Deep Reinforcement Learning for Automated Stock Trading: Inclusion of Short Selling

Eeshaan Asodekar(✉), Arpan Nookala, Sayali Ayre, and Anant V. Nimkar

Sardar Patel Institute of Technology, Mumbai, India
{eeshaan.asodekar,ganesharpan.nookala,sayali.ayre,
anant_nimkar}@spit.ac.in

Abstract. Multiple facets of the financial industry, such as algorithmic trading, have greatly benefited from their unison with cutting-edge machine learning research in recent years. However, despite significant research efforts directed towards leveraging supervised learning methods alone for designing superior algorithmic trading strategies, existing studies continue to confront significant hurdles like striking the optimum balance of risk and return, incorporating real-world complexities, and minimizing max drawdown periods. This research work proposes a modified deep reinforcement learning (DRL) approach to automated stock trading with the inclusion of short selling, a new thresholding framework, and employs turbulence as a safety switch. The DRL agents' performance is evaluated on the U.S. stock market's DJIA index constituents. The modified DRL agents are shown to outperform previous DRL approaches and the DJIA index, in terms of absolute returns, risk-adjusted returns, and lower max drawdowns, while giving insights into the effects of short selling inclusion and proposed thresholding.

Keywords: Machine learning · Deep reinforcement learning · Actor-critic framework · Markov Decision Process · Automated stock trading

1 Introduction

The optimal stock trading strategy problem, and the search for a robust and perennially profitable automated stock trading strategy have intrigued researchers and industry practitioners alike, even in decades prior to modern advances in machine learning and the exponential growth of affordable computing power. The primary aim of the optimal stock trading strategy is the maximization of investment returns, on the basis of projected returns and the underlying risk of a given set of stocks, by optimizing the weighted allocation of capital. A reliable and sustainably profitable equity trading strategy is instrumental to asset management companies and Quant funds.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022
M. Ceci et al. (Eds.): ISMIS 2022, LNAI 13515, pp. 187–197, 2022.
https://doi.org/10.1007/978-3-031-16564-1_18

The traditional approach proposed by Markowitz [10] consists of two phases. The expected returns and covariance matrix for a stock are computed. Then, optimal weights are computed by either maximizing the return for a target risk level for the portfolio, or by minimizing the risk level for a target level of return. However, if the investor desires to update the decisions taken at each step and consider factors such as transaction cost, this technique might be challenging to execute. Another disadvantage of this approach is its inability to incorporate technical indicators, which are used to identify stock trading opportunities based on statistical trends. A second approach to stock trading is to construe it as a Markov Decision Process (MDP) [13] and then narrow it down on the optimal strategy using Dynamic Programming. However, the scalability of this paradigm is limited due to the mammoth state spaces needed to encompass the stock market and its complexities using dynamic programming. High-frequency trading (HFT) is another approach that gained a foothold in the last decade. HFT uses sophisticated computer algorithms to trade humongous quantities of stock in fractions of a second, based on analysis of multiple markets and financial indicators. However, because of increased competition, reduced volatility and trading volumes, stricter regulations, and high maintenance costs, there has been a significant decline in the profitability of HFTs. These factors aggravated conditions for HFTs, which were already operating within a zero-sum game.

A recent approach is the application of machine learning (ML) [7] and deep learning (DL) algorithms [16] to predict stock prices and infer market conditions. Although the results are nearly satisfactory and have improved over the years, there are significant deficiencies in this approach of utilizing supervised learning methods alone to forecast stock prices with the intention of making a profit. Firstly, accurately forecasting stock prices does not automatically translate into the system making a profit. Secondly, it is excruciatingly complex to train a supervised learning model to consider all the complications, uncertainties and latency of the financial markets. Thirdly, profitable trading strategies vary significantly according to market situations, and thus the supervised learning system must master not a single, but multiple strategies, therefore there is not a single strategy that the model must optimize for.

This paper proposes a modified deep reinforcement learning (DRL) approach to the optimal stock trading strategy problem. The main contributions of this work are as follows. Firstly, we propose a modified DRL approach with the addition of short selling to the action space of the agents, thereby increasing the accrued alpha. Secondly, in order to maximize the net profit, a thresholding control is implemented, which significantly decreases the possibility of the model acting on weak or stray sell signals. Lastly, we also discuss the use of the turbulence as a safety switch to avert excruciating losses during extreme events like market crashes and recessions.

The rest of the paper is organized as follows. Section 2 presents a literature survey the various approaches to the optimal stock trading strategy problem. Section 3 gives background on the three DRL algorithms used. Section 4 presents the proposed modified DRL approach to automated stock trading. Section 5

delineates the performance metrics and presents the results along with discussion. Section 6 gives concluding remarks and future scope of work.

2 Related Works

It is only in the latter half of this decade that prodigious computing power has become available to the world. As a result, most algorithmic trading models built by mathematicians, physicists, and financial experts to tackle the optimal stock trading strategy problem do not use machine learning. Such traditional trading models use methods based on complex mathematical analysis, and utilize mean reversion and trend following methods.

Most academic research on applying machine learning (ML) [7] methods to algorithmic trading focus on forecasting or prediction of financial securities using deep learning (DL) methodologies [16]. The overarching idea of this approach is that if the price or value of a specific financial security or asset is known in advance, to a reasonable degree of accuracy, then a party may make strategically beneficial and optimum trading decisions to profit from market movements. Various supervised learning approaches [16] with significant modifications since their inception have already been investigated.

Alternatively, fewer research works [1] have explored Reinforcement Learning (RL) approaches, with some researchers proposing significant modifications to the vanilla approaches [4] to tackle the challenge of algorithmic trading and portfolio management. Moody et al. [12] asserted the superiority of reinforcement learning techniques over traditional supervised learning techniques, and have employed direct reinforcement learning and recurrent reinforcement learning models to train intraday trading agents. Researchers have also employed an adaptive RL model [5] as a basis for automated trading systems.

Deep reinforcement learning (DRL) methodologies [2,3], which are a nascent branch of machine learning, are now being explored to tackle the automated stock trading problem in recent times by researchers. Deng et al. [6] introduced contemporary DL into the traditional DRL framework and have made the system technical indicator free. Another research work explored the DRL approach by training a Deep Deterministic Policy Gradient (DDPG) agent to learn a stock trading strategy [18] with promising results. More recently, an Ensemble Strategy [17] combining three actor-critic DRL algorithms was proposed by researchers.

3 Background

This section gives a distilled description of the three DRL algorithms employed to achieve the objective of inter-day profit maximization, and the four technical indicator inputs given to the DRL agents.

3.1 Deep Reinforcement Learning Algorithms

1. Advantage Actor Critic (A2C)
   A2C [11] is a synchronous, deterministic, actor-critic algorithm that employs an advantage function to reduce the variance of the policy gradient, thereby making the model more robust. It waits for each actor to complete their segment of experience before updating, averaging over all actors.
2. Deep Deterministic Policy Gradient (DDPG)
   DDPG [9] learns a Q-function and a policy at the same time. It is specifically adapted for environments with continuous action spaces, thereby making it a felicitous algorithm for stock trading, and also maximizes investment returns. It learns the Q-function using off-policy data and the Bellman equation and then utilizes the Q-function to learn the policy.
3. Proximal Policy Optimization (PPO)
   PPO [14] is an on-policy algorithm, which guarantees that the new updated policy does not differ significantly from the preceding policy. This results in less variance in training at the expense of some bias, but it assures smoother training and ensures that the agent does not incur cataclysmic drops in performance. Thus, PPO is robust, stable, and executes with celerity.

3.2 Technical Indicators

Technical indicators are pattern based signals derived by analyzing historical stock data. Each of these indicators gives unique insights into stock price movements based on statistical trends.

1. Moving Average Convergence Divergence (MACD): A strong indicator of trend-following momentum, it juxtaposes multiple moving averages of the price of a stock, each of a different time frame.
2. Commodity Channel Index (CCI): An indicator of irrational behaviour, representing potential opportunities wherein a stock is being overbought due to unreasonable optimism or being oversold due to undue pessimism.
3. Relative Strength Index (RSI): A momentum indicator which evaluates recent stock price changes to signal a bullish or bearish momentum.
4. Average Directional Index (ADX): A tool used to gauge the strength of stock prices' bullish or bearish trends.

4 Deep Reinforcement Learning Approach to Automated Stock Trading

The following section details the formulation of the deep reinforcement learning paradigm applied to automated stock trading and expounds the proposed modifications to the DRL agents.

4.1 Stock Trading Problem Formulation

To account for the stochasticity and the uncertainties of the real-world financial markets, a Markov Decision Process is used as the framework as follows:

- State $s = [b, c, p]$: an information vector, which encapsulates the current cash balance b, the current holdings c, and the stock price p, for all N stocks under consideration, where $b \in Q$, $c \in Z$, $p \in Z^+$
- Action a: a vector of actions allowed for the agent, for all the N stocks under consideration. The allowed actions on each stock include buying, holding, selling, shorting, and covering a short.
- Reward r : The reward signal is the difference between the value of the portfolio holding on the n^{th} and the $(n+1)^{th}$ day i.e. inter-day profit
- Policy $\pi(s)$: The current trading strategy, at state s, which is the probability distribution of actions available to the agent at state s.
- Q-value Q_π(s, a): The expected reward for taking a specific action a, at state s, by following policy π.

At any given time step n, the agent has the following actions at its disposal for all the N stocks under consideration:

- Buying $p[d]$ shares results in $c_{t+1}[d] = c_t[d] + p[d]$, and $d = 1, ..., N$
- Holding, $c_{t+1}[d] = c_t[d]$ shares
- Selling $q[d] \in [1, c[d]]$ shares results in $c_{t+1}[d] = c_t[d] - q[d]$, where $q[d] \in Z^+$
- Shorting $r\prime[d]$ shares results in $c_{t+1}[d] = -(r_t[d] + r\prime[d])$
- Covering $u[d]$ shares, thereby reducing or eliminating the shorting position to $c_{t+1}[d] = -(r_t[d] - u[d])$.

Thus, depending upon the action taken by the agent in the time step t, the portfolio value at $t + 1$ would be a spectrum of values, all depending upon the combination of the actions taken by the agent in the preceding time step, t.

4.2 Dataset

The agents are trained on the U.S. equity market's DJIA index's constituents, from January 2000 to December 2016 (in-sample) and the performance is evaluated from January 2017 to June 2021 (out-of-sample). The input data to the DRL agents was the daily end of day data (viz. open, high, low, adjusted close, and volume data) and the four technical indicators (viz. MACD, CCI, RSI, and ADX) as given in Sect. 3.2. The given time-span of the data is specifically chosen such that the agents get the experience of all the stages of the equity market cycle; from the early phase booms, to consolidation and steady growth in the mid-phase, to a crash or a meltdown towards the ending of the cycle.

4.3 Trading Objectives

The primary underlying objective of the agent is inter-day net profit maximization. The reward function is defined such that the reward at time step $t+1$ is the difference between the value of the agent's long and short positions from time step t and $t + 1$, thereby maximizing daily positive change in portfolio value.

4.4 Financial Markets' Constraints

The trading costs and commissions are incorporated in the proposed scheme, and it is assumed to be 0.1% for each transaction, to reflect the trading costs incurred in the financial markets. Also, the constituents of the *Dow Jones Industrial Average* (DJIA) are deliberately chosen since they have sufficient market liquidity.

4.5 Shorting Thresholding

To decrease the likelihood of the agents acting on stray or weak selling(and thus, shorting) signals, a thresholding mechanism is put into place. k is a value used by the agents as a decision parameter and ranges from 1 to -1. A positive value of k results in a long position being taken or a short being covered, wherein the size of the position is determined by the magnitude of k. Whereas a negative value of k results in a shorting position being taken or a long position being sold. The thresholding mechanism allows the agent to take a shorting position only when the k value goes beyond a certain cutoff value.

4.6 Turbulence as a Safety Switch

Turbulence [8] is frequently used as an indicator for market volatility and extreme price fluctuations, which occurs during sudden events like wars, bubbles bursting, and financial crises. The agents are hard-coded to liquidate all positions when the turbulence reaches a certain threshold and the effect of which are delineated in the results. Mathematically, turbulence is represented as:

$$d_t = (y_t - \mu)\sum^{-1}(y_t - \mu)' \tag{1}$$

where d_t denotes the turbulence for a particular time period t (scalar), y_t denotes a vector of asset returns for period t ($1 \times$ n vector), μ denotes the average vector of historical returns ($1 \times$ n vector), $\sum$ denotes historical returns' covariance matrix (n $\times$ n matrix).

5 Results and Discussion

This section describes the performance metrics of evaluation, and showcases the results obtained and the significant alpha delivered by the three modified DRL agents in comparison to vanilla DRL agents and benchmarks. All models are given access to one million USD of capital at the beginning of their evaluation period. We also present a rationale on the underlying factors contributing to the superior performance of the modified DRL agents proposed.

5.1 Performance Metrics

Performance Metrics used in the evaluation are: firstly, the Cumulative returns, which is the difference between the final value of the portfolio and the initial value during evaluation. Secondly, the annual returns are the geometric average of the returns earned by the agent every year, during evaluation. Thirdly, the Sharpe ratio [15], which gives key insights into risk-adjusted returns, and is used widely in the industry. Lastly, Max Drawdown, which is the biggest loss experienced by a portfolio from its maximum to its minimum value before a new maximum value is reached again.

5.2 Performance Evaluation of Modified DRL Agents

The agents' training is continued in the performance evaluation stage to help acclimatize to ongoing market conditions. It can be observed from Table 1 that the three DRL agents with the proposed modifications outperform the vanilla agents, the ensemble approach [17], and the DJIA index in terms of the cumulative returns, annual returns, Sharpe ratio, and max drawdowns (Fig. 1). The benchmark approaches give cumulative returns of 70%–91%, while vanilla DRL methods give 72%–81%, and the proposed DRL models give 96%–146%. A similar outperformance can be seen from Table 1 for the remaining three performance metrics. All the three modified DRL agents beat the DJIA and the ensemble method, during the bull market periods as indicated by higher returns, as well as during bear market periods, as indicated by lower max drawdowns.

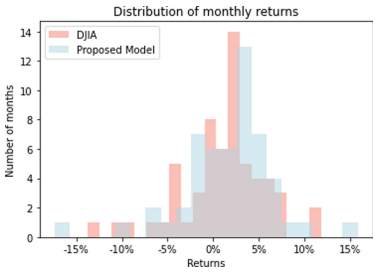

Fig. 1. Monthly returns of Proposed Model (A2C and k=-0.4) vs DJIA

Table 1. Performance of the 3 modified DRL agents over the evaluation period

Models	Proposed models			Vanilla DRL models			Benchmarks	
	PPO	A2C	DDPG	PPO	A2C	DDPG	DJIA index	Ensemble
Cumulative return (%)	121.97	96.71	146.61	72.84	75.33	81.26	70.32	91.46
Annual return (%)	20.21	16.82	23.04	13.39	13.77	14.64	13.19	16.513
Sharpe ratio	1.03	1.12	0.93	0.66	0.76	0.64	0.77	0.83
Max drawdown (%)	−21.82	−18.32	−33.39	−37.07	−28.27	−40.51	−34.78	−32.684

Short Selling Inclusion. The inclusion of short selling adds two more action dimensions to the actions space of the DRL agent, and the results imply its effective adaptation by automated trading DRL agents. It can be observed from the results showcased in Table 1, all the agents with proposed modifications beat

their vanilla counterparts and the ensemble approach, both of which did not have the shorting action, thereby asserting the value of the addition of shorting and covering in the action space of the agent.

Comparison of the Three DRL Agents. The DDPG agent was the best performing in terms of returns, but the A2C agent was the best performing in terms of the Sharpe ratio and the max drawdowns. It was noticed that as soon as the markets started to rebound after the March 2020 Covid-19 recession, A2C and DDPG agent's performance paled in comparison to that of the PPO agent, which was a departure form the previous trend. This leads to the inference that the A2C and DDPG agents give better returns during the middle and the matured phases of an equity market cycle. However, the PPO agent is seen to be performing better during the early to mid-phase of the equity market cycle.

Thresholding Shorting Signals. To decrease the likelihood of the agents acting on a stray or weak selling(and thus, shorting) signal, the thresholding mechanism was put into place to allow the agent to take a shorting position only when the sell signal beyond a certain cutoff value k. There is a significant decrease in the losses due to shorting positions which the model acted upon due to weaker sell signals, as evidenced by the results of this modification as showcased in Table 2. We can observe increasing annual returns and cumulative returns, and decreasing max drawdowns as we decrease the k value from -0.2

Table 2. Variation of returns and max drawdown as shorting signals are thresholded

k	A2C			PPO			DDPG		
	Annual return	Cumulative return	Max drawdown	Annual return	Cumulative return	Max drawdown	Annual return	Cumulative Return	Max drawdown
−0.9	13.46%	76.68%	−31.61%	13.97%	76.69%	−31.61%	20%	123.44%	−36.42%
−0.8	14%	109.16%	−28.61%	18.47%	109.16%	−28.61%	15.19%	85.06%	−37.17%
−0.7	25.03%	55.87%	−45.18%	10.74%	55.87%	−45.18%	15.33%	86.08%	−24.30%
−0.6	13.29%	17.21%	−58.37%	3.72%	17.21%	−58.37%	20.31%	123.64%	−36.94%
−0.5	11.42%	72.84%	−37.07%	13.39%	72.84%	−37.07%	13.06%	70.64%	−36.94%
−0.4	20.89%	54.07%	−41.95%	10.44%	54.07%	−41.95%	10.94%	57.11%	−37.72%
−0.3	17.64%	80.40%	−38.89%	14.52%	80.40%	−38.89%	23.04%	146.61%	−33.39%
−0.2	13.77%	39.79%	−44.30%	8%	39.79%	−44.30%	14.64%	81.26%	−40.11%

Table 3. A2C turbulence thresholding with shorting cutoff k = 0.4

Volatility threshold	Annual Return	Cumulative return	Max drawdown	Annual volatility	Sharpe ratio
50	0.06%	0.28%	−0.65%	0.49%	0.13
100	10.10%	52.05%	−11.30%	9.38%	1.07
150	16.82%	96.71%	−18.32%	14.91%	1.12
200	19.89%	120.28%	−31.47%	20.57%	0.99
250	16.42%	93.80%	−23.65%	19.39%	0.88
300	17.50%	101.75%	−28.19%	20.66%	0.88

to −0.6; beyond which the returns start to decrease because of an excessively high threshold for shorting, making the agents long-only in most situations.

Experimentation with Turbulence as a Safety Switch. The agents are hard-coded to liquidate all positions if the turbulence in the equity markets exceeds a certain set limit. If the limit is set to be too large, the agents are allowed to trade even when the markets are in a downturn incurring excruciating losses. At the same time, if the level is too low the agents liquidate all positions when the markets get even a little turbulent in a long bull run. Few selected results are presented in Table 3, which lead to the conclusion that a threshold of 100 to 150 serves the best in acting as a safety switch whilst still allowing the agents to take advantage of market movements, but a threshold in the range of 50 proves to be very restrictive and sensitive. A threshold of 200 and beyond is equivalent to not putting any safety switch for liquidation, as those levels of turbulence are seldom met, and significant losses would have already been incurred in the downturn when these levels are breached. It can be observed from Table 3 that the best results in terms of the returns, max drawdown, volatility, and Sharpe ratio are obtained when the turbulence threshold is set at 150, beyond which although there is an increase in the returns but volatility increases significantly, and a decreased Sharpe ratio is obtained. Thus, it can be inferred that the use of the market volatility as a safety switch proves to be critical especially during panic periods and the early phases of a recession, forcing the agents to cut their losses. This modification proves to be significant in reducing the max downturn during equity market crashes, thereby working as an effective safety switch.

6 Conclusion

In this paper, we have proposed a modified deep reinforcement learning (DRL) approach for the agents to learn a robust and consistently profitable automated stock trading strategy. The modifications being the addition of shorting and covering to the action space of the DRL agents, along with thresholding of the shorting action, and the use of the turbulence as a safety switch for three DRL algorithms, namely Proximal Policy Optimization (PPO), Advantage Actor Critic (A2C), and Deep Deterministic Policy Gradient (DDPG). The inclusion of shorting in the action space of the agents allows them to gain a significant edge over the previously researched DRL approaches, and their non-shorting counterparts as well as the benchmark indices. The thresholding of the sell signals for shorting significantly decreases the losses caused when the agents act on stray or weak selling signals. Also, the use of turbulence as a safety switch greatly cuts the losses that would have accumulated when a excruciating market correction occurs. All the modifications lead to higher returns, Sharpe ratio, and lower max drawdowns compared to past approaches.

In the future, we aim to explore the incorporation of fundamental analysis and macroeconomic parameters, use more esoteric ML techniques such as Autoencoders, and use Natural language Processing to give equity analysts'

research reports as input. We would also like to train and evaluate the agents in an intra-day trading setting, and increase the stock-space to larger indexes' constituents, and expose the agents to the data of the financial markets of the major economies of the world. All with the goal of moving towards a closed loop, perpetually profitable automated stock trading system.

References

1. Bertoluzzo, F., Corazza, M.: Testing different reinforcement learning configurations for financial trading: introduction and applications. Procedia Econ. Fin. **3**, 68–77 (2012)
2. Boukas, I., et al.: A deep reinforcement learning framework for continuous intraday market bidding. Mach. Learn. **110**(9), 2335–2387 (2021). https://doi.org/10.1007/s10994-021-06020-8
3. Chen, L., Gao, Q.: Application of deep reinforcement learning on automated stock trading. In: 2019 IEEE 10th International Conference on Software Engineering and Service Science (ICSESS), pp. 29–33 (2019)
4. Dang, Q.-V.: Reinforcement learning in stock trading. In: Le Thi, H.A., Le, H.M., Pham Dinh, T., Nguyen, N.T. (eds.) ICCSAMA 2019. AISC, vol. 1121, pp. 311–322. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-38364-0_28
5. Dempster, M., Leemans, V.: An automated fx trading system using adaptive reinforcement learning. Expert Syst. Appl. **30**(3), 543–552 (2006)
6. Deng, Y., Bao, F., Kong, Y., Ren, Z., Dai, Q.: Deep direct reinforcement learning for financial signal representation and trading. IEEE Trans. Neural Networks Learn. Syst. **28**(3), 653–664 (2017)
7. Jiao, Y., Jakubowicz, J.: Predicting stock movement direction with machine learning: An extensive study on s&p 500 stocks. In: 2017 IEEE International Conference on Big Data (Big Data), pp. 4705–4713 (2017)
8. Kritzman, M., Li, Y.: Skulls, financial turbulence, and risk management. Financ. Anal. J. **66**(5), 30–41 (2010)
9. Lillicrap, T.P., et al.: Continuous control with deep reinforcement learning. In: ICLR, Conference Track Proceedings (2016)
10. Markowitz, H.: Portfolio selection. J. Finan. **7**(1), 77–91 (1952)
11. Mnih, V., et al.: Asynchronous methods for deep reinforcement learning. In: Proceedings of the 33rd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 48, pp. 1928–1937. PMLR (2016)
12. Moody, J.E., Saffell, M.: Learning to trade via direct reinforcement. IEEE Trans. Neural Networks **12**(4), 875–89 (2001)
13. Neuneier, R.: Optimal asset allocation using adaptive dynamic programming. In: Touretzky, D., Mozer, M., Hasselmo, M. (eds.) Advances in Neural Information Processing Systems, vol. 8. MIT Press (1995)
14. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv:abs/1707.06347 (2017)
15. Sharpe, W.F.: The sharpe ratio. J. Portfolio Manag. **21**(1), 49–58 (1994)
16. Shen, J., Shafiq, M.O.: Short-term stock market price trend prediction using a comprehensive deep learning system. J. Big Data **7**(1), 1–33 (2020). https://doi.org/10.1186/s40537-020-00333-6

17. Yang, H., Liu, X.Y., Zhong, S., Walid, A.: Deep reinforcement learning for automated stock trading: an ensemble strategy. In: Proceedings of the First ACM International Conference on AI in Finance, ICAIF 2020 (2020)
18. Xiong, Z., Liu, X.Y., Zhong, S., Yang, H., Walid, A.: Practical deep reinforcement learning approach for stock trading. In: NeurIPS Workshop on Challenges and Opportunitiesfor AI in Financial Services: the Impact of Fairness, Explainability, Accuracy, and Privacy (2018)