University of Victoria
ECE 559B Deep Reinforcement Learning

Final Project Report

Project name: Portfolio Management using Deep Deterministic Policy Gradient

Student Name: Huan-Rui Zhang

Abstract

In this project, a trained DDPG agent is proposed to maximize the profit in the stock market of the Vanguard 500 Index Fund ETF (NYSE: VOO). Both the technical indicators and economic indicators were utilized to train the DDPG agent. To reduce the amount of data, only the closing price in the market information is used. Unlike most of the research in the portfolio management area, the author of this report wants to explore the benefits of the economic indicators to portfolio management, as economic indicators were related to the economic situations. Two popular portfolio management benchmark metrics were used to measure the trained DDPG agent in this project, which are maximum drawdown (MDD) and Sharpe ratio (SR). The simulation results showed that the proposed DDPG agent has 2.1% to 4.7% improvement in terms of MDD comparing to VOO. The proposed DDPG agent also has a competitive performance in terms of SR. Further improvement may be done by using the stochastic policy and fine-tuning the size of the actor network and the critic network in the future work.

1. Introduction

Deep Reinforcement Learning has been applied in the stock market for several years. The data for the stock market can be classified into structured data and unstructured data [1]. The structured data consists of market information, technical indicators, and economic indicators, as shown in figure 1. The unstructured data has news, social network, and blogs. Since unstructured data are originally higher dimensional text information, they should be pre-processed into digit number by using the Natural Language Processing (NLP) techniques and be reduced the dimension of the data with the Principal Component Analysis (PCA) techniques. This report will heavily use the structured data and will describe the background of each structured data in the following paragraphs. The unstructured data will not be utilized in this report, but it is a highly potential research area for the stock market research, as it is

related to the application of data science.

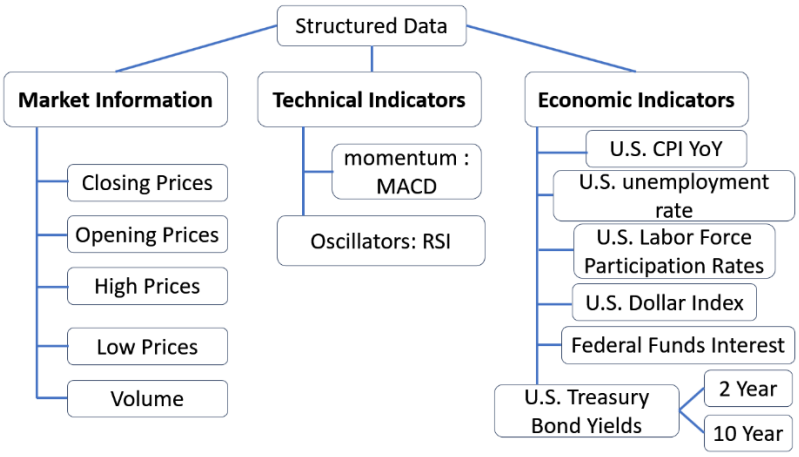

Figure 1. Structured data for the stock market analysis.

The price of the stock market influenced by many factors, including the historical stock market data, and fundamental factors, and traders' psychological behaviors [2]. It is infeasible to use all the available features to predict the stock market. A better way to predict the stock market is to select the most important features from the three types of structured data; the basic features are the information of the stock market, such as the open, high, low, closing prices, and volume. The closing price is the most widely used information in portfolio management or the stock price predictions areas. The technical indicators are calculated from the historical price by using statistical and mathematical formulas. Different indicators represent the different tendencies. The popular technical indicators are the Moving Average Convergence Divergence (MACD) and the Relative Strength Index (RSI). The MACD is a momentum indicator shows the tendency of the stock price, while the RSI is an oscillator measures the stock is overbought or oversold. A better way to use the technical indicators is to understand the limitation of each indicator and choose different with a combination to minimize the limitation.

In terms of the economic indicators, the economic indicators cover from macroeconomic factors to microeconomic factors. The microeconomic factors are information for a company. While the macroeconomic factors are related to a country or the global economy. There are several macroeconomic

factors can be used in the stock market analysis applications, such as the Year-of-Year in Consumer Price Index (CPI YoY), unemployment rates, labor force participation rates, United States (U.S) dollar index, and federal funds interest rate. Some researchers have explored the relationship between macroeconomic features and the stock market return in other countries [3]. The CPI YoY is an indicator that shows the fluctuations of the living cost by year, which is highly related to the inflation situation.

The detailed formula of MACD and RSI are listed below [4]. The MACD is computed from the Exponential Moving Average (EMA). The EMA is calculated by

$$EMA_N(today) = ClosingPrice * k + EMA_N(yesterday) * (1 - k) \quad (1)$$

where $k = \frac{2}{(N+1)}$ denotes the multiplier and the N represents the number of days in EMA.

The MACD is the difference between the 12-EMA and the 26-day EMA.

$$MACD(today) = EMA_{12}(today) - EMA_{26}(today) \quad (2)$$

The RSI is computed from the average gain and average loss.

$$RSI = 100 - (\frac{100}{1+\frac{average\ gain}{average\ loss}}) \quad (3)$$

Value/policy iteration and policy gradient have been combined with deep learning (DL) algorithms by using the Recurrent Neural Network (RNN) [5]; the paper argued to use a non-Markovian decision process (MDP) is more suitable for learning an optimal policy. This project is mainly focused on the RL problems in MDP, hence the scenario of [5] is not feasible in the scope of this project. A Deep Deterministic Policy Gradient (DDPG) has been proposed to deal with the continuous action domain by combining the Deep Q-Learning (DQN) and actor-critic methods [6]. Recently, a robust version of the DDPG algorithm has been published [7], which is more robust to unpredictable and new events in the stock market. The robust DDPG algorithm also achieved a higher rate of return compared to the DDPG algorithm.

The rest of this report is organized as follows. In section 2, the MDP problem formulation will be

described. The problem classification and discussion through critical and analytical thinking will be provided in section 3. In section 4, the solution methods and workflow diagrams will be shown. Section 5 lists the implementation and results of this project, including the learning process and the behavior of the learned policy. Finally, the conclusion and future works will be given in section 6.

2. MDP Problem Formulation

The goal of this task is to build up a DDPG agent to maximize the profit in Vanguard S&P 500 Exchange Traded Fund (ETF) (NYSE: VOO). There are four assumptions in this project:

Assumption 1: The liquidity of VOO is high enough that the agent can carried at the last price when an action is taken.

Assumption 2: The investment made by this agent has no influence on the market.

Assumption 3: Zero trading fee

Assumption 4: Zero expense ratio (0.03% in real world)

The agent, environment, state/action space, reward, transition dynamics are listed below.

Agent: A trading software

Environment: Vanguard 500 Index Fund ETF (NYSE: VOO)

State (All in continuous space):

- Cash: 0 ~ 20000 USD
- Position: 0 ~ 20
- Market information: Closing price from 200 ~ 450
- Technical indicators: MACD, RSI
- Economic indicators: CPI YoY, U.S. unemployment rate, U.S. labor force participation rate, U.S. Dollar Index, Federal Funds Interest, U.S. 2-year and 10-year Treasury Bond Yields

Actions: sell, hold, buy in continuous space -20 ~ 20

Reward:

Profit=$Cash_{t+1}$+($Position_{t+1}$*$Closing\_price_t$) - Original_cash -(Original_position*Original_price) (4)

Profit = -10000 for action is out of the range or (position + Action) < 0 or ((cash - (closing_price * Action)) >= 0)

Transition dynamics: $P_r(S_{t+1},\ R_{t+1}\ |\ S_t,\ A_t)$ is shown in figure 2.

The ideas of choosing this MDP problem formulation will be described in the following paragraphs. The cash will be initialized to 10000 U.S. dollars. While the agent interacting with the environment, the cash varies from 0 to 20000 U.S. dollars. The position means the number of VOO that the agent has bought, ranging from 0 to 20. I only used the closing price from the market information, which ranges from 200 to 450 in the date ranging from January 1, 2017 to November 20, 2023. I used the two technical indicators, MACD and RSI, to measure the tendency and signal overbought/oversold of VOO.

In terms of the economic indicators, the CPI YoY, the U.S. unemployment rate, U.S. labor force participation rates are good hints to understand the inflation situation and economic situation in U.S. The U.S. dollar index and U.S. federal funds interest can also influence the stock market; the higher U.S. dollar index means more funds come to the U.S. market. The U.S. Federal funds interest relates to the inflation, economic, political factors, which will indirectly influence the stock market. The difference between the 2-year U.S. treasury bond yield and 10-year U.S. treasury bond yield is usually referred the "yield curve spread" or simply the "yield spread". The yield of the 10-year U.S. treasury bond is usually higher than the 2-year U.S. treasury bond. When the yield of the 2-year U.S. treasury bond higher than the 10-year U.S. treasury bond, this is called an inverted yield curve. The inverted yield curve implies that the market is becoming more pessimistic about the economic prospects for the near future, which will also affect the stock market.

The actions are continuous space from -20 to 20, which is wanting to gain more flexibility in buying or

selling the stock since the price of VOO is not cheap even only buy/sell one position of VOO. The reward function calculates the difference between the value of current asset and the original asset by considering both the cash and stock. If the action is unreasonable, then the reward will be -10000. The transition dynamics are shown in figure 2. The agent chooses an action *At* based on the state *St* and then obtain the reward *Rt+1*. The next state *St+1* will be updated; the cash of the next state is calculated by the original cash minus the bought position multiply with the last closing price. The position of the next state is the original position adds the position bought/sold in action *At*. The next state of market information, technical information, and economic information updates to the information of next trading day.

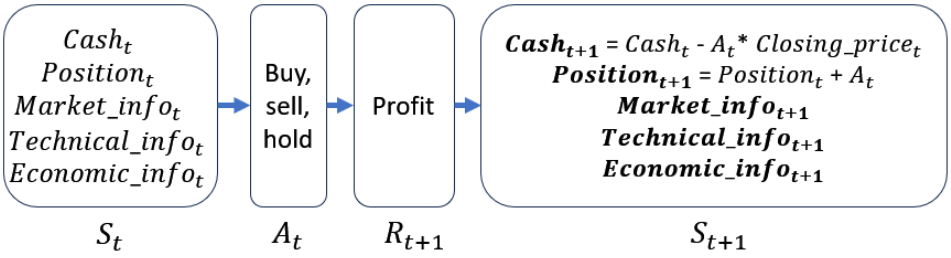

Figure 2. Transition dynamics for this project.

3. Problem Classification

Below summarize the problem classifications of this project.

Continuous task: There is no terminal state of the stock market in this application because stock market opens every trading day. However, the task can be regarded as an episodic task if a terminal state is defined, for instance, every ten trading days is an episodic task. If a terminal state existed, then the DDPG is no longer suitable for this problem, the Monte Carlo (MC) methods will be involved in that case. To clarify, I consider the problem with this project is a continuous task and without any terminal state.

Deterministic policy: This project learns the deterministic policy from experience without knowing the

system dynamics and policy by using DDPG algorithms. The advantages of deterministic policy include easier to interpret and understand and consistent actions for the same set of circumstances. However, the stock market is known for its inherent uncertainties, and market conditions can change rapidly. Although stochastic policies may be more suitable for capturing the dynamic nature of financial markets, the author does not confident to implement the agent by using REINFORCE before the deadline of the project. Therefore, the use of REINFORCE in this topic will be listed in the future works.

Continuous space: The action spaces and state space are all in continuous space. It is unpreventable to use continuous state space because the structured data in the stock market is all in continuous space. The motivation of using the continuous action spaces is to have more efficiently utilize the cash in buying or selling the stock. Imaging that an investor wants to get 200 U.S. dollars back by selling the position of VOO, but the price of VOO is 400 U.S. dollars per position now; if the action space is finite, the investor will need to decide to sell one position or not. If the action space is finite, the DQN will be used in this problem instead of the DDPG. With the benefit of continuous action space, the investor can sell whatever amount of the position of VOO. The same benefit of investors who want to buy VOO but only have 200 U.S. dollars, the investors can buy 0.5 positions of VOO in continuous action space. We can see, hence, the benefits of utilizing the continuous action spaces above achieved better utilization in the fund to investors.

4. Solution Methods

The DDPG algorithm is chosen to be the solution algorithm for this project. DDPG has strengths in continuous action spaces, addressing the maximization bias of DQN, continuous task by temporal differences (TD) method. The drawback of DDPG is that it may stick with the local optimum policy since the DDPG uses two neural networks. There are several schemes to mitigate the possibility of obtaining the local optimum policy: hyperparameter tuning, adding regularization terms, and adjusting

the learning rate during the training phase.

The workflow diagram of this project is shown in figure 3. The DDPG agent consists of a critic network, and an actor network. The critic network is a three layer fully connected layer with 300, 200, and 1 neuron amount. The actor network is also a three layer fully connected layer with 48, 48, and 1 neuron amount. The first two fully connected layers of both networks are followed by a hyperbolic tangent function as the activation function, which is a non-linear activation function. A non-linear activation function is good at dealing with the approximation problem. The critic network takes the state, reward, and the TD error of previous state as its input to evaluate the current TD error and send the current TD error to the actor network. The actor network uses the TD error from the critic network and the state of the environment to derive the corresponding action. The action generated by the actor network will be applied to the environment by the DDPG agent. After the environment changing to the next state, the above workflow will start again. The environment shows in figure 3 is the VOO market. In fact, I also collected the information about the economic indicators which is not part of VOO market information. As section 2 mentioned, the economic indicators are my states either. The reason why I did not use the convolutional neural network (CNN) in my critic network or actor network is that the CNN is good at dealing with the data has spatial relationships such as images. In this project, the information on the state and reward has no relationship in spatial domain.

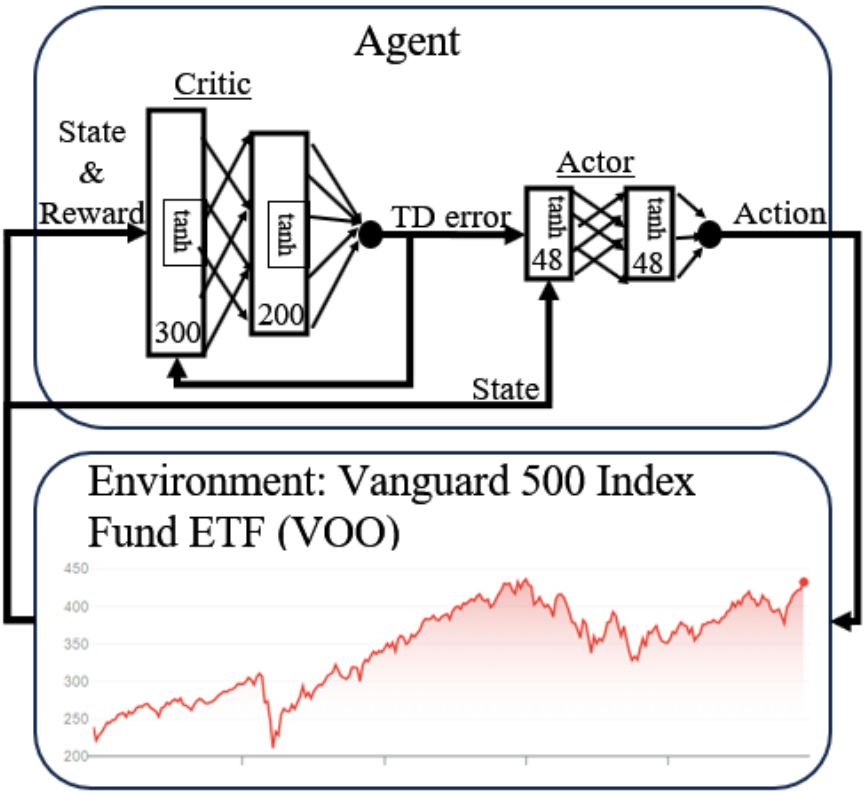

Figure 3. The workflow diagram of this project.

5. Implementation and Result

The general flow of this project is illustrated in figure 4. The initial state of cash is set to be 10000 U.S. dollars; the initial position is generated randomly; the market information, technical information, and economic information is set according to the randomly initialize trading day. Since the historical data of the market information, technical information, and economic information are limited, I selected the historical data on the day from January 1, 2017 to December 31, 2020 as the training dataset; while the testing dataset of the historical data is in the day from January 1, 2021 to November 20, 2023, as illustrated in figure 5. The model of actor network and critic network will be initialized with random values between -1 to 1.

I trained this DDPG agent through over 10000 episodes or 10 hours, with each episode has a random starting date of the trading and a maximum step in 20; as the amount of the historical data is limited. The agent also evaluates one time every 50 episodes. The evaluation during training is to let me know the setting is suitable for this project or not.

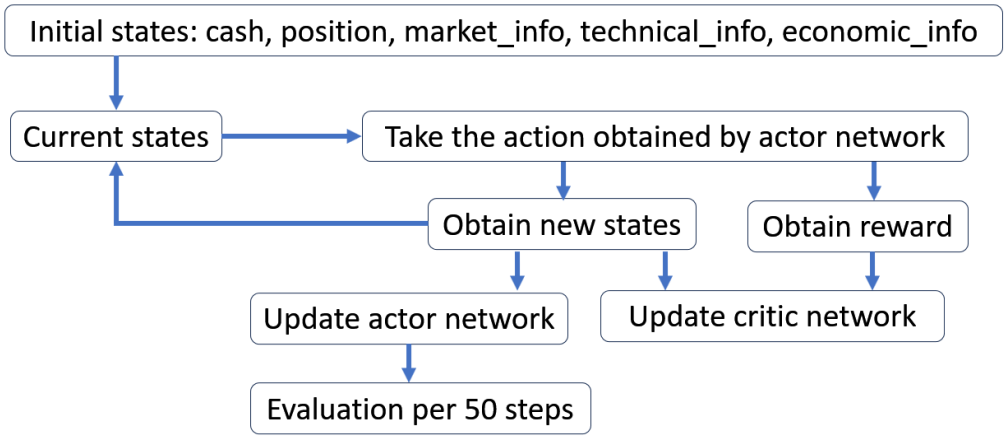

Figure 4. The general flow for training the DDPG agent in this project.

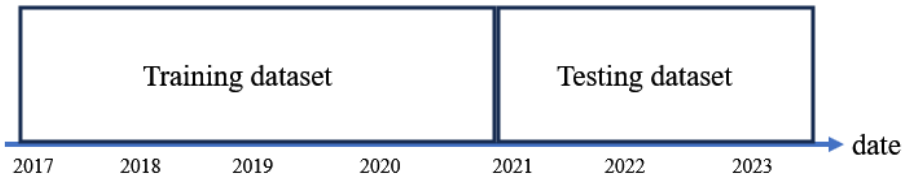

Figure 5. Illustration of the structured data chosen for the training dataset and testing dataset.

Figure 6 illustrates the learning process of training the DDPG agent. The red star point indicates the evaluation results every 50 episodes. The blue curve is the average reward of each episode, which plunged when the action is unreasonable. The yellow curve denotes the return of the episode.

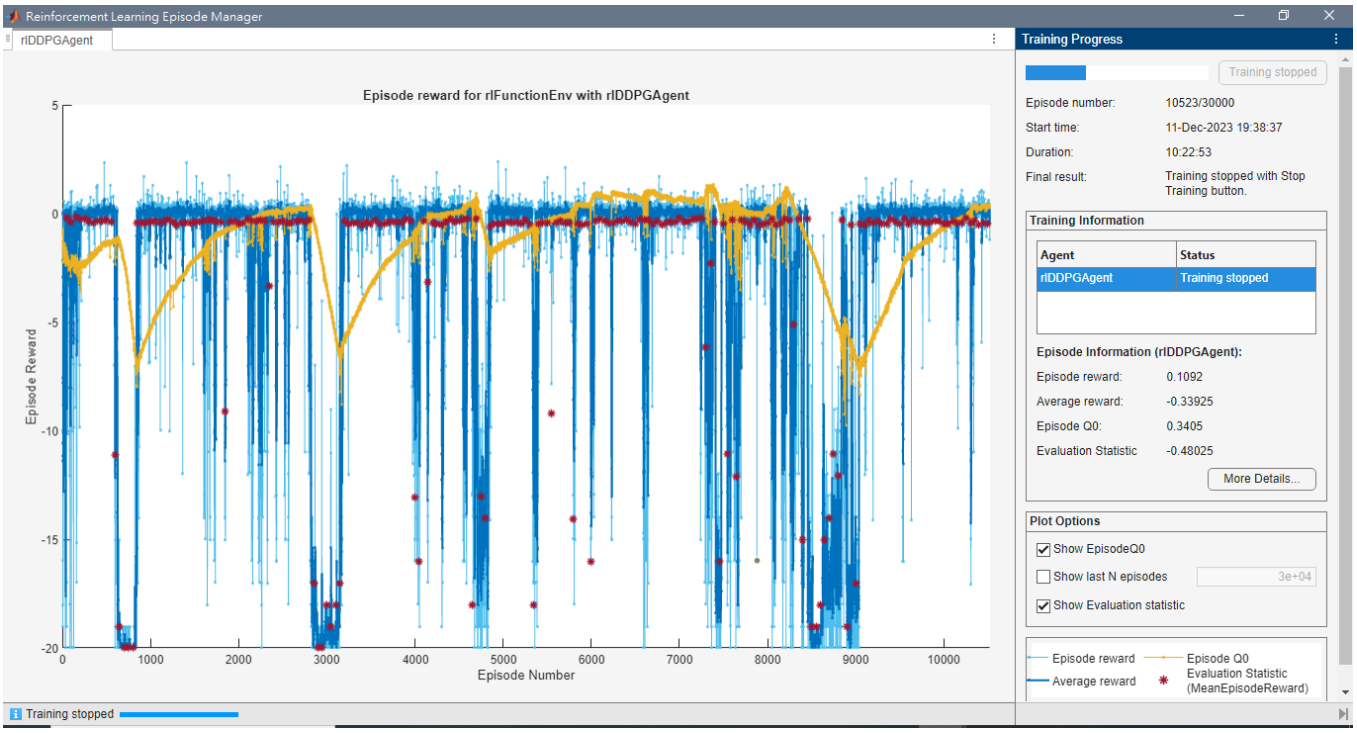

Figure 6. Illustration of the learning process of the trained DDPG agent in this project in terms of reward.

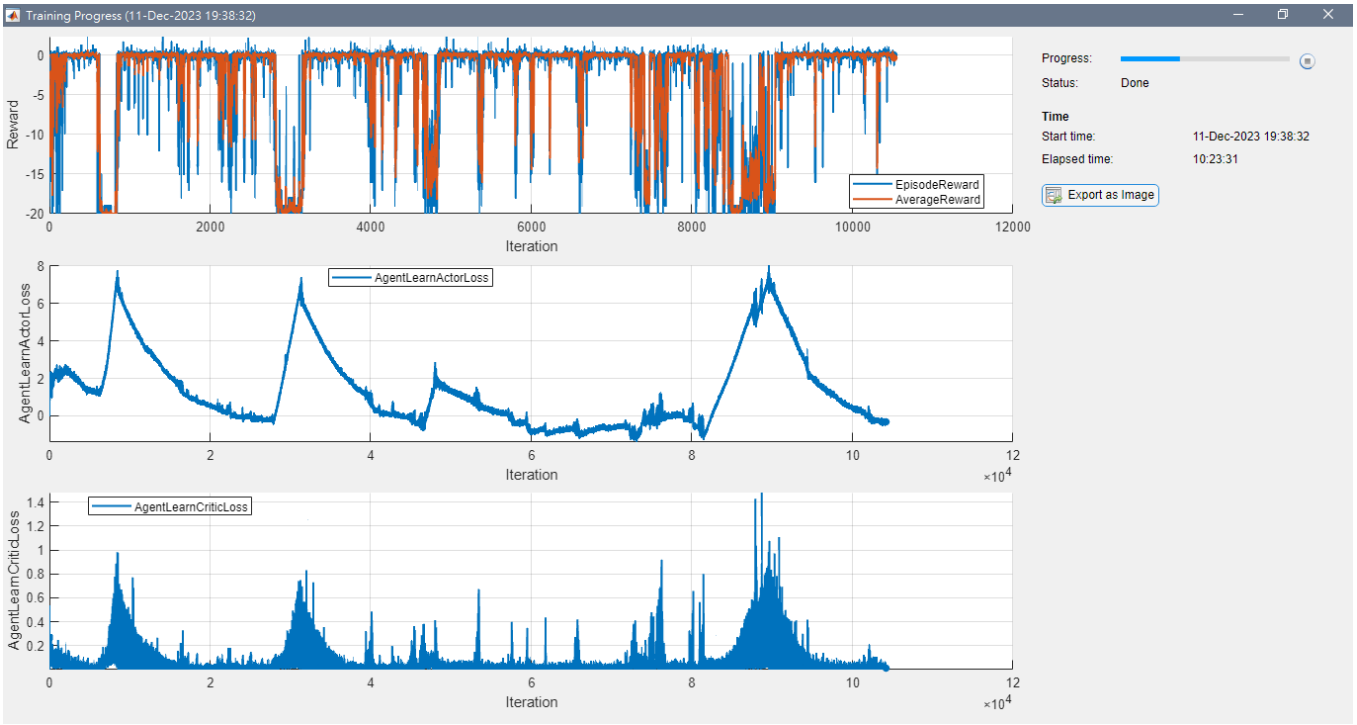

Figure 7. Illustration of the learning process of the trained DDPG agent in this project in terms of loss.

The illustration for the loss function of the learning process is shown in figure 7. The loss function of the actor network and critic network surged once the action is unreasonable. After meeting the unreasonable action, the agent adapted on the network which reduced the loss. The final losses of the two networks are almost close to zero when it stopped by me.

To benchmark the performance of the trained DDPG agent, two popular metrics, maximum drawdown (MDD) and Sharpe ratio (SR), are widely used in stock market [7].

Maximum drawdown (MDD) represents the loss from the highest price to a lowest price. The formula for MDD is written below.

$$MDD = \frac{Troguh\ Value - Peak\ Value}{Peak\ Value} * 100\% \quad (5)$$

where the trough value must happen after the date of the peak value. The lower MDD value means the less loss will suffer.

Sharpe ratio (SR) measures a risk adjusted return and is defined as

$$SR = \frac{R_p - R_f}{\sigma_p} \quad (6)$$

where R_p and R_f denotes the return of a portfolio and the risk-free rate, respectively. The σ_p is the standard deviation of the portfolio's excess return. The higher SR value represents the better return compared to risk-free asset. In this project, I assume the risk-free rate is zero. In a typical case, the risk-free asset is money. Some researchers set the risk-free rate to be the U.S. federal funds interest rate.

Table 1 shows the MDD and SR for both VOO and the trained DDPG agent in different trading periods. The better one among each metric will be indicated in boldfaced. The results of table 1 indicate the trained DDPG agent in this project has better MDD over the four trading periods. The SR of VOO has a 0.2064 gap with the SR of the trained DDPG agent in the trading period starting from January 3, 2022, while the SR of the VOO and the trained DDPG are very close to each other in the other trading periods.

The behaviors of the learned policy are illustrated from figure 7 to figure 10, where each figure is end at

the trading date of November 20, 2023. The orange curve denotes the VOO closing price in each trading day. The dashed blue curve represents the original asset of the agent. The solid blue curve indicates the asset value of the agent in each trading day. The tendencies shown in figure 7 to figure 10 are consistent with the results shown in table 1.

Table 1. The MDD and SR of the testing dataset for different start trading day until November 20, 2023.

	MDD for VOO	MDD for DDPG	SR for VOO	SR for DDPG
January 4, 2021	-25.3%	**-22.9%**	**1.2938**	1.2880
January 3, 2022	-25.3%	**-20.6%**	-0.0838	**0.1226**
September 30, 2022	-22.1%	**-20.0%**	1.3991	**1.4440**
January 3, 2023	-17.1%	**-14.6%**	**1.2750**	1.2356

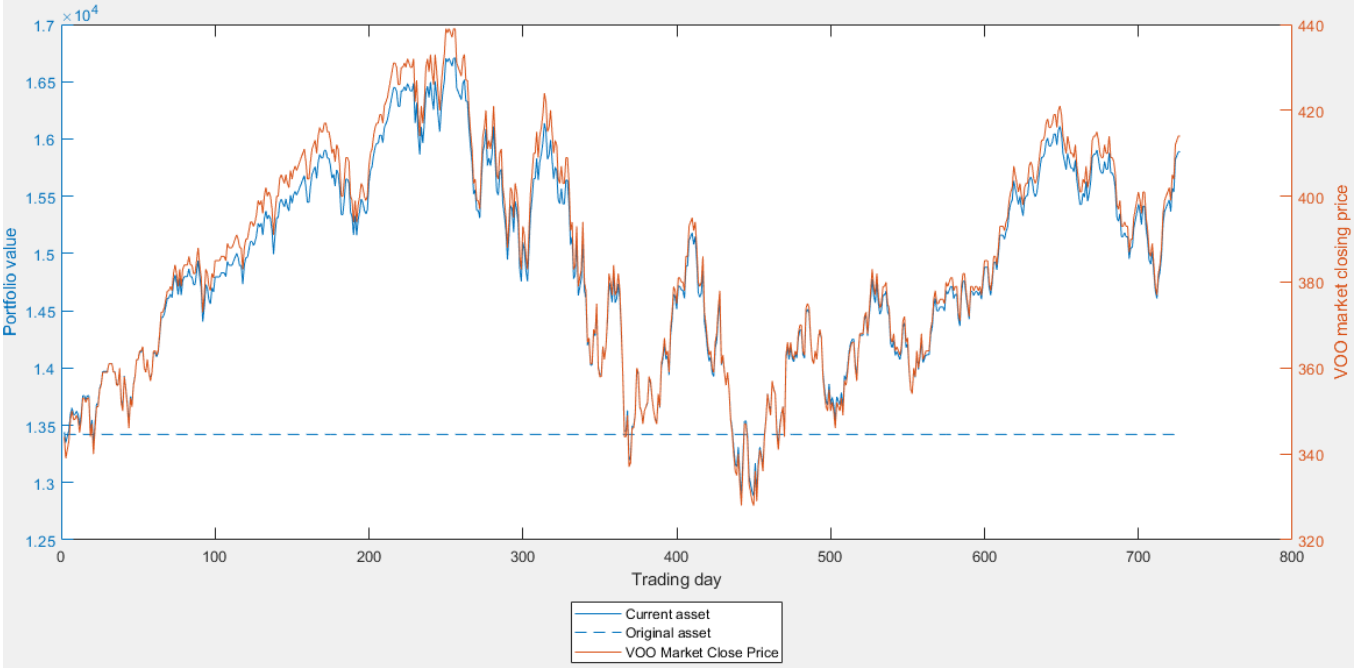

Figure 7. Illustration of the behavior of the learned policy of the trained DDPG agent in the trading period starting from January 4, 2021 to November 20, 2023.

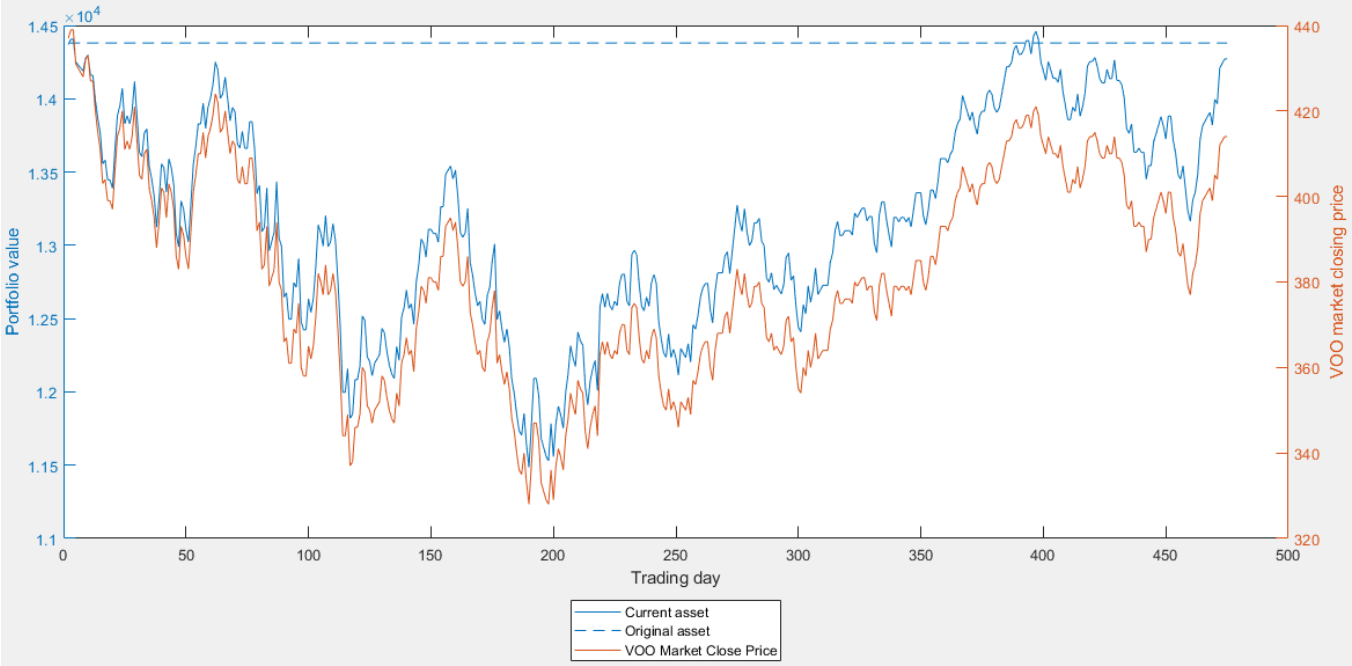

Figure 8. Illustration of the behavior of the learned policy of the trained DDPG agent in the trading period starting from January 3, 2022 to November 20, 2023.

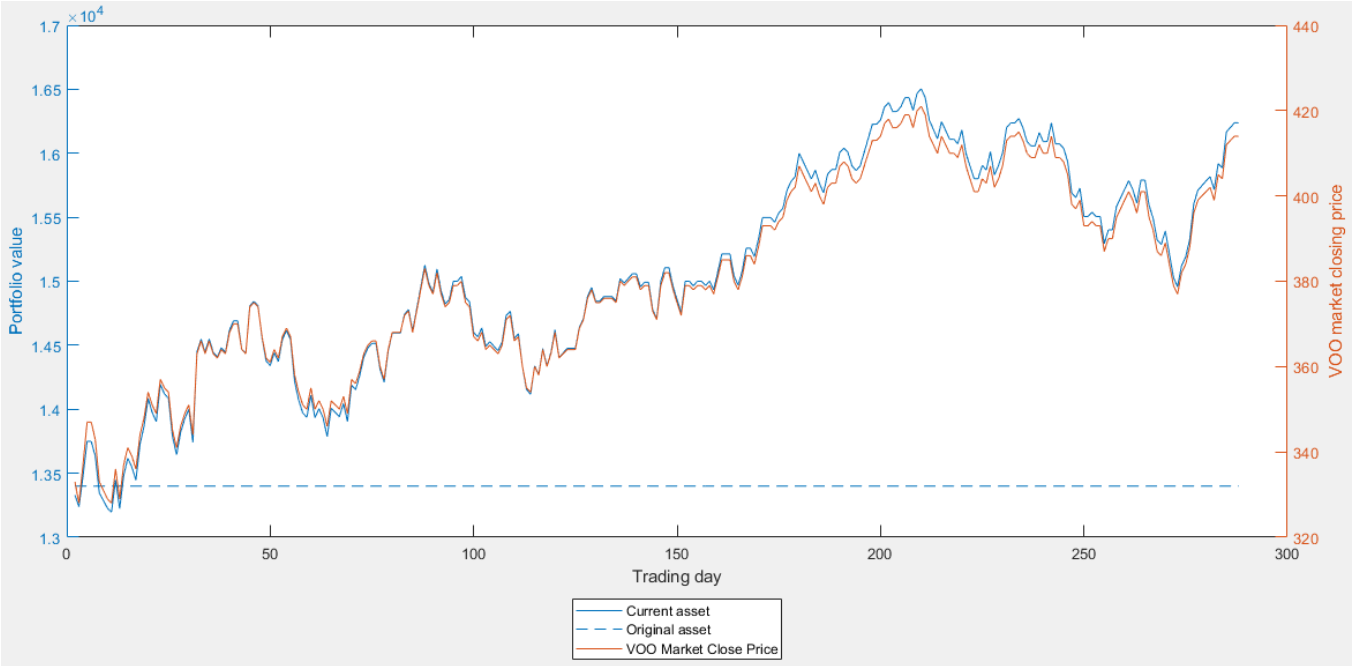

Figure 9. Illustration of the behavior of the learned policy of the trained DDPG agent in the trading period starting from September 30, 2022 to November 20, 2023.

The trained DDPG agent performed better than the VOO in the trading period starting from January 3,

2022 to November 20, 2023, as we can see there is an obvious gap between these two curves in figure 8. Based on the results in table 1 and figure 7 to figure 10, the trained DDPG performs better than the VOO in terms of MDD. The trained DDPG also shows a competitive performance in terms of SR even there are two trading periods have a slightly lower SR.

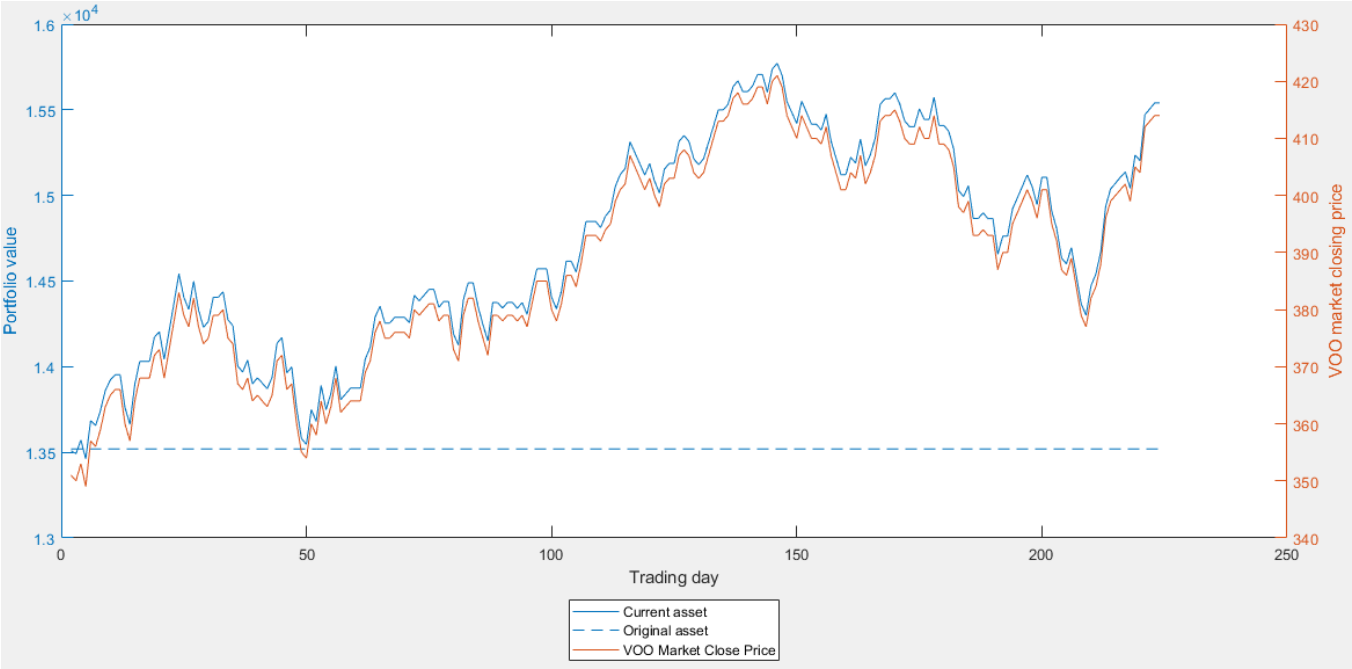

Figure 10. Illustration of the behavior of the learned policy of the trained DDPG agent in the trading period starting from January 3, 2023 to November 20, 2023.

6. Conclusion

Compared to Vanguard 500 Index Fund ETF, the proposed DDPG agent has 2.1% to 4.7% improvement in maximum drawdown by using the closing price, technical indicators, and economic indicators. In terms of the Sharpe ratio, the proposed DDPG agent achieved a competitive performance. The future improvement in this study is to apply the stochastic policy and to fine-tune the size of the actor network and critic network.

References

[1] O Bustos, A. Pomares-Quimbaya, "Stock market movement forecast: A Systematic review," *Expert Systems with Applications*, Volume 156, 2020.

[2] Htun HH, Biehl M, Petkov N. "Survey of feature selection and extraction techniques for stock market prediction," *Financial Innovation*, 9(1): 26., 2023.

[3] Alhomadi, Abraham, "Forecasting Stock Market Prices: A Machine Learning Approach," Utah State University, Logan, State of Utah, All Graduate Plan B and other Reports 1610, 2021.

[4] Nti, I.K., Adekoya, A.F. & Weyori, B.A. "A systematic review of fundamental and technical analysis of stock market predictions," *Artificial Intelligence Review*, 53, 3007–3057, 2020.

[5] Yuh-Jong Hu and Shang-Jen Lin, "Deep Reinforcement Learning for Optimizing Finance Portfolio Management," in *2019 Amity International Conference on Artificial Intelligence (AICAI)*, Dubai, United Arab Emirates, 2019, pp. 14-20

[6] Lillicrap, Timothy P., et al. "Continuous control with deep reinforcement learning." in *2016 International Conference on Learning Representations (ICLR)*, Caribe Hilton, San Juan, Puerto Rico.

[7] Mohammad Amin Masoudi, "Robust Deep Reinforcement Learning for Portfolio Management," M.S. thesis, Master of Science in Management and University of Ottawa, 2021. [Online]. Available: https://ruor.uottawa.ca/bitstream/10393/42743/1/Masoudi_Mohammad_Amin_2021_thesis.pdf