Hindawi
Complexity
Volume 2019, Article ID 3582516, 20 pages
https://doi.org/10.1155/2019/3582516

WILEY | Hindawi

Research Article

Optimizing the Pairs-Trading Strategy Using Deep Reinforcement Learning with Trading and Stop-Loss Boundaries

Taewook Kim [1,2] **and Ha Young Kim** [3]

[1] *Qraft Technologies, Inc., Ttukseom-ro 1-gil, Sungdong-gu, Seoul 04778, Republic of Korea*
[2] *Department of Financial Engineering, Ajou University, Worldcupro 206, Yeongtong-gu, Suwon 16499, Republic of Korea*
[3] *Graduate School of Information, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul, 03722, Republic of Korea*

Correspondence should be addressed to Ha Young Kim; haimkgetup@gmail.com

Received 6 February 2019; Revised 14 April 2019; Accepted 11 June 2019; Published 12 November 2019

Guest Editor: Benjamin M. Tabak

Many researchers have tried to optimize pairs trading as the numbers of opportunities for arbitrage profit have gradually decreased. Pairs trading is a market-neutral strategy; it profits if the given condition is satisfied within a given trading window, and if not, there is a risk of loss. In this study, we propose an optimized pairs-trading strategy using deep reinforcement learning—particularly with the deep Q-network—utilizing various trading and stop-loss boundaries. More specifically, if spreads hit trading thresholds and reverse to the mean, the agent receives a positive reward. However, if spreads hit stop-loss thresholds or fail to reverse to the mean after hitting the trading thresholds, the agent receives a negative reward. The agent is trained to select the optimum level of discretized trading and stop-loss boundaries given a spread to maximize the expected sum of discounted future profits. Pairs are selected from stocks on the S&P 500 Index using a cointegration test. We compared our proposed method with traditional pairs-trading strategies which use constant trading and stop-loss boundaries. We find that our proposed model is trained well and outperforms traditional pairs-trading strategies.

1. Introduction

Pairs trading is a method for obtaining arbitrage profit when there is a statistical difference between two stocks with similar characteristics that are cointegrated or highly correlated. This is possible because of the statistical reason that spreads made by two stocks have a mean reversion in the long run [1]. In the early days, pairs-trading methods were popular because of the opportunity to obtain arbitrage profit [1–4]. However, as many investors including hedge funds sought these arbitrage opportunities by executing the pairs-trading strategy, its profitability began to deteriorate [5, 6]. To overcome these shortcomings, significant research has been conducted to improve the pairs-trading strategy [7–10].

The mechanism of pairs trading is as follows. First, a pair of stocks with similar trends is identified. Second, regression analysis such as ordinary least squares (OLS), total least squares (TLS), and error correction models (ECM) is used to calculate the spread of these stocks. Finally, if the spread hits preset boundaries, investors will open a portfolio which takes a long position on the undervalued stock and shorts the overvalued stock. Subsequently, if the spread reverses to the mean, investors will close the portfolios which are opposite position to the open portfolio. In this case, the investor obtains an arbitrage profit by executing this strategy. However, there is a risk when the spread does not reverse to the mean. In such a situation, investors are at high risk because they cannot close the portfolio. By setting a stop-loss boundary, investors can hedge the risk [11–13].

Many researchers have applied various statistical methods to improve the efficiency and performance of pairs trading. In particular, they focused on using the spread as a trading signal. The study in [1] collected pairs of stocks based on minimizing the sum of squared deviations between the two stocks and then executed the trading strategy if the difference between the pairs is twice the standard deviation of the spread. They used normalized US stock price data from 1962 to 2002 to test the profitability of pairs trading. The

study in [14] used the cointegration approach to protect the pairs-trading strategy from severe losses. They applied an OLS method to create a spread and set various conditions that translated into trading actions. From these models, they achieved a trading strategy with a minimum level of profits protected from risk of loss. The results showed about an 11% annualized excess return over the entire period. The research in [15] compared the distance and cointegration approaches for each high-frequency and daily dataset to check whether it is profitable for Norwegian seafood companies. The performance is similar between two approaches. Reference [16] used a Kalman filter to calculate spread, which was then used as a high-frequency trading signal, on the shares constituting the KOSPI 100 Index. He found that the pairs-trading strategy's performance was significant on the KOSPI and was better during daily market conditions at market opening and closing. Moreover, [7] optimized a pairs-trading system as a stochastic control problem. They used the Ornstein-Uhlenbeck process to calculate spread as a trading signal and tested their model with simulated data; the results showed that their strategy performs well. In addition, [17] suggested the Ornstein-Uhlenbeck process to make a market microstructure noise used as a trading signal in pairs trading strategy. The performance is better under this method than in traditional estimators such as ARIMA(1,1) and maximum likelihood. Reference [18] applied a cointegration method to Chinese commodity futures from 2006 to 2016 to check whether pairs trading was suitable in that market. They used OLS regression to create spreads from the pairs. Furthermore, [10] applied a cointegration test to assorted pairs of stocks and a vector error-correction model to create a trading signal.

It is important to set a boundary to optimize the pairs-trading strategy. This boundary is a criterion for deciding whether to execute a pairs-trading strategy. If a low boundary is set, many strategies will be executed, but profits will be lower; if a high boundary is set, investors will get high returns when the strategy is executed. However, all this assumes that mean reversion occurs. If the spread does not return to the average in the specified trading window, losses will be incurred. If a low boundary is set, the loss will be small. However, if the strategy is executed with a high boundary, the loss will increase. Therefore, the performance of pair trading depends on how the boundary is set. Reference [14] suggested taking a minimum-profit condition, which could be efficient to reduce losses in a pairs-trading system. They set a trading rule with a diverse open condition: for example, if the spread is above 0.3, 0.5, 0.75, 1.0, and 1.5 standard deviations. They used the daily closing prices from January 2, 2001, to August 30, 2002, of two stocks, the Australia New Zealand Bank and the Adelaide Bank. The results showed that, as the open condition value decreases, the number of trades and profits increases. Also [19] suggested optimal preset boundaries calculated from estimated parameters for the average trade duration, intertrade interval, and number of trades and used them to maximize the minimum total profit. They used the daily closing price data from January 2, 2004, to June 30, 2005, of seven pairs of stocks on the Australian Stock Exchange. The results showed that their proposed method was efficient in making profits using the pairs-trading strategy. Reference [18] examined whether the pairs-trading strategy could be applied to the daily return of Chinese commodity futures from 2006 to 2016 using three methods: classical, closed-loop, and dynamic stop-loss. The closed-loop method takes only a stop-profit barrier which executes the strategy and does not consider the risk if spreads revert to the mean. The classical method adds stop-loss boundaries to the closed-loop method. The dynamic stop-loss method uses a variety of stop-profit and stop-loss barriers to fit the spreads if the spread is larger than the standard deviation, which is set using criteria based on the historical average of spreads. The results showed that these methods obtained an annualized return of over 15%, especially the closed-loop method, which yielded the highest profit of 26.94%. In addition, [20] experimented with fixed optimal threshold selection, conditional volatility, percentile, spectral analysis, and neural network thresholds in pairs-trading strategy. Of these, the neural network threshold has outperformed all other strategies.

Following the success of reinforcement learning, demonstrated by its successful performance at Atari games [21], many researchers have attempted to apply this algorithm to the financial trading system. Reference [22] proposed a deep Q-trading system using reinforcement learning methods. They applied Q-learning to a trading system to trade automatically. They set a delta price using data from the past 120 days, had three discrete action spaces (buy, hold, and sell), and used long-term profit as a reward. They used daily data from January 01, 2001, to December 31, 2015, of the Hang Seng Index and the S&P 500 Index. The experimental results showed that their proposed method outperformed buy-and-hold strategies and recurrent reinforcement learning methods. Reference [23] proposed three steps to apply reinforcement learning to the financial trading system. First, they reduced relative replay size to fit financial trading. Second, they proposed an action-augmentation technique that provides more feedback from the action to the agent. Third, they used long sequences as reinforcement data to conduct recurrent neural network training. The experimental data comprised tick-by-tick data of 12 forex currency pairs from January 2012 to December 2017. The results showed that the action-augmentation technique yielded more profit than an epsilon-greedy policy. Reference [10] used an N-armed bandit problem to optimize the pairs-trading strategy. They took the spread using an error-correction model and found the parameters using a grid-search algorithm. They compared their proposed model with a constant parameter model, which was similar to a traditional pairs-trading strategy. They used intraday one-minute data of some stocks in the FactSet database from June 2015 to January 2016. The performance of their proposed model was better than the constant-parameter model.

We investigate not only the dynamic boundary based on a spread in each trading window—which can achieve higher profit than the fixed boundary used in traditional pairs trading strategy—but also if it is possible to train deep reinforcement learning methods to follow this mechanism. To this end, we propose a new method to optimize the pairs trading strategy using deep reinforcement learning, especially deep Q-networks, since pairs trading strategy can

be thought of as a game. After opening a portfolio position, the profit can be set whether portfolio is closed, stop-loss position. Therefore, if we set this strategy as a game by setting boundaries which are optimized in spreads in trading window, we can achieve more profit than traditional pairs trading strategies. In particular, we set the pairs-trading system to be a kind of game and obtain the optimal boundaries, trading thresholds, and stop-loss thresholds according to the calculated spread. The reason for this construction is that if the portfolio is opened and closed in the trading window in the calculated spread, it will be unconditionally profitable if the portfolio is closed. If the portfolio reaches the stop-loss boundary or does not converge to the mean, losses may occur. We therefore set the DQN to learn by positively rewarding it if it takes a closed position and negatively rewarding it if it reaches the stop-loss or exit thresholds. We conducted the following experiments to verify that our proposed method is optimized compared to the conventional method. First, we used different spreads calculated using OLS and TLS to see how the results differ depending on the spread used for input. Second, depending on the formation window and trading window, the spread and hedge ratio will be varied. We therefore set a total of six window sizes for selecting the optimal window size which had the best performance. Finally, we compared the proposed method with the traditional pairs-trading strategy using the test data with the optimal window size. In this experiment, we use the daily adjusted closing prices from January 2, 1990, to July 31, 2018, of 50 stocks in the S&P 500 Index. Experimental results show that our proposed method outperforms the traditional pairs-trading strategy across all the pairs. In addition, we can confirm that the performance measure varies according to the spread.

The main contributions of this study are as follows. First, we propose a novel method to optimize pairs trading strategy using deep reinforcement learning, especially deep Q-networks with trading and stop-loss boundaries. The experimental results show that our method can be applied in the pairs trading system and also to various other fields, including finance and economics, when there is a need to optimize a rule-based strategy to be more efficient. Second, we propose an optimized dynamic boundary based on a spread in each trading window. Our proposed method outperforms traditional pairs trading strategy which set a fixed boundary. Last, we find that our method outperforms traditional pairs trading strategy in all pairs based on constituent stocks in S&P 500. Since our method selects optimal boundaries based on spreads, it can be applied to other stock markets such as KOSPI, Nikkei, and Hang Seng. It should be noted that the present work is a part of the Master thesis [24].

The rest of this paper is organized as follows. Section 2 explains the technical background. Section 3 describes the materials and methods. Section 4 shows the results and provides a discussion of the experiments. Section 5 provides our conclusions to this study.

2. Technical Background

2.1. The Traditional Pairs-Trading Strategy. Pairs trading is a representative market-neutral trading strategy which simultaneously longs an undervalued stock and shorts an overvalued stock. This strategy is a form of statistical arbitrage trading that assumes the movements of the prices of the two assets will be similar to previous trends [1]. It follows the assumption that asset prices will return to the long-term equilibrium. This strategy started from the idea that arbitrage opportunities exist when the price gap between two assets expands to or past a certain level. It is also based on the belief that historical price movements will not change significantly in the future.

In Figure 1, the graph drawn in blue is a spread made of two stocks that are cointegrated, the red lines are the trading boundaries, and the green lines are the stop-loss boundaries. When this spread reaches the trading boundaries, the portfolio is opened and only closed when the spread returns to the average. However, losses are incurred when prices reach the stop-loss boundaries after the portfolio is opened and do not return to the average. Furthermore, after the portfolio is opened, if the trading signal is not reversed to mean during the trading window, the portfolio is closed by force; this is called the exit position of the portfolio.

2.1.1. The Cointegration Test. There are many approaches for pair selection such as the discrete approach [11, 25–27], the cointegration approach [10, 16, 27], and the stochastic approach [7, 8]. In this study, we use the cointegration approach to choose pairs which have long-term equilibrium. Generally, a linear combination of nonstationary variables is also a nonstationary relationship. Assume that x_t and y_t have unit roots; as previously mentioned, the linear combination of these variables follows nonstationary conditions.

$$\begin{aligned} x_t &\sim I(1), \\ y_t &\sim I(1) \end{aligned} \tag{1}$$

$$y_t = \alpha + \beta x_t + \varepsilon_t \tag{2}$$

However, it can be a stationary relationship if the nonstationary variables are cointegrated. In this case, this regression must be checked to determine whether it is a spurious regression or cointegrated. Johansen's method is widely used to test for cointegration [28]. In this method, the number of cointegration relations and the parameters of the model are estimated and tested using maximum likelihood estimation (MLE). Since all variables are regarded as endogenous variables, there is no need to select dependent variables and multiple cointegration relationships are identified. In addition, we use MLE to estimate the cointegration relation with the vector autoregression model and to determine the cointegration coefficient based on the likelihood-ratio test. There is therefore an advantage in performing various hypothesis tests related to the estimation of cointegration parameters and the setting of other models when there is cointegration, and not merely to test for cointegration.

2.2. Spread Calculation

2.2.1. Ordinary Least Squares. In regression analysis, OLS is widely used to estimate parameters by minimizing the sum

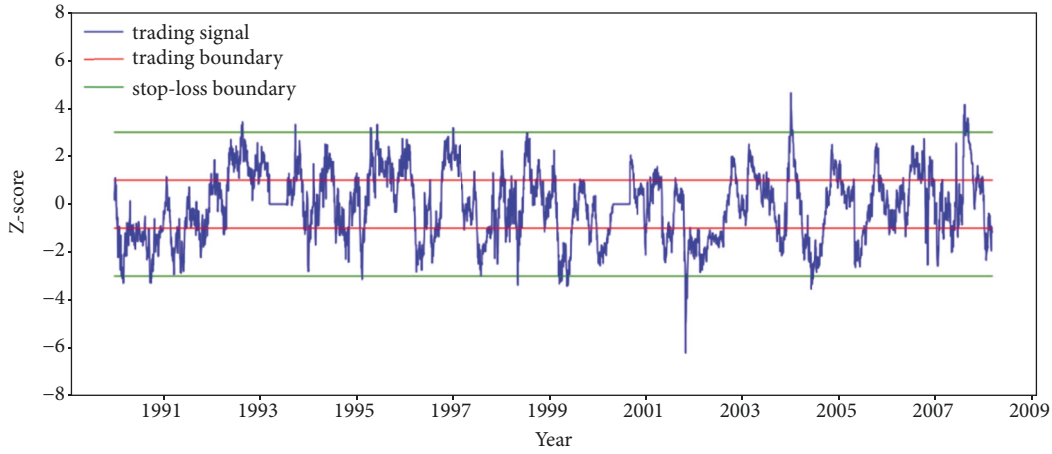


Figure 1: The traditional pairs-trading strategy.

of the squared errors [29]. Assume that x_i, y_i, and ε_i are an independent variable, a dependent variable, and an error term. We can estimate β from the following equation by taking a partial derivative:

$$y_i = \beta x_i + \varepsilon_i \sim N\left(0, \sigma_\varepsilon^2\right) \tag{3}$$

$$\sum_{i=1}^{n} \left(y_i - \beta x_i\right)^2 \tag{4}$$

$$\beta = \left(\sum_{i=1}^{n} x_i' x_i\right)^{-1} \sum_{i=1}^{n} x_i' y_i \tag{5}$$

The value obtained from equation (5) is used for the number of stock orders. The epsilon value is also used as a trading signal through Z-scoring, in the state composed of the formation-window size.

2.2.2. Total Least Squares. TLS estimates parameters to minimize the sum of the measured distance and the vertical distance between regression lines [30]. Since the vertical distance does not change when the X and Y coordinates are changed, the value of β is calculated consistently. In the TLS method, the observed values of X_i and Y_i have the following error terms:

$$Y_i = y_i + e_i \sim N\left(0, \sigma_e^2\right) \tag{6}$$

$$X_i = x_i + u_i \sim N\left(0, \sigma_u^2\right) \tag{7}$$

where x_i and y_i are true values and e_i and u_i are error terms following independent identical distributions. It is assumed that there is linear combination of true values. For convenience, we represent the error variance ratio in equation (10):

$$y_i = \beta_0 + \beta_1 x_i \tag{8}$$

$$Y_i = \beta_0 + \beta_1 x_i + e_i \sim N\left(0, \sigma_e^2\right) \tag{9}$$

$$\tau = \frac{\operatorname{var}\left(Y_i \mid x_i\right)}{\operatorname{var}\left(X_i \mid x_i\right)} = \frac{\sigma_e^2}{\sigma_u^2} \tag{10}$$

The orthogonal regression estimator is calculated by minimizing the sum of the measured distance and the vertical distance between regression lines in equation (11):

$$\sum_{i=1}^{n} \left\{ \frac{\left(Y_i - \beta_0 + \beta_1 x_i\right)^2}{\tau} + \left(X_i - x_i\right)^2 \right\} \tag{11}$$

$$\beta_1 = \frac{s_{YY}^2 - \tau s_{XX}^2 + \left\{\left(s_{YY}^2 - \tau s_{XX}^2\right)^2 + 4\tau s_{XY}^2\right\}^{1/2}}{2s_{XY}} \tag{12}$$

The value obtained from equation (12) is used in the same way as that obtained from equation (5) and the epsilon value is also used as a trading signal through the Z-score in the state composed of the formation-window size.

2.3. Reinforcement Learning and the Deep Q-Network. The idea of reinforcement learning is to find an optimal policy which maximizes the expected sum of discounted future rewards [31]. These rewards come from selecting the optimal value of each action, called the optimal Q-value. Reinforcement learning basically solves the problem defined by the Markov decision process (MDP). It consists of a tuple (S, A, P, R, γ), where S is a finite set of states, A is a finite set of actions, P is a state transition probability matrix, R is a reward function, and γ is a discount factor. In environment ε, agent-observed state s_t at time t, action a_t is selected. From the results of these sequences, environmental feedback is provided to the agent in the form of reward r_t and next state s_{t+1}. An action is selected by the action-value function $Q_\pi(s, a)$ that represents the expected sum of discounted future rewards.

$$Q_\pi\left(s_t, a_t\right) = \mathbb{E}_\pi\left[\sum_{i=t}^{T} \gamma^{i-t} r_i \mid s_t, a_t, \pi\right] \tag{13}$$

In this action-value function $Q_\pi(s_t, a_t)$, we find an optimal action-value function $Q^*(s_t, a_t)$, following an optimal policy

which maximizes the expected sum of discounted future rewards.

$$Q^*(s_t, a_t) = \max_{\pi} Q_\pi(s_t, a_t) \tag{14}$$

This optimal action-value function can be formulated as the Bellman equation.

$$Q^*(s_t, a_t) = \max_{a_{t+1}} [r_t + \gamma Q(s_{t+1}, a_{t+1})] \tag{15}$$

The DQN uses a nonlinear function approximator to estimate the action value function. This network is trained by minimizing a sequence of loss functions $L_t(\theta_t)$, which changes with each sequence of t. The weight of θ_t is updated as the sequence progresses:

$$L_t(\theta_t) = \mathbb{E}_{(s,a)\sim\rho(\bullet)}\left[(y_t - Q(s_t, a_t; \theta_t))^2\right] \tag{16}$$

$$y_t = \max_{a_{t+1}} [r_t + \gamma Q(s_{t+1}, a_{t+1}; \theta_{t-1}) \mid s_t, a_t] \tag{17}$$

3. Materials and Methods

3.1. Data. In this study, 50 stocks from the S&P 500 Index were selected based on their trading volume and market capitalization. To carry out the experiment, the data must cover the same period. Therefore, corresponding stocks were selected, leaving a total of 25 stocks. Table 1 represents the dataset of stock names, abbreviations of those stocks, and their respective sectors. We collected the adjusted daily closing prices using Thomson Reuters' database. The period of the training dataset is from January 2, 1990, to December 31, 2008, comprising 4792 data points; the test dataset covers the period from January 2, 2009, to July 31, 2018, comprising 2411 data points. From these datasets, a pair of stocks will be selected during the training dataset period using the cointegration test.

3.2. Selecting Pairs Using the Cointegration Test. It is necessary to pair stocks which have long-run statistical relationships or similar price movements. It is possible to determine the degree to which two stocks have had similar price movements through the correlation value. Furthermore, the long-term equilibrium of a pair of stocks is an important characteristic for the execution of pairs trading. In this study, we used the cointegration approach to select pairs of stocks. Through Johansen's method, we selected 11 pairs of stocks that have long-run equilibria. Table 2 shows the resulting pairs of stocks that were identified based on t-statistics and Figure 2 shows price movements of the cointegrated stocks XOM and CVX. Using this dataset, we will verify whether our proposed method has better performance than the traditional pairs-trading method.

3.3. Trading Signal. After selecting the pairs, it is necessary to extract the signal for trading. To extract signals, we opt for the OLS or TLS methods. First, because the stock price follows a random walk [32], we need to ensure that it follows the $I(1)$ process through the augmented Dickey-Fuller test. Subsequently, the $I(0)$ process should be created using the logarithmic difference in stock prices which is then applied to the OLS and TLS methods. In equation (18), α_1 is a constant value, β_1 is a hedge ratio (which is used as trading size), ε_t is the error term, and $\log P_{A,t}$ and $\log P_{B,t}$ are the logarithmic differences in the stock prices A and B at time t. We convert values of ε_t into a Z-score used as a trading signal. For example, if the trading signal reaches the threshold, we short one share of the overvalued stock (represented as $\log P_{A,t}$) and long β_1 shares of the undervalued stock (represented as $\log P_{B,t}$). The hedge ratio is determined based on the window size. We set a total of six discrete window sizes to obtain the optimal window size for the experiment. Trading windows are constituted using half of the formation-window size. The spread obtained here is used as a state when applying reinforcement learning (i.e., as an input of the DQN).

$$\log P_{B,t} = \alpha_1 + \beta_1 \log P_{A,t} + \varepsilon_t \tag{18}$$

3.4. Proposed Method: Optimized Pairs-Trading Strategy Using the DQN Method. In this study, we optimize the pairs-trading strategy with a type of game using the DQN. We will attempt to implement an optimal pairs-trading strategy by taking optimal trading and stop-loss boundaries that correspond to the given spread, since performance depends on how trading and stop-loss boundaries are set in pairs trading [14]. Figure 3 shows the mechanism of our proposed pairs-trading strategy. Throughout the cointegration test, we identify pairs and, using regression analysis, obtain a hedge ratio used as trading volume and a spread used as a trading signal and state. In the case of the DQN, two hidden layers are set up and the number of neurons is optimized by taking half of input size through trial and error. Action values consist of the six discrete spaces in Table 3. Each value of a_t has values for trading and stop-loss boundaries.

A pairs-trading system can make a profit if the spread touches the threshold and returns to the average such that the portfolio is closed in each trading window. On the other hand, if the trading boundary is touched and the stop-loss boundary is reached, the system tries to minimize losses by stopping trades. If the spread touches the trading boundary but fails to return to the average, the strategy may end up with a profit or a loss. In this study, the pairs-trading strategy is therefore considered as a kind of game; closing a portfolio yields a positive reward and a portfolio that reaches its stop-loss threshold yields a negative reward. Although an exited portfolio may possibly generate a positive profit, there is also a possibility that losses will occur and it is therefore set to yield a negative reward. We set the other conditions (such as the maintenance of the portfolio or not to execute the portfolio) to zero so as to concentrate on the close, stop-loss, and exit positions.

$$W_t = \left| v_{A,t} \times \frac{S_{A,t'} - S_{A,t}}{S_{A,t}} + v_{B,t} \times \frac{S_{B,t'} - S_{B,t}}{S_{B,t}} \right| t < t' \tag{19}$$

$$R_t = \begin{cases} 1000 \times W_t & \text{if portfolio closed} \\ -1000 \times W_t & \text{if portfolio reaches stop} - \text{loss threshold} \\ -500 \times W_t & \text{if portfolio exited} \end{cases} \tag{20}$$

Table 1: The 25 stocks on the S&P 500 Index used in this study.

No.	Ticker	Stock	Sector
1	AAPL	Apple Inc.	Technology
2	MSFT	Microsoft Corporation	Technology
3	BRKa	Berkshire Hathaway Inc.	Financial Services
4	JPM	JPMorgan Chase & Co.	Financial Services
5	JNJ	Johnson & Johnson	Healthcare
6	XOM	Exxon Mobil Corporation	Energy
7	BAC	Bank of America Corporation	Financial Services
8	WFC	Wells Fargo & Company	Financial Services
9	WMT	Walmart Inc.	Consumer Defensive
10	UNH	UnitedHealth Group Incorporated	Healthcare
11	CVX	Chevron Corporation	Energy
12	T	AT&T Inc.	Communication Services
13	PFE	Pfizer Inc.	Healthcare
14	ADBE	Adobe Systems Incorporated	Technology
15	MCD	McDonald's Corporation	Consumer Cyclical
16	MDT	Medtronic plc	Healthcare
17	MMM	3M Company	Industrials
18	HON	Honeywell International Inc.	Industrials
19	GE	General Electric Company	Industrials
20	ABT	Abbott Laboratories	Healthcare
21	MO	Altria Group, Inc.	Consumer Defensive
22	UNP	Union Pacific Corporation	Industrials
23	TXN	Texas Instruments Incorporated	Technology
24	UTX	United Technologies Corporation	Industrials
25	LLY	Eli Lilly and Company	Healthcare

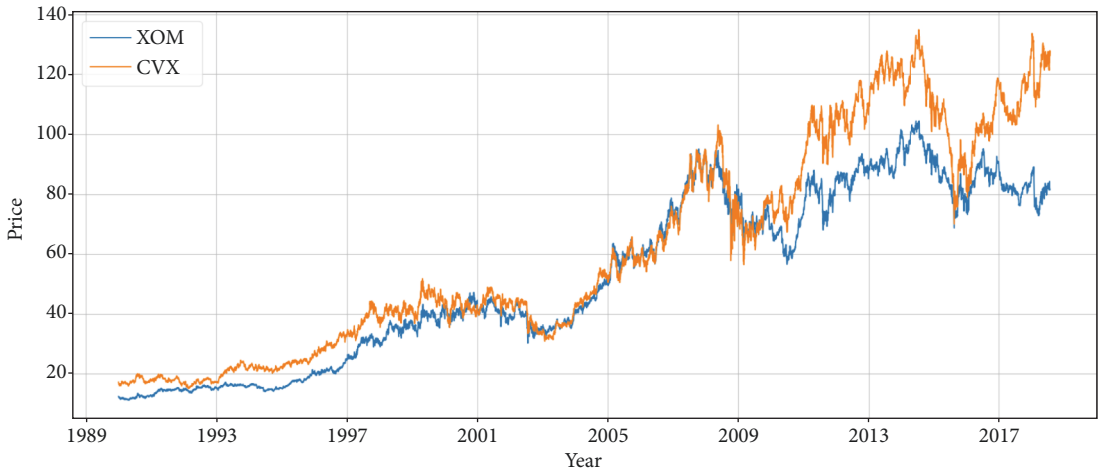


Figure 2: Cointegrated stock price movements.

We fix the values of portfolio close, stop-loss, and exit to +1000, −1000, and −500, respectively. When we update the Q-values, we must consider the reward as a significant component of efficiently training the DQN. We therefore set the reward value to have a range similar to that of the Q-value. Additionally, we included the corresponding profit or loss value to reflect that weight after the trading ended. In equation (19), $v_{A,t}$ and $v_{B,t}$ are the stock orders of stocks A and B at time t, $S_{A,t}$ and $S_{B,t}$ are the stock prices of A and B at time t, and $S_{A,t'}$ and $S_{B,t'}$ are the stock prices of A and B at time t'.

Algorithm 1 shows the process of our proposed method. Before we start our proposed method, we set a replay memory and batch size and select pairs using the cointegration test. At each epoch, we initialized total profit to 1.0. In the training scheme, we set a state which has spreads within the formation window and select actions which are used as

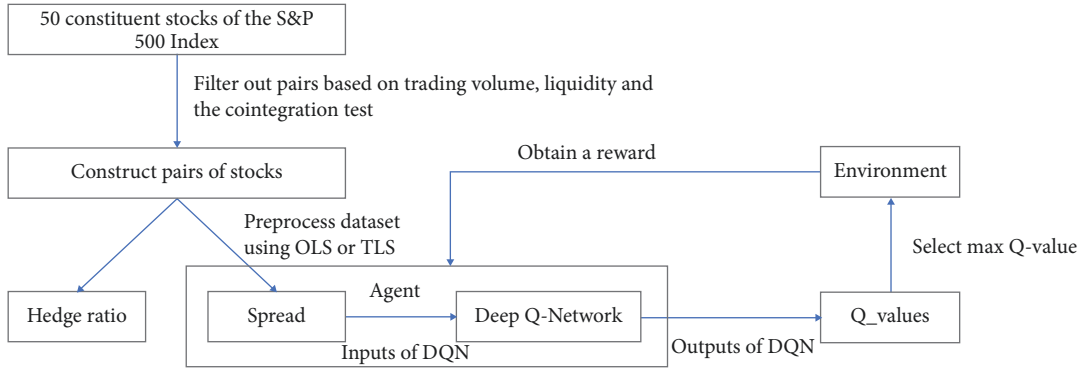


Figure 3: Steps for proposed pairs-trading strategy using the DQN method.

Initialize replay memory $\boldsymbol{D}$ and batch size $\boldsymbol{N}$
Initialize deep Q-network
Select pairs using cointegration test
(1) **For** each epoch **do**
(2) Profit = 1.0
(3) **For** steps t = 1, . . . until end of training data set **do**
(4) Calculate spreads using OLS or TLS methods
(5) Obtain initial state by converting spread to Z-score based on formation window $\boldsymbol{s_t}$
(6) Using epsilon-greedy method, select a random action $\boldsymbol{a_t}$
(7) Otherwise select $\boldsymbol{a_t} = \boldsymbol{argmax_a Q(s_t, a)}$
(8) Execute traditional pairs-trading strategy based on the action selected
(9) Obtain reward $\boldsymbol{r_t}$ by performing the pairs-trading strategy
(10) Set next state $\boldsymbol{s_{t+1}}$
(11) Store transition $(\boldsymbol{s_t, a_t, r_t, s_{t+1}})$ in $\boldsymbol{D}$
(12) Sample minibatch of transition $(\boldsymbol{s_t, a_t, r_t, s_{t+1}})$ from $\boldsymbol{D}$.
(13) $$\boldsymbol{y_t} = \begin{cases} \boldsymbol{r_t} & \boldsymbol{if\ s_{t+1} = terminal} \\ \boldsymbol{r_t + \gamma max_{a'} Q(s_{t+1}, a')} & \boldsymbol{if\ s_{t+1} = non - terminal} \end{cases}$$
(14) Update Q-network by performing a gradient descent step on $\{\boldsymbol{y_t - Q(s_t, a)}\}^2$
(15) **End**
(16) **End**

Algorithm 1: Optimized pairs-trading system using DQN.

trading and stop-loss boundaries. Throughout the trading window, we executed a strategy similar to a traditional pairs-trading strategy using the action selected. After executing the strategy, we obtain a reward based on the results of the portfolio. Finally, for the Q-learning process, we update the Q-networks by performing a gradient descent step.

3.5. Performance Measure. We check our experiment results based on profit, maximum drawdown, and the Sharpe ratio. Profit is commonly used as a performance measure for trading strategies. It is calculated as the sum of returns taking into consideration trading cost. Since many trades can increase total profit, it is necessary to determine the total profit taking into consideration transaction costs depending on trading volume. In this study, we set a trading cost of 5 bp; equation (21) is almost the same as equation (19), but it does not include absolute value, and C is trading cost. Maximum drawdown represents the maximum cumulative loss from the highest to the lowest values of the portfolio during a given investment period where $P(t)$ is the value of the portfolio and T is the terminal time value. The Sharpe ratio is an indicator of the degree of excess profits from investing in risky assets used in evaluating portfolios [33]. In equation (23), R_p is the expected sum of portfolio returns and R_f is the risk-free rate; we set this value to 0 and σ_p is the standard deviation of portfolio returns.

$$Profit = \sum_{t=1}^{T} \left[\left(v_{1,t} * \frac{S_{1,t'} - S_{1,t}}{S_{1,t}} + v_{2,t} * \frac{S_{2,t'} - S_{2,t}}{S_{2,t}} \right) - C * \left(v_{1,t} + v_{2,t} \right) \right] \tag{21}$$

$$MDD(T) = \max_{\tau \in (0,T)} \left(\max_{t \in (0,\tau)} P(t) - p(\tau) \right) \tag{22}$$

Table 2: Summary statistics for pairs verified using cointegration tests.

No.	Pairs	t-statistic	Correlation
1	MSFT/JPM	−3.5423∗∗	0.9165
2	MSFT/TXN	−3.448∗∗	0.8641
3	BRKa/ABT	−3.5148∗∗	0.9493
4	BRKa/UTX	−3.3992∗∗	0.9609
5	JPM/T	−3.5882∗∗	0.8486
6	JPM/HON	−5.8209∗∗∗	0.9250
7	JPM/GE	−3.4494∗∗	0.9105
8	JNJ/WFC	−3.5696∗∗	0.9693
9	XOM/CVX	−4.05∗∗∗	0.9879
10	HON/TXN	−4.0625∗∗∗	0.7469
11	GE/TXN	−3.467∗∗	0.9148

Note: ∗∗∗ and ∗∗ denote a rejection of the null hypothesis at the 1% and 5% significance levels, respectively.

$$Sharpe\ ratio = \frac{R_p - R_f}{\sigma_p} \tag{23}$$

The Materials and Methods section should contain sufficient details so that all procedures can be repeated. It may be divided into headed subsections if several methods are described.

4. Results and Discussion

We use the stock pair XOM and CVX, which rejects the null hypothesis at the 1% significance level, to verify whether our proposed model is trained well. The lengths of the window sizes such as the formation window and trading window are selected from the performance results with the training dataset. From these results, we select an optimized window size and compare our proposed model with traditional pairs trading, which takes a constant set of actions with the test dataset.

4.1. Training Results. To find the optimum window size for the optimized pairs-trading system, we experimented with six cases. We performed the experiments based on six window sizes, and the results for each window size are calculated by averaging the top-5 results for a total of 11 pairs. From Tables 4 and 5, we can find that the best performance is obtained when the formation and training windows are 30 and 15, respectively, based on the profit generated by both the OLS and TLS methods. When we trained our networks, we set a positive reward for taking more closed positions and fewer stop-loss and exit positions. We can find the lowest ratio of portfolio closed positions based on the number of open positions, which in the formation and trading windows are for 30 and 15 days (0.68). Contrary to this result, the highest ratios of the number of closed positions in the formation and trading windows are for 120 and 60 days (0.73). However, the highest profits reported in the formation and trading windows are for 30 and 15 days. This can be explained when we check the ratio of the number of stop-loss portfolios. The formation and trading window sizes are 30 and 15 days and the ratio of portfolio stop-loss position is 0.13, but the formation and trading window sizes are 0.20. This result indicates that it is important to reduce the stop-loss position while increasing the closed position. In addition, we can see that the trading signals made with the TLS method are better than those made with the OLS method in all six of the discrete window sizes. The reason for this is based on the difference between the hedge ratios of the two methods. In OLS, when one side is the reference, the relative change of the other side is estimated. Since the assumption is that there is no error component on the reference side and there is an error only on the other side, the hedge ratio varies depending on the side used as the reference. However, in TLS, hedging ratios are the same regardless of which side is used as the reference. For this reason, the experimental results confirm that the TLS method is better able to determine when to execute the pairs-trading strategy. From these results, we take the optimum window size when we verify our proposed method in the test dataset. However, we first need to ensure that the model we proposed is well-trained.

It is important to check whether our reinforcement learning algorithm is trained well. Reference [21] suggested that a steadily increasing average of Q-values is evidence that the DQN is learning well. Figure 4(a) shows the average Q-values of HON and TXN as training progressed. We find that the average Q-values steadily increased, indicating that our proposed model is properly trained. In addition, we provide a positive reward when the portfolio closes and a negative reward when the portfolio reaches the stop-loss threshold or exits. Figure 4(b) shows the ratio of the number of portfolio positions as training progressed. The ratio of closed to open portfolio positions increased and the ratio of portfolios reaching their stop-loss thresholds to open portfolio positions decreased. We also find that the ratio of portfolio exits to open portfolio positions slightly increased. It is possible that the rewards given for an open portfolio position compared to those given for a closed portfolio position are relatively small. The DQN is therefore trained to prevent portfolios from reaching their stop-loss thresholds (the more important objective) over exiting them. This result can also serve as a basis for judging whether the proposed model is being trained properly.

Tables 6 and 7 represent the performance results of XOM and CVX in the training dataset. We call our proposed model pairs-trading DQN (PTDQN) and traditional pairs trading with constant action values as pairs trading with action 0 (PTA0) to pairs trading with action 5 (PTA5). From this result, we can confirm that our proposed method is more profitable than the constant pairs-trading strategies. In addition, we can see that the TLS method has a higher profitability compared to the OLS method. From PTA0 to PTA5, the trading boundary and the stop-loss boundary grew larger; the numbers of open and closed portfolios and portfolios that reached their stop-loss thresholds are reduced. In other words, there is less opportunity for profit, but the probability of loss is also reduced. It is important not only to take a lot of closed positions, but also to take the best action to open and close the portfolio. For example, if a portfolio is

Table 3: Setting a discrete action space.

	Action					
	A0	A1	A2	A3	A4	A5
Trading boundary	± 0.5	± 1.0	± 1.5	± 2.0	± 2.5	± 3.0
Stop-loss boundary	± 2.5	± 3.0	± 3.5	± 4.0	± 4.5	± 5.0

Table 4: Results of applying the DQN method to each window size using OLS.

Formation window	Trading window	MDD	Sharpe ratio	Profit	# of open portfolios	# of closed portfolios	# of stop-loss portfolios	# of portfolio exits
30	15	−0.3682	0.1197	2.7344	328	225	44	58
60	30	−0.3779	0.1327	2.5627	210	147	41	21
90	45	−0.4052	0.1409	2.4112	160	114	34	11
120	60	−0.4383	0.1165	2.0287	134	98	28	8
150	75	−0.4395	0.1244	2.0098	110	80	24	6
180	90	−0.5045	0.1180	1.9390	100	73	21	5

Table 5: Results of applying the DQN method to each window size using TLS.

Formation window	Trading window	MDD	Sharpe ratio	Profit	# of open portfolios	# of closed portfolios	# of stop-loss portfolios	# of exited portfolios
30	15	−0.4422	0.1061	2.9436	320	229	46	44
60	30	−0.5031	0.1143	2.5806	204	144	42	17
90	45	−0.5824	0.1072	2.4588	155	110	36	9
120	60	−0.5768	0.1181	2.4378	136	98	31	6
150	75	−0.5805	0.1245	2.4127	110	79	26	5
180	90	−0.5467	0.1209	2.3570	100	72	23	4

Table 6: Average top-5 performance results for XOM and CVX using OLS within the training period.

Model	MDD	Sharpe ratio	Profit	# of open portfolios	# of closed portfolios	# of stop-loss portfolios	# of exited portfolios
PTDQN	−0.0842	0.1835	3.4068	469	336	64	96
PTA0	−0.2014	0.1452	2.5934	565	382	132	50
PTA1	−0.1431	0.1773	2.7603	409	279	45	84
PTA2	−0.1234	0.1955	2.6307	325	191	16	118
PTA3	−0.2586	0.0861	1.3850	208	86	2	120
PTA4	−0.2591	0.0803	1.1933	124	39	2	83
PTA5	−0.2448	−0.0638	0.8588	47	11	0	36

Table 7: Average top-5 performance results for XOM and CVX using TLS within the training period.

Model	MDD	Sharpe ratio	Profit	# of open portfolios	# of closed portfolios	# of stop-loss portfolios	# of exited portfolios
PTDQN	−0.0944	0.2133	4.8760	541	399	104	63
PTA0	−0.1210	0.1522	4.1948	579	413	125	41
PTA1	−0.1015	0.1650	3.8834	430	310	50	70
PTA2	−0.1483	0.1722	3.3425	320	209	13	98
PTA3	−0.1386	0.1771	2.4385	217	101	3	113
PTA4	−0.1749	0.1602	1.6852	119	38	2	79
PTA5	−0.2862	0.0137	1.0362	55	10	0	45

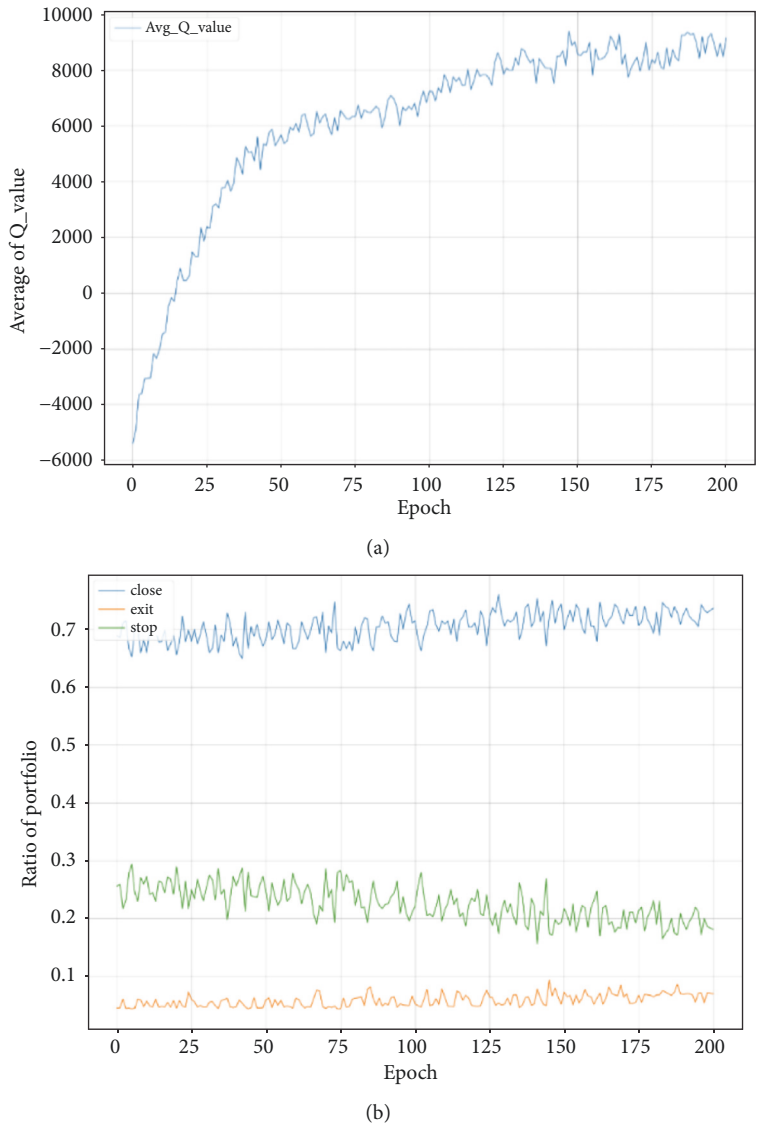


Figure 4: Verification that our proposed model is well-trained with HON and TXN using TLS. (a) Average of Q-values. (b) Ratio of portfolios.

opened and closed by a boundary corresponding to action 0 within the same spread and if a portfolio is opened and closed by a boundary corresponding to action 1, the corresponding profit is different. Assuming that the mean reversion is certain to occur, if we take the maximum boundary condition to open a portfolio, we will obtain a larger profit than when we take a smaller boundary condition. We can see that the PTDQN returns are higher than the strategy with the highest return among the traditional pairs trading strategies that take the constant action. Figures 5–8 show the changes in trading and stop-loss boundaries and the highest profit for constant action when applying the DQN method during the training period using OLS and TLS.

Figures 5 and 6 show comparisons of PTDQN and PTA1 using the TLS method. Figure 5 consists of the spread, trading, and stop-loss boundaries. We find that trading and stop-loss boundaries have different values in PTDQN, showing that it has learned to find the optimal boundary according to each spread. In contrast to PTDQN, PTA1 in Figure 6 has constant trading and stop-loss boundaries. Figures 7 and 8 exhibit the same features we see in Figures 5 and 6. The difference between these methods lies in the spreads: different results can be obtained depending on the spreads used. Making better spreads can therefore improve performance.

Figures 9 and 10 represent the profit corresponding to DQN and constant actions using TLS and OLS. Reference [34] suggested that an average value over multiple trials should be presented to show the reproducibility of deep reinforcement learning because there may be different results from high variances across trials and random seeds. We therefore conducted five trials with different random seeds. The profit graph of DQN represents the average profit of these trials and the filled region between the maximum and minimum profit values. We can see that PTDQN had a higher profit than the traditional pairs-trading strategies during

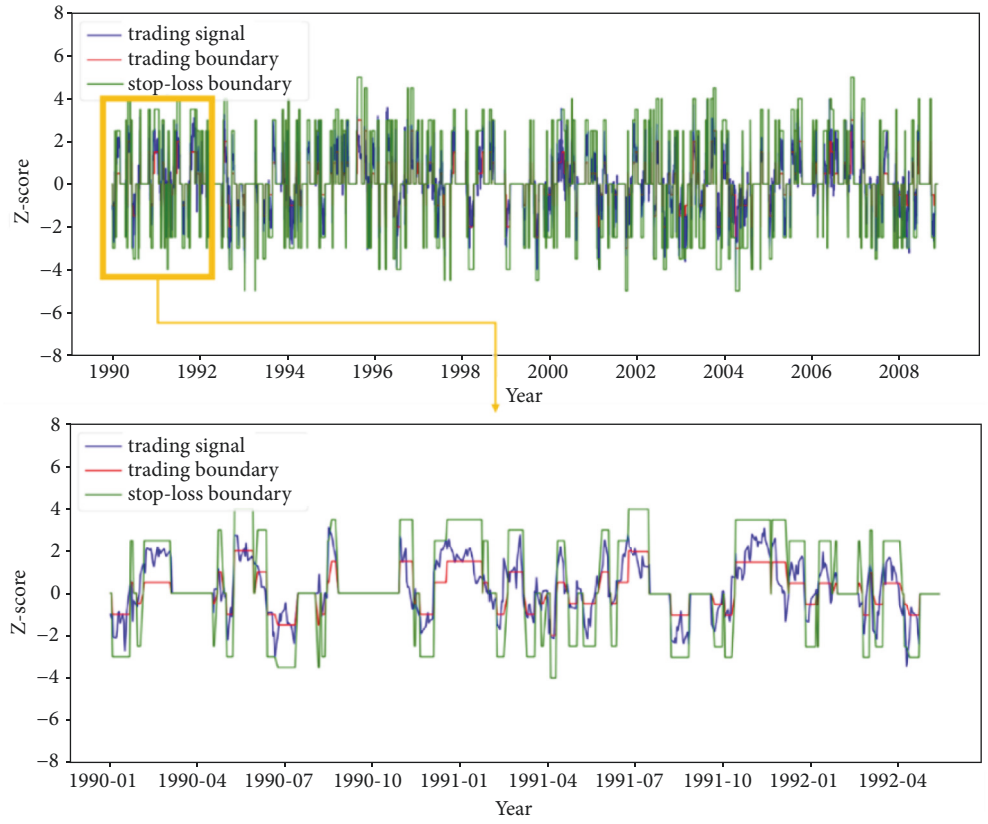


Figure 5: An example of optimizing pairs trading using PTDQN based on a training scheme using TLS.

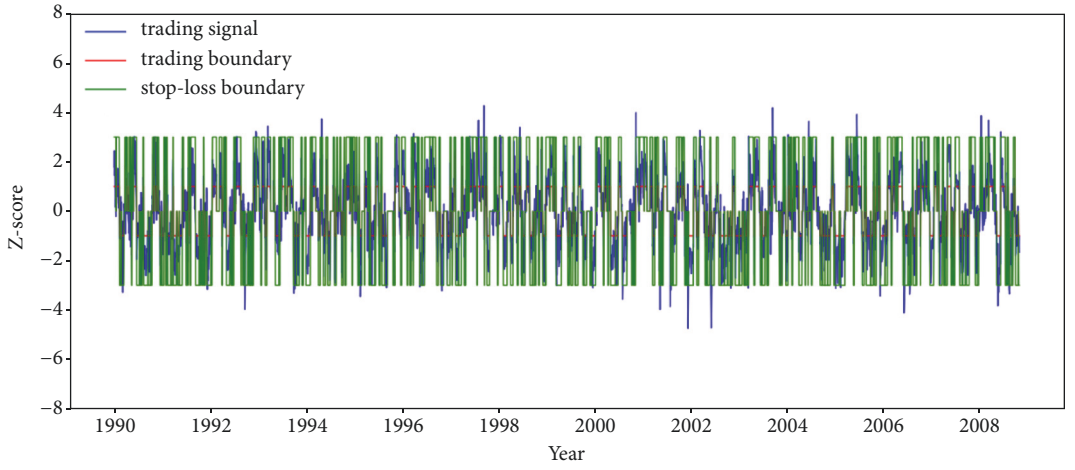


Figure 6: An example of PTA1 based on a training scheme using TLS.

the training period. This means that, even with the same spread, we can see how profit will change as the boundaries are changed. In other words, finding the optimal boundary for the spread is an important factor in optimizing the profitability of pairs trading.

4.2. Test Results. Tables 8 and 9 show the average performance measures of each pair tested by applying the top-5 trained models. We can see that the constant action with the highest returns for each pair is different, and the TLS method is higher in all pairs than the OLS method based on profit, as shown above. We also find that PTDQN has better performance than traditional pairs-trading strategies. The pair with the highest profit using the proposed method is HON and TXN (3.2755); it also shows the biggest difference between the DQN method and the optimal constant action (0.9377). We find that the proposed method has a higher Sharpe ratio in all pairs except for MO and UTX when the

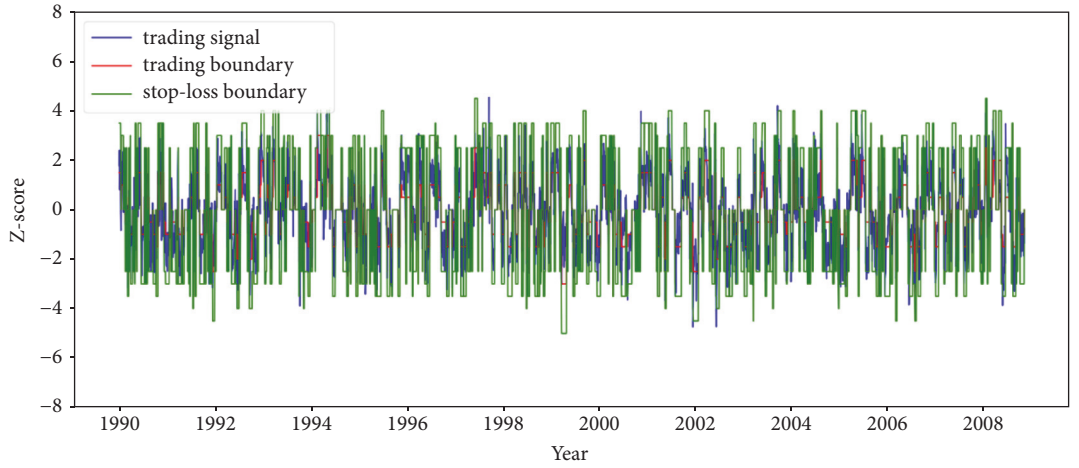


Figure 7: An example of optimizing PTDQN based on a training scheme using OLS.

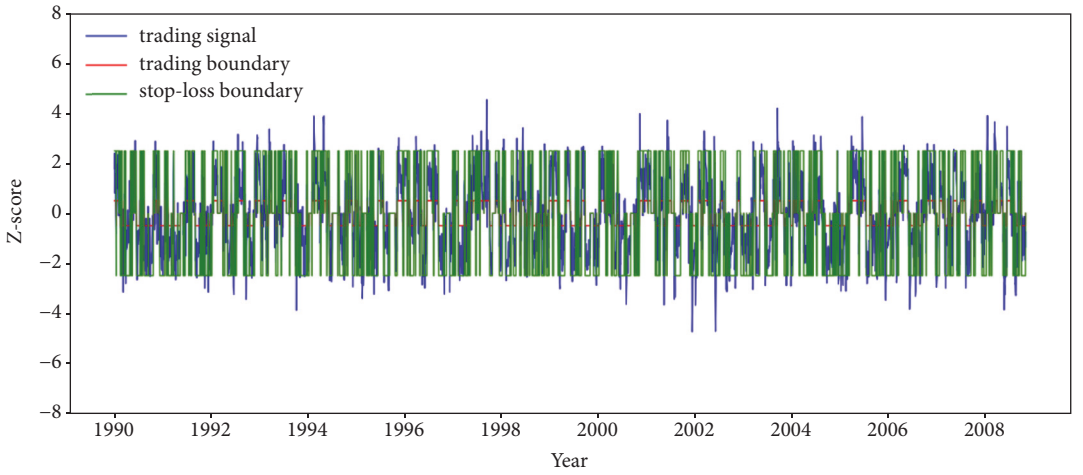


Figure 8: An example of PTA0 based on a training scheme using OLS.

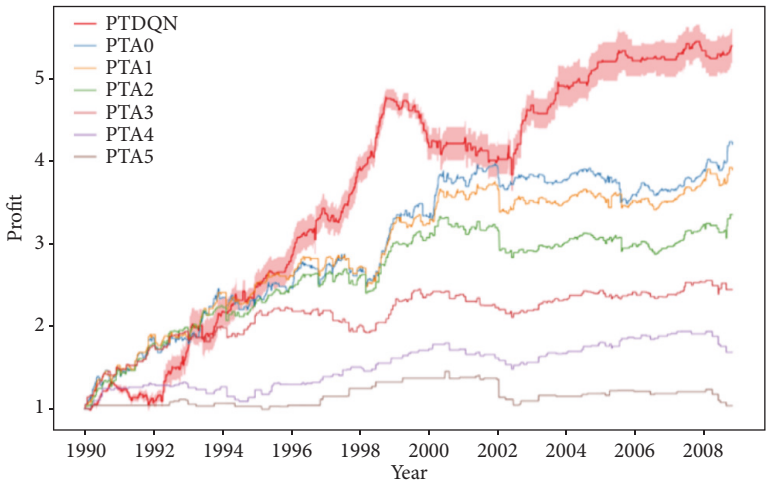


Figure 9: Average top-5 profits generated by PTDQN and traditional pairs-trading strategies using TLS in training periods.

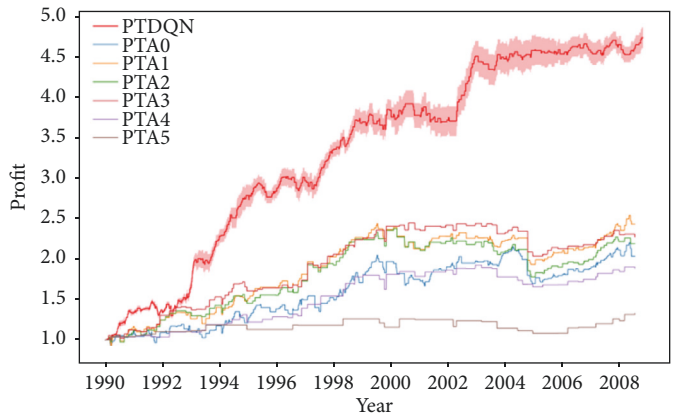


Figure 10: Average top-5 profits generated by PTDQN and traditional pairs-trading strategies using OLS in training periods.

TLS method is used. If we add the Sharpe ratio in addition to the total profit as an objective function, we can build a more optimized pairs-trading system. Based on these results, we can ensure the robustness of our proposed method for our dataset. The proposed method can be applied to other pairs of stocks found in other global markets.

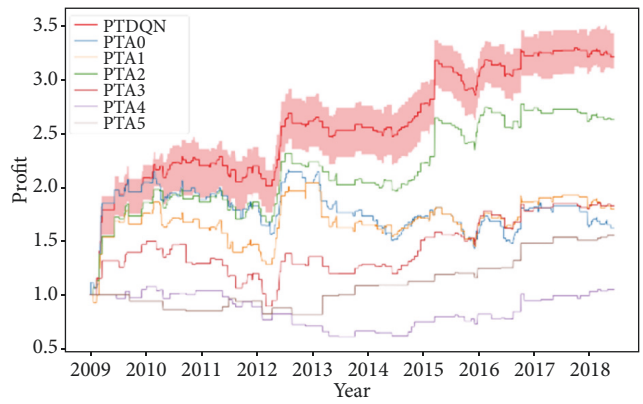
In Figure 11, we can see that our proposed method, PTDQN, outperforms the traditional pairs trading strategies that have constant actions in test dataset. The crucial aspect of this method is the selection of optimal boundary in the spread that makes the highest profit in constant action, which is like a constant boundary. Therefore, the trend is the same as traditional pairs trading strategies; however, when the optimal boundaries which have the highest profit in the spread are combined, PTDQN is found to have higher profit than traditional pairs trading strategies. This method can therefore be applied in various fields when there is a need to optimize the efficiency of a rule-based strategy [35, 36]. In this study, we consider spread and boundaries to be the important factors of pairs trading strategy. Therefore, we tried to optimize pairs trading strategy with various trading and stop-loss boundaries using deep reinforcement learning and our method outperforms rule-based strategies. By optimizing key parameters in rule-based methods, it can improve the performances.

Pairs trading uses two types of stock which have the same trends. However, it can be broken due to various factors such as economic issues and company risk. In this situation, the spread between two stocks is extremely large. Although this situation cannot be avoided, we hedge this risk by taking a dynamic boundary. In this sense, taking the lowest stop-loss boundary is the best choice since it can be overcome with the least loss. By taking the dynamic boundary using the deep reinforcement learning method, we can see that not only profits are increased, but losses are also minimized as compared to taking a fixed boundary.

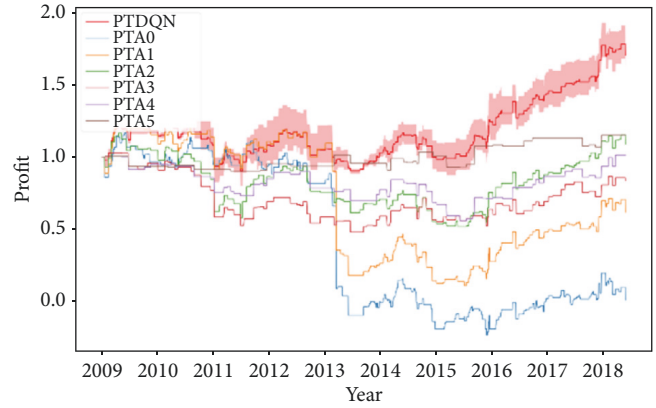
5. Conclusions

We propose a novel approach to optimize pairs trading strategy using a deep reinforcement learning method, especially deep Q-networks. There are two key research questions posed. First, if we set a dynamic boundary based on a spread in each trading window, can it achieve higher profit than traditional pairs trading strategy? Second, is it possible that deep reinforcement learning method can be trained to follow this mechanism? To investigate these questions, we collected pairs selected using the cointegration test. We experimented with how the results varied according to the spread and the method used. We therefore set different spreads using OLS and TLS methods as the input of the DQN and the trading signal. To conduct this experiment, we set up a formation window and a trading window. The hedge ratio, which is an important factor in determining how much stock to take, depends on this value. We therefore applied the OLS and TLS methods and experimented to find the optimal window size by varying the formation window and the trading window.

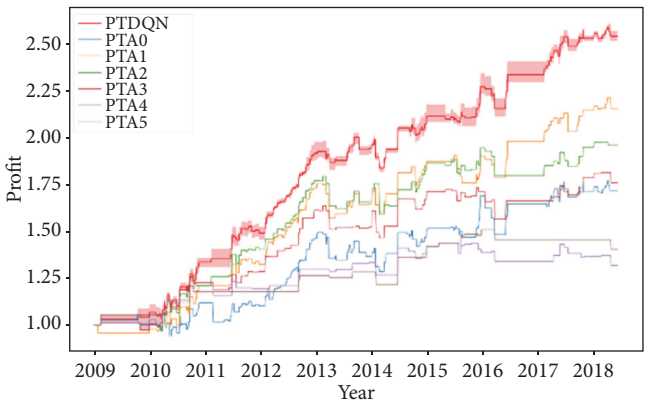
Tables 6 and 7 show the average performance values of the formation windows and trading windows in the training dataset. The results show that all six window sizes were higher when TLS spreads were used than in OLS spreads. In addition, we can see that profitability gradually increases as the estimation windows and trading windows of methods using TLS and OLS decreased. The reason is that although the ratio of closed position portfolio is the lowest in what we set formation and trading windows, the ratio of stop-loss position portfolio is also the lowest compared with other formation and trading windows. It means that reducing stop-loss position portfolio is important as well as increasing closed position portfolio to make a profit. Using the optimal window size, we then check whether our DQN is properly trained. At each epoch, we find that the average Q-value steadily increased, the ratio of closed portfolios increased, and the ratio of portfolios that reached their stop-loss thresholds decreased, confirming that our DQN is trained well. Based on these results, we find that our proposed model using the test dataset with a formation window of 30 and a trading window of 15 had results that were superior to those of traditional pairs-trading strategies in the out-of-sample dataset. In Figure 11, we can see that the profit path of PTDQN is similar PTA0 to PTA5, but better than that from



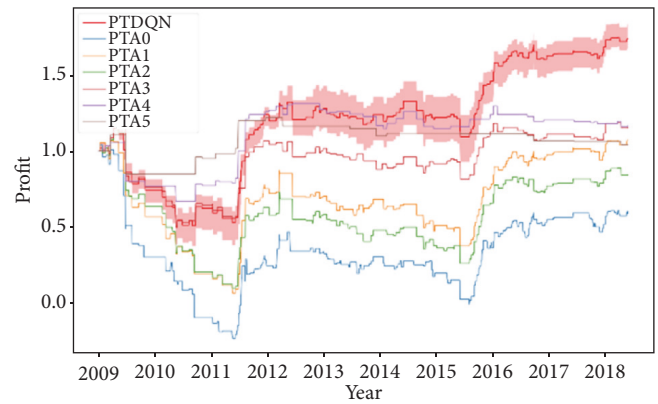
(a) MSFT/JPM



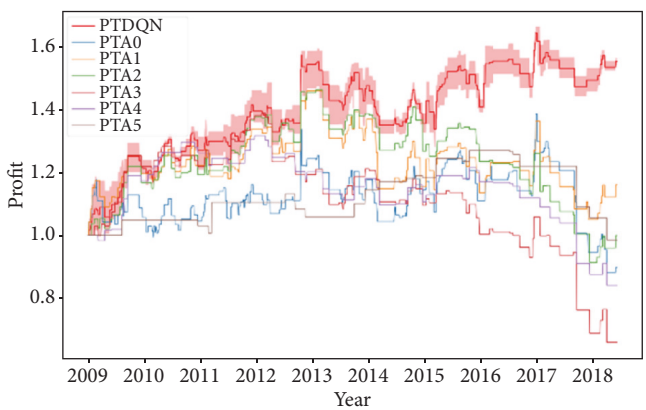
(b) MSFT/TXN



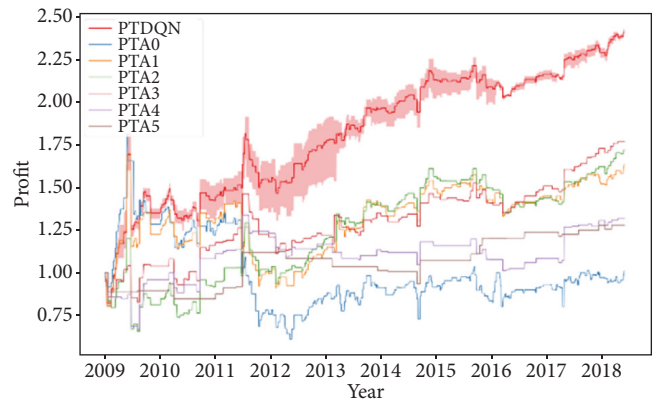
(c) BRKa/ABT



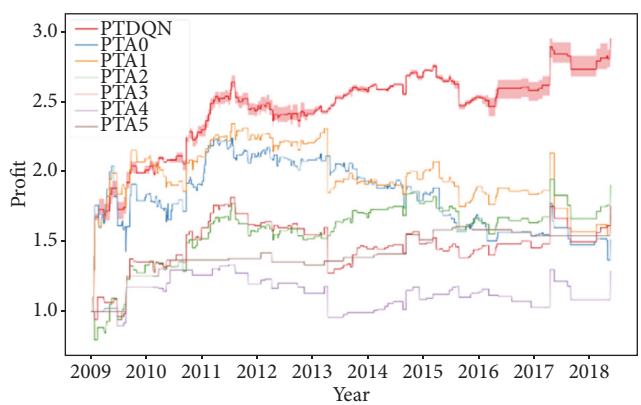
(d) BRKa/UTX



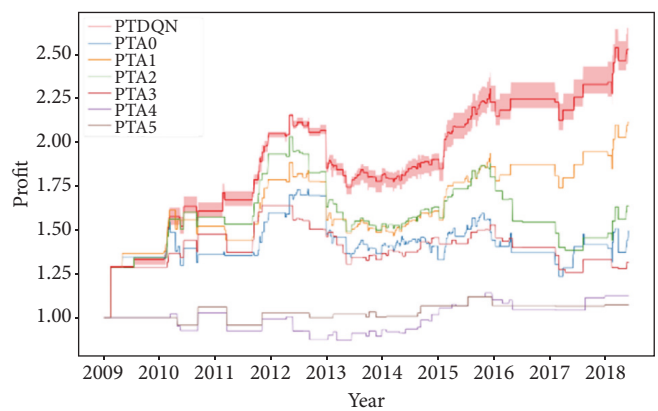
(e) JPM/T



(f) JPM/HON

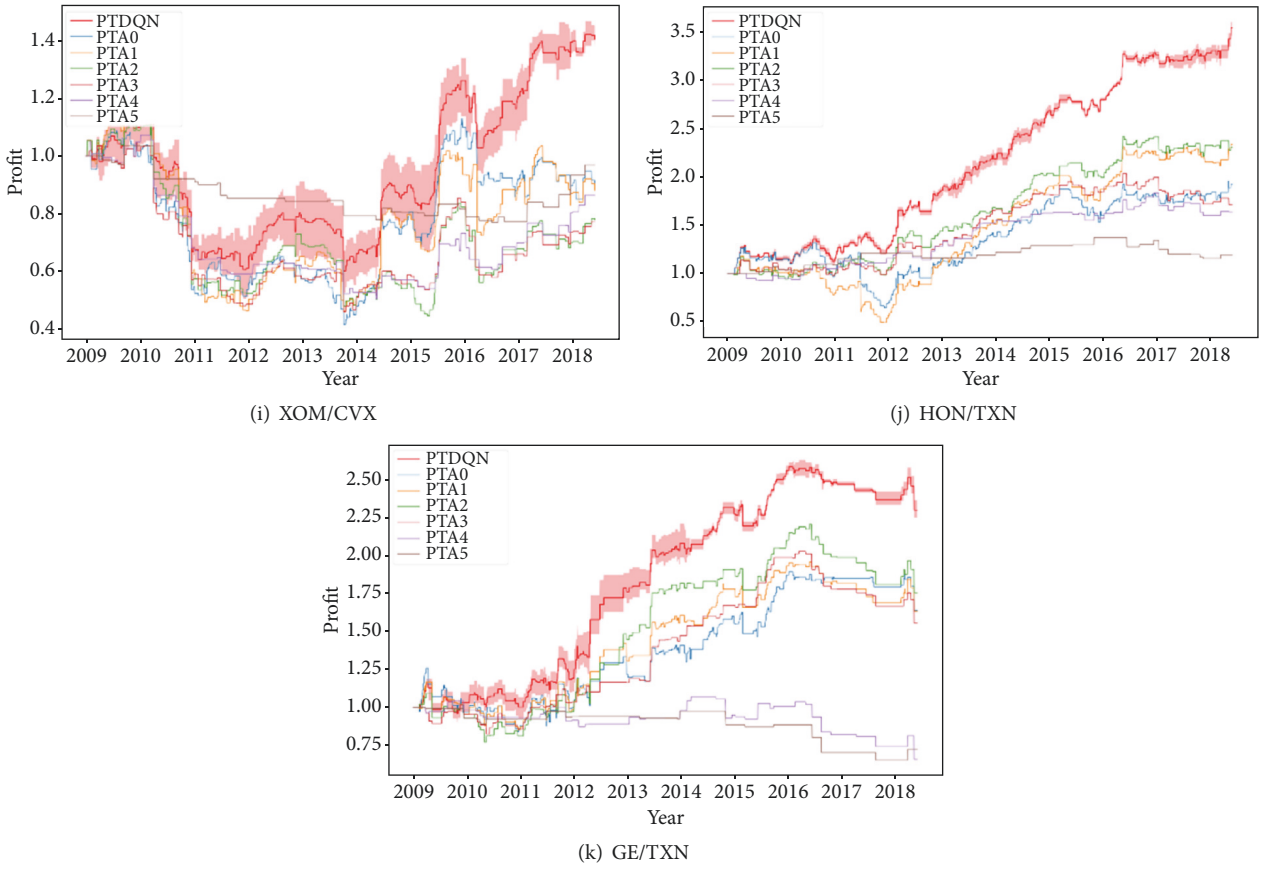


(g) JPM/GE



(h) JNJ/WFC

Figure 11: Continued.



(i) XOM/CVX

(j) HON/TXN

(k) GE/TXN

Figure 11: Average top-5 profits of PTDQN and PTA0 to PTA5 using TLS with the test dataset.

other methods. This shows that taking dynamic boundaries based on our method is efficient in optimizing the pairs trading strategy. During economic issues uncertainties, it can be a risk to manage the pairs trading strategies including our proposed method. However, we set a reward function if spread is suddenly high, and our network is trained to prevent this situation by taking less stop-loss boundary since it is trained to maximize the expected sum of future rewards. Therefore, our proposed method can minimize the risk when the economic risks appeared compared with traditional pairs trading strategy with fixed boundary.

From the experimental results, we show that our method can be applied in the pairs trading system. It can be applied in various fields, including finance and economics, when there is a need to optimize the efficiency of a rule-based strategy. Furthermore, we find that our method outperforms the traditional pairs trading strategy in all pairs based on constituent stocks in S&P 500. If we select appropriate pairs which are cointegrated, we can apply our methods to other markets such as KOSPI, Nikkei, and Hang Seng. The study focused on only spreads made by two stocks, which have long-term equilibrium patterns. Since our method selects optimal boundaries based on spreads, it can be applied to other stock markets such as KOSPI, Nikkei, and Hang Seng.

In future works, we can develop our proposed model as follows. First, as profit was set as the objective function in this study, the performance of the model is lower than traditional pairs trading when based on other performance measures. It can therefore be possible to create a better-optimized pairs-trading strategy by including all these other performance indicators as part of the objective function. Second, we can use other statistical methods such as the Kalman filter and error-correction models to use diversified spreads. Finally, it is possible to create a more-optimized pairs-trading strategy by continuously changing the discrete set of window sizes and boundaries. We will solve these difficulties in future studies.

Data Availability

The data used to support the findings of this study have been deposited in the figshare repository (DOI: 10.6084/m9.figshare.7667645).

Disclosure

The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. This work represents a part of the study

Table 8: Average top-5 performance results of the proposed method and the traditional pairs-trading strategy in the out-of-sample dataset using TLS.

Pairs	Model	MDD	Sharpe ratio	Profit	# of open portfolios	# of closed portfolios	# of stop-loss portfolios	# of exited portfolios
MSFT/JPM	PTDQN	−0.1122	0.2294	3.0446	186	126	38	62
	PTA0	−0.3411	0.0742	1.6236	211	136	57	18
	PTA1	−0.2907	0.0979	1.8001	162	104	26	32
	PTA2	−0.1507	0.1936	2.6303	131	64	7	60
	PTA3	−0.4032	0.1542	1.8282	97	39	1	57
	PTA4	−0.4340	0.0400	1.0480	55	13	0	42
	PTA5	−0.1836	0.3098	1.5524	30	7	0	23
MSFT/TXN	PTDQN	−0.3420	0.1001	1.5423	204	132	47	65
	PTA0	−1.2094	−0.0571	0.0013	244	152	76	16
	PTA1	−0.9225	−0.0177	0.6131	178	110	25	43
	PTA2	−0.5574	0.0351	1.0887	134	68	8	58
	PTA3	−0.5375	−0.0128	0.8326	97	34	1	62
	PTA4	−0.4485	0.0260	1.0118	66	15	1	50
	PTA5	−0.1048	0.1233	1.1502	32	5	0	27
BRKa/ABT	PTDQN	−0.0740	0.3159	2.3655	162	111	30	43
	PTA0	−0.1392	0.1554	1.7157	182	128	35	18
	PTA1	−0.1048	0.2464	2.1508	138	96	15	27
	PTA2	−0.1133	0.2538	1.9578	108	64	3	40
	PTA3	−0.1040	0.2480	1.7576	76	35	1	40
	PTA4	−0.0829	0.2087	1.3171	44	13	0	31
	PTA5	−0.0704	0.4366	1.4013	19	7	0	12
BRKa/UTX	PTDQN	−0.5401	0.1174	1.5744	167	105	35	58
	PTA0	−1.2143	−0.0199	0.5918	192	117	55	19
	PTA1	−0.9340	0.0346	1.0701	147	89	12	45
	PTA2	−0.9099	−0.0009	0.8435	122	60	5	57
	PTA3	−0.5673	0.0473	1.1520	89	32	1	56
	PTA4	−0.3641	0.0694	1.1628	53	9	0	44
	PTA5	−0.2309	0.0408	1.0405	18	3	0	15
JPM/T	PTDQN	−0.1384	0.1283	1.4653	175	113	42	53
	PTA0	−0.3630	0.0071	0.8968	205	129	60	15
	PTA1	−0.2801	0.0460	1.1595	144	94	17	32
	PTA2	−0.3750	0.0192	0.9987	119	62	5	51
	PTA3	−0.5241	−0.0717	0.6609	92	35	0	56
	PTA4	−0.3607	−0.0550	0.8411	56	18	0	38
	PTA5	−0.2235	0.0061	0.9851	22	6	0	16
JPM/HON	PTDQN	−0.1872	0.1523	2.2510	223	155	39	62
	PTA0	−0.6769	0.0190	1.0077	274	180	70	23
	PTA1	−0.4644	0.0622	1.6331	201	139	24	38
	PTA2	−0.4537	0.0840	1.7165	149	87	2	60
	PTA3	−0.2410	0.1414	1.7648	107	43	0	64
	PTA4	−0.3313	0.0879	1.3150	62	16	0	46
	PTA5	−0.1693	0.1803	1.2777	28	7	0	21

Table 8: Continued.

Pairs	Model	MDD	Sharpe ratio	Profit	# of open portfolios	# of closed portfolios	# of stop-loss portfolios	# of exited portfolios
JPM/GE	PTDQN	−0.1098	0.2123	2.8250	193	124	46	65
	PTA0	−0.3897	0.0507	1.5137	224	142	65	17
	PTA1	−0.3404	0.0640	1.6912	163	109	18	36
	PTA2	−0.1628	0.1284	1.9032	132	73	6	53
	PTA3	−0.2980	0.1142	1.7555	106	38	1	67
	PTA4	−0.2817	0.0790	1.2884	55	13	0	42
	PTA5	−0.0612	0.4776	1.7489	21	6	0	15
JNJ/WFC	PTDQN	−0.1576	0.2437	2.3741	143	100	28	38
	PTA0	−0.2872	0.0892	1.4932	164	115	37	12
	PTA1	−0.2219	0.1948	2.1147	127	90	15	21
	PTA2	−0.3188	0.1322	1.6362	99	55	5	38
	PTA3	−0.2324	0.1084	1.3141	68	27	0	41
	PTA4	−0.1532	0.1043	1.1228	40	14	0	26
	PTA5	−0.0970	0.1203	1.0734	16	6	0	10
XOM/CVX	PTDQN	−0.4265	0.0605	1.1924	218	135	45	77
	PTA0	−0.6189	0.0236	0.8812	256	161	67	28
	PTA1	−0.5999	0.0154	0.8809	197	118	25	54
	PTA2	−0.6034	−0.0073	0.7792	153	70	8	75
	PTA3	−0.5628	−0.0224	0.7734	114	38	2	74
	PTA4	−0.5311	−0.0200	0.8643	70	18	1	51
	PTA5	−0.2583	0.0060	0.9692	31	4	0	27
HON/TXN	PTDQN	−0.0874	0.2679	3.2755	233	164	49	63
	PTA0	−0.5108	0.1080	1.9219	276	186	66	23
	PTA1	−0.5841	0.1625	2.3378	207	140	28	38
	PTA2	−0.1926	0.2086	2.3096	158	92	4	62
	PTA3	−0.1611	0.1557	1.7100	114	49	2	63
	PTA4	−0.1254	0.2289	1.6374	69	23	0	46
	PTA5	−0.1578	0.1924	1.1925	28	9	0	19
GE/TXN	PTDQN	−0.1133	0.1871	2.1398	172	117	30	48
	PTA0	−0.3348	0.0967	1.6398	201	136	44	21
	PTA1	−0.1656	0.1070	1.6355	153	101	19	33
	PTA2	−0.2043	0.1388	1.7568	117	68	8	41
	PTA3	−0.2335	0.1591	1.5555	89	39	2	48
	PTA4	−0.3847	−0.1355	0.6570	45	7	0	38
	PTA5	−0.3489	−0.2730	0.7218	21	2	0	19
MO/UTX	PTDQN	−0.5264	0.0840	1.2940	150	88	35	58
	PTA0	−1.0950	−0.0272	0.6231	178	102	56	19
	PTA1	−0.7205	0.0286	1.0362	125	73	12	39
	PTA2	−0.8361	−0.0040	0.8658	105	51	3	50
	PTA3	−0.4311	0.0052	0.9323	79	24	0	54
	PTA4	−0.3916	0.1141	1.2129	48	12	0	36
	PTA5	−0.1311	0.2948	1.1276	14	3	0	11

Table 9: Average top-5 performance results of the proposed method and the traditional pairs-trading strategy in the out-of-sample dataset using OLS.

Pairs	Model	MDD	Sharpe ratio	Profit	# of open portfolios	# of closed portfolios	# of stop-loss portfolios	# of exited portfolios
MSFT/JPM	PTDQN	−0.2096	0.1228	1.9255	215	137	54	62
	PTA0	−0.3618	0.0492	1.3365	225	141	61	23
	PTA1	−0.5036	0.0188	1.0185	168	102	28	38
	PTA2	−0.4045	0.0611	1.3591	124	59	8	57
	PTA3	−0.5055	−0.0094	0.8636	97	33	3	61
	PTA4	−0.4195	−0.0009	0.9459	58	12	1	45
	PTA5	−0.2018	0.1236	1.1593	29	6	0	23
MSFT/TXN	PTDQN	−0.2878	0.0698	1.3466	244	153	65	68
	PTA0	−0.5271	0.0070	0.8489	252	156	72	24
	PTA1	−0.4721	0.0255	1.0286	187	117	26	44
	PTA2	−0.3816	0.0215	0.9912	145	71	10	64
	PTA3	−0.6553	−0.1015	0.5053	104	30	2	72
	PTA4	−0.2719	0.0422	1.0532	63	16	1	46
	PTA5	−0.1850	0.0068	0.9785	34	7	0	27
BRKa/ABT	PTDQN	−0.1282	0.1644	1.5076	180	109	48	57
	PTA0	−0.5073	−0.0265	0.7070	183	112	48	22
	PTA1	−0.2649	0.0453	1.0786	139	80	13	46
	PTA2	−0.2246	0.1056	1.2942	121	60	4	56
	PTA3	−0.1686	0.1241	1.2718	91	38	1	52
	PTA4	−0.1483	0.0176	0.9778	49	12	0	37
	PTA5	−0.1602	0.0004	0.9830	16	2	0	14
BRKa/UTX	PTDQN	−0.5231	0.0816	1.2976	215	132	57	69
	PTA0	−1.1928	−0.0647	0.3332	216	133	57	25
	PTA1	−0.8697	−0.0157	0.7445	167	100	15	51
	PTA2	−0.7815	−0.0071	0.8391	135	70	5	60
	PTA3	−0.3573	0.0315	1.0292	94	36	0	58
	PTA4	−0.2096	0.0684	1.0857	52	11	0	41
	PTA5	−0.1317	−0.1174	0.9312	16	2	0	14
JPM/T	PTDQN	−0.1338	0.1391	1.4547	205	127	60	50
	PTA0	−0.3588	0.0069	0.9054	208	130	61	16
	PTA1	−0.2535	0.0405	1.0902	151	96	19	35
	PTA2	−0.1872	0.0542	1.1198	119	66	5	48
	PTA3	−0.2574	0.0336	1.0502	94	39	0	55
	PTA4	−0.2212	0.0345	1.0312	57	20	0	37
	PTA5	−0.2348	−0.1922	0.8299	20	5	0	15
JPM/HON	PTDQN	−0.3869	0.1071	1.5175	250	162	57	68
	PTA0	−0.7141	0.0181	0.9444	256	166	59	30
	PTA1	−0.5065	0.0702	1.3071	198	127	22	49
	PTA2	−0.4649	0.1071	1.4260	152	84	3	65
	PTA3	−0.4871	0.0763	1.2098	102	44	0	58
	PTA4	−0.3503	−0.0694	0.8178	50	13	0	37
	PTA5	−0.2980	−0.1721	0.8040	23	6	0	17

Table 9: Continued.

Pairs	Model	MDD	Sharpe ratio	Profit	# of open portfolios	# of closed portfolios	# of stop-loss portfolios	# of exited portfolios
JPM/GE	PTDQN	−0.1195	0.1443	1.7682	226	133	64	69
	PTA0	−0.4379	0.0036	0.8549	232	137	66	29
	PTA1	−0.1523	0.0987	1.4814	165	98	16	51
	PTA2	−0.1738	0.1264	1.5661	134	62	5	67
	PTA3	−0.2680	0.0729	1.2026	93	29	0	64
	PTA4	−0.2104	0.1298	1.3242	51	12	0	39
	PTA5	−0.1461	−0.0423	0.9586	18	3	0	15
JNJ/WFC	PTDQN	−0.1890	0.1266	1.7194	202	130	47	56
	PTA0	−0.8705	−0.0326	0.4635	207	131	53	22
	PTA1	−0.6189	−0.0134	0.7318	150	91	19	39
	PTA2	−0.4763	0.0309	1.0563	124	57	4	62
	PTA3	−0.2318	0.1447	1.6072	97	33	2	62
	PTA4	−0.2415	0.0549	1.0632	50	13	0	37
	PTA5	−0.0880	0.2468	1.1886	20	4	0	16
XOM/CVX	PTDQN	−0.3316	0.0265	1.1517	141	81	23	43
	PTA0	−0.7629	−0.0547	0.4186	240	149	61	30
	PTA1	−0.5648	0.0132	0.8754	193	114	23	56
	PTA2	−0.6977	−0.0387	0.6655	154	70	7	77
	PTA3	−0.5235	0.0277	0.9865	117	38	1	78
	PTA4	−0.4781	−0.0577	0.8117	63	12	1	50
	PTA5	−0.3787	−0.1492	0.8090	29	3	0	26
HON/TXN	PTDQN	−0.1339	0.1534	1.8852	270	175	64	69
	PTA0	−0.4135	0.0212	0.9455	276	177	70	28
	PTA1	−0.2758	0.0666	1.3216	207	124	27	55
	PTA2	−0.2614	0.1054	1.5031	159	84	5	69
	PTA3	−0.1759	0.1413	1.5617	117	45	2	70
	PTA4	−0.0834	0.2650	1.7044	66	23	0	43
	PTA5	−0.0664	0.4606	1.6830	30	13	0	17
GE/TXN	PTDQN	−0.1676	0.1263	1.6411	206	140	43	62
	PTA0	−0.6133	0.0178	0.9742	211	144	44	23
	PTA1	−0.3085	0.0586	1.2743	166	109	19	38
	PTA2	−0.2402	0.0585	1.2216	128	68	5	55
	PTA3	−0.3190	−0.0013	0.9193	91	31	2	58
	PTA4	−0.2493	−0.0285	0.9117	49	8	0	41
	PTA5	−0.0862	0.1417	1.0936	23	4	0	19
MO/UTX	PTDQN	−0.3181	0.0524	1.1402	188	117	49	59
	PTA0	−0.4688	0.0041	0.8667	195	121	52	21
	PTA1	−0.6166	−0.0230	0.7470	144	84	13	46
	PTA2	−0.5034	−0.0076	0.8666	115	51	4	59
	PTA3	−0.2833	0.0457	1.0873	88	32	0	56
	PTA4	−0.2901	0.0356	1.0280	44	12	0	32
	PTA5	−0.1500	0.0992	1.0297	13	2	0	11

conducted as a Master Thesis in Financial Engineering during 2016 and 2018 at the University of Ajou, Republic of Korea.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea Government (MSIT: Ministry of Science and ICT) (No. NRF-2017R1C1B5018038).

References

[1] E. Gatev, W. N. Goetzmann, and K. G. Rouwenhorst, "Pairs trading: performance of a relative-value arbitrage rule," *Yale ICF Working Paper No. 08-03*, 1998, https://ssrn.com/abstract=141615 or http://dx.doi.org/10.2139/ssrn.141615.

[2] R. J. Elliott, J. van der Hoek, and W. P. Malcolm, "Pairs trading," *Quantitative Finance*, vol. 5, no. 3, pp. 271–276, 2005.

[3] S. Andrade, V. Di Pietro, and M. Seasholes, "Understanding the profitability of pairs trading," 2005.

[4] G. Hong and R. Susmel, "Pairs-trading in the Asian ADR market," Univ. Houston, Unpubl. Manuscr., 2003.

[5] E. Gatev, W. N. Goetzmann, and K. G. Rouwenhorst, "Pairs trading: performance of a relative-value arbitrage rule," *Review of Financial Studies* , vol. 19, no. 3, pp. 797–827, 2006.

[6] B. Do and R. Faff, "Does simple pairs trading still work?" *Financial Analysts Journal*, vol. 66, no. 4, pp. 83–95, 2018.

[7] S. Mudchanatongsuk, J. A. Primbs, and W. Wong, "Optimal pairs trading: A stochastic control approach," in *Proceedings of the 2008 American Control Conference, ACC*, pp. 1035–1039, USA, June 2008.

[8] A. Tourin and R. Yan, "Dynamic pairs trading using the stochastic control approach," *Journal of Economic Dynamics & Control*, vol. 37, no. 10, pp. 1972–1981, 2013.

[9] Z. Zeng and C. Lee, "Pairs trading: optimal thresholds and profitability," *Quantitative Finance*, vol. 14, no. 11, pp. 1881–1893, 2014.

[10] S. Fallahpour, H. Hakimian, K. Taheri, and E. Ramezanifar, "Pairs trading strategy optimization using the reinforcement learning method: a cointegration approach," *Soft Computing*, vol. 20, no. 12, pp. 5051–5066, 2016.

[11] P. Nath, "High frequency pairs trading with U.S. treasury securities: risks and rewards for hedge funds," *SSRN Electronic Journal*, 2004.

[12] T. Leung and X. Li, "Optimal mean reversion trading with transaction costs and stop-loss exit," *International Journal of Theoretical and Applied Finance*, vol. 18, no. 3, 2013.

[13] E. Ekström, C. Lindberg, and J. Tysk, "Optimal liquidation of a pairs trade," in *Advanced Mathematical Methods for Finance*, pp. 247–255, Springer, Heidelberg, 2011.

[14] Y. Lin, M. McCrae, and C. Gulati, "Loss protection in pairs trading through minimum profit bounds: A cointegration approach," *Journal of Applied Mathematics and Decision Sciences*, vol. 2006, pp. 1–14, 2006.

[15] A. Mikkelsen, "Pairs trading: the case of Norwegian seafood companies," *Applied Economics*, vol. 50, no. 3, pp. 303–318, 2017.

[16] K. Kim, "Performance analysis of pairs trading strategy utilizing high frequency data with an application to KOSPI 100 Equities," *SSRN Electronic Journal*, p. 24, 2011.

[17] V. Holý and P. Tomanová, *Estimation of Ornstein-Uhlenbeck Process Using Ultra-High-Frequency Data with Application to Intraday Pairs Trading Strategy*, 2018.

[18] D. Chen, J. Cui, Y. Gao, and L. Wu, "Pairs trading in Chinese commodity futures markets: an adaptive cointegration approach," *Accounting & Finance*, vol. 57, no. 5, pp. 1237–1264, 2017.

[19] H. Puspaningrum, Y. Lin, and C. M. Gulati, "Finding the optimal pre-set boundaries for pairs trading strategy based on cointegration technique," *Journal of Statistical Theory and Practice*, vol. 4, no. 3, pp. 391–419, 2010.

[20] A. A. Roa, "Pairs trading: optimal thershold strategies," 2018.

[21] V. Mnih, K. Kavukcuoglu, D. Silver et al., "Playing atari with deep reinforcement learning," https://arxiv.org/abs/1312.5602, 2013.

[22] Y. Wang, D. Wang, S. Zhang, Y. Feng, S. Li, and Q. Zhou, "Deep Q-trading," 2017, http://cslt.riit.tsinghua.edu.cn/.

[23] C.-Y. Huang, "Financial trading as a game: a deep reinforcement learning approach," 2018, https://arxiv.org/abs/1807.02787.

[24] T. Kim, *Optimizing the pairs trading strategy using Deep reinforcement learning [M.S. thesis]*, Ajou University, Suwon, Republic of Korea, 2019.

[25] B. Do, R. Faff, and K. Hamza, "A new approach to modeling and estimation for pairs trading," in *Proceedings of the 2006 Financial Management Association European Conference*, 2006.

[26] R. D. Dittmar, C. J. Neely, and P. A. Weller, "Is technical analysis in the foreign exchange market profitable? A genetic programming approach," *Journal of Financial and Quantitative Analysis*, vol. 43, p. 43, 1997.

[27] H. Rad, R. K. Low, and R. Faff, "The profitability of pairs trading strategies: distance, cointegration and copula methods," *Quantitative Finance*, vol. 16, no. 10, pp. 1541–1558, 2016.

[28] S. Johansen, "Statistical analysis of cointegration vectors," *Journal of Economic Dynamics and Control*, vol. 12, no. 2-3, pp. 231–254, 1988.

[29] M. H. Kutner, C. J. Nachtsheim, J. Neter, and W. Li, "Applied linear statistical models," 1996.

[30] G. H. Golub and C. F. Van Loan, "An analysis of the total least squares problem," *SIAM Journal on Numerical Analysis*, vol. 17, no. 6, pp. 883–893, 1980.

[31] R. S. Sutton and A. G. Barto, "Introduction to reinforcement learning," *Learning*, 1998.

[32] E. F. Fama, "Random walks in stock market prices," *Financial Analysts Journal*, vol. 51, no. 1, pp. 75–80, 1995.

[33] W. F. Sharpe, "The sharpe ratio," *The Journal of Portfolio Management*, 1994.

[34] P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, and D. Meger, "Deep reinforcement learning that matters," in *Proceedings of the Thirthy-Second AAAI Conference On Artificial Intelligence (AAAI)*, 2018.

[35] Y. H. Li, X. M. Lu, and N. C. Kar, "Rule-based control strategy with novel parameters optimization using NSGA-II for power-split PHEV operation cost minimization," *IEEE Transactions on Vehicular Technology*, vol. 63, no. 7, pp. 3051–3061, 2014.

[36] L. Dymova, P. Sevastianov, and K. Kaczmarek, "A stock trading expert system based on the rule-base evidential reasoning using Level 2 Quotes," *Expert Systems with Applications*, vol. 39, no. 8, pp. 7150–7157, 2012.