JOIM
www.joim.com

Journal Of Investment Management, Vol. 22, No. 4, (2024), pp. 16–25

HEDGING BARRIER OPTIONS USING REINFORCEMENT LEARNING

Jacky Chen[a,c], *Yu Fu*[a,d], *John Hull*[a,*], *Zissis Poulos*[a,b], *Zeyu Wang*[a,d], *and Jun Yuan*[a,c]

We examine the use of reinforcement learning (RL) to hedge barrier options. We find that, when the hedger's objective is to minimize value at risk or conditional value at risk, RL is an attractive alternative to traditional hedging approaches. RL requires an assumption about the stochastic process followed by the underlying asset during the life of the exotic option, but our tests show that the results from using RL are fairly robust to this assumption. We do not consider transaction costs in this research. However, we show that RL involves less trading than traditional hedging approaches. As a result, the existence of transaction costs can be expected to increase the attractiveness of RL.

1 Introduction

Exotic options trade less actively than vanilla options, but attract higher bid-offer spreads. They are therefore potentially profitable to a market maker if hedging is handled well. Some exotic options such as Asian options are relatively easy to hedge while others are more difficult. Barrier options, which will be the focus of this paper, present particular problems to the hedger because the delta of the option usually exhibits a discontinuity when the asset price crosses the barrier.

As an example, consider a down-and-in put option where the barrier is at 10, the strike price is 11, and the time to maturity is 10 trading days. Assume geometric Brownian motion for the asset price with zero risk-free rate, zero dividend yield, and volatility equal to 30%. When the asset price is 10.1 and the barrier has not been passed, the delta is −1.308. When the asset price moves to 9.9 so that the instrument becomes a regular put, delta jumps by 0.35 to −0.958. (By contrast, for a vanilla put option, delta changes from −0.919 to −0.958 as the asset price changes from 10.1 to 9.9.)

One heuristic sometimes used by practitioners to handle barrier options is known as the "barrier

[a]Financial Innovation Hub (FinHub), a research center at the Joseph L. Rotman School of Management, University of Toronto.
[b]Assistant professor at York University in Toronto.
[c]We are adjunct professors at Rotman.
[d]We have research positions at Rotman.
*Corresponding author.

shift." The trader hedges as though the barrier is different from that in the contract. An argument consistent with this is provided by Broadie *et al.* (1997). Option pricing formulas assume that there is continuous monitoring to determine whether a barrier has been hit. When there is discrete monitoring (e.g., once a day), as is usually the case, Broadie *et al*. show that it is correct to value down (up) barrier options by decreasing (increasing) the barrier level in a model that assumes continuous monitoring.

A recent attempt to develop a discrete hedging strategy for barrier options is provided by Baule and Rosenthal (2022). These authors investigate the performance of different hedging strategies when the asset price is close to the barrier. However, they assume a mean–variance objective function for the hedger. Hedgers are naturally more concerned about outcomes where losses rather than gains result. We therefore choose a different setup from Baule and Rosenthal. Specifically, we consider two one-sided objective functions: value at risk (VaR) and conditional value at risk (CVaR). We assume hedges are rebalanced daily and use reinforcement learning to investigate whether multi-period strategies give better results than traditional hedging approaches for these objective functions.

Reinforcement learning has been used for hedging in other contexts. Buehler *et al*. (2019, 2022), Halperin (2017), and Cao *et al*. (2021) consider how reinforcement learning can improve the delta hedging of an option portfolio, particularly when trades in the underlying asset are subject to transaction costs. Cao *et al*. (2023) use reinforcement learning for gamma and vega hedging when a market maker is faced with hedging a portfolio of options and new options arrive stochastically. RL has also been used in portfolio management by Zhang and Aaraba (2022). Our current research complements that of Wu and Jaimungal (2023) and Jaimungal *et al*. (2022) who consider how reinforcement learning can be used in conjunction with objective functions that reward gains while protecting against downside risks.

RL approaches have a number of potential advantages over traditional hedging approaches. As we will show, RL has the advantage that it requires less trading on average than other hedging approaches and therefore leads to less transaction costs being incurred. It gives the user flexibility as far as the objective function is concerned. Furthermore, the multi-period approach that underlies reinforcement learning is consistent with the way the performance of a trader is usually assessed. We show that RL strategies are robust to the assumptions made about the process for the underlying asset and perform relatively well in stressed environments.

The rest of this paper is organized as follows. Section 2 introduces the RL approach. Section 3 presents our main results. Section 4 considers the use of vanilla options for hedging. Conclusions are in Section 5.

2 The RL Approach

Reinforcement learning is a procedure for choosing the actions that should be taken when particular states are encountered in multi-period decision-making. In our case, the actions are the hedging decisions (taken once a day) and the states are the time to maturity, the current holdings, and the underlying asset price. Typically, the decision-maker starts with a random policy and iteratively updates the policy so that the objective function is improved.

The methods we use are those described in Cao *et al*. (2023) and will not be explained in detail here. They involve the use of an actor neural network (NN) in conjunction with a critic NN. The actor NN implements the current hedging strategy at each iteration. The critic NN estimates the P&L

distribution at the hedging horizon for the current hedging strategy and computes gradients that lead to changes in the actions so that the objective function (VaR or CVaR in our case) is improved. We use a quantile representation of the P&L distribution at the hedging horizon with 1% intervals so that the distribution is represented by 100 points. The error between the empirical P&L distribution and the critic NN's estimate is measured using the Huber quantile loss.

The calculated gradients are used to iteratively update the actor NN's parameters, such that after each iteration the actor hedging policy improves. In early training stages, the critic's estimate is generally poor and thus the actor's hedging policy is far from optimal. As the critic improves so does the actor's policy to the point where both NNs converge to an optimum.

Our focus in this research is on non-breached barrier states. Upon breach, we switch to Black–Scholes–Merton delta hedging till expiry. This setup results in variable episode lengths for RL training, reflecting the steps before a breach. Two terminal conditions of an episode exist: (1) breach before expiry, and (2) mature without breach. For the latter, we follow the usual simulation and daily reward scheme as per Cao *et al.* (2023). For the former, the last step reward is the step P&L plus the cumulative P&L from Delta hedging post-breach.

It takes approximately one hour to train the RL strategy, involving 50,000 episode simulations using a single Nvidia RTX 4090 card.

3 Hedging Using the Underlying Asset

Barrier options are popular exotic options. When the asset price crosses the barrier level, the nature of the option changes. For example, a down-and-in-put option becomes a vanilla put option when the barrier is breached from above. Barrier options are often used in structured products. For example, in Japan a popular structured product called Uridashi includes barrier options. Barrier options are less expensive than the corresponding vanilla options because they provide the same payoff as the vanilla options only in some circumstances. (In the case of an "in" barrier option, the barrier must be hit for the option to become a vanilla option; in the case of an "out" barrier option, crossing the barrier leads to the option disappearing.) However, the barriers make it harder for market makers to hedge their risks when the asset price is close to the barrier. This is illustrated in Figure 1 which shows the delta of the down-and-in put option considered in the introduction is a discontinuous function of the asset price.

In this section, we compare reinforcement learning (RL) with traditional hedging approaches. It sometimes assumed that a hedger's objective should be to minimize the mean of the loss (gain) plus a constant times the standard deviation of the loss (gain). This assumes (somewhat unrealistically) that the hedger is equally averse to results in the two tails of the loss (gain) distribution. One advantage of RL is that the hedger has a great deal of freedom in the choice of the objective. We assume that the hedger's main concern is to avoid large losses. We consider two different objective

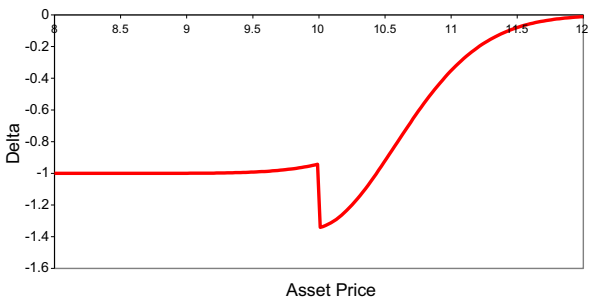

Figure 1 Delta given by the Black–Scholes–Merton model for a down-and-in-put as a function of the asset price. The barrier is at 10, the strike price is 11, volatility is 30%, and time to maturity is 10 days.

functions (both to be minimized). These are:

(a) VaR95: This is the 95th percentile of the loss (gain) distribution over the life of the option; and
(b) CVaR95: This is the expected loss over the life of the option when the loss is larger than the 95th percentile of the loss (gain) distribution.

RL procedures can be used for other objective functions. For example, they could be used for objective functions suggested in Wu and Jaimungal (2023), where results in the gain tail of the distribution are rewarded while those in the loss tail are penalized.

We consider three different strategies for hedging a short down-and-in put option for a particular objective function. These are:

(a) Delta hedging: Each day the trader takes a position in the underlying asset to neutralize the delta of the option;
(b) Myopic hedging: Each day the trader looks one day ahead and chooses the strategy that is optimal for the objective function being assumed; and
(c) RL Hedging: The trader uses reinforcement learning to choose a strategy, involving daily rebalancing, that is optimal for the objective function over the life of the option

When the barrier is breached so that the position being hedged is a short vanilla put option, it is assumed that the hedger switches to delta hedging.

3.1 Illustrative results

We illustrate our results by considering a situation where the asset price is 10.6, the barrier is at 10, the time to maturity is 20 days, the underlying put option is on 100 units of the asset, and hedges are adjusted once a day. We assume geometric Brownian motion for asset price movements with a volatility of 30%. For convenience, we assume a risk-free rate and dividend yield of zero. VaR95 and CVaR95 for a number of different strike prices are shown in Table 1.

Table 1 Values of the objective function when a short position in a down-and-in put barrier option is hedged in three different ways. The initial asset price is 10.6, the barrier is 10, and the volatility is 30%.

Strike price	Objective function	Delta	Myopic	RL
10.2	VaR95	11.60	11.64	10.80
	CVaR95	16.91	16.59	15.82
10.4	VaR95	13.62	14.07	13.07
	CVaR95	20.76	19.66	18.56
10.6	VaR95	16.65	17.02	14.41
	CVaR95	26.16	23.96	20.56
10.8	VaR95	19.38	20.91	16.63
	CVaR95	32.69	29.29	24.75
11.0	VaR95	23.13	25.85	18.90
	CVaR95	39.72	35.57	27.98

The table shows that RL improves the objective function when compared with simpler strategies. The amount of improvement increases as the strike price increases. This is because the hedger's potential liability from a breach of the barrier increases as the strike price increases and there is more to be gained from handling the hedging well. Interestingly, myopic hedging is an improvement over delta hedging when CVaR95 is used as the objective function, but the reverse is true for VaR95.

Figure 2 shows the terminal P&L distributions for the three hedging strategies. It illustrates that RL clearly outperforms at the left tail. This is achieved by accepting a slightly lower expected return.

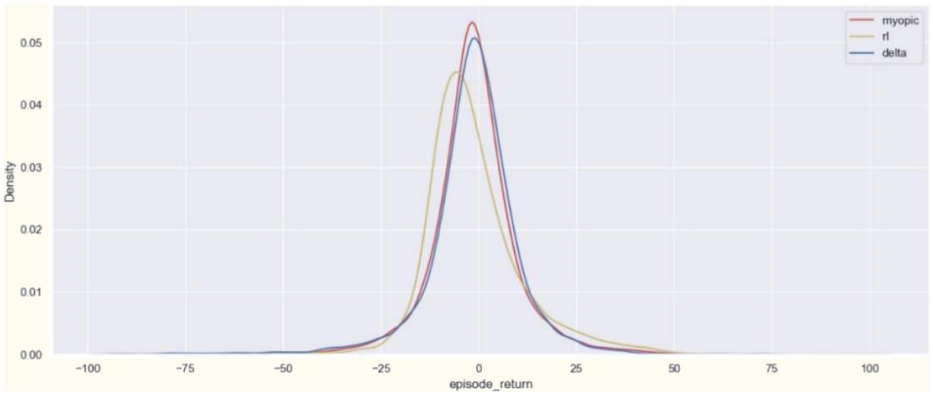

Figure 2 Terminal P&L distributions when a short position in a down-and-in put barrier option is hedged in three different ways. The initial asset price is 10.6, the barrier is 10, the strike price is 10.8 and the volatility is 30%.

Figure 3 examines the RL hedging strategy as a function of asset price and time to maturity. The top row compares the asset holdings of RL strategies with delta hedging. (Higher asset holdings are indicated by darker shades.) It shows that there tends to be less variation in asset holdings when the RL strategy is used than when delta hedging is used. This is indicative of less frequent trading, a point discussed further in Section 3.3.

The hedging ratio heat maps in the lower part of Figure 3 show that there is a tendency for RL to hedge less than delta as the asset price nears the barrier and to hedge more than delta as it

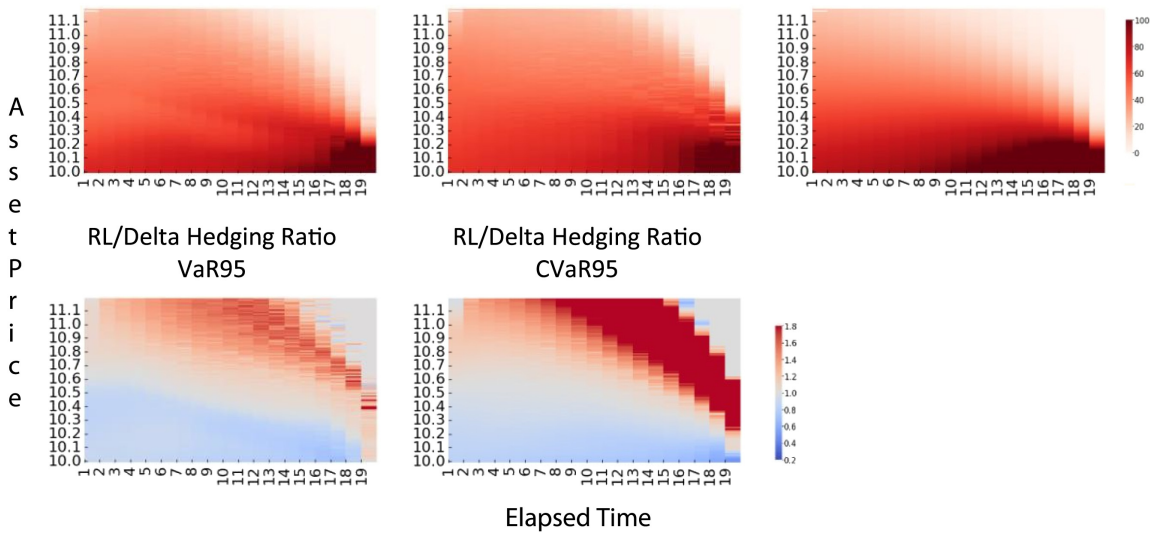

Figure 3 RL and Delta hedging strategies for a short down-and-in put option. In the top row, the heat maps show the short positions in different states. The bottom two heat maps show the ratio of short positions under RL to that under delta. (We apply a floor of 0.5 share on the asset holding to avoid zero denominator in division when calculating the hedging ratio.) In these ratio heat maps, white indicates a ratio of 1, blue indicates a ratio less than 1, and red indicates a ratio greater than 1. The initial asset price is 10.6, the barrier is 10, the strike is 10.6, initial time to maturity is 20 days, and the volatility is 30%.

moves away. This tendency is more pronounced for CVaR95 than VaR95. The RL strategy recognizes that delta will decline if the barrier is breached and therefore under-hedges relative to delta when near the barrier. It also recognizes that the cost of over-hedging when away from the barrier is small when compared to potential benefits. In the upper right corner, where the asset price surpasses a certain level and time to maturity is less than five days, both strategies involve virtually no hedging.

Figure 4 considers different P&L ranges: the top 10% cumulative loss, mid-range (quantile 45% to 55%), and top 10% cumulative profit. We select a representative asset price path closest to the highest density path at each time step using a Euclidean distance measure. For the top 10% cumulative loss, the asset price begins high and gradually descends, often crossing the barrier near expiry, leading to a significant loss, as shown in the representative path. RL's over-hedging for these paths, as depicted in Figure 3, somewhat mitigates these extreme losses. In the mid P&L range, the asset price exhibits two modes, either descending and breaching the barrier mid-life or ascending. In both cases, RL and Delta hedging perform similarly, with minor P&L differences, but RL hedging holdings are less variable. For the top 10% cumulative profit, the asset price oscillates above the barrier, corresponding to the under-hedging area in Figure 3. The RL portfolio benefits from adopting a more static hedging approach, characterized by fewer variations in the holding of the underlying asset.

3.2 Robustness tests

A potential disadvantage of RL is that it requires an assumption about the stochastic process that will apply to the asset price during the whole life of the exotic option. Delta hedging and myopic hedging are more flexible in that they can be based on the latest information on parameters of the stochastic process such as volatility.[1] Our robustness test assumes that RL is implemented on the assumption that volatility will be 30% when in fact other volatilities are observed. It is assumed that delta and myopic hedging are implemented using the observed volatilities.

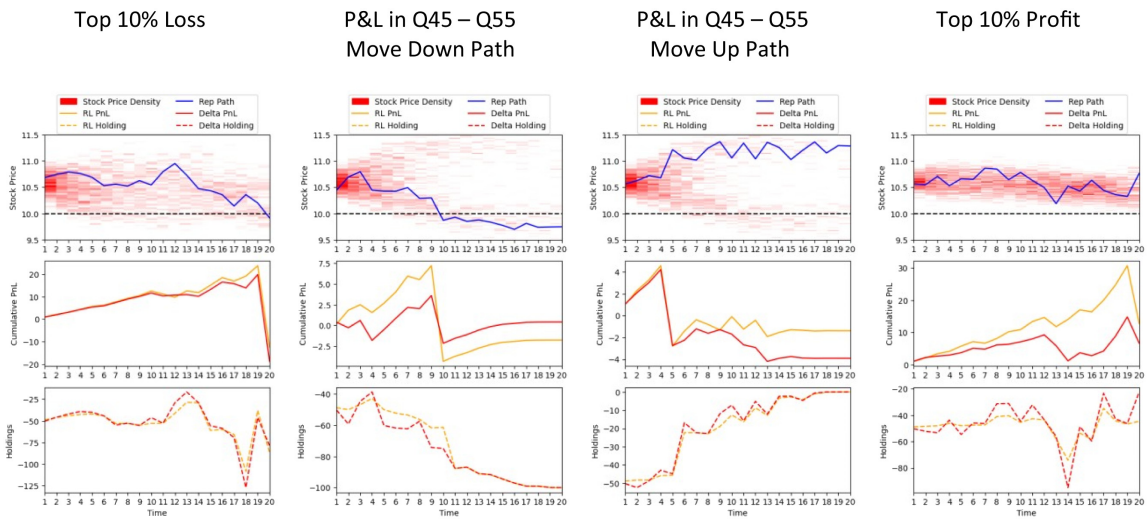

Figure 4 The upper charts show representative asset price trajectories and density distributions for different P&L Quantile Ranges. The middle charts show the cumulative P&L for the quantile ranges and the lower charts show the holdings.

Table 2 Values of the objective function when a short position in a down-and-in put barrier option is hedged in three different ways. The initial asset price is 10.6, the barrier is 10, and the strike is 10.8. Training for RL uses 30% volatility.

Volatility	Objective function	Delta	Myopic	RL
10%	VaR95	6.64	6.76	7.18
	CVaR95	16.56	16.34	15.31
20%	VaR95	17.31	19.87	15.39
	CVaR95	29.59	26.65	22.82
30%	VaR95	19.38	20.91	16.63
	CVaR95	32.69	29.29	24.75
40%	VaR95	21.38	22.86	19.68
	CVaR95	35.77	32.39	29.50
50%	VaR95	24.03	24.46	24.93
	CVaR95	37.91	35.71	34.98

In Table 2, we explore scenarios where the observed values of implied volatility are between 10% and 50%. This means that there is a potential mismatch between the volatility used during training and the actual market volatilities. Despite this mismatch, the results indicate that RL's hedging performance compares well with the delta and myopic strategies. For observed volatilities of 20% and 40%, RL produces better results for both objective functions. For the more extreme volatilities of 10% and 50%, RL underperforms slightly when VaR95 is the objective function but still produces improvements for CVaR95. Note that as the observed volatilities increase, the values of the objective functions increase for all hedging procedures. Hedging cannot completely neutralize the impact of a wider range of asset price movements.

3.3 Volume of trading

A key advantage of RL over delta and myopic hedging is that it involves less trading. To give a simple example of how this happens, assume a two-period binomial model. Suppose that the asset price moves so that delta increases from 0.5 to 0.7 in the first period and that the next asset price movement will lead, with equal probability, to delta returning to 0.5 or increasing to 0.9. Suppose further that the hedger wants to be fully delta-hedged at the end of the second period. When delta hedging is used the expected total of the changes to delta during the two periods is

$$0.2 + 0.5 \times (0.2 + 0.2) = 0.40$$

An RL hedger, looking two periods ahead, might choose not to hedge on the first day. The expected total of changes to delta during the two days is then

$$0.0 + 0.5 \times (0.0 + 0.4) = 0.20$$

This shows that the RL hedger does half as much trading on average.

In practice delta is a non-linear function of the asset price. This can give extra potential benefits from RL. If we change the above example so that delta increases to 0.8 or reduces to 0.5 with equal probability during the second period, the expected amount of trading done by the RL hedger is only 43% of that done by delta hedging.

Table 3 compares the volume of trading when RL is used with the volume of trading for delta and myopic hedging for the examples in Table 1. It can be seen that RL leads to between 20% and 30% less trading than the other strategies. Table 1 assumes zero transaction costs. The improvements from using RL can therefore be expected

Table 3 Average number of units of the asset traded during life of option for the examples in Table 1 (option is on 100 units of the asset).

	Delta	Myopic	RL
VaR95	254.85	259.65	184.95
CVaR95	254.85	238.50	181.62

to be even greater than those in Table 1 when transaction costs are considered. The results in Cao *et al*. (2021) and Cao (2023) illustrate the benefits of RL when there are transaction costs.

3.4 Stress testing results

We now evaluate the performance of different hedging strategies during extremely stressful periods. Scenarios that banks often utilize for stress testing purposes are those encountered during the Covid pandemic. Specifically, banks consider what are termed the L-Shape and V-Shape scenarios. In the L-Shape scenario, asset price movements closely resemble those of the S&P500 index from February 19th to March 18th, 2020. In the V-Shape scenario, they closely resemble those during the March 13th to April 13th – period where the S&P500 experienced a rebound after the central bank announced monetary support on March 23rd. Figure 5 shows the price changes observed during these two windows.

We continue to assume a Black–Scholes–Merton model with a 30% constant volatility. Table 4 presents our findings. It shows that under both the L-Shape and V-Shape scenarios RL outperforms both the Myopic and Delta strategies by about 20%. One interesting observation is that CVaR95 RL outperforms VaR95 RL in terms of P&L for both scenarios. This illustrates a general result that CVaR95 tends to perform better in severely adverse scenarios.

Table 4 Losses when a short position in a down-and-in put barrier option is hedged in several different ways. The initial asset price is 10.6, the barrier is 8.7, the strike is 10.8 and the volatility is 30%.

Scenario window	Delta	VaR95 myopic	CVaR95 myopic	VaR95 RL	CVaR95 RL
L-Shape	150.69	158.56	148.42	125.25	116.29
V-Shape	210.64	226.50	213.00	179.64	163.14

4 Using Vanilla Options

We now consider how the hedging of barrier options can be improved by using vanilla options. We consider the same 20-day options as in Table 1. We assume that the hedger can bring at-the-money options with five days to maturity into the portfolio each day. We consider two alternatives:

(a) The hedger uses RL to determine the option position taken each day.

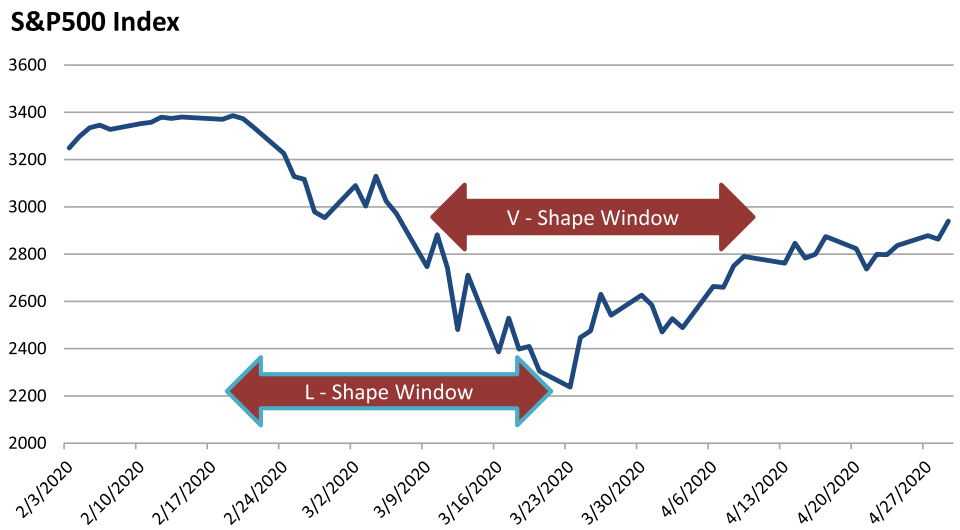

Figure 5 S&P500 price changes during the Covid period.

Table 5 Value of objective function when RL is used to determine positions in a vanilla option for hedging. Results are compared with the strategy where gamma is fully hedged each day and the strategy in Table 1 where RL is used to determine positions in the underlying asset. The position is made delta-neutral each day when options are used for hedging.

Strike price	Objective function	RL: vanilla option with delta neutrality	Gamma plus delta neutrality	RL using underlying (Table 1)
10.2	VaR95	8.01	9.77	10.80
	CVaR95	13.64	14.8	15.82
10.4	VaR95	9.02	11.09	13.07
	CVaR95	14.72	17.52	18.56
10.6	VaR95	11.98	13.48	14.41
	CVaR95	18.47	21.61	20.56
10.8	VaR95	13.99	15.96	16.63
	CVaR95	23.29	26.73	24.75
11.0	VaR95	16.22	18.99	18.90
	CVaR95	28.49	32.51	27.98

(b) The hedger takes the option position to neutralize gamma each day.

In both cases, after the option position has been taken, the underlying asset is traded to neutralize delta.

The results are shown in Table 5. Just as Table 1 shows that using RL to determine positions taken in the underlying asset is an improvement on delta hedging, Table 5 shows that using RL to take positions in options is an improvement on gamma hedging. Both alternatives are improvements on the best alternative for using only the underlying for hedging.

As illustrated in Table 3, RL leads to a non-trivial saving in transactions when the underlying is used for hedging. In our tests we found that a similar result holds when options are used for hedging. Option transactions are less than half as great when strategy (a) rather than strategy (b) is used. The size of transaction costs associated with trading options is much higher than that associated with trading the underlying asset. We find that when transaction costs are taken into account, RL using only the underlying can give better results than RL using options.

5 Conclusions

Traditionally, hedging derivatives has involved taking actions to manage the current values of the Greek letters (delta, gamma, vega, etc.). Reinforcement learning (RL) is a tool that allows the hedger to develop strategies that look several periods ahead. It has a number of advantages. It leads to strategies that involve less trading and therefore save transaction costs. It takes account of the frequency with which hedges will be adjusted. (Delta hedging assumes continual rebalancing.) It allows the hedger to specify an objective function that reflects more fully her risk preferences. (Potential gains and potential losses do not need to be treated symmetrically.) It can also be used to align the trader's objectives with the way the trader's performance is assessed. (For example, if the trader's performance is assessed over the next month, the hedging strategy can be based on a one-month time horizon.)

In earlier research, we have demonstrated the superior performance of hedging strategies derived from RL when there are transaction costs. This paper demonstrates that RL can be useful in hedging exotic options where delta is liable to be discontinuous even if there are no transaction costs. The existence of transaction costs makes RL a potentially even more attractive tool for the exotic options we consider. The performance of RL strategies is robust to the assumptions about the stochastic process followed by the underlying asset and our tests indicate that the RL-generated strategies handle stressed scenarios such as those seen in February–April, 2020, better than traditional strategies.

The results in this paper assume that the hedging strategy is based on a simple model reflecting market data at the start of the hedging period. A potential improvement involves allowing the strategy to be updated to reflect the latest market information. One approach is to use simulated training datasets in conjunction with real-world datasets. This will increase the relevance of the training scenarios and could make the strategies more robust in the dynamic market conditions of the real world. As more real-world data is accumulated, the adaptability and performance will potentially be enhanced. This continuous learning process will potentially allow the RL agent to evolve and improve its strategies over time, taking into account the latest market trends and anomalies.

Acknowledgments

This research was supported by the Financial Innovation Hub (FinHub) at the Joseph L. Rotman School of Management, University of Toronto.

Endnote

[1] Note that we assume no updating of the RL strategy during the life of the hedge.

References

Baule, R. and Rosenthal, P. (2022). "Time-Discrete Hedging of Down-and-Out Puts with Overnight Trading Gaps," *Journal of Risk and Financial Management* **15**, 29, https://doi.org/10.3390/jrfm15010029.

Broadie, M., Glasserman, P., and Kou, S. (1997). "A Continuity Correction for Discrete Barrier Options," *Mathematical Finance* **7**, 325–349.

Buehler, H., Gonon, L., Teichmann, J., and Wood, B. (2019). "Deep Hedging," *Quantitative Finance* **19**, 1271–1291, doi: 10.1080/14697688.2019.1571683.

Buelher, H., Murray, P., and Wood, B. (2022). Deep Bellman Hedging, arXiv.2207.00932.

Cao, J., Chen, J., Hull, J., and Poulos, Z. (2021). "Deep Hedging of Derivatives Using Reinforcement Learning," *Journal of Financial Data Science* **3**(1), 10–27.

Cao, J., Chen, J., Farghadani, S., Hull, J., Poulos, Z., Wang, Z., and Yuan, J. (2023). "Gamma and Vega Hedging Using Deep Distributional Reinforcement Learning," *Frontiers in Artificial Intelligence*. Section: Artificial Intelligence in Finance **6**, doi: 10.3389/frai.2023.1129370.

Halperin, I. (2017). "QLBS: Q-Learner in the Black-Scholes (-Merton) Worlds," arXiv.1712.04609, doi: 10.2139/ssrn.
3087076.

Jaimungal, S., Pesenti, S. M., Wang, Y. S., and Tatsat, H. (2022). "Robust Risk-Aware Reinforcement Learning," *SIAM Journal on Financial Mathematics* **13**(1), 213–226.

Wu, D. and Jaimungal, S. (2023). "Robust Risk-Aware Option Hedging," arXiv:2303.15216.

Zhang, C. and Aaraba, A. (2022). "Dynamic Optimal Portfolio Construction with Reinforcement Learning," Available at SSRN : https://ssrn.com/abstract=422316 or http://dx.doi.org/10.2139/ssrn.4221316.

Keywords: Reinforcement learning; barrier options; hedging