Enhancing Hedging Strategies with Deep Reinforcement Learning and Implied Volatility Surfaces

Carlos Octavio Perez Mendoza

A Thesis

in the Department

of

Mathematics and Statistics

Presented in Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy (Mathematics) at

Concordia University

Montréal, Québec, Canada

March 2025

Concordia University

School of Graduate Studies

This is to certify that the thesis prepared

By: **Mr. Carlos Octavio Perez Mendoza**

Entitled: **Enhancing Hedging Strategies with Deep Reinforcement Learning and Implied Volatility Surfaces**

and submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy (Mathematics)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the Final Examining Committee:

_____ Chair
Dr. Diego Elias Damasceno Costa

_____ External Examiner
Dr. Petter Kolm

_____ Arm's Length Examiner
Dr. Michel Denault

_____ Examiner
Dr. Cody Hyndman

_____ Examiner
Dr. Melina Mailhot

_____ Supervisor
Dr. Frédéric Godin

Approved by _____
Dr. Marco Bertola , Chair of Department
Department of Mathematics and Statistics

March 27, 2025

_____
Dr. Pascale Sicotte, Dean of Faculty
Faculty of Arts and Science

Abstract

Title: Enhancing Hedging Strategies with Deep Reinforcement Learning and Implied Volatility Surfaces

Carlos Octavio Perez Mendoza, Ph.D.

Concordia University, 2025

This thesis explores the use of deep reinforcement learning (DRL) to enhance dynamic option hedging by incorporating forward-looking market information, mitigating speculation, and optimizing portfolio rebalancing frequency. The first paper, *Enhancing Deep Hedging of Options with Implied Volatility Surface Feedback Information*, introduces a DRL-based hedging framework that leverages implied volatility surface data, improving hedging performance over traditional methods. The second paper, *Is the Difference between Deep Hedging and Delta Hedging a Statistical Arbitrage?*, examines whether deep hedging introduces speculative behavior in incomplete markets, demonstrating that proper risk measure selection prevents unwanted speculation. The third paper, *Implied-Volatility-Surface-Informed Deep Hedging with Options*, extends deep hedging by integrating implied volatility surface-informed decisions, no-trade regions, and multiple hedging instruments, improving cost efficiency and adaptability. This research contributes by defining frameworks that enhance existing techniques for managing risk in financial markets.

Keywords: Deep reinforcement learning, optimal hedging, implied volatility surfaces, arbitrage.

Acknowledgments

I would like to express my deepest gratitude to Frédéric Godin, my doctoral supervisor, who has guided and shaped my early steps in research throughout my PhD. I am truly grateful for his dedication, kindness, and unwavering support in my doctoral journey. I also deeply appreciate his invaluable contributions, from conceptualization to his guidance in implementing solutions to overcome the challenges we encountered in each of our projects.

I would also like to extend my heartfelt thanks to Geneviève Gauthier and Pascal François for their invaluable guidance and collaboration in my research projects. Their expertise, insightful feedback, and contributions to the manuscripts have been instrumental in bringing these projects to completion. I am deeply grateful for the countless hours of joint effort, their commitment, and their dedication.

Finally, and most importantly, I want to express my deepest appreciation to my family. To my partner, Maryam, who has stood by my side unconditionally throughout this journey, offering unwavering support and companionship. To my parents, Lilian and Octavio, without whom reaching this milestone would not have been possible, and to my siblings, Gustavo and Leonardo, for their constant encouragement. This thesis would not have been possible without your love, support, and patience.

Contribution of Authors

This manuscript-based thesis consists of three papers, each forming a main chapter.

Chapter 2: Carlos Octavio Perez Mendoza wrote the manuscript and conducted all numerical experiments. Pascal François, Geneviève Gauthier, and Frédéric Godin provided supervision, manuscript review, and editing support. The preprint version presenting the results is entitled *Enhancing Deep Hedging of Options with Implied Volatility Surface Feedback Information*.

Chapter 3: Carlos Octavio Perez Mendoza wrote the manuscript and conducted all numerical experiments. Pascal François, Geneviève Gauthier, and Frédéric Godin provided supervision, manuscript review, and editing support. The results were published in an article entitled *Is the Difference between Deep Hedging and Delta Hedging a Statistical Arbitrage?*.

Chapter 4: Carlos Octavio Perez Mendoza wrote the manuscript and conducted all numerical experiments. Pascal François, Geneviève Gauthier, and Frédéric Godin provided supervision, manuscript review, and editing support. The preprint version presenting the results is entitled *Implied-Volatility-Surface-Informed Deep Hedging with Options*.

Frédéric Godin served as the main supervisor, providing guidance on methodology and conceptualization. All authors contributed to the development of methodologies and research design. The final manuscript was reviewed and approved by all authors.

Contents

List of Figures xii

List of Tables xv

Introduction 1

Chapter 2 3

2.1 Introduction 4

2.2 The hedging problem 6

2.2.1 The hedging optimization problem 6

2.2.2 Reinforcement learning and deep hedging 8

2.2.2.1 Neural network architecture 9

2.2.2.2 Neural network optimization 10

2.3 Joint market dynamics 11

2.3.1 Daily implied volatility surface representation 11

2.3.2 Joint Implied Volatility and Return (JIVR) 11

2.4 Numerical study 12

2.4.1 Stochastic market generator 12

2.4.1.1 Market simulator . . . 12
2.4.1.2 Market parameters for numerical experiments . . . 13
2.4.2 Benchmarks . . . 13
2.4.3 Neural network settings . . . 14
2.4.3.1 Neural network architecture . . . 14
2.4.3.2 State space selection . . . 15
2.4.4 Benchmarking of hedging strategies . . . 19
2.4.4.1 Benchmarking over randomized economic conditions . . . 19
2.4.4.2 Impact of the state of the economy on performance . . . 21
2.4.4.3 Impact of moneyness level on performance . . . 23
2.4.4.4 Impact of option maturity on performance . . . 25
2.4.4.5 Impact of transaction costs on performance . . . 26
2.4.5 Global importance of IV surface features . . . 29
2.4.6 Backtesting . . . 31
2.5 Conclusion . . . 33
2.6 Appendix . . . 34
2.6.1 Details for the MSGD training approach . . . 34
2.6.2 Joint Implied Volatility and Return model . . . 35
2.6.2.1 Daily implied volatility surface . . . 35
2.6.2.2 Joint Implied Volatility and Return . . . 35
2.6.3 Benchmarks . . . 37
2.6.3.1 Black-Scholes model . . . 37

2.6.3.2 Leland model . . . 37
2.6.3.3 Smile-implied model . . . 38
2.6.4 Network fine-tuning . . . 38
2.6.4.1 Bounded strategies - leverage constraint . . . 38
2.6.4.2 Network architecture selection . . . 40
2.6.4.3 Dropout parameter selection . . . 42
2.6.4.4 Quadratic hedging problem . . . 44
2.6.4.5 JIVR Model parameters . . . 47
Chapter 3 **48**
3.1 Introduction . . . 49
3.2 Market model for hedging . . . 51
3.3 Hedging strategies . . . 52
3.3.1 Deep hedging . . . 52
3.3.2 Delta hedging . . . 52
3.3.3 Statistical arbitrage . . . 53
3.4 Numerical study . . . 54
3.4.1 Stochastic market dynamics . . . 54
3.4.2 Comparative analysis of deep hedging and delta hedging strategies . . . 55
3.4.3 Robustness assessment . . . 59
3.4.3.1 Robustness assessment across option maturities and moneyness levels . . . 59

3.4.3.2 Robustness assessment across different economic conditions 60
3.4.3.3 The case of a straddle option 61
3.5 Conclusion 63
3.6 Appendix 64
3.6.1 The deep hedging algorithm 64
3.6.2 Details for the MSGD training approach 65
Chapter 4 **67**
4.1 Introduction 68
4.2 The hedging problem 71
4.2.1 The hedging optimization problem 71
4.2.2 Reinforcement learning and deep hedging 73
4.2.2.1 Neural network architecture 74
4.2.3 Neural network optimization 75
4.3 Market simulator 75
4.3.1 Daily implied volatility surface 76
4.3.2 Joint Implied Volatility and Return 76
4.4 Numerical study 77
4.4.1 Market settings for numerical experiments 77
4.4.2 Benchmarks 78
4.4.3 Neural network settings 79

4.4.3.1 Neural network architecture . . . 79
4.4.3.2 State space . . . 79
4.4.4 Benchmarking of hedging strategies . . . 80
4.4.4.1 Benchmarking in the absence of transaction costs . . . 80
4.4.4.2 Benchmarking in the presence of transaction costs . . . 83
4.4.4.3 Impact of no-trade regions . . . 88
4.4.5 Assessing the presence of speculative components in hedging positions 90
4.4.5.1 Risk premium and good deals . . . 90
4.4.6 Statistical study and sensitivity analysis of hedging strategies . . . 91
4.4.6.1 Statistical analysis of hedging option positions: Benchmarks vs RL agents . . . 91
4.4.6.2 Sensitivity analysis . . . 94
4.4.7 Backtesting . . . 95
4.5 Conclusion . . . 97
4.6 Appendix . . . 98
4.6.1 Neural network settings . . . 98
4.6.1.1 Network architecture . . . 98
4.6.1.2 Details for the MSGD training approach . . . 99
4.6.2 Joint Implied Volatility and Return model . . . 101
4.6.2.1 Daily implied volatility surface . . . 101
4.6.2.2 Joint Implied Volatility and Return . . . 101
4.6.3 Benchmarks . . . 103

4.6.3.1 Leland Model . . . 103
4.6.3.2 Delta gamma hedging . . . 103
4.6.4 Soft constraint regularization . . . 104
4.6.5 Impact of state variable inclusion on hedging performance . . . 106
4.6.6 Statistical arbitrage . . . 106
4.6.7 Systematic outperformance of RL agents . . . 108
4.6.8 JIVR Model parameters . . . 109

Conclusion **111**

Bibliography **113**

List of Figures

Figure 2.1 Optimal penalty function value for a short position in an ATM call option with maturity of 63 days under various state spaces and transaction cost levels. 17

Figure 2.2 Hedging metrics for a short position in an ATM call option with maturity of 63 days under different states of the economy. 22

Figure 2.3 Hedging metrics for a short position in OTM, ATM and ITM call options with a maturity of 63 days. 24

Figure 2.4 Hedging metrics for a short position in ATM call options with maturities of 21, 63 and 126 days. 25

Figure 2.5 Hedging metrics for a short position in an ATM call option with a maturity of 63 days under different transaction cost levels. 27

Figure 2.6 Normalized global importance of features when hedging a European call options with a maturity $N = 63$ days, across various moneyness levels. . 30

Figure 2.7 Global importance of features when hedging a European ATM call options, across various maturities. 31

Figure 2.8 Cumulative P&L for ATM call options with a maturity of 63 days under real asset price dynamics. 32

Figure 2.9 Doubling strategy dynamics for a short position in a ATM call option with a maturity of 63 days. 40

Figure 2.10 Network performance for a short position in a ATM call option with a maturity of 63 days. 41

Figure 2.11 Neural network performance for a short position in an ATM call option with a maturity of 63 days: sensitivity to the state of the economy. 42
Figure 2.12 RNN-FNN performance for a short position in an ATM call option with a maturity of 63 days: the effect of the dropout parameter in the training phase. 43
Figure 2.13 RNN-FNN loss function for a short position ATM call option with maturity $N = 63$ days and 16 time steps for rebalancing. 46
Figure 3.1 P&L distribution of the strategy δ^-. 57
Figure 3.2 Risk for the differential strategy, $\rho(-V_T^{\delta^-}(0))$, evaluated across different maturities and moneyness levels. 60
Figure 3.3 Risk for the differential strategy, $\rho(-V_T^{\delta^-}(0))$, evaluated across different market conditions. 61
Figure 4.1 Hedging error distribution in the absence of transaction costs. 83
Figure 4.2 Hedging error distribution in the presence of transaction costs. 87
Figure 4.3 Evolution of rebalancing day proportions and average hedging costs at different transaction cost levels. 89
Figure 4.4 Scatter plot from ranked data of risk premium and hedging option positions. 91
Figure 4.5 Pearson correlation between DG and RL agent's hedging option positions. 92
Figure 4.6 Distribution of hedging option positions. 93
Figure 4.7 Marginal impact on hedging positions with respect to IV coefficients and underlying asset volatility. 95
Figure 4.8 Cumulative P&L for a ATM straddle instruments with a maturity of 63 days under real asset price dynamics. 96
Figure 4.9 Hedging error distribution under real asset price dynamics. 97
Figure 4.10 Optimal penalty function and soft constraint values for various penalization parameter values, applied to a straddle with a maturity of $T = 63$ days. 105

Figure 4.11 P&L distribution of the strategy ϕ^-. 108
Figure 4.12 Empirical distribution of penalty functions for a straddle with maturity of $T = 63$ days. 109

List of Tables

Table 2.1 Aggregated hedging metrics for a short position in an ATM call option with maturity of 63 days. 20
Table 2.3 Clusters of dates representing different time periods 21
Table 2.4 Hedging costs when hedging a short ATM call option position with a maturity of $N = 63$ days under various transaction cost levels. 28
Table 2.6 RNN-FNN hedging error statistics of a short position in a ATM call option with maturity of 63 days. 39
Table 2.8 RNN-FNN hedging error statistics for a short position ATM call option with two different maturities and rebalancing periods under the MSE as penalty function. 45
Table 2.10 Estimated Gaussian copula parameters 47
Table 2.11 JIVR model parameter estimates 47
Table 3.1 Performance assessment for deep hedging, delta hedging and their difference over a short position on an ATM call option with maturity $T = 63$ days. 56
Table 3.2 Statistical relationships between positions of delta hedging and deep hedging. 58
Table 3.3 Performance assessment for deep hedging, delta hedging and their difference over a short position on an ATM straddle strategy with maturity $T = 63$ days. 62
Table 4.1 State variables 80

Table 4.2 Hedging performance metrics under the assumption of zero transaction costs. 81
Table 4.3 Optimal rebalancing threshold l values of DG and RL strategies. . . . 84
Table 4.4 Optimal risk measure values of deep hedging, delta hedging, and delta gamma hedging. 86
Table 4.5 Optimal risk measure values for different state space configurations. . 106
Table 4.6 Statistical arbitrage statistic 107
Table 4.7 Estimated Gaussian copula parameters 109
Table 4.8 JIVR model parameter estimates 110

Introduction

The evolution of hedging strategies in financial markets has been significantly influenced by advances in machine learning and computational finance. Traditional hedging approaches, such as delta and delta-gamma hedging, rely on parametric models and assumptions that often simplify market dynamics. While these methods are widely used, their effectiveness can be limited in complex and volatile market conditions. The emergence of deep hedging, introduced by Buehler et al. (2019), provides a data-driven alternative that learns optimal hedging strategies from market data. By leveraging deep reinforcement learning (DRL), deep hedging adapts dynamically to changing conditions without explicitly specifying an underlying stochastic model. While deep hedging has shown significant flexibility and adaptability (e.g., Cao et al. (2020), Carbonneau (2021), and Cao et al. (2023)), the integration of forward-looking information into its framework remains largely unexplored.

This thesis is guided by three main objectives: (1) improving hedging performance through DRL algorithms that leverage implied volatility surfaces, demonstrating superior adaptability compared to traditional delta and delta-gamma hedging; (2) addressing concerns regarding about the speculative components inherent in deep hedging, particularly its potential to inadvertently generate statistical arbitrage under specific market conditions; and (3) examining the dynamics of DRL-generated hedging policies by employing global feature importance techniques and advanced statistical methods to gain a more comprehensive understanding of the underlying decision-making process.

The thesis is structured into three main chapters, each presented as an independent paper. Chapter 2 presents a deep hedging framework for European options using policy gradient reinforcement learning. A key innovation is the inclusion of IV surface data as an additional input to the hedging agent. By leveraging this information, the model refines its risk assessment, leading to improved hedging performance relative to both traditional delta hedging and standard deep hedging approaches. Empirical results from simulations and backtesting demonstrate that incorporating IV surfaces enhances the model's ability to adapt to different market conditions.

Chapter 3 investigates the relationship between deep hedging and statistical arbitrage in an incomplete market setting. While it has been shown that the difference between deep hedging and replicating portfolio strategies could introduce speculative components, we test this claim within a GARCH-based market model. The findings suggest that deep hedging may introduce a speculative component if the risk measure does not sufficiently penalize adverse outcomes. However, selecting an appropriate risk measure mitigates this effect.

Chapter 4 introduces an enhanced deep hedging approach designed to hedge portfolios of options using multiple hedging instruments. The model dynamically integrates IV surface evolution to refine decision-making and risk assessment. The study also highlights the benefits of optimizing hedging frequency, demonstrating that a well-calibrated deep hedging model can achieve superior performance while reducing unnecessary rebalancing.

The bibliography for all papers is presented at the end of the thesis.

Chapter 2

Enhancing Deep Hedging of Options with Implied Volatility Surface Feedback Information

Abstract

We present a dynamic hedging scheme for S&P 500 options, where rebalancing decisions are enhanced by integrating information about the implied volatility surface dynamics. The optimal hedging strategy is obtained through a deep policy gradient-type reinforcement learning algorithm, with a novel hybrid neural network architecture improving the training performance. The favorable inclusion of forward-looking information embedded in the volatility surface allows our procedure to outperform several conventional benchmarks such as practitioner and smiled-implied delta hedging procedures, both in simulation and backtesting experiments.

JEL classification: C45, C61, G32.

Keywords: Deep reinforcement learning, optimal hedging, implied volatility surfaces.

2.1 Introduction

Since the advent of the Black and Scholes (1973) framework, dynamic hedging has become a standard financial risk management tool for managing the risk associated with options portfolios. The Black and Scholes (1973) framework has the remarkable property that delta hedging –a hedging strategy invested exclusively in the underlying asset and the money market account– achieves the perfect replication of a European-style contingent claim. In practice, this property is lost due to frictions. Most notably, considering the infrequent rebalancing of the hedging portfolio, classic delta hedging, which is inherently local, can no longer protect against infinitesimal shocks in the underlying asset price. Such an imperfect hedge inevitably yields a hedging error that has to be managed.

Several works have subsequently extended the idealized setting of the Black-Scholes framework to account for imperfect hedging, incorporating features such as discrete-time rebalancing (Boyle and Emanuel, 1980), transaction costs (Leland, 1985; Boyle and Vorst, 1992; Toft, 1996; Meindl and Primbs, 2008; Zakamouline, 2009; Lai and Lim, 2009), trading constraints (Edirisinghe et al., 1993) or liquidity costs (Frey, 1998; Cetin et al., 2010; Guéant and Pu, 2017). Another very important avenue for the development of a hedging procedure is the computation of the delta (and other "Greek" sensitivity parameters) which purely relies on market data and does not require stringent postulates about stochastic dynamics of associated risk factors: see among others, Bates (2005), Alexander and Nogueira (2007), and François and Stentoft (2021). This approach utilizes implied volatility (IV) surfaces to derive the Greeks, specifically the option delta and gamma, based on a mild scale-invariance assumption for the underlying asset return distribution.

This paper builds on existing literature by developing a data-driven approach that incorporates IV surface information to derive optimal hedging positions. Unlike traditional methods focusing on local conditions, we adopt a global perspective, minimizing the risk metric associated with the terminal hedging error in a multi-period horizon framework. We leverage

developments in risk-aware reinforcement learning (RL) to find the sequence of hedge ratios optimizing the hedger's total risk until the option expiry, a setup analogous to Buehler et al. (2019)'s deep hedging approach. The novelty of our work lies in integrating factors that influence IV surface dynamics into the state variable set for determining hedging positions. The hedging agent therefore relies on market expectations for the underlying return distributions over several temporal horizons, producing genuine optimal forward-looking multi-stage hedging decisions informed by IV surfaces. To conduct our numerical experiments, we use the JIVR model of François et al. (2023), which is a parsimonious, tractable and data-driven econometric model representing the joint dynamics of the underlying asset return and five interpretable factors driving the IV surface of the S&P 500. Notably, the model has been calibrated on a data period spanning more than 25 years of daily option data, reflecting market behavior in a wide array of scenarios, including several financial crises.

As an additional contribution, we propose a novel hybrid neural network architecture combining feedforward and long-short-term memory (LSTM) layers, which is shown the improve training performance over conventional architecture.

Hedging procedures with risk-aware reinforcement learning procedures have received substantial attention from the literature recently, see for instance Halperin (2019), Cao et al. (2020), Du et al. (2020), Carbonneau and Godin (2021), Carbonneau (2021), Horvath et al. (2021), Imaki et al. (2021), Lütkebohmert et al. (2022), Cao et al. (2023), Carbonneau and Godin (2023), Marzban et al. (2023a), Mikkilä and Kanniainen (2023), Pickard and Lawryshyn (2023), Raj et al. (2023), Wu and Jaimungal (2023) and Neagu et al. (2024). Nevertheless, to the best of our knowledge, our paper is the first to conduct multi-stage RL-based hedging that incorporates IV surface information directly as state variables.

Our RL approach is shown to substantially outperform delta-based hedging strategies acting as benchmarks. In particular, RL agents trained with asymmetric objective functions such as the conditional value-at-risk (CVaR) or the semi-mean squared-error (SMSE) offer superior

tradeoffs between profitability and downside risk. The outperformance of RL agents is even more pronounced in the presence of transaction costs, as such agents manage to develop hedging strategies that remain efficient in terms of risk mitigation while generating lower turnover. A feature importance analysis highlights that the conditional variance of the underlying asset returns, the level of the IV surface and its slope all significantly influence hedging performance, irrespective of the risk metric employed within the objective function.

The rest of the paper is organized as follows. Section 2.2 frames the hedging problem in terms of a deep reinforcement learning framework. Section 2.3 provides the components of the JIVR model. Section 2.4 presents the numerical results, assessments, and global feature importance analysis.[1] Section 2.5 concludes.

2.2 The hedging problem

The mathematical formulation of the hedging problem considered herein, along with the solution approach based on deep reinforcement learning, are described in this section.

2.2.1 The hedging optimization problem

This paper introduces a dynamic hedging strategy for European-style options that leverages insights provided by the implied volatility surface. The approach aims to minimize some risk measure applied to the terminal hedging error.

The European option payoff $\Psi(S_T)$ depends on the price of the underlying asset at maturity, denoted as T trading days. The hedging strategy involves managing a self-financing portfolio composed of both the underlying asset and a risk-free asset, with daily rebalancing. The strategy is represented by the predictable process $\{(\phi_t, \delta_t)\}_{t=1}^{T}$, where ϕ_t is the cash held at time $t-1$ and carried forward to the next period, and δ_t denotes the number of shares of the

[1]The Python code to replicate the numerical experiments from this paper can be found at the following link: https://github.com/cpmendoza/DeepHedging_JIVR.git.

risky asset S held during the interval $(t-1,t]$. The time-t portfolio value is

$$V_t^\delta = \phi_t \mathrm{e}^{r_t \Delta} + \delta_t S_t \mathrm{e}^{q_t \Delta}$$

where r_t is the time-t annualized continuously compounded risk-free rate, q_t is the annualized underlying asset dividend yield, both on the interval $(t-1,t]$, and $\Delta = \frac{1}{252}$. To account for transaction costs the self-financing condition entails

$$\phi_{t+1} + \delta_{t+1} S_t = V_t^\delta - \kappa S_t \mid \delta_{t+1} - \delta_t \mid, \tag{2.1}$$

where κ is the rate of proportional transaction costs.

The optimal hedging problem is an optimization task where an agent seeks to minimize the risk exposure associated with a short position in the option. More precisely, it is a sequential decision problem where the agent looks for the best sequence of actions $\delta = \{\delta_t\}_{t=1}^T$ that minimizes a penalty function ρ applied to the hedging error at maturity for a short position, defined as

$$\xi_T^\delta = \Psi(S_T) - V_T^\delta.$$

Note that ξ_T^δ is a loss variable, with profits being represented by $-\xi_T^\delta$. The problem is

$$\delta^* = \arg\min_\delta \rho\left(\xi_T^\delta\right), \tag{2.2}$$

where ρ is a risk measure, acting as the penalty function. Each time-t action $(\phi_{t+1}, \delta_{t+1})$ is of feedback-type, with such decision being a function of current available information on the market: $\delta_{t+1} = \tilde{\delta}(X_t)$ for some function $\tilde{\delta}$ with state variables vector X_t. Section 2.4.3.2 further describes these state variables that include the underlying asset current value as well as some information about the implied volatility surface, among others. Due to Equation (2.1), ϕ_{t+1} is fully characterized when δ_{t+1} is specified, and as such the time-t action to be

chosen is simply δ_{t+1}.

In this paper we consider three penalty functions that are very popular in the literature:

- Mean Square Error (MSE): $\rho(\xi_T^\delta) = \mathbb{E}[(\xi_T^\delta)^2]$.
- Semi Mean-Square Error (SMSE): $\rho(\xi_T^\delta) = \mathbb{E}[(\xi_T^\delta)^2 \mathbb{1}_{\{\xi_T^\delta \geq 0\}}]$.
- Conditional Value-at-Risk (CVaR_α): $\rho(\xi_T^\delta) = \mathbb{E}[\xi_T^\delta \mid \xi_T^\delta \geq \text{VaR}_\alpha(\xi_T^\delta)]$, where $\alpha \in (0,1)$ and $\text{VaR}_\alpha(\xi_T^\delta)$ is the Value-at-Risk defined as $\text{VaR}_\alpha(\xi_T^\delta) = \min_c\{c : \mathbb{P}(\xi_T^\delta \leq c) \geq \alpha\}$ and $\alpha = 95\%$ or 99% in this work.

The MSE was first proposed in the seminal variance-optimal hedging framework of Schweizer (1995), which was later extended to the multivariate case by Rémillard and Rubenthaler (2013). The SMSE is a particular case of the asymetric polynomial penalty considered for instance subsequently in Pham (2000), François et al. (2014) and Carbonneau and Godin (2023). It provides the advantage over the MSE to avoid penalizing hedging gains. Lastly, we consider CVaR as a standard metric for measuring potential catastrophic tail events, frequently mandated by financial regulators for use in financial institutions. This metric has also been explored in global hedging contexts in Melnikov and Smirnov (2012), Godin (2016), Buehler et al. (2019), Carbonneau and Godin (2021) or Cao et al. (2023), among others.

Historically, an alternative method for addressing the problem described in Equation (2.2) employs backward recursion within dynamic programming frameworks. However, this approach is hindered by the curse of dimensionality, limiting its practical applicability. To overcome these limitations, we tackle Problem (2.2) using reinforcement learning.

2.2.2 Reinforcement learning and deep hedging

The optimal hedging problem (2.2) can be formulated as a reinforcement learning (RL) problem because it is a feedback sequential decision-making task. In this framework, an agent learns a policy (the investment strategy δ) which dictates trading actions to be applied as a function of state variables to minimize the hedging objective highlighted in Equation (2.2).

More precisely, the problem consists in learning the mapping $\tilde{\delta}$.

Consistently with Buehler et al. (2019), the problem established in (2.2) is solved by direct estimation of the policy through a deep policy gradient approach that relies on the estimation of the mapping $\tilde{\delta}$ by an Artificial Neural Network (ANN). Denoting by $\tilde{\delta}_\theta$ the policy obtained when $\tilde{\delta}$ is estimated by an ANN with parameters θ, the objective function we need to minimize is thus

$$\mathcal{O}(\theta) = \rho\left(\xi_T^{\tilde{\delta}_\theta}\right). \tag{2.3}$$

2.2.2.1 Neural network architecture

We propose a nonstandard architecture which consists in a Recurrent Neural Network with a Feedforward Connection (RNN-FNN) that combines the traditional Long Short-Term Memory Network (LSTM) and the Feedforward Neural Network (FFNN) architectures. The inclusion of LSTM layers mitigates problems related to vanishing gradients.[2] The RNN-FNN network is defined as a composition of LSTM cells $\{C_l\}_{l=1}^{L_1}$ and FFNN layers $\{\mathcal{L}_j\}_{j=1}^{L_2}$ under the following functional representation:

$$\tilde{\delta}_\theta(X_t) = (\underbrace{\mathcal{L}_J \circ \mathcal{L}_{L_2} \circ \mathcal{L}_{L_2-1} \circ ... \circ \mathcal{L}_1}_{\text{FFNN layers}} \circ \underbrace{C_{L_1} \circ C_{L_1-1} ... \circ C_1}_{\text{LSTM cells}})(X_t).$$

The LSTM cell C_l is a mapping that transforms a vector $Z_t^{(C,l-1)}$ of dimension $d^{(C,l-1)}$ into a vector $Z_t^{(C,l)}$ of dimension $d^{(C,l)}$ based on the following equations, considering $Z_t^{(C,0)} = X_t$:

$$\begin{aligned}
i^{(l)} &= \text{sigm}(W_i^{(l)} Z_t^{(C,l-1)} + b_i^{(l)}),\\
o^{(l)} &= \text{sigm}(W_o^{(l)} Z_t^{(C,l-1)} + b_o^{(l)}),\\
c^{(l)} &= i^{(l)} \odot \tanh(W_c^{(l)} Z_t^{(C,l-1)} + b_c^{(l)}),\\
Z_t^{(C,l)} &= o_t^{(l)} \odot \tanh(c^{(l)}),
\end{aligned}$$

[2]Vanishing gradients arise when the gradients of the penalty function become extremely small, leading to slow or halted training (details can be found in Goodfellow et al. (2016)).

where sigm$(\cdot)$ and tanh$(\cdot)$ are respectively the sigmoid and hyperbolic tangent functions applied element-wise and $\odot$ is the Hadamard product. Layer $\mathcal{L}_j$ represents a FFNN layer that maps the input vector $Z_t^{(\mathcal{L},j-1)}$ of dimension $d^{(\mathcal{L},j-1)}$ into a vector $Z_t^{(\mathcal{L},j)}$ of dimension $d^{(\mathcal{L},j)}$ by applying a linear transformation $T_{\mathcal{L}_j}(Z_t^{(\mathcal{L},j-1)}) = W_{\mathcal{L}_j} Z_t^{(\mathcal{L},j-1)} + b_{\mathcal{L}_j}$ and, subsequently, an element-wise non-linear activation function $g_{\mathcal{L}_j}$, i.e., $\mathcal{L}_j(Z_t^{(\mathcal{L},j-1)}) = (g_{\mathcal{L}_j} \circ T_{\mathcal{L}_j})(Z_t^{(\mathcal{L},j-1)})$ for $j \in \{1, ..., L_2, J\}$, considering $Z_t^{(\mathcal{L},0)} = Z_t^{(C,L_1)}$.

The trainable parameters θ of the RNN-FNN network are listed below:

- If $L_1 \geq l \geq 1$: $W_i^{(l)}$, $W_o^{(l)}$, $W_c^{(l)} \in \mathbb{R}^{d^{(C,l)} \times d^{(C,l-1)}}$ and $b_i^{(l)}$, $b_o^{(l)}$, $b_c^{(l)} \in \mathbb{R}^{d^{(C,l)} \times 1}$ with $d^{(C,0)}$ defined as the original dimension of the network input.
- If $L_2 \geq j \geq 1$: $W_{\mathcal{L}_j} \in \mathbb{R}^{d^{(\mathcal{L},j)} \times d^{(\mathcal{L},j-1)}}$ and $b_{\mathcal{L}_j} \in \mathbb{R}^{d^{(\mathcal{L},j)}}$ with $d^{(\mathcal{L},0)} = d^{(C,L_1)}$.
- If $j = J$: $W_{\mathcal{L}_J} \in \mathbb{R}^{1 \times d^{(\mathcal{L},L_2)}}$ and $b_{\mathcal{L}_J} \in \mathbb{R}$.

The selected hyperparameter values for our experiments are detailed in Section 2.4.3.1.

2.2.2.2 Neural network optimization

The RNN-FNN network is optimized with the Mini-batch Stochastic Gradient Descent method (MSGD). This training procedure relies on updating iteratively all the trainable parameters of the network based on the recursive equation

$$\theta_{j+1} = \theta_j - \eta_j \nabla_\theta \hat{\mathcal{O}}(\theta_j), \tag{2.4}$$

where θ_j is the set of parameters obtained after iteration j, η_j is the learning rate (step size) which determines the magnitude of change in parameters on each time step,[3] ∇_θ is the gradient operator with respect to θ and $\hat{\mathcal{O}}$ is the Monte Carlo estimate of the objective function (2.3) computed on a mini-batch. Additional details are provided in Appendix 2.6.1.

[3]This parameter can be either deterministic or adaptive, i.e., it may be adjusted during the training period. For more details on this, please refer to Goodfellow et al. (2016).

2.3 Joint market dynamics

This section describes the market dynamics that represent the joint evolution of the S&P 500 index price and its associated Implied Volatility (IV) surface. Such model is used to construct the state space of the hedging problem.

2.3.1 Daily implied volatility surface representation

On any given day, the cross-section of option prices on the S&P 500 index is captured through the IV surface model introduced by François et al. (2022) which characterizes the entire surface parsimoniously with a linear combination of five factors. More precisely, the time-t IV of an option with time-to-maturity $\tau_t = \frac{T-t}{252}$ years and moneyness $M_t = \frac{1}{\sqrt{\tau_t}} \log \frac{S_t e^{(r_t - q_t)\tau_t}}{K}$, where K is the strike price, is

$$\sigma(M_t, \tau_t, \beta_t) = \sum_{i=1}^{5} \beta_{t,i} f_i(M_t, \tau_t) \tag{2.5}$$

where $\beta_t = (\beta_{t,1}, \beta_{t,2}, \beta_{t,3}, \beta_{t,4}, \beta_{t,5})$ stands for the IV factor coefficients at time t and the functions $\{f_i\}_{i=1}^5$ represent the long-term at-the-money (ATM) level, the time-to-maturity slope, the moneyness slope, the smile attenuation and the smirk, respectively (see Appendix 2.6.2.1 for their specification).

Following the same data processing and estimation procedure outlined in the aforementioned study, we extract the daily time series of the IV factor coefficients spanning from January 4, 1996, to December 31, 2020.[4]

2.3.2 Joint Implied Volatility and Return (JIVR)

The JIVR model proposed by François et al. (2023) leverages the IV representation (2.5) and provides explicit joint dynamics for the IV surface and the S&P 500 index price.

The first building block of the JIVR model represents the daily underlying asset excess

[4]The sample comes from OptionMetrics database, to which the conventional data exclusion filters are applied.

log-return, $R_{t+1} = \log\left(\frac{S_{t+1}}{S_t}\right) - (r_t - q_t)\Delta$. It integrates an NGARCH(1,1) process with Normal Inverse Gaussian (NIG) innovations to capture volatility clusters and reproduce fat tailed asymmetric returns, while borrowing information from the volatility surface to anchor the evolution of the conditional variance process $h_{t,R}$. The second building block includes a multivariate heteroskedastic autoregressive processes with non-Gaussian innovations for all implied volatility factor coefficients. The multivariate time series representation of the JIVR model is presented in detail in Appendix 2.6.2.2.

The full model is characterized by the current market conditions $(S_t, \{\beta_{t,i}\}_{i=1}^5, \beta_{t-1,2}, h_{t,R}, \{h_{t,i}\}_{i=1}^5)$, which are respectively the underlying S&P 500 index price, IV factor coefficients, and conditional variances for the S&P 500 return and for such coefficients.

Following François et al. (2023), the maximum likelihood estimation is applied to S&P 500 excess returns alongside the time series estimates of surface coefficients $\{\beta_{t,i}\}_{i=1}^5$. As a byproduct, we obtain daily estimates of conditional variance series $h_{t,R}$ and $\{h_{t,i}\}_{i=1}^5$ for the time period extending between January 4, 1996 and December 31, 2020.

2.4 Numerical study

In this section, simulation and backtesting experiments are conducted to evaluate the performance of the proposed hedging approach.

2.4.1 Stochastic market generator

2.4.1.1 Market simulator

The JIVR model is used in subsequent simulation experiments to generate paths of the variables pertaining to market dynamics $(S_t, \{\beta_{t,i}\}_{i=1}^5, h_{t,R}, \{h_{t,i}\}_{i=1}^5)$, which drive the hedging decisions. The initial conditions $(\{\beta_{0,i}\}_{i=1}^5, h_{0,R}, \{h_{0,i}\}_{i=1}^5)$ are randomly chosen among values prevailing in our data sample extending between January 4, 1996 and December 31, 2020. This constitutes a wide variety of states of the economy. Following the determination of

initial values, the simulation progresses through all time steps in two distinct phases. First, a sequence of NIG innovations $\{\epsilon_t\}_{t=1}^{T}$ is simulated using the Monte Carlo method considering the dependence structure of contemporaneous innovations. Secondly, equations of Section 2.6.2.2 are used to obtain values for $(\{R_t, \beta_{t,i}\}_{i=1}^{5}, h_{t,R}, \{h_{t,i}\}_{i=1}^{5})$ for $t = 1, \ldots, T$ based on the simulated innovations.

2.4.1.2 Market parameters for numerical experiments

The initial underlying asset value is normalized to $S_0 = 100$ for simplicity. The options being hedged are assumed to be European call options ($\Psi(S_T) = \max(S_T - K, 0)$) with maturities $T \in \{21, 63, 126\}$ days for short-term, medium-term and long-term maturities, and strikes $K \in \{90, 100, 110\}$ for in-the-money (ITM), at-the-money (ATM) and out-of-the-money (OTM) options, respectively.[5] Various levels of proportional transaction cost are considered, namely $\kappa \in \{0\%, 0.05\%, 0.5\%, 1\%\}$. The initial value of the hedging strategy, V_0^{δ}, is set to the price of the option being hedged, which is provided by the prevailing implied volatility surface. Other parameters of the model are specified in Section 2.6.2.2.

2.4.2 Benchmarks

We benchmark RL hedging strategies against the performance of three standard models: (i) the practitioner's Black-Scholes delta hedging (BS) which applies the current implied volatility into the Black-Scholes formula to obtain the hedged option's delta; (ii) the Leland (1985) delta (BSL), which modifies the BS delta to reflect proportional transaction costs; and (iii) the Smile-implied delta (SI) introduced by Bates (2005) and whose closed-form expression for the IV model (2.5) is provided by François et al. (2022). The explicit formulas for these three benchmarks are detailed in Appendix 2.6.3.

[5] Unreported numerical results of European put options exhibit a similar pattern due to the Put-Call parity formula.

2.4.3 Neural network settings

2.4.3.1 Neural network architecture

We employ the RNN-FNN architecture from Section 2.2.2.1 with two LSTM cells ($L_1 = 2$) of width 56 ($d_i = 56$ for $i = 1, 2$) and two FFNN-hidden layers ($L_2 = 2$) of width 56 with ReLU activation function (i.e., $g_{\mathcal{L}_i}(x) = \max(0, x)$ for $i = 1, 2$). In the context of the output FFNN layer $\mathcal{L}_J$, which maps the output of hidden layers $Z_t^{(\mathcal{L}, L_2)}$ into the position in the underlying asset δ_{t+1}, the standard deep hedging framework typically employs a linear activation function. However, this activation function tends to induce RL agents to adopt doubling strategies, increasing the position in the underlying asset several orders of magnitude over the current position on each period after any loss until the cumulative loss amount is fully recovered. Such strategies are definitely undesirable from the perspective of sound risk management. Therefore, we opt to introduce a leverage constraint through the output layer activation function.

This leverage constraint is a dynamic upper bound on the activation function, denoted as $g_{\mathcal{L}_J}(Z, t) = \min(Z, B_t(Z))$, with $Z = T_{\mathcal{L}_J}(Z_t^{(\mathcal{L}, L_2)})$ representing the typical deep hedging underlying asset position, and $B_t(Z)$ the highest position in the index that can be held in the portfolio. Such an upper bound avoids excessive leverage and limits the borrowing capacity, i.e., the cash held satisfies $\phi_{t+1}(X_t) \geq -B$ for all X_t and t, and $B > 0$.[6] The latter, in conjunction with the self-financing constraint (2.1), establishes the dynamic upper bound as

$$B_t(Z) = \begin{cases} \frac{V_0 + B}{S_0} & \text{if } t = 0 \\ \frac{V_t + B + \kappa S_t \delta_t}{S_t(1+\kappa)} & \text{if } t > 0 \text{ and } Z \geq \delta_t \\ \frac{V_t + B - \kappa S_t \delta_t}{S_t(1-\kappa)} & \text{if } t > 0 \text{ and } Z < \delta_t. \end{cases} \tag{2.6}$$

Appendix 2.6.4.1 provides numerical experiments that illustrate the necessity of including a leverage constraint in the network architecture to alleviate the aforementioned issues.

[6] This type of leverage condition has been previously employed in the literature. For instance, the seminal paper of Harrison and Pliska (1981) assumes a restricted borrowing capacity to maintain a positive wealth throughout the entire hedging period.

Our numerical experiments demonstrate the superiority of the RNN-FNN architecture over standalone LSTM and FFNN networks, considering a leverage constraint of $B = 100$ for all agents. These results are presented in Appendix 2.6.4.2. Additionally, agents are trained as described in Section 2.2.2.2 on a training set of 400,000 independent simulated paths with mini-batch size of 1,000 and a learning rate of 0.0005 that is progressively adapted by the ADAM optimization algorithm.[7] In addition, we include a regularization method called dropout with parameter $p = 0.5$ to reduce the likelihood of overfitting and enhance the model performance on unseen data.[8] The training procedure is implemented in Python, using Tensorflow and considering the Glorot and Bengio (2010) random initialization of the initial parameters of the neural network. Numerical results are obtained from a test set of 100,000 independent paths.

2.4.3.2 State space selection

The nature of the hedging problem, combined with the JIVR model, establishes a decision framework where the optimal decision is entirely defined by the state variables at time t, denoted as $X_t = (V_t^\delta, \delta_t, \tau_t, S_t, \{\beta_{t,i}\}_{i=1}^5, h_{t,R}, \{h_{t,i}\}_{i=1}^5)$.

However, implementing numerical methods within the deep policy gradient approach across the entire state space may lead to overfitting. In fact, other studies utilizing the RL framework have reported optimal results with a reduced state space. For instance, Buehler et al. (2019) and Cao et al. (2023) employ the deep hedging algorithm without including the portfolio value V_t^δ in the state space configuration. Similarly, Carbonneau (2021) employs the deep hedging framework in a frictionless market without including δ_t in the state space.

We explore three different state space configurations. The first configuration aims to replicate

[7] ADAM is a dynamic learning rate algorithm engineered to accelerate training speeds in deep neural networks and achieve rapid convergence, details can be seen in Goodfellow et al. (2016).

[8] Full details of the dropout regularization method can be seen in Goodfellow et al. (2016). The selection of $p = 0.5$ stems from our numerical experiments detailed in Appendix 2.6.4.3, indicating that this value consistently outperforms others across all penalty functions.

the full state space, denoted by

$$(V_t^{\delta}, \delta_t, \tau_t, S_t, \{\beta_{t,i}\}_{i=1}^5, h_{t,R}, \{h_{t,i}\}_{i=1}^5), \tag{2.7}$$

typically considered in a dynamic optimization task with the portfolio value as a state variable. The second state space does not consider the portfolio value and also omits the variance of the IV coefficients $\{h_{t,i}\}_{i=1}^5$ under the intuition that these coefficients have a second-order effect, with the IV surface coefficients $\{\beta_{t,i}\}_{i=1}^5$ capturing most of the variability. Hence, the reduced state space is defined as

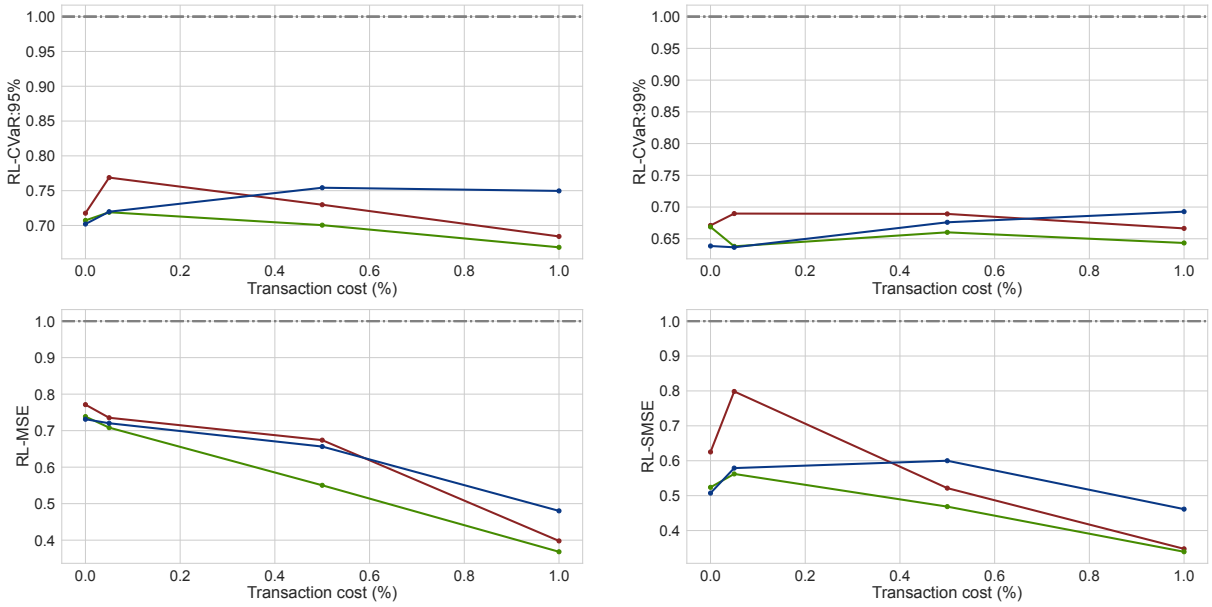
$$(\delta_t, \tau_t, S_t, \{\beta_{t,i}\}_{i=1}^5, h_{t,R}).$$

The current position δ_t is required for the RL agent to learn about the transaction cost associated with the next rebalancing. This state space component is no longer useful when $\kappa = 0$. For that reason, we remove this component in absence of transaction cost. In such case, the reduced state space is denoted by

$$(\tau_t, S_t, \{\beta_{t,i}\}_{i=1}^5, h_{t,R}).$$

Our numerical experiments consider the three aforementioned state spaces to hedge a European ATM call option with maturity $N = 63$ days, while taking into account four different transaction cost rates $\kappa \in \{0\%, 0.05\%, 0.5\%, 1\%\}$. The impact of the option moneyness and maturity on hedging performance are studied in later subsections. Additionally, our numerical experiments involve four RL agents: RL-CVaR$_{95\%}$, RL-CVaR$_{99\%}$, RL-MSE, and RL-SMSE. These agents aim to minimize different penalty functions.

Figure 2.1: Optimal penalty function value for a short position in an ATM call option with maturity of 63 days under various state spaces and transaction cost levels.



Results are computed using 100,000 out-of-sample paths according to the conditions outlined in Section 2.4.3.1. Each panel illustrates the optimal penalty function value of an RL agent considering four transaction cost levels. These values are normalized by the estimated values of each penalty function obtained with BSL delta hedging for each transaction cost level. Agents are trained under the specified penalty functions considering three state spaces: full state space $(V_t^\delta, \delta_t, \tau_t, S_t, \{\beta_{t,i}\}_{i=1}^5, h_{t,R}, \{h_{t,i}\}_{i=1}^5)$ (red curve), and the two reduced state spaces $(\delta_t, \tau_t, S_t, \{\beta_{t,i}\}_{i=1}^5, h_{t,R}, \{h_{t,i}\}_{i=1}^5)$ (green curve) and $(\tau_t, S_t, \{\beta_{t,i}\}_{i=1}^5, h_{t,R}, \{h_{t,i}\}_{i=1}^5)$ (blue curve).

Figure 2.1 illustrates the estimated value of the four penalty functions in proportion to that obtained with the BSL delta, $\rho(\xi_T^\delta)/\rho(\xi_T^{BSL})$, across different transaction cost rates for all agents. These numerical results reveal that RL agents outperform BSL delta hedging for all considered state space configurations, as the relative values of the penalty functions are substantially lower than the reference line value of 1.

Moreover, these results confirm that including the portfolio value in the full state space is unnecessary for our approach, as metrics under the reduced state space (green curve) show lower values than those achieved by agents using the full state space (red curve). This is consistent with the work of Buehler et al. (2019), Cao et al. (2020), Buehler et al. (2022), and

Cao et al. (2023), where RL techniques are applied in the hedging context without considering the portfolio value as a state variable.

In the absence of transaction costs, Figure 2.1 confirm that the inclusion of δ_t in the reduced state space is unnecessary, as the incremental performance with and without it is negligible. Conversely, in the presence of transaction costs, agents trained under the reduced state space $(\delta_t, \tau_t, S_t, \{\beta_{t,i}\}_{i=1}^5, h_{t,R})$ demonstrate superior performance (green curve), and the outperformance with respect to benchmarks becomes more pronounced as the transaction cost rate increases. This observation underscores the importance of considering previous positions in the underlying when optimizing rebalancing actions, as the former is indicative of transaction costs to be paid for the various possible actions.

The superiority of RL agents trained on the reduced state space can be explained by the bias-variance dilemma, where the informational content provided by some of the variables (reduction in bias) is insufficient to compensate for additional complexity and instability (variance) they cause during training. This is seemingly why removing the IV parameters' variances $\{h_{t,i}\}_{i=1}^5$ from the state space increases the performance in this experiment. Our numerical results indicate that the network architecture considered for the agents acting in our proposed high-dimensional environment does not require the full state space. In fact, the performance of the agents deteriorates for some transaction cost levels when the volatilities of the IV coefficients are included, as shown in Figure 2.1, and the time required to converge to optimal solutions during training increases by an average of 190% compared to RL agents under the reduced state space across all experiments.

In all subsequent experiments, we consider the reduced state space $(\delta_t, \tau_t, S_t, \{\beta_{t,i}\}_{i=1}^5, h_{t,R})$ in the presence of transaction costs and the state space $(\tau_t, S_t, \{\beta_{t,i}\}_{i=1}^5, h_{t,R})$ in the absence of transaction costs to enhance parsimony.

2.4.4 Benchmarking of hedging strategies

In this section, we compare the performance of RL agents with classic hedging approaches. Aggregated results are presented in Section 2.4.4.1, while the segmentation among various states of the economy is shown in Section 2.4.4.2. The following subsections break down the performance of the RL agents with respect to moneyness and maturity of the hedged option, or to transaction cost levels, demonstrating the robustness of the RL approach. Finally, Section 2.4.6 outlines a comparison of performance over historical paths spanning from January 5, 1996, to December 31, 2020.

2.4.4.1 Benchmarking over randomized economic conditions

We begin by comparing the hedging performance of the benchmarks and RL agents trained under the four penalty functions: $\text{CVaR}_{95\%}$, $\text{CVaR}_{99\%}$, MSE, and SMSE. This comparison is performed in terms of estimated values of all penalty functions and the average Profit and Loss, Avg P&L $= \mathbb{E}[-\xi_T^\delta]$, across all paths in a test set. Additionally, we employ the CVaR deviation measure, defined as $\text{CVaR}_\alpha(\xi_T^\delta - E[\xi_T^\delta])$, as a deviation metric for the hedging error in the test set. In such test set, initial economic conditions (the initial value of state variables in the path) are sampled randomly among historical values from our sample. It therefore reflects aggregate performance across various economic conditions. Our analysis focuses on hedging a European ATM call option with a maturity of $N = 63$, under the assumption of no transaction costs, which is $\kappa = 0$.

Table 2.1: Aggregated hedging metrics for a short position in an ATM call option with maturity of 63 days.

	Benchmark		Reinforcement Learning			
Metric	BS	SI	$\text{CVaR}_{95\%}$	$\text{CVaR}_{99\%}$	MSE	SMSE
Avg P&L	0.356	**0.498**	0.380	0.321	0.285	0.374
$\text{CVaR}_\alpha(\xi_T^\delta - E[\xi_T^\delta])$	2.159	3.005	**1.646**	1.720	1.766	1.652
$\text{CVaR}_{95\%}$	1.803	2.507	**1.266**	1.399	1.481	1.278
$\text{CVaR}_{99\%}$	3.442	4.351	2.312	**2.198**	2.625	2.245
MSE	0.898	1.818	1.243	1.362	**0.657**	1.018
SMSE	0.298	0.564	0.184	0.214	0.183	**0.151**
Avg P&L/$\text{CVaR}_\alpha(\xi_T^\delta - E[\xi_T^\delta])$	0.165	0.166	**0.231**	0.187	0.161	0.227

Results are computed using 100,000 out-of-sample paths in the absence of transaction costs ($\kappa = 0\%$). Agents are trained under the reduced state space $(\tau_t, S_t, \{\beta_{t,i}\}_{i=1}^5, h_{t,R})$ according to the conditions outlined in Section 2.4.3.1. BS stands for Black-Scholes delta hedging, whereas SI is the smile-implied delta hedging.

Table 2.1 presents hedging performance metrics attained by both benchmarks and RL agents. Every RL agent achieves the lowest value for the corresponding metric which they used as objective function during training, which is expected. Furthermore, the numerical results indicate that RL agents provide hedging strategies that are much less risky than benchmarks, as evidenced by metrics such as $\text{CVaR}_{95\%}$, $\text{CVaR}_{99\%}$, and SMSE. In particular, when computing the variation rate between the estimated value of each of these metrics obtained by RL agents relative to the value obtained by BS delta, we observe an average reduction rate across RL agents of 24%, 31%, and 38%, respectively. Similarly, when comparing with SI delta strategies, we observe average reductions across RL agents of 45%, 46%, and 67%, respectively. Moreover, in terms of the MSE metric, the MSE agent demonstrates significant superiority over benchmarks, reducing 26% and 63% compared to BS delta and SI delta strategies, respectively.

Regarding average Profit and Loss (P&L), SI delta yields higher profitability; however, it also entails increased risk. In contrast, RL agents trained using $\text{CVaR}_{95\%}$, $\text{CVaR}_{99\%}$, and SMSE as penalty functions obtain lower values compared to SI delta strategies, but they reach a more favorable trade-off between profitability and risk management as shown by the ratio Avg P&L/$\text{CVaR}_\alpha(\xi_T^\delta - E[\xi_T^\delta])$, which yields greater values for these agents. Additionally, RL agents achieve lower risk than both the BS and SI delta hedging strategies, which highlights enhanced risk management.

2.4.4.2 Impact of the state of the economy on performance

In this section, we analyze the performance metrics of hedging strategies across clusters of paths representing different economic states, as detailed in Table 2.3. A path is assigned to a cluster if its state variable initial values are drawn from the subset of dates corresponding to that cluster. This approach helps isolate the impact of economic conditions on performance and assess the robustness of RL hedging strategies. Again, we hedge an ATM European call option with a maturity of $N = 63$ days under the assumption of no transaction cost.

Table 2.3: Clusters of dates representing different time periods

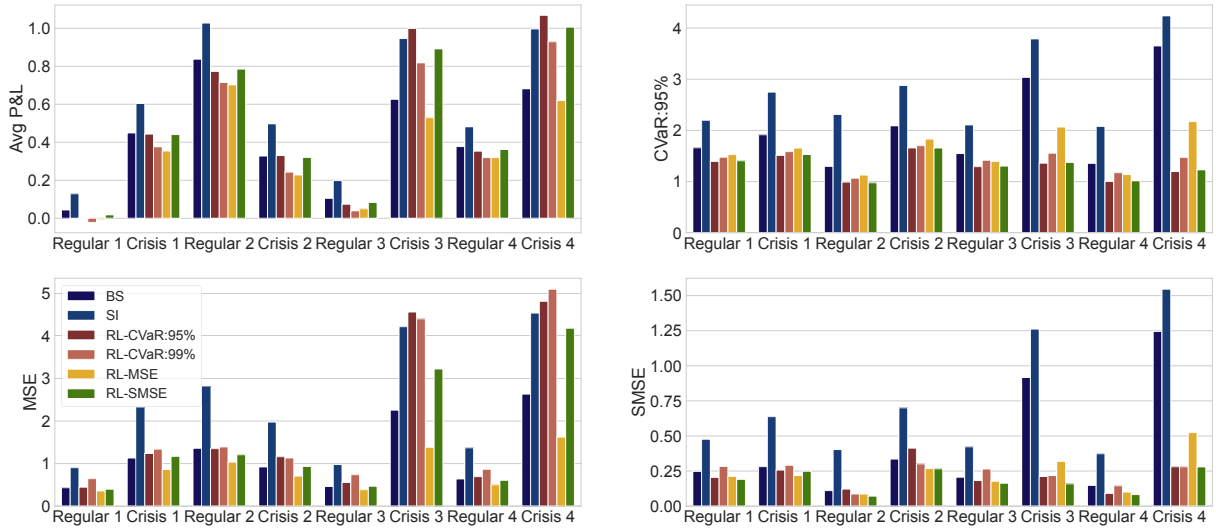
Period	Time frames	Avg option price
Regular 1	05/01/1996 - 28/02/1997	\$3.849
Crisis 1 (Asian financial crisis)	03/03/1997 - 31/12/1998	\$5.351
Regular 2	04/01/1999 - 31/12/1999	\$5.301
Crisis 2 (Dot-com bubble crisis)	03/01/2000 - 31/12/2002	\$5.758
Regular 3	02/01/2003 - 31/12/2007	\$3.877
Crisis 3 (Global financial crisis)	02/01/2008 - 31/12/2009	\$8.009
Regular 4	04/01/2010 - 31/01/2020	\$3.728
Crisis 4 (Covid-19 pandemic crisis)	03/02/2020 - 31/12/2020	\$6.486

These periods aim to approximate different states of the economy to highlight the performance of our approach within time windows capturing the financial fluctuations characteristic of each economic crisis. The column "Avg option price" displays the average option price of an ATM call option with maturity of 63 days per cluster.

Our numerical results, illustrated in Figure 2.2, indicate that both benchmarks and RL agents are sensitive to the environment, showcasing better performance during regular periods

compared to financial crises in terms of risk, as shown by the two right plots showing the $\text{CVaR}_{95\%}$ and SMSE metrics. Additionally, both approaches tend to offer more profitable strategies during crisis periods and periods of high volatility (for instance, even though Regular period 2 is not labeled as a crisis periods, it is characterized by economic recovery with significant market fluctuations). This trend can be explained by the fact that higher volatility during these periods leads to a higher risk premium, increasing the initial portfolio value and thus the final P&L, as highlighted by the Avg P&L metric (see the top left panel of Figure 2.2). Regarding the MSE metric, the results align with the Avg P&L metric, with both approaches showing superior performance during regular periods (except for Regular period 2 which exhibited high volatility), as higher profits are detrimental for the MSE metric penalizing upside risk.

Figure 2.2: Hedging metrics for a short position in an ATM call option with maturity of 63 days under different states of the economy.



Results are computed using 100,000 out-of-sample paths in the absence of transaction costs ($\kappa = 0\%$). Agents are trained under the reduced state space $(\tau_t, S_t, \{\beta_{t,i}\}_{i=1}^{5}, h_{t,R})$ according to the conditions outlined in Section 2.4.3.1. The results are organized chronologically across the periods outlined in Table 2.3.

The results that are segregated by time period align with the aggregated findings. Indeed, for any of the $\text{CVaR}_{95\%}$, MSE, and SMSE risk metrics, the RL agent which was trained with

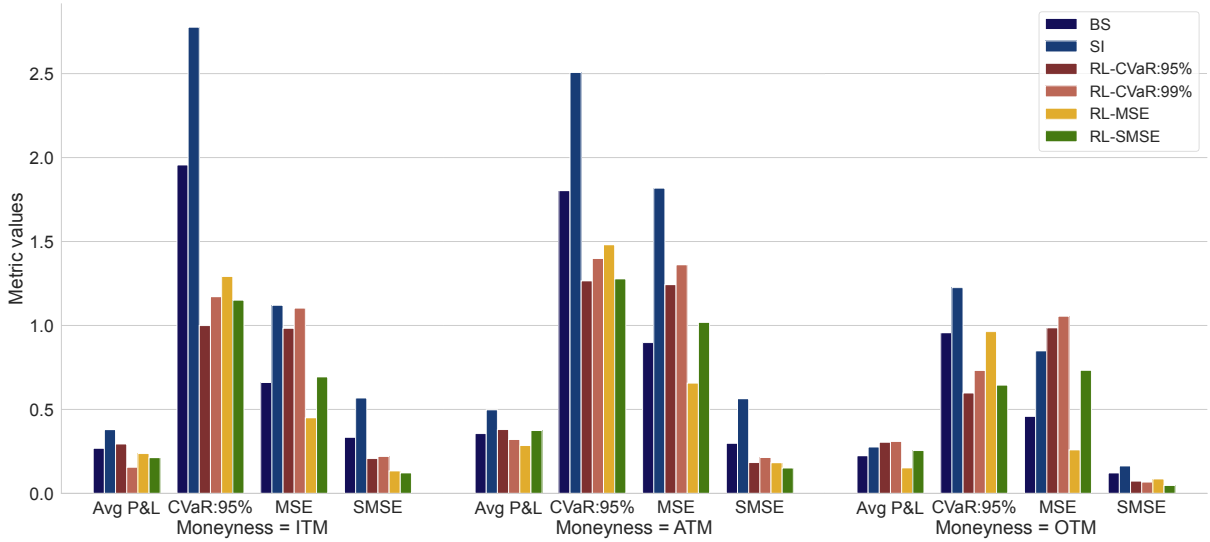
such risk metric as its objective function consistently exhibits lower values than benchmarks across all periods. These differences are particularly notable during crisis periods, where BS delta and SI delta tend to be riskier, while RL agents demonstrate greater stability. The higher values in the MSE statistic of agents trained under CVaR and SMSE can be attributed to the fact that these objective functions only penalize hedging losses, allowing agents to seek positive returns on average. This is consistent with the Avg P&L panel, where CVaR and SMSE agents tend to display more profitable strategies than the MSE agent.

These results demonstrate the robustness of the RL approach across various environments, exhibiting greater stability in scenarios with extreme behavior and improving hedging performance in terms of the penalty functions considered in the hedging problem. Additionally, they support our initial findings, indicating that RL agents do not significantly sacrifice profitability even during extreme market conditions, a phenomenon that could be attributed to higher initial option prices during these periods, as shown in Table 2.3.

2.4.4.3 Impact of moneyness level on performance

We investigate the robustness of our approach with respect to the moneyness level of the option being hedged by including OTM and ITM call options with a maturity of $N = 63$ days in our analysis. Our findings, depicted in Figure 2.3, demonstrate that RL consistently outperforms benchmarks with respect to the objective function considered during the training process, regardless of the option moneyness.

Figure 2.3: Hedging metrics for a short position in OTM, ATM and ITM call options with a maturity of 63 days.



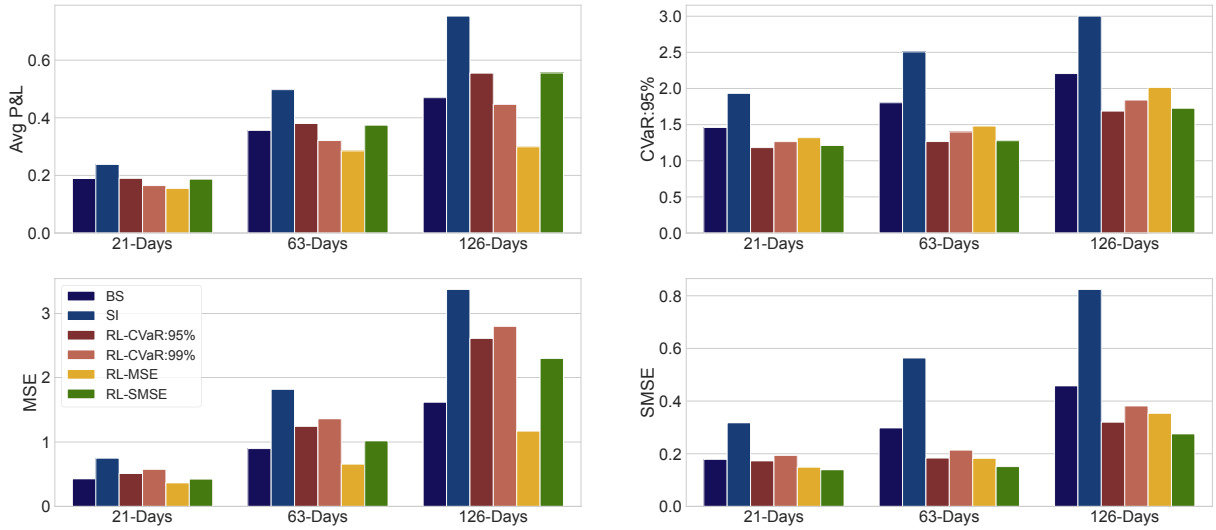
Results are computed using 100,000 out-of-sample paths in the absence of transaction costs ($\kappa = 0\%$). Agents are trained under the reduced state space $(\tau_t, S_t, \{\beta_{t,i}\}_{i=1}^{5}, h_{t,R})$ according to the conditions outlined in Section 2.4.3.1. The average option price stands at \$0.59 for OTM options, \$3.89 for ATM options, and \$11.37 for ITM options.

In line with our previous experiments, we observe that SI delta tends to offer more profitable strategies compared to benchmarks, while BS delta demonstrates a similar level of profitability for ATM and ITM options. However, this trend does not hold for OTM options, where RL agents not only exhibit better risk management but also yield more profitable strategies on average (refer to CVaR and SMSE agents for OTM options in Figure 2.3). The discrepancy can be attributed to the fact that RL agents trained under CVaR and SMSE do not track the option value nor penalize gains at maturity, allowing agents to profit from OTM paths at maturity. In contrast, BS and BSL are option tracking methodologies that aim to replicate the option value at maturity, thereby reducing potential gains for OTM options at maturity. The latter is consistent with the RL agent trained under the MSE, which achieves the minimum MSE value for OTM options and the lowest Avg P&L, as it penalizes both gains and losses.

2.4.4.4 *Impact of option maturity on performance*

As a third test to assess the robustness of our approach, we compare the performance of RL agents against benchmarks when hedging ATM call options with different maturities: 21, 63, and 126 days, in absence of transaction costs. Our results, illustrated in Figure 2.4, show that the average profitability of hedging strategies increases with option maturity (see Avg P&L in the top left panel). However, the risk also increases with maturity for all hedging strategies (refer to $\text{CVaR}_{95\%}$ and SMSE depicted in the panels positioned at the top and bottom right corners), as expected.

Figure 2.4: Hedging metrics for a short position in ATM call options with maturities of 21, 63 and 126 days.



Results are computed using 100,000 out-of-sample paths in the absence of transaction costs ($\kappa = 0\%$). Agents are trained under the reduced state space $(\tau_t, S_t, \{\beta_{t,i}\}_{i=1}^{5}, h_{t,R})$ according to the conditions outlined in Section 2.4.3.1. Average prices are \$2.15, \$3.89 and \$5.65 for options with maturities of 21 days, 63 days and 126 days, respectively.

These results also demonstrate consistent behavior across all maturities for all hedging metrics, displaying a uniform distribution among the different strategies across all maturity levels. In line with our previous experiments, RL agents consistently outperform benchmarks with respect to asymmetric penalty functions for all maturities, regardless of the penalty function used during training. Furthermore, in the case of the MSE metric (bottom left panel in

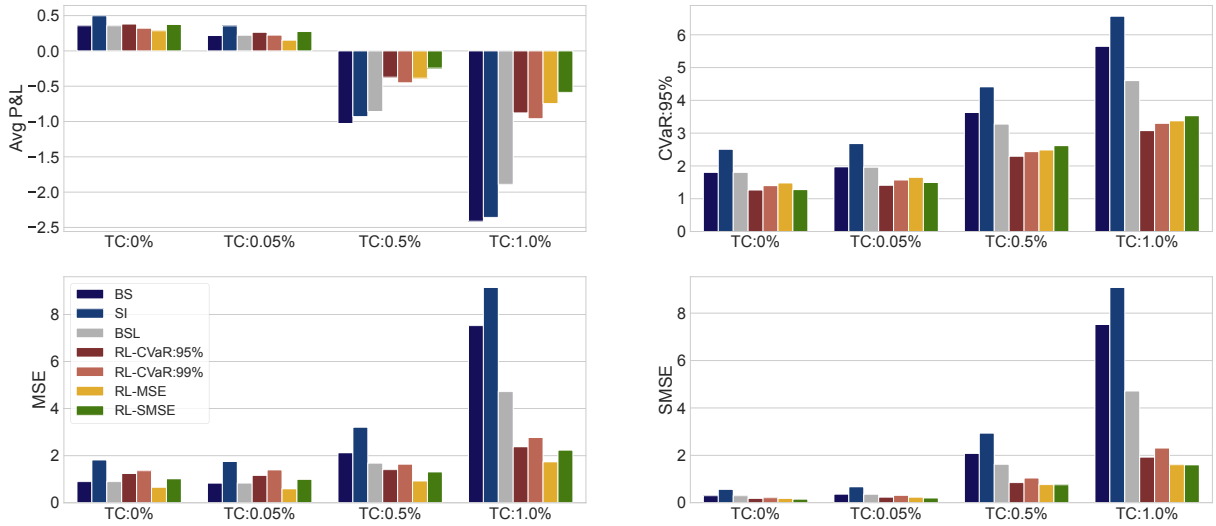
Figure 2.4), the RL agent trained under that penalty function displays the lowest value across all maturities.

2.4.4.5 Impact of transaction costs on performance

We now investigate scenarios where hedgers encounter transaction costs, which is meant to reflect more realistic hedging scenarios. Specifically, we analyze the hedging effectiveness of RL agents in comparison to benchmarks, including the BSL delta hedging. Our examination focuses on hedging an ATM European call option with a maturity of $N = 63$ days across varying levels of transaction costs assumed to be 0.05%, 0.5% and 1%.

Figure 2.5 illustrates the performance of benchmarks and RL agents, with the four panels depicting the Avg P&L, $\text{CVaR}_{95\%}$, MSE, and SMSE metrics. In general, as expected, the performance of each strategy tends to deteriorate as transaction costs increase. However, performance drops are more pronounced for benchmarks than for RL agents. As anticipated, RL agents consistently outperform benchmarks in terms of downside risk (observe $\text{CVaR}_{95\%}$ and SMSE metrics depicted in the panels positioned at the top and bottom right corners of Figure 2.5) across all transaction cost levels, regardless of the penalty function used during training. Furthermore, while BSL delta hedging adjusts the strategies in terms of the transaction cost level and performs better than the other benchmarks, RL strategies display superior metrics, highlighting their better capacity to adapt to different transaction cost levels.

Figure 2.5: Hedging metrics for a short position in an ATM call option with a maturity of 63 days under different transaction cost levels.



Results are computed using 100,000 out-of-sample paths under the reduced state space $(\tau_t, S_t, \{\beta_{t,i}\}_{i=1}^5, h_{t,R})$ without transaction costs, and under $(\delta_t, \tau_t, S_t, \{\beta_{t,i}\}_{i=1}^5, h_{t,R})$ in the presence of transaction cost. Agents are trained according to the conditions outlined in Section 2.4.3.1. The average option price is \$3.89 with a standard deviation of \$1.29. The acronym TC stands for transaction cost. BS stands for Black-Scholes delta hedging, BSL for Leland delta hedging and SI for smile-implied delta hedging.

In contrast to our previous results where SI delta provides more profitability, the results regarding the Avg P&L metric (see top left panel in Figure 2.5) indicate that this profitability is influenced by the inclusion of transaction costs, making the SI delta hedging less profitable strategy as transaction costs increase, whereas RL agents tend to exhibit much lower average losses due to their adaptability. In terms of MSE, benchmarks exhibit a more sensitive behavior than the RL agents regarding the increase in transaction costs, especially when the transaction costs are set at 0.5% and 1% (see the bottom-left panel of Figure 2.5), irrespective of the penalty function considered during training.

Table 2.4 displays descriptive statistics about hedging costs, defined as the sum of discounted transaction costs over a path,

$$\sum_{t=0}^{T-1} e^{-r\Delta t} \kappa S_t \mid \delta_{t+1} - \delta_t \mid, \tag{2.8}$$

across different transaction cost levels. In the cases of BS delta and SI delta strategies, which overlook transaction costs, we observe a higher cost of hedging due to significant variations in the position of the underlying asset during rebalancing across different time steps, thereby leading to lower profitability as transaction costs increase. This trend is evident in the case of SI delta which incurs the highest average hedging cost, followed by BS delta and BSL delta, when the transaction cost rate is set to 0.5% and 1%. These results consistently align with the performance differences observed in Figure 2.5, where benchmarks are associated with the highest losses.

Table 2.4: Hedging costs when hedging a short ATM call option position with a maturity of $N = 63$ days under various transaction cost levels.

		Benchmark - Delta			Reinforcement Learning			
κ	Metric	BS	SI	Leland	$\text{CVaR}_{95\%}$	$\text{CVaR}_{99\%}$	MSE	SMSE
0.05%	Mean	0.138	0.142	0.136	0.112	**0.108**	0.150	0.113
	Std	0.041	0.048	0.040	0.024	**0.021**	0.042	0.025
0.5%	Mean	1.376	1.420	1.223	0.751	0.788	0.767	**0.674**
	Std	0.412	0.482	0.317	0.137	0.137	0.205	**0.127**
1%	Mean	2.753	2.839	2.260	1.245	1.259	1.139	**1.031**
	Std	0.824	0.964	0.535	0.203	0.212	0.254	**0.188**

The average cost of hedging is computed by evaluating the hedging cost across 100,000 out-of-sample independent paths. Transaction cost levels are assumed to be proportional to the trade size. The lowest values across all hedging strategies are highlighted in bold. The average option price is \$3.89 with a standard deviation of \$1.29. Std stands for standard deviation.

Furthermore, although none of the agents directly aim to minimize transaction costs, we observe that RL agents with asymmetric penalty functions tend to offer hedging strategies with lower costs, as shown by the highlighted values in Table 2.4. This underscores the robustness of RL agents in addressing transaction costs, as they minimize the penalty function and thereby indirectly reduce the cost of hedging. Conversely, the RL agent trained under

the MSE criterion tends to exhibit higher average costs compared to the other agents. This can be attributed to the MSE agent penalizing both negative and positive returns, creating additional turnover in some situations to avoid large profits.

2.4.5 Global importance of IV surface features

We now investigate to what extent the risk factors characterizing the IV surface described in Equation (2.5) contribute to total performance of our model. We employ the Shapley Additive Global Importance (SAGE) methodology, introduced by Covert et al. (2020), to evaluate their impact. Specifically, let $\rho\left(\xi_T^{\tilde{\delta}_\theta(\mathcal{X})}\right)$ be the risk measure when the model is trained with the state space $\mathcal{X}$. The amount of risk reduction achieved by adding the state variables $(\{\beta_{t,i}\}_{i=1}^5, h_{t,R})$ to a baseline model $\tilde{\delta}_\theta$ with state variables (τ_t, S_t) is

$$\rho\left(\xi_T^{\tilde{\delta}_\theta(\tau_t,S_t)}\right) - \rho\left(\xi_T^{\tilde{\delta}_\theta(\tau_t,S_t,\{\beta_{t,i}\}_{i=1}^5,h_{t,R})}\right) = \sum_{j\in\{\{\beta_{t,i}\}_{i=1}^5,h_{t,R}\}} \mathcal{C}_j$$

where $\mathcal{C}_j$, the contribution of feature j to the total risk reduction, is

$$\mathcal{C}_j = \sum_{\mathcal{X}\subseteq\{\tau_t,S_t,\{\beta_{t,i}\}_{i=1}^5,h_{t,R}\}\setminus\{j\}} \frac{|\mathcal{X}|!(\bigtriangledown - |\mathcal{X}|)!}{6!}\left[\rho\left(\xi_T^{\tilde{\delta}_\theta(\mathcal{X})}\right) - \rho\left(\xi_T^{\tilde{\delta}_\theta(\mathcal{X},|)}\right)\right].$$

Exact contributions $\mathcal{C}_j$ cannot be negative, although their estimates sometimes are. Negative values may occur for the contributions we present that are evaluated out-of-sample.

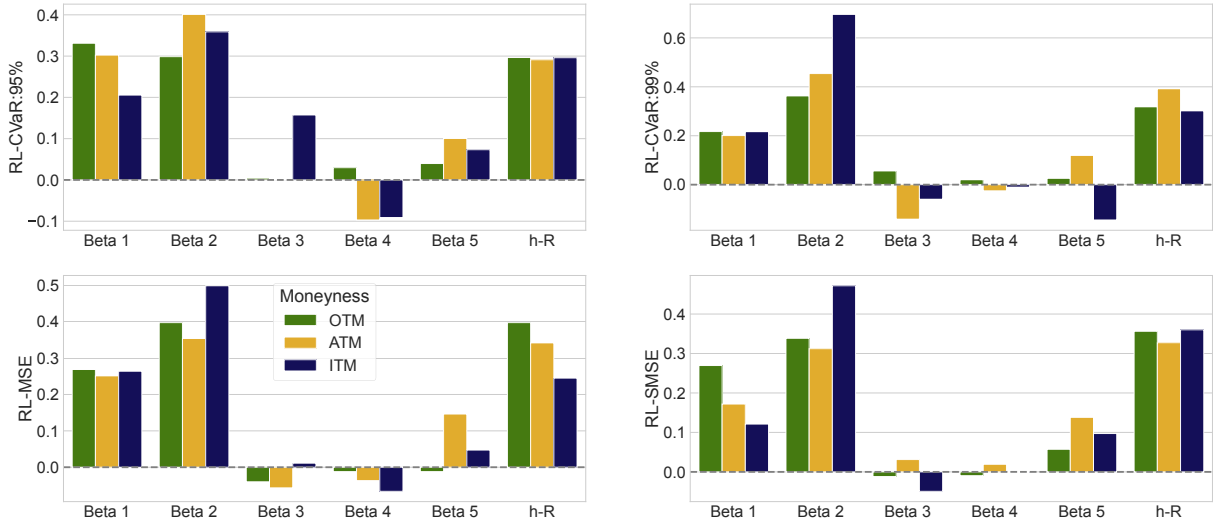
Our numerical experiments involve training RL agents using four penalty functions: $\text{CVaR}_{95\%}$, $\text{CVaR}_{99\%}$, MSE, and SMSE. We analyze the global importance of these state variables across different moneyness and maturities to comprehend their impact on the performance of the RL agents. The global importance of state variables is normalized by the risk reduction achieved by the respective RL agent to present contributions in the same order of magnitude: the relative global importance is

$$\frac{\mathcal{C}_j}{\rho\left(\xi_T^{\tilde{\delta}_\theta(\tau_t,S_t)}\right)-\rho\left(\xi_T^{\tilde{\delta}_\theta(\tau_t,S_t,\{\beta_{t,i}\}_{i=1}^5,h_{t,R})}\right)}, \quad \text{for } j\in\{h_{t,R}\}\cup\{\beta_{t,i}\}_{i=1}^5. \tag{2.9}$$

The relative global importance of the IV characteristic $\beta_{t,i}$ and the return conditional variance $h_{t,R}$ to the risk reduction depends on the moneyness and the time-to-maturity of the option to be hedged, as well as on the choice of risk measure.

Figure 2.6 studies the case of 63-day-to-maturity call options. It illustrates the relative contribution of each state variable across moneyness levels. Overall, the conditional variance of the underlying asset returns, the long-term ATM level β_1 and the time-to-maturity slope β_2 of the IV surface play a major role, no matter what risk measure or moneyness is considered. This underscores that RL agents utilize both the historical variance process and market expectations of future volatility to adjust positions in the underlying asset. The moneyness slope, the smile attenuation and the smirk have a second order effect.

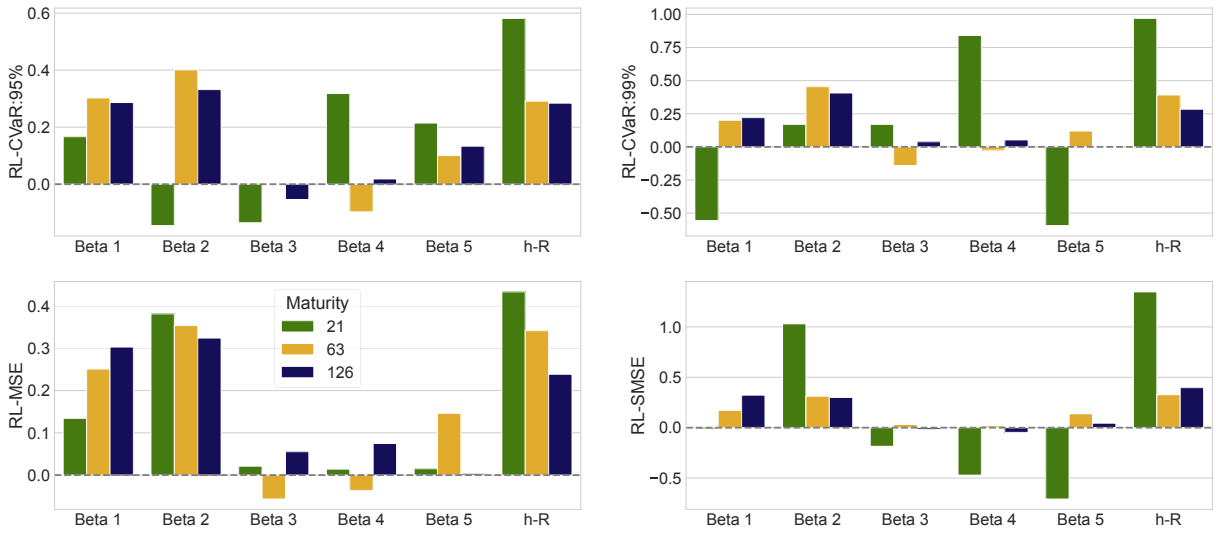
Figure 2.6: Normalized global importance of features when hedging a European call options with a maturity $N = 63$ days, across various moneyness levels.



Results are computed using 100,000 out-of-sample paths in the absence of transaction costs ($\kappa = 0\%$). Each panel illustrates the Shapley values for all state variables $(\{\beta_{t,i}\}_{i=1}^5, h_{t,R})$ and different moneyness: OTM, ATM, and ITM. These results are shown for the four RL agents: RL-CVaR$_{95\%}$, RL-CVaR$_{99\%}$, RL-MSE, and RL-SMSE. The Shapley values are normalized by the risk reduction achieved by the respective agent according to Equation (2.9).

Figure 2.7 presents contributions to hedging risk reduction for three option maturities, namely 21, 63 or 126 business days. It underscores the persistent significance of the conditional variance process $\{h_{t,R}\}$ across the various option maturities. Its contribution is even more important for short-term maturities. This reflects the fact that $h_{t,R}$ has a direct impact on immediate market shocks. Both β_4 and β_5 are additional contributors mostly for short-term options when CVaR risk measures are considered. Intuitively, high values of the smile attenuation (β_4) and the smirk (β_5) indicate a steep slope and a strong smirk for the short term smile which, in turn, induces a high exposure to tail risk for short term options.

Figure 2.7: Global importance of features when hedging a European ATM call options, across various maturities.



Results are computed using 100,000 out-of-sample paths in the absence of transaction costs ($\kappa = 0\%$). Each panel illustrates the Shapley values for all features under different maturities: 21, 63, and 126 business days. These results are shown for the four RL agents: $\text{CVaR}_{95\%}$, $\text{CVaR}_{99\%}$, MSE, and SMSE. The Shapley values are normalized by the risk reduction achieved by the respective agent according to Equation (2.9).

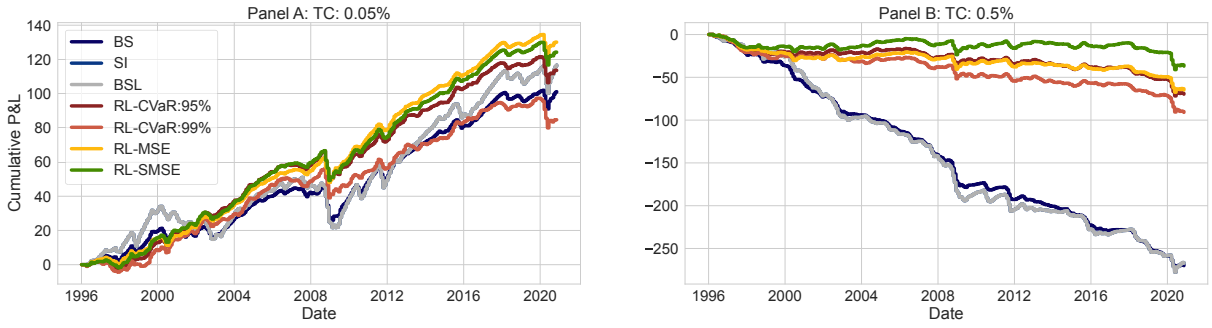
2.4.6 Backtesting

In this section, we benchmark our approach using a historical path of the JIVR model spanning from January 5, 1996, to December 31, 2020, to assess the effectiveness of RL agents. Unlike our above experiments, which use simulated paths to test hedging strategies, this experiment examines their performance based on the historical series $\{R_t, \beta_t\}$. Specifically,

we evaluate the hedging performance considering a new European ATM call option with a maturity of 63 days every 21 business days along this historical path. The option prices, serving as initial hedging portfolio values, are determined with the prevailing IV surface on the day the hedge is initiated.

To assess the robustness of the model under more general market conditions, we conduct a comparison of cumulative P&Ls, which are computed as the cumulative sums of the P&L achieved by each strategy at the maturity of each option during the analyzed period. As illustrated in Figure 2.8, which depicts the evolution of the cumulative P&L across two panels, each representing a different transaction cost level, the gap in cumulative P&L between RL agents and benchmarks significantly widens as transaction cost rates increase. This, again, highlights the adaptability of the RL approach to various market conditions.

Figure 2.8: Cumulative P&L for ATM call options with a maturity of 63 days under real asset price dynamics.



Results are computed based on the observed P&L from hedging 296 synthetic ATM Call options under real market conditions observed from May 1, 1996, to December 31, 2020. A new option is considered every 21 business days. Agents are trained under the reduced state space $(\delta_t, \tau_t, S_t, \{\beta_{t,i}\}_{i=1}^{5}, h_{t,R})$ according to the conditions outlined in Section 2.4.3.1.

In contrast to the findings presented in Section 2.4.4.1, where SI delta yielded the highest profitability among all strategies, the RL approach achieved the highest profitability in historical paths. Specifically, agents trained under the MSE and SMSE emerged as the most profitable ones.

Moreover, RL agents consistently outperform benchmarks across most of the transaction cost

levels. However, when the transaction cost is small (left panel), Benchmarks exhibit a better cumulative P&L compared to the RL agent trained under the $\text{CVaR}_{99\%}$ penalty function at the end of the period, December 31, 2020. Nevertheless, Benchmarks also demonstrate more variability and larger losses, such as those observed between 2008 and 2012. In contrast, RL agents show a smaller decrease in cumulative P&L, indicating more resilience during crisis periods. In general, the observed performance of the RL agents under historical data is consistent with our findings under simulated data, indicating the robustness of our approach.

2.5 Conclusion

This study introduces a novel deep hedging framework that integrates forward-looking volatility information through a functional representation of the IV surface, combined with backward-looking conventional features. Our implementation employs deep policy gradient methods and utilizes a unique neural network architecture consisting of LSTM cells and FFNN layers to enhance training efficiency. Additionally, the architecture incorporates a budget constraint mitigating the incentive to gamble and enabling the agent to learn hedging strategies that are more effective for risk management.

Our approach consistently outperforms traditional benchmarks both in the absence and presence of transaction costs. The stability of our approach is assessed across various economic states using simulated data, demonstrating greater robustness than benchmarks in extreme scenarios such as economic crises.

The global importance analysis of IV features confirms the significant enhancement of hedging performance in terms of risk reduction relative to the penalty functions. Our analysis underscores the critical importance of key factors such as the underlying asset return conditional variance process (h_R), the long-term ATM IV level (β_1) and the time-to-maturity slope (β_2). RL agents utilize both the historical variance process and market expectations of future volatility to adjust positions in the underlying asset.

2.6 Appendix

2.6.1 Details for the MSGD training approach

The MSGD method estimates the penalty function $\mathcal{O}(\theta)$, which is typically unknown, through small samples of the hedging error called batches. Let $\mathbb{B}_j = \{\xi_{T,i}^{\tilde{\delta}_{\theta_j}}\}_{i=1}^{N_{\text{batch}}}$ be the j-th batch where N_{batch} is the batch size and $\xi_{T,i}^{\tilde{\delta}_{\theta_j}}$ denotes the hedging error of the i-th path in the j-th batch defined as

$$\xi_{T,i}^{\tilde{\delta}_{\theta_j}} = \Psi(S_{T,(j-1)N_{\text{batch}}+i}) - V_{T,i}^{\tilde{\delta}_{\theta_j}} \quad \text{for} \quad i \in \{1,\dots,N_{\text{batch}}\},\, j \in \{1,...N\},$$

where $S_{T,(j-1)N_{\text{batch}}+i}$ is the price of the underlying asset at time T in the $((j-1)N_{\text{batch}}+i)$-th simulated path, $V_{T,i}^{\tilde{\delta}_{\theta_j}}$ is the terminal value of the hedging strategy for that path when $\theta = \theta_j$ and the simulated states are X_i.

The penalty function estimation for the batch $\mathbb{B}$ is

$$\begin{aligned}
\hat{C}^{(\text{MSE})}(\theta_j, \mathbb{B}_j) &= \frac{1}{N_{\text{batch}}} \sum_{i=1}^{N_{\text{batch}}} \left(\xi_{T,i}^{\tilde{\delta}_{\theta_j}}\right)^2, \\
\hat{C}^{(\text{SMSE})}(\theta_j, \mathbb{B}_j) &= \frac{1}{N_{\text{batch}}} \sum_{i=1}^{N_{\text{batch}}} \left(\xi_{T,i}^{\tilde{\delta}_{\theta_j}}\right)^2 \mathbb{1}_{\left\{\xi_{T,i}^{\tilde{\delta}_{\theta_j}} \geq 0\right\}}, \\
\hat{C}^{(\text{CVaR})}(\theta_j, \mathbb{B}_j) &= \widehat{\text{VaR}}_\alpha(\mathbb{B}_j) + \frac{1}{(1-\alpha)N_{\text{batch}}} \sum_{i=1}^{N_{\text{batch}}} \max\left(\xi_{T,i}^{\tilde{\delta}_{\theta_j}} - \widehat{\text{VaR}}_\alpha(\mathbb{B}_j), 0\right),
\end{aligned}$$

where $\widehat{\text{VaR}}_\alpha(\mathbb{B}_j) = \xi_{T,\lceil \alpha \cdot N_{\text{batch}} \rceil}^{\tilde{\delta}_{\theta_j}}$ is the estimation of the VaR obtained from the ordered sample $\{\xi_{T,[i]}^{\tilde{\delta}_{\theta_j}}\}_{i=1}^{N_{\text{batch}}}$ and $\lceil \cdot \rceil$ is the ceiling function. These empirical approximations are used to estimate the gradient of the penalty function required in Equation (2.4).[9]

The selection of batch size plays a key role in the MSGD training approach as we empirically

[9] In particular, the gradient of these estimations has analytical expressions for FFNN, LSTM networks and thus for RNN-FNN networks. Details about gradient of the empirical objective function are provided in Goodfellow et al. (2016).

measure tail risk. A larger batch size provides a more accurate gradient estimate by averaging gradients computed over more simulated paths, thereby promoting stable convergence during training. However, large batch sizes may introduce certain disadvantages such as increased memory requirements, slower convergence, and potential generalization issues. We adopt the batch size used in Carbonneau (2021) ($N_{\text{batch}} = 1{,}000$), which achieves a good balance between accuracy and convergence.

2.6.2 Joint Implied Volatility and Return model

2.6.2.1 Daily implied volatility surface

The functional representation of the IV surface model introduced by François et al. (2022) is

$$\begin{aligned}\sigma(M_t, \tau_t, \beta_t) = & \underbrace{\beta_{t,1}}_{f_1\text{: Long-term ATM IV}} + \beta_{t,2} \underbrace{e^{-\sqrt{\tau_t/T_{conv}}}}_{f_2\text{: Time-to-maturity slope}} + \beta_{t,3} \underbrace{\left(M_t \mathbb{1}_{\{M_t \geq 0\}} + \frac{e^{2M_t} - 1}{e^{2M_t} + 1} \mathbb{1}_{\{M_t < 0\}} \right)}_{f_3\text{: Moneyness slope}} \\ & + \beta_{t,4} \underbrace{\left(1 - e^{-M_t^2}\right) \log(\tau_t / T_{max})}_{f_4\text{: Smile attenuation}} + \beta_{t,5} \underbrace{\left(1 - e^{(3M_t)^3}\right) \log(\tau_t / T_{max}) \mathbb{1}_{\{M_t < 0\}}}_{f_5\text{: Smirk}}, \quad \tau_t \in [T_{min}, T_{max}]\end{aligned} \tag{2.10}$$

where T_{max} is set to 5 years, $T_{min} = 6/252$ and $T_{conv} = 0.25$. As in Dumas et al. (1998), a minimum threshold of 0.01 is applied to the volatility surface to prevent negative values.

2.6.2.2 Joint Implied Volatility and Return

The JIVR model proposed by François et al. (2023) has 6 components, one for the underlying asset excess returns and the other 5 for fluctuations of the IV surface coefficients. The S&P 500 excess return follows:

$$\begin{aligned} R_{t+1} &= \xi_{t+1} - \psi(\sqrt{h_{t+1,R}\Delta}) + \sqrt{h_{t+1,R}\Delta}\epsilon_{t+1,R}, \\ h_{t+1,R} &= Y_t + \kappa_R(h_{t,R} - Y_t) + a_R h_{t,R}(\epsilon_{t,R}^2 - 1 - 2\gamma_R\epsilon_{t,R}), \\ Y_t &= \left(\omega_R\, \sigma\left(0, \frac{1}{12}, \beta_t\right)\right)^2, \end{aligned} \tag{2.11}$$

where the equity risk premium is

$$\xi_{t+1} = \psi(-\lambda\sqrt{h_{t+1,R}\Delta}) - \psi((1-\lambda)\sqrt{h_{t+1,R}\Delta}) + \psi(\sqrt{h_{t+1,R}\Delta}), \tag{2.12}$$

the process $\{\epsilon_{t,R}\}_{t=0}^{T}$ is a sequence of iid standardized NIG random variables with parameters ζ_R and φ_R,[10] and ψ is their cumulant generating function.[11] Parameters of the excess return component of the model are thus $\Theta_R = (\lambda, \kappa_R, \gamma_R, a_R, \omega_R, \zeta_R, \varphi_R)$.

The evolution of the long-term factor (β_1) is

$$\begin{aligned}
\beta_{t+1,1} &= \alpha_1 + \sum_{i=1}^{5} \theta_{1,j}\beta_{t,j} + \sqrt{h_{t+1,1}\Delta}\epsilon_{t+1,1}, \\
h_{t+1,1} &= U_t + \kappa_1(h_{t,1} - U_t) + a_1 h_{t,1}(\epsilon_{t,1}^2 - 1 - 2\gamma_1\epsilon_{t,1}), \\
U_t &= \left(\omega_1 \cdot \sigma\left(0, \frac{1}{12}, \beta_t\right)\right)^2.
\end{aligned} \tag{2.13}$$

For the other 4 coefficients, $i \in \{2,3,4,5\}$, the time evolution satisfies

$$\begin{aligned}
\beta_{t+1,i} &= \alpha_i + \sum_{j=1}^{5} \theta_{i,j}\beta_{t,j} + \nu\beta_{t-1,2}\mathbb{1}_{\{i=2\}} + \sqrt{h_{t+1,i}\Delta}\epsilon_{t+1,i}, \\
h_{t+1,i} &= \sigma_i^2 + \kappa_i(h_{t,i} - \sigma_i^2) + a_i h_{t,i}(\epsilon_{t,i}^2 - 1 - 2\gamma_i\epsilon_{t,i}),
\end{aligned} \tag{2.14}$$

[10] The standard NIG random variable ϵ follows the probability density function with parameters ζ and φ:

$$f(x) = \frac{B_1\left(\sqrt{\frac{\varphi^6}{\varphi^2+\zeta^2} + (\varphi^2+\zeta^2)\left(x + \frac{\varphi^2\zeta}{\varphi^2+\zeta^2}\right)^2}\right)}{\pi\sqrt{\frac{1}{\varphi^2+\zeta^2} + \frac{\varphi^2+\zeta^2}{\varphi^6}\left(x + \frac{\varphi^2\zeta}{\varphi^2+\zeta^2}\right)^2}} e^{\left(\frac{\varphi^4}{\varphi^2+\zeta^2} + \zeta\left(x + \frac{\varphi^2\zeta}{\varphi^2+\zeta^2}\right)\right)},$$

where $B_1(\cdot)$ denotes the modified Bessel function of the second kind with index 1. The common $(\alpha, \beta, \delta, \mu)$-specification can be obtained by replacing β and γ ($\gamma = \sqrt{\alpha^2 - \beta^2}$), with ζ and φ, respectively, and imposing a null mean and unit variance to express δ and μ interms of α, β.

[11] For $-\sqrt{\zeta^2+\varphi^2} - \zeta < z < \sqrt{\zeta^2+\varphi^2} - \zeta$, the cumulant generating function is given by

$$\psi(z) = \frac{\varphi^2}{\varphi^2+\zeta^2}\left(-\zeta z + \varphi^2 - \varphi\sqrt{\varphi^2+\zeta^2-(\varphi+\zeta)^2}\right).$$

where $\{\epsilon_{t,i}\}_{i=1}^{5}$ are time-independent standardized NIG random variables with parameters $\{(\zeta_i, \varphi_i)\}_{i=1}^{5}$, respectively. Parameters are $\{\omega_1, \nu, \Theta_1, \Theta_2, \Theta_3, \Theta_4, \Theta_5\}$ with $\Theta_i = (\alpha_i, \{\theta_{i,1}\}_{i=1}^{5}, \sigma_i, \kappa_i, a_i, \gamma_i, \zeta_i, \varphi_i)\}_{i=1}^{5}$.

The JIVR model also imposes a dependence structure on contemporaneous innovations $\epsilon_t = (\epsilon_{t,R}, \epsilon_{t,1}, ..., \epsilon_{t,5})$ through a Gaussian copula parameterized in terms of a covariance matrix Σ of dimension 6×6.

Estimates of all JIVR model parameters are taken from Table 5 and Table 6 of François et al. (2023). In all experiments, the annualized continuously compounded risk-free rate and dividend yield are assumed to be constant with values set to $r = 2.66\%$ and $q = 1.77\%$, respectively.

2.6.3 Benchmarks

The three benchmarks outlined in this appendix operate under the premise that the implied volatilities follow the IV model introduced in Equation (2.5).

2.6.3.1 Black-Scholes model

The Black and Scholes (1973) pricing formula for an European call option is:

$$\text{Call}_t = S_t \cdot \mathrm{e}^{-q_t \tau_t} \Phi(d_t) - K \cdot e^{-r_t \tau_t} \Phi(d_t - \sigma(M_t, \tau_t, \beta_t)\sqrt{\tau_t})$$

where $d_t = \frac{\log\left(\frac{S_t}{K}\right) + \left(r_t - q_t + \frac{1}{2}\sigma(M_t, \tau_t, \beta_t)^2\right)\tau_t}{\sigma(M_t, \tau_t, \beta_t)\sqrt{\tau_t}}$, σ_t is the implied volatility of the option and Φ is the cumulative distribution function of the standard normal distribution. Moreover, the Black-Scholes delta is

$$\Delta_t^{BS} = \mathrm{e}^{-q_t \tau_t} \Phi(d_t).$$

2.6.3.2 Leland model

The Leland delta, as proposed by Leland (1985), presents a variation of the option replication approach outlined in the work of Black and Scholes (1973), incorporating parameters such as

the transaction cost proportion κ and the rebalancing frequency λ.

$$\Delta_t^L = \mathrm{e}^{-q_t\tau_t}\Phi\left(\tilde{d}_t\right),$$

where $\tilde{d}_t = \frac{\log\left(\frac{S_t}{K}\right)+\left(r_t-q_t+\frac{1}{2}\tilde{\sigma}_t^2\right)\tau_t}{\tilde{\sigma}_t\sqrt{\tau_t}}$ with $\tilde{\sigma}_t^2 = \sigma(M_t,\tau_t,\beta_t)^2\left[1+\sqrt{2/\pi}\frac{2\kappa}{\sigma(M_t,\tau_t,\beta_t)\sqrt{\lambda}}\right]$.

2.6.3.3 Smile-implied model

François et al. (2022) provide the delta associated to the IV surface under model (2.10):

$$\Delta_t^{SI} = e^{-q_t\tau_t}\left(\Phi(\lceil_{t,1}) + \phi(\lceil_{t,1})\frac{\partial\sigma}{\partial M}\right)$$

with

$$\lceil_{t,1} = \frac{M_t}{\sigma(M_t,\tau_t,\beta_t)} + \frac{1}{2}\sigma(M_t,\tau_t,\beta_t)\sqrt{\tau_t},$$

and

$$\begin{aligned}\frac{\partial\sigma}{\partial M} = {} & \beta_{t,3}\mathbb{1}_{\{M_t\geq 0\}} + \beta_{t,3}\left(1-\left(\frac{e^{2M_t}-1}{e^{2M_t}+1}\right)^2\right)\mathbb{1}_{\{M_t<0\}} + \beta_{t,4}2M_te^{-M_t^2}\log\left(\frac{T}{T_{max}}\right) \\ & - \beta_{t,5}81M_t^2e^{27M_t^3}\log\left(\frac{T}{T_{max}}\right)\mathbb{1}_{\{M_t<0\}}.\end{aligned}$$

2.6.4 Network fine-tuning

2.6.4.1 Bounded strategies - leverage constraint

This section presents a comparison of the hedging performance of RL agents trained under the full state space considering the $\text{CVaR}_{95\%}$ as penalty function. The comparison is conducted considering two RL agents: (i) RL-$\text{CVaR}_{95\%}$-LC, agent subject to a leverage constraint equivalent to the initial value of the underlying asset, set at $B = 100$, and (ii) RL-$\text{CVaR}_{95\%}$, agent operating without leverage constraints. This comparison evaluates the estimated values of all penalty functions and the average Profit and Loss (Avg P&L). This analysis aims to elucidate the impact of boundary conditions on RL agent behavior and performance when

hedging of a European ATM call option with a maturity of $N = 63$ days in the absence of transaction costs, i.e., $\kappa = 0\%$).

Table 2.6 outlines that the agent without a leverage constraint shows more profitable strategies; however, it also leads to very large losses. For instance, out-of-sample $\text{CVaR}_{99\%}$ and SMSE metrics exhibit higher values for this agent, compared to its counterpart with a leverage constraint.

Table 2.6: RNN-FNN hedging error statistics of a short position in a ATM call option with maturity of 63 days.

Function	RL-$\text{CVaR}_{95\%}$-LC	RL-$\text{CVaR}_{95\%}$
Avg P&L	0.440	**0.715**
$\text{CVaR}_{95\%}$	1.294	**1.242**
$\text{CVaR}_{99\%}$	**2.502**	2.622
MSE	**1.205**	14.650
SMSE	**0.241**	0.327

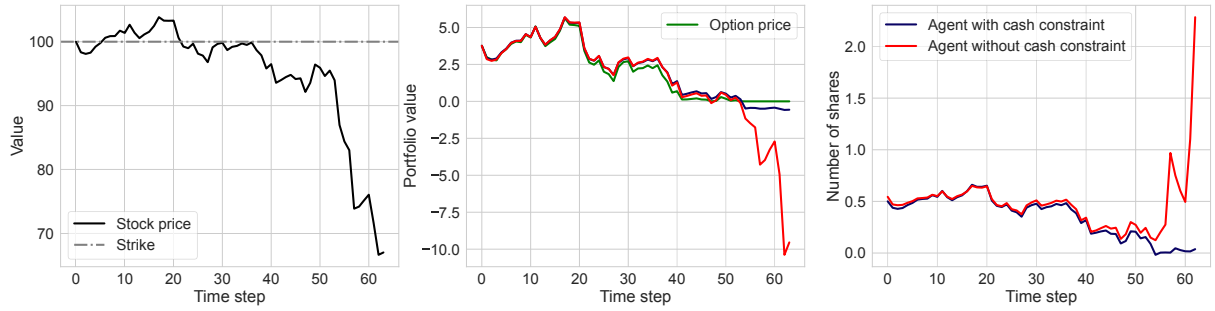
Results are computed using 100,000 out-of-sample paths in the absence of transaction costs ($\kappa = 0\%$). Agents are trained under the full state space $(V_t^\delta, \delta_t, \tau_t, S_t, \{\beta_{t,i}\}_{i=1}^5, h_{t,R}, \{h_{t,i}\}_{i=1}^5)$ according to the conditions outlined in Section 2.4.3.1. The average option price is \$3.89 with a standard deviation of \$1.29. RL-$\text{CVaR}_{95\%}$-LC denotes the RL agent trained with a leverage constraint, while RL-$\text{CVaR}_{95\%}$ refers to the agent without it. Best performances are highlighted in bold.

Moreover, our numerical experiments demonstrate that the agent without leverage constraints learns doubling strategies, which are incompatible with sound risk management practices. For instance, Figure 2.9 illustrates such behavior through three panels associated with the hedging process of a deep OTM path (first panel).[12] We observe that the agent, without a leverage constraint, tends to increase its position in the underlying (third panel) when a loss in the portfolio value is observed (second panel), aiming to recover the loss over the long run

[12]For the purpose of this experiment, a deep OTM path refers to the trajectory of the underlying asset that keeps the option significantly out-of-the-money.

with doubling strategies. Conversely, agents trained with a leverage constraint control their position in the underlying asset, resulting in the learning of different and less risky strategies.

Figure 2.9: Doubling strategy dynamics for a short position in a ATM call option with a maturity of 63 days.



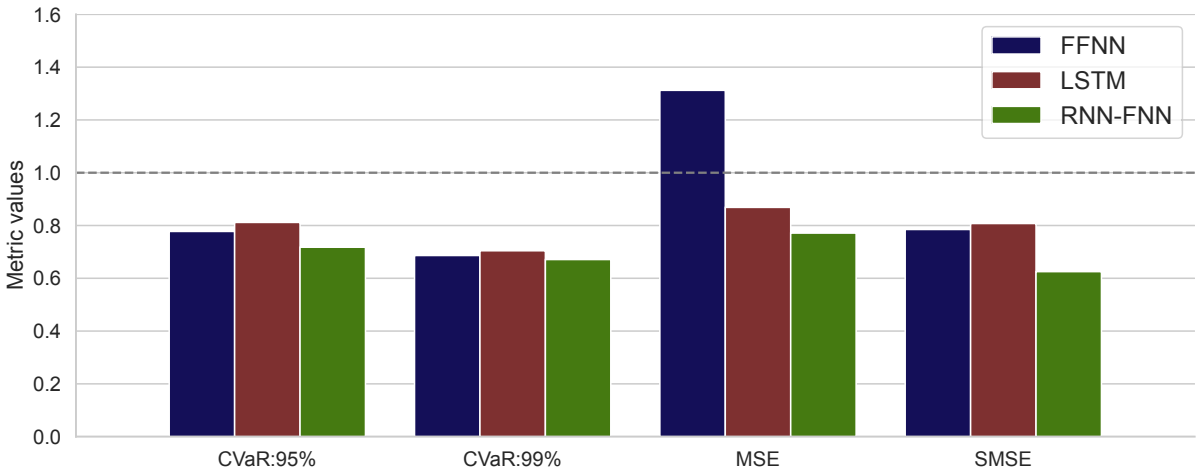
Results are obtained using a deep OTM path in the absence of transaction costs ($\kappa = 0\%$), with agents trained according to the conditions outlined in Section 2.4.3.1. The training encompasses the full state space and utilizes $\text{CVaR}_{95\%}$ as a penalty function. The agent referred to as "Agent with cash constraint" incorporates a leverage constraint of $B = 100$, whereas the agent labeled "Agent without cash constraint" does not.

2.6.4.2 Network architecture selection

In this section, we investigate the superiority of the RNN-FNN architecture compared to conventional architectures introduced in deep hedging literature, such as the FFNN and the LSTM architectures. In line with our previous experiments, we consider four penalty functions to train the agents under the full state space: $\text{CVaR}_{95\%}$, $\text{CVaR}_{99\%}$, MSE, and SMSE. Again, our experiment focuses on hedging a short position of a European ATM call option with a maturity of $N = 63$ days, assuming no transaction costs.

The superiority of the RNN-FNN over the LSTM and the FFNN is evaluated based on the optimal value of each penalty function. Figure 2.10 illustrates the optimal values of each penalty function for each architecture, normalized by the estimated value obtained with the BS delta. Notably, the RNN-FNN setup considered in this paper significantly outperforms both the benchmark and other architectures. Conversely, the FFNN does not surpass the benchmark for the MSE, and the LSTM exhibits almost the same performance as the benchmark.

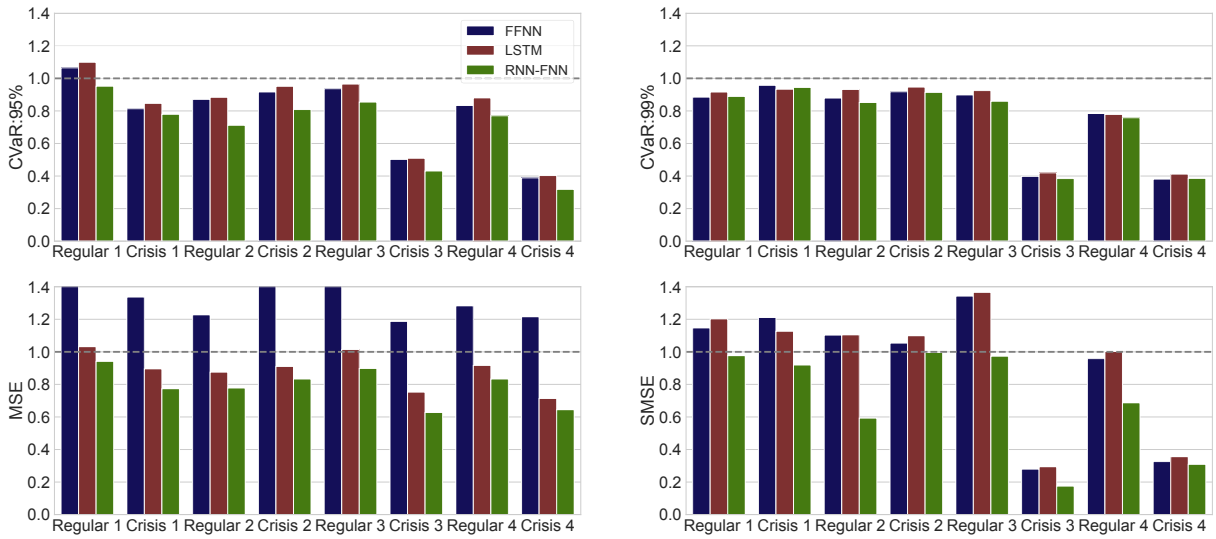
Figure 2.10: Network performance for a short position in a ATM call option with a maturity of 63 days.



Results are computed using 100,000 out-of-sample paths in the absence of transaction costs ($\kappa = 0\%$). Agents are trained under the reduced state space $(\tau_t, S_t, \{\beta_{t,i}\}_{i=1}^5, h_{t,R}, \{h_{t,i}\}_{i=1}^5)$ according to the conditions outlined in Section 2.4.3.1. The setup for the different networks follows the architecture described in Section 2.2.2.1, with $L_1 = 0$ and $L_2 = 4$ for the FFNN, $L_1 = 4$ and $L_2 = 0$ for the LSTM, and $L_1 = 2$ and $L_2 = 2$ for RNN-FNN. Results show optimal values obtained from agents trained under $\text{CVaR}_{95\%}$, $\text{CVaR}_{99\%}$, MSE, and SMSE for the three networks. These values are normalized by the estimated values of each penalty function obtained with the BS delta.

A second test to demonstrate the superiority of our architecture involves computing the optimal values of the penalty functions over various clusters of paths. The objective is to isolate the impact of different state of the economy on performance and assess the robustness of our architecture. Figure 2.11 displays the optimal values of each architecture normalized by the estimated values obtained by the BS delta for all penalty functions. Notably, RNN-FNN agents outperform both the benchmark and other architectures across all penalty functions, regardless of the economic conditions under which the simulations were conducted. Conversely, FFNN and LSTM agents fail to outperform the benchmark across all states of the economy when the agents are trained under the $\text{CVaR}_{95\%}$, MSE and SMSE penalty functions, as shown in the top-left, bottom-left and bottom-right panels, respectively.

Figure 2.11: Neural network performance for a short position in an ATM call option with a maturity of 63 days: sensitivity to the state of the economy.



Results are computed using 100,000 out-of-sample paths in the absence of transaction costs ($\kappa = 0\%$). Agents are trained under the reduced state space $(\tau_t, S_t, \{\beta_{t,i}\}_{i=1}^5, h_{t,R}, \{h_{t,i}\}_{i=1}^5)$ according to the conditions outlined in Section 2.4.3.1. Each panel illustrates the values obtained by different architectures for the following hedging metrics: Avg P&L, $\text{CVaR}_{95\%}$, MSE, and SMSE. These metrics are normalized by the estimated values of each penalty function obtained with the BS delta. The setup for the different networks follows the architecture described in Section 2.2.2.1, with $L_1 = 0$ and $L_2 = 4$ for the FFNN, $L_1 = 4$ and $L_2 = 0$ for the LSTM, and $L_1 = 2$ and $L_2 = 2$ for RNN-FNN.[13]

Results obtained under various economic conditions unequivocally demonstrate the superiority of the RNN-FNN network across all performance metrics. Moreover, the RNN-FNN not only offers better performance in terms of risk management but also reduces computational costs and implementation complexity, with an average reduction in training time of 46%.

2.6.4.3 Dropout parameter selection

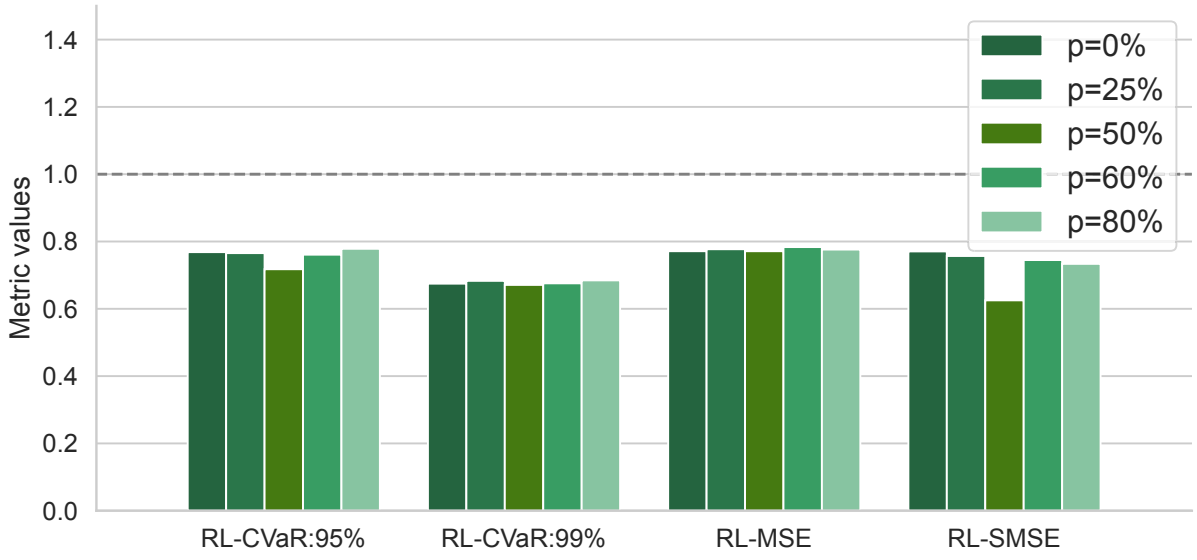
The process of selecting the dropout parameter for the regularization method involved evaluating the performance of four agents across a range of potential parameter values. These agents are trained using the full state space, considering four penalty functions: $\text{CVaR}_{95\%}$, $\text{CVaR}_{99\%}$, MSE, and SMSE. The performance of the agents is measured in terms of the

[13] The periods aim to approximate different states of the economy considering the time frames specified in Table 2.3.

estimated values of the penalty functions after hedging a short position of a European call option with a maturity of $N = 63$ days, with no transaction costs.

Figure 2.12 illustrates a comparison of the optimal values of penalty functions normalized by the estimated values obtained using the BS delta. It is noteworthy that the RNN-FNN network consistently outperforms the benchmark regardless of the dropout parameter value. Additionally, the selection of the dropout parameter remains consistent across all objective functions, with a dropout probability (p) of 50% yielding optimal performance. Consequently, for all our experiments, a dropout regularization parameter of $p = 50\%$ is adopted. This parameter drives the likelihood of randomly dropping out a fraction of the units within a neural network during training, thereby generating varied architectures for each training iteration. As demonstrated in Warde-Farley et al. (2013), this method not only effectively mitigates overfitting but also boosts performance, which is consistent with our numerical results.

Figure 2.12: RNN-FNN performance for a short position in an ATM call option with a maturity of 63 days: the effect of the dropout parameter in the training phase.



Results are computed using 100,000 out-of-sample paths in the absence of transaction costs ($\kappa = 0\%$). Agents are trained under the reduced state space $(\tau_t, S_t, \{\beta_{t,i}\}_{i=1}^5, h_{t,R}, \{h_{t,i}\}_{i=1}^5)$ according to the conditions outlined in Section 2.4.3.1. All the metrics are expressed in proportion of values obtained under the Black-Scholes delta hedging strategy.

2.6.4.4 Quadratic hedging problem

This appendix aims to compare the closed-form solution of the quadratic hedging problem as outlined by Godin (2019) with our own approach within the framework of Black-Scholes market dynamics, where log-returns are assumed to adhere to a Gaussian distribution. Table 2.8 illustrates performance metrics using three distinct penalty functions from two experiments analyzing the hedging error of an ATM call option with a strike price of K=100, with maturities of 63 days and 252 days. The experiments involve 16 time steps and 5 time steps for rebalancing, respectively.

The outcomes of the 63-day maturity ATM option reveal that the RL agents outperform the two benchmarks, Black-Scholes delta (BS) and the quadratic hedging solution (QH), examined in this study. Additionally, the performance gap between the RL agent trained with the full state space (RL-Full) and the agent trained with the reduced state space (RL-Reduced) is negligible. In fact, the Kolmogorov-Smirnov test fails to reject the null hypothesis that the hedging errors of both agents are equally distributed, with a confidence level of 99.9%.[14]

[14]The Kolmogorov-Smirnov test, outlined in Darling (1957), is a non-parametric statistical test used to determine whether two samples differ significantly.

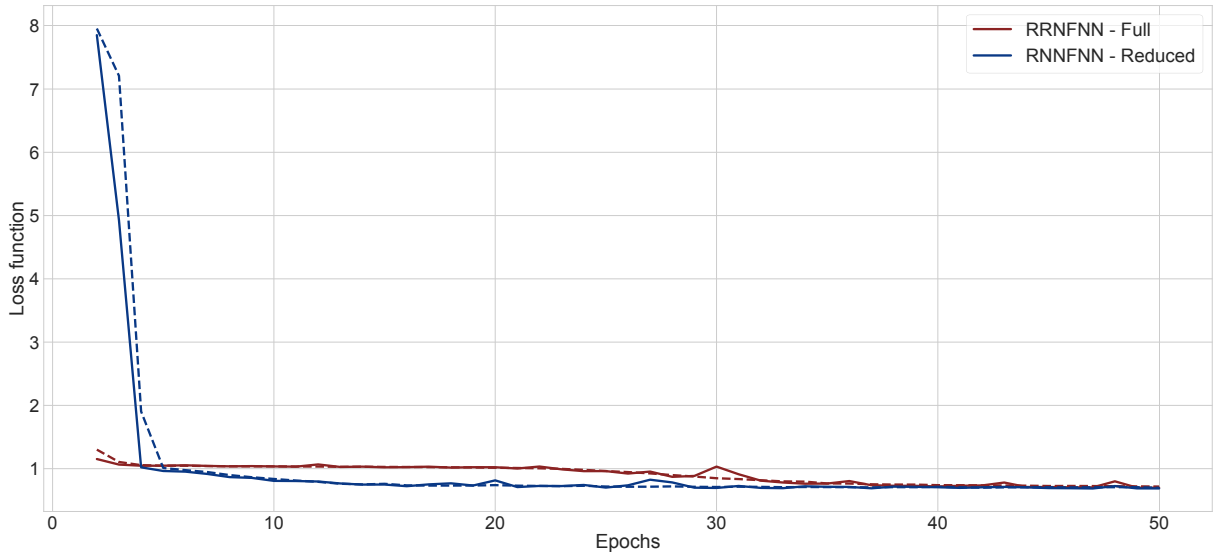
Table 2.8: RNN-FNN hedging error statistics for a short position ATM call option with two different maturities and rebalancing periods under the MSE as penalty function.

	Maturity: 63, Time steps:16				Maturity: 252, Time steps:5			
Function	BS	QH	RL-F	RL-R	BS	QH	RL-F	RL-R
Avg P&L	0.005	-0.081	-0.014	-0.009	0.185	-0.335	-0.279	-0.216
$\text{CVaR}_{95\%}$	1.942	2.619	1.897	1.931	7.291	6.662	6.596	6.514
$\text{CVaR}_{99\%}$	2.896	3.748	2.808	2.881	10.62	10.488	9.913	9.624
MSE	0.684	1.272	0.681	0.683	8.691	7.609	7.865	7.834
SMSE	0.367	0.647	0.350	0.359	5.292	3.721	3.852	3.941

These results are computed considering the hedging error of 99,000 out-of-sample independent paths from the Black-Scholes market with yearly parameters $\mu = 0.0892$ and and $\sigma = 0.1952$. The RNN-FNN is trained based on 400,000 independent paths under the same scheme.

Conversely, results of the 252-day maturity ATM option reveal that the QH approach exhibits slightly superior performance in MSE, as anticipated due to its closed-form nature. However, the performance of RL agents demonstrates stronger potential in terms of risk management, evident from their ability to yield the lowest CVaR values and closely approximate the MSE of the closed-form solution. Moreover, the agent trained with the reduced state space (RL-Reduced) exhibits enhanced performance compared to its counterpart trained with the full state space.

Figure 2.13: RNN-FNN loss function for a short position ATM call option with maturity $N = 63$ days and 16 time steps for rebalancing.



Results are computed considering the MSE as the loss functions and the hedging error of 99,000 out-of-sample independent paths with random initialization for the validation loss curve (full line) and 400,000 independent paths fot the training loss curve (dash line). Simulations are performed based on the Black-Scholes model and agents are trained considering the cash constraint of $B = 100$.

Consistent with the findings detailed in Section 2.4.3.2 regarding the dynamics of the JIVR model, the RL agent trained with the reduced state space exhibits improved the rate of convergence during the training phase. For instance, as depicted in Figure 2.13, the penalty curve evolution for the RL-MSE agent over 50 epochs illustrates this trend. This approach effectively reduces computational costs during training and accelerates convergence to optimal performance. These numerical outcomes confirm that our method achieves robust performance without necessitating the inclusion of portfolio value, even within the Quadratic Hedging framework with Black-Scholes market dynamics.

2.6.4.5 JIVR Model parameters

Table 2.10: Estimated Gaussian copula parameters

	$\epsilon_{t,R}$	$\epsilon_{t,1}$	$\epsilon_{t,2}$	$\epsilon_{t,3}$	$\epsilon_{t,4}$	$\epsilon_{t,5}$
$\epsilon_{t,R}$	1.000					
$\epsilon_{t,1}$	-0.550	1.000				
$\epsilon_{t,2}$	-0.690	0.140	1.000			
$\epsilon_{t,3}$	0.030	-0.030	-0.0100	1.000		
$\epsilon_{t,4}$	-0.220	0.250	0.120	0.280	1.000	
$\epsilon_{t,5}$	-0.340	0.170	0.370	0.130	-0.050	1.000

Table 2.11: JIVR model parameter estimates

Parameter	β_1	β_2	β_3	β_4	β_5		S&P500
α	0.000899	0.008400	0.000770	-0.001393	0.000657	λ	2.711279
θ_1	0.996290	-0.013869		0.002841			
θ_2	0.003669	0.877813	0.001300				
θ_3		-0.032640	0.997071	0.003722	-0.004198		
θ_4				0.980269			
θ_5		-0.047789			0.986019		
ν		0.089445					
$\sigma\sqrt{252}$		0.380279	0.052198	0.048641	0.051536		
ω	0.267589						0.977291
κ	0.838220	0.965751	0.974251	0.945377	0.980844		0.888977
a	0.134152	0.098272	0.092646	0.102201	0.100502		0.056087
γ	-0.111813	-1.482862	0.096766	0.060558	-0.102996		2.507796
ζ	0.143760	0.852943	0.029109	-0.159051	0.092664		-0.641306
φ	1.351070	1.538928	2.284780	1.449977	1.428477		2.039669

Chapter 3

Is the Difference between Deep Hedging and Delta Hedging a Statistical Arbitrage?

Abstract

Horikawa and Nakagawa (2024) claim that in a complete market admitting statistical arbitrage, the difference between the deep hedging and the replicating portfolio hedging positions is a statistical arbitrage. Deep hedging can thus include an undesirable speculative component. We test whether this remains true in a GARCH-based incomplete market dynamics. We observe that the difference between deep hedging and delta hedging is a speculative overlay if the risk measure considered does not put sufficient relative weight on adverse outcomes. Nevertheless, a suitable choice of risk measure can prevent the deep hedging agent from engaging in speculation.

JEL classification: C45, C61, G32.

Keywords: Deep reinforcement learning, optimal hedging, arbitrage.

3.1 Introduction

The seminal paper of Buehler et al. (2019), which proposes to use deep reinforcement learning (RL) methods to obtain optimal hedging procedures for financial derivatives, initiated a recent strand of literature.[15] Deep RL methods are particulary well-suited to solve dynamic hedging problems because these methods can handle the curse of dimensionality, a problem that more traditional approaches (e.g., finite elements dynamic programming) struggle to overcome. They can also work with very general dynamics for asset prices, not being limited by mathematical tractability issues.

While the ability of deep hedging strategies to outperform standard counterparts is well-documented, the existing literature has not yet extensively analyzed the structure of optimal policies and explained how such incremental performance is attained. Neagu et al. (2024) make a step in that direction by investigating the impact of the various features on optimal risk management decisions in the presence of illiquidity market impacts.

In their recent work, Horikawa and Nakagawa (2024) investigate complete markets that allow for statistical arbitrage with respect to a specific risk measure ρ. They assert that, within this framework, deep hedging strategies that minimize the chosen risk metric combine the traditional delta-hedging approach with a statistical arbitrage overlay. In a vector auto-regressive stochastic volatility model and in a GAN-simulated market model, Buehler et al. (2021) find that the optimal hedging strategy maximizing the entropy utility can also incorporate a statistical arbitrage component. Such claims raise concerns about the suitability of the deep hedging approach; incorporating speculative or arbitrage-like components that do not contribute to reducing the risk exposure within hedging portfolios would be deemed undesirable in practice. Our objective is therefore to assess empirically whether deep

[15] See for instance Halperin (2019), Cao et al. (2020), Du et al. (2020), Carbonneau and Godin (2021), Carbonneau (2021), Horvath et al. (2021), Imaki et al. (2021), Lütkebohmert et al. (2022), Cao et al. (2023), Carbonneau and Godin (2023), Marzban et al. (2023b), Mikkilä and Kanniainen (2023), Raj et al. (2023) and Wu and Jaimungal (2023). See also Hambly et al. (2023) and Pickard and Lawryshyn (2023) for related surveys.

hedging policies minimizing conventional risk metrics still contain a speculative component in incomplete market settings, which would generalize the complete market conclusion of Horikawa and Nakagawa (2024). We use a GARCH-based market setting as an illustrative example. A GARCH model is chosen due to its simplicity and its ability to reflect one of the most natural causes of market incompleteness in practice, namely time-varying volatility. Among the GARCH family members, we picked the GJR-GARCH because it captures well the negative skewness typically observed for S&P 500 index returns.

This study uses the CVaR (Rockafellar and Uryasev, 2002) as the risk measure driving the optimization of the hedging strategy. The CVaR metric accounts for a large spectrum of risk preferences depending on how the confidence level is set. As far as market risk management by regulated financial institutions is concerned, the confidence level is usually set at a very high level to account for a conservative risk management against extreme scenarios. But more generally, the confidence level reflects the manager's attitude toward risk, as illustrated for example by Su and Li (2024): When the confidence level is set to zero, the CVaR reduces to the expectation operator and the manager is risk-neutral; when it approaches 1, the CVaR becomes the maximum operator and the extremely conservative manager only focuses on the worst-case scenario. The CVaR is widely used in practice to measure risk, and its use differentiates our work from Buehler et al. (2021) who rely on the entropy risk measure which is not widely adopted by practitioners.

The paper is divided as follows. Section 3.2 provides the hedging problem formulation. Section 3.3 describes the deep hedging framework used to solve the problem, and discusses the delta hedging benchmark. Numerical experiments assessing the behavior of the deep versus delta hedging difference strategy are provided in Section 3.4.[16] Section 3.5 concludes.

[16]The Python code which allows replicating the numerical experiments from this paper can be found at https://github.com/cpmendoza/DeepHedging_StatisticalArbitrage.git.

3.2 Market model for hedging

This paper considers dynamic risk management strategies for European call options, which involve the construction of a self-financing portfolio composed of the underlying asset and a cash account. The portfolio is rebalanced daily to optimally offset the net risk exposure at the option maturity, denoted as T days. The time-t underlying asset price is S_t. The trading strategy is represented by the predictable process $\delta = \{\delta_t\}_{t=1}^{T}$, where δ_t is the number of underlying asset shares held during the interval $(t-1, t]$. The time-t discounted gain made by the hedging portfolio is $G_t^\delta = \sum_{k=1}^{t} \delta_k(\beta^k S_k \mathrm{e}^{q\Lambda} - \beta^{k-1} S_{k-1})$ with $\beta = \mathrm{e}^{-r\Lambda}$, where r is the annualized continuously compounded risk-free rate, q is the annualized underlying asset dividend yield, and the period length is $\Lambda = \frac{1}{252}$ years. The time-t self-financing portfolio value is

$$V_t^\delta(V_0) = \beta^{-t}(V_0 + G_t^\delta), \tag{3.1}$$

where V_0 is the initial portfolio value that we set to the option price.

The hedging problem is a sequential decision problem where the holder of a short position in a call option seeks for the best sequence of actions δ that minimizes the risk associated with the hedging error

$$\xi_T^\delta = \max(S_T - K, 0) - V_T^\delta(V_0), \tag{3.2}$$

where K is the call option strike price. The hedging problem is formulated as

$$\delta^* = \arg\min_{\delta} \rho\left(\xi_T^\delta\right), \tag{3.3}$$

where ρ is the risk measure used by the agent to quantify risk. In this paper we consider the Conditional Value-at-Risk (CVaR_α) defined as $\rho(\xi_T^\delta) = \mathbb{E}[\xi_T^\delta \mid \xi_T^\delta \geq \mathrm{VaR}_\alpha(\xi_T^\delta)]$, where $\alpha \in (0,1)$ and $\mathrm{VaR}_\alpha(\xi_T^\delta)$ is the Value-at-Risk defined as $\mathrm{VaR}_\alpha(\xi_T^\delta) = \min_c\{c : \mathbb{P}(\xi_T^\delta \leq c) \geq \alpha\}$. The CVaR is a commonly used objective function in the deep hedging literature, see for instance Carbonneau and Godin (2021), Cao et al. (2023) or Wu and Jaimungal (2023). In

addition, an appealing feature of the CVaR is that it allows to finetune the investor's attitude towards risk through the confidence level. A high value of α puts more emphasis on risk reduction, whereas a low value of α penalizes losses and rewards gains.

Each time-t action δ_{t+1} is a feedback-type decision, being a function of the information currently available on the market: $\delta_{t+1} = \tilde{\delta}(X_t)$ for some function $\tilde{\delta}$ of the state variable vector X_t.

3.3 Hedging strategies

3.3.1 Deep hedging

The deep hedging (DH) framework, introduced by Buehler et al. (2019), provides a solution to the hedging problem (3.3) by leveraging RL techniques. The DH policy $\tilde{\delta}$ is approximated with a neural network δ_θ^{DH} with parameters θ, which returns a hedging position δ_{t+1} when provided with time-t input features X_t. The objective function to be minimized is thus

$$\mathcal{O}(\theta) = \rho\left(\xi_T^{\delta_\theta^{DH}}\right). \tag{3.4}$$

The approach to obtain optimized parameters θ is standard and based on mini-batch stochastic gradient descent. All details pertaining to the optimization procedure and the considered architecture for the neural network are provided in Appendix 3.6.1 and Appendix 3.6.2. Note that agents are trained on training sets of 400,000 independent simulated paths, but numerical results are obtained from test sets of 100,000 independent paths, which provide out-of-sample results.

3.3.2 Delta hedging

Delta hedging aims to reduce the risk associated with price movements of an underlying asset by adjusting the hedging portfolio positions in the underlying asset based on the sensitivity (Δ) of the option price to changes in the price of the underlying asset. Specifically, the time-t position in the underlying asset is defined as the time-t sensitivity Δ_t, which is the partial

derivative of the time-t option price with respect to the underlying asset value.

3.3.3 Statistical arbitrage

Statistical arbitrage strategies, also known as "good deals" according to the terminology of Cochrane and Saa-Requejo (2000), are profit-seeking trading strategies that capitalize on statistical anomalies in the market. Bondarenko (2003) defines statistical arbitrage as a trading strategy that makes a profit on average without requiring any initial capital investment. Assa and Karai (2013) extend this definition to offer a more nuanced and comprehensive evaluation, ensuring that the trading strategy is not only profitable on average, but also resilient in terms of risk management. As in Assa and Karai (2013), we say that δ is a statistical arbitrage opportunity if

$$\rho\left(-V_T^{\delta}(0)\right) < 0, \tag{3.5}$$

that is, if the trading strategy δ which requires zero initial investment is deemed strictly less risky than a null investment according to risk measure ρ. Such definition is also in line with that of Buehler et al. (2021), who focus on the case of the entropy risk measure.

Horikawa and Nakagawa (2024) claim that in a complete market model that admits statistical arbitrage, the difference between the deep hedging and the delta hedging strategies denoted by

$$\delta^{-} = \delta^{DH} - \Delta, \tag{3.6}$$

is a statistical arbitrage strategy according to risk measure ρ. We wish to further extend their study and examine if the trading strategy δ^{-} behaves like a statistical arbitrage in more general incomplete market dynamics, using a GARCH-based market model as a representative candidate for illustration. In other words, we investigate whether the deep hedging approach typically incorporates a speculative arbitrage-like component aimed at exploiting the structure of the risk measure considered.

3.4 Numerical study

3.4.1 Stochastic market dynamics

We consider market dynamics based on a GARCH process to represent the underlying asset log-returns. The GJR-GARCH(1,1) model introduced by Glosten et al. (1993) captures time-varying volatility and accounts for the leverage effect. For $t = 1, \dots, T$, log-returns in the model follow

$$R_t = \mu + \sigma_t \epsilon_t, \quad \sigma_{t+1}^2 = \omega + \sigma_t^2(\alpha + \gamma \mathbb{1}_{\{\epsilon_t < 0\}})\epsilon_t^2 + \beta \sigma_t^2, \tag{3.7}$$

where $\mu, \gamma \in \mathbb{R}$, ω, α, β are positive, $\mathbb{1}_A$ is the dummy variable indicating if event A occurs and $\{\epsilon_t\}_{t=1}^T$ are independent standard normal random variables. Parameter estimates are obtained through maximum likelihood on a daily time series of the S&P 500 index extending from January 4, 2016, to December 31, 2020. Estimated parameters are $\mu = 0.06\%$, $\omega = 0.01\%$, $\alpha = 0.11$, $\gamma = 0.20$ and $\beta = 0.78$. Furthermore, in all experiments, the annualized continuously compounded risk-free rate and dividend yield are assumed to be constant with values set to $r = 1.67\%$ and $q = 1.65\%$, respectively. These values represent the historical averages of the 1-year zero-coupon yield and the annualized S&P 500 dividend yield over the period extending from 2016 to 2020.

The initial option price is computed using Monte Carlo simulation based on the risk-neutral valuation formula

$$\text{Call}_0 = \mathrm{e}^{-rT\Lambda} \mathbb{E}^{\mathbb{Q}}[\max(S_T - K, 0)], \tag{3.8}$$

where $\mathbb{Q}$ is a risk-neutral measure.[17]

[17] The $\mathbb{Q}$ risk-neutral dynamics of the GARCH model are defined by the following equations:

$$R_t = (r - q)\Lambda - \frac{\sigma_t^2}{2} + \sigma_t \tilde{\epsilon}_t, \quad \sigma_{t+1}^2 = \omega + \sigma_t^2(\alpha + \gamma \mathbb{1}_{\{\tilde{\epsilon}_t - \eta_t < 0\}})(\tilde{\epsilon}_t - \eta_t)^2 + \beta \sigma_t^2,$$

where $\eta_t = (\mu - (r - q)\Lambda + \sigma_t^2/2)/\sigma_t$ and $\{\tilde{\epsilon}_t\}_{t=1}^T$ are independent standard normal random variables under $\mathbb{Q}$.

The call option delta is also calculated through Monte-Carlo simulation based on the relationship

$$\Delta_t = \mathrm{e}^{-r\tau\Lambda}\mathbb{E}^{\mathbb{Q}}\left[\frac{S_{t+\tau}}{S_t}\mathbb{1}_{\{S_{t+\tau}>K\}} \mid \mathcal{F}_t\right] \tag{3.9}$$

where $\tau = T - t$ and $\mathcal{F}_t$ represents the available information at time t, i.e., that being generated by the state X_t.

In this model the state space considered for the RL approach is represented by the vector $X_t = (V_t^{\delta}, \log(S_t), \sigma_{t+1}, \tau)$.

3.4.2 Comparative analysis of deep hedging and delta hedging strategies

In this section, we study the relationship between delta hedging and deep hedging. More specifically, we investigate whether the difference between both strategies represents a speculative overlay reminiscent of a statistical arbitrage. The comparison is conducted by hedging the at-the-money (ATM) European call option, with $S_0 = K = 100$, and maturity $T = 63$ days. The leverage constraint is $B = 100$.

Table 3.1 presents the hedging performance of the deep hedging agents trained with the CVaR_α risk measure. We consider the following confidence levels α: 1%, 5%, 10%, 20%, 50%, 85%, 90%, and 95%. High values of α only put weight on the most adverse outcomes and entail focusing purely on risk reduction. Conversely, low values for α both penalize losses and reward gains, which leads to seeking risk-reward trade-offs. As such, the CVaR_α with a low confidence level does not align with the objective of limiting the variability of the hedging error. Since the CVaR_α is an increasing function of α, there are more statistical arbitrage strategies becoming available as α decreases.

Columns labeled "Base strategies" display the risk measure applied to the hedging error for the deep hedging strategy, and the difference between the risk provided by deep hedging and that of delta hedging. Columns labeled "Difference strategy" provide statistics (hedging error risk and expectation of net cash flow) of the trading strategy δ^- representing the differential position between deep hedging and delta hedging. Since such strategy reflects a long position

on the deep hedge and a short position on the delta hedge, the option payoffs from the two (long and short) positions cancel out. Hence, performance is assessed by looking at $V_T^{\delta^-}(0)$ rather than $\xi_T^{\delta^-}$.

Table 3.1: Performance assessment for deep hedging, delta hedging and their difference over a short position on an ATM call option with maturity $T = 63$ days.

	Base strategies		Difference strategy	
Metric	$\rho(\xi_T^{\delta^{DH}})$	$\rho(\xi_T^{\delta^{DH}}) - \rho(\xi_T^{\Delta})$	$\rho(-V_T^{\delta^-}(0))$	$\mathbb{E}[V_T^{\delta^-}(0)]$
$\text{CVaR}_{1\%}$	-1.550	-1.436	-1.441	1.507
$\text{CVaR}_{5\%}$	-1.484	-1.421	-1.300	1.507
$\text{CVaR}_{10\%}$	-1.398	-1.394	-1.167	1.506
$\text{CVaR}_{20\%}$	-1.203	-1.320	-0.927	1.506
$\text{CVaR}_{50\%}$	-0.221	-0.806	0.003	1.505
$\text{CVaR}_{85\%}$	1.979	-0.004	1.280	-0.031
$\text{CVaR}_{90\%}$	2.412	-0.115	1.433	-0.126
$\text{CVaR}_{95\%}$	3.481	-0.081	1.810	-0.254

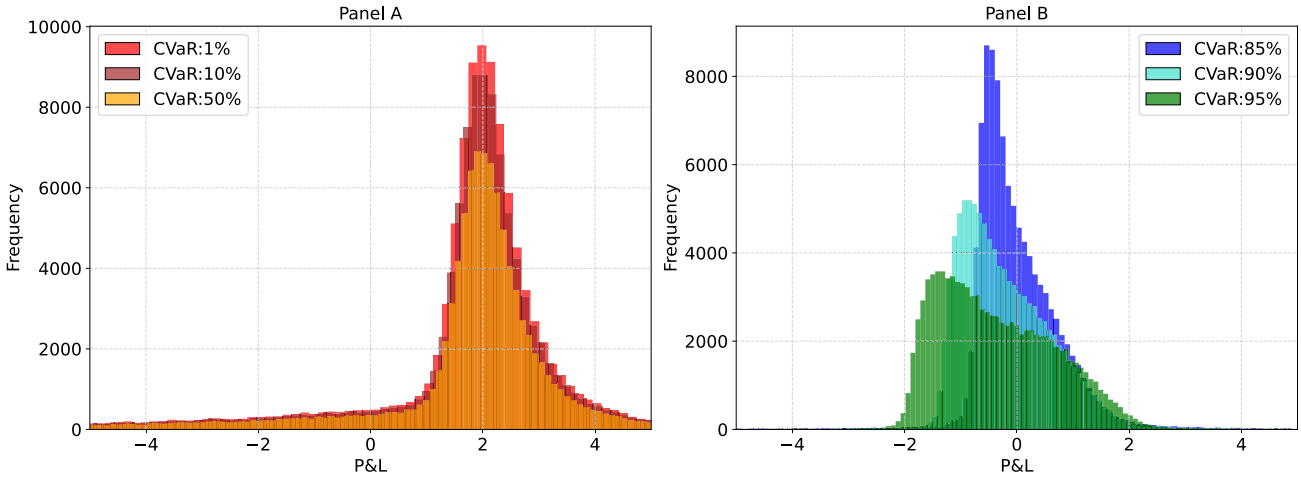
Results are computed using 100,000 out-of-sample paths. The initial price of the option is 3.16. ξ_T^{δ} is the hedging error for trading strategy δ, with δ^{DH} being deep hedging and Δ being delta hedging. The strategy δ^- uses the underlying asset position defined by the difference between that of the deep hedging and the delta hedging strategies.

For all confidence levels α below 50%, the strategy δ^- exhibits both positive average profitability $\mathbb{E}[V_T^{\delta^-}(0)]$ and a CVaR value $\rho(-V_T^{\delta^-}(0))$ that is negative. This corresponds to a formal statistical arbitrage strategy. Moreover, for the case $\alpha = 50\%$, even if the risk measure is positive, it is nevertheless negligible in comparison to expected profits. The strategy δ^-, though not a statistical arbitrage from the definition, is very close to one. Conversely, difference strategies δ^- using $\alpha \geq 85\%$ clearly do not qualify as statistical arbitrage; the associated risk measure is high and the average profitability is negative.

The distribution of the profit and loss (P&L) for the trading strategy δ^-, which is $V_T^{\delta^-}(0)$, is depicted in Figure 3.1 for various confidence levels. This confirms that difference strategies

associated with low confidence levels α (1%, 10% and 50%) are exactly or similar to statistical arbitrage with very high average profits and a very fat left tail (high extreme loss potential). The deep hedging agent is incorporating a strong speculative element in its trading strategy, which is unsuitable in practice. Conversely, the strategies associated with higher values for α do not exhibit characteristics of a statistical arbitrage and do not lead to concerns about the suitability of the deep hedging strategy.

Figure 3.1: P&L distribution of the strategy δ^{-}.



Distributions are computed using 100,000 out-of-sample paths. The P&L is simply defined by the portfolio value $V_T^{\delta^-}(0)$ at maturity.

We analyze the statistical relationship between deep hedging and delta hedging strategies through (i) sample Spearman (rank) correlations between underlying asset positions of both strategies, and (ii) the regression model

$$\delta^{DH} = \kappa_0 + \kappa_1 \Delta + \epsilon, \tag{3.10}$$

with δ^{DH} and Δ being positions produced by the deep hedging and delta hedging strategies, respectively. Metrics (regressions and correlations) are computed across all rebalancing points of all paths in the test sets.

Table 3.2 provides the Spearman correlation coefficient ϱ, which evaluates monotonic relationships between strategies, and the coefficient of determination R^2, which measures the strength of the linear association between the strategies. These metrics are computed for the various CVaR confidence levels.

Table 3.2: Statistical relationships between positions of delta hedging and deep hedging.

	Statistics	
Metric	ϱ	R^2
$\text{CVaR}_{1\%}$	-0.270	0.003
$\text{CVaR}_{5\%}$	-0.271	0.003
$\text{CVaR}_{10\%}$	-0.272	0.003
$\text{CVaR}_{20\%}$	-0.273	0.003
$\text{CVaR}_{50\%}$	-0.273	0.003
$\text{CVaR}_{85\%}$	0.939	0.773
$\text{CVaR}_{90\%}$	0.963	0.816
$\text{CVaR}_{95\%}$	0.969	0.808

Results are for a short position on the ATM call option with a maturity of $N = 63$ days. They are computed using 100,000 out-of-sample paths. The metric ϱ denotes the (unconditional) Spearman correlation between underlying asset positions of the delta hedging strategy and the deep hedging strategy across all rebalancing days while R^2 represents the R^2 statistic obtained after regressing deep hedging positions onto delta hedging positions.

Results presented in Table 3.2 show strong monotonic and linear association between deep and delta hedging positions for high confidence levels $\alpha = 85\%$, 90% or 95%. The deep hedging strategies can therefore be seen as alterations of the delta hedging procedure that improve hedging performance. Conversely, for low confidence levels (50% or below), deep hedging positions seem completely unrelated to delta hedging positions, indicating that the agent has mostly abandoned its hedging objective and is rather attempting to speculate or conduct statistical arbitrage.

It is important to note that the hedging strategy associated with $\alpha = 50\%$ exhibits a behavior that is quite similar to a statistical arbitrage even if it does not qualify as a formal statistical arbitrage. It indeed is highly speculative with large expected profits, a large left tail for the P&L and a negative correlation with delta hedging positions. This indicates that the Buehler et al. (2021) approach, which consists in using a change of measure under which statistical arbitrage strategies are removed, could be insufficient to prevent speculative behavior from the hedging agent.

3.4.3 Robustness assessment

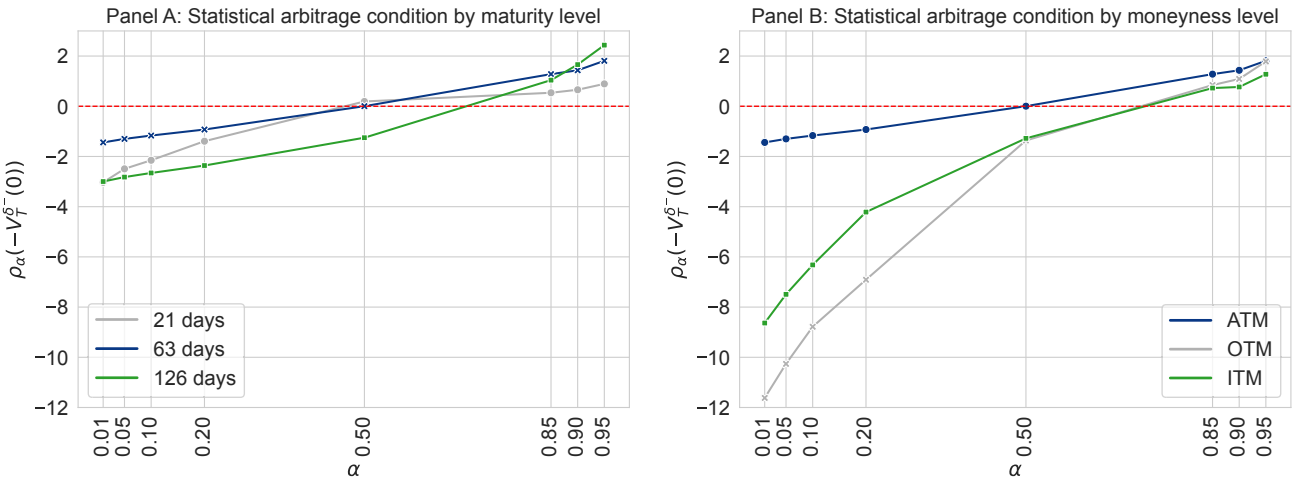
This section assesses the robustness of the above findings with respect to modifications of the baseline setup. In particular, we test the presence of statistical arbitrage over various option maturities and moneyness levels, for parameters associated with various economic periods, and for a straddle option instead of a vanilla one.

3.4.3.1 Robustness assessment across option maturities and moneyness levels

The arbitrage condition (3.5) for the trading strategy δ^- is assessed along two dimensions: the option maturity and the moneyless level. More precisely, the first dimension examines maturities of 21, 63, and 126 days, representing short-, medium-, and long-term maturities, when hedging an ATM call option. The second dimension considers out-of-the-money (OTM) and in-the-money (ITM) options with strike prices of 110 and 90, respectively, while assuming a hedged option maturity of 63 days.

Numerical results presented in Figure 3.2 demonstrate that the differential position between deep hedging and delta hedging qualifies as statistical arbitrage for small values of α, regardless of the maturity (Panel A) or the option moneyness (Panel B).

Figure 3.2: Risk for the differential strategy, $\rho(-V_T^{\delta^-}(0))$, evaluated across different maturities and moneyness levels.



Results are computed using 100,000 out-of-sample paths. Panel A presents $\rho(-V_T^{\delta^-}(0))$ for an ATM call option across different maturity levels. Panel B examines ATM, OTM, and ITM options with a 63-day maturity and strike prices of 100, 110, and 90, respectively.

The susceptibility to statistical arbitrage varies across the different option configurations. In Panel A, we observe that higher values of α are needed to eliminate statistical arbitrage for longer maturity options. This can be explained through time diversification of risk, with repeated speculative actions becoming less and less risky in aggregate as time goes by due to the law of large numbers. In Panel B, we see that the moneyness considered during the training process also influences the presence of statistical arbitrage, with the ATM option being the least susceptible to statistical arbitrage. Nevertheless, even for ITM/OTM options, the speculative component can be avoided if α is sufficiently large. For instance the trading strategy δ^- does not qualify as statistical arbitrage when $\alpha = 85\%$, 90% or 95%, regardless of the option configuration considered.

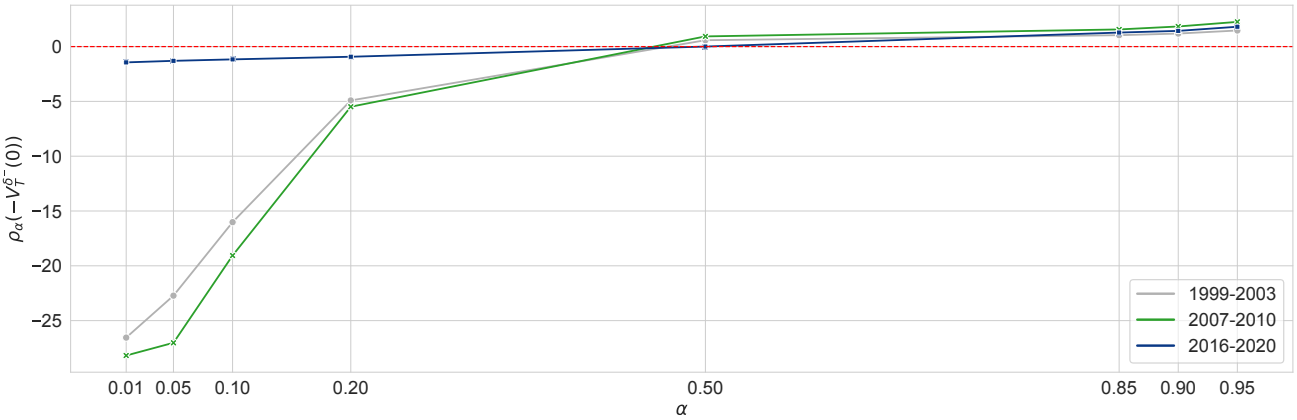
3.4.3.2 Robustness assessment across different economic conditions

As a second experiment, we examine whether our findings remain consistent under different market conditions. Specifically, we investigate whether the differential position between deep hedging and delta hedging qualifies as statistical arbitrage when simulated market dynamics

are designed to replicate various economic conditions, including periods of economic crises. In addition to the period considered in the analysis of Section 3.4.2, two more periods are included: the first, from January 4, 1999, to December 31, 2003, which includes the Dot-com bubble, and the second, from January 3, 2007, to December 31, 2010, which spans the Global Financial Crisis.

Figure 3.3 illustrates the values of $\rho(-V_T^{\delta^-}(0))$, the risk for the differential strategy δ^-, across different confidence levels and under three different economic conditions, when hedging an ATM call option with a maturity of 63 days. Our results show that RL agents may exploit this speculative component more aggressively during crisis periods at low confidence levels. However, consistently with above experiments, if the confidence level is sufficiently high the speculative component can be avoided regardless of economic conditions.

Figure 3.3: Risk for the differential strategy, $\rho(-V_T^{\delta^-}(0))$, evaluated across different market conditions.



Results are computed using 100,000 out-of-sample paths for an ATM call option with a 63-day maturity. The economic periods considered are approximated to cover three crisis-related intervals: the Dot-com bubble (1999–2003), the Global Financial Crisis (2007–2010), and the onset of the COVID-19 crisis (2016–2020)

3.4.3.3 The case of a straddle option

We extend our analysis to a different derivative instrument by examining an options portfolio consisting of a straddle strategy. The straddle is composed of a long ATM call option and a long ATM put option. Following the approach in Section 3.4.2, we evaluate whether the

trading strategy δ^- qualifies as a statistical arbitrage when used to hedge a straddle with a maturity of 63 days.

Table 3.3 presents results pertaining to the following two key aspects: the hedging performance of the deep hedging agents trained using CVaR_α, and the statistical arbitrage assessment for the differential position between the deep hedging and delta hedging strategies across various confidence levels α. Our findings align with those of Table 3.1 which are obtained by hedging a vanilla call option. Specifically, the strategy δ^- qualifies as statistical arbitrage for all confidence levels below 50%, while its behavior deviates significantly from that of a statistical arbitrage at higher confidence levels (85%, 90%, and 95%).

Table 3.3: Performance assessment for deep hedging, delta hedging and their difference over a short position on an ATM straddle strategy with maturity $T = 63$ days.

	Base strategies		Difference strategy	
Metric	$\rho(\xi_T^{\delta^{DH}})$	$\rho(\xi_T^{\delta^{DH}}) - \rho(\xi_T^{\Delta})$	$\rho(-V_T^{\delta^-}(0))$	$\mathbb{E}[V_T^{\delta^-}(0)]$
$\text{CVaR}_{1\%}$	-3.301	-2.872	-2.882	3.014
$\text{CVaR}_{5\%}$	-3.167	-2.839	-2.600	3.014
$\text{CVaR}_{10\%}$	-2.994	-2.785	-2.334	3.013
$\text{CVaR}_{20\%}$	-2.603	-2.634	-1.854	3.013
$\text{CVaR}_{50\%}$	-0.633	-1.602	0.007	3.012
$\text{CVaR}_{85\%}$	3.756	-0.016	2.559	-0.062
$\text{CVaR}_{90\%}$	4.632	-0.230	2.865	-0.253
$\text{CVaR}_{95\%}$	6.769	-0.163	3.620	-0.509

Results are computed using 100,000 out-of-sample paths. The initial price of the straddle strategy is 5.42. ξ_T^δ is the hedging error for trading strategy δ, with δ^{DH} being deep hedging and Δ being delta hedging. The strategy δ^- uses the underlying asset position defined by the difference between that of the deep hedging and the delta hedging strategies.

3.5 Conclusion

Consider the trading strategy whose underlying asset positions correspond to the difference between these of deep hedging and delta hedging. What if there exist market models under which such strategy is a statistical arbitrage? This would raise concerns about the suitability of deep hedging procedures, as it raises the possibility that typical deep hedging strategies could consist of conventional hedging strategies that are enhanced with speculative overlays which are unrelated to hedging.

Our study shows that these concerns can be mitigated; if the risk measure considered in the hedging optimization problem does not sufficiently penalize losses relative to rewards provided for gains, the deep hedging strategy attaches a statistical arbitrage strategy overlay to the delta hedging strategy. Nevertheless, when using a proper risk measure (the CVaR with sufficiently high α in our case) within the optimization problem, the difference between deep hedging and delta hedging does not exhibit statistical arbitrage-like behavior and cannot be interpreted as a speculative strategy reaping profits while exploiting blind spots of the chosen risk measure.

Moreover, our robustness assessments highlight that susceptibility to statistical arbitrage may be influenced by option characteristics and economic conditions. Longer maturities and specific moneyness levels are more vulnerable at low confidence levels, while crises amplify speculative components. However, at high confidence levels (e.g., 85%, 90%, 95%), the differential position between deep hedging and delta hedging consistently departs from a statistical arbitrage, regardless of aforementioned factors. These findings emphasize the importance of selecting a sufficiently high α to mitigate speculative behavior in deep hedging strategies.

The two main conclusions from this study are therefore that (i) the objective function of the deep hedging problem must be carefully selected to prevent the hedging agent from abandoning its hedging objective and pursuing speculative behavior, and (ii) deep hedging

can soundly achieve its hedging objectives when provided with a suitable risk measure. A possibility could be to use risk measures that do not provide any reward for gains, such as the semi-RMSE used in Carbonneau and Godin (2023). However, this would come with the cost of negatively impacting the profitability of the strategy. More research is therefore required to determine what risk measure could be used in the objective function to produce sound hedging behavior.

3.6 Appendix

3.6.1 The deep hedging algorithm

The neural network is optimized with the Mini-batch Stochastic Gradient Descent method (MSGD). This training procedure relies on updating iteratively all the trainable parameters of the network based on the recursive equation

$$\theta_{j+1} = \theta_j - \eta_j \nabla_\theta \hat{\mathcal{O}}(\theta_j), \tag{3.11}$$

where θ_j is the set of parameters obtained after iteration j, η_j is the learning rate (step size) which determines the magnitude of change in parameters on each time step, ∇_θ is the gradient operator with respect to θ and $\hat{\mathcal{O}}$ is the Monte Carlo estimate of the objective function (3.4) computed on a mini-batch. Automatic differentiation packages are used to compute the gradient of $\hat{\mathcal{O}}$. Additional details are provided in Appendix 3.6.2.

For the neural network, we employ a fully-connected Feedforward Neural Network (FFNN) architecture with four hidden layers of width 56 using a ReLU activation function. The output FFNN layer, which maps the output of the hidden layer Z into the position in the underlying asset position δ_{t+1}^{DH}, is equipped with a dynamic upper bound on the activation function to preclude excessive leverage. Indeed, agents have finite borrowing capacity in practice. We impose that the time-t cash account value ϕ_t satisfies $\phi_t \geq -B$ for all t and for

some threshold $B > 0$. This is achieved by setting the final output layer activation to

$$f(Z, t) = \min\left(Z, (V_t + B)/S_t\right), \tag{3.12}$$

which ensures that the cash amount borrowed in the portfolio remains below $B > 0$ (see François et al. (2024)).

Agents are trained on training sets of 400,000 independent simulated paths with mini-batch size of 1,000 and a learning rate of 0.0005 that is progressively adapted with the ADAM (Kingma and Ba, 2014) optimization algorithm. The training procedure is implemented in Python, using Tensorflow and considering the Glorot and Bengio (2010) random initialization of the initial parameters of the neural network.

3.6.2 Details for the MSGD training approach

The MSGD method estimates the penalty function $\mathcal{O}(\theta)$, which is typically unknown, through small samples of the hedging error called batches. Let $\mathbb{B}_j = \{\xi_{T,i}^{\delta_{\theta_j}^{DH}}\}_{i=1}^{N_{\text{batch}}}$ be the j-th batch where $\xi_{T,i}^{\delta_{\theta_j}^{DH}}$ denotes the hedging error of the i-th simulated path in the j-th batch defined as

$$\xi_{T,i}^{\delta_{\theta_j}^{DH}} = \max(S_{T,ij} - K, 0) - V_{T,i}^{\delta_{\theta_j}^{DH}}(V_0),$$

where $S_{T,ij}$ is the price of the underlying asset at time T in the i-th simulated path, and $V_{T,i}^{\delta_{\theta_j}^{DH}}$ is the terminal value of the hedging strategy for that path when $\theta = \theta_j$. The penalty function estimation for the batch $\mathbb{B}$ is

$$\hat{C}^{(\text{CVaR})}(\theta_j, \mathbb{B}_j) = \widehat{\text{VaR}}_\alpha(\mathbb{B}_j) + \frac{1}{(1-\alpha)N_{\text{batch}}} \sum_{i=1}^{N_{\text{batch}}} \max\left(\xi_{T,i}^{\delta_{\theta_j}^{DH}} - \widehat{\text{VaR}}_\alpha(\mathbb{B}_j), 0\right),$$

where $\widehat{\text{VaR}}_\alpha(\mathbb{B}_j) = \xi_{T,\lceil \alpha \cdot N_{\text{batch}} \rceil}^{\delta_{\theta_j}^{DH}}$ is the estimation of the VaR obtained from the ordered sample $\{\xi_{T,[i]}^{\delta_{\theta_j}^{DH}}\}_{i=1}^{N_{\text{batch}}}$ and $\lceil \cdot \rceil$ is the ceiling function. These empirical approximations are used to

estimate the gradient of the penalty function required in Equation (3.11).[18]

[18]Details about gradient of the empirical objective function are provided in Goodfellow et al. (2016).

Chapter 4

Implied-Volatility-Surface-Informed Deep Hedging with Options

Abstract

We propose an enhanced deep hedging framework to hedge portfolios of options, which integrates implied volatility surface-informed decisions with multiple hedging instruments. In presence of transaction fees, a state-dependent no-trade region provides an optimal rebalancing frequency and improve hedging performance. By leveraging information from the evolving implied volatility surfaces, our approach consistently outperforms traditional delta and delta gamma hedging approaches across diverse market conditions from 1996 to 2020. The inclusion of no-trade regions drives optimal practitioner delta gamma solutions towards minimal rebalancing frequencies, similar to static hedging. In contrast, deep hedging strategies show superior adaptability, delivering enhanced performance in both simulated environments and backtesting.

JEL classification: C45, C61, G32.

Keywords: Deep reinforcement learning, optimal hedging, implied volatility surfaces.

4.1 Introduction

This paper presents a reinforcement learning-based hedging framework that enhances deep hedging with market-implied expectations. By incorporating the joint dynamics of implied volatility and asset returns within a data-driven market simulator, we optimize hedging decisions using multiple instruments. This approach extends the deep hedging paradigm by integrating forward-looking information from the implied volatility surface, improving adaptability to market fluctuations. This framework provides a scalable and dynamic solution to hedge portfolios of options more effectively.[19]

A commonly used approach for hedging portfolios of derivatives is delta-gamma hedging, which adjusts portfolio positions based on sensitivities to movements in the underlying asset's price. While effective in mitigating risk, this strategy often incurs significant transaction costs, particularly when multiple hedging instruments, such as options, are involved. These costs make it challenging to implement hedging strategies that are both risk-efficient and cost-effective.

Methodologies such as those proposed by Coleman et al. (2007) and Kélani and Quittard-Pinon (2017) address these challenges by minimizing local risk while incorporating standard options as hedging instruments. The latter approach further accounts for transaction costs. While these approaches provide values insights, the adequate incorporation of market expectations remains unexplored. Additionally, complementing these frameworks, deep hedging, introduced by Buehler et al. (2019), offers a data-driven alternative that leverages deep reinforcement learning (DRL) to dynamically adapt to evolving market conditions, including shifting expectations and historical market patterns. Despite its demonstrated flexibility and adaptability (e.g., Cao et al. (2020), Carbonneau (2021), and Cao et al. (2023)), integrating forward-looking information into the deep hedging framework remains an open area of research.

Recent developments in deep hedging methodologies offer a promising avenue to address these

[19] In this paper, we consider a basket of European options written on the same underlying asset.

limitations. For instance, François et al. (2024) makes a first attempt in this direction, showing how deep hedging strategies can mitigate transaction costs across varying market conditions while effectively incorporating insights from implied volatility (IV) surface dynamics. While this research focused on hedging vanilla options using instruments such as the risk-free rate and the underlying asset, the potential benefits of expanding the hedging set with additional instruments and IV-informed policies remain largely unexplored.

In this paper, we propose a dynamic hedging framework to optimize hedging decisions while minimizing the risk metric associated with terminal hedging error. Our framework extends the deep hedging paradigm to accommodate a broader class of derivative instruments by dynamically trading both the underlying asset and a liquid derivative instrument. Following François et al. (2024), our hedging decisions are informed by factors driving the dynamics of the IV surface, enhancing the model's adaptability to market conditions and improving hedging efficiency, both in the presence and absence of transaction fees..

To optimize rebalancing decisions, we incorporate no-trade regions, a concept widely studied in portfolio optimization under transaction costs. For instance, Constantinides (1986) introduces the idea that proportional transaction costs create regions where rebalancing is suboptimal, further developed by Davis and Norman (1990) and Balduzzi and Lynch (1999), who focus on portfolio allocation rather than rebalancing costs. These regions indicate when rebalancing becomes cost-effective, balancing transaction costs with desired adjustments. In hedging, optimal rebalancing based on delta changes has been explored by Henrotte (1993), Toft (1996), and Martellini (2000), while Hodges and Neuberger (1989) examines no-trade bands around delta, showing that exact replication is often infeasible or costly. They demonstrate that higher transaction costs or risk aversion require tighter rebalancing regions. Building on these ideas, our framework defines the no-trade region as a state-dependent subset, using a distance measure in the portfolio allocation space to determine when rebalancing is optimal relative to the objective function.

To ensure our strategies remain optimal and aligned with sound risk management, we control for speculative behaviors in our hedging strategies. François et al. (2024) show that RL strategies can introduce doubling behaviors, where agents increase exposure to recover from losses. To mitigate this, our framework incorporates a soft tracking error constraint, aligning hedging decisions with risk management goals. Additionally, Buehler et al. (2021) and Horikawa and Nakagawa (2024) highlight that deep hedging can unintentionally include statistical arbitrage overlays. We test our approach to ensure that any performance improvements are not driven by by speculative-like components.

We utilize the JIVR model, introduced by François et al. (2023), to capture the joint dynamics of S&P 500 log-returns and implied volatility surfaces, which inform hedging strategy decisions. These decisions are optimized by minimizing a risk measure applied to terminal hedging errors. As an illustrative example, we hedge a straddle portfolio using the risk-free instrument, the underlying asset, and a vanilla European call options as hedging instruments.

Through statistical analysis and sensitivity assessments, we investigate the relationship between key state variables and hedge ratios, further validating the robustness of our methodology. The performance of our framework is assessed in both a simulated environment and through backtesting. Both evaluations demonstrate the superiority of out approach in comparison to traditional delta and delta gamma hedging. Our deep reinforcement learning framework consistently outperforms these methods across a range of market conditions, providing a more resilient, cost-effective, and practical solution for risk management in options trading.

The paper is organized as follows. Section 4.2 frames the hedging problem in terms of a deep reinforcement learning framework. Section 4.3 provides the components of the market simulator. Section 4.4 presents the numerical results, assessments, and global feature importance analysis. Section 4.5 concludes.

4.2 The hedging problem

In this section, we present the mathematical formulation of the hedging problem, along with the computational scheme to obtain the numerical solution.

4.2.1 The hedging optimization problem

We propose dynamic hedging strategies for managing portfolios of options. Our approach focuses on minimizing a risk measure applied to terminal hedging error while considering variable market conditions and accounting for transaction costs.

The goal is to hedge a short position of a portfolio of contingent claims written on the same underlying asset, S_t, over the hedging period $0, \ldots, T$. The time-t market portfolio value is expressed as $\mathcal{P}_t = \Psi_t(S_t)$ for some function Ψ_t. For illustrative purposes, we use a straddle portfolio with maturity T in our numerical examples. In this case, the value $\mathcal{P}_T$ represents the portfolio's payoff, which is given by the mapping $\Psi_T(S_T) = \max(S_T - K, 0) + \max(K - S_T, 0)$ with K the strike price.

The hedging strategy involves managing a self-financing portfolio composed of the risk-free asset, the underlying asset, and a hedging option. Specifically, the hedging option is a European option on the same underlying asset with a longer maturity $T^* > T$. The strategy is represented by the predictable process $\{\phi_t\}_{t=1}^{T}$, with $\phi_t = (\phi_t^{(r)}, \phi_t^{(S)}, \phi_t^{(\mathrm{O})})$ where $\phi_t^{(r)}$ is the cash held at time $t-1$ and carried forward to the next period, $\phi_t^{(S)}$ denotes the number of shares of the risky asset S and $\phi_t^{(\mathrm{O})}$ the number of shares of the hedging option, both held during the interval $(t-1, t]$. The time-t hedging portfolio value is

$$V_t^{\phi} = \phi_t^{(r)} \mathrm{e}^{r_t \Delta} + \phi_t^{(S)} S_t \mathrm{e}^{q_t \Delta} + \phi_t^{(\mathrm{O})} \mathrm{O}_t(T^*)$$

where $\mathrm{O}_t(T^*)$ is the time-t hedging option value, $\Delta = \frac{1}{252}$ represents the time increment in years, r_t is the time-t annualized continuously compounded risk-free rate and q_t is the annualized underlying asset dividend yield, both on the interval $(t-1, t]$. To account for

transaction costs the self-financing condition entails that for $t = 0, \ldots, T-1$,

$$\phi_{t+1}^{(r)} + \phi_{t+1}^{(S)} S_t + \phi_{t+1}^{(C)} \mathrm{O}_t(T^*) = V_t^{\phi} - \kappa_1 S_t \mid \phi_{t+1}^{(S)} - \phi_t^{(S)} \mid -\kappa_2 \mathrm{O}_t(T^*) \mid \phi_{t+1}^{(\mathrm{O})} - \phi_t^{(\mathrm{O})} \mid, \quad (4.1)$$

where κ_1 and κ_2 represent the proportional transaction cost rates for the underlying asset and the hedging option, respectively. In a practical financial context, transaction costs for options are typically higher than those for the underlying asset, consequently, we assume $\kappa_1 << \kappa_2$.

The optimal sequence of actions $\phi = \{\phi_t\}_{t=1}^{T}$ corresponds to those that minimize the application of a risk measure ρ to the hedging error at maturity for a short position in the option portfolio:

$$\xi_T^{\phi} = \mathcal{P}_T - V_T^{\phi}.$$

A positive value in ξ_T^{ϕ} implies that the hedging strategy does not have enough funds to cover the portfolio value $\mathcal{P}_T$. Therefore the hedging problem is

$$\phi^* = \arg\min_{\phi} \left\{ \rho\left(\xi_T^{\phi}\right) \right\}. \quad (4.2)$$

Each time-t action ϕ_{t+1} is a function of current available information on the market: $\phi_{t+1} = \tilde{\phi}(X_t)$ for some function $\tilde{\phi}$ with state variables vector X_t. Due to Equation (4.1), $\phi_{t+1}^{(r)}$ is fully determined when $\phi_{t+1}^{(S)}$ and $\phi_{t+1}^{(\mathrm{O})}$ are specified, and as such the time-t action to be chosen is $(\phi_{t+1}^{(S)}, \phi_{t+1}^{(\mathrm{O})})$.

This paper examines three widely recognized risk measures in the literature:

- Mean Square Error (MSE): $\rho\left(\xi_T^{\phi}\right) = \mathbb{E}\left[\left(\xi_T^{\phi}\right)^2\right]$.
- Semi Mean-Square Error (SMSE): $\rho\left(\xi_T^{\phi}\right) = \mathbb{E}\left[\left(\xi_T^{\phi}\right)^2 \mathbb{1}_{\{\xi_T^{\phi} \geq 0\}}\right]$.
- Conditional Value-at-Risk (CVaR_α): $\rho\left(\xi_T^{\phi}\right) = \mathbb{E}\left[\xi_T^{\phi} \middle| \xi_T^{\phi} \geq \mathrm{VaR}_\alpha\left(\xi_T^{\phi}\right)\right]$, where $\mathrm{VaR}_\alpha\left(\xi_T^{\phi}\right)$ is the Value-at-Risk defined as $\mathrm{VaR}_\alpha\left(\xi_T^{\phi}\right) = \min_c \left\{ c : \mathbb{P}\left(\xi_T^{\phi} \leq c\right) \geq \alpha \right\}$, and $\alpha \in (0, 1)$.

4.2.2 Reinforcement learning and deep hedging

The problem described in Equation (4.2) is addressed by directly estimating the policy function (the investment strategy $\tilde{\phi}$) using a policy gradient method. This approach leverages a parametric representation of the policy function through an Artificial Neural Network (ANN). Specifically, a parameter vector θ is introduced to define the policy $\tilde{\phi}$, which is optimized to minimize the risk measure ρ applied to the hedging error at maturity. Representing the policy generated by the ANN as $\tilde{\phi}_\theta$, the hedging strategy defined as $\phi_{t+1} = \tilde{\phi}_\theta(X_t)$. The approximate optimization problem considered is therefore

$$\arg\min_\theta \left\{ \rho\left(\xi_T^{\tilde{\phi}_\theta}\right) \right\}.$$

Given the inherent continuity of ANNs, the mapping $\phi_{t+1} = \tilde{\phi}_\theta(X_t)$ may lead to frequent small adjustments in the hedging position, potentially increasing long-term transaction costs. To mitigate this effect, we introduce a no-trade region, within which there is no rebalancing.

At time t, the no-trade region is determined by the distance between the current portfolio position, ϕ_t, and the next position proposed by the ANN, $\tilde{\phi}_\theta(X_t)$. Specifically, rebalancing occurs only if the cumulative deviation in positions across hedging instruments exceeds a threshold l:

$$\phi_{t+1} = \begin{cases} \phi_t, & \text{if } |\phi_t^{(S)} - \tilde{\phi}_\theta^{(S)}(X_t)| + |\phi_t^{(\mathrm{O})} - \tilde{\phi}_\theta^{(\mathrm{O})}(X_t)| \leq l, \\ \tilde{\phi}_\theta(X_t), & \text{otherwise.} \end{cases} \tag{4.3}$$

This formulation expresses the no-trade region in terms of the number of shares, providing a measure of the distance at which rebalancing becomes cost-effective, capturing the trade-off between transaction costs and maintaining proximity to the desired portfolio adjustments. In this framework, both the ANN parameters θ and the rebalancing threshold l are treated as learnable parameters, allowing the model to jointly optimize the size of rebalancing actions

and decisions of whether or not to rebalance.

As shown in François et al. (2024), the policy $\tilde{\phi}_\theta$ may inadvertently incorporate speculative elements, such as doubling strategies, where agents continuously increase their exposure in an attempt to recover successive losses. Such strategies are undesirable as they deviate from sound risk management principles. To reduce the likelihood of encountering this problem, we introduce a soft tracking error constraint that penalizes the network during training if the time-t tracking error, $\xi_t^{(\tilde{\phi}_\theta, l)} = \mathcal{P}_t - V_t^{(\tilde{\phi}_\theta, l)}$, exceeds the initial hedging portfolio value at any time t. This constraint is defined as:

$$SC(\theta,\, l) = \mathbb{P}\left(\max_{t\in\{0,\ldots,T\}} \left[P_t - V_t^{(\tilde{\phi}_\theta, l)}\right] > V_0\right). \tag{4.4}$$

This design leaves gains unpenalized, consistent with the asymmetric nature of rational agents. As a result, the penalty function employed in our approach is

$$\mathcal{O}_\lambda(\theta,\, l) = \rho\left(\xi_T^{(\tilde{\phi}_\theta, l)}\right) + \lambda \cdot SC(\theta,\, l), \tag{4.5}$$

where λ is a penalty parameter that controls the weight of the soft constraint in the optimization process. Its optimal value is determined independently using a validation set as part of the model selection procedure.

4.2.2.1 Neural network architecture

We employ a Recurrent Neural Network with a Feedforward Connection (RNN-FNN), integrating Long Short-Term Memory (LSTM) networks with Feedforward Neural Network (FFNN) architectures. This hybrid design has demonstrated superior training performance compared to conventional ANN architectures, as shown in Fecamp et al. (2020) and François et al. (2024). The RNN-FNN network is defined as a composition of LSTM cells $\{C_l\}_{l=1}^{L_1}$ and

FFNN layers $\{\mathcal{L}_j\}_{j=1}^{L_2}$ under the following functional representation:

$$\tilde{\phi}_\theta(X_t) = (\underbrace{\mathcal{L}_J \circ \mathcal{L}_{L_2} \circ \mathcal{L}_{L_2-1} \circ ... \circ \mathcal{L}_1}_{\text{FFNN layers}} \circ \underbrace{C_{L_1} \circ C_{L_1-1} ... \circ C_1}_{\text{LSTM cells}})(X_t).$$

The explicit formulas for this ANN are detailed in Appendix 4.6.1.1.

4.2.3 Neural network optimization

The RNN-FNN network $\tilde{\phi}_\theta(\cdot)$, along with the rebalancing threshold l, are optimized with the Mini-batch Stochastic Gradient Descent method (MSGD). This training procedure relies on updating iteratively all the trainable parameters of the optimization problem based on the recursive equations

$$\theta_{j+1} = \theta_j - \eta_j^{(1)} \frac{\partial}{\partial \theta} \hat{\mathcal{O}}_\lambda(\theta, l), \tag{4.6}$$

$$l_{j+1} = l_j - \eta_j^{(2)} \frac{\partial}{\partial l} \hat{\mathcal{O}}_\lambda(\theta, l), \tag{4.7}$$

where $\eta_j^{(1)}$ and $\eta_j^{(2)}$ are the learning rates that determine the magnitude of the change of the parameters per time-step, these rates are dynamically adjusted using the Adam optimization algorithm.[20] Additionally, $\hat{\mathcal{O}}(\theta, l)$ is the Monte-Carlo estimate of the penalty function defined at Equation (4.5). Further details can be found in Appendix 4.6.1.2.

4.3 Market simulator

Our approach incorporates a market simulator to replicate the joint dynamics of the S&P 500 price and its associated IV dynamics. Indeed, optimal actions are characterized by the behavior of the underlying asset and hedging instrument prices. Using a simulator provides the advantage of generating a large diversity of scenarios, enabling RL agents to explore the state space while identifying optimal policies. This alleviates the issue of scarcity in real market data.

[20] Adam is an adaptive learning rate method designed to accelerate training in deep neural networks and promote rapid convergence, as detailed in Kingma and Ba (2015).

The optimization of hedging strategies in our framework requires specifying the joint dynamics of the underlying asset and of options on such asset. As such, we leverage the JIVR model from François et al. (2023), which models the temporal dynamics of S&P 500 returns and various factors driving the IV surface, along with their interdependencies. The JIVR has the advantage of being data-driven, allowing to replicate multiple realistic shapes of the IV surface encountered in practice. It has been calibrated on a extensive data sample including multiple crises; it can therefore reflect a broad array of economic conditions. Finally, the multi-factor nature of the model leads to a flexible relationship between the underlying asset price and volatility surfaces. Such feature allows reflecting self-contained properties of the option market, consistently with the "instrumental approach" of option pricing detailed in Rebonato (2005). This section describes the JIVR model providing joint dynamics of the S&P 500 index price and its associated IV surface.

4.3.1 Daily implied volatility surface

The time-t IV of an option with time-to-maturity $\tau_t = \frac{T-t}{252}$ years and moneyness $M_t = \frac{1}{\sqrt{\tau_t}} \log \frac{S_t e^{(r_t - q_t)\tau_t}}{K}$ is given by:

$$\sigma(M_t, \tau_t, \beta_t) = \sum_{i=1}^{5} \beta_{t,i} f_i(M_t, \tau_t). \tag{4.8}$$

The vector $\beta_t = (\beta_{t,1}, \beta_{t,2}, \beta_{t,3}, \beta_{t,4}, \beta_{t,5})$ represents the IV factor coefficients at time t, while the functions $\{f_i\}_{i=1}^5$ capture the effects of the long-term at-the-money (ATM) level, time-to-maturity slope, moneyness slope, smile attenuation, and smirk, respectively. A detailed description of the functional components of the IV surface, $\{f_i\}_{i=1}^5$, can be found in Appendix 4.6.2.1.

4.3.2 Joint Implied Volatility and Return

The JIVR model introduced by François et al. (2023) builds upon the IV representation in Equation (4.8), offering an explicit formulation for the joint dynamics of the IV surface and

the S&P 500 price. More precisely, this joint representation is based on an econometric model for (i) the underlying asset returns, and (ii) fluctuations of the IV surface coefficients β_t along a mean-reversion component for their volatility h_t. The multivariate time series formulation of the JIVR model is provided in detail in Appendix 4.6.2.2.

The JIVR model is used in subsequent simulation experiments to generate paths of the state variables $(S_t, \{\beta_{t,i}\}_{i=1}^5, h_{t,R}, \{h_{t,i}\}_{i=1}^5)$, which drive the market dynamics, where $h_{t,R}$ and $\{h_{t,i}\}_{i=1}^5$ are volatilities for the S&P 500 and each of the IV factors. Estimates of model parameters and volatility series $\{\hat{h}_{t,i}\}_{t=1}^N$ with $i \in \{1, \ldots, 5, R\}$ are taken from François et al. (2023), who apply maximum likelihood on a multivariate time series made of S&P 500 returns and surface coefficients estimates $\{\hat{\beta}_t\}_{t=1}^N$, with sample dates extending between January 4, 1996 and December 31, 2020.

4.4 Numerical study

4.4.1 Market settings for numerical experiments

We model a discrete-time financial market with daily trading opportunities over a time horizon of T days. The initial conditions of the JIVR model, $(\{\beta_{0,i}\}_{i=1}^5, h_{0,R}, \{h_{0,i}\}_{i=1}^5)$, are randomly sampled from the estimated values in our data set, covering the period from January 4, 1996, to December 31, 2020. Across all experiments, the annualized continuously compounded risk-free rate and dividend yield are assumed to remain constant, with values fixed at $r = 2.66\%$ and $q = 1.77\%$, respectively.[21]

The initial value of the underlying asset is set to $S_0 = 100$ for simplicity. The hedged portfolio is an ATM straddle with a maturity of $T = 63$ days. At any time $t < T$, the portfolio value $\mathcal{P}_t$ is determined using the IV surface prevailing at that moment. At maturity, at time $t = T$, $\mathcal{P}_T$ represents the final portfolio cash flow.

We assume the use of the risk-free asset, the underlying asset, and an ATM European call

[21]The annualized average rates of the S&P 500 dividend yield (1.77%) and the zero-coupon yield (2.66%) are calculated using OptionMetrics data over the sample period from January 4, 1996, to December 31, 2020.

option with a maturity longer than that of the straddle, specifically an ATM call option with a maturity of $T^* = 84$ days, as the hedging instruments. The time-to-maturity of the hedging option decreases over time and is not reset to 84 days on each rebalancing day. The positions in both the underlying asset and the hedging option are rebalanced daily.

The hedge follows the self-financing dynamics described in Equation (4.1), incorporating different levels of proportional transaction costs for the hedging option. The proportional transaction cost for call options on the S&P 500 index has an average value of 0.95%, as reported in the study of Chaudhury (2019). To assess the impact of transaction costs on rebalancing the hedging option, we consider several values we consider several values around the same range, specifically $\kappa_2 \in \{0.5\%, 1\%, 1.5\%, 2\%\}$. In contrast, the transaction cost for the underlying asset is almost negligible, with values around 0.047%, according to Bazzana and Collini (2020). For illustrative purposes, we set κ_1 to 0.05%. The initial value of the hedging portfolio is set equal to the price of the straddle, *i.e*, $V_0 = P_0$.

4.4.2 Benchmarks

We benchmark the performance of our framework against several established approaches: (i) the RL method proposed by François et al. (2024), which incorporates IV-informed decisions using only the underlying asset as a hedging instrument, (ii) delta hedging (D), where only the underlying asset is used for hedging, and (iii) delta gamma (DG) hedging, which includes one additional hedging option in the portfolio.

For the second and third benchmarks, the delta and gamma of financial instruments are computed using the *practitioner's* approach. This involves inserting the IV for each instrument into the closed-form expressions for Black-Scholes' delta and gamma. In the case of delta hedging, the delta is adjusted based on the correction introduced by Leland (1985), which accounts for the impact of proportional transaction costs on the underlying asset position. This adjusted delta reverts to the standard Black-Scholes delta when no transaction costs are applied. In both benchmarks, the volatility parameter is updated daily according to the

prevailing IV surface, which aligns the hedging strategies with dynamic market conditions. The explicit formulas for these two benchmarks are provided in Appendix 4.6.3.

For all three strategies, we further enhance the performance by incorporating the no-trade region, as defined in Equation (4.3), to ensure a fair and consistent comparison.[22] Additionally, the inclusion of the rebalancing threshold l improves the performance of each strategy, with l optimized based on the risk measure used to benchmark our framework (further details are provided subsequently).

4.4.3 Neural network settings

4.4.3.1 Neural network architecture

We consider a RNN-FNN architecture with two LSTM cells ($L_1 = 2$) of width 56 ($d_i = 56$ for $i = 1, 2$), two FFNN-hidden layers ($L_2 = 2$) of width 56 with ReLU activation function (i.e. $g_{\mathcal{L}_i}(X) = \max(0, X)$ for $i = 1, 2$),[23] and one two-dimensional output FFNN layer with a linear activation function. Numerical experiments suggest that $\lambda = 1$ is a relevant choice. A detailed description of the experimental procedure can be found in Appendix 4.6.4.

Agents are trained as described in Section 2.2.2.2 on a training set of 400,000 independent simulated paths with mini-batch size of 1000 and a learning rate of 0.0005. In addition, we include dropout regularization method with parameter $p = 0.5$ as in François et al. (2024). The training procedure is implemented in Python, using Tensorflow and considering the Glorot and Bengio (2010) random initialization of the initial parameters of the neural network. Numerical results are obtained from a test set of 100,000 independent paths.

4.4.3.2 State space

The state space considered in our RL framework includes the state variables generated by the JIVR model, along with a new set of state variables associated with the straddle and

[22] The optimization process is carried out as detailed in Section 2.2.2.2, following Equation (4.7), using Mini-batch Stochastic Gradient Descent.

[23] The rectified linear unit or ReLU function is commonly used in deep learning to reduce the probability of gradient vanishing and introduce sparsity into the inference process.

hedging portfolio. These variables are detailed in Table 4.1.

Table 4.1: State variables

Notation	Description
S_t	Underlying asset price
τ_t	Time-to-maturity of the straddle
$\{\beta_{t,i}\}_{i=1}^5$	IV factors described in Section 4.3.1
$\{h_{t,i}\}_{i=1}^5$	IV coefficients' variances
$h_{t,R}$	Conditional underlying asset return variance
$\mathcal{P}_t$	Straddle value
$\Delta_t^{\mathcal{P}}$	Delta of the straddle
$\gamma_t^{\mathcal{P}}$	Gamma of the straddle
O_t	Hedging option price
$V_t^{(\tilde{\phi}_\theta, l)}$	Hedging portfolio value
$\phi_t^{(S)}$	Underlying asset position
$\phi_t^{(\mathrm{O})}$	Hedging option position

In our illustrative example, the straddle serves as a contract-specific reinforcement learning task, as defined in Peng et al. (2024), where the optimization problem is solved for a specific option with given contract parameters. While the state variables associated with the target portfolio ($\mathcal{P}_t$, Δ_t^P, and γ_t^P) are not required in this setting, our numerical experiments demonstrate that in practice their inclusion enhances performance across all risk measures (details in Appendix 4.6.5). This improvement is likely due to the suboptimal convergence of ANNs in finite settings. Furthermore, incorporating these state variables extends our framework to a more general contract-unified reinforcement learning task, allowing for the optimization of portfolios with any combination of options and contract parameters.

4.4.4 Benchmarking of hedging strategies

4.4.4.1 Benchmarking in the absence of transaction costs

We begin by evaluating the hedging performance of both benchmark methods and RL agents trained using three different risk measures: MSE, SMSE, and $\mathrm{CVaR}_{95\%}$. This evaluation considers the estimated values of each risk measure alongside the sample average of the

hedging error, mean $\left(\xi_T^{(\tilde{\phi}_\theta, l)}\right) = \frac{1}{N}\sum_{i=1}^{N} \xi_{T,i}^{(\tilde{\phi}_\theta, l)}$, where $\xi_{T,i}^{(\tilde{\phi}_\theta, l)}$ represents the i-th terminal hedging error in the test set of size N. Additionally, we incorporate the sample standard deviation of the terminal hedging error, std $\left(\xi_T^{(\tilde{\phi}_\theta, l)}\right)$, as a metric to quantify the variability of hedging errors within the test set. Our analysis is conducted under the assumption of zero transaction costs, i.e., $\kappa_1 = \kappa_2 = 0$.

Table 4.2 presents the optimal values of the risk measures for the various hedging strategies in two cases. In the first case, the hedging instruments are limited to the risk-free asset and the underlying asset (columns labeled as S_t). In the second scenario, an ATM call option is introduced as an additional hedging instrument (columns labeled as $S_t + \mathrm{O}_t$). The columns under RL correspond to different risk measures used as objective functions during training, while each row represents the performance metric computed from test set hedging errors. In both cases, RL strategies consistently outperform the benchmarks and achieve the optimal values when the performance assessment metric matches the risk measure used during training.

Table 4.2: Hedging performance metrics under the assumption of zero transaction costs.

Instruments	S_t				S_t+O_t			
Strategy	D	RL			DG	RL		
		MSE	SMSE	$\mathrm{CVaR}_{95\%}$		MSE	SMSE	$\mathrm{CVaR}_{95\%}$
mean $\left(\xi_T^{(\tilde{\phi}_\theta, l)}\right)$	**-0.713**	-0.543	-0.656	-0.681	-0.069	-0.035	-0.089	-0.087
std $\left(\xi_T^{(\tilde{\phi}_\theta, l)}\right)$	1.756	1.392	1.526	1.702	0.811	**0.324**	0.325	0.326
MSE	3.593	2.232	2.760	3.362	0.663	**0.106**	0.114	0.114
SMSE	1.193	0.546	0.424	0.596	0.338	0.038	**0.025**	0.027
$\mathrm{CVaR}_{95\%}$	3.606	2.549	2.208	2.031	1.927	0.648	0.516	**0.514**

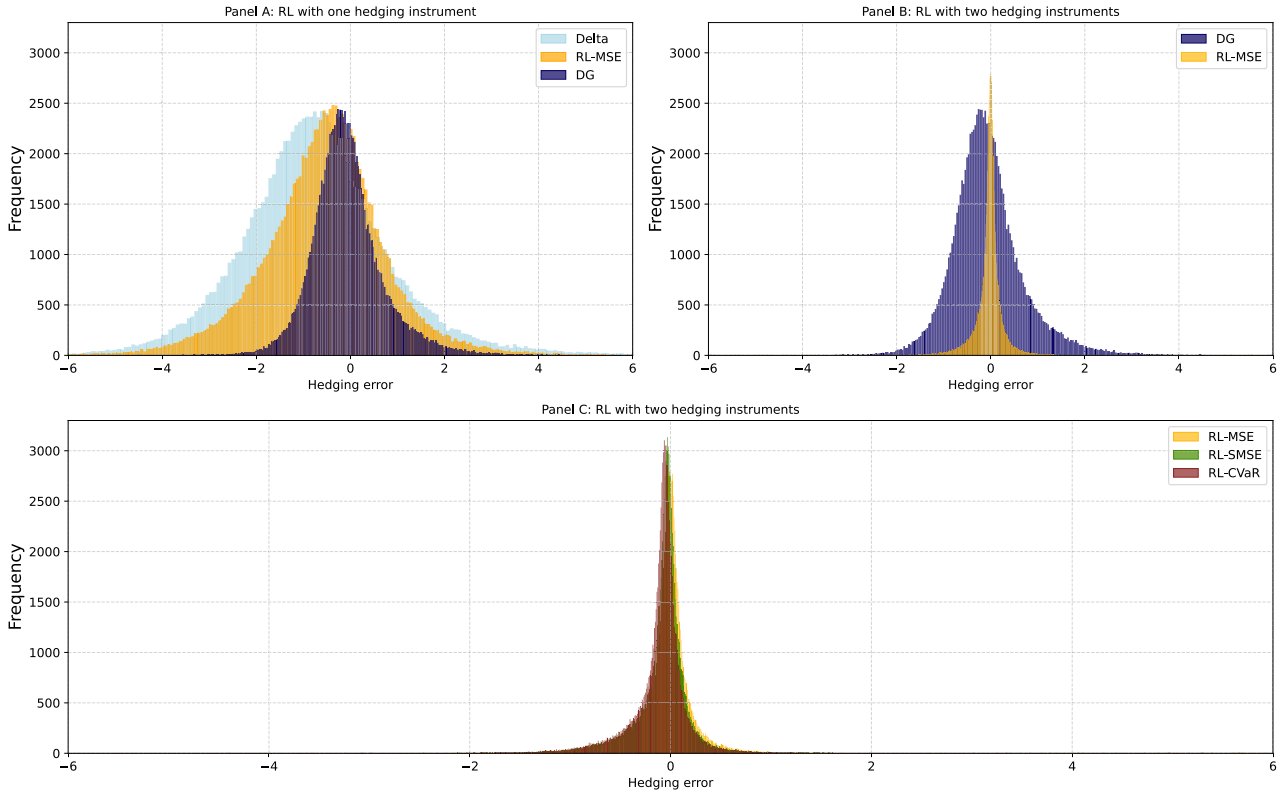
Results are computed using 100,000 out-of-sample paths in the absence of transaction costs ($\kappa_1 = \kappa_2 = 0$). Agents are trained according to the conditions outlined in Section 4.4.3. D stands for delta hedging, whereas DG refers to delta gamma hedging. The average initial straddle value is \$7.55.

Moreover, our numerical results highlight the benefits of incorporating a second hedging instrument. Specifically, all strategies that include an option as an additional hedging instrument exhibit lower risk by standard deviation, MSE, SMSE, and $\text{CVaR}_{95\%}$, compared to those relying solely on a single hedging instrument, including RL-based strategies. Notably, the delta gamma hedging strategy achieves a reduction in standard deviation of at least 42%, calculated by comparing the minimum standard deviation among single-instrument strategies, 1.392, with that of the DG strategy, 0.811. Similarly, delta gamma hedging results in a 70% reduction in MSE, a 20% decrease in SMSE, and a 5% reduction in CVaR compared to the lowest values of these performance metrics across all strategies under the column S_t.

Furthermore, RL agents utilizing multiple hedging instruments yield significantly less risky strategies than the DG strategy, as evidenced by a reduction in standard deviation of at least 60%, calculated by comparing the maximum standard deviation among RL strategies with two hedging instruments, 0.326, to that of the DG strategy, 0.811. Likewise, this advantage is further supported by other performance metrics, such as $\text{CVaR}_{95\%}$ and SMSE, which show reductions of at least 92% and 73%, respectively.

Figure 4.1 illustrates the distribution of hedging errors for the various strategies. Panel A compares the hedging error distributions of the benchmark and RL agents, each using only the underlying asset as hedging instruments, against the traditional DG hedging strategy. In this case, the inclusion of an option in the hedging strategy helps reduce the loss distribution tail. Panel B compares the DG strategy with the RL-MSE strategy, both incorporating two hedging instruments, highlighting the superior performance of the RL approach, as it significantly reduces the variance, consistent with the results shown in Table 4.2. Finally, Panel C presents a comparison of the three RL agents, demonstrating that the distribution achieved with asymmetric risk measures exhibits greater skewness.

Figure 4.1: Hedging error distribution in the absence of transaction costs.



Results are computed using 100,000 out-of-sample paths according to the conditions outlined. The initial average straddle value is $7.55.

4.4.4.2 Benchmarking in the presence of transaction costs

In this analysis, we incorporate the no-trade region, as defined by Equation (4.3), to determine the optimal rebalancing frequency while accounting for transaction costs. The rebalancing threshold l for the benchmarks is estimated using Equation (4.7) on the training set, considering the possible combinations of risk measures and transaction cost levels as independent optimization problems. For the RL strategies, this parameter is jointly estimated alongside the other ANN parameters during the training process. Table 4.3 presents the optimal values of the rebalancing threshold l across different transaction cost levels for both the DG and RL strategies, considering all risk measures.

Table 4.3: Optimal rebalancing threshold l values of DG and RL strategies.

Risk measure		MSE		SMSE		$\text{CVaR}_{95\%}$	
κ_1	κ_2	DG_l	RL_l	DG_l	RL_l	DG_l	RL_l
0%	0%	0.0	0.0	0.0	0.0	0.0	0.0
0.05%	0.5%	0.904	0.013	1.777	0.019	2.011	0.018
0.05%	1%	1.107	0.017	2.425	0.023	2.520	0.026
0.05%	1.5%	1.205	0.032	2.502	0.032	2.671	0.033
0.05%	2%	1.498	0.033	2.517	0.034	2.706	0.034

Optimal values are computed across different transaction cost levels using 100,000 out-of-sample paths.

The numerical results presented in Table 4.3 show a monotonic increase in rebalancing threshold values as transaction costs rise, a pattern consistently observed across all risk measures for both benchmarks and RL agents. This trend reflects broader no-trade regions at higher transaction cost levels, suggesting that small adjustments increasingly degrade hedging performance as transaction costs increase, regardless of the risk measure or approach. The incorporation of no-trade regions proves beneficial, as evidenced by the non-zero rebalancing threshold values when transaction costs are introduced into the hedging problem. In contrast, the zero threshold values obtained through the optimization process, when transaction costs are absent, are expected. This is because, under these conditions, rebalancing does not negatively affect hedging performance.

Notably, the optimal rebalancing thresholds for RL agents are consistently lower than for non-RL strategies (columns 2, 4 and 6), often approaching zero across all transaction cost levels. This suggests that the no-trade region may function as a noise-reduction mechanism within the RL framework. In this context, the likelihood of observing identical hedging positions between consecutive time steps in the traditional deep hedging framework is typically zero due to the continuity of the ANN. This results in small adjustments that, over time, increase

transaction costs and ultimately degrade hedging performance. However, the no-trade region effectively addresses this behavior by preventing such adjustments.

In terms of hedging performance, Table 4.4 presents the optimal values of the risk measures for two cases: when the hedging instruments are restricted to the risk-free asset and the underlying asset (column labeled S_t), and when an ATM call option is added as an additional hedging instrument (column labeled $S_t + \mathrm{O}_t$). This comparison is illustrated across two panels: Panel A considers strategies without a no-trade region (i.e., $l = 0$), while Panel B incorporates the no-trade region, highlighting its impact on the results.

The impact of the no-trade region can be assessed by comparing the performance of each strategy between Panel A and Panel B. For benchmark strategies, our numerical results indicate that in the case of DG hedging, incorporating a no-trade region significantly enhances hedging performance across all risk measures, particularly as transaction costs for the hedging option increase. For instance, when optimizing the rebalancing threshold using MSE, the optimal MSE for DG strategies decreases by 15%, from 0.837 (Panel A) to 0.711 (Panel B), when the hedging option's transaction cost is set to $\kappa_2 = 0.5\%$. The improvement becomes even more significant as transaction costs increase to $\kappa_2 = 2\%$, where the MSE drops by 38%, from 1.957 to 1.197, after incorporating the no-trade region.

On the other hand, for delta hedging, the improvement from introducing a rebalancing threshold is minimal, as transaction costs on the underlying asset remain nearly negligible. Similarly, for RL agents, the performance gains from the no-trade region are less pronounced, confirming that it acts as a regularization mechanism that smooths the ANN mapping. This aligns with the consistently low rebalancing threshold values reported in Table 4.3.

Table 4.4: Optimal risk measure values of deep hedging, delta hedging, and delta gamma hedging.

Risk measure		MSE				SMSE				$\text{CVaR}_{95\%}$			
Instruments		S_t		S_t+O_t		S_t		S_t+O_t		S_t		S_t+O_t	
κ_1	κ_2	D_l	RL_l	DG_l	RL_l	D_l	RL_l	DG_l	RL_l	D_l	RL_l	DG_l	RL_l
Panel A ($l = 0$)													
0%	0%	3.593	2.232	0.663	**0.106**	1.193	0.424	0.338	**0.025**	3.606	2.031	1.927	**0.514**
0.05%	0.5%			0.837	**0.124**			0.723	**0.058**			2.581	**0.704**
0.05%	1%	3.384	2.145	1.092	**0.144**	1.395	0.693	1.018	**0.099**	3.880	2.281	2.857	**0.784**
0.05%	1.5%			1.465	**0.165**			1.414	**0.132**			3.159	**0.952**
0.05%	2%			1.957	**0.193**			1.919	**0.151**			3.487	**1.010**
Panel B ($l \neq 0$)													
0%	0%	3.593	2.232	0.663	**0.106**	1.193	0.424	0.338	**0.025**	3.606	2.031	1.927	**0.514**
0.05%	0.5%			0.711	**0.122**			0.490	**0.052**			1.935	**0.647**
0.05%	1%	3.383	2.145	0.821	**0.136**	1.361	0.689	0.616	**0.069**	3.842	2.278	2.015	**0.733**
0.05%	1.5%			0.986	**0.156**			0.803	**0.098**			2.213	**0.863**
0.05%	2%			1.197	**0.179**			1.025	**0.141**			2.429	**0.972**

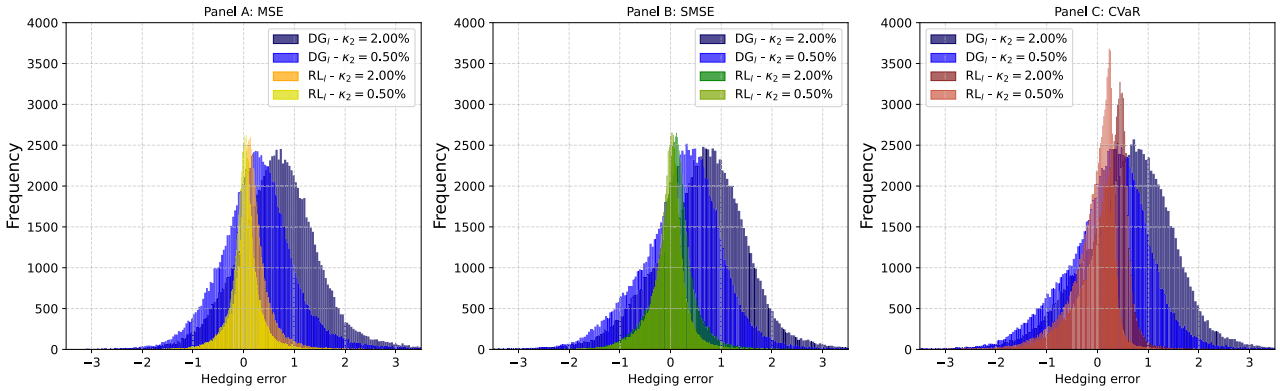
Optimal values of risk measures are computed using 100,000 out-of-sample paths. The initial average straddle value is $P_0 = 7.55$.

Considering that the no-trade region has a favorable impact on hedging performance, we now focus on the analysis of Panel B, which incorporates this feature. The results indicate that RL agents consistently outperform the benchmarks across all risk measures in both cases: with and without a hedging option. As observed previously, the inclusion of the ATM call option significantly enhances hedging performance. This improvement is particularly evident for the DG strategy, which achieves better metrics than RL agents without the hedging option when evaluated using the MSE risk measure. However, this advantage vanishes for asymmetric risk measures when the transaction costs associated with the hedging option become excessively high, for example 1.5% or greater for SMSE and 2% for CVaR. In such cases, DG strategies fail to outperform RL agents using a single risky hedging instrument. This can be attributed to the increased transaction costs associated with trading the hedging options, which offset the potential benefits of including them in the DG portfolio.

Furthermore, the advantages of incorporating a hedging option are particularly evident for RL agents. As shown in the column labeled $S_t + \mathrm{O}_t$, RL agents consistently outperform all benchmarks across a range of transaction cost levels and risk measures. This can be obseverved by the values highlighted in bold at each risk merasure column. For example, the RL agent trained using MSE with two hedging instruments (columns $S_t + \mathrm{O}_t$ under MSE) and optimized with the no-trade region (Panel B) achieve an MSE of 0.106. In comparison, other benchmarks yield significantly higher values: 3.593 for delta hedging, 2.232 for RL strategies with only one hedging instrument, and 0.663 for delta-gamma hedging when transaction costs are set to 0%. This improvement becomes even more pronounced as transaction costs increase. A similar trend is observed across other risk measures.

Moreover, to further emphasize the benefits of using RL over DG, Figure 4.2 presents the histogram of hedging error distributions at maturity for both DG and RL strategies, considering two different combinations of transaction cost levels. As shown in the figure, RL agents consistently generate narrower distributions across all risk measures compared to the DG strategy, demonstrating greater resilience to increases in transaction costs. This characteristic is particularly advantageous from a risk management perspective, as it reflects enhanced stability in performance despite increasing costs.

Figure 4.2: Hedging error distribution in the presence of transaction costs.



Results are computed using 100,000 out-of-sample paths according to the conditions outlined in Section 4.4.3. The transaction cost for the underlying asset is set to $\kappa_1 = 0.05\%$. The initial average straddle value is $P_0 = 7.55$.

4.4.4.3 Impact of no-trade regions

Since the no-trade region is determined by the rebalancing threshold, we analyze its impact by examining how the rebalancing threshold influences both the rebalancing frequency defined as the proportion of days on which portfolio positions are adjusted along a given path,

$$\mathrm{RF}_l = \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{1}_{\{\phi_{t+1} \neq \phi_t\}}, \tag{4.9}$$

and the hedging cost as the sum of discounted transaction costs over a given path,

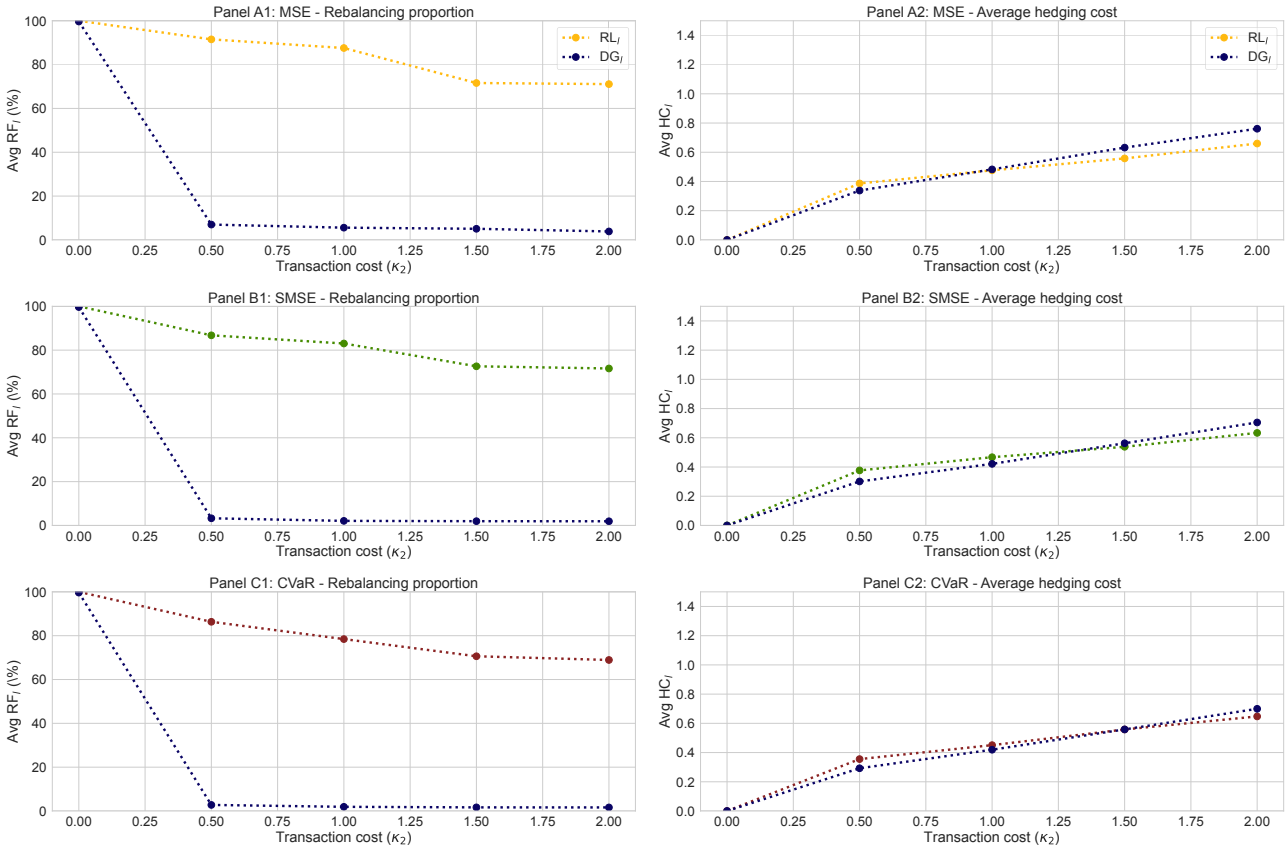
$$\mathrm{HC}_l = \sum_{t=0}^{T-1} e^{-r\Delta t} \mathcal{HC}_t \tag{4.10}$$

where the transaction cost at time t, $\mathcal{HC}_t$, is given by:

$$\mathcal{HC}_t = \kappa_1 S_t \mid \phi_{t+1}^{(S)} - \phi_t^{(S)} \mid + \kappa_2 \mathrm{O}_t(T^*) \mid \phi_{t+1}^{(\mathrm{O})} - \phi_t^{(\mathrm{O})} \mid . \tag{4.11}$$

This analysis enables the evaluation of the trade-off between portfolio adjustment frequency and the associated transaction costs. The impact of the rebalancing threshold on both rebalancing frequency and hedging cost is illustrated in Figure 4.3 across all risk measures and transaction cost levels.

Figure 4.3: Evolution of rebalancing day proportions and average hedging costs at different transaction cost levels.



Results are computed over 100,000 out-of-sample paths according to the conditions outlined in Section 4.4.3.

Results depicted in Figure 4.3 show that RL agents yield a higher average rebalancing frequency compared to DG strategies, which tend to behave more like semi-static approaches with fewer rebalancing days. This finding aligns with the observations of Carr and Wu (2014), who show that increasing the rebalancing frequency does not necessarily improve the performance of option tracking frameworks such as delta hedging in the presence of transaction cost.

In terms of hedging cost, although the difference between strategies is minimal, RL agents exhibit greater robustness to increasing κ_2. Despite their higher rebalancing frequency, their performance deteriorates less noticeably as transaction costs rise, demonstrating their ability

to adapt and maintain effective hedging even in high-cost environments.

4.4.5 Assessing the presence of speculative components in hedging positions

In this section, we explore whether the risk management strategies incorporate speculative elements, such as "good deals" which refer to capitalizing on the benefits derived from the risk premium. This scenario would be considered undesirable in practice, as it deviates from sound risk management practices.

4.4.5.1 Risk premium and good deals

As a second test, we analyze if RL agents exploit "good deals" in terms of seeking benefit form the risk premium offered by the hedging option, specically an ATM call option with maturity of $T^* > T$ days and strike K^*. In this analysis we consider the risk premium (RP) as the difference between the expected payoff at time t and the option price at time t, *i.e.*,

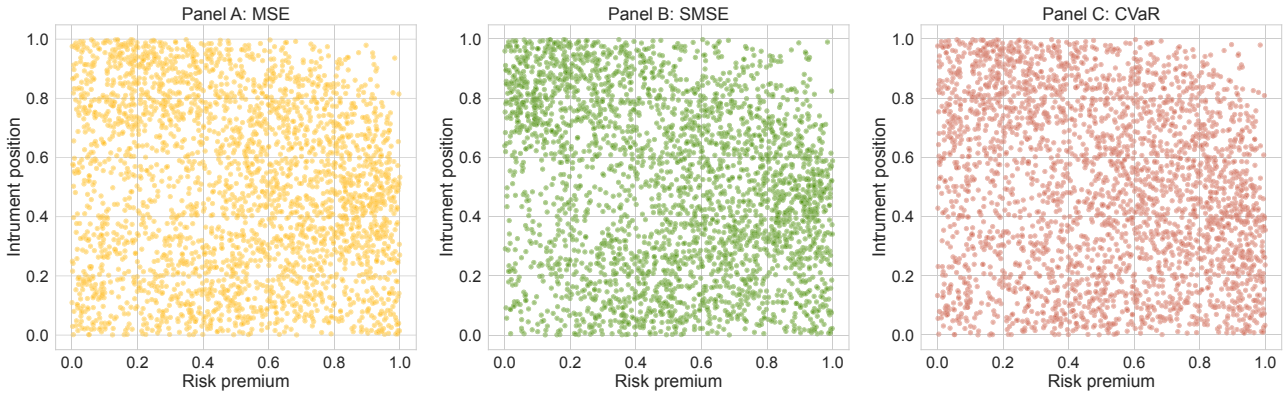
$$\mathrm{RP}_t = \exp(-r_t(T^* - t))\mathbb{E}[\max(S_{T^*} - K^*, 0) \mid \mathcal{F}_t] - \mathrm{O}_t(K^*, T^*)\,, \tag{4.12}$$

where $\mathcal{F}_t$ denotes the information available at time t. The risk premium is calculated using a stochastic-on-stochastic simulation approach, where the present value of the expected payoff is simulated at each time step, nested within the simulated paths. In this framework, state vectors are randomly sampled from the test set to serve as initial points for the simulation. By repeatedly simulating from various initial conditions, the risk premium captures the dynamic behavior of the system under diverse potential paths.

In this analysis, we examine whether there is a statistical relationship between the risk premium RP_t and the position in the hedging instrument $\phi_{t+1}^{(\mathrm{O})}$, with the goal of determining whether RL agents capitalize on RP_t. Figure 4.4 presents the scatter plot of ranked data between these two variables, using a sample of 20,000 data points from 100,000 out-of-sample paths across all risk measures. The scatter plot does not reveal any strong dependence patterns, suggesting weak or no significant relationship. This finding is further supported by

the sample correlations, which range from -0.001 to -0.006 across all scenarios. These results imply that RL agents do not appear to engage in strategies specifically aimed at capturing risk premium benefits.

Figure 4.4: Scatter plot from ranked data of risk premium and hedging option positions.



Results are computed using a sample of 20,000 data points from 100,000 out-of-sample paths. Transaction cost levels are set to 0%.

As a complementary analysis, we assess whether our approach incorporates speculative elements, such as statistical arbitrage overlays, that may not align with sound risk management practices. Our results suggest that RL agents do not adopt in such strategies, regardless of the risk measure used during the optimization task. Further details can be found in Appendix 4.6.6.

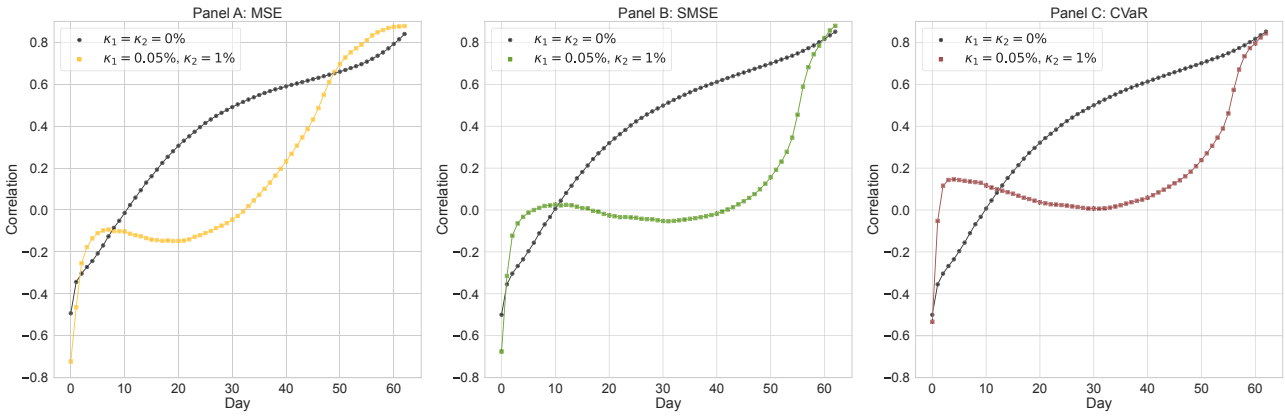
4.4.6 Statistical study and sensitivity analysis of hedging strategies

4.4.6.1 Statistical analysis of hedging option positions: Benchmarks vs RL agents

We start by analyzing the relationship between the hedging option positions recommended by the DG strategy and those generated by RL agents. This analysis aims to understand how the outperformance documented in Sections 4.4.4.1 and 4.4.4.2 is achieved by RL agents by examining the positions taken by the hedger. Figure 4.5 presents the sample correlation between the hedging option positions for DG and RL agents, $\phi^{(\mathrm{O,DG})}$ and $\phi^{(\mathrm{O,RL})}$, when both are optimized using three risk measures: MSE, SMSE, and $\mathrm{CVaR}_{95\%}$. The sample correlation is computed daily over the entire hedging period, considering two scenarios: one without

transaction costs and another with transaction costs set to $\kappa_1 = 0.05\%$ and $\kappa_2 = 1\%$, for illustration purposes.

Figure 4.5: Pearson correlation between DG and RL agent's hedging option positions.

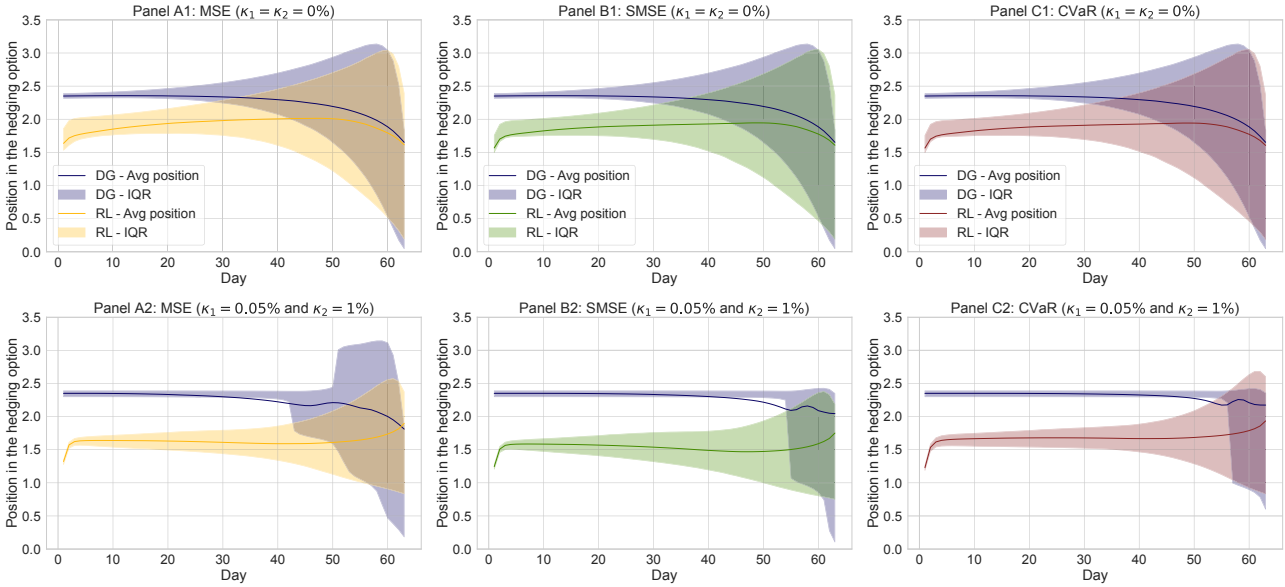


Results are based on a sample of 100,000 out-of-sample paths. Agents are trained under the conditions described in Section 4.4.3.

Our numerical results reveal a clear pattern across all risk measures, with a significant difference in the correlation between the RL agent and DG hedging strategies, particularly at the beginning of the hedging horizon. At the onset, the RL agent adopts a fundamentally different approach, leading to low or even negative correlation with the DG strategy. As the hedging period progresses, and in the absence of transaction costs, the correlation gradually increases, albeit slowly, reaching approximately 50% by mid-period. This slow convergence is striking, as it suggests that the RL agent continues to employ a hedging strategy that differs from the traditional DG approach. The eventual convergence towards higher correlation levels is expected, as the payoff structure becomes clearer close to maturity. However, the inclusion of transaction costs results in the RL agent maintaining a distinct approach, with the correlation remaining near zero for a considerable period. This suggests that the RL agent has learned to manage hedging costs more effectively, avoiding myopic decision-making, in contrast to DG, which focuses on option tracking. This behavior is likely driven by the optimization objective of the RL agent, which explicitly targets minimizing the terminal hedging error rather than tracking the option price, as is the case in DG hedging.

Additionally, a potential secondary source of divergence between these strategies may stem from differences in rebalancing size. While the frequency of rebalancing influences the timing of adjustments, the magnitude of these adjustments could also play a key role in differentiating the hedging behaviors of the various strategies. Figure 4.6 illustrates the average hedging option position, along with the interquartile range, over time for all risk measures. The analysis is presented for two scenarios: one without transaction costs (first row), and another with transaction costs set to $\kappa_1 = 0.05\%$ and $\kappa_2 = 1\%$ (second row).

Figure 4.6: Distribution of hedging option positions.



Results are computed over 100,000 out-of-sample paths according to the conditions outlined in Section 4.4.3.1. IQR stands for the interquartile range, representing the range between the 25th and 75th percentiles.

Our findings reveal that RL agents tend to have lower option positions during the initial stages of the hedging period, a tendency that becomes increasingly evident with the introduction of transaction costs. This behavior likely stems from the significant trading costs associated with the hedging option, suggesting that RL agents employ more frequent rebalancing with smaller positions early in the period, gradually increasing their hedge sizes over time. This suggests that the agent initially maintains a lower exposure to the volatility risk premium, gradually increasing its hedge positions over time. By deferring full engagement with the

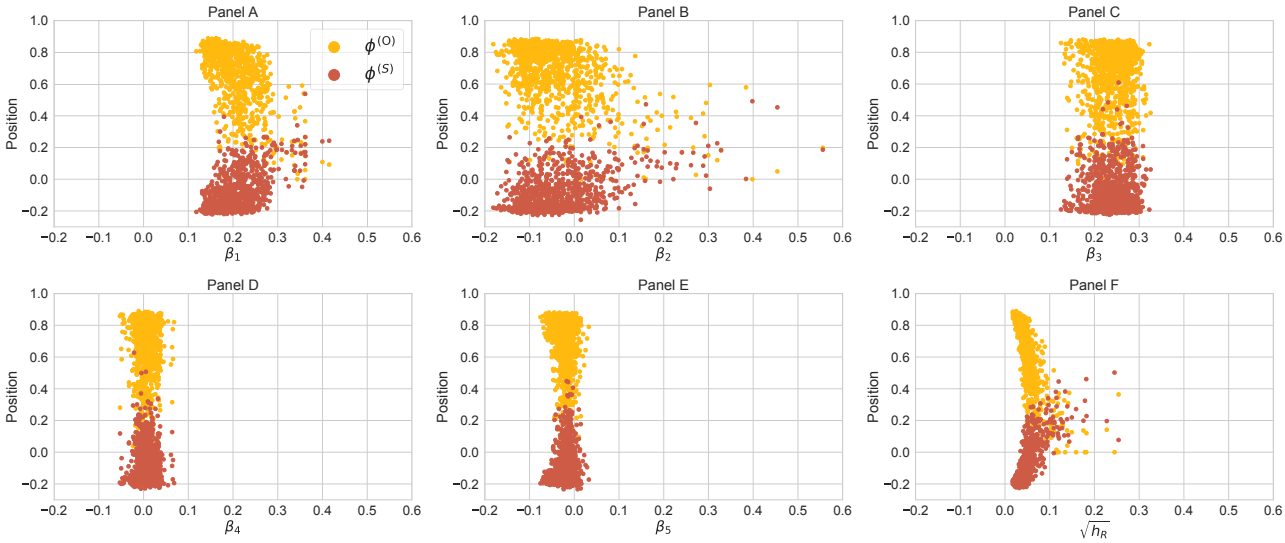
hedge, the agent seeks to balance cost efficiency with effective risk management. In contrast, DG strategies tend to overutilize options early on to fully neutralize gamma risk. However, this approach incurs prolonged exposure to the volatility premium, which proves suboptimal.

4.4.6.2 Sensitivity analysis

We analyze the sensitivity of the RL agents' positions to variations in the risk factors defining the IV surface, exploring how the agents leverage information from its shape. We begin by examining the policy behavior of the RL positions across different initial scenarios for the state variables $(\{\beta_{t,i}\}_{i=1}^{5}, h_{t,R})$. To evaluate the impact of each state variable, we sort the sample of initial state vectors in our test set according to each variable and observe the resulting hedging positions in the same order. This approach accounts for the interdependence between these state variables and the other components of the state vector, as detailed in Table 4.1, and illustrates how variations in the selected variable influence the direction of the hedging positions. We focus on the initial state vector to ensure comparable market conditions in terms of initial underlying asset price and maturity, i.e., at $T = 63$ days-to-maturity.

Figure 4.7 illustrates the hedging positions of the RL agent trained with the MSE risk measure while assuming no transaction costs. This scenario is selected for simplicity, as François et al. (2024) demonstrate that RL agents trained under MSE, SMSE, and CVaR exhibit sensitivity to similar state variables, even though the degree of sensitivity may differ across the risk measures. Each panel displays the hedging positions when the initial state vectors are sorted by each of the state variables, $(\{\beta_{t,i}\}_{i=1}^{5}, h_{t,R})$.

Figure 4.7: Marginal impact on hedging positions with respect to IV coefficients and underlying asset volatility.



Results are computed using a sample of 20,000 data points from 100,000 out-of-sample paths for an ATM straddle with maturity of $T = 63$ days. Transaction cost levels are set to 0%.

These empirical results suggest that the position in the hedging option exhibits a more pronounced decreasing trend with respect to the conditional variance of the underlying asset returns, the long-term ATM level β_1, the time-to-maturity slope β_2, and the smirk β_5 of the IV surface. As noted in François et al. (2024), this highlights that RL agents utilize both the historical variance process and market expectations of future volatility to adjust their positions. Overall, the results indicate that the RL agent reduces its exposure to the hedging option when higher risk is perceived at the onset.

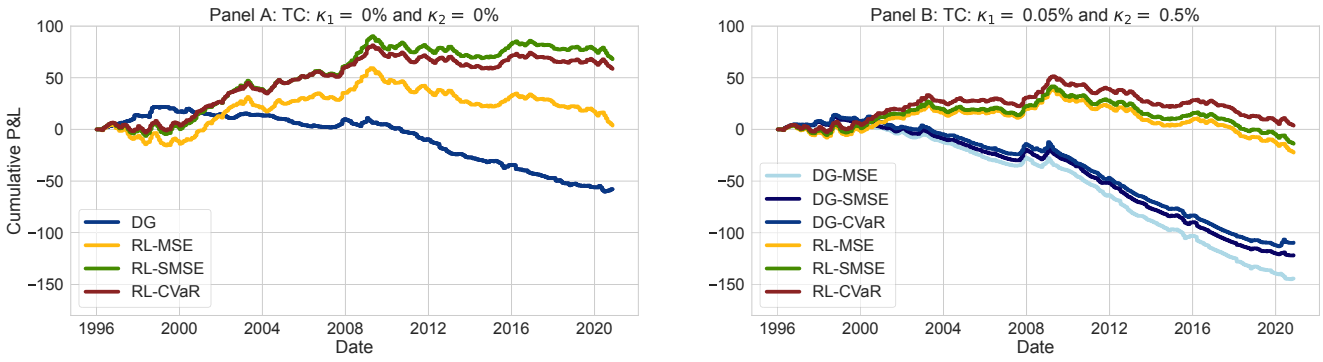
4.4.7 Backtesting

In this section, we benchmark our approach using historical paths from the JIVR model, spanning from January 5, 1996, to December 31, 2020, to evaluate the effectiveness of the RL agents. This experiment assesses the performance of risk management strategies based on the historical series (R_t, β_t). Specifically, we evaluate the hedging performance by introducing a new ATM straddle instrument with a maturity of 63 days every 21 business days along the

historical paths. The initial values for the hedging portfolio are determined by the straddle prices, which are calculated using the prevailing implied volatility surface on the day the hedge is initiated.

To assess the robustness of our approach in more general market conditions, we compare cumulative P&Ls, which are computed as the running totals of the P&L achieved by each strategy at the maturity of each instrument during the analysis period. Figure 4.8 displays the cumulative P&L evolution across two panels corresponding to different transaction cost levels.

Figure 4.8: Cumulative P&L for a ATM straddle instruments with a maturity of 63 days under real asset price dynamics.



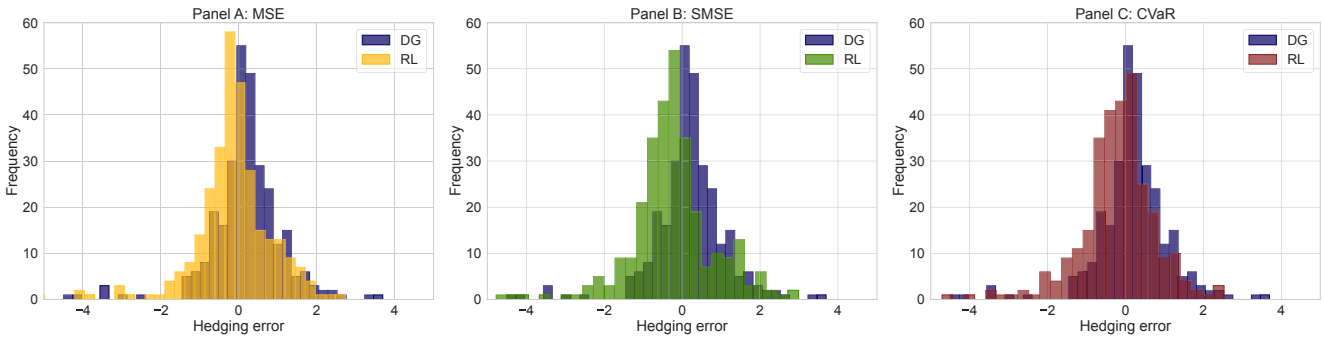
Results are computed based on the observed P&L from hedging 296 straddle instruments under real market conditions observed from May 1, 1996, to December 31, 2020. A new ATM straddle is considered every 21 business days.

As illustrated in Figure 4.8, RL strategies consistently outperform the benchmarks in both scenarios, with and without transaction costs. Notably, the gap between the cumulative P&L of RL agents and the benchmarks widens significantly as transaction costs increase, highlighting the adaptability of the RL approach to diverse market conditions. Additionally, RL strategies optimized using the MSE function yield lower cumulative P&L compared to those optimized with asymmetric risk measures, reflecting the inherent differences in the objectives of these risk measures.

For hedging errors, we examine their distribution under real asset price dynamics by analyzing

the errors generated by 296 ATM straddles over the period from May 1, 1996, to December 31, 2020. Figure 4.9 displays the histogram of hedging errors for both benchmarks and RL agents across all risk measures, with transaction costs excluded for simplicity.

Figure 4.9: Hedging error distribution under real asset price dynamics.



Results are computed based on the observed P&L from hedging 296 straddle instruments under real market conditions observed from May 1, 1996, to December 31, 2020. Transaction cost levels are set to 0%.

As shown in Figure 4.9, RL agents demonstrate not only superior cumulative P&L over the analysis period but also a less pronounced right tail in the hedging error distribution. Furthermore, the hedging error distributions produced by RL agents are shifted more toward the left, underscoring their ability to effectively manage risk. These findings highlight the robustness of the RL approach, with the observed performance under historical data providing evidence of its reliability.

4.5 Conclusion

This study presents a deep hedging framework for portfolios of options using multiple hedging instruments. The implementation incorporates state-dependent no-trade regions to optimize rebalancing frequency in the presence of transaction costs. The hedging policies leverage forward-looking volatility information through a functional representation of the IV surface, combined with traditional backward-looking features. Furthermore, the optimization framework includes a soft constraint to discourage speculative behavior, enabling the agent to develop hedging strategies that prioritize effective risk management.

Our approach consistently outperforms traditional benchmarks in both the absence and presence of transaction costs, underscoring the hedging performance benefits of incorporating additional instruments, such as options. Moreover, the inclusion of no-trade regions enhances performance for both RL and DG strategies. Specifically, RL policies are smoothed, avoiding unnecessary rebalancing, while DG strategies converge toward semi-static hedging approaches.

Our findings highlight a significant difference in correlation between RL and DG hedging strategies, indicating that the RL agent maintains a distinct approach. With transaction costs, the RL agent keeps correlation near zero for an extended period, suggesting an alternative cost management strategy that avoids myopic decision-making, unlike DG's focus on option tracking.

The sensitivity analysis of RL policies with respect to IV features reveals that RL agents effectively integrate both historical variance and market expectations of future volatility into their hedging decisions. The observed decreasing trend in hedging option exposure in response to higher conditional variance, long-term ATM levels, time-to-maturity slopes, and IV smirk highlights the agents' ability to dynamically mitigate risk, serving as a protective mechanism against volatility fluctuations.

4.6 Appendix

4.6.1 Neural network settings

4.6.1.1 Network architecture

The RNN-FNN network is defined as a composition of LSTM cells $\{C_l\}_{l=1}^{L_1}$ and FFNN layers $\{\mathcal{L}_j\}_{j=1}^{L_2}$, represented by the following functional form:

$$\tilde{\phi}_\theta(X_t) = (\underbrace{\mathcal{L}_J \circ \mathcal{L}_{L_2} \circ \mathcal{L}_{L_2-1} \circ ... \circ \mathcal{L}_1}_{\text{FFNN layers}} \circ \underbrace{C_{L_1} \circ C_{L_1-1} ... \circ C_1}_{\text{LSTM cells}})(X_t).$$

Each LSTM cell C_l maps a vector $Z_t^{(C,l-1)}$ of dimension $d^{(C,l-1)}$ to a new vector $Z_t^{(C,l)}$ of dimension $d^{(C,l)}$ based on the following equations, starting with $Z_t^{(C,0)} = X_t$:

$$\begin{aligned} i^{(l)} &= \text{sigm}(W_i^{(l)} Z_t^{(C,l-1)} + b_i^{(l)}), \\ o^{(l)} &= \text{sigm}(W_o^{(l)} Z_t^{(C,l-1)} + b_o^{(l)}), \\ c^{(l)} &= i^{(l)} \odot \tanh(W_c^{(l)} Z_t^{(C,l-1)} + b_c^{(l)}), \\ Z_t^{(C,l)} &= o_t^{(l)} \odot \tanh(c^{(l)}), \end{aligned}$$

where $\text{sigm}(\cdot)$ and $\tanh(\cdot)$ represent the sigmoid and hyperbolic tangent functions, applied element-wise, and $\odot$ denotes the Hadamard product. The FFNN layer $\mathcal{L}_j$ maps the input vector $Z_t^{(\mathcal{L},j-1)}$ of dimension $d^{(\mathcal{L},j-1)}$ to an output vector $Z_t^{(\mathcal{L},j)}$ of dimension $d^{(\mathcal{L},j)}$ by applying a linear transformation $T_{\mathcal{L}_j}(Z_t^{(\mathcal{L},j-1)}) = W_{\mathcal{L}_j} Z_t^{(\mathcal{L},j-1)} + b_{\mathcal{L}_j}$ followed by a non-linear activation function $g_{\mathcal{L}_j}$. Thus, the operation for layer $\mathcal{L}_j$ is expressed as:

$$\mathcal{L}_j(Z_t^{(\mathcal{L},j-1)}) = (g_{\mathcal{L}_j} \circ T_{\mathcal{L}_j})(Z_t^{(\mathcal{L},j-1)})$$

for $j \in \{1, ..., L_2, J\}$, where the initial input to the first FFNN layer is $Z_t^{(\mathcal{L},0)} = Z_t^{(C,L_1)}$.

The trainable parameters θ of the RNN-FNN network are defined as follows:

- For $L_1 \geq l \geq 1$: $W_i^{(l)}$, $W_o^{(l)}$, $W_c^{(l)} \in \mathbb{R}^{d^{(C,l)} \times d^{(C,l-1)}}$ and $b_i^{(l)}$, $b_o^{(l)}$, $b_c^{(l)} \in \mathbb{R}^{d^{(C,l)} \times 1}$, with $d^{(C,0)}$ representing the original dimension of the input vector.
- For $L_2 \geq j \geq 1$: $W_{\mathcal{L}_j} \in \mathbb{R}^{d^{(\mathcal{L},j)} \times d^{(\mathcal{L},j-1)}}$ and $b_{\mathcal{L}_j} \in \mathbb{R}^{d^{(\mathcal{L},j)}}$, with $d^{(\mathcal{L},0)} = d^{(C,L_1)}$.
- For $j = J$: $W_{\mathcal{L}_J} \in \mathbb{R}^{2 \times d^{(\mathcal{L},L_2)}}$ and $b_{\mathcal{L}_J} \in \mathbb{R}$.

The hyperparameter values chosen for our experiments are specified in Section 4.4.3.1

4.6.1.2 Details for the MSGD training approach

The MSGD method estimates the penalty function $\mathcal{O}_\lambda(\theta, l)$, which is typically unknown, by using small samples of the hedging error, referred to as batches. Let $\mathbb{B}_j = \left\{ \xi_{T,i}^{(\tilde{\phi}_{\theta_j}, l_j)} \right\}_{i=1}^{N_{\text{batch}}}$

be the j-th batch, where N_{batch} is the batch size and $\xi_{T,i}^{(\tilde{\phi}_{\theta_j},l_j)}$ denotes the hedging error for the i-th path in the j-th batch. This is defined as

$$\xi_{T,i}^{(\tilde{\phi}_{\theta_j},l_j)} = \Psi(S_{T,(j-1)N_{\text{batch}}+i}) - V_{T,i}^{(\tilde{\phi}_{\theta_j},l_j)} \quad \text{for} \quad i \in \{1,\ldots,N_{\text{batch}}\},\, j \in \{1,\ldots,N\},$$

where $S_{T,(j-1)N_{\text{batch}}+i}$ is the price of the underlying asset at time T in the $((j-1)N_{\text{batch}}+i)$-th simulated path, and $V_{T,i}^{(\tilde{\phi}_{\theta_j},l_j)}$ represents the terminal value of the hedging strategy for the path when $\theta = \theta_j$ and $l = l_j$, with the simulated states being X_i.

The penalty function estimation for batch $\mathbb{B}$ is as follows:

$$\begin{aligned}
\hat{\mathcal{O}}_\lambda^{(\text{MSE})}(\theta_j,l_j,\mathbb{B}_j) &= \frac{1}{N_{\text{batch}}}\sum_{i=1}^{N_{\text{batch}}}\left(\xi_{T,i}^{(\tilde{\phi}_{\theta_j},l_j)}\right)^2 + \lambda\cdot\widehat{SC}(\theta_j,l_j,\mathbb{B}_j),\\
\hat{\mathcal{O}}_\lambda^{(\text{SMSE})}(\theta_j,l_j,\mathbb{B}_j) &= \frac{1}{N_{\text{batch}}}\sum_{i=1}^{N_{\text{batch}}}\left(\xi_{T,i}^{(\tilde{\phi}_{\theta_j},l_j)}\right)^2 \mathbb{1}_{\left\{\xi_{T,i}^{(\tilde{\phi}_{\theta_j},l_j)}\geq 0\right\}} + \lambda\cdot\widehat{SC}(\theta_j,l_j,\mathbb{B}_j),\\
\hat{\mathcal{O}}_\lambda^{(\text{CVaR})}(\theta_j,l_j,\mathbb{B}_j) &= \widehat{\text{VaR}}_\alpha(\mathbb{B}_j) + \frac{1}{(1-\alpha)N_{\text{batch}}}\sum_{i=1}^{N_{\text{batch}}}\max\left(\xi_{T,i}^{(\tilde{\phi}_{\theta_j},l_j)} - \widehat{\text{VaR}}_\alpha(\mathbb{B}_j),0\right)\\
&\quad + \lambda\cdot\widehat{SC}(\theta_j,l_j,\mathbb{B}_j),
\end{aligned}$$

where

$$\widehat{SC}(\theta_j,l_j,\mathbb{B}_j) = \frac{1}{N_{\text{batch}}}\sum_{i=1}^{N_{\text{batch}}}\mathbb{1}_{\left\{\max_{t\in\{0,\ldots,T\}}\left[P_{t,i}-V_{t,i}^{(\tilde{\phi}_{\theta_j},l_j)}\right]>V_{0,i}^{(\tilde{\phi}_{\theta_j},l_j)}\right\}},$$

and $\widehat{\text{VaR}}_\alpha(\mathbb{B}_j) = \xi_{T,\lceil\alpha\cdot N_{\text{batch}}\rceil}^{(\tilde{\phi}_{\theta_j},l_j)}$ is the value-at-risk estimation derived from the ordered sample $\left\{\xi_{T,[i]}^{(\tilde{\phi}_{\theta_j},l_j)}\right\}_{i=1}^{N_{\text{batch}}}$, where $\lceil\cdot\rceil$ is the ceiling function. These empirical approximations are used to estimate the gradient of the penalty function, which is required in Equations (4.6) and (4.7). The gradient of these empirical objective functions has analytical expressions for FFNN, LSTM networks, and RNN-FNN networks. Detailed calculations of the gradient of the empirical objective function are provided in Goodfellow et al. (2016). We use the batch size from François et al. (2024) ($N_{\text{batch}} = 1{,}000$).

4.6.2 Joint Implied Volatility and Return model

4.6.2.1 Daily implied volatility surface

The full functional representation of the IV surface model introduced by François et al. (2022) is given by:

$$\sigma(M_t, \tau_t, \beta_t) = \underbrace{\beta_{t,1}}_{f_1\text{: Long-term ATM IV}} + \beta_{t,2} \underbrace{e^{-\sqrt{\tau_t/T_{conv}}}}_{f_2\text{: Time-to-maturity slope}} + \beta_{t,3} \underbrace{\left(M_t \mathbb{1}_{\{M_t \geq 0\}} + \frac{e^{2M_t} - 1}{e^{2M_t} + 1} \mathbb{1}_{\{M_t < 0\}} \right)}_{f_3\text{: Moneyness slope}}$$
$$+ \beta_{t,4} \underbrace{\left(1 - e^{-M_t^2}\right) \log(\tau_t / T_{max})}_{f_4\text{: Smile attenuation}} + \beta_{t,5} \underbrace{\left(1 - e^{(3M_t)^3}\right) \log(\tau_t / T_{max}) \mathbb{1}_{\{M_t < 0\}}}_{f_5\text{: Smirk}}, \quad \tau_t \in [T_{min}, T_{max}] \tag{4.13}$$

As in François et al. (2022), T_{max} is set to 5 years, $T_{min} = 6/252$ and T_{conv} to 0.25.

4.6.2.2 Joint Implied Volatility and Return

The multivariate time series representation of the JIVR model introduced by François et al. (2023) includes two components: one for the returns of the underlying asset and another for the fluctuations of the IV surface coefficients. The first component follows an NGARCH(1,1) process with NIG innovations, and is expressed as:

$$\begin{aligned} R_{t+1} &= \xi_{t+1} - \psi(\sqrt{h_{t+1,R}\Delta}) + \sqrt{h_{t+1,R}\Delta}\epsilon_{t+1,R}, \\ h_{t+1,R} &= Y_t + \kappa_R(h_{t,R} - Y_t) + a_R h_{t,R}(\epsilon_{t,R}^2 - 1 - 2\gamma_R \epsilon_{t,R}), \\ Y_t &= \left(\omega_R \, \sigma\left(0, \frac{1}{12}, \beta_t\right)\right)^2, \end{aligned} \tag{4.14}$$

where the equity risk premium is given by:

$$\xi_{t+1} = \psi(-\lambda\sqrt{h_{t+1,R}\Delta}) - \psi((1-\lambda)\sqrt{h_{t+1,R}\Delta}) + \psi(\sqrt{h_{t+1,R}\Delta}), \tag{4.15}$$

and the process $\{\epsilon_{t,R}\}_{t=0}^{T}$ is a sequence of iid standardized NIG random variables with parameters ζ_R and φ_R. The standard NIG random variable ϵ is fully defined by its probability

density function, with parameters ζ and φ.[24] Parameters for the excess return component of the model are $\Theta_R = (\lambda, \kappa_R, \gamma_R, a_R, \omega_R, \zeta_R, \varphi_R)$.

The second component of the model consists of five heteroskedastic autoregressive processes, each having NIG innovations for the coefficients of the implied volatility factors. The evolution of the long-term factor β_1 is modeled as:

$$\begin{aligned}
\beta_{t+1,1} &= \alpha_1 + \sum_{i=1}^{5} \theta_{1,j}\beta_{t,j} + \sqrt{h_{t+1,1}\Delta}\epsilon_{t+1,1}, \\
h_{t+1,1} &= U_t + \kappa_1(h_{t,1} - U_t) + a_1 h_{t,1}(\epsilon_{t,1}^2 - 1 - 2\gamma_1\epsilon_{t,1}), \\
U_t &= \left(\omega_1 \cdot \sigma\left(0, \frac{1}{12}, \beta_t\right)\right)^2.
\end{aligned} \tag{4.16}$$

For the evolution of the other four coefficients, for $i \in \{2, 3, 4, 5\}$, we have:

$$\begin{aligned}
\beta_{t+1,i} &= \alpha_i + \sum_{j=1}^{5} \theta_{i,j}\beta_{t,j} + \nu\beta_{t-1,2}\mathbb{I}_{\{i=2\}} + \sqrt{h_{t+1,i}\Delta}\epsilon_{t+1,i}, \\
h_{t+1,i} &= \sigma_i^2 + \kappa_i(h_{t,i} - \sigma_i^2) + a_i h_{t,i}(\epsilon_{t,i}^2 - 1 - 2\gamma_i\epsilon_{t,i}),
\end{aligned} \tag{4.17}$$

where $\{\epsilon_{t,i}\}_{i=1}^5$ are time-independent standardized NIG random variables with parameters $\{(\zeta_i, \varphi_i)\}_{i=1}^5$. Parameters for the various IV coefficient marginal processes are denoted

$$\{\Theta_i = (\omega_1, \alpha_i, \theta_{i,1}, \theta_{i,2}, \theta_{i,3}, \theta_{i,4}, \theta_{i,5}, \nu, \sigma_i, \kappa_i, a_i, \gamma_i, \zeta_i, \varphi_i)\}_{i=1}^5.$$

[24] The standard NIG random variable ϵ is fully characterized by the following probability density function with parameters ζ and φ:

$$f(x) = \frac{B_1\left(\sqrt{\frac{\varphi^6}{\varphi^2+\zeta^2} + (\varphi^2+\zeta^2)\left(x + \frac{\varphi^2\zeta}{\varphi^2+\zeta^2}\right)^2}\right)}{\pi\sqrt{\frac{1}{\varphi^2+\zeta^2} + \frac{\varphi^2+\zeta^2}{\varphi^6}\left(x + \frac{\varphi^2\zeta}{\varphi^2+\zeta^2}\right)^2}} e^{\left(\frac{\varphi^4}{\varphi^2+\zeta^2} + \zeta\left(x + \frac{\varphi^2\zeta}{\varphi^2+\zeta^2}\right)\right)},$$

where $B_1(\cdot)$ denotes the modified Bessel function of the second kind with index 1. The common $(\alpha, \beta, \delta, \mu)$-specification can be obtained by replacing β and γ ($\gamma = \sqrt{\alpha^2 - \beta^2}$), with ζ and φ, respectively, and imposing a null mean and unit variance to express δ and μ in terms of α, β. .

Additionally, the JIVR model imposes a dependence structure on the contemporaneous innovations $\epsilon_t = (\epsilon_{t,R}, \epsilon_{t,1}, ..., \epsilon_{t,5})$ through a Gaussian copula, which is parameterized using a covariance matrix Σ of dimension 6×6. Parameter estimates for the entire JIVR model are sourced from Table 5 and Table 6 of François et al. (2023).

4.6.3 Benchmarks

The benchmarks presented in this appendix assume that implied volatilities adhere to the IV model specified in Equation (4.8).

4.6.3.1 Leland Model

The Leland delta hedging strategy, introduced by Leland (1985), modifies the classical option replication framework of Black and Scholes (1973) by incorporating transaction costs, represented by the proportion κ, and the rebalancing frequency λ. The hedging position in the underlying asset is given by:

$$\phi_{t+1}^{(S)} = \mathrm{e}^{-q_t \tau_t} \Phi\left(\tilde{d}_t\right),$$

where

$$\tilde{d}_t = \frac{\log\left(\frac{S_t}{K}\right) + \left(r_t - q_t + \frac{1}{2}\tilde{\sigma}_t^2\right)\tau_t}{\tilde{\sigma}_t \sqrt{\tau_t}}$$

with the adjusted volatility

$$\tilde{\sigma}_t^2 = \sigma(M_t, \tau_t, \beta_t)^2 \left[1 + \sqrt{\frac{2}{\pi}} \frac{2\kappa}{\sigma(M_t, \tau_t, \beta_t)\sqrt{\lambda}}\right].$$

Here, Φ denotes the cumulative distribution function of the standard normal distribution.

4.6.3.2 Delta gamma hedging

The delta gamma hedging strategy involves both the underlying asset S and an additional hedging instrument, O. This setup allows for neutralizing both the delta and gamma of the portfolio. The trading strategy ϕ is fully determined by the process $(\phi^{(S)}, \phi^{(\mathrm{O})})$, expressed as:

$$(\phi_{t+1}^{(S)}, \phi_{t+1}^{(\mathrm{O})}) = \left(\Delta_t^P - \frac{\Gamma_t^P}{\Gamma_t^{(\mathrm{O})}} \Delta_t^{(\mathrm{O})}, \frac{\Gamma_t^P}{\Gamma_t^{(\mathrm{O})}} \right),$$

where Δ_t^P and Γ_t^P represent the Black-Scholes delta and gamma of the hedged portfolio, while $\Delta_t^{(\mathrm{O})}$ and $\Gamma_t^{(\mathrm{O})}$ correspond to the Black-Scholes delta and gamma of the hedging option. The explicit formulas for the Black-Scholes delta and gamma are given as follows:

$$\Delta = \mathrm{e}^{-q_t \tau_t} \Phi(d_t), \quad \Gamma = \mathrm{e}^{-q\tau_t} \frac{\varphi(d_t)}{S_t \sigma_t \sqrt{\tau_t}},$$

where $d_t = \frac{\log\left(\frac{S_t}{K}\right) + \left(r_t - q_t + \frac{1}{2}\sigma(M_t, \tau_t, \beta_t)^2\right)\tau_t}{\sigma(M_t, \tau_t, \beta_t)\sqrt{\tau_t}}$, σ_t is the implied volatility of the option, and Φ and φ represent the cumulative distribution function and probability density function of the standard normal distribution, respectively.

4.6.4 Soft constraint regularization

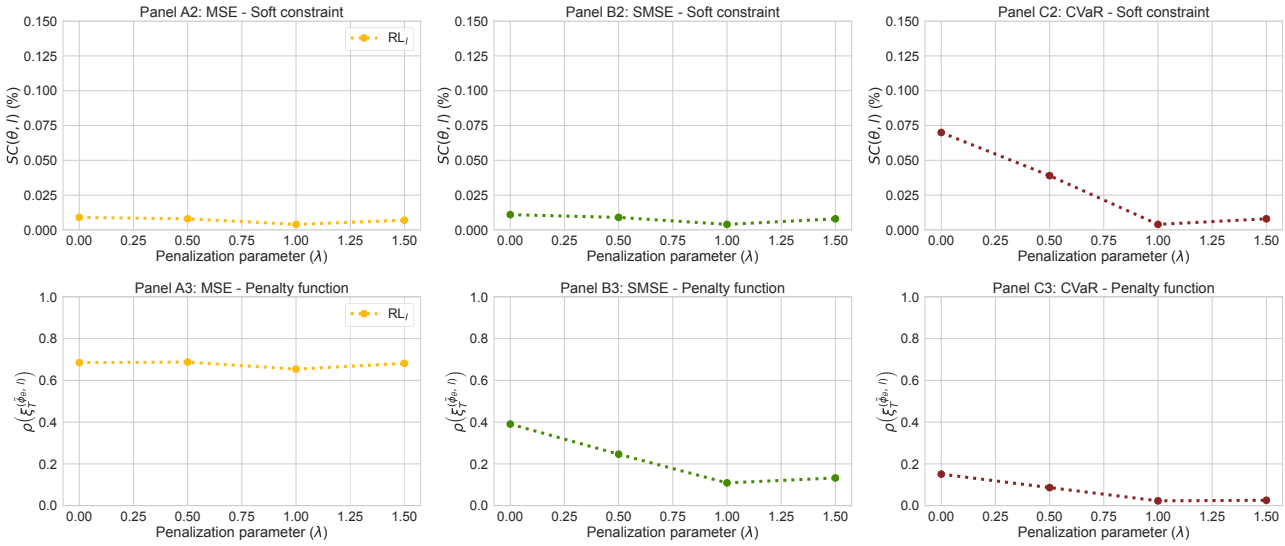
The estimation of the penalization parameter λ introduced in Equation (4.5), which governs the weight of the soft constraint in the optimization process, is approached as a model selection problem. In this framework, the model is trained multiple times using fixed values of λ, iterating across a predefined grid of λ values. The optimal λ is then selected based on an evaluation conducted on the validation set,[25] considering two key factors: the soft constraint value and the penalty function value.

To find the optimal value of λ, we hedge an ATM straddle instrument with a maturity of $T = 63$ days in the absence of transaction costs ($\kappa_1 = \kappa_2 = 0\%$). The hedging instruments used in this experiment include the risk-free asset, the underlying asset, and an ATM call option with a maturity of $T^* = 84$ days (four months). The optimization of hedging strategies considers three penalty functions: MSE, SMSE, and $\mathrm{CVaR}_{95\%}$. This process is repeated for various values of the penalization parameter λ: 0, 0.5, 1, and 1.5. Figure 4.10 illustrates the

[25]The validation set consists of 100,000 independent simulated paths, generated as outlined in Section 4.4.1. This set is distinct from the training and test sets described in Section 4.4.3.1.

optimal values of the soft constraint and the penalty functions for various λ values, evaluated on a validation set.

Figure 4.10: Optimal penalty function and soft constraint values for various penalization parameter values, applied to a straddle with a maturity of $T = 63$ days.



Results are computed over 100,000 out-of-sample paths according to the conditions outlined in Section 4.4.3.1 using an ATM call option with maturity $T_1 = 84$ days as the hedging instrument under different transaction costs levels.

The results illustrated in Figure 4.10 highlight the heightened sensitivity of asymmetric penalty functions to variations in the penalization parameter λ. The SMSE penalty function exhibits significant sensitivity of ρ, achieving its minimum value at $\lambda = 1$, which aligns with the corresponding minimum value of the soft constraint. For the CVaR penalty function, the soft constraint demonstrates greater sensitivity compared to the penalty function itself, indicating that CVaR is more susceptible to deviations from the hedged portfolio value in the absence of the soft constraint. The minimum value of the soft constraint for CVaR also occurs at $\lambda = 1$, corresponding to the stabilization point of the penalty function. In contrast, the MSE penalty function is mildly affected by the soft constraint, yet its minimum value is also observed at $\lambda = 1$, mirroring the behavior of the other penalty functions.

Based on these findings, we select $\lambda = 1$ for our subsequent experiments. This value ensures

soft constraint levels remain below 0.025% across all penalty functions.

4.6.5 Impact of state variable inclusion on hedging performance

To evaluate the impact of including state variables $\mathcal{P}_t$, Δ_t^P, and γ_t^P in the reinforcement learning framework, we conduct additional numerical experiments. Specifically, we compare the performance of RL agents trained with and without these variables across various risk measures. The evaluation involves hedging a straddle position with a maturity of $T = 63$ days, ATM call option with a maturity of $T^* = 84$ days as the hedging instrument. Table 4.5 demonstrates that the inclusion of state variables consistently improves hedging performance, likely because they provide additional structure and information that compensate for the suboptimal convergence issues typically encountered in finite settings.

Table 4.5: Optimal risk measure values for different state space configurations.

State space	MSE	SMSE	$\text{CVaR}_{95\%}$
$\mathcal{S}\backslash\{\mathcal{P}_t, \Delta_t^P, \gamma_t^P\}$	0.199	0.086	0.693
$\mathcal{S}\backslash\{\mathcal{P}_t\}$	0.131	0.059	0.687
$\mathcal{S}$	**0.106**	**0.025**	**0.514**

Optimal values are computed using 100,000 out-of-sample paths. Transaction cost levels are set to $\kappa_1 = \kappa_2 = 0\%$. The full state space, as described in Table 4.1, is denoted by $\mathcal{S}$.

4.6.6 Statistical arbitrage

In this analysis, we explore whether our framework can incorporate a speculative layer, such as statistical arbitrage, which capitalizes on the underlying structure of the risk measure that informs the hedging optimization problem.

In line with the definition provided by Assa and Karai (2013) and following studies such as those by Buehler et al. (2021), Horikawa and Nakagawa (2024) and François et al. (2025), we define statistical arbitrage strategies as profit-seeking trading strategies that exploit statistical anomalies in the market. Specifically, we assess whether the difference between the RL

strategies, ϕ^{RL}, and DG strategies, ϕ^{DG}, denoted by

$$\phi^{-} = \phi^{RL} - \phi^{DG}, \tag{4.18}$$

exhibits characteristics of statistical arbitrage with respect to a risk measure ρ by evaluating the condition

$$\rho\left(-V_T^{\phi^{-}}(0)\right) < 0. \tag{4.19}$$

This condition implies that the trading strategy requires zero initial investment and is considered strictly less risky than a null investment according to the risk measure ρ. We aim to investigate whether the trading strategy ϕ^{-} behaves like statistical arbitrage in our framework. Specifically, we examine the difference strategy to determine if RL simply adds a speculative component to the DG strategy, or if there is another underlying mechanism at play. This analysis is conducted with the risk metrics ρ set to $\text{CVaR}_{95\%}$ and SMSE.

Table 4.6 presents the hedging error risk associated with the trading strategy ϕ^{-}, which represents the differential position between the RL and DG strategies. This analysis is conducted across the strategies obtained under different risk measures while hedging an ATM straddle intrument with a maturity of $T = 63$ days.

Table 4.6: Statistical arbitrage statistic

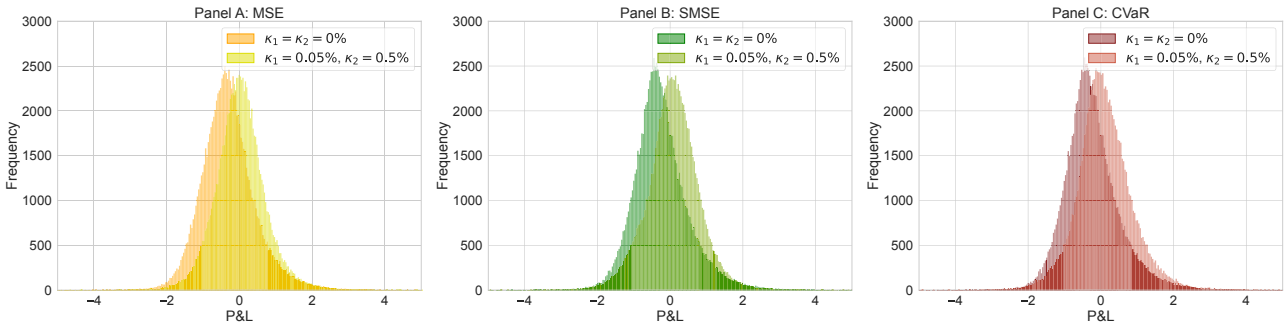
	$\rho\left(-V_T^{\phi^{-}}(0)\right)$				
Risk measure	$\kappa_1 = \kappa_1 = 0\%$	κ_2=0.5%	κ_2=1%	κ_2=1.5%	κ_2=2%
SMSE	1.719	1.597	1.691	1.805	1.882
$\text{CVaR}_{95\%}$	1.721	1.583	1.644	1.782	1.767

Results are computed over 100,000 out-of-sample paths according to the conditions outlined in Section 4.4.3.1. The transaction cost for the underlying asset is set to $\kappa_1 = 0.05\%$.

Our numerical results show no evidence of statistical arbitrage, as all hedging error risks

produce positive values. To further illustrate the absence of arbitrage-like behavior, Figure 4.11 presents the profit and losses (P&L) of the strategy ϕ^- at time T with no initial investment, considering two scenarios: one without transaction costs and another with transaction cost levels set at 0.05% for κ_1 and 0.5% for κ_2. The three panels display distributions that are either symmetric around zero or shifted to the left, indicating the absence of profit-seeking trading strategies. This reinforces the conclusion that the RL strategies within our framework are solely focused on hedging, without introducing speculative overlays.

Figure 4.11: P&L distribution of the strategy ϕ^-.

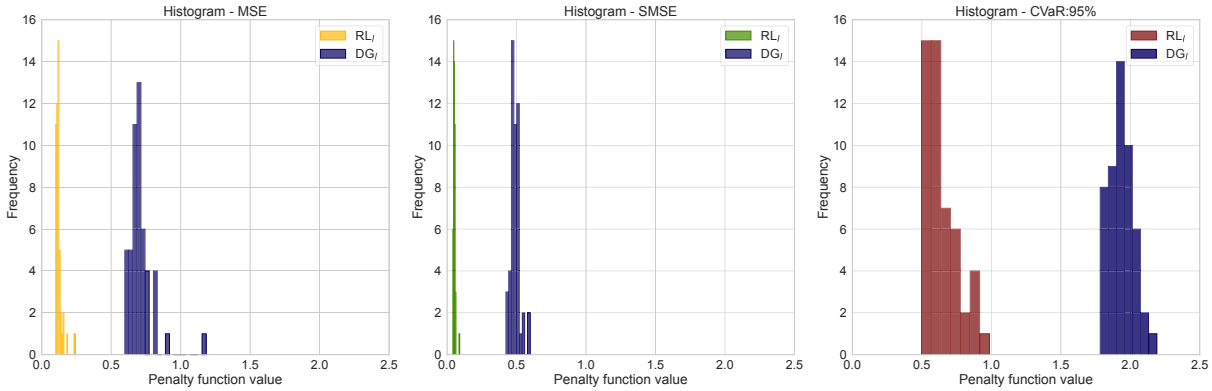


Distributions are computed using 100,000 out-of-sample paths. The P&L is simply defined by the portfolio value $V_T^{\phi^-}(0)$ at maturity.

4.6.7 Systematic outperformance of RL agents

We validate the outperformance of RL agents by hedging a straddle instrument with a maturity of $T = 63$ days, incorporating an ATM call option with a maturity of $T^* = 84$ days as a hedging instrument. In this validation, we analyze the empirical distribution of each penalty function under transaction cost levels set to $\kappa_1 = 0.05\%$ and $\kappa_2 = 0.5\%$ for simplicity. The empirical distributions are derived by bootstrapping the hedging error over 100,000 paths, with batches of size 1,000. As shown in Figure 4.12, the RL approach consistently outperforms the delta gamma strategy, as evidenced by the non-overlapping empirical distributions.

Figure 4.12: Empirical distribution of penalty functions for a straddle with maturity of $T = 63$ days.



Results are computed using bootstrapping with a sample size of 1,000 over 100,000 out-of-sample paths according to the conditions outlined in Section 4.4.3.1 using an ATM call option with maturity $T_1 = 84$ days as the hedging instrument. Transaction cost levels are set to 0.05% for κ_1 and 0.5% for κ_2.

4.6.8 JIVR Model parameters

Table 4.7: Estimated Gaussian copula parameters

	$\epsilon_{t,R}$	$\epsilon_{t,1}$	$\epsilon_{t,2}$	$\epsilon_{t,3}$	$\epsilon_{t,4}$	$\epsilon_{t,5}$
$\epsilon_{t,R}$	1.000					
$\epsilon_{t,1}$	-0.550	1.000				
$\epsilon_{t,2}$	-0.690	0.140	1.000			
$\epsilon_{t,3}$	0.030	-0.030	-0.010	1.000		
$\epsilon_{t,4}$	-0.220	0.250	0.120	0.280	1.000	
$\epsilon_{t,5}$	-0.340	0.170	0.370	0.130	-0.050	1.000

Table 4.8: JIVR model parameter estimates

Parameter	β_1	β_2	β_3	β_4	β_5		S&P500
α	0.000899	0.008400	0.000770	-0.001393	0.000657	λ	2.711279
θ_1	0.996290	-0.013869		0.002841			
θ_2	0.003669	0.877813	0.001300				
θ_3		-0.032640	0.997071	0.003722	-0.004198		
θ_4				0.980269			
θ_5		-0.047789			0.986019		
ν		0.089445					
$\sigma\sqrt{252}$		0.380279	0.052198	0.048641	0.051536		
ω	0.267589						0.977291
κ	0.838220	0.965751	0.974251	0.945377	0.980844		0.888977
a	0.134152	0.098272	0.092646	0.102201	0.100502		0.056087
γ	-0.111813	-1.482862	0.096766	0.060558	-0.102996		2.507796
ζ	0.143760	0.852943	0.029109	-0.159051	0.092664		-0.641306
φ	1.351070	1.538928	2.284780	1.449977	1.428477		2.039669

Conclusion

This thesis investigates the application of deep reinforcement learning (DRL) techniques for hedging financial derivatives in incomplete markets. Across the three papers presented, we extend the deep hedging framework of Buehler et al. (2019) by developing and analyzing trading strategies that enhance hedging performance across various market conditions, risk measures, and hedging instruments.

The first paper introduces a novel deep hedging framework that integrates forward-looking volatility information through a functional representation of the implied volatility surface, combined with conventional historical features. Our implementation employs deep policy gradient methods and a neural network architecture incorporating LSTM cells and feedforward layers, enhancing training efficiency and risk management effectiveness. The results demonstrate that our approach consistently outperforms traditional benchmarks both in the presence and absence of transaction costs. Additionally, global importance analysis confirms that incorporating implied volatility features significantly enhances hedging performance, with key factors such as conditional variance and the long-term at-the-money implied volatility level playing a crucial role in decision-making.

The second paper explores whether deep hedging strategies embed speculative overlays that could lead to statistical arbitrage when compared to conventional delta hedging. Our findings reveal that if the risk measure used in the hedging optimization does not sufficiently penalize losses relative to gains, deep hedging can incorporate speculative elements. However, when

using an appropriate risk measure, such as CVaR with a sufficiently high confidence level, deep hedging strategies do not exhibit statistical arbitrage-like behavior. This highlights the critical role of carefully selecting risk measures to ensure that deep hedging remains a sound risk management strategy. Our analysis further indicates that susceptibility to statistical arbitrage is influenced by factors such as option maturity, moneyness, and economic conditions, reinforcing the importance of proper risk measure selection.

The third paper extends the deep hedging framework to portfolios of options, incorporating multiple hedging instruments and state-dependent no-trade regions to optimize rebalancing frequency in the presence of transaction costs. The results confirm that deep reinforcement learning strategies benefit from the inclusion of additional hedging instruments and that state-dependent no-trade regions improve performance by reducing unnecessary rebalancing. Furthermore, the study demonstrates that deep hedging agents dynamically adjust positions based on both historical variance and market expectations of future volatility, showcasing a nuanced and adaptive approach to risk management.

Overall, this thesis contributes to the advancement of deep reinforcement learning techniques for financial derivative hedging by addressing key aspects such as the integration of implied volatility information, the impact of risk measures on hedging behavior, and the optimization of hedging strategies with multiple instruments and transaction cost considerations. The findings reinforce the viability of deep hedging as a robust and adaptive approach for managing financial risk in incomplete markets.

Bibliography

Alexander, C. and Nogueira, L. M. (2007). Model-free hedge ratios and scale-invariant models. *Journal of Banking & Finance*, 31(6):1839–1861.

Assa, H. and Karai, K. M. (2013). Hedging, Pareto optimality, and good deals. *Journal of Optimization Theory and Applications*, 157:900–917.

Balduzzi, P. and Lynch, A. W. (1999). Transaction costs and predictability: Some utility cost calculations. *Journal of Financial Economics*, 52(1):47–78.

Bates, D. S. (2005). Hedging the smirk. *Finance Research Letters*, 2(4):195–200.

Bazzana, F. and Collini, A. (2020). How does HFT activity impact market volatility and the bid-ask spread after an exogenous shock? An empirical analysis on S&P 500 ETF. *The North American Journal of Economics and Finance*, 54:101240.

Black, F. and Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy*, 81(3):637–654.

Bondarenko, O. (2003). Statistical arbitrage and securities prices. *The Review of Financial Studies*, 16(3):875–919.

Boyle, P. P. and Emanuel, D. (1980). Discretely adjusted option hedges. *Journal of Financial Economics*, 8(3):259–282.

Boyle, P. P. and Vorst, T. (1992). Option replication in discrete time with transaction costs. *The Journal of Finance*, 47(1):271–293.

Buehler, H., Gonon, L., Teichmann, J., and Wood, B. (2019). Deep hedging. *Quantitative Finance*, 19(8):1271–1291.

Buehler, H., Gonon, L., Teichmann, J., Wood, B., Mohan, B., and Kochems, J. (2022). Deep hedging: Hedging derivatives under generic market frictions using reinforcement learning. In *Swiss Finance Institute Research Paper*. [Sl]: SSRN.

Buehler, H., Murray, P., Pakkanen, M. S., and Wood, B. (2021). Deep hedging: Learning risk-neutral implied volatility dynamics. *arXiv preprint arXiv:2103.11948*.

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*, 6:1129370.

Cao, J., Chen, J., Hull, J., and Poulos, Z. (2020). Deep hedging of derivatives using reinforcement learning. *The Journal of Financial Data Science*.

Carbonneau, A. (2021). Deep hedging of long-term financial derivatives. *Insurance: Mathematics and Economics*, 99:327–340.

Carbonneau, A. and Godin, F. (2021). Equal risk pricing of derivatives with deep hedging. *Quantitative Finance*, 21(4):593–608.

Carbonneau, A. and Godin, F. (2023). Deep equal risk pricing of financial derivatives with non-translation invariant risk measures. *Risks*, 11(8):140.

Carr, P. and Wu, L. (2014). Static hedging of standard options. *Journal of Financial Econometrics*, 12(1):3–46.

Cetin, U., Soner, H. M., and Touzi, N. (2010). Option hedging for small investors under liquidity costs. *Finance and Stochastics*, 14:317–341.

Chaudhury, M. (2019). *Option bid-ask spread and liquidity*. SSRN.

Cochrane, J. H. and Saa-Requejo, J. (2000). Beyond arbitrage: Good-deal asset price bounds

in incomplete markets. *Journal of Political Economy*, 108(1):79–119.

Coleman, T. F., Kim, Y., Li, Y., and Patron, M. (2007). Robustly hedging variable annuities with guarantees under jump and volatility risks. *Journal of Risk and Insurance*, 74(2):347–376.

Constantinides, G. M. (1986). Capital market equilibrium with transaction costs. *Journal of Political Economy*, 94(4):842–862.

Covert, I., Lundberg, S. M., and Lee, S.-I. (2020). Understanding global feature contributions with additive importance measures. *Advances in Neural Information Processing Systems*, 33:17212–17223.

Darling, D. A. (1957). The Kolmogorov-Smirnov, Cramér-von Mises tests. *The Annals of Mathematical Statistics*, 28(4):823–838.

Davis, M. H. A. and Norman, A. R. (1990). Portfolio selection with transaction costs. *Mathematics of Operations Research*, 15(4):676–713.

Du, J., Jin, M., Kolm, P. N., Ritter, G., Wang, Y., and Zhang, B. (2020). Deep reinforcement learning for option replication and hedging. *The Journal of Financial Data Science*, 2(4):44–57.

Dumas, B., Fleming, J., and Whaley, R. E. (1998). Implied volatility functions: Empirical tests. *The Journal of Finance*, 53(6):2059–2106.

Edirisinghe, C., Naik, V., and Uppal, R. (1993). Optimal replication of options with transactions costs and trading restrictions. *Journal of Financial and Quantitative Analysis*, 28(1):117–138.

Fecamp, S., Mikael, J., and Warin, X. (2020). Deep learning for discrete-time hedging in incomplete markets. *Journal of Computational Finance*, 25(2).

François, P. and Stentoft, L. (2021). Smile-implied hedging with volatility risk. *Journal of Futures Markets*, 41(8):1220–1240.

François, P., Galarneau-Vincent, R., Gauthier, G., and Godin, F. (2022). Venturing into uncharted territory: An extensible implied volatility surface model. *Journal of Futures Markets*, 42(10):1912–1940.

François, P., Galarneau-Vincent, R., Gauthier, G., and Godin, F. (2023). Joint dynamics for the underlying asset and its implied volatility surface: A new methodology for option risk management. *SSRN*.

François, P., Gauthier, G., and Godin, F. (2014). Optimal hedging when the underlying asset follows a regime-switching Markov process. *European Journal of Operational Research*, 237(1):312–322.

François, P., Gauthier, G., Godin, F., and Mendoza, C. O. P. (2025). Is the difference between deep hedging and delta hedging a statistical arbitrage? *Finance Research Letters*, 73:106590.

François, P., Gauthier, G., Godin, F., and Pérez Mendoza, C. O. (2024). Enhancing deep hedging of options with implied volatility surface feedback information. *Working paper*.

Frey, R. (1998). Perfect option hedging for a large trader. *Finance and Stochastics*, 2:115–141.

Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings.

Glosten, L. R., Jagannathan, R., and Runkle, D. E. (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. *The Journal of Finance*, 48(5):1779–1801.

Godin, F. (2016). Minimizing CVaR in global dynamic hedging with transaction costs. *Quantitative Finance*, 16(3):461–475.

Godin, F. (2019). A closed-form solution for the global quadratic hedging of options under geometric Gaussian random walks. *The Journal of Derivatives*, 26(3):97–107.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT press.

Guéant, O. and Pu, J. (2017). Option pricing and hedging with execution costs and market impact. *Mathematical finance*, 27(3):803–831.

Halperin, I. (2019). The QLBS Q-learner goes NuQLear: Fitted Q iteration, inverse RL, and option portfolios. *Quantitative Finance*, 19(9):1543–1553.

Hambly, B., Xu, R., and Yang, H. (2023). Recent advances in reinforcement learning in finance. *Mathematical Finance*, 33(3):437–503.

Harrison, J. M. and Pliska, S. R. (1981). Martingales and stochastic integrals in the theory of continuous trading. *Stochastic Processes and their Applications*, 11(3):215–260.

Henrotte, P. (1993). Transaction costs and duplication strategies. *Graduate School of Business, Stanford University*.

Hodges, S. D. and Neuberger, A. (1989). Optimal replication of contingent claims under transaction costs. *Review Futures Market*, 8:222–239.

Horikawa, H. and Nakagawa, K. (2024). Relationship between deep hedging and delta hedging: Leveraging a statistical arbitrage strategy. *Finance Research Letters*, page 105101.

Horvath, B., Teichmann, J., and Žurič, Ž. (2021). Deep hedging under rough volatility. *Risks*, 9(7):138.

Imaki, S., Imajo, K., Ito, K., Minami, K., and Nakagawa, K. (2021). No-transaction band network: A neural network architecture for efficient deep hedging. *arXiv preprint arXiv:2103.01775*.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA,*

May 7-9, 2015, Conference Track Proceedings.

Kélani, A. and Quittard-Pinon, F. (2017). Pricing and hedging variable annuities in a lévy market: A risk management perspective. *The Journal of Risk and Insurance*, 84(1):209–238.

Lai, T. L. and Lim, T. W. (2009). Option hedging theory under transaction costs. *Journal of Economic Dynamics and Control*, 33(12):1945–1961.

Leland, H. E. (1985). Option pricing and replication with transactions costs. *The Journal of Finance*, 40(5):1283–1301.

Lütkebohmert, E., Schmidt, T., and Sester, J. (2022). Robust deep hedging. *Quantitative Finance*, 22(8):1465–1480.

Martellini, L. (2000). Efficient option replication in the presence of transactions costs. *Review of Derivatives Research*, 4:107–131.

Marzban, S., Delage, E., and Li, J. Y.-M. (2023a). Deep reinforcement learning for option pricing and hedging under dynamic expectile risk measures. *Quantitative Finance*, 23(10):1411–1430.

Marzban, S., Delage, E., and Li, J. Y.-M. (2023b). Deep reinforcement learning for option pricing and hedging under dynamic expectile risk measures. *Quantitative Finance*, 23(10):1411–1430.

Meindl, P. J. and Primbs, J. A. (2008). Dynamic hedging of single and multi-dimensional options with transaction costs: A generalized utility maximization approach. *Quantitative Finance*, 8(3):299–312.

Melnikov, A. and Smirnov, I. (2012). Dynamic hedging of Conditional Value-at-Risk. *Insurance: Mathematics and Economics*, 51(1):182–190.

Mikkilä, O. and Kanniainen, J. (2023). Empirical deep hedging. *Quantitative Finance*, 23(1):111–122.

Neagu, A., Godin, F., Simard, C., Kosseim, L., et al. (2024). Deep hedging with market impact. *To appear in proceedings of the 37th Canadian Conference on Artificial Intelligence (CAIAC 2024)*.

Peng, X., Zhou, X., Xiao, B., and Wu, Y. (2024). A risk sensitive contract-unified reinforcement learning approach for option hedging.

Pham, H. (2000). Dynamic L^p-hedging in discrete time under cone constraints. *SIAM Journal on Control and Optimization*, 38(3):665–682.

Pickard, R. and Lawryshyn, Y. (2023). Deep reinforcement learning for dynamic stock option hedging: A review. *Mathematics*, 11(24):4943.

Raj, S., Kerenidis, I., Shekhar, A., Wood, B., Dee, J., Chakrabarti, S., Chen, R., Herman, D., Hu, S., Minssen, P., et al. (2023). Quantum deep hedging. *Quantum*, 7:1191.

Rebonato, R. (2005). *Volatility and correlation: The perfect hedger and the fox*. John Wiley & Sons.

Rockafellar, R. T. and Uryasev, S. (2002). Conditional value-at-risk for general loss distributions. *Journal of Banking & Finance*, 26(7):1443–1471.

Rémillard, B. and Rubenthaler, S. (2013). Optimal hedging in discrete time. *Quantitative Finance*, 13(6):819–825.

Schweizer, M. (1995). Variance-optimal hedging in discrete time. *Mathematics of Operations Research*, 20(1):1–32.

Su, X. and Li, Y. (2024). Robust portfolio selection with subjective risk aversion under dependence uncertainty. *Economic Modelling*, 132:106667.

Toft, K. B. (1996). On the mean-variance tradeoff in option replication with transactions costs. *Journal of Financial and Quantitative Analysis*, 31(2):233–263.

Warde-Farley, D., Goodfellow, I. J., Courville, A., and Bengio, Y. (2013). An empirical

analysis of dropout in piecewise linear networks. *arXiv preprint arXiv:1312.6197*.

Wu, D. and Jaimungal, S. (2023). Robust risk-aware option hedging. *Applied Mathematical Finance*, 30(3):153–174.

Zakamouline, V. (2009). The best hedging strategy in the presence of transaction costs. *International Journal of Theoretical and Applied Finance*, 12(06):833–860.