Higher-Order Greeks Hedging of American Options with Deep Reinforcement Learning

Alexander Tsoskounoglou[a], Sven Karbach[a,b,1], Drona Kandhai[a,b,1] and Michel Vellekoop[c]

[a]Informatics Institute, University of Amsterdam, LAB42, Science Park 900, Amsterdam, 1098 XH, The Netherlands
[b]Korteweg-de Vries Institute for Mathematics, University of Amsterdam, Science Park 105–107, Amsterdam, 1098 XG, The Netherlands
[c]Faculty of Economics and Business, University of Amsterdam, Section Quantitative Economics, Roetersstraat 11, Amsterdam, 1001 NJ, The Netherlands

ARTICLE INFO

Keywords:
Deep Reinforcement Learning
American Options
Heston Model
Transaction Costs
Hedging
PDE pricing

ABSTRACT

This paper studies a hedging framework for American options that addresses the practical challenges posed by early exercise, transaction costs, and higher-order risk exposure. We combine PDE-based option pricing under the Heston stochastic volatility model using the modified Craig–Sneyd scheme with a distributional deep reinforcement learning (DRL) approach to learn adaptive hedging strategies. Unlike traditional methods, our DRL agent optimizes directly for tail-risk objectives, including Value-at-Risk and Conditional Value-at-Risk, while accounting for transaction cost constraints and limited re-balancing frequency. Through extensive numerical experiments, we demonstrate that the proposed DRL framework consistently outperforms Delta-Gamma and Delta-Vega hedging, particularly in high-cost regimes where complete higher-order neutrality is impractical. Specifically, our method reduces VaR by approximately 17.1% compared to Delta-Gamma strategies and up to 39.4% relative to Delta-Vega approaches. Moreover, the learned strategies remain robust under misspecification of key Heston parameters such as the asset-volatility correlation. The robustness results support the practical applicability of our approach, enabling effective higher-order risk management even in the presence of model uncertainty.

1. Introduction

Hedging American options presents substantial challenges due to early exercise features, transaction costs and frequent rebalancing requirements. These complexities further increase when looking at portfolios of American options or the sequential issuance of option liabilities. In volatile markets, managing higher-order Greeks, particularly Gamma and Vega, is essential for further reducing risk, as Delta hedging alone is insufficient. Traditional methods for pricing American options based on binomial trees [8] or finite difference methods [3] have proven effective under both constant and stochastic volatility models, including the Heston framework [10]. However, when directly applied to hedging, these approaches typically assume frictionless markets and continuous trading. In practice, this leads to unrealistic strategies characterized by excessive rebalancing and elevated transaction costs, while still leaving significant tail-risk exposures unaddressed. Consequently, practical hedging demands cost-aware adaptive methods that dynamically balance risk exposure and trading efficiency.

Recent advancements in *Deep Reinforcement Learning* (DRL) have shown to address the complexities of hedging under realistic market conditions effectively [15, 13]. Unlike classical methods, DRL-based strategies explicitly incorporate market frictions during policy training, enabling adaptive hedging that balances risk minimization, in particular the tail risk measured via Value-at-Risk (VaR) or Conditional VaR (CVaR), against transaction costs. However, most existing DRL literature focuses on European-style options [7, 6, 5], where early exercise is not permitted, and the associated option evaluation is comparatively simpler. Robust DRL frameworks specifically tailored for American-style options, particularly put options, which are frequently exercised early, remain scarce [16]. Moreover, existing studies often concentrate solely on Delta hedging or rely on simplified asset dynamics. As a result, the accurate and cost-aware management of higher-order Greeks under realistic conditions, including transaction costs and stochastic volatility, has not been comprehensively addressed in the literature [4, 19, 14, 9].

1.1. Main Contributions

To address these gaps, this paper proposes an advanced DRL framework specifically designed for dynamic hedging of American options under realistic market conditions, including stochastic volatility and transaction costs. The principal contributions of this paper are:

- **Distributional DRL Algorithm:** We extend the D4PG-QR algorithm [6] to optimize higher-order Greeks hedging policies for American options explicitly with respect to tail-risk metrics such as VaR and CVaR, and transaction costs, offering improved risk management over standard mean-based DRL approaches.
- **Detailed Benchmarking and Robustness Analysis:** The proposed DRL policies are rigorously benchmarked against classical Delta, Delta-Gamma, and Delta-Vega strategies across varying transaction cost regimes and models. Moreover, we demonstrate the robustness of the learned hedging strategy to model parameter misspecifications in the Heston stochastic volatility model.

[1]The authors acknowledge support from the AI4FinTech initiative at the University of Amsterdam.

contact: alexander.tsoskounoglou@gmail.com, sven@karbach.org, b.d.kandhai@uva.nl, m.h.vellekoop@uva.nl

In addition, the paper makes two computational contributions:

- **Accurate Heston Simulation under QE Scheme:** In contrast to prior work relying on simplified volatility dynamics, we employ the *robust Quadratic-Exponential* scheme [1, 17] to simulate asset paths in the Heston stochastic volatility model. This approach preserves the essential distributional characteristics of the underlying Cox–Ingersoll–Ross (CIR) process, ensuring that volatility paths remain strictly positive and follow realistic trajectories. This is crucial for accurate use of calibrated Heston parameters, as naïve schemes like Euler-Maruyama introduce biases which results in inconsistencies between calibration and simulations.
- **Integration of Advanced PDE Pricing Methods:** We incorporate the *Modified Craig–Sneyd Alternating Direction Implicit* method [12] for efficient and accurate pricing of American options in the Heston framework. This method yields reliable Greek sensitivities required for effective higher-order hedging.

In this paper, we conduct extensive numerical experiments and demonstrate that the proposed DRL framework significantly outperforms classical Greek-based hedging strategies, improving both risk reduction and cost efficiency, especially so in high transaction cost regimes. In particular, our DRL approach reduces VaR by approximately 17.1% compared to Delta-Gamma hedging and up to 39.4% relative to Delta-Vega hedging, showing the effectiveness of adaptive, risk-aware policy learning (see Table 1 and Table 3 below).

1.2. Structure of the Paper

The remainder of this article is organized as follows. Section 2 formalize the hedging problem, details the stochastic models, option–pricing schemes, and the distributional DRL algorithm. Section 3 presents the main empirical results obtained under a realistic market regime with positive interest rates and negative market price of vol-risk, benchmarking the learned strategy against Delta, Delta–Gamma, and Delta–Vega hedges. Complementary experiments carried out in a drift-free setting are reported in Section 4 to isolate the influence of early exercise incentives. Section 5 then assesses the robustness of the policy to misspecification of Heston parameters. Finally, Section 6 summarizes the key findings, and Section 7 outlines avenues for future research.

2. Hedging Problem and DRL Framework

The core objective addressed in this paper is the development of an optimal dynamic hedging strategy for a portfolio of American-style put options, with particular emphasis on managing higher-order risks, specifically Gamma and Vega. The hedging instruments to do so are the underlying asset, as well as additional American put options written on the same underlying asset.

Focusing on the higher-order aspect of the portfolio, we employ a Deep Reinforcement Learning framework that directly optimizes the sequence of hedging decisions. The DRL agent learns the optimal dynamic policy throughout the option's life-cycle including the initial hedge position α_0 at time zero. The market dynamics are simulated with a stochastic process of choice, using a set of parameters, which are assumed to remain constant during the hedging period. More precisely, the underlying asset price process $(S_t)_{t\geq 0}$ and its instantaneous variance process $(v_t)_{t\geq 0}$ are simulated under either a stochastic volatility model, in our case the Heston model, or the Geometric Brownian Motion (GBM) as a baseline. At each discrete re-balancing interval, the agent determines the optimal quantity of hedging instruments to minimize risk exposure measured by a risk metric such as Value-at-Risk over the full hedging horizon. Option prices and Greek sensitivities are computed at each time step based on the simulated market state (S_t, v_t), in addition to the fixed set of initially calibrated model parameters.

2.1. Modeling the Underlying (Physical Measure)

In this study, we employ two primary models to simulate (underlying) asset price paths: the classical Black–Scholes (BS) model based on a GBM and the Heston stochastic volatility model.

Black–Scholes Model Under the physical measure $\mathbb{P}$ the GBM dynamics are

$$\mathrm{d}S_t = S_t\,(r+\mu-q)\,\mathrm{d}t + S_t\,\sigma\,\mathrm{d}W_t, \quad t>0,$$

where σ is the constant volatility, r is the risk-free rate, q is the continuous dividend yield, μ represents the excess mean rate of return on equity under $\mathbb{P}$, and $(W_t)_{t\geq 0}$ is a standard Wiener process under $\mathbb{P}$.

Heston Model The Heston stochastic volatility model reads

$$\begin{aligned}\mathrm{d}S_t &= S_t\,(r+\mu-q)\,\mathrm{d}t + S_t\sqrt{v_t}\,\mathrm{d}W_t^S, \quad t>0,\\ \mathrm{d}v_t &= \kappa_{\mathbb{P}}\big(\theta_{\mathbb{P}}-v_t\big)\,\mathrm{d}t + \xi\sqrt{v_t}\,\mathrm{d}W_t^v, \quad t>0,\end{aligned}$$

where $(W_t^S)_{t\geq 0}$ and $(W_t^v)_{t\geq 0}$ are standard Wiener processes under $\mathbb{P}$ with correlation ρ, i.e., $\mathrm{d}\langle W^S, W^v\rangle_t = \rho\,\mathrm{d}t$. The parameters ξ, $\kappa_{\mathbb{P}}$, and $\theta_{\mathbb{P}}$ denote the vol-of-vol, the speed of mean reversion, and the long-run mean of the variance process, respectively.

Simulation Methods. For the BS model we sample paths from its closed form solution. In contrast, Heston model paths are simulated using the robust Quadratic-Exponential (QE) scheme [1, 17], which is well-known for its numerical stability and its ability to preserve key distributional properties of the Heston model, including accurate moment matching of the CIR stochastic volatility process.

2.2. American Option Pricing

The valuation of American put options in the BS model is done via the binomial tree approach [8] using a 200-step tree and the valuation of American put options under the Heston model is performed using the modified Craig–Sneyd

finite-difference scheme [12, 11]. This choice is justified due to its numerical stability, efficiency, and the accurate computation of the options sensitivities (Greeks), which is a crucial property for our higher-order hedging set-up. The scheme numerically solves the following partial differential inequality subject to an early exercise condition:

$$\frac{\partial U}{\partial t} + \frac{1}{2}\upsilon S^2 \frac{\partial^2 U}{\partial S^2} + \rho\xi S\upsilon \frac{\partial^2 U}{\partial S \partial \upsilon} + \frac{1}{2}\xi^2 \upsilon \frac{\partial^2 U}{\partial \upsilon^2} + (r-q)S\frac{\partial U}{\partial S} + \kappa_{\mathbb{Q}}(\theta_{\mathbb{Q}} - \upsilon)\frac{\partial U}{\partial \upsilon} - rU \leq 0,$$

and such that $U \geq (K - S)^+$. In our implementation, a 50×25 grid in the stock and volatility dimensions provides a reasonable trade-off between computational cost and pricing accuracy, which allows us to simulate a high number of option prices required for the training and evaluating of the DRL agent (see [18] for more details).

2.3. DRL Framework for Hedging

The hedging task is formulated as a *Markov Decision Process* and is solved through a distributional deep-reinforcement learning algorithm. In particular, we define the state s_t, action a_t, and reward $R(s_t, a_t)$ at time t, which we use to manage risk exposure under market frictions, i.e. the exposure to the two Greeks Gamma and Vega under transaction costs. Delta hedging is assumed to be frictionless and Δ is absent from the state space s_t. In practice, the net delta of a call–put combination is typically small, as the call's positive delta largely offsets the put's negative delta.

State Space. The state at time t is defined by the vector,

$$s_t = \{S_t,\ \Gamma^{\text{Tot}},\ \mathcal{V}_t^{\text{Tot}},\ \Gamma_t^{\text{Hedg}},\ \mathcal{V}_t^{\text{Hedg}}\},$$

where:

- S_t is the current price of the underlying asset,
- Γ_t^{Tot} and $\mathcal{V}_t^{\text{Tot}}$ denote the total aggregated Gamma and Vega exposure of the portfolio at time t, respectively. The total portfolio consists of the Delta portfolio, the liability portfolio, and the hedging portfolio,
- Γ_t^{Hedg} and $\mathcal{V}_t^{\text{Hedg}}$ represent the Gamma and Vega of the available hedging option, which has 30 days to maturity and strike $K = S_t$ at time t.

Note that the instantaneous variance process υ_t does not appear explicitly in the state representation, but enters implicitly through the Gamma and Vega sensitivities.

Action Space. We model a continuous action space $\mathcal{A} := [X_{\min}, X_{\max}]$. The action $a_t \in \mathcal{A}$ denotes the number of option shares to purchase at time t for hedging purposes. These options are written on the same underlying asset as the rest of the portfolio, but can have different maturity from it. By dynamically adjusting the hedge position, the agent aims to reduce the portfolio's Gamma and Vega exposures, achieving either partial or full hedging. The bounds $X_{\min}$ and $X_{\max}$ correspond to hedging 0% and 100% of the Gamma or Vega risk. Specifically, $X_{\min}$ is set based on the lower of the Gamma or Vega exposure, while $X_{\max}$ reflects the higher of the two.

Reward Function. The per-step reward is defined as:

$$R(s_t, a_t) = \text{PnL}_{t\to t+\Delta t}^{\text{Tot}} - \text{TC}_t, \tag{1}$$

where $\text{PnL}_{t\to t+\Delta t}^{\text{Tot}}$ denotes the profit and loss of the total portfolio over the interval $[t, t+\Delta t]$, and TC_t represents the transaction costs incurred at time t. The reward thus explicitly captures both portfolio profit/loss and trading costs. The DRL agent learns to optimize the cumulative reward over time, with a focus not only on expected return, but also on tail-risk management. To this end, we employ a distributional reinforcement learning approach based on quantile regression [6, 7], enabling direct optimization of tail-risk objectives such as VaR and CVaR and Mean-Std.

Policy and Discount Factor. The hedging policy $\pi(a \mid s)$ is a conditional distribution that assigns to every observed market state $s_t \in \mathcal{S}$ a feasible hedge adjustment $a_t \in \mathcal{A}$. Given such policy, the agent generates a path which contains a set of states and decisions $\{(s_t, a_t)\}_{t=0}^{T-1}$ and stores a reward stream $R(s_t, a_t)$ defined previously. For any coherent objective (risk) function $\mathcal{G}$, the optimisation problem becomes

$$\pi^* = \arg\max_{\pi}\ \mathcal{G}\Big[\sum_{t=1}^{T} \gamma^t R(s_t, a_t)\Big], \qquad a_t \sim \pi(\cdot \mid s_{t-1}).$$

For example, under a $\text{VaR}_{95\%}$ criterion this becomes $\mathcal{G} = \mathcal{F}_{0.95}^{-1}$. The discount factor $\gamma \in (0, 1]$ controls how strongly rewards realised later in the hedge horizon influence the risk measure; we set $\gamma = 1$, treating profit-and-loss at every step up to the $T = 30$-day maturity with equal importance.

Objective Functions We evaluate risk using three commonly applied tail-sensitive metrics, each serving as a potential training and evaluation objective for the DRL agent. Let PnL denote the cumulative profit and loss of the total portfolio. The objective functions are as follows:

$$\begin{aligned}\text{Mean-Std} &= \mathbb{E}[\text{PnL}_t] - 1.645 \cdot \text{Std}[\text{PnL}_t]\\ \text{VaR}_{95\%} &= \mathcal{F}_{0.05}^{-1}[\text{PnL}_t]\\ \text{CVaR}_{95\%} &= \mathbb{E}[\text{PnL}_t \mid \text{PnL}_t \leq \text{VaR}_{95\%}]\end{aligned}$$

Here, Mean-Std penalizes variability around the expected return symmetrically by subtracting a multiple of the standard deviation. In a normal distribution a z score of -1.645 corresponds to the 95th quantile, and as shown in Figure 1 in our case is very close to the 95th as well. $\text{VaR}_{95\%}$ targets the 95th percentile of the PnL distribution, identifying extreme losses with a fixed quantile cutoff. $\text{CVaR}_{95\%}$ measures the expected loss conditional on exceeding the VaR threshold, thus providing a deeper assessment of tail risk.

2.4. Distributional Actor–Critic Algorithm

We employ the D4PG-QR algorithm [2, 6], which integrates continuous control with distributional reinforcement learning. By modeling the full return distribution via quantile regression, this approach enables direct optimization of

contact: alexander.tsoskounoglou@gmail.com, sven@karbach.org, b.d.kandhai@uva.nl, m.h.vellekoop@uva.nl

VaR and CVaR, outperforming standard mean-based reinforcement learning methods.

- **Actor network**: outputs the continuous hedge adjustment a_t given the current state s_t.
- **Critic network**: estimates the distribution of future returns using quantile regression, enabling explicit optimization of risk-sensitive objectives.

Training Procedure. To create the simulated hedging environment, first, the underlying asset paths are generated (using either the Quadratic-Exponential scheme for the Heston model or the closed-form solution for the BS). Second, for each time step, PDE-based option prices and Greeks are computed for the option portfolios. Third, the DRL agent learns optimal hedging through the following iterative procedure:

1. Transitions (s_t, a_t, r_t) are collected and stored in a replay buffer.
2. The critic network is updated off-policy by minimizing the distributional Bellman error via quantile regression.
3. The actor network is updated using the policy gradient, guided by feedback from the distributional critic.

Evaluation and Robustness Testing. The trained DRL policy is evaluated out-of-sample on 8,000 episodes, which are simulated from the underlying models, to assess performance and generalization. Specifically, we test:

- Hedging performance relative to classical Delta, Delta-Gamma, and Delta-Vega strategies under various transaction cost regimes and objective functions.
- Robustness under misspecification of Heston model parameters: initial variance v_0, long-term variance $\theta_{\mathbb{P}}$, mean reversion speed $\kappa_{\mathbb{P}}$, vol-of-vol ξ and correlation ρ.

3. Experiments and Results

The framework required approximately 150,000 CPU hours for development and testing. Each Heston stochastic volatility scenario (e.g., 2% transaction cost optimized with $\text{CVaR}_{95\%}$ objective) consumed around 1,440 CPU hours for simulation and training.

Transitioning from a constant volatility (BS) environment to the Heston model notably increased computational requirements, necessitating a doubling of training episodes from 32,000 to 65,000 and scaling parallelization from a single DRL agent to roughly 21 agents. This scaling addressed variability in hedging performance across training runs (see Figure 2).

The option valuation step required approximately 10 million PDE-based option calculations in total across all simulated paths, underscoring the computational intensity inherent in large-scale, finite-difference-based simulations for precise option pricing and Greeks estimation.

Although further optimizations could significantly speed up both valuation and agent training, such improvements remain beyond the present scope, as this study emphasizes analyzing results derived from the proposed methodology.

3.1. Experiment Setup

The experiments investigate hedging strategies for American put options under stochastic conditions. Specifically, we consider:

- An initial American put option with a time to maturity (TTM) of 30 days.
- Subsequent American put option arrivals, each with a 60-day TTM and an at-the-money strike price ($K = S_t$), following a Poisson process with intensity $\lambda = 1$ and equal probability (50%) of being either short or long positions.
- Transaction costs represented by bid-ask spreads for option trading set at three levels: $\text{spread}_t \in \{0.5\%, 1\%, 2\%\}$ $\forall t = 0, \dots, T$ for buying/selling 30-day American put options, where we note again that Δ-hedging is assumed to be frictionless.

We adopt the Heston stochastic volatility model with the following set of reasonable (potentially market calibrated) parameters using a positive risk-free rate and equity drift, also incorporating a volatility risk premium of $\lambda_v = 1.5$, and

$$\begin{array}{llll} r = 2\% & \mu = 3\% & q = 0\% & \kappa_{\mathbb{P}} = 1.0 \\ \theta_{\mathbb{P}} = 0.30^2 & \xi = 0.3 & \rho = -0.7 & v_0 = 0.30^2 \end{array}$$

The mean rate of return for S_t is $r + \mu - q = 5\%$ and the market price of volatility risk introduces a linear adjustment to the Heston dynamics, modifying the mean-reversion term as follows:

$$\kappa_{\mathbb{P}}\big(\theta_{\mathbb{P}} - v_t\big)\,\mathrm{d}t = \kappa_{\mathbb{Q}}\big(\theta_{\mathbb{Q}} - v_t\big)\,\mathrm{d}t - \lambda_v\, v_t\,\mathrm{d}t,$$

$$\kappa_{\mathbb{Q}} = \kappa_{\mathbb{P}} + \lambda_v, \quad \theta_{\mathbb{Q}} = \frac{\kappa_{\mathbb{P}}\,\theta_{\mathbb{P}}}{\kappa_{\mathbb{P}} + \lambda_v}\,.$$

Incorporating a positive risk-free-rate into our framework makes the early exercise feature of an American put option economically advantageous, in contrast to scenarios where $r = q = 0\%$, in which the early exercise is not advantageous. Our goal is to perform our analysis for scenarios where hedging American options poses additional challenges. In contrast, for the GBM benchmark we set the volatility to $\sigma = 0.30$ and take $r = \mu = q = 0$, matching the specification in Cao [6]. This yields a like-for-like comparison and serves as a validation of our DRL implementation.

3.2. Black-Scholes Results

In this constant-volatility framework, we compare three hedging strategies: the standard *Delta* hedge, a *Delta–Gamma* hedge that enforces full Gamma neutrality at each rebalancing step, and a *DRL* policy trained on the same GBM paths. Table 1 presents tail-risk metrics (Mean–Std, VaR95, and CVaR95) along with the expected trading costs for the RL agent, under three different bid–ask spreads. Since volatility is constant, there is no Vega risk, and Gamma exposure becomes the sole second-order concern. As a

Table 1
Results under Black-Scholes model (constant volatility case)

Obj. Function	Objective Function Value Delta	Delta-Gamma	RL	$\mathbb{E}[TC^{RL}]$
0.5% Transaction Cost				
Mean-Std	26.62	5.49	**5.43**	2.92
VaR95	24.71	5.51	**5.38**	2.88
CVaR95	39.29	6.65	**6.33**	3.06
1% Transaction Cost				
Mean-Std	26.62	9.73	**8.65**	4.94
VaR95	24.71	9.92	**8.60**	4.42
CVaR95	39.29	11.26	**10.22**	4.64
2% Transaction Cost				
Mean-Std	26.62	18.61	**13.69**	6.09
VaR95	24.71	18.93	**14.08**	6.09
CVaR95	39.29	20.87	**16.83**	6.76

Hedging performance under constant volatility using the BS model. This table compares Delta, Delta-Gamma, and DRL-based (D4PG-QR) strategies across transaction cost levels of 0.5%, 1%, and 2%. Delta hedging is assumed to be frictionless, thus risk values remain constant through different transaction costs. The Delta-Gamma strategy fully neutralizes Gamma each step and incurs expected transaction costs of 3.14, 6.85, and 13.7, respectively. The DRL agent adaptively hedges approximately 80%, 60%, and 30% of Gamma exposure, dynamically balancing cost and risk.All results are averaged over 8,000 simulated hedging episodes.

result, this setting isolates the Delta-Gamma hedging trade-off.

The results demonstrate that DRL learns a stable trade-off between cost and risk. In high-cost environments, it selectively tolerates residual Gamma exposure to avoid unnecessary rebalancing. For 0.5% transaction costs on average the RL agent reduces 2.75% the risk metric, for 1% it reduces it for 11.22% and for 2% it reduces it by 23.81%. Those reductions showcase the effectiveness of using an adaptive hedging methodology, even under constant volatility, where vega exposure is zero. As shown in Table 1, the RL agent reduces the gamma hedging intensity as transaction costs increase which is consistent with real trading heuristics: hedge less when it's expensive, and accept moderate exposure when full neutrality becomes inefficient. Our GBM results for American options are aligned with Cao's [6] GBM results for European options.

3.3. Heston Results

Learning an optimal hedging policy under stochastic volatility is substantially more computationally challenging, as the agent must learn to manage both Gamma and Vega exposures. This increases the dimensionality of the state space and introduces greater complexity in the underlying market dynamics and hedging decisions. The added difficulty is reflected in the comparison between DRL-based strategies under constant volatility (Table 1) and stochastic volatility (Table 2), where the presence of volatility uncertainty leads to higher residual risk and transaction costs.

The key objectives of our study are as follows: (i) benchmark the performance of DRL-based hedging policies against classical Greeks-based strategies (*Delta*, *Delta-Gamma*, and *Delta-Vega*); (ii) examine how increasing transaction costs influence hedging intensity and the trade-off between cost and risk reduction; and (iii) quantify the added learning complexity introduced by stochastic volatility and its effect on the agent's ability to develop efficient hedging strategies.

Comparison to Baselines. The Delta, Delta-Gamma, and Delta-Vega strategies represent classical hedging approaches in which the corresponding Greek exposures are fully neutralized at each time step. As Delta hedging is assumed to be frictionless, the expected transaction costs for the Delta-Gamma and Delta-Vega strategies arise solely from purchasing additional options to hedge Gamma and Vega exposures.

In Table 2 both Delta-Gamma and Delta-Vega strategies achieve better risk metrics than pure Delta hedging, confirming the effectiveness of higher-order risk management. Most importantly, under stochastic volatility the DRL-based strategy consistently outperforms all classical methodologies across all transaction cost levels and metrics. Unlike static strategies, the DRL agent dynamically adjusts the percentage of Gamma and Vega exposures hedged over time, optimizing the trade-off between risk reduction and incurring transaction costs. Among the baselines, Delta-Vega consistently performs the worst, as fully hedging Vega exposure with shorter maturity options is very inefficient. The DRL policy learns to adaptively reduce hedging intensity as transaction costs increase to prevent higher transaction costs incurred from increasing tail metrics.

Effect of Increasing Spread. Although transaction cost levels (spreads) increase linearly from 0.5% to 1% and 2%, the expected transaction costs incurred by the RL agent do not rise proportionally. This non-linear increase indicates that the DRL policy adapts by reducing hedging intensity, specifically, lowering Gamma and Vega hedge ratios, to control costs.

In high-cost environments, fully neutralizing higher-order Greeks becomes prohibitively expensive and leads to suboptimal outcomes. The RL agent learns to balance this trade-off by selectively hedging when the increased costs result in a reduction of the objective function (risk). As excessive trading can amplify tail risk, the DRL approach yields more favorable overall risk-cost profiles under high transaction cost regimes.

3.3.1. Hedging Dynamics Visualization

The hedging behavior of the DRL agent under the Heston model is visualized in Figure 1, which depicts results for a 30-day American put option with sequential option arrivals, optimized using a $\text{VaR}_{95\%}$ objective and a 2% transaction cost spread. Subplots (1,1) and (2,1) illustrate how the agent

Table 2
Hedging Results under Stochastic Volatility (Heston) with $r=2\%$ and mean rate of return 5%

Objective Function	Delta	Objective Function Value for Delta-Gamma	Delta-Vega	RL	RL Gamma Hedge Ratio	RL Vega Hedge Ratio	Expected RL Transaction Cost
0.5% Transaction Cost							
Mean-Std[2]	55.81	32.02	40.89	**25.93**			
VaR95	54.54	31.68	38.93	**25.12**	0.40	0.30	3.44
CVaR95[2]	77.17	44.48	57.41	**35.61**			
1% Transaction Cost							
Mean-Std[2]	55.81	35.22	46.86	**29.43**			
VaR95	54.54	35.03	44.79	**28.55**	0.26	0.24	5.61
CVaR95[2]	77.17	47.90	64.02	**38.68**			
2% Transaction Cost							
Mean-Std[2]	55.81	41.74	59.11	**34.41**			
VaR95	54.54	41.96	57.43	**34.78**	0.24	0.18	9.59
CVaR95[2]	77.17	54.83	77.47	**45.61**			

Out-of-sample hedging performance under the physical measure $\mathbb{P}$ (Heston model, $r = 2\%$, mean rate of return 5%, volatility risk premium $\lambda_v = 1.5$). Columns report the risk metric (**Mean–Std**, **VaR95**, **CVaR95**) for the four strategies *Delta*, *Delta–Gamma*, *Delta–Vega* and *DRL*, evaluated at bid–ask spreads of 0.5%, 1% and 2%. The final three columns give the RL agent's average Gamma-hedge ratio, Vega-hedge ratio and expected transaction costs. For reference, the rule-based benchmarks incur expected costs of Delta−Gamma = (3.13, 6.26, 12.52) and Delta−Vega = (5.81, 11.61, 23.23). Results are averaged over 8 000 simulated episodes.

dynamically reduces Gamma and Vega exposures over time, while subplots (1,2) and (2,2) show the respective Greek contributions from each portfolio component (if results for the 0.5% transaction cost scenario were plotted, they would show more aggressive Gamma and Vega hedging behavior).

Note that the liability portfolio receives randomly arriving options with 60-day time-to-maturity (TTM), whereas the hedging instruments have a fixed 30-day TTM. As a result, the Gamma exposure of the hedge portfolio tends to exceed that of the liabilities, while the opposite holds for Vega. This asymmetry introduces additional complexity into the optimization problem and underscores the flexibility of the DRL agent in adapting its hedging policy to account for non-stationary risk exposures and instrument mismatches.

In addition, subplot (3,1) displays the average PnL of each constituent portfolio, illustrating how the liability and hedge components can exhibit substantial fluctuations that largely offset one another. Subplots (3,2) and (4,1) reveal a near-linear relationship between transaction cost and portfolio risk, indicating that the DRL strategy maintains stable cost-risk trade-offs over the hedging horizon. Subplot (4,2) shows the cumulative PnL distribution at maturity T, highlighting how each optimization objective, Mean-Std, VaR, and CVaR, corresponds to a different part of the PnL distribution.

All together, the visualizations provide compelling evidence of the DRL agent's ability to perform dynamic, cost-aware hedging under stochastic volatility and sequential portfolio construction, adapting effectively to evolving market and risk conditions.

3.3.2. Training Stability and Agent Selection

Figure 2 shows training and validation results for 21 DRL agents trained under the Heston model with a 2% transaction cost and VaR_{95} objective. Between agents there is identical architecture, hyperparameters, and training data, with performance variaablity being attributed only to the NNs weights initialization seed. Most agents converge to stable, effective hedging strategies that outperform Delta-Gamma across all risk metrics; with only 1/21 failing to learn an optimal control policy.

The variability in the outcome stems from sensitivity to weight initialization and the inherent stochasticity of policy search and gradient descent in high-dimensional spaces. To reduce uncertainty in the learned policy, we trained multiple agents and selected the best-performing policy based on out of sample results. While most trained agents consistently outperform classical benchmarks, ensuring consistent convergence to the best case, remains a nontrivial challenge.

In practice, robust policy convergence is as critical as model architecture. Thus, future work may mitigate variability in outcome by employing warm-starts, evolutionary policies, and adaptive scheduling during training.

4. Experiments under Zero Mean Rate of Return for Equity

In this section, we conduct a series of controlled experiments where the risk-free rate, equity drift, and equity risk premium are all set to zero, i.e., $r = q = \mu = \lambda_v = 0\%$. Under these assumptions, the early exercise feature of an American put option lacks economic incentive. Table 3

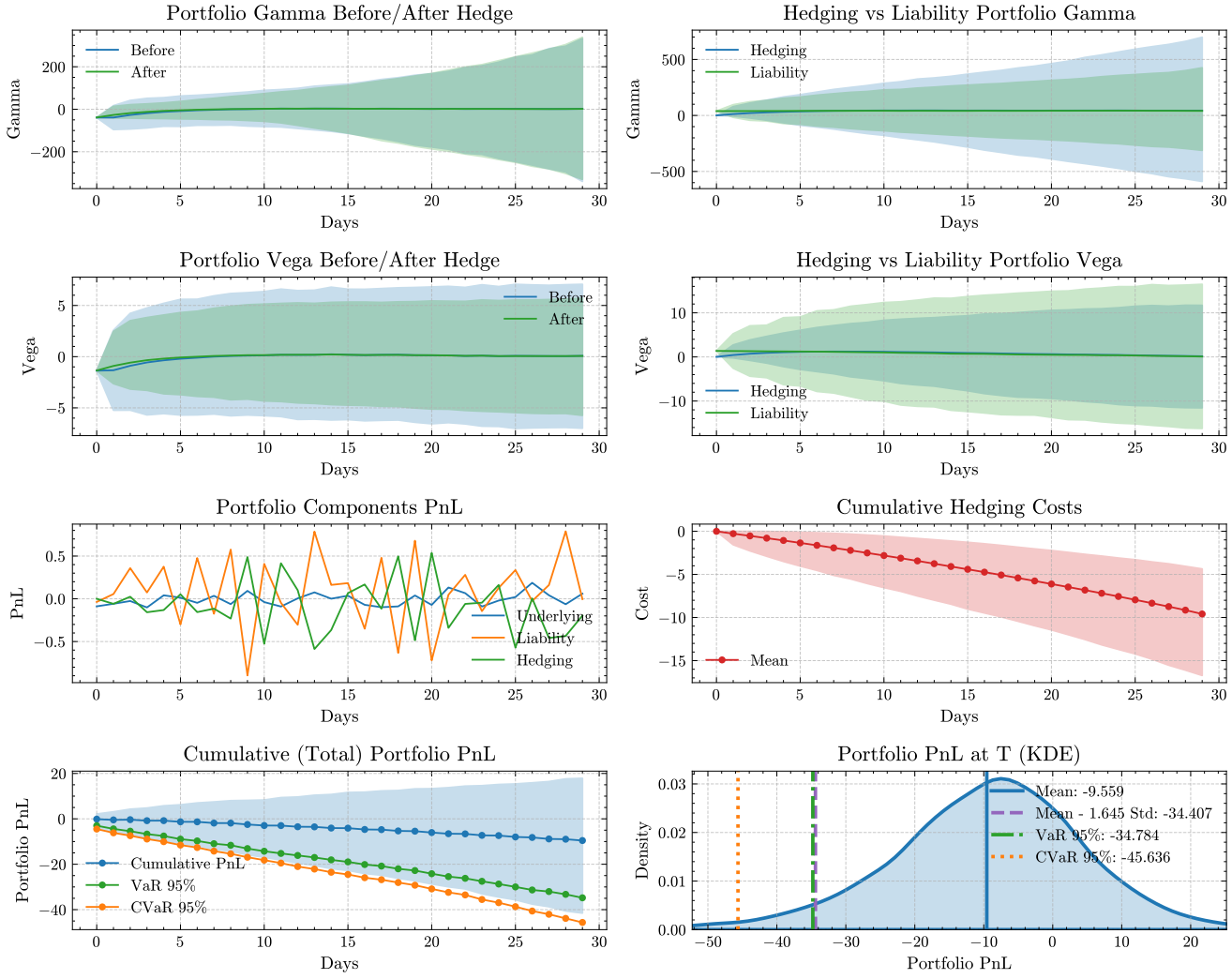

Figure 1: Visualization of DRL hedging dynamics under the Heston model for a 30-day American put option, optimized for $\text{VaR}_{95\%}$ with a 2% transaction cost under P. The agent responds to sequential option arrivals (60-day maturity) and adjusts a 30-day hedging option accordingly. Subplots (1,1) and (2,1) show the evolution of Gamma and Vega exposures, while (1,2) and (2,2) break these exposures down into liability and hedge components. Subplots (3,1) and (3,2) show average PnL evolution and its relation to transaction cost. Subplot (4,1) relates cost to risk, and (4,2) compares final PnL distributions under different optimization objectives. Shaded areas denote 95% interquartile ranges over 8,000 episodes.

reports the results under these assumptions, effectively isolating the valuation dynamics from drift and the market price of volatility risk.

This setup enables a focused assessment of the early exercise component's influence on the performance of the reinforcement learning (RL) algorithm. It also facilitates an indirect comparison with the findings of Cao [6], albeit with the use of a different stochastic volatility model (Cao employs SABR with $r = q = \mu = 0\%$) and numerical method used for option evaluations.

In Table 3 the RL-based approach continues to outperform rule-based strategies, having slightly worse risk metrics consistently across transaction costs. Additionally, Delta-Vega hedging exhibits the largest relative deterioration in risk, with increases of 7.92% in VaR95 under 2% transaction costs relative to Table 2. Pure Delta hedging benefits modestly from positive drift, whereas Delta-Gamma hedging shows negligible differences.

Furthermore, the RL strategy under positive drift and negative λ exhibits higher vega hedging percentages in Table 2 versus Table 3, coresponding to the increased volatility risk which a negative λ introduces. Lastly, RL also realizes cost efficiencies, with expected transaction costs reduced by approximately 8–9% relative to the neutral-drift case. The results of Table 2 versus Table 3 highlight the robustness of the DRL framework across different drift regimes.

Choosing an Objective Function. In this section, we train DRL agents with all three objective functions. Based on those results, Table 4 bellow reveals that the choice of objective function does not significantly impact the risk

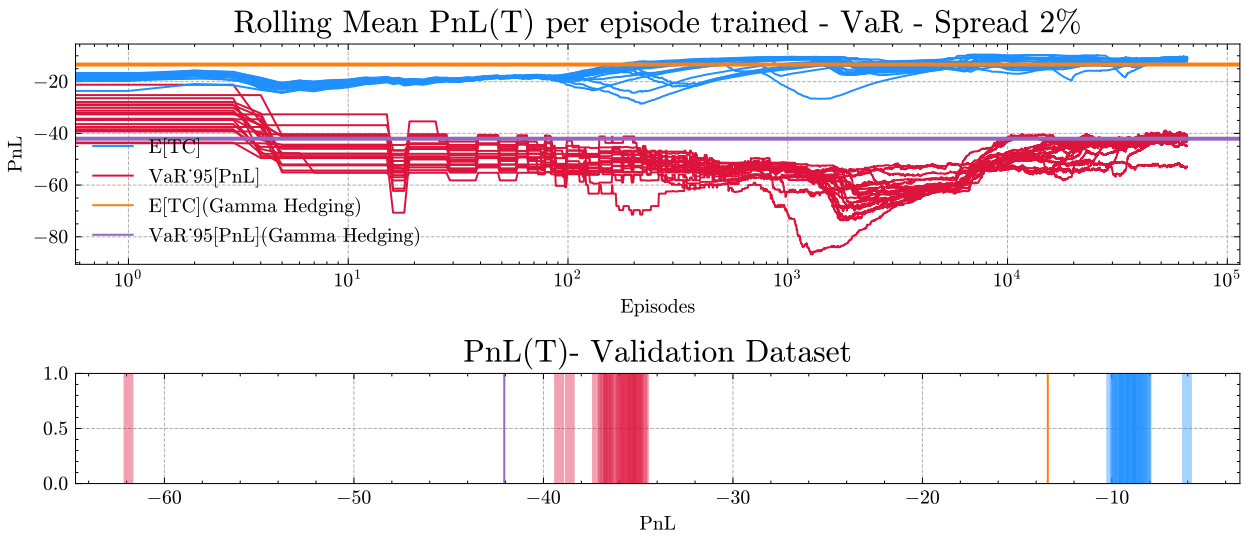

Figure 2: Training and validation performance across 21 independently initialized DRL agents trained under the Heston model using a $\text{VaR}_{95\%}$ objective with a 2% transaction cost. The top panel shows the rolling average PnL and $\text{VaR}_{95\%}$ during training; the bottom panel shows the corresponding validation performance over 8000 episodes. Each line represents one agent initialized with a different random seed, trained and evaluated on the same dataset. The wide dispersion in final outcomes highlights the sensitivity of DRL performance to initialization, underscoring the need for multiple training runs and agent selection in high-dimensional hedging problems.

Table 3
Hedging Results under Heston's stochastic volatility model (with zero drift)

Objective Function	Delta	Objective Function Value for Delta-Gamma	Delta-Vega	RL	RL Gamma Hedge Ratio	RL Vega Hedge Ratio	Expected RL Transaction Cost
0.5% Transaction Cost							
Mean-Std	56.54	31.72	43.10	**26.51**	0.43	0.26	3.81
VaR95	55.93	31.02	41.22	**25.57**	0.46	0.25	3.57
CVaR95	80.67	44.01	60.38	**35.41**	0.45	0.26	3.72
1% Transaction Cost							
Mean-Std	56.54	35.15	49.51	**29.55**	0.32	0.19	5.76
VaR95	55.93	34.72	47.96	**29.00**	0.36	0.20	5.87
CVaR95	80.67	47.74	67.45	**40.36**	0.38	0.17	5.91
2% Transaction Cost							
Mean-Std	56.54	42.15	62.66	**34.77**	0.29	0.18	9.88
VaR95	55.93	42.09	61.98	**34.84**	0.23	0.16	8.78
CVaR95[3]	80.67	55.32	81.83	**45.43**	0.23	0.16	8.78

Comparison of hedging strategies under stochastic volatility using the Heston model. Risk and cost outcomes are reported for classical Delta, Delta-Gamma, Delta-Vega, and DRL-based hedging across three transaction cost levels (0.5%, 1%, 2%). DRL policies are trained using the D4PG-QR algorithm with Mean-Std, $\text{VaR}_{95\%}$ or $\text{CVaR}_{95\%}$ objectives. For each strategy and objective, we report the corresponding loss metric value. The final three columns summarize the DRL agent's average Gamma hedge ratio, Vega hedge ratio, and expected transaction cost over 8,000 evaluation paths. Gamma/Vega hedge ratios denote the average proportion of risk neutralized across the hedging horizon. The expected transaction costs for Delta-Gamma are 3.35, 6.69, and 13.39 for the respective transaction cost regimes; for Delta-Vega, the values are 6.23, 12.45, and 24.92. Delta hedging is assumed to be frictionless, thus risk values remain constant through different transaction costs regimes.

metrics which the agent was not directly optimized for. Thus, in our experiments agents trained to minimize VaR95 also achieved strong performance in terms of CVaR95 and Mean-Std, and vice versa. This observation is consistent with Figure 1, subplot (4,2), which shows that all three optimization objectives effectively targeting neighboring regions in the resulting PnL distribution (specially in the case of Mean-Std and VaR95). Consequently, to reduce training costs in the more demanding stochastic volatility setting with early-exercise features, we restricted training to the VaR95 objective and only report those results in Table 2.

Table 5
Our Results compared to Cao's [6] results

	Our Results (Heston, Amer. Put)			Cao's Results (SABR, Euro. Call)		
Objective Function	**Delta-Gamma Value**	**RL Value**	**% Reduction**	**Delta-Gamma Value**	**RL Value**	**% Reduction**
0.5% Transaction Cost						
Mean-Std	31.72	26.51	**16.43%**	19.46	17.76	**8.74%**
VaR95	31.02	25.57	**17.57%**	19.23	19.31	**-0.42%**
CVaR95	44.01	35.41	**19.54%**	27.53	26.06	**5.34%**
1% Transaction Cost						
Mean-Std	35.15	29.55	**15.93%**	23.06	20.03	**13.13%**
VaR95	34.72	29.00	**16.48%**	23.02	20.22	**12.16%**
CVaR95	47.74	40.36	**15.46%**	31.55	26.81	**15.03%**
2% Transaction Cost						
Mean-Std	42.15	34.77	**17.51%**	30.51	24.17	**20.77%**
VaR95	42.09	34.84	**17.24%**	30.67	23.85	**22.23%**
CVaR95[3]	55.32	45.43	**17.88%**	39.79	31.57	**20.67%**

Comparison of DRL hedging results for American puts under the Heston model (our work) versus European calls under the SABR model as presented by Cao [6]. The percentage reduction is computed as $\big(\text{Delta-Gamma} - \text{RL}\big)/\text{Delta-Gamma} \times 100\%$, and reflects the improvement in the objective function value achieved by the DRL policy relative to the Delta-Gamma baseline.

Table 4
Performance of 1%-transaction costs-trained agents across risk metrics

Objective F.	**Mean–Std**	**VaR$_{95\%}$**	**CVaR$_{95\%}$**	$\mathbb{E}$[TC]
Mean–Std	29.55	29.17	39.92	5.76
VaR95	29.27	29.00	38.92	5.87
CVaR95	30.07	29.74	40.36	6.16

These results suggest that the DRL policy exhibits robustness across different risk metrics, even when trained with respect to only one. Nevertheless, VaR emerges as a particularly practical objective. It is less sensitive to extreme outliers during training than CVaR. Which could explain why VaR optimized agent outperforms the CVaR optimized agent in the CVaR metric in Table 4. Additionally, the VaR objective function corresponds to a fixed quantile of the PnL distribution, offering a more stable and interpretable target compared to the Mean-Std objective.

4.1. Comparison to Cao's [6] Results

Table 5 compares our DRL hedging results for American put options under Heston dynamics with zero drift and market price vol-premium, with those reported by Cao [6], who studied European call options under the SABR model and also zero drift and zero market price vol-premium. Despite differences in option type, stochastic model, and hedging setup, the relative improvements achieved by the DRL agent are comparable in magnitude across many objective functions and transaction cost levels. With the biggest difference being that our method achieves consistent reductions in risk across all objectives and transaction costs, where in Cao's case the RL's risk reduction increase with transaction costs. This could be attributed to the different stochastic model of choice or/and the option valuation methodology.

5. Robustness to Model Misspecification

In real-world applications, Heston model parameters are rarely calibrated with perfect accuracy, resulting in discrepancies between the assumed model dynamics and the true behavior of the market. This section evaluates the robustness of the deep reinforcement learning hedging policy under such parameter mismatches. Where the realized market conditions (parameters) vary from the initial assumed (calibrated) ones. For these experiments we assumed $r = \mu = q = \lambda_v = 0$ and parameters $\kappa_{\mathbb{P}}, \theta_{\mathbb{P}}, \xi, \rho$ are assumed to be constant for the duration of the simulation.

Robustness is a critical requirement for practical deployment, as it ensures that the learned policy maintains effective risk mitigation even when model assumptions are violated. A robust strategy should generalize well to moderate shifts in key parameters, avoiding catastrophic performance degradation due to calibration errors or regime shifts in volatility dynamics.

As shown in Table 6 and Figure 3 in general small shifts in Heston parameters lead to small differences in the value of CVaR_{95} indicating that the DRL hedging strategy generally maintains its expected risk metrics and transaction costs. This robustness is particularly evident for perturbations in the parameters $\{\theta_{\mathbb{P}}, \rho, \kappa_{\mathbb{P}}\}$, where the impact on risk and cost remains modest. In contrast, shifts in $\{v_0, \xi\}$ lead to more

[2] CVaR95 and Mean-Std Metrics are reported based on the performance of an agent optimized with the VaR95 objective function.

Table 6
Robustness testing over different bumped parameters

Parameter	CVaR_{95}	transaction costs	Observations
Initial Volatility (v_0)			
$v_0 = 0.1$	29.92	6.24	Reduced risk, reduced cost.
$v_0 = 0.2$	39.79	9.01	Reduced risk, reduced cost.
$v_0 = 0.3$	45.15	9.37	–
$v_0 = 0.4$	60.09	16.28	Larger risk, larger cost.
$v_0 = 0.5$	72.44	24.84	Larger risk, larger cost.
Volatility-of-Volatility (ξ)			
$\xi = 0.1$	33.65	9.75	Reduced risk, larger cost.
$\xi = 0.2$	40.12	9.51	Reduced risk, similar cost.
$\xi = 0.3$	45.15	9.37	–
$\xi = 0.4$	59.15	9.38	Larger risk, similar cost.
$\xi = 0.5$	68.98	9.46	Larger risk, similar cost.
Mean Reversion Speed ($\kappa_{\mathbb{P}}$)			
$\kappa_{\mathbb{P}} = 0.5$	51.71	9.38	Larger risk, similar cost.
$\kappa_{\mathbb{P}} = 0.8$	50.18	9.38	Larger risk, similar cost.
$\kappa_{\mathbb{P}} = 1.0$	45.15	9.37	–
$\kappa_{\mathbb{P}} = 1.2$	48.35	9.38	Larger risk, similar cost.
$\kappa_{\mathbb{P}} = 1.5$	47.05	9.38	Larger risk, similar cost.
Correlation (ρ)			
$\rho = -0.9$	47.65	9.03	Larger risk, reduced cost.
$\rho = -0.8$	48.48	9.21	Larger risk, similar cost.
$\rho = -0.7$	45.15	9.37	–
$\rho = -0.6$	49.72	9.54	Larger risk, similar cost.
$\rho = -0.5$	50.16	9.68	Larger risk, similar cost.
Long Term Variance ($\theta_{\mathbb{P}}$)			
$\theta_{\mathbb{P}} = 0.03$	49.49	9.31	Larger risk, similar cost.
$\theta_{\mathbb{P}} = 0.06$	49.34	9.35	Larger risk, similar cost.
$\theta_{\mathbb{P}} = 0.09$	45.15	9.37	–
$\theta_{\mathbb{P}} = 0.12$	49.18	9.43	Larger risk, similar cost.
$\theta_{\mathbb{P}} = 0.15$	49.17	9.51	Larger risk, similar cost.

Robustness evaluation of DRL hedging performance under Heston model parameter misspecification. Each row shows the resulting $\text{CVaR}_{95\%}$ and expected transaction cost (transaction costs) when a single parameter is perturbed from its baseline value: $\{v_0 = 0.3,\ \xi = 0.3,\ \kappa_{\mathbb{P}} = 1.0,\ \rho = -0.7,\ \theta_{\mathbb{P}} = 0.09\}$. The third column provides qualitative observations. Results are averaged over 8,000 episodes using the same DRL policy trained under baseline parameters with a 2% transaction cost and $\text{CVaR}_{95\%}$ objective. The 95% confidence intervals for baseline $\text{CVaR}_{95\%}$ and transaction costs are 1.59 and 0.36, respectively.

pronounced changes in both the risk and cost profiles, indicating greater sensitivity to initial variance v_0 and volatility-of-volatility ξ.

Most notably, initial variance v_0 exhibits a monotonic relationship with cost: higher than expected v_0 increases both expected costs and tail risk, and lower than expected v_0 results to lower costs and tail risk. One potential approach would be to use variance swaps or volatility futures to reduce

[3]The CVaR_{95} metric is from the VaR_{95}-optimized agent, which outperforms the CVaR_{95}-optimized agent on its own metric, as shown in Table 4.

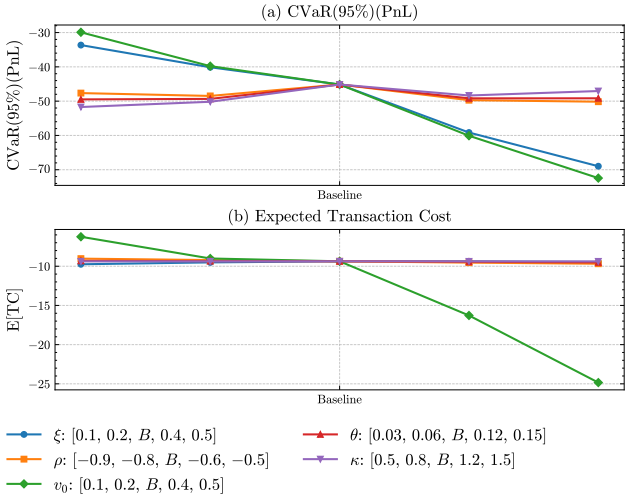

Figure 3: Robustness Testing on Heston model: Out of sample evaluation of the $\text{CVaR}_{95\%}$, (2% transaction costs) trained agent on bumped parameter values

the sensitivity to changes in realized volatility. Table 6 showcases the significance of correctly estimating v_0 and ξ at the time of hedge initiation.

6. Conclusion

We proposed a deep reinforcement learning based hedging framework for managing Gamma and Vega exposure in American put options under stochastic volatility. The approach combines the Quadratic-Exponential scheme for accurate simulation of Heston paths with the Modified Craig–Sneyd finite difference scheme for precise valuation of American options, and a cutting-edge DRL framework (D4PG-QR) for option hedging. To the best of our knowledge, this is the first study to address higher-order hedging of American options within the Heston framework using DRL.

Our results demonstrate that the DRL-based strategy consistently achieves lower risk and transaction costs compared to fully hedged Gamma/Vega strategies. By selectively hedging higher-order exposures, the agent dynamically optimizes the trade-off between tail-risk reduction and cost efficiency.

We further assess the method's robustness to parameter misspecification and observe that the learned policy demonstrates stable performance, reinforcing its practical viability in realistic market conditions with imperfect calibration. Additionally, under the physical measure $\mathbb{P}$ and a positive risk-free rate, which incentivizes optimal early exercise for American options in specific scenarios, the RL agent achieves a slightly lower risk profile at equivalent or lower hedging costs. This suggests that the RL agent offers robust risk and cost reduction across different drift regimes for the underlying and the risk-free-rate.

contact: alexander.tsoskounoglou@gmail.com, sven@karbach.org, b.d.kandhai@uva.nl, m.h.vellekoop@uva.nl

An analysis of the training process in Section 3.3.2 reveals sensitivity to the NNs weights initialization, highlighting the need for subsequent work on ensuring robust optimization with a single agent, but with the current setup, multiple training runs are required to select the best-performing agent.

Finally, we provide a comparison with the work of Cao [6]. Despite differences in modeling assumptions, we observe a somewhat comparable relative reduction in risk, reinforcing the effectiveness of a DRL (D4PG-QR) hedging approach.

7. Future Research Directions

While our results demonstrate the effectiveness of DRL for higher-order hedging under stochastic volatility, several extensions remain open for exploration.

First, instead of starting with a single option, future work could start from a given portfolio of options, e.g. calls, puts, European, American, swaps and futures across various strikes and maturities. This would better reflect real-world trading environments and increase policy applicability.

Second, expanding the action space to include additional hedging instruments, such as variance swaps or options with different maturities and strikes, could enhance the agent's ability to control complex sensitivities, including higher-order Greeks.

Third, more realistic modeling of transaction costs; varying by strike, maturity, and market conditions; would increase exactness. Similarly, simulating the arrival of new liabilities using historical trade data would align training environments more closely with institutional settings.

Fourth, incorporating jump-diffusion models would allow the framework to better capture short-term price shocks and improve hedging accuracy for short-dated options. Additionally, modeling parameter uncertainty during training (e.g., randomizing Heston parameters) could improve robustness to market regime changes.

From a computational perspective, improvements in PDE interpolation schemes could significantly reduce evaluation overhead. This includes employing Chebyshev interpolation or structured nearest-neighbor grids to accelerate price and Greek lookups without compromising accuracy.

Lastly, algorithmic performance could benefit from more tuned DRL architectures. Enhancements might include integrating additional higher-order Greeks into the state space (Charm $\left(\frac{\partial \Delta}{\partial t}\right)$, Veta $\left(\frac{\partial \mathcal{V}}{\partial t}\right)$, Color $\left(\frac{\partial \Gamma}{\partial t}\right)$), employing scheduling techniques for learning rate decay, or using better initialization strategies for the NNs in the actor-critic NNs. Such enhancements are more geared to increase the quality and the speed of training DRL agents.

Code and Data Availability

This study relies exclusively on simulated data. The source code and scripts used to generate the results are publicly available under github.com/ithakis/DRL-Hedging-AmericanOptions.

References

[1] Andersen, L.B., 2007. Efficient simulation of the Heston stochastic volatility model. Available at SSRN 946405 .

[2] Barth-Maron, G., Hoffman, M.W., Budden, D., Dabney, W., Horgan, D., Tb, D., Muldal, A., Heess, N., Lillicrap, T., 2018. Distributed distributional deterministic policy gradients. arXiv preprint arXiv:1804.08617 .

[3] Brennan, M.J., Schwartz, E.S., 1978. Finite difference methods and jump processes arising in the pricing of contingent claims: A synthesis. The Journal of Financial and Quantitative Analysis 13, 461–474. URL: http://www.jstor.org/stable/2330152.

[4] Bühler, H., Gonon, L., Teichmann, J., Wood, B., 2019. Deep hedging. Quantitative Finance 19, 1271–1291. URL: https://arxiv.org/abs/1802.03042, doi:10.1080/14697688.2019.1571683.

[5] Cannelli, L., Nuti, G., Sala, M., Szehr, O., 2023. Hedging using reinforcement learning: Contextual k-armed bandit versus q-learning. The Journal of Finance and Data Science 9, 100101.

[6] Cao, J., Chen, J., Farghadani, S., Hull, J., Poulos, Z., Wang, Z., Yuan, J., 2023. Gamma and vega hedging using deep distributional reinforcement learning. Frontiers in Artificial Intelligence 6, 1129370.

[7] Cao, J., Chen, J., Hull, J., Poulos, Z., 2020. Deep hedging of derivatives using reinforcement learning. The Journal of Financial Data Science 3, 10–27. URL: http://dx.doi.org/10.3905/jfds.2020.1.052, doi:10.3905/jfds.2020.1.052.

[8] Cox, J.C., Ross, S.A., Rubinstein, M., 1979. Option pricing: A simplified approach. Journal of financial Economics 7, 229–263.

[9] Giurca, A., Borovkova, S., 2021. Delta hedging of derivatives using deep reinforcement learning. SSRN Electronic Journal URL: https://ssrn.com/abstract=3847272, doi:10.2139/ssrn.3847272.

[10] Heston, S.L., 1993. A closed-form solution for options with stochastic volatility with applications to bond and currency options. The review of financial studies 6, 327–343.

[11] in 't Hout, K.J., Foulon, S., 2011. ADI finite difference schemes for option pricing in the Heston model with correlation. URL: https://arxiv.org/abs/0811.3427, arXiv:0811.3427.

[12] In't Hout, K., Welfert, B., 2009. Unconditional stability of second-order ADI schemes applied to multi-dimensional diffusion equations with mixed derivative terms. Applied Numerical Mathematics 59, 677–692.

[13] Mikkilä, O., Kanniainen, J., 2023. Empirical deep hedging. Quantitative Finance 23, 111–122.

[14] Murray, P., Wood, B., Bühler, H., Wiese, M., Pakkanen, M., 2022. Deep hedging: Continuous reinforcement learning for hedging of general portfolios across multiple risk aversions, in: Proceedings of the Third ACM International Conference on AI in Finance, ACM. pp. 1–9. URL: https://dl.acm.org/doi/10.1145/3533271.3561731, doi:10.1145/3533271.3561731.

[15] Pickard, R., Lawryshyn, Y., 2023. Deep reinforcement learning for dynamic stock option hedging: A review. Mathematics 11. URL: https://www.mdpi.com/2227-7390/11/24/4943, doi:10.3390/math11244943.

[16] Pickard, R., Wredenhagen, F., Jesus, J.D., Schlener, M., Lawryshyn, Y., 2024. Hedging American put options with deep reinforcement learning. URL: https://arxiv.org/abs/2405.06774, arXiv:2405.06774.

[17] Tsoskounoglou, A., 2024. Simulating the Heston model with the Quadratic Exponential scheme. URL: https://medium.com/@alexander.tsoskounoglou/simulating-the-heston-model-with-quadratic-exponential-50cf2b1366b5. accessed: 2025-02-28.

[18] Tsoskounoglou, A., 2025. Dynamic Higher-Order Hedging with DRL: Risk and Cost Optimization for American Options. Master's thesis. University of Amsterdam. URL: https://dspace.uba.uva.nl/server/api/core/bitstreams/11c39c60-0f3e-41fc-a85a-561cef3184c7/content. available online.

[19] Zheng, L., Liu, J., Wang, Y., Zhang, J., Li, X., 2023. Option dynamic hedging using reinforcement learning. Journal of Financial Data Science 5, 44–57. URL: https://jfds.pm-research.com/content/5/2/44, doi:10.3905/jfds.2022.1.097.

contact: alexander.tsoskounoglou@gmail.com, sven@karbach.org, b.d.kandhai@uva.nl, m.h.vellekoop@uva.nl