EX-DRL: Hedging Against Heavy Losses with EXtreme Distributional Reinforcement Learning

Parvin Malekzadeh
University of Toronto
Canada
p.malekzadeh@mail.utoronto.ca

Zissis Poulos
York University
Canada
zpoulos@yorku.ca

Jacky Chen
University of Toronto
Canada
jackyjc.chen@rotman.utoronto.ca

Zeyu Wang
University of Toronto
Canada
colinzeyu.wang@utoronto.ca

Konstantinos N Plataniotis
University of Toronto
Canada
kostas@ece.utoronto.ca

Abstract

Recent advancements in Distributional Reinforcement Learning (DRL) for modeling loss distributions have shown promise in developing hedging strategies in derivatives markets. A common approach in DRL involves learning the quantiles of loss distributions at specified levels using Quantile Regression (QR). This method is particularly effective in option hedging due to its direct quantile-based risk assessment, such as Value at Risk (VaR) and Conditional Value at Risk (CVaR). However, these risk measures depend on the accurate estimation of extreme quantiles in the loss distribution's tail, which can be imprecise in QR-based DRL due to the rarity and extremity of tail data, as highlighted in the literature. To address this issue, we propose EXtreme DRL (EX-DRL), which enhances extreme quantile prediction by modeling the tail of the loss distribution with a Generalized Pareto Distribution (GPD). This method introduces supplementary data to mitigate the scarcity of extreme quantile observations, thereby improving estimation accuracy through QR. Comprehensive experiments on gamma hedging options demonstrate that EX-DRL improves existing QR-based models by providing more precise estimates of extreme quantiles, thereby improving the computation and reliability of risk metrics for complex financial risk management. The implementation is available here.

CCS Concepts

• **Computing methodologies → Reinforcement learning**; • **Applied computing → Economics**.

Keywords

Distributional Reinforcement Learning, Generalized Pareto Distribution, Option Hedging

ACM Reference Format:
Parvin Malekzadeh, Zissis Poulos, Jacky Chen, Zeyu Wang, and Konstantinos N Plataniotis. 2024. EX-DRL: Hedging Against Heavy Losses with EXtreme Distributional Reinforcement Learning. In *5th ACM International Conference on AI in Finance (ICAIF '24), November 14–17, 2024, Brooklyn, NY, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3677052.3698668

1 Introduction

Option trading desks, especially in high-volume environments like banks, must engage in frequent (e.g., daily) hedging to manage risk and maintain profitability. Specifically, traders try to minimize the sensitivity of their option portfolios to sharp or broad movements in the underlying stock price, as well as to the stock's volatility. These sensitivities, known as the 'Greeks,' when reduced or eliminated, incur costs and directly affect the expected value and distribution of losses from hedging over time. Ideally, traders aim for minimal expected losses and narrower loss distributions, indicating reduced risk exposure. However, the typical hedging strategies employed are 'myopic'—they make decisions based on current conditions without considering future implications, leading to suboptimal risk management, particularly in the presence of trading frictions.

To alleviate this problem, a recent thread of research has proposed an alternative take by exploring the application of Reinforcement Learning (RL) to option hedging [7, 8, 10, 30, 37], since RL can efficiently learn long-horizon strategies in a stochastic environment [25, 27]. More recent advancements, as demonstrated in [9], show that Distributional RL (DRL) [3, 12]—which models the entire loss (negative reward) distribution for every potential hedging state and action—can surpass classical RL strategies for European options, which typically focuses on expected loss as a minimization criterion.

Quantile Regression (QR) is typically used in DRL algorithms to approximate the loss distribution. Specifically, the loss is modeled as a quantile distribution, represented by a uniform mixture of Dirac deltas at locations corresponding to quantile values. These quantile values are derived using the QR loss function, which is applied within the framework of the distributional Bellman equation. QR is conceptually straightforward and directly assesses risk measures such as Value-at-Risk (VaR) and Conditional Value-at-Risk (CVaR) through the quantile-based representation it creates. However, as highlighted in several studies [2, 4, 31], the accuracy

ICAIF '24, November 14–17, 2024, Brooklyn, NY, USA

ACM ISBN 979-8-4007-1081-0/24/11
https://doi.org/10.1145/3677052.3698668

of these quantile value estimates, especially those representing extreme values, relies on observations from the distribution's tails. These observations are infrequent yet carry high severity, making the estimation of extreme quantile values inaccurate and prone to under-coverage bias compared to the desired coverage level in practice [2]. Consequently, sparse data in the extreme tail ends of the distribution render the resulting risk measures unreliable.

To enhance the ability of DRL to model the tails of loss distributions effectively while maintaining scalability and fidelity in representing the rest of the distribution, this paper presents a hybrid model. In this model, the tail of the loss distribution is separately modeled using a Generalized Pareto Distribution (GPD), recognized for its efficiency in capturing extreme values in time series analysis [4, 31, 35], while the main body of the distribution remains a quantile distribution. Termed Extreme-DRL (EX-DRL), our proposed approach is compatible with any QR-based DRL method and substantially enhances the estimation of extreme quantiles, which is crucial for hedging portfolios of options. In our experiments, we combine our hybrid model with QR-Distributed Distributional Deterministic Policy Gradients (QR-D4PG) [9], a recently introduced QR-based DRL algorithm. We refer to the proposed algorithm as EX-D4PG and compare its performance to that of QR-D4PG [9]. The findings across a wide range of market scenarios are:

- EX-D4PG learns a better policy by improved estimates of VaR and CVaR, highlighting the agent's enhanced capability to minimize extreme losses and strengthen risk management.
- EX-D4PG achieves a better Profit and Loss (PnL) distribution with a higher mean and a narrower spread than QR-D4PG.
- The hedging policy of the EX-D4PG agent demonstrates increased sensitivity to changes in the volatility of the underlying assets, enabling more responsive adjustments.

The rest of the paper is organized as follows. Section 2 discusses related work in RL-based option hedging and GPD-based extreme modelling. Section 3 provides background on QR-based DRL and GPDs. Section 4 explains how DRL can be used for option hedging. Finally, Section 5 describes the proposed EX-DRL framework in detail, and Section 6 presents empirical results.

2 Related Work

The concept of using RL as an effective optimization framework in financial applications dates back to 2009 [20]. However, it was recently that a broader array of RL-based option hedging strategies began to emerge [7, 10, 13, 15, 16, 19, 28, 30, 37], leveraging advances in deep learning and modern RL algorithms. In terms of the optimization mechanism, these methods vary: some [16] adopt value-based algorithms that directly optimize the expected loss, others [13, 37] directly optimize the hedging policy using policy-based methods, while the majority [10, 15, 19, 30] implement actor-critic methods that optimize both the hedging policy and the expected loss. These diverse approaches have solidified RL as a compelling alternative to traditional hedging strategies. A critical design decision in these methods is the selection of the objective function, which significantly influences the reward function. Common choices include VaR or a convex risk measure such as CVaR and dynamic expectiles. The selection among these objective functions and optimization mechanisms presents distinct trade-offs in terms of scalability, optimality, and time-consistency.

While RL has brought improvements in several domains within risk management, risk-averse formulations based on DRL have exhibited superior scalability in multi-option hedging and have empirically outperformed traditional Greek-neutral hedging strategies under various volatility scenarios [9, 34]. In DRL, the entire loss distribution is estimated using a categorical [3], normal [24], or quantile representation combined with QR [12]. Quantile representation provides the flexibility to stochastically adjust the quantile values at desired levels and has been successfully applied towards learning risk-averse policies in domains such as Atari games and sports [21–23, 26]. However, these applications typically do not encounter events of extreme risk, which are prevalent in option hedging. Consequently, while DRL approaches provide a broad framework to incorporate various risk-return measures, they tend to yield inaccuracies in the tails of the return distribution. This, in turn, affects the accuracy of extreme quantile estimates, including VaR and CVaR, which are of paramount importance in financial risk management.

In the context of modeling extreme values, GPD from extreme value theory [33], offers a way to describe the asymptotic behavior of the distribution of data extremes. A large body of research has leveraged the benefits of GPD to model extreme events in time series and financial analysis [1, 4, 17, 32, 38]. However, its applicability in the context of DRL-based hedging remains under-explored. Complementary to our framework is the formulation by [14]. Both methods utilize GPDs to model the tails of distributions in DRL. However, in [14], the methodology relies on Monte Carlo simulations to estimate the loss and finite difference approximations to compute gradients, which is a suitable approach for low-dimensional policy spaces. Our method, on the other hand, uses deep neural networks to parameterize the policy and improve it via distributional Bellman updates, making it suitable for multi-option hedging scenarios involving large action spaces (thus high-dimensional policy spaces).

3 Background

In this section, we provide a brief background on DRL, focusing on QR-based algorithms, and then discuss GPDs before turning to our specific use case of hedging derivatives portfolios.

3.1 Distributional Reinforcement Learning

We consider the problem of sequential decision-making of an agent acting in an environment over discrete time steps t within a time horizon T, i.e., $t \in \{1, 2, ..., T\}$. We formulate this as a Markov Decision Process (MDP) specified by $(\mathcal{S}, \mathcal{A}, P_\mathcal{S}, R, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $P_\mathcal{S}(.|s, a)$ is the Markovian transition function that assigns to each state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$ the probability of reaching the next state $s' \in \mathcal{S}$, $R(s, a)$ is the reward function, and $0 < \gamma < 1$ denotes the discount factor. The policy of an agent is formulated as $a = \pi(s)$, which maps state s to action a and we assume to be deterministic.

In DRL, the goal is to learn the return distribution, which is the distribution of the random variable return $Z^\pi(s, a) = \sum_{t=1}^{T} \gamma^{t-1} R(s_t, a_t)$, where $s_1 = s$, $a_1 = a$, $s_t \sim P_\mathcal{S}(.|s_{t-1}, a_{t-1})$, and $a_t = \pi(s_t)$. It has been shown that the distribution of Z^{π} can be updated using the distributional Bellman equation [3] as:

$$Z^{\pi}(s,a) \overset{\text{D}}{=} T^{\pi}Z(s,a) \overset{\text{D}}{:=} R(s,a) + \gamma Z^{\pi}(s',a'), \tag{1}$$

where $\overset{\text{D}}{=}$ denotes equality in distribution, $s' \sim P_S(.|s,a)$, and $a' = \pi(s')$. In the rest of the paper, we simplify notation by using R instead of $R(s,a)$.

QR-based DRL: One of the popular methods to model and learn the distribution of return $Z^{\pi}(s,a)$ in DRL is the QR framework, which was first proposed by [12]. QR methods approximate the distribution of $Z^{\pi}(s,a)$ using a discrete set of N quantile values $\{\theta_w^{\tau_n}(s,a)\}_{n=0}^{N-1}$, parameterized with w, at specific quantile levels $\{\tau_n = \frac{n}{N}\}_{n=0}^{N-1}$:

$$Z_w(s,a) \sim \sum_{n=0}^{N-1} \frac{1}{N} \delta_{\theta_w^{\tau_n}(s,a)}, \tag{2}$$

where $\delta_{\theta_w^{\tau_n}(s,a)}$ is the Dirac delta function located at $\theta_w^{\tau_n}(s,a)$.[1] QR-based DRL learns the parameters w by minimizing the QR loss:

$$\mathcal{L}^{\text{QR}}(w) = \sum_{n=0}^{N-1} \mathbb{E}_{z \sim \hat{Z}(s,a)} \left[\rho_{\tau_n}\left(z - \theta_w^{\tau_n}(s,a)\right)\right], \tag{3}$$

where $\rho_{\tau_n}(e) = e \times (\tau_n - \mathbb{I}_{\{e<0\}})$. The QR loss can be minimized stochasticity by sampling M values $\{z_m\}_{m=1}^{M}$ from the distribution of the target $\hat{Z}(s,a)$, which serves as the benchmark for the distribution of $Z^{\pi}(s,a)$ during the learning updates. Given a transition tuple (s,a,R,s',a'), DRL methods, including QR, derive the target distribution (i.e., the distribution of $\hat{Z}(s,a)$) based on the distribution of $Z(s',a')$ using distributional Bellman equation. QR models the distribution of $Z(s',a')$ as a mixture of N quantile values $\phi_{w'}^{\tau_n}(s',a')$ parametrized with w':

$$Z_{w'}(s',a') \sim \sum_{n=0}^{N-1} \frac{1}{N} \delta_{\phi_{w'}^{\tau_n}(s',a')}. \tag{4}$$

Using the distributional Bellman equation, the target distribution in QR approaches is thus computed based on w' as

$$\hat{Z}_{w'}(s,a) \overset{\text{D}}{=} R + \gamma Z_{w'}(s',a') \sim \sum_{n=0}^{N-1} \frac{1}{N} \delta_{y_n}, \tag{5}$$

where $y_n = R + \gamma \phi_{w'}^{\tau_n}(s',a')$. Thus, sampling M values $\{z_m\}_{m=1}^{M}$ from $\hat{Z}_{w'}(s,a)$ corresponds to sampling from $\{y_n\}_{n=1}^{N}$. After updating w via a gradient step on $\mathcal{L}^{\text{QR}}(w)$, the parameter w' defining the target distribution is updated with the new value of w. Doing so ensures that the improvements made to the distribution's understanding of the return $Z_w(s,a)$ are carried forward to the distribution of the target $\hat{Z}_{w'}(s,a)$.

QR algorithms were initially applied in value-based DRL algorithms [11, 12], where only the return distribution is learned via the distributional Bellman equation, and actions are selected based on this learned distribution. However, value-based methods are limited to discrete state and action spaces and cannot be applied to continuous spaces. Recently, the use of QR in actor-critic DRL methods [9, 34] has been proposed, where both the return distribution and the policy are parameterized and learned, making it applicable to continuous state and action spaces.

3.2 Generalized Pareto Distribution

Originating from the peaks over thresholds approach in extreme value theory [33], the GPD is commonly employed to model the right-tail probabilities of various distributions. However, it can also be adapted to analyze extremely low values or the left tail, which is our focus here. Below, we detail the adaptation of the GPD for modeling the left tail, accompanied by a formal specification:

Theorem 3.1. *Let $X_1, X_2, \ldots$ be independent random variables with Cumulative Distribution Function (CDF) F. Given a small threshold u, the CDF of threshold exceedance $u - X$ conditioned on $X < u$ follows a GPD distribution [33],*

$$p(u - X \leq x | X < u) = H(x), \tag{6}$$

where $H(x)$ is the CDF of the GPD:

$$H(x) = \begin{cases} 1 - (1 + \frac{\varepsilon x}{\sigma})^{-1/\varepsilon} & \text{if } \varepsilon \neq 0 \\ 1 - e^{-\frac{x}{\sigma}} & \text{if } \varepsilon = 0, \end{cases} \tag{7}$$

where $-\infty < \varepsilon < \infty$ and $0 < \sigma < \infty$ are the shape and the scale parameters of GPD.

When $\varepsilon = 0$, the GPD is light-tailed with an exponential decay tail. If $\varepsilon < 0$, the distribution has a bounded upper limit. When $\varepsilon > 0$, the GPD is considered heavy-tailed [36]. In this case, the distribution exhibits a power-law tail, meaning that the probabilities of observing very small values decay polynomially rather than exponentially. This type of tail behavior is associated with 'fat tails,' indicating a higher likelihood of extreme events compared to a normal distribution. Research has demonstrated the efficacy of the heavy-tailed GPD for capturing extreme values theoretically [4, 31, 35]; therefore, it has been leveraged for measuring extreme events in time series and financial analysis [1, 17, 32, 38].

4 Formulating Hedging with Distributional Reinforcement Learning

In this section, we explain how we formulate the problem of hedging an options portfolio using DRL.

The environment, similar to that in [9], simulates a 30-day hedging problem that the trading desk of a bank will have to manage to hedge their options exposure resulting from client orders. We assume that the trader rebalances her position at fixed intervals of $dt = 1$ day leading up to the maturity of the option at $T = 30$ days. Moreover, we set $\gamma = 1$ as T is relatively short.

The price of the underlying asset, S, is assumed to follow a geometric Brownian motion and risk-neutral behavior governed by:

$$dS = (r - q)S\,dt + \sigma S\,dz, \tag{8}$$

where dz is a Wiener process, and σ is the volatility of the underlying asset. The risk-free interest rate r and the dividend yield q are assumed to be constant. Additionally, we assume that the real-world expected return μ is also constant. The real-world process for the asset price is the same as that given above, with $r - q$ replaced by μ.

[1] Throughout this paper, we use δ_x to denote a Dirac delta function located at x.

The value of the option is given by the Black–Scholes–Merton option pricing model developed by [6] and [29]:

$$S_0 N(d_1) e^{-qT} - K e^{-rT} N(d_2), \tag{9}$$

where

$$d_1 = \frac{\ln\left(\frac{S_0}{K}\right) + \left(r - q + \frac{\sigma^2}{2}\right) T}{\sigma\sqrt{T}}, \tag{10}$$

$$d_2 = d_1 - \sigma\sqrt{T}, \tag{11}$$

and S_0 represents the initial price of the underlying asset, K is the strike price of the option, and $N(\cdot)$ denotes the CDF of the standard normal distribution.

Delta (Δ) is the first partial derivative of the option price with respect to the asset price, and gamma (Γ) is the second partial derivative with respect to the asset price. Under this model, a natural idea is to regard the value of a European option as a function of S and σ for the calculation of these partial derivatives:

$$\Delta = N(d_1) e^{-qT}, \tag{12}$$

$$\text{and } \Gamma = \frac{N'(d_1) e^{-qT}}{S_0 \sigma \sqrt{T}}. \tag{13}$$

The delta and gamma of a portfolio are calculated by summing those for the individual options in the portfolio.

State: Our state includes the information that a trader would typically required when making a hedging decision. The state variable s_t at time step t consists of:

- The price of the underlying asset.
- The gamma of the portfolio.
- The gamma of the at-the-money option used for hedging.

Action: The agent is allowed to trade an at-the-money call option once every day to maximize the utility against the objective function. Daily, the agent's action a_t is to determine the proportion of maximum hedging done against the total gamma exposure. In other words, the portfolio gamma after hedging relative to gamma before hedging has to be between 0 and 1. This constraint effectively forces the agent to consider hedging as opposed to speculating. Once the agent determines the gamma hedging ratio for the day, the resulting delta will be rebalanced to 0 automatically.

Reward: Each day, we assume new options orders arrive following a Poisson process with a 50/50 chance of being long or short. Following market convention, the options provide the holder the right to buy or sell 100 units of the underlying assets. The reward (negative cost) $R(s_t, a_t)$ is therefore

$$R(s_t, a_t) = -\kappa \left|V_t H_t\right| + \left(P_t - P_{t-1}\right), \tag{14}$$

where V_t is the value of the option used for hedging at time t, H_t is the position taken in the option used for hedging at time t, κ is the transaction cost associated with the option used for hedging as a proportion of the value of the option, and P_t and P_{t-1} are the portfolio's market values at time t and $t-1$, respectively, that have not previously expired. If an option expires at time t, its value at time t is set equal to its intrinsic value $\max(S-K, 0)$. We assume no funding costs are required to finance these options, which is not unreasonable given the short (30-day) time horizon considered.

Risk Measure: Traditionally, financial risk management focuses on minimizing risk measures derived from the loss distribution, particularly managing risks associated with high-cost outcomes in the right tail of the loss distribution. However, to align with DRL standard, we redefine our approach based on the distribution of return $Z^\pi(s,a) = \sum_{t=1}^{T} \gamma^{t-1} R(s_t, a_t)$, effectively shifting our focus to the left side of this distribution where potential low-return, high-risk scenarios occur. We adapt VaR and CVaR to this context for a specific quantile level $\alpha \in (0,1)$, denoted as VaRα and CVaRα. VaRα reflects the quantile at which the CDF of $Z^\pi(s,a)$ reaches $1-\alpha$, and CVaRα calculates the expected returns below VaR$_\alpha$.

QR is particularly useful here as it directly measures VaR and CVaR through the quantile-based representations of the return distribution specified in Eq. (2). Specifically, VaR$\alpha(Z_w(s,a))$ directly corresponds to the quantile value $\theta_w^{1-\alpha}(s,a)$, and CVaRα is calculated as the average of quantile values $\theta_w^{\tau_n}(s,a)$ that are less than or equal to VaR$\alpha(Z_w(s,a))$.

However, estimating VaR and CVaR via QR becomes challenging when α approaches 1. This difficulty arises because observations for extreme quantile values in the tail of the distribution are scarce in samples from the target distribution (i.e., $\{y_n\}_{n=1}^N$), leading to volatile estimates of these values (i.e., $\theta_w^{1-\alpha}$) and consequently, the risk measures [2, 4]. Evaluating the probability of extreme events has become a pivotal issue in financial risk management, often requiring high α values (e.g., $0.95, 0.99$). While increasing the number of quantiles N can theoretically enhance the estimation of extreme quantiles by providing more observations at extreme values, this approach has its challenges. Each additional quantile represents new parameters to estimate, potentially leading to a high-dimensional parameter space. This parameter increase complicates the model, heightens the risk of overfitting, and escalates the computational demands. These complexities highlight the need for more robust methods to handle extreme values in DRL for risk management.

Given the success of the QR framework in risk assessment and the effectiveness of the GPD in modeling distribution tails, the subsequent section will introduce our framework, EX-DRL, for options hedging. This framework leverages GPD in conjunction with QR to accurately capture the extreme quantile values of the distribution of return $Z(s,a)$.

5 EX-DRL: EXtreme Distributional Reinforcement Learning

EX-DRL enriches samples in the tail regions of the return distribution, by not imposing rigid distributional assumptions on the central part of the distribution of the target random variable $\hat{Z}(s,a)$ and focusing instead on the extreme values in its left tail modeled with a heavy-tailed GPD. This provides more robust estimates for extreme quantile values and allows us to obtain quantile values directly using the QR method while enhancing their estimation through extensive sampling from the GPD, without the need to increase the number of quantiles N.

To benefit from directly measuring VaR and CVaR from the estimated return distribution, we model the return as $Z_w(s,a)$, whose distribution is a mixture of N quantile values $\{\theta_w^{\tau_n}(s,a)\}_{n=0}^{N-1}$ similar to QR-based DRL, as expressed in Eq. (2). However, the distribution of the target random variable $\hat{Z}(s,a)$ is represented as a mixture distribution consisting of the tail distribution for extreme values, represented using a heavy-tailed GPD, and an unknown body distribution for non-extreme values modeled as a quantile distribution.

Given that the distribution of target $\hat{Z}(s,a)$ is derived from the distribution of $Z(s',a')$, where s' and a' are the successor state and action, it follows that $Z(s',a')$ must also represent a mixture distribution. To facilitate this, we introduce a threshold $u(s',a')$, defining the demarcation between extreme and non-extreme values. This threshold is crucial as it determines the separation point in the distribution of $Z(s',a')$, which is then partitioned into two distinct components:

(1) **Random Variable $Z^{Tail}(s',a')$ Representing the Tail of the Return Distribution**: Represents the behavior of extreme values below $u(s',a')$. It employs a GPD random variable, denoted as $GPD_{\psi}(s',a')$, where the scale parameter $\sigma_{\psi_1}(s',a')$ and the shape $\varepsilon_{\psi_2}(s',a')$ parameter are defined by parameters ψ_1 and ψ_2, respectively. Here, $\psi = \{\psi_1, \psi_2\}$ represents the combined parameterization for both the scale and shape. This GPD is negated and shifted so that its endpoint aligns with the threshold $u(s',a')$:

$$Z_{\psi}^{Tail}(s',a') \overset{\text{D}}{=} -GPD_{\psi}(s',a') + u(s',a'). \tag{15}$$

(2) **Random Variable $Z^{Body}(s',a')$ Representing the Body of the Return Distribution**: Encompasses the bulk of the distribution greater than or equal to $u(s',a')$ and is modeled as a mixture of $N_{Body} < N$ quantiles $\{\phi_{w'}^{\tau_j}(s',a')\}_{j=0}^{N_{Body-1}}$ greater than or equal to the threshold $u(s',a')$, as

$$Z_{w'}^{Body}(s',a') \sim \sum_{j=0}^{N_{Body}-1} \frac{1}{N}\delta_{\phi_{w'}^{\tau_j}(s',a')}. \tag{16}$$

The relative proportions of $Z_{w'}^{Body}(s',a')$ and $Z_{\psi}^{Tail}(s',a')$ are adjusted such that the entire distribution of $Z(s',a')$ sums to unity, integrating both components seamlessly. Thus, the composition of $Z(s',a')$ is parameterized with w' and ψ and is expressed as

$$Z_{\psi,w'}(s',a') \overset{\text{D}}{=} (1-\beta)Z_{\psi}^{Tail}(s',a') + Z_{w'}^{Body}(s',a'), \tag{17}$$

where $\beta \in (0,1)$ defines the proportion of the distribution considered as the body of the distribution, referred to as the ‘body proportion.’

Given this mixture model configuration in Eq. (17), the distribution of target $\hat{Z}_{\psi,w'}(s,a)$ is achieved using the distributional Bellman equation as

$$\hat{Z}_{\psi,w'}(s,a) \overset{\text{D}}{=} R + \gamma Z_{\psi,w'}(s',a') \tag{18}$$

$$\overset{\text{D}}{=} R + \gamma\left((1-\beta)Z_{\psi}^{Tail}(s',a') + Z_{w'}^{Body}(s',a')\right). \tag{19}$$

The proposed mixture model for the distributions of $Z_{\psi,w'}(s',a')$ and $\hat{Z}_{\psi,w'}(s,a)$ is illustrated in Figure 1.

5.1 Sampling from the target distribution

We now focus on learning the parameters w of the quantiles $\{\theta_w^{\tau_n}(s,a)\}_{n=0}^{N-1}$ of $Z_w(s,a)$ using samples from $\hat{Z}_{\psi,w'}(s,a)$ defined in Eq. (19). To achieve this, we first sample from the tail and body components of distribution of $Z_{\psi,w'}(s',a')$ as follows:

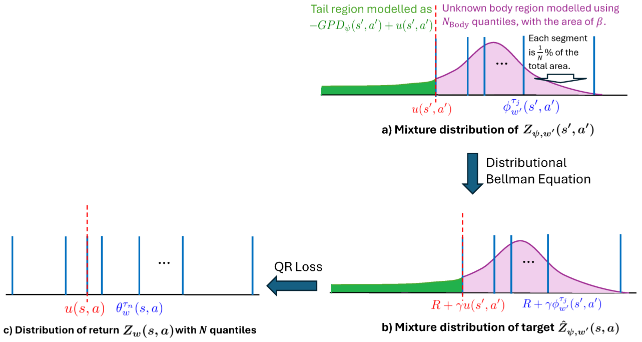

Tail region modelled as $-GPD_{\psi}(s',a') + u(s',a')$
Unknown body region modelled using N_{Body} quantiles, with the area of β.
Each segment is $\frac{1}{N}$% of the total area.
...
$u(s',a')$
$\phi_{w'}^{\tau_j}(s',a')$
a) Mixture distribution of $Z_{\psi,w'}(s',a')$
Distributional Bellman Equation
QR Loss
...
$R + \gamma u(s',a')$
$R + \gamma\phi_{w'}^{\tau_j}(s',a')$
b) Mixture distribution of target $\hat{Z}_{\psi,w'}(s,a)$
...
$u(s,a)$
$\theta_w^{\tau_n}(s,a)$
c) Distribution of return $Z_w(s,a)$ with N quantiles

Figure 1: The proposed mixture distribution for the target distribution $\hat{Z}_{\psi,w'}(s,a)$.

(1) **Sampling from $Z_{\psi}^{Tail}(s',a')$:** To better estimate quantiles in the tail of the distribution, we generate a specific number M_{Tail} of samples from $Z_{\psi}^{Tail}(s',a') \overset{\text{D}}{=} -GPD_{\psi}(s',a') + u(s',a')$. We achieve this by first generating M_{Tail} samples from $GPD_{\psi}(s',a')$, then negating each sample and shifting it by $u(s',a')$.

(2) **Sampling from $Z_{w'}^{Body}(s',a')$:** Given the proportional weight of the body and tail components in the distribution of $\hat{Z}_{\psi,w'}(s,a)$, the number of samples from $Z_{w'}^{Body}(s',a')$ must be $M_{\text{Body}} = \text{round}(\frac{M_{\text{Tail}} \times \beta}{1-\beta})$. Sampling from $Z_{w'}^{Body}(s',a')$ corresponds selecting M_{Body} samples from $\{\phi_{w'}^{\tau_j}(s',a')\}_{j=0}^{N_{Body}-1}$, as specified in Eq. (16).

By combining samples from $Z_{\psi}^{Tail}(s',a')$ and $Z_{w'}^{Body}(s',a')$ and applying the distributional Bellman equation, we obtain samples $\left\{\{z_k^{\text{Tail}}\}_{k=1}^{M_{\text{Tail}}}, \{z_l^{\text{Body}}\}_{l=1}^{M_{\text{Body}}}\right\}$ from the target distribution $\hat{Z}_{\psi,w'}(s,a)$.

5.2 Updating return distribution parameters

Given the samples $\left\{\{z_k^{\text{Tail}}\}_{k=1}^{M_{\text{Tail}}}, \{z_l^{\text{Body}}\}_{l=1}^{M_{\text{Body}}}\right\}$, the quantile parameters w of the distribution of $Z_w(s,a)$ are updated using the QR loss as:

$$\begin{aligned}\mathcal{L}^{\text{QR}}(w) \quad = \quad & \sum_{n=0}^{N-1}\left[\frac{1}{M_{\text{Body}}}\sum_{l=1}^{M_{\text{Body}}}\rho_{\tau_n}\left(z_l^{\text{Body}} - \theta_w^{\tau_n}(s,a)\right)\right. \\ & \left. + \quad \frac{1}{M_{\text{Tail}}}\sum_{k=1}^{M_{\text{Tail}}}\rho_{\tau_n}\left(z_k^{\text{Tail}} - \theta_w^{\tau_n}(s,a)\right)\right].\end{aligned} \tag{20}$$

The QR loss function integrates contributions from both the body and tail of the return distribution. By explicitly incorporating M_{Tail} samples from the tail, the model better addresses rare but potentially significant events, which are often under-represented in standard QR-based methods. Thus, this approach leverages the QR method to directly provide quantile values, while the GPD model of the tail enhances the accuracy of these estimates by generating M_{Tail} samples in the extreme tail segment of the return distribution.

5.3 Updating target distribution's parameters

In the previous subsection, we utilized samples from the target random variable $\hat{Z}_{\psi,w'}$ to update the quantile parameters w. With these updated parameters w (and thus the updated quantile values $\theta_w^{\tau_n}$), we now focus on updating parameters of the target distribution. This update specifically targets the parameters ψ of the distribution of tail Z_ψ^{Tail} and the parameters w' of the distribution of body $Z_{w'}^{Body}$.

To accomplish this, we consider any tuple $(s'', a'' = \pi(s''))$ where $s'' \in \mathcal{S}$, and proceed as follows:

(1) **Updating Parameters of the Distribution of $Z_\psi^{Tail}(s'', a'')$:** As $Z_\psi^{Tail}(s'', a'')$ is modeled using $GPD_\psi(s'', a'')$, updating its distribution parameters ψ involves adjusting the parameters of the distribution of $GPD_\psi(s'', a'')$. To accomplish this, we follow these steps:
- Identify the quantiles values $\{\theta_w^{\tau_i}(s'', a'')\}_{i=0}^{N_{\text{Tail}}-1}$ among $\{\theta_w^{\tau_n}(s'', a'')\}_{n=0}^{N-1}$ that are below the threshold $u(s'', a'')$.
- Transform these values by negating them and shifting them by $u(s'', a'')$ to ensure they fall within the support range of the GPD.
- Apply Maximum Likelihood Estimation (MLE) to these transformed values to update the parameters ψ.

(2) **Updating Parameters of the Distribution of $Z_{w'}^{Body}(s'', a'')$:** Consistent with QR-based practices, we update the parameters w' of using the most recent values w.

5.4 Selection of threshold u

Until now, we have assumed the existence of a unique threshold $u(s'', a'')$ for each tuple (s'', a''). Depending on the specific application, a fixed proportion $\beta \in (0, 1)$ is used to define the segment of the mixed distribution of $Z_{\psi,w'}(s'', a'')$ considered as the body, occurring at or above $u(s'', a'')$. Thus, the threshold $u(s'', a'')$ corresponds to the quantile value where the CDF of $Z_{\psi,w'}(s'', a'')$ is $1 - \beta$, i.e., $u(s'', a'') = \phi_{w'}^{1-\beta}(s'', a'')$. This method ensures that $u(s'', a'')$ accurately separates the most extreme values from the body based on the predefined proportion β.

5.5 Hedging with EX-DRL

Our proposed EX-DRL approach is compatible with any QR-based method. As described in Section 4, our option hedging problem employs continuous state and action spaces; thus, we implement EX-DRL within an actor-critic fashion. Here, the distribution of return $Z_w(s, a)$ acts as the critic, modeled via a neural network with parameters w. Simultaneously, the policy $\pi_\xi(s)$, parameterized by ξ, is represented by a separate neural network (the actor). Both actor and critic parameters are iteratively updated until convergence.

Drawing inspiration from the effectiveness of QR-D4PG [9], a recent actor-critic QR-based DRL method, we propose an extension, EX-D4PG, which replaces the quantile distribution used for the target distribution in QR-D4PG with our mixture model target distribution. Unlike QR-D4PG, EX-D4PG's critic not only reflects the usual dynamics of the return distribution but also emphasizes extreme event modeling. During training, the critic's parameters (i.e., parameters w of the distribution of $Z_w(s, a)$) are updated using gradient descent on the QR loss defined in Eq. (20). Concurrently, the actor network adjusts its parameters ξ through gradient ascent to maximize its risk measure, i.e., VaR$\alpha(Z_w(s, a))$ or CVaR$\alpha(Z_w(s, a))$. We should point out that QR-D4PG keeps α constant in VaRα and CVaRα for each hedging period, and therefore constitutes a time-inconsistent optimization framework. Specifically, it follows the stochastic control formulation presented in [5]. The proposed extension, EX-D4PG, inherits QR-D4PG's properties, but the general method we propose, EX-DRL, is orthogonal and can be applied to both time-consistent and time-inconsistent control problems.

Compared to QR-D4PG, our method introduces two additional neural networks, parameterized by $\psi = \{\psi_1, \psi_2\}$, which models the scale σ_{ψ_1} and shape ε_{ψ_2} parameters of the GPD, respectively. As noted in [36], a GPD with $\varepsilon \geq 1$ results in an infinite mean, which would imply that the CVaR could also be infinite. Therefore, to effectively adapt to markets with heavy-tailed distributions, our neural network is specifically designed, through its final sigmoid activation function, to ensure that $0 < \varepsilon_{\psi_2} < 1$, maintaining a finite expectation for the CVaR and ensuring stability in estimation under extreme market conditions. The pseudocode of our proposed EX-D4PG is given in Algorithm 1.

6 Experimental Results

To evaluate the performance of EX-D4PG in capturing extreme losses, we conducted a series of hedging experiments and compared the results with those of QR-D4PG [9].

6.0.1 ***Setup.*** The experimental parameters mirror those used in the gamma hedging experiments described in [9], enabling direct comparisons between the EX-D4PG and QR-D4PG agents. In this setup, client orders arrive according to a Poisson process at an intensity of 1.0 option per day. Each order typically involves a 60-day option on 100 units of the underlying asset, with an equal likelihood of being a long or short position. We assume transaction costs at 1% of the option price. Hedging is conducted using a 30-day at-the-money option. The initial stock price is set at $10, with an annual volatility of 0.3. We evaluated the performance across 5,000 test scenarios, with results averaged over 3 runs. The hedge ratio is calculated as the reduction in gamma exposure post-hedging relative to pre-hedging:

$$\text{Gamma Hedge Ratio} = 1 - \frac{\sum \text{sign}(\Gamma)\Gamma^+}{\sum \text{sign}(\Gamma)|\Gamma|}. \tag{21}$$

A higher gamma hedge ratio indicates that the agent is more likely to hedge away all its gamma exposure, and a gamma hedge ratio of 1 means the agent always fully hedges its gamma exposure.

To determine the value of the body proportion β for our mixed distribution model, we tested values $\{0.94, 0.95, 0.96, 0.99\}$ across various volatility levels. The value $\beta = 0.95$ was chosen for its superior performance across several test scenarios. The hyperparameters of the actor and critic neural networks are the same as those used in our baseline QR-D4PG [9]. For the additional neural networks modeling the scale σ_{ψ_1} and shape ε_{ψ_2} parameters of the GPD, we use hidden layers of size $(512, 512, 256)$, with tanh activation for the first layer and elu for the subsequent layers, sharing

Algorithm 1 The proposed EX-D4PG framework.

```
Input: $N$: number of quantiles, $M_{\text{Tail}}$: number of samples from the target distribution's tail, $\beta$: body proportion, $\alpha$: learning rate
Initialize: $w$: critic's parameters, $w'$: target's body parameters, $\psi$: target's tail (i.e., GPD) parameters, $\xi$: actor's parameters
while not converged do
    for t=1, 2, ..., T do
        Make transition $(s, a, R, s', a' = \pi_\xi(s'))$
        Select threshold $u(s', a') = \phi_{w'}^{1-\beta}(s', a')$.
        Sampling from the target distribution:
        Sample $\{x_k\}_{k=1}^{M_{\text{Tail}}}$ from $GPD_\psi(s', a')$
        Define $\{y_k^{\text{Tail}} = -x_k + u(s', a')\}_{k=1}^{M_{\text{Tail}}}$
        Generate tail samples $\{z_k^{\text{Tail}} = R + \gamma y_k^{\text{Tail}}\}_{k=1}^{M_{\text{Tail}}}$
        Define $M_{\text{Body}} = \text{round}(\frac{M_{\text{Tail}} \times \beta}{1-\beta})$
        Sample $\{y_l^{\text{Body}}\}_{l=1}^{M_{\text{Body}}}$ from $\{\phi_{w'}^{\tau_j}(s', a')\}_{j=0}^{N_{Body}-1}$ for $\tau_j = \frac{j}{N} + 1 - \beta$
        Create body samples $\{z_l^{\text{Body}} = R + \gamma y_l^{\text{Body}}\}_{l=1}^{M_{\text{Body}}}$
        Update critic's parameters:
        Compute QR loss $\mathcal{L}^{\text{QR}}(w)$ according to Eq. (20).
        Update $w$ as $w \leftarrow w - \alpha \nabla_w \mathcal{L}^{\text{QR}}(w)$
        Update actor's parameters:
        $\xi \leftarrow \xi + \alpha \nabla_\xi \text{(C)VaR}_\alpha(Z_w(s, \pi_\xi(s)))$
        Update the target distribution's parameters:
        Consider any tuple $(s'', a'' = \pi_\xi(s''))$
        Generate $\{\theta_w^{\tau_i}(s'', a'')\}_{i=0}^{N_{\text{Tail}}-1}$ for $\tau_i = \frac{i}{N}$
        Define $L(\psi)$ as the log-likelihood of $GPD_\psi(s'', a'')$ fitted to $\{\theta_w^{\tau_i}(s'', a'')\}_{i=0}^{N_{\text{Tail}}-1}$
        Update target's tail (GPD) parameters as $\psi \leftarrow \psi + \alpha \nabla_\psi L(\psi)$
        Update target's body parameters as $w' \leftarrow w$
    end for
end while
```

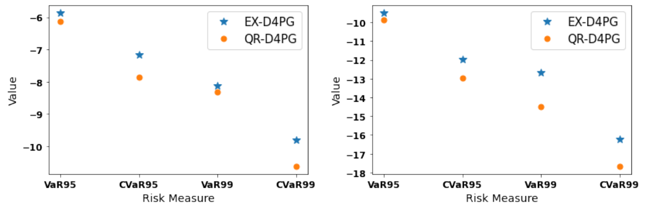
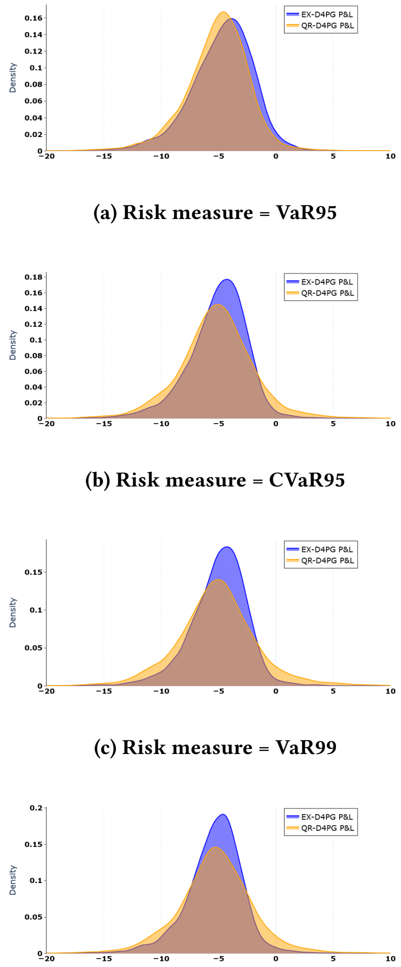


Figure 2: VaR and CVaR values when volatility = 0.3 (left) and volatility = 0.5 (right).

the same backbone for both σ_{ψ_1} and ε_{ψ_2}. The last layer of both networks is of size 1. In the shape network's final layer, we applied a sigmoid activation function to ensure $0 < \varepsilon_{\psi_2} < 1$ for modeling the heavy tail. For the scale network, we applied a softplus activation function in the final layer to ensure $\sigma_{\psi_1} > 0$. Further details about the implementation are available here.



(a) Risk measure = VaR95

(b) Risk measure = CVaR95

(c) Risk measure = VaR99

(d) Risk measure = CVaR99

Figure 3: PnL distributions for different risk measures when volatility = 0.5.

6.0.2 ***Results.*** Figure 2 compares the performance of QR-D4PG and EX-D4PG at volatility levels of 0.3 and 0.5 for the underlying asset. The figure shows that the EX-D4PG outperforms QR-D4PG across all evaluated risk measures, i.e., VaRα or CVaRα for $\alpha = \{0.95, 0.99\}$.
Further analysis of the PnL distribution, as depicted in Figure 3, is supported by the corresponding statistics in Table 1. These results reveal that the EX-D4PG model incurs fewer extreme losses compared to QR-D4PG. Additionally, the PnL distribution for EX-D4PG is more on the positive side of the graph, indicating improved mean performance and exhibiting a narrower spread of results, as evidenced by reduced standard deviations.

Moreover, we tested the performance of both RL agents under varying volatility conditions, ranging from 0.1 to 0.8, using VaR95 and CVaR95 as the risk measures. The results are shown in Figure 4. The EX-D4PG agent performs comparably to the QR-D4PG

Table 1: PnL distribution statistics for different risk measures when volatility = 0.5.

Risk Measure	Metric	EX-D4PG	QR-D4PG
VaR95			
	Mean	-4.67	-5.10
	Standard Deviation	2.84	2.94
	VaR95	-9.51	-9.86
CVaR95			
	Mean	-5.07	-5.22
	Standard Deviation	2.72	3.44
	VaR95	-9.68	-10.66
VaR99			
	Mean	-5.00	-5.08
	Standard Deviation	2.62	3.82
	VaR99	-12.66	-14.47
CVaR99			
	Mean	-5.35	-5.25
	Standard Deviation	2.65	3.50
	VaR99	-13.00	-14.58

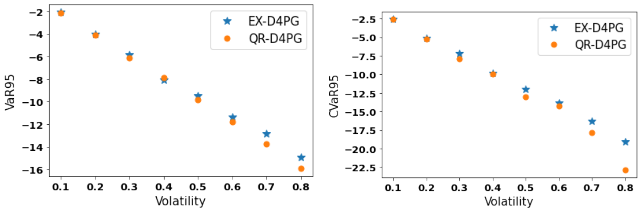


Figure 4: Comparison of VaR95 (left) and CVaR95 (right) at different volatility levels.

at low to medium volatility levels. However, as volatility increases, EX-D4PG demonstrates markedly better performance, underscoring the effectiveness of explicitly modeling the tail distribution in enhancing the agent's performance under extreme conditions.

Finally, Figure 5 analyzes the hedging behavior of the RL agents by comparing the gamma hedge ratio, as defined in Eq. (21). The results show that while both agents generally under-hedge their gamma exposure (indicated by a gamma hedge ratio of less than 1), the EX-D4PG agent demonstrates more adaptive hedging behavior. It adjusts its hedge ratio downward as volatility increases, which helps to reduce hedging costs and improve overall financial performance.

It is important to note from Figure 5 that QR-D4PG slightly outperforms EX-D4PG at volatility levels of 0.1 and 0.4. This is due to using a fixed value of $\beta = 0.95$ across all levels to balance model performance with generalizability. Optimizing β individually for each level could enhance EX-D4PG's performance, consistently surpassing QR-D4PG. Due to space constraints only outcomes for a fixed β are shown. Additionally, this sets a clear baseline for future work to explore adaptive strategies, further tailoring β to individual scenarios. Such iterative improvements are fundamental in advancing research in this field.

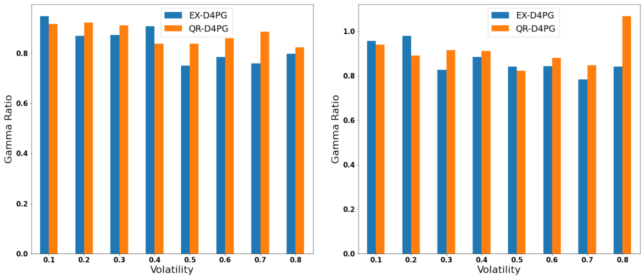


Figure 5: Comparison of gamma hedge ratio using VaR95 (left) and CVaR95 (right) at various volatility levels.

7 Conclusion and Future Work

In this paper, we introduced EX-DRL, a novel DRL algorithm designed to enhance the estimation of extreme quantiles in option hedging. EX-DRL employs a GPD to model the tail of the target distribution and is compatible with QR-based DRL methods, thereby enabling direct estimation of risk measures such as VaR and CVaR. For experimental validation, we integrated the GPD-based tail modeling within a recently proposed QR-based method, QR-D4PG, and termed this integration EX-D4PG. The results confirmed that EX-D4PG outperforms QR-D4PG in precision in extreme quantile predictions, thus providing a more reliable measure of extreme losses.

While our approach currently employs MLE for parameter estimation of the GPD, recent studies [18, 36] suggested more advanced estimation techniques that could potentially enhance parameter accuracy. As a direction for future research, integrating these advanced estimation methods into our EX-DRL framework could further refine the accuracy of extreme quantile estimation.Another natural extension of the experiment would be to consider both delta and gamma hedging jointly by allowing the RL agent to trade both the underlying asset and at-the-money options as hedging instruments.

References

[1] S Abilasha, Sahely Bhadra, Ahmed Zaheer Dadarkar, and P Deepak. 2022. Deep Extreme Mixture Model for Time Series Forecasting. In *CIKM*. 1726–1735.

[2] Yu Bai, Song Mei, Huan Wang, and Caiming Xiong. 2021. Understanding the under-coverage bias in uncertainty estimation. *Advances in Neural Information Processing Systems* 34 (2021), 18307–18319.

[3] Marc G Bellemare, Will Dabney, and Rémi Munos. 2017. A distributional perspective on reinforcement learning. In *International conference on machine learning*. PMLR, 449–458.

[4] Boris Beranger, Simone A Padoan, and Scott A Sisson. 2021. Estimation and uncertainty quantification for extreme quantile regions. *Extremes* 24, 2 (2021), 349–375.

[5] Tomas Björk and Agatha Murgoci. 2014. A theory of Markovian time-inconsistent stochastic control in discrete time. *Finance and Stochastics* 18 (2014), 545–592.

[6] Fischer Black and Myron Scholes. 1973. The pricing of options and corporate liabilities. *Journal of political economy* 81, 3 (1973), 637–654.

[7] Hans Buehler, Lukas Gonon, Josef Teichmann, and Ben Wood. 2019. Deep hedging. *Quantitative Finance* 19, 8 (2019), 1271–1291.

[8] Hans Buehler, Phillip Murray, and Ben Wood. 2022. Deep bellman hedging. *arXiv preprint arXiv:2207.00932* (2022).

[9] Jay Cao, Jacky Chen, Soroush Farghadani, John Hull, Zissis Poulos, Zeyu Wang, and Jun Yuan. 2023. Gamma and vega hedging using deep distributional reinforcement learning. *Frontiers in Artificial Intelligence* 6 (2023), 1129370.

[10] Jay Cao, Jacky Chen, John Hull, and Zissis Poulos. 2021. Deep Hedging of Derivatives Using Reinforcement Learning. *The Journal of Financial Data Science* 3, 1 (2021), 10–27.

[11] Will Dabney, Georg Ostrovski, David Silver, and Rémi Munos. 2018. Implicit quantile networks for distributional reinforcement learning. In *International*

conference on machine learning. PMLR, 1096–1105.

[12] Will Dabney, Mark Rowland, Marc Bellemare, and Rémi Munos. 2018. Distributional reinforcement learning with quantile regression. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

[13] Roberto Daluiso, Marco Pinciroli, Michele Trapletti, and Edoardo Vittori. 2023. Cva hedging with reinforcement learning. In *Proceedings of the Fourth ACM International Conference on AI in Finance*. 261–269.

[14] Parisa Davar, Frédéric Godin, and Jose Garrido. 2024. Catastrophic-risk-aware reinforcement learning with extreme-value-theory-based policy gradients. *arXiv preprint arXiv:2406.15612* (2024).

[15] Kang Gao, Stephen Weston, Perukrishnen Vytelingum, Namid Stillman, Wayne Luk, and Ce Guo. 2023. Deeper Hedging: A New Agent-based Model for Effective Deep Hedging. In *Proceedings of the Fourth ACM International Conference on AI in Finance*. 270–278.

[16] Igor Halperin. 2020. QLBS: Q-Learner in the Black-Scholes (-Merton) Worlds. *The Journal of Derivatives* 28, 1 (2020), 99–122.

[17] Yi He, Liang Peng, Dabao Zhang, and Zifeng Zhao. 2022. Risk analysis via generalized Pareto distributions. *Journal of Business & Economic Statistics* 40, 2 (2022), 852–867.

[18] Yan Huang, Fuyu Du, Jian Chen, Yan Chen, Qicong Wang, and Maozhen Li. 2019. Generalized Pareto model based on particle swarm optimization for anomaly detection. *IEEE Access* 7 (2019), 176329–176338.

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

[20] Yuxi Li, Csaba Szepesvari, and Dale Schuurmans. 2009. Learning exercise policies for american options. In *Artificial intelligence and statistics*. PMLR, 352–359.

[21] Guiliang Liu, Yudong Luo, Oliver Schulte, and Pascal Poupart. 2022. Uncertainty-aware reinforcement learning for risk-sensitive player evaluation in sports game. *Advances in Neural Information Processing Systems* 35 (2022), 20218–20231.

[22] Yudong Luo, Guiliang Liu, Pascal Poupart, and Yangchen Pan. 2023. An alternative to variance: Gini deviation for risk-averse policy gradient. *Advances in Neural Information Processing Systems* 36 (2023), 60922–60946.

[23] Yecheng Ma, Dinesh Jayaraman, and Osbert Bastani. 2021. Conservative offline distributional reinforcement learning. *Advances in neural information processing systems* 34 (2021), 19235–19247.

[24] Parvin Malekzadeh, Ming Hou, and Konstantinos N Plataniotis. 2023. A unified uncertainty-aware exploration: Combining epistemic and aleatory uncertainty. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[25] Parvin Malekzadeh and Konstantinos N Plataniotis. 2024. Active Inference and Reinforcement Learning: A Unified Inference on Continuous State and Action Spaces under Partial Observability. *Neural Computation* (2024), 1–64.

[26] Parvin Malekzadeh, Konstantinos N Plataniotis, Zissis Poulos, and Zeyu Wang. 2024. A Robust Quantile Huber Loss with Interpretable Parameter Adjustment in Distributional Reinforcement Learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing*. 6120–6124.

[27] Parvin Malekzadeh, Mohammad Salimibeni, Arash Mohammadi, Akbar Assa, and Konstantinos N Plataniotis. 2020. MM-KTD: multiple model kalman temporal differences for reinforcement learning. *IEEE Access* 8 (2020), 128716–128729.

[28] Saeed Marzban, Erick Delage, and Jonathan Yu-Meng Li. 2022. Equal risk pricing and hedging of financial derivatives with convex risk measures. *Quantitative Finance* 22, 1 (2022), 47–73.

[29] Robert C Merton. 1973. Theory of rational option pricing. *The Bell Journal of economics and management science* (1973), 141–183.

[30] Phillip Murray, Ben Wood, Hans Buehler, Magnus Wiese, and Mikko Pakkanen. 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*. 361–368.

[31] Qiyun Pan, Young Myoung Ko, and Eunshin Byon. 2020. Uncertainty quantification for extreme quantile estimation with stochastic computer models. *IEEE Transactions on Reliability* 70, 1 (2020), 134–145.

[32] Shige Peng, Shuzhen Yang, and Jianfeng Yao. 2023. Improving value-at-risk prediction under model uncertainty. *Journal of Financial Econometrics* 21, 1 (2023), 228–259.

[33] James Pickands III. 1975. Statistical inference using extreme order statistics. *the Annals of Statistics* (1975), 119–131.

[34] Anil Sharma, Freeman Chen, Jaesun Noh, Julio DeJesus, and Mario Schlener. 2024. Hedging Beyond the Mean: A Distributional Reinforcement Learning Perspective for Hedging Portfolios with Structured Products. *arXiv preprint arXiv:2407.10903* (2024).

[35] Abhay K Singh, David E Allen, and Powell J Robert. 2013. Extreme market risk and extreme value theory. *Mathematics and computers in simulation* 94 (2013), 310–328.

[36] Dylan Troop, Frédéric Godin, and Jia Yuan Yu. 2021. Bias-corrected peaks-over-threshold estimation of the cvar. In *Uncertainty in Artificial Intelligence*. PMLR, 1809–1818.

[37] Edoardo Vittori, Michele Trapletti, and Marcello Restelli. 2020. Option hedging with risk averse reinforcement learning. In *Proceedings of the first ACM international conference on AI in finance*. 1–8.

[38] Wen Xu, Huixia Judy Wang, and Deyuan Li. 2022. Extreme quantile estimation based on the tail single-index model. *Statistica Sinica* 32, 2 (2022), 893–914.