error in Actor-Critic Methods," *Proceedings of Machine Learning Research,* vol. 80, pp. 1587-1596, 2018.

[18] L.-C. Cheng and J.-S. Sun, "Multiagent-based deep reinforcement learning framework for multi-asset adaptive trading and portfolio management," *Neurocomputing,* vol. 594, 2024.