A Review of Reinforcement Learning in Financial Transaction Optimization

Abstract

Reinforcement learning (RL) is transforming financial trading by enabling dynamic decision-making in areas like high-frequency trading, investment portfolio optimization, and option pricing. Algorithms such as Deep Q-Learning and PPO optimize trade execution, while multi-agent RL enhances asset allocation strategies, balancing risk and reward. In risk management, RL methods like A2C improve pricing models and hedging strategies. However, challenges like data scarcity, non-stationary markets, and computational complexity limit broader adoption. Future research should explore interdisciplinary approaches, including game theory, federated learning, and synthetic data generation, while developing adaptive RL models tailored to financial applications. RL's potential to revolutionize trading continues to grow as computational and algorithmic advancements progress.

Keywords

reinforcement learning (RL), finance, deep learning, financial trading optimization

1. Introduction

The financial industry operates in a highly dynamic and complex environment, where effective decision-making is critical (Greenberg et al. 2019)[1]. Traditional methods often struggle to adapt to the real-time, nonlinear, and volatile nature of financial markets. Reinforcement learning (RL), a subset of machine learning, has emerged as a powerful tool capable of addressing these challenges by making sequential decisions and optimizing strategies based on feedback from the environment. RL's ability to learn from interaction and adapt to changing conditions makes it particularly suitable for financial applications, where dynamic optimization is essential.

This review focuses on the application of Reinforcement Learning in financial trading optimization, a field where RL techniques have shown great promise. Key areas of exploration include high-frequency trading (Cao et al. 2024)[2], investment portfolio optimization (Hu et al. 2019)[3], and option pricing (Stoiljkovic et al. 2023)[4]. In high-frequency trading, RL algorithms optimize trade execution by adapting to rapidly fluctuating markets, enabling traders to maximize returns and minimize risks. Investment portfolio optimization leverages RL's ability to dynamically allocate assets, balancing risk and reward in response to evolving market conditions. Meanwhile, RL-based methods in option pricing and risk management offer innovative approaches to nonlinear pricing models and adaptive hedging strategies.

Despite its potential, the application of RL in finance faces several challenges. These

include the scarcity of high-quality, labeled financial data (Liu et al. 2022)[5], the non-stationary nature of markets (Marinescu et al. 2017)[6], and the computational complexity (Mishra et al. 2022)[7] of training and deploying RL models. Addressing these issues requires interdisciplinary approaches, such as integrating game theory (Yang et al. 2020)[8], leveraging federated learning for collaborative model training (Yu et al. 2020)[9], and generating synthetic data using advanced techniques like Generative Adversarial Networks (GANs) (Liu et al. 2022)[5].

This review aims to provide a comprehensive overview of RL's role in financial trading optimization, summarizing current advancements, identifying key challenges, and exploring future directions. By highlighting both the opportunities and limitations, this review seeks to contribute to the growing body of knowledge and encourage further research in this transformative field.

2. Fundamentals of Reinforcement Learning

2.1 Key Concepts

Reinforcement learning (RL)[10] is a machine learning paradigm where agents learn to make decisions by interacting with an environment to maximize cumulative rewards. Unlike supervised learning, which relies on labeled data, RL focuses on learning from feedback in the form of rewards or penalties(*Araújo* et al. 1998)[11]. Several key concepts underpin RL, providing the foundation for its application in financial trading optimization:

Agent and Environment. The agent represents the decision-maker, such as a trading algorithm, while the environment embodies the system it interacts with, such as financial markets. The agent observes the environment's state, performs actions, and receives feedback in the form of rewards.

State, Action, Reward and Policy. State (S) captures the relevant information about the environment at a given time, such as stock prices, market indicators, or portfolio compositions. Actions (A) are decisions made by the agent, such as buying, selling, or holding an asset. Reward (R) quantifies the outcome of an action, guiding the agent's learning process. For instance, profit from a trade can serve as a reward signal. Policy (π) defines the agent's strategy, mapping states to actions. Policies can be deterministic, where a specific action is chosen for a given state, or stochastic, where actions are selected based on probability distributions.

Value Function (V) and Q-Function (Q). Value Function (V) represents the expected cumulative reward starting from a given state under a specific policy. Q-Function (Q) extends the value function by incorporating the expected reward of taking a specific action in a given state, followed by the policy.

Exploration vs. Exploitation. A fundamental challenge in RL is balancing exploration (trying new actions to discover better strategies) and exploitation (choos-

ing actions based on the current policy to maximize rewards). Financial applications require careful tuning of this trade-off to adapt to rapidly changing market dynamics.

Markov Decision Process (MDP). Financial problems are often modeled as MDPs (Bauerle et al. 2011)[12], characterized by a set of states, actions, transition probabilities, and rewards. MDPs assume the Markov property, where the future state depends only on the current state and action, simplifying the modeling of complex systems like financial markets.

These foundational concepts enable RL agents to learn optimal strategies through iterative interaction with financial environments. They form the basis for advanced RL algorithms, allowing applications like high-frequency trading, portfolio management, and option pricing to thrive in dynamic and uncertain markets.

2.2 Algorithms Overview

Reinforcement learning (RL) algorithms form the backbone of financial trading optimization, enabling agents to learn optimal decision-making strategies in dynamic and uncertain environments. These algorithms can be broadly categorized into value-based, policy-based, and actor-critic methods (Liu et al. 2020)[13], each with unique characteristics and advantages.

Value-based algorithms (Mckenzie et al. 2022)[14] focus on learning the optimal action-value function, $Q(s,a)$, which represents the expected cumulative reward for taking action a in state s. A prominent example is Q-Learning, a model-free RL algorithm that iteratively updates $Q(s,a)$ using the Bellman equation (Lu et al. 2022)[15]. Variants like Deep Q-Networks (DQN) extend Q-Learning by incorporating deep neural networks to approximate the action-value function (Boppiniti & Sai Teja 2021)[16], allowing effective handling of high-dimensional state spaces, such as those in financial trading. DQN has been particularly useful in high-frequency trading (Cao et al. 2024)[2], where agents must make quick decisions in complex environments.

Policy-based algorithms (Wang et al. 2019)[17] directly optimize the policy $\pi(a|s)$, which maps states to actions, by maximizing the expected cumulative reward. These methods are particularly effective in continuous action spaces (Zhong et al. 2019)[18]. Proximal Policy Optimization (PPO) (schulman et al. 2017)[19] and Trust Region Policy Optimization (TRPO) (Schulman & John 2015)[20] are popular policy-based algorithms that ensure stable learning by limiting abrupt policy updates. PPO has been widely adopted for portfolio optimization (Scaletta & Gioele 2024)[21], enabling agents to allocate assets dynamically while balancing risk and reward.

Actor-critic algorithms (Konda et al. 1999)[22] combine value-based and policy-based methods, utilizing two components: the actor, which updates the policy, and the critic, which evaluates the policy using a value function. Algorithms like Advantage Actor-Critic (A2C) (Yang et al. 2020)[23] and Deep Deterministic Policy

Gradient (DDPG) (Khemlichi et al. 2021)[24] have shown great promise in financial applications. For instance, A2C has been used in option pricing to improve nonlinear pricing models and hedging strategies (Kang et al. 2024)[25], while DDPG is well-suited for continuous trading tasks, such as portfolio rebalancing (Sadriu et al. 2022)[26].

In financial markets, where multiple agents interact, multi-agent RL (MARL) (Karpe et al. 2020)[27] is increasingly relevant. MARL algorithms extend traditional RL methods to cooperative or competitive settings, enabling dynamic modeling of market participants' behavior. These frameworks are particularly useful in simulating realistic trading environments and developing robust trading strategies.

While these algorithms have demonstrated significant potential in financial trading, challenges remain. Non-stationary environments (Marinescu et al. 2017)[6], data scarcity (Liu et al. 2022)[5], and computational complexity demand robust and adaptive algorithmic designs (Mishra et al. 2022)[7]. Future advancements, such as integrating game theory (Yang et al. 2020)[8], federated learning (Yu et al. 2020)[9], and synthetic data generation (Liu et al. 2022)[5], are expected to further enhance the applicability of RL in finance.

2.3 Modeling Financial Trading as an MDP

Markov Decision Processes (MDPs) provide a structured framework for modeling financial trading (Bauerle et al. 2011)[12], capturing the sequential decision-making nature of trading tasks. An MDP is defined by a tuple (S,A,P,R,γ), where S represents states, A denotes actions, P specifies state transition probabilities, R defines the reward function, and γ is the discount factor. Applying MDPs to financial trading involves defining these components in the context of market dynamics and trading objectives.

The state(S) space represents all relevant market information that an RL agent uses to make trading decisions. States can include features such as historical price data, order book depth, technical indicators, and macroeconomic variables. For example, in high-frequency trading, states often consist of time-series data capturing asset prices and trading volumes. In portfolio optimization, states may reflect current portfolio composition and market trends.

Actions(A) in financial trading represent the set of decisions available to the agent, such as buying, selling, or holding an asset. In discrete action spaces, actions are predefined as specific trading operations, whereas in continuous action spaces, they can include fractional trading volumes or asset weight adjustments in a portfolio. Selecting the appropriate action space is critical for aligning the MDP with the underlying trading problem.

The reward(R) function quantifies the agent's performance, guiding its learning process. In trading, rewards are often tied to financial outcomes, such as realized profits, risk-adjusted returns, or transaction costs. For example, a reward can be

defined as the net profit from a trade after accounting for slippage and transaction fees. Designing an effective reward function is essential for achieving the desired trading objectives, such as maximizing returns while minimizing risks.

The state transition probabilities(P) represent the likelihood of moving from one state to another after taking a specific action. Financial markets exhibit inherent randomness and non-stationarity, making exact modeling of P challenging. Instead, RL algorithms often rely on sampling-based methods, using historical data to approximate transitions. Advanced techniques, such as Generative Adversarial Networks (GANs), can also be employed to simulate realistic market dynamics.

The discount factor(γ) determines the agent's preference for short-term versus long-term rewards. In financial trading, a lower γ encourages short-term profitability, suitable for high-frequency trading, while a higher γ promotes long-term strategies, such as portfolio rebalancing.

Modeling financial trading as an MDP involves assumptions like the Markov property, which states that the future state depends only on the current state and action. However, real-world financial markets often exhibit complex dependencies, such as autocorrelations (Challet et al. 2005)[28] and external influences, which violate this assumption. Addressing these challenges requires incorporating advanced feature engineering (Shui et al. 2024)[29], state representation learning (Park et al. 2021)[30], and adaptive modeling techniques (Routledge & Bryan R 1999)[31]. By framing financial trading as an MDP, reinforcement learning algorithms can effectively optimize trading strategies, enabling agents to adapt to complex market environments and achieve superior financial outcomes.

3. Current Research Trends and Developments

3.1 High-Frequency Trading

High-frequency trading (HFT) (Cao et al. 2024)[2] is a domain within financial trading characterized by rapid execution of orders and high turnover rates, often operating on millisecond or microsecond scales. Reinforcement learning (RL) has emerged as a powerful tool in HFT, enabling agents to make optimal decisions in highly dynamic and competitive markets.

HFT is fraught with challenges such as extreme market volatility, sparse rewards, and non-stationary environments (Zhang et al. 2024)[32]. The high dimensionality of order book data, the need for low-latency execution, and the adversarial nature of competing algorithms further complicate the development of effective strategies. These challenges necessitate robust and adaptive RL models capable of handling real-time decision-making and learning.

HFT can be naturally framed as a Markov Decision Process (MDP), where:States represent market features, such as order book depth, price trends, and volume imbal-

ances. Actions include placing, canceling, or modifying orders, with considerations for price and volume. Rewards reflect trading objectives, such as net profit, inventory control, or market impact minimization.

Deep Q-Networks (DQN) (Cao et al. 2024)[2] have been extensively applied to HFT tasks, leveraging neural networks to approximate the action-value function. Modifications such as Double DQN and Prioritized Experience Replay enhance stability and efficiency in learning. These approaches are effective in capturing short-term trading opportunities by evaluating the potential outcomes of discrete actions. Proximal Policy Optimization (PPO) and Advantage Actor-Critic (A2C) have also been applied to HFT (Motard & Pierre 2022)[33], particularly in continuous action spaces. These algorithms enable more nuanced decision-making, such as dynamically adjusting order placement strategies based on real-time market signals.

Given the adversarial nature of HFT, where multiple agents compete, reinforcement learning frameworks incorporating game-theoretic elements have gained traction (Yang et al. 2020)[8]. These approaches model the strategic interactions between trading algorithms, enabling the development of resilient strategies that can adapt to evolving market conditions.

RL-powered HFT systems have demonstrated the ability to optimize execution strategies, reduce market impact, and enhance profitability. By learning patterns from historical and real-time data, RL agents outperform traditional heuristic-based methods, making them invaluable tools in modern financial markets.

Despite its potential, the application of RL in HFT requires addressing latency constraints (Menkveld et al. 2013)[34], ethical concerns (Roncella et al. 2022)[35], and regulatory compliance (Bell et al. 2014)[36], making it a highly specialized and evolving field.

3.2 Investment Portfolio Optimization

Investment portfolio optimization is a core problem in finance (Hu et al. 2019)[3], involving the allocation of assets to maximize returns while minimizing risks. Reinforcement learning (RL) provides a flexible framework for this task, enabling dynamic adjustments to portfolio weights based on evolving market conditions.

Traditional portfolio optimization methods, such as mean-variance analysis, rely on static assumptions about returns and covariance, which may not hold in dynamic markets. RL addresses these limitations by learning adaptive strategies directly from data. However, challenges such as high-dimensional state spaces, sparse rewards, and transaction costs must be carefully managed to achieve practical results.

In this context, the Markov Decision Process (MDP) elements are defined as: States represent market conditions, including historical asset prices, volatility, and macroeconomic indicators. Actions correspond to adjustments in portfolio weights, such as increasing or decreasing investment in specific assets. Rewards reflect financial

objectives, such as risk-adjusted returns (e.g., Sharpe ratio) or net portfolio growth after accounting for transaction costs.

Deep Deterministic Policy Gradient (DDPG) (Khemlichi et al. 2021)[24] and Proximal Policy Optimization (PPO) (Scaletta & Gioele 2024)[21] are widely used for portfolio optimization. These algorithms handle continuous action spaces, allowing precise adjustments to asset weights. RL agents can dynamically rebalance portfolios, respond to market trends, and improve diversification compared to static strategies.

RL has demonstrated superior performance in portfolio optimization by adapting to non-stationary environments and capturing complex relationships between assets. As financial data availability and computational power continue to grow, RL-based approaches are poised to become integral tools in modern investment strategies. However, real-world deployment requires careful consideration of interpretability, risk constraints, and regulatory compliance.

3.3 Risk Management and Option Pricing

Risk management and option pricing are critical aspects of financial decision-making, requiring sophisticated tools to model uncertainties and optimize outcomes. Reinforcement learning (RL) has emerged as a powerful method for addressing these challenges by leveraging data-driven decision-making frameworks.

Risk management involves identifying, assessing, and mitigating financial risks. RL approaches can optimize hedging strategies (Cao et al. 2021)[37] by dynamically adjusting positions based on market conditions. For example, an agent can minimize portfolio risk by learning to balance exposure to volatile assets while considering transaction costs and liquidity constraints (Shin et al. 2019)[38]. Algorithms like Deep Q-Networks (DQN) (Park et al. 2020)[39] and Proximal Policy Optimization (PPO) (Liu & Peng 2023)[40] are commonly applied to dynamically adjust hedging strategies, demonstrating resilience in uncertain and non-stationary markets.

In option pricing, traditional models like Black-Scholes rely on simplifying assumptions, such as constant volatility and log-normal price distributions, which often fail in real-world scenarios. RL can enhance option pricing (Stoiljkovic et al. 2023)[4] by directly learning from market data, capturing complex patterns such as stochastic volatility and jumps in asset prices. Advantage Actor-Critic (A2C) (Yang et al. 2020)[23] and Deep Deterministic Policy Gradient (DDPG) (Khemlichi et al. 2021)[24] are effective for these tasks, enabling agents to price options accurately while simultaneously developing optimal hedging strategies.

RL-powered approaches in risk management and option pricing can adapt to real-time market changes, outperforming static methods. These models enhance precision in pricing exotic options, reduce hedging costs, and improve the robustness of risk management frameworks. By learning directly from data, RL mitigates the reliance on rigid assumptions, providing more flexible and realistic solutions.

The integration of RL into risk management and option pricing faces challenges such as sparse rewards, high-dimensional state spaces, and regulatory constraints. Future advancements in interpretability, hybrid modeling (combining RL with traditional financial models), and computational efficiency will further enhance the adoption of RL in these critical financial domains (Bai et al. 2024)[41].

4. Challenges

4.1 Data Complexity and Scarcity

Data complexity and scarcity are significant challenges in applying reinforcement learning (RL) to financial trading and decision-making. The nature of financial markets, characterized by high-dimensional, noisy, and non-stationary data, creates hurdles for effective RL model training and deployment.

Financial data often involve intricate relationships between variables, such as price movements, trading volumes, and macroeconomic indicators. These relationships are nonlinear and influenced by external factors like geopolitical events or economic policies. Furthermore, high-frequency trading data may include millions of observations within a short timeframe, leading to computational bottlenecks and the need for advanced feature extraction techniques (Seddon et al. 2017)[42]. Capturing meaningful patterns in such complex data requires sophisticated models that can handle temporal dependencies and dynamic behaviors.

Despite the abundance of financial data, high-quality labeled data for specific tasks, such as rare market conditions or exotic asset classes, can be scarce. For example, tail events like market crashes occur infrequently, making it challenging to train RL agents to handle such scenarios effectively. Additionally, access to proprietary datasets, such as detailed order book information, is often restricted, limiting the scope of RL applications for academic and smaller-scale practitioners.

To address the challenges of data complexity and scarcity in financial reinforcement learning (RL), researchers and practitioners use several techniques. Data augmentation methods, such as bootstrapping (Semenoglou et al. 2023)[43], synthetic data generation (Liu et al. 2022)[5], and adversarial learning, enhance the training datasets, with Generative Adversarial Networks (GANs) (Naritomi et al. 2020)[44] being particularly effective in simulating realistic market conditions. Transfer learning (Lanzetta & Vincenzo 2024)[45] helps mitigate data scarcity by applying pretrained RL models from similar tasks to new domains, leveraging existing knowledge. Robust feature engineering (Salamkar & Muneer Ahmed 2023)[46], which combines domain expertise with techniques like embeddings and dimensionality reduction, ensures higher-quality input features for RL models. Additionally, simulation environments provide controlled settings to train RL agents (Mascioli et al. 2024)[47], enabling them to learn complex strategies without solely depending on historical data. These approaches collectively enhance RL model performance and adaptabil-

ity in financial applications.

Addressing data complexity and scarcity is critical for the success of RL in financial applications. Advances in data generation, preprocessing, and augmentation techniques will not only improve model performance but also broaden the applicability of RL across diverse financial domains. However, practitioners must carefully validate models trained on augmented or simulated data to ensure robustness and reliability in real-world trading scenarios.

4.2 Stability and Generalization

Stability and generalization (Boyan et al. 1994)[48] are critical considerations when applying reinforcement learning (RL) to financial tasks, where the unpredictability and non-stationarity of markets pose significant challenges. Stability ensures that the RL model learns consistently without oscillating or diverging during training, while generalization ensures that the model performs effectively across unseen market conditions and various financial instruments.

Financial environments are inherently noisy and adversarial, leading to unstable learning dynamics for RL agents. Overfitting to specific patterns in historical data is common (Tuite et al. 2012)[49], causing poor performance in live trading scenarios. Furthermore, the sensitivity of RL algorithms to hyperparameters (Eimer et al. 2023)[50], reward design, and market shifts exacerbates instability during training.

Markets are non-stationary (Marinescu et al. 2017)[6], meaning that relationships between variables can shift due to factors such as economic changes or policy decisions. RL models trained on historical data may fail to generalize to new conditions (Zhang et al. 2018)[51], leading to suboptimal or even detrimental trading decisions. Additionally, financial datasets often exhibit data imbalance (Akyildirim et al. 2021)[52], with rare but impactful events such as market crashes being underrepresented.

Techniques to enhance stability and generalization in reinforcement learning (RL) include regularization (Wang et al. 2020)[53] and early stopping (Raskutti et al. 2014)[54], such as dropout and weight decay, which prevent overfitting and ensure consistent learning. Ensemble methods, which combine decisions from multiple RL agents, reduce variance and improve robustness in dynamic markets. Robust training techniques, like adversarial training (Bai et al. 2021)[55] and distributional RL (Singh et al. 2021)[56], bolster resilience against noisy or adversarial data. Dynamic adaptation methods, such as meta-reinforcement learning (Nagabandi et al. 2018)[57], allow agents to respond effectively to shifting market conditions. Additionally, validation on diverse datasets, encompassing both synthetic and historical data, ensures that models generalize well across varied market scenarios. Together, these strategies enable RL models to deliver reliable, scalable performance in complex financial environments.

4.3 Algorithmic Complexity and Computational Cost

Algorithmic complexity and computational cost are significant barriers to deploying reinforcement learning (RL) in financial applications. Advanced RL algorithms, especially deep reinforcement learning (DRL) models, involve high-dimensional state and action spaces that demand substantial computational resources for training. Techniques like policy gradient methods or actor-critic frameworks require iterative updates with large-scale data, which can be computationally expensive.

The cost is further amplified in financial contexts where agents need to process continuous streams of market data and adapt in near real-time. This requires specialized hardware, such as GPUs or TPUs, and efficient algorithmic implementations. Additionally, the fine-tuning of hyperparameters and the necessity of extensive simulations for training add to the overall complexity.

To address these challenges, researchers are exploring lightweight RL algorithms (Savaglio et al. 2019)][58], parallel processing (Nair et al. 2015)[59], and distributed computing frameworks (Liang et al. 2018)[60] to reduce training times and costs. Simplified model architectures, combined with transfer learning or pretraining techniques, also help alleviate computational burdens. While algorithmic complexity and computational cost remain obstacles, these advancements are paving the way for more practical and scalable RL solutions in finance.

5. Outlook

5.1 Interdisciplinary Integration

Interdisciplinary integration plays a crucial role in advancing reinforcement learning (RL) applications in the financial sector (Zhang et al. 2021)[61]. By combining insights from finance, computer science, mathematics, and behavioral economics, researchers and practitioners develop more robust and practical RL models tailored to the complexities of financial markets.

In finance, RL benefits from well-established theories such as the Efficient Market Hypothesis (Odermatt et al. 2021)[62] and Modern Portfolio Theory (Jang et al. 2023)[63], which provide foundational frameworks for designing reward functions and constraints. At the same time, computer science contributes cutting-edge advancements in deep learning, optimization algorithms, and high-performance computing, enabling RL models to process and analyze vast amounts of market data in real time. Mathematics plays a pivotal role in defining the stochastic and dynamic aspects of financial problems, often represented as Markov Decision Processes or stochastic differential equations, while behavioral economics adds a human-centric perspective, allowing RL agents to incorporate investor sentiment and irrational market behaviors into their decision-making processes.

A notable example of interdisciplinary integration is the application of RL in algorithmic trading (Théate et al. 2021)[64], where financial theories guide the strat-

egy, computer science ensures scalability, and mathematical modeling accounts for market volatility. Similarly, RL-driven portfolio optimization benefits from economic theories of risk and return, while advanced computational tools streamline the learning and execution process.

Such integration ensures RL models are not only theoretically sound but also practically relevant, allowing for innovative solutions that address the unique challenges of financial markets. By leveraging interdisciplinary approaches, the field continues to push the boundaries of RL applications in finance, paving the way for more intelligent, adaptive, and efficient systems.

5.2 Novel Algorithm Design

Novel algorithm design is pivotal for advancing reinforcement learning (RL) applications in finance (Huang et al. 2024)[65], addressing the unique demands of dynamic and complex market environments. Tailored algorithms are developed to enhance performance, efficiency, and robustness in financial tasks.

For instance, hybrid algorithms that combine traditional financial models with deep RL frameworks (Deep & Akash TTU 2024)[66] have emerged as powerful tools. An example is the integration of Monte Carlo methods with RL to improve exploration in environments with sparse rewards, such as long-term portfolio optimization. Similarly, distributional RL has been adapted to capture the probabilistic nature of financial returns (Pacheco Aznar & David 2023)[67], offering better risk management insights compared to standard RL approaches.

Another innovation involves multi-agent RL (Karpe et al. 2020)[27], where multiple agents interact within a simulated financial market to mimic real-world scenarios such as competition among traders or collaborative risk-sharing. These approaches enable more realistic and effective training, allowing algorithms to adapt to complex interdependencies and market dynamics. Such novel designs not only improve RL's applicability to financial problems but also inspire broader advancements in algorithmic development, setting a foundation for more sophisticated financial decision-making systems.

6. Summary

This review explored the application of reinforcement learning (RL) in financial domains, emphasizing its transformative potential and associated challenges. Beginning with an overview of key concepts and algorithms, we examined how financial tasks such as trading, portfolio optimization, and risk management are modeled as Markov Decision Processes, enabling RL to adapt to dynamic market conditions. This review mainly highlighted the practical implementation of RL in high-frequency trading, portfolio allocation, and hedging strategies, showcasing the advantages of adaptability and data-driven decision-making.

We also addressed some critical challenges, including data complexity and scarcity, stability and generalization, and computational costs, outlining strategies like data augmentation, robust feature engineering, ensemble methods, and lightweight algorithm designs to overcome these barriers. Interdisciplinary integration and novel algorithm development were identified as key drivers for advancing RL's efficacy in the financial sector, blending insights from finance, computer science, and mathematics to create robust and efficient solutions.

While RL demonstrates significant promise in improving financial decision-making, challenges such as interpretability, scalability, and real-world validation persist. Future research should focus on enhancing algorithmic robustness, exploring new methodologies for market simulation, and fostering collaboration across disciplines. By addressing these challenges, RL has the potential to redefine the landscape of financial technology and innovation.

References

[1] A. E. Greenberg and H. E. Hershfield, "Financial decision making," *Consumer Psychology Review*, vol. 2, no. 1, pp. 17–29, 2019.

[2] G. Cao, Y. Zhang, Q. Lou, and G. Wang, "Optimization of high-frequency trading strategies using deep reinforcement learning," *Journal of Artificial Intelligence General science (JAIGS) ISSN: 3006-4023*, vol. 6, no. 1, pp. 230–257, 2024.

[3] Y.-J. Hu and S.-J. Lin, "Deep reinforcement learning for optimizing finance portfolio management," in *2019 amity international conference on artificial intelligence (AICAI)*, pp. 14–20, IEEE, 2019.

[4] Z. Stoiljkovic, "Applying reinforcement learning to option pricing and hedging," *arXiv preprint arXiv:2310.04336*, 2023.

[5] C. Liu, C. Ventre, and M. Polukarov, "Synthetic data augmentation for deep reinforcement learning in financial trading," in *Proceedings of the third ACM international conference on AI in finance*, pp. 343–351, 2022.

[6] A. Marinescu, I. Dusparic, and S. Clarke, "Prediction-based multi-agent reinforcement learning in inherently non-stationary environments," *ACM Transactions on Autonomous and Adaptive Systems (TAAS)*, vol. 12, no. 2, pp. 1–23, 2017.

[7] S. Mishra, "A reinforcement learning approach for training complex decision making models," *Journal of AI-Assisted Scientific Discovery*, vol. 2, no. 2, pp. 329–352, 2022.

[8] Y. Yang and J. Wang, "An overview of multi-agent reinforcement learning from game theoretical perspective," *arXiv preprint arXiv:2011.00583*, 2020.

[9] S. Yu, X. Chen, Z. Zhou, X. Gong, and D. Wu, "When deep reinforcement learning meets federated learning: Intelligent multitimescale resource management for multiaccess edge computing in 5g ultradense network," *IEEE Internet of Things Journal*, vol. 8, no. 4, pp. 2238–2251, 2020.

[10] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[11] A. F. Araújo and A. P. Braga, "Reward-penalty reinforcement learning scheme for planning and reactive behaviour," in *SMC'98 Conference Proceedings. 1998 IEEE International Conference on Systems, Man, and Cybernetics (Cat. No. 98CH36218)*, vol. 2, pp. 1485–1490, IEEE, 1998.

[12] N. Bäuerle and U. Rieder, *Markov decision processes with applications to finance*. Springer Science & Business Media, 2011.

[13] Y.-t. Liu, J.-m. Yang, L. Chen, T. Guo, and Y. Jiang, "Overview of reinforcement learning based on value and policy," in *2020 Chinese Control And Decision Conference (CCDC)*, pp. 598–603, IEEE, 2020.

[14] M. C. McKenzie and M. D. McDonnell, "Modern value based reinforcement learning: A chronological review," *IEEE Access*, vol. 10, pp. 134704–134725, 2022.

[15] F. Lu, J. Mathias, S. Meyn, and K. Kalsi, "Model-free characterizations of the hamilton-jacobi-bellman equation and convex q-learning in continuous time," *arXiv preprint arXiv:2210.08131*, 2022.

[16] S. T. Boppiniti, "Evolution of reinforcement learning: From q-learning to deep," *Available at SSRN 5061696*, 2021.

[17] X. Wang, Y. Gu, Y. Cheng, A. Liu, and C. P. Chen, "Approximate policy-based accelerated deep reinforcement learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 6, pp. 1820–1830, 2019.

[18] S. Zhong, Q. Liu, Z. Zhang, and Q. Fu, "Efficient reinforcement learning in continuous state and action spaces with dyna and policy approximation," *Frontiers of Computer Science*, vol. 13, pp. 106–126, 2019.

[19] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[20] J. Schulman, "Trust region policy optimization," *arXiv preprint arXiv:1502.05477*, 2015.

[21] G. Scaletta, *Deep Reinforcement Learning for Portfolio Optimization.* PhD thesis, Politecnico di Torino, 2024.

[22] V. Konda and J. Tsitsiklis, "Actor-critic algorithms," *Advances in neural information processing systems*, vol. 12, 1999.

[23] H. Yang, X.-Y. Liu, S. Zhong, and A. Walid, "Deep reinforcement learning for automated stock trading: An ensemble strategy," in *Proceedings of the first ACM international conference on AI in finance*, pp. 1–8, 2020.

[24] F. Khemlichi, H. Chougrad, Y. I. Khamlichi, A. El Boushaki, and S. E. B. Ali, "Deep deterministic policy gradient for portfolio management," in *2020 6th IEEE Congress on Information Science and Technology (CiSt)*, pp. 424–429, IEEE, 2021.

[25] C. Kang, J. Woo, and J. W.-K. Hong, "A2c reinforcement learning for cryptocurrency trading and asset management," in *2024 IEEE International Conference on Blockchain and Cryptocurrency (ICBC)*, pp. 1–7, IEEE, 2024.

[26] L. Sadriu, "Deep reinforcement learning approach to portfolio optimization," 2022.

[27] M. Karpe, J. Fang, Z. Ma, and C. Wang, "Multi-agent reinforcement learning in a realistic limit order book market simulation," in *Proceedings of the first ACM international conference on AI in finance*, pp. 1–7, 2020.

[28] D. Challet and T. Galla, "Price return autocorrelation and predictability in agent-based models of financial markets," *Quantitative Finance*, vol. 5, no. 6, pp. 569–576, 2005.

[29] H. Shui, X. Sha, B. Chen, and J. Wu, "Stock weighted average price prediction based on feature engineering and lightgbm model," in *Proceedings of the 2024 International Conference on Digital Society and Artificial Intelligence*, pp. 336–340, 2024.

[30] D.-Y. Park and K.-H. Lee, "Practical algorithmic trading using state representation learning and imitative reinforcement learning," *IEEE Access*, vol. 9, pp. 152310–152321, 2021.

[31] B. R. Routledge, "Adaptive learning in financial markets," *The Review of Financial Studies*, vol. 12, no. 5, pp. 1165–1202, 1999.

[32] L. Zhang and L. Hua, "Major issues in high-frequency financial data analysis: A survey of solutions," *Available at SSRN 4834362*, 2024.

[33] P. Motard, "Hierarchical reinforcement learning for algorithmic trading," 2022.

[34] A. J. Menkveld and M. A. Zoican, "Need for speed? low latency trading and adverse selection," *Manuscript*, vol. 7, 2013.

[35] A. Roncella and I. Ferrero, "The ethics of financial market making and its implications for high-frequency trading," *Journal of business ethics*, vol. 181, no. 1, pp. 139–151, 2022.

[36] H. A. Bell and H. Searles, "An analysis of global hft regulation," *George Mason University WorkingPaper*, no. 14-11, 2014.

[37] J. Cao, J. Chen, J. Hull, and Z. Poulos, "Deep hedging of derivatives using reinforcement learning," *arXiv preprint arXiv:2103.16409*, 2021.

[38] W. Shin, S.-J. Bu, and S.-B. Cho, "Automatic financial trading agent for low-risk portfolio management using deep reinforcement learning," *arXiv preprint arXiv:1909.03278*, 2019.

[39] H. Park, M. K. Sim, and D. G. Choi, "An intelligent financial portfolio trading strategy using deep q-learning," *Expert Systems with Applications*, vol. 158, p. 113573, 2020.

[40] P. Liu, "A review on derivative hedging using reinforcement learning," *Journal of Financial Data Science*, p. 1, 2023.

[41] Y. Bai, Y. Gao, R. Wan, S. Zhang, and R. Song, "A review of reinforcement learning in financial applications," *Annual Review of Statistics and Its Application*, vol. 12, 2024.

[42] J. J. Seddon and W. L. Currie, "A model for unpacking big data analytics in high-frequency trading," *Journal of Business Research*, vol. 70, pp. 300–307, 2017.

[43] A.-A. Semenoglou, E. Spiliotis, and V. Assimakopoulos, "Data augmentation for univariate time series forecasting with neural networks," *Pattern Recognition*, vol. 134, p. 109132, 2023.

[44] Y. Naritomi and T. Adachi, "Data augmentation of high frequency financial data using generative adversarial network," in *2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, pp. 641–648, IEEE, 2020.

[45] V. Lanzetta, "Transfer learning for financial data predictions: a systematic review," *arXiv preprint arXiv:2409.17183*, 2024.

[46] M. A. Salamkar, "Feature engineering: Using ai techniques for automated feature extraction and selection in large datasets," *Journal of Artificial Intelligence Research and Applications*, vol. 3, no. 2, pp. 1130–1148, 2023.

[47] C. Mascioli, A. Gu, Y. Wang, M. Chakraborty, and M. Wellman, "A financial market simulation environment for trading agents using deep reinforcement learning," in *Proceedings of the 5th ACM International Conference on AI in Finance*, pp. 117–125, 2024.

[48] J. Boyan and A. Moore, "Generalization in reinforcement learning: Safely approximating the value function," *Advances in neural information processing systems*, vol. 7, 1994.

[49] C. Tuite, A. Agapitos, M. O'Neill, and A. Brabazon, "Tackling overfitting in evolutionary-driven financial model induction," *Natural Computing in Computational Finance: Volume 4*, pp. 141–161, 2012.

[50] T. Eimer, M. Lindauer, and R. Raileanu, "Hyperparameters in reinforcement learning and how to tune them," in *International Conference on Machine Learning*, pp. 9104–9149, PMLR, 2023.

[51] C. Zhang, O. Vinyals, R. Munos, and S. Bengio, "A study on overfitting in deep reinforcement learning," *arXiv preprint arXiv:1804.06893*, 2018.

[52] E. Akyildirim, A. Sensoy, G. Gulay, S. Corbet, and H. N. Salari, "Big data analytics, order imbalance and the predictability of stock returns," *Journal of Multinational Financial Management*, vol. 62, p. 100717, 2021.

[53] K. Wang, B. Kang, J. Shao, and J. Feng, "Improving generalization in reinforcement learning with mixture regularization," *Advances in Neural Information Processing Systems*, vol. 33, pp. 7968–7978, 2020.

[54] G. Raskutti, M. J. Wainwright, and B. Yu, "Early stopping and non-parametric regression: an optimal data-dependent stopping rule," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 335–366, 2014.

[55] T. Bai, J. Luo, J. Zhao, B. Wen, and Q. Wang, "Recent advances in adversarial training for adversarial robustness," *arXiv preprint arXiv:2102.01356*, 2021.

[56] R. Singh, Q. Zhang, and Y. Chen, "Improving robustness via risk averse distributional reinforcement learning," in *Learning for Dynamics and Control*, pp. 958–968, PMLR, 2020.

[57] A. Nagabandi, I. Clavera, S. Liu, R. S. Fearing, P. Abbeel, S. Levine, and C. Finn, "Learning to adapt in dynamic, real-world environments through meta-reinforcement learning," *arXiv preprint arXiv:1803.11347*, 2018.

[58] C. Savaglio, P. Pace, G. Aloi, A. Liotta, and G. Fortino, "Lightweight reinforcement learning for energy efficient communications in wireless sensor networks," *IEEE Access*, vol. 7, pp. 29355–29364, 2019.

[59] A. Nair, P. Srinivasan, S. Blackwell, C. Alcicek, R. Fearon, A. De Maria, V. Panneershelvam, M. Suleyman, C. Beattie, S. Petersen, *et al.*, "Massively parallel methods for deep reinforcement learning," *arXiv preprint arXiv:1507.04296*, 2015.

[60] E. Liang, R. Liaw, R. Nishihara, P. Moritz, R. Fox, K. Goldberg, J. Gonzalez, M. Jordan, and I. Stoica, "Rllib: Abstractions for distributed reinforcement learning," in *International conference on machine learning*, pp. 3053–3062, PMLR, 2018.

[61] W. Zhang, A. Valencia, and N.-B. Chang, "Synergistic integration between machine learning and agent-based modeling: A multidisciplinary review," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 5, pp. 2170–2190, 2021.

[62] L. Odermatt, J. Beqiraj, and J. Osterrieder, "Deep reinforcement learning for finance and the efficient market hypothesis," *Available at SSRN 3865019*, 2021.

[63] J. Jang and N. Seong, "Deep reinforcement learning for stock portfolio optimization by connecting with modern portfolio theory," *Expert Systems with Applications*, vol. 218, p. 119556, 2023.

[64] T. Théate and D. Ernst, "An application of deep reinforcement learning to algorithmic trading," *Expert Systems with Applications*, vol. 173, p. 114632, 2021.

[65] Y. Huang, X. Wan, L. Zhang, and X. Lu, "A novel deep reinforcement learning framework with bilstm-attention networks for algorithmic trading," *Expert Systems with Applications*, vol. 240, p. 122581, 2024.

[66] A. T. Deep, "Advanced financial market forecasting: integrating monte carlo simulations with ensemble machine learning models," 2024.

[67] D. Pacheco Aznar, "Portfolio management: A deep distributional rl approach," *Available at SSRN*, 2023.