DISTRIBUTIONAL REINFORCEMENT LEARNING BASED ON HISTORICAL INFORMATION FOR OPTION HEDGING

Anonymous authors
Paper under double-blind review

ABSTRACT

Options are widely used financial derivatives for risk management and corporate operations. Option hedging aims to mitigate investment risks from asset price fluctuations by buying and selling other financial products. Traditional hedging strategies based on the Black-Scholes model face practical limitations due to the assumptions of constant volatility and the neglect of transaction costs. Recently, reinforcement learning(RL) has gained attention in the study of option hedging strategies, but several challenges remain: current methods rely on real-time market data (e.g., underlying asset prices, holdings, remaining option term) to determine optimal positions, underutilizing the potential value of historical data; existing approaches focus on the expected hedging cost, overlooking the comprehensive distribution of costs; In the aspect of training data generation, commonly used single simulation methods perform well under specific conditions but struggle to ensure the robustness of the model across diverse datasets. To address these issues, we propose a novel distributional RL option hedging method that incorporates historical information. Historical states are included in the state variables, with a gated recurrent unit (GRU) network layer extracting historical information. This is then combined with current information from fully connected layers to inform subsequent network layers, ensuring the agent considers both current and historical market information when learning hedging strategies. The output of the value network is set as a series of quantiles, with the Quantile Huber Loss function fitting their distribution to evaluate strategies based on distribution rather than expected value. To diversify data sources, we use a combination of the Black-Scholes model, the Binomial model, and the Heston model to simulate a large volume of option data. Experimental results show that our method significantly reduces hedging costs and demonstrates strong adaptability and practicality under various market conditions.

1 INTRODUCTION

Options are tradable financial derivatives that grant the holder the right, but not the obligation, to buy or sell a certain asset at a specified price at a future date. They play a significant role in risk management and corporate operations(Xiao et al., 2021). In options-related trading, option hedging is one of the most closely watched issues(Mandelli et al., 2023), primarily aimed at controlling investment risk. Option hedging primarily refers to the process where the option seller takes certain measures to reduce potential risks after earning a premium from selling the option. When the option expires, if the buyer exercises the option, the seller must sell or buy the asset at the agreed price, which may be unfavorable for the seller. To mitigate this potential risk, sellers often adopt option hedging strategies, involving the buying or selling of other related financial instruments to offset the risk caused by asset price fluctuations.

Traditional hedging strategies, such as delta hedging based on the Black-Scholes (BS) model(Black & Scholes, 1973), use delta values to obtain the optimal position of the underlying asset. However, the assumptions of the BS model regarding constant volatility, continuous trading, and frictionless markets do not hold in reality, as the volatility of the underlying asset is constantly changing and

there are transaction costs and the inability to trade continuously in the actual market. Therefore, such methods have significant limitations in practical applications.

To address the aforementioned issues, in recent years, using reinforcement learning(RL) methods for option hedging has become a research hotspot. RL leverages large amounts of data and continuously interacts with the environment, allowing for the adjustment of hedging strategies based on environmental feedback, which resolves the issue that traditional option hedging models cannot fully reflect real market conditions(Hambly et al., 2023).

In the design of option hedging based on RL, the underlying asset price, the holding of the underlying asset, etc. are typically considered as the states, while the trading volume or change in the holding of the underlying asset is treated as the action. The profit and loss(P&L) or changes in cash flow during the option hedging process are considered as the reward. The basic framework is illustrated in Figure 1. In this process, the setting of state variables, the estimation of the value function, and the collection of training data are three main issues.

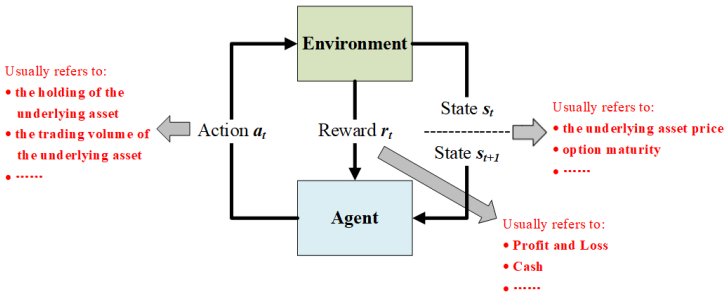

Figure 1: The basic framework of RL for option hedging

In terms of state variables, previous research has primarily focused on the current market information (such as the price of the underlying asset, the holding of the underlying asset, and the time to expiration when hedging). This means that the agent decides the optimal holding for hedging based only on the current market information, as illustrated in the left of Figure 2. The hedging decision at time t_1 is made based only on the state at s_1; the hedging decision at time t_2 is made based only on the state at s_2, independent of the state at s_1, and so on. However, in actual hedging, considering historical market information is also important. This means that at time t_2, the hedging decision should take into account not only the market information at t_2 but also before t_2, and so forth.

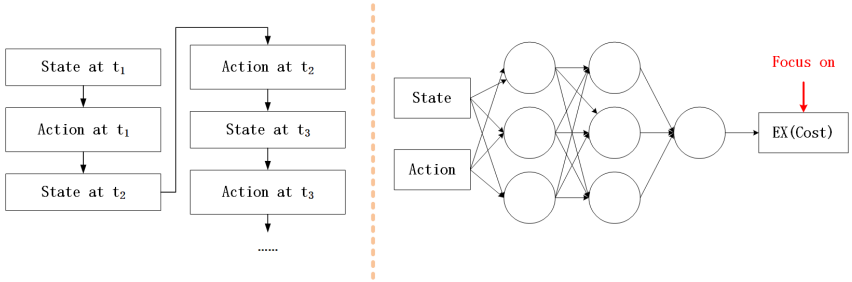

Figure 2: The left is previous decision-making process.The right is previous value network for option hedging.

In terms of value function, previous research has primarily focused on the expected value of hedging costs (as shown in the right of Figure 2). While this approach of focusing solely on the expected value of hedging costs provides an average view of hedging costs under a given strategy, it fails to reflect the overall distribution of hedging costs. In practice, the distribution of hedging costs is crucial for evaluating the effectiveness of a hedging strategy. Therefore, ignoring the cost distribution when evaluating hedging strategies may increase the risk of strategy execution.

Regarding data, RL methods necessitates extensive data. in practice, due to factors such as different issuance times and durations of options, the available option data is insufficient to meet this requirement. As a result, research often uses finance parametric methods to generate data. However, existing studies typically use a single method to generate data for training. Such approach can lead to a classic problem in financial engineering: the model's performance will primarily reflect accuracy on this type of simulated data, while performance may degrade when using other data(Lillicrap et al., 2019).

To solve the issues in option hedging based on RL mentioned above, we propose a new distributional RL method based on historical information for option hedging. The main contributions includes:

(1)**Incorporating Historical Information into State Variables:** Addressing the issue where past research only considers current market information as the state, we include historical information into the state. We introduce GRU and leverage its memory function to extract historical information. The extracted historical information is then combined with the current information extracted by fully connected layer and fed into subsequent network layers. This enables the agent to consider both current and historical information when learning the option hedging strategy.

(2)**Estimating the Distribution of Hedging Costs:** Tackling the problem where past research mainly focuses on the expected value of hedging costs, we employ a set of quantiles to estimate the hedging costs in the value network instead of estimating the expected value. In other words, we treat the output of the value network as a distribution and use quantile regression combined with the Huber loss function to fit quantiles. This approach aims to estimate the distribution of hedging costs.

(3)**Generating Simulated Data with Multiple Methods:** To address the issue of using a single method to generate simulated data in past research, we combine multiple option pricing methods to generate simulated option data for training. By doing so, the agent is exposed to a richer variety of samples during the learning process, enhancing its adaptability to different market conditions.

2 Related Work

Halperin (2020) was the first to apply RL to address dynamic option hedging. He proposed QLBS model based on Q-Learning(Watkins, 1989), which provides an effective method for option pricing and hedging without the constraint of continuous time. However, this approach is only effective in finite state and action space and still assumes frictionless markets. To avoid the curse of dimensionality, Kolm & Ritter (2019) used SARSA(Sutton & Barto, 2018) to consider nonlinear trading costs in continuous state space, making agent effective for hedging in environments that more closely resemble real markets. With the advancement of RL, Du et al. (2020) applied more advanced algorithms such as DQN(Mnih et al., 2013) and PPO(Schulman et al., 2017) for discrete option hedging, also addressing the issue of integer stocks and incorporating strike prices as additional state variables. Furthermore, in their model training, the chosen strike prices were no longer fixed but rather within a given range. This allowed the agent to train on a range of strike prices without retraining for each specific strike price within that range. Although discretizing actions simplifies training process, it also introduces significant errors. To address it, Cao et al. (2020) employed DDPG(Lillicrap et al., 2019), which allows for continuous state and action spaces for option hedging, providing more precise action outputs. They still used the current information as state variables. Additionally, they used two value networks to estimate the expected value and its square of the hedging cost to calculate the utility function. Their experiments under P&L and cash-flow showed the former was more effective than the latter. In response to same problem, Mikkilä (2020) used TD3, and add moneyness and implied volatility to state variables. Addtionally, they introduced sim-to-real method(training on simulated data and testing on real data). Similarly, Giurca & Borovkova (2021) also used sim-to-real in DQN and DDPG. Both of them found that RL also outperformed traditional methods in actual data. Due to the uncertainty in financial markets, Zheng et al. (2023) incorporated uncertainty into DDPG to develop more robust hedging strategies. Their results showed that accounting for uncertainty provided better solutions compared to Cao et al. (2020). Similarly, to further align with real market, Neagu et al. (2024) integrated market shocks into the state space. Their findings indicated that incorporating market shocks into the state space helps better adapt to the market environment. Previous research on option hedging relied directly or indirectly used parametric models to generate data for training, to address this issue, Mikkilä & Kanniainen (2023) directly trained the model using actual data. Their results indicate that training DRL models directly on real data performs better.

Contrary to them, Cannelli et al. (2023) solved this problem by using the more efficient CMAB algorithm, but the data was still generated by GBM and BS models. In practice, effective option hedging requires more than just current market conditions, historical market dynamics are crucial as well. However, existing methods using RL often rely solely on current information, ignoring historical data. Furthermore, when evaluating hedging strategies, it's important to consider both the expected hedging costs and the associated risks, but current methods tend to focus on the expected value of costs while overlooking their distributional characteristics. Additionally, most research uses the BS model for simulating option data, which lacks diversity and may result in strategies that perform poorly on other types of data. Therefore, it is important to develop RL methods for option hedging that incorporate historical information, cost distribution, and data diversity.

3 MODEL

3.1 DISTRIBUTIONAL REINFORCEMENT LEARNING BASED ON HISTORICAL INFORMATION

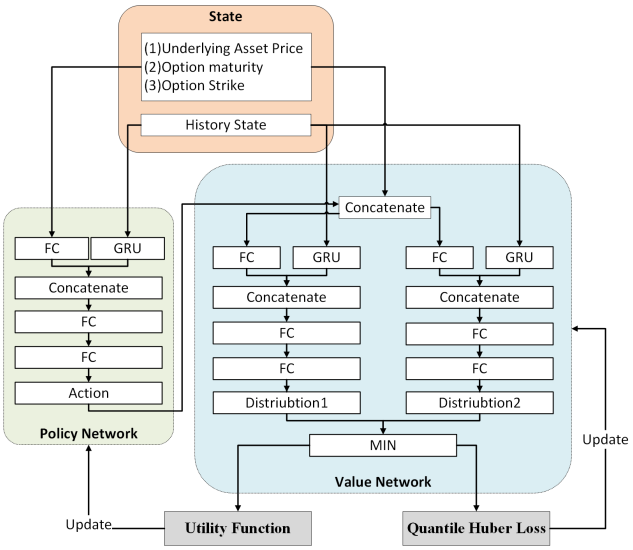

Figure 3: DRL Framework based on Historical Information for Option Hedging

Our model architecture is shown in Figure 3. When applying RL to option hedging, choosing the appropriate state and action variables as well as the reward function is crucial for the agent to learn option hedging. For illustrative purposes, some symbol definitions are shown in Table 1.

Table 1: Symbol Explanation

symbol	explanation	symbol	explanation
s_t	the underlying asset price at time t	o_t	option price at time t
m_t	option expiration time at time t	K	strike
p_t	holdings of the underlying asset at time t	$hstate_t$	historical state at time t

In the selection of state variables, firstly, the underlying asset price is closely related to the option. If the underlying asset price continues to rise, the likelihood of the buyer exercising the option at expiration will also continue to increase, which directly affects the performance of the option hedging strategy. Secondly, the holding of the underlying asset also affects the performance of the option hedging strategy. Although the seller can hold a certain amount of the underlying asset to offset the potential risk brought by the option during a price increase, holding too much of the underlying asset will increase costs to a certain extent, while holding too little may result in insufficient hedging. Moreover, the time to expiration is also a crucial factor not to be neglected in option hedging. Therefore, considering these three characteristics is our primary concern. Additionally, we believe that in option hedging, it is essential to consider not only the current market conditions but also past

market trends, as historical data provides valuable information for anticipating future market movements. Consequently, we have also included historical states in the state variables of our model.In summary, the state at time t is represented as $state_t$:

$$state_t = (s_t, m_t, p_t, hstate_t) \tag{1}$$

In terms of selecting action variables, since we are focusing on reducing potential losses from selling options by buying and selling the underlying asset, the action at time t is represented as $action_t$:

$$action_t = p_{t+1} \in [0, 100] \tag{2}$$

Using the example of the SSE 50ETF, where one option contract represents 10,000 units of the underlying asset. $action_t = 0$ indicates that the holdings of the underlying asset are 0, and $action_t = 50$ indicates that the holdings of the underlying asset are 5,000.

To simplify the calculation process, we set the risk-free rate to 0 and the transaction cost as a percentage of the transaction amount for the underlying asset. Based on these assumptions, the agent mainly focuses on P&L during the option hedging process, which includes two components:

(1)**P&L Due to the Underlying Asset**. Given a certain amount of the underlying asset held, if the price of the underlying asset rises, it results in gains, while if the price falls, it results in losses. Additionally, trading the underlying asset incurs transaction costs. Thus, this part of P&L can be calculated as(per is percentage of transaction costs to the trading volume of the underlying asset):

$$asset_t = p_t \times (s_{t+1} - s_t) - per \times |p_{t+1} - p_t| \tag{3}$$

(2)**P&L Due to the Option**. If the option has not yet expired, changes in the option's price can result in P&L. The reward calculation for this section follows formula (4). If the option expires and the underlying asset price is above the strike price, the buyer will exercise the option, resulting in a loss for the seller, additionally, if the seller still holds the underlying asset, trading this remaining amount incurs transaction costs. The P&L calculation for this section follows formula (5).

$$option_t = o_t - o_{t+1} \tag{4}$$

$$option_t = o_t - max(s_t - K, 0) \tag{5}$$

Based on the components outlined above, the total reward can be calculated as:

$$R_t = asset_t + option_t \tag{6}$$

3.2 Historical Information Fusion

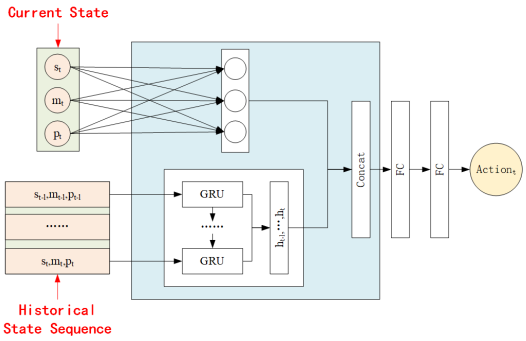

Figure 4: The Fusion of Current and Historical Information in our model

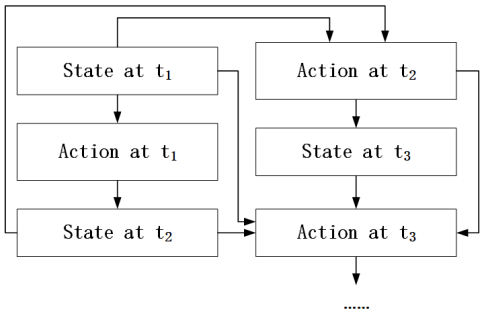

Figure 5: Decision-Making Process in Our Method

In the process of option hedging, it is crucial to focus not only on current market conditions but also on historical market dynamics. Therefore, we hope that the agent can consider both current and historical information during the learning process. Hence, we include not only the current underlying asset price, holdings, and option expiration as state variables in the network input but

also incorporate historical states. Additionally, we introduce a GRU(Chung et al., 2014) to extract historical sequence information, thereby providing the agent with a memory capability during the process of learning option hedging. We choose GRU rather than LSTM because they have the similar effect, but GRU has fewer parameters. Otherwise, What we need to emphasize that our method is designed to better handle historical temporal data, enabling agent to leverage dependencies in time series, is concerned with the same data. This is different from the experience replay buffer, which reduces the correlation between different data by storing the data and then randomly sampling them for model training. To implement this, the specific steps are as follows:

First, the data at time t is input into a fully connected network layer to extract the current information. The result of this extraction is denoted as X_t:

$$n_t = (s_t, m_t, p_t) \tag{7}$$

$$X_t = FC(n_t) = W_n n_t + b_n \tag{8}$$

Where n_t represents the input variables composed of the data at time t, W_n and b_n denote trainable weights and bias vector in FC layer for current feature extraction.

Next, the historical data sequence at time t is input into GRU to extract the historical information. The result of this extraction is denoted as H_t:

$$hstate_t = (n_{t-l}, \ldots, n_t) \tag{9}$$

$$z_t = \sigma(W_z \cdot [h_{t-1}, n_t]) \tag{10}$$

$$r_t = \sigma(W_r \cdot [h_{t-1}, n_t]) \tag{11}$$

$$\hat{h}_t = tanh\left(W \cdot [r_t \odot h_{t-1}, n_t]\right) \tag{12}$$

$$h_t = z_t \odot h_{t-1} + (1 - z_t) \odot \hat{h}_t \tag{13}$$

$$H_t = [h_{t-l}, \ldots, h_t] \tag{14}$$

where $hstate_t$ represents the input variables which consist of the historical data sequence at time t, and H_t denotes the hidden state extracted by the GRU at time t.

Finally, the extracted current information X_t and historical information H_t are combined to form C_t and fed into subsequent layers for further computation.

$$C_t = concat(X_t, H_t) \tag{15}$$

Through the above methods, the intelligent agent no longer makes decisions based solely on current market information when hedging, but also considers historical market information in the decision-making process, thus achieving the goal of simultaneously considering both when hedging. As shown in Figure 5, at time $t3$, the agent considers information from time $t1$, $t2$, and $t3$ when hedging.

3.3 Distribution based on Quantile Regression

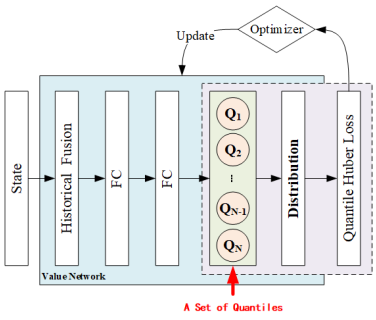

Figure 6: Implementation Method of Distributional RL in Our Method

When evaluating option hedging strategies, it is not enough to only focus on the expected value of hedging costs, as the effectiveness of the strategy depends not only on the expected value of hedging

costs but also on some other important factors. Therefore, we hope that the value network can consider the effectiveness of the strategy from a distribution perspective, so that the agent can not only consider the overall situation, but also other information about the distribution, thus achieving a more comprehensive evaluation of the strategy.

To this end, we set the output of the value network as a set of quantiles (as shown in Figure 6), treat this set of quantiles as the distribution, and then fit these quantiles through quantile regression.To meet the above requirements, we used the Quantile Huber loss function when fitting quantiles in the value network, which is expressed as:

$$loss_{\tau}^{\kappa}(y, y') = \begin{cases} (1-\tau)\mathcal{L}_{\kappa}(y, y'), & if y < y' \\ \tau \cdot \mathcal{L}_{\kappa}(y, y'), & otherwise \end{cases} \tag{16}$$

$$\mathcal{L}_{\kappa}(y, y') = \begin{cases} \frac{1}{2}(y - y')^2, & if |y - y'| < k \\ k\left(|y - y'| - \frac{1}{2}k\right), & otherwise \end{cases} \tag{17}$$

At the same time, in the process of option hedging, in addition to considering hedging costs, risk levels also need to be taken into account. The optimal strategy should balance both the average and volatility of asset returns. Therefore, we adopted a common utility function in investment, the mean-variance utility functionMarkowitz (1952)(such as the formula (18)), as the objective function in the strategy network, to maximize expected returns while reducing their volatility.

$$L_{actor} = Max[E(R) - risk \times V(R)] \tag{18}$$

Here, risk is the risk preference parameter, with a higher value indicating a greater aversion to risk.

Since the output of the value network is a set of quantiles, and the value network assists the strategy network in learning strategies, we take the mean and standard deviation of the output quantiles as the mean and standard deviation in the utility function, respectively. Finally, we obtain the objective function of the strategy network as a formula, which uses gradient rise to achieve the goal of maximizing the expected value of returns while reducing the volatility of returns.

$$L_{actor} = Max\left(\frac{1}{N}\sum_{i=1}^{N} y_c^i - risk * \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(y_c^i - \frac{1}{N}\sum_{i=1}^{N} y_c^i)^2}\right) \tag{19}$$

Here, N is the number of quantiles and y_c^i is the i-th quantile of the value network output.

4 EXPERIMENTS

4.1 DATA

In the actual market, although there are options with the same strike price and expiration date, they cannot be considered identical due to factors such as the time of issuance and different remaining periods. This results in a relatively limited amount of data for options with specific strike prices and expiration dates in practice. However, in RL, the process by which the agent learns an option hedging strategy requires a large amount of data. Therefore, we are considering using financial models to simulate a substantial amount of underlying asset data and option data.

For the underlying data, we use geometric Brownian motion to simulate its price. For option data, if only one method is used, the option hedging strategy learned by the agent may exhibit superior performance for the data generated by that method, but performance may decline for data generated by other methods. Therefore, we use three methods to simulate option prices: the BS model(Black & Scholes, 1973), the BI model(Cox et al., 1979), and the Heston model(Heston, 2015). The data generated by these three methods are then combined and provided to the agent for learning.

4.2 EXPERIMENTS

Considering the presence of transaction costs in actual trading, we set the transaction cost as a proportional cost in the experiment, as 1% of the total transaction amount for each underlying asset

trade. The evaluation metrics include Mean and Std of Return and Reward, and Percent, which respectively represent the mean and standard deviation of the total hedging cost and daily hedging cost for testing paths at the end of hedging, and the proportion of test paths that outperform delta hedging using this method. We then used the results from the test paths to plot boxplots comparing the performance of different hedging methods. Finally, we used Gaussian kernel density estimation to show the P&L distributions under different methods, illustrating the characteristics of the P&L distributions for each hedging method under the same market conditions.

4.2.1 COMPARATIVE EXPERIMENT

To illustrate the efficacy of our model for option hedging, we compared its test results with the BS-delta hedging and the DDPGCao et al. (2020). The performance of each model was evaluated using the five aforementioned metrics, and the results are presented in Table 2. Additionally, boxplots illustrating the performance of the three hedging strategies are presented in the left of Figure 7.

Table 2: The evaluation of the metrics in comparative experiment.

Model	**Return**		**Reward**		**Percent**
	Mean	**Std**	**Mean**	**Std**	
BS-Delta	-242.6	84.19	-12.10	21.74	
DDPGCao et al. (2020)	-170.87	96.82	-8.54	24.44	71.8%
ours	-117.13	154.26	-5.90	32.31	77.4%

From Table 2, it can be observed that, in the cases we considered, compared to delta hedging, although using RL to option hedging increases the standard deviation, it significantly reduces the hedging costs. Additionally, though our model increases the standard deviation to some extent compared to the DDPG model, it further reduces the hedging costs. However, we found that the RL method is not superior to BS-delta hedging in all cases. Thus, we also compared the proportion of hedging results from the DDPG model and our model that outperformed BS-delta hedging. The results show that the proportion of paths outperforming BS-delta hedging is 71.8% for the DDPG model and 77.4% for our model, indicating that our model better adapts to different market conditions. Next, we used Gaussian kernel density estimation to analyze the P&L distributions for three methods, as shown in the right of Figure 7. It is evident that in terms of profits, our method has a significantly higher probability density compared to the other two methods, especially in the range of returns from 0 to 200. Regarding losses, our method's probability density is lower than the other two methods in most cases, indicating that our method generally achieves higher profits and effectively reduces losses.

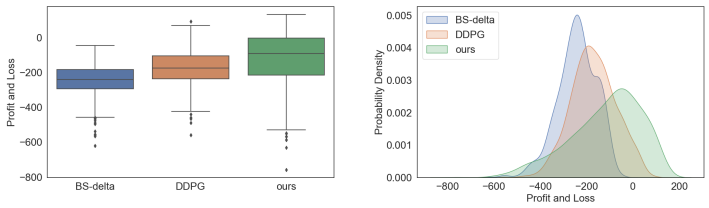

Figure 7: Boxplots and Probability density plot of hedging costs in comparative experiments

4.2.2 ABLATION EXPERIMENT

To verify the importance of considering historical information and focusing on cost distribution when hedging, we also conducted ablation experiments, and the results are shown in Table 3, where NoHist is the result without considering historical information, NoDist is the result of the value network only considering the expected value of hedging costs, BS-Sim is the result of simulating data only using BS model, and All is the result without melting any part.

Table 3: The evaluation of the metrics in ablation experiment.

Model	Return		Reward		Percent
	Mean	Std	Mean	Std	
NoHist	-138.6	143.75	-6.93	37.03	71.8%
NoDist	-147.74	163.85	-7.39	42.38	70.9%
BS-Sim	-127.13	130.31	-6.36	30.05	76.5%
All	-117.13	154.26	-5.90	32.31	77.4%

From Table 3, it can be observed that when historical information is ignored, or cost distribution is not considered, the average hedging cost significantly increases, and the proportion of hedging results better than delta hedging decreases. When only the BS model is used to generate data, although the standard deviation decreases slightly, the average hedging cost increases, and the proportion of results better than delta hedging also decreases. Next, the boxplots and the probability density plots are shown in Figure 8. It can be seen that when historical information is not considered, the cost distribution is ignored in value network evaluation strategies, or only the BS model is used to generate simulation data, the average P&L decrease. Overall, the distribution shifts to the left to some extent, indicating an increased probability of negative P&L and a decreased probability of positive P&L.

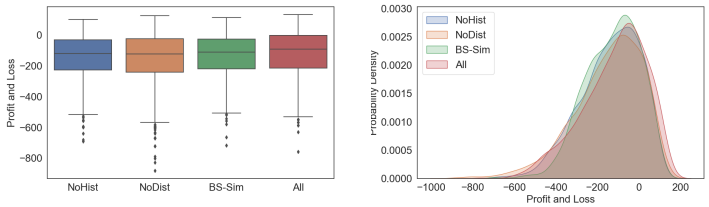

Figure 8: Boxplots and Probability density plot of hedging costs in ablation experiments

In summary, considering historical information and evaluating the P&L distribution is indispensable in option hedging. Moreover, when using RL methods for option hedging, generating option-related simulation data using various methods can enhance the agent's ability to adapt to market diversity.

4.3 Results and Analysis

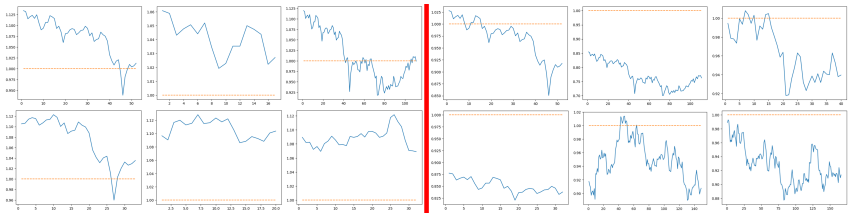

Figure 9: The change of moneyness for some SSE 50ETF options.

When analyzing the hedging effectiveness of our method, we randomly selected some SSE 50ETF options for experiments. These options included different strike prices and expiration dates. By observing the changes in the moneyness of these options, we found that the options hedged more effectively under our method compared to delta hedging were mainly in-the-money during their validity period (as shown in the left of Figure 9). Conversely, the options that performed better under delta hedging were mainly out-of-the-money during their validity period (as shown in the right of Figure 9). To further analyze the hedging process, we selected three SSE 50ETF options that were mostly in-the-money during their validity period to demonstrate the changes in cumulative P&L. The cumulative P&L have three sources: option P&L, underlying asset P&L, and transaction

costs. Additionally, we also displayed the changes in the underlying asset price, option price, and delta of these options during their validity period.

The code for the first option contract is 10005257, with an initial underlying price of 2.647, a strike price of 2.4, and a validity period of 36 days(including 26 trading days). Observing the hedging results (Table 4) and the process (Figure 10) under three methods, it was found that there were significant losses at the beginning and end of the hedging period for all three methods. This is because the underlying asset was not held before the hedging started. To mitigate the loss from selling the option, a substantial cost was incurred at the beginning of the hedging period to purchase the underlying asset. At the end of the hedging period, the underlying price was above the strike price, leading the buyer to exercise the option, resulting in significant losses for the seller. Additionally, as shown in Table 4, the total P&L without considering transaction costs for three methods were 88.877, 92.000, and 92.305, respectively, indicating they all effectively hedged the option when transaction costs were not considered. However, when transaction costs were taken into account, delta hedging incurred higher transaction costs, while the two RL methods had similar hedging effectiveness, each reducing by approximately 119, leading to lower total P&L.

The code for the second SSE 50ETF option contract is 10004596, with an initial underlying price of 2.651, a strike price of 2.5, and a validity period of 55 days(including 35 trading days). According to the hedging results for this option under three methods (Table 5), without considering transaction costs, the total P&L obtained by three methods were -41.912, 390, and 410.688, respectively. Compared to delta hedging, both RL methods achieved higher returns, with our method yielding slightly higher profits. When transaction costs were taken into account, both RL methods significantly reduced transaction costs (by approximately 50%) compared to delta hedging. Consequently, the final P&L were greatly improved, achieving results superior to delta hedging. From Figure 11, it is noted that both our method and DDPG method incurred more losses in the 9-17 day period when hedging this option. This could be due to the option being primarily out-of-the-money during this time, resulting in two RL methods being less effective than delta hedging. However, as the underlying asset's price continued to rise later on, the RL methods achieved high returns, which offset most of the losses incurred during that period, ultimately leading to a performance superior to delta hedging.

The code for the third SSE 50ETF option contract is 10003801, with an initial underlying price of 3.247, a strike price of 2.9, and a validity period of 169 days(including 111 trading days). According to the hedging results for this option under three methods (Table 6), without considering transaction costs, total P&L obtained by three methods were 271.882, 750.000, and 3274.05, respectively. Compared to delta hedging, both RL methods increased returns, with our method achieving significantly higher profits. Observing the changes in the underlying P&L during the hedging process for this option (Figure 12) and costs were taken into account, both RL methods significantly reduced transaction costs compared to delta hedging. Although our method had higher transaction costs than the DDPG method, the losses on the underlying assets were much smaller than those with the DDPG method, ultimately resulting in better performance than both delta hedging and DDPG hedging.

Based on above analysis, we believe that our approach is expected to lower costs and achieve more positive returns for in-the-money options, surpassing delta hedging. Yet, for out-of-the-money options, it's less effective than delta hedging. In such cases, we recommend alternative hedging strategies.

5 Conclusion

This paper proposes a new distributional reinforcement learning method based on historical information for option hedging. First, the historical state is included in the state variable, and the GRU is used to extract historical information, which is combined with the current information extracted from the full connection layer and then provided to the subsequent network layer for learning, to ensure that the agent can consider both current and historical market information when learning option hedging; Secondly, we set the output of the value network as a set of quantiles, and use Quantile Huber Loss function to fit the distribution of its evaluation, to evaluate the advantages and disadvantages of the strategy from the distribution rather than the expected value; Then, in order to make the data sources of agents diverse in the process of learning strategies, we use BS model, BI model and Heston model to simulate a large number of option data. Finally, the experimental results show that this method can significantly reduce hedging costs.

REFERENCES

Fischer Black and Myron Scholes. The pricing of options and corporate liabilities. *Journal of Political Economy*, 81(3):637–654, 1973. doi: 10.1086/260062. URL https://doi.org/10.1086/260062.

Loris Cannelli, Giuseppe Nuti, Marzio Sala, and Oleg Szehr. Hedging using reinforcement learning: Contextual k-armed bandit versus q-learning. *The Journal of Finance and Data Science*, 9: 100101, 2023. ISSN 2405-9188. doi: https://doi.org/10.1016/j.jfds.2023.100101. URL https://www.sciencedirect.com/science/article/pii/S240591882300017X.

Jay Cao, Jacky Chen, John Hull, and Zissis Poulos. Deep hedging of derivatives using reinforcement learning. *The Journal of Financial Data Science*, 3(1):10–27, December 2020. ISSN 2640-3943. doi: 10.3905/jfds.2020.1.052. URL http://dx.doi.org/10.3905/jfds.2020.1.052.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling, 2014. URL https://arxiv.org/abs/1412.3555.

John C. Cox, Stephen A. Ross, and Mark Rubinstein. Option pricing: A simplified approach. *Journal of Financial Economics*, 7(3):229–263, 1979. ISSN 0304-405X. doi: https://doi.org/10.1016/0304-405X(79)90015-1. URL https://www.sciencedirect.com/science/article/pii/0304405X79900151.

Jiayi Du, Muyang Jin, Petter Kolm, Gordon Ritter, Yixuan Wang, and Bofei Zhang. Deep reinforcement learning for option replication and hedging. *The Journal of Financial Data Science*, 2: 44–57, 10 2020. doi: 10.3905/jfds.2020.1.045.

Alexandru Giurca and Svetlana Borovkova. Delta hedging of derivatives using deep reinforcement learning. *Available at SSRN 3847272*, 2021.

Igor Halperin. Qlbs: Q-learner in the black-scholes(-merton) worlds. *The Journal of Derivatives*, 28 (1):99–122, 2020. doi: 10.3905/jod.2020.1.108. URL https://www.pm-research.com/content/iijderiv/28/1/99.

Ben Hambly, Renyuan Xu, and Huining Yang. Recent advances in reinforcement learning in finance. *Mathematical Finance*, 33(3):437–503, 2023. doi: https://doi.org/10.1111/mafi.12382. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/mafi.12382.

Steven L. Heston. A Closed-Form Solution for Options with Stochastic Volatility with Applications to Bond and Currency Options. *The Review of Financial Studies*, 6(2):327–343, 04 2015. ISSN 0893-9454. doi: 10.1093/rfs/6.2.327. URL https://doi.org/10.1093/rfs/6.2.327.

Petter N Kolm and Gordon Ritter. Dynamic replication and hedging: A reinforcement learning approach. *The Journal of Financial Data Science*, 1(1):159–171, 2019.

Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning, 2019. URL https://arxiv.org/abs/1509.02971.

Francesco Mandelli, Marco Pinciroli, Michele Trapletti, and Edoardo Vittori. Reinforcement learning for credit index option hedging, 2023. URL https://arxiv.org/abs/2307.09844.

Harry M Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):77–91, 1952.

Oskari Mikkilä. Optimal hedging with continuous action reinforcement learning. Master's thesis, 2020.

Oskari Mikkilä and Juho Kanniainen. Empirical deep hedging. *Quantitative Finance*, 23(1):111–122, 2023. doi: 10.1080/14697688.2022.2136037. URL https://doi.org/10.1080/14697688.2022.2136037.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning, 2013. URL https://arxiv.org/abs/1312.5602.

Andrei Neagu, Frédéric Godin, Clarence Simard, and Leila Kosseim. Deep hedging with market impact, 2024. URL https://arxiv.org/abs/2402.13326.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Christopher John Cornish Hellaby Watkins. Learning from delayed rewards. 1989.

Bo Xiao, Wuguannan Yao, and Xiang Zhou. Optimal option hedging with policy gradient. In *2021 International Conference on Data Mining Workshops (ICDMW)*, pp. 1112–1119, 2021. doi: 10.1109/ICDMW53433.2021.00145.

Cong Zheng, Jiafa He, and Can Yang. Option dynamic hedging using reinforcement learning, 2023. URL https://arxiv.org/abs/2306.10743.

A APPENDIX

A.1 SSE 50ETF-10005257

Table 4: The hedging results for SSE 50ETF option-10005257

Model	Option P&L	Underlying P&L	Cost	Total P&L
delta	372.000	-283.123	645.717	-556.840
DDPG(Cao et al., 2020)		-280.000	526.600	-434.600
ours		-279.695	526.603	-434.298

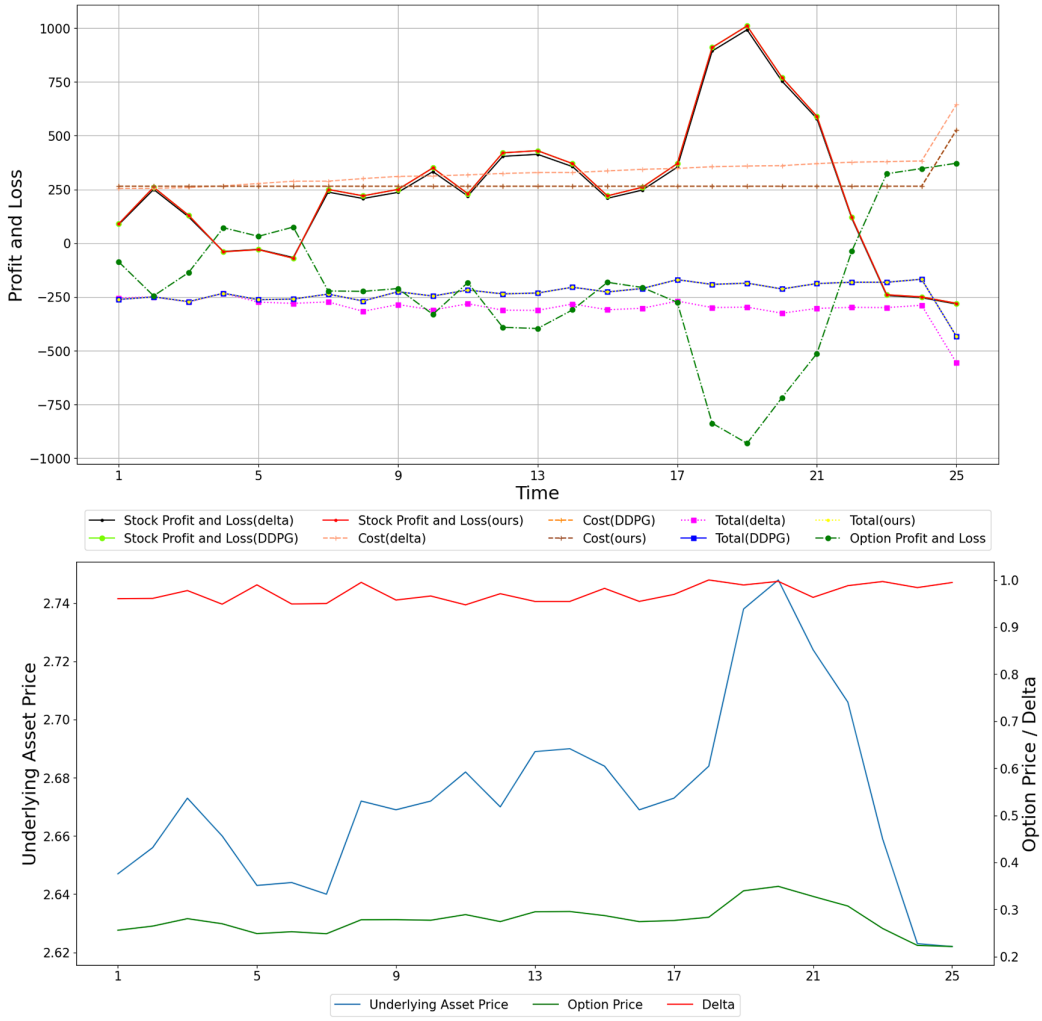

Figure 10: The hedging results for SSE 50ETF-10005257. The first figure displays the changes of accumulated P&L for the underlying assets, accumulated P&L for the options and the cumulative transaction costs under delta-hedging, DDPG model(Cao et al., 2020) and our model. The second figure illustrates the changes of the underlying asset prices, option prices, and delta values for three options throughout its validity periods.

A.2 SSE 50ETF-10004596

Table 5: The hedging results for SSE 50ETF option-10004596

Model	Option P&L	Underlying P&L	Cost	Total P&L
delta	1150.000	-1191.912	1043.721	-1085.633
DDPG(Cao et al., 2020)		-760.000	522.600	-132.600
ours		-739.312	530.889	-120.200

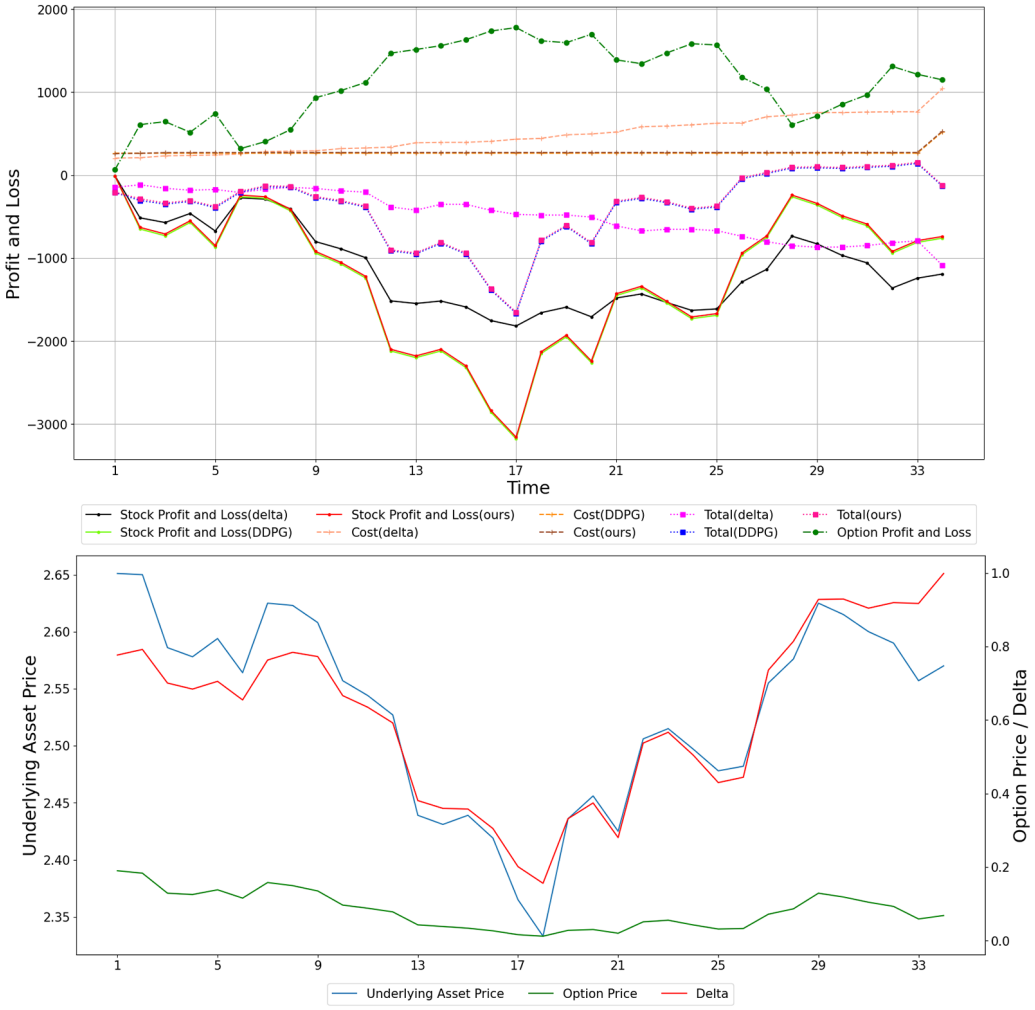

Figure 11: The hedging results for SSE 50ETF-10004596. The first figure displays the changes of accumulated P&L for the underlying assets, accumulated P&L for the options and the cumulative transaction costs under delta-hedging, DDPG model(Cao et al., 2020) and our model. The second figure illustrates the changes of the underlying asset prices, option prices, and delta values for three options throughout its validity periods.

A.3 SSE 50ETF-1003801

Table 6: The hedging results for SSE 50ETF option-10003801

Model	Option P&L	Underlying P&L	Cost	Total P&L
delta	4240.000	-3968.118	1918.169	-1646.287
DDPG(Cao et al., 2020)		-3490.000	614.500	135.500
ours		-965.950	893.316	2380.734

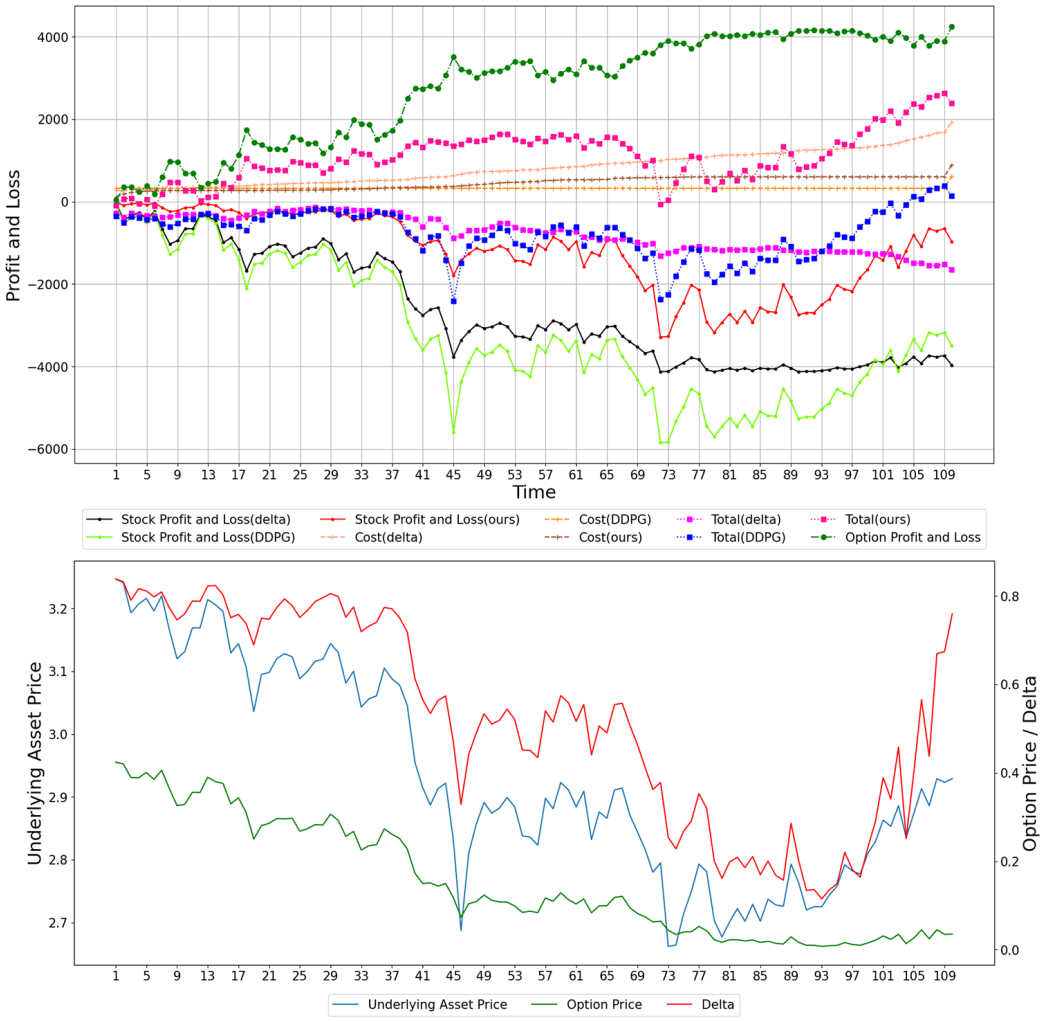

Figure 12: The hedging results for SSE 50ETF-10003801. The first figure displays the changes of accumulated P&L for the underlying assets, accumulated P&L for the options and the cumulative transaction costs under delta-hedging, DDPG model(Cao et al., 2020) and our model. The second figure illustrates the changes of the underlying asset prices, option prices, and delta values for three options throughout its validity periods.