Risk-Averse Policy Gradient for Tail Risk Optimization Using Extreme Value Theory

Parisa Davar

A Thesis

in

The Department

of

Mathematics and Statistics

Presented in Partial Fulfillment of the Requirements

for the Degree of

Master of Science (Mathematics) at

Concordia University

Montréal, Québec, Canada

July 2024

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Parisa Davar**

Entitled: **Risk-Averse Policy Gradient for Tail Risk Optimization Using Extreme Value Theory**

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

Master of Science (Mathematics)

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

Signed by the Final Examining Committee:

_____ Examiner
Dr. Mélina Mailhot

_____ Examiner
Dr. Yang Lu

_____ Thesis Co-Supervisor
Prof. Jose Garrido

_____ Thesis Co-Supervisor
Dr. Frédéric Godin

Approved by _____
Dr. Marco Bertola, Chair of Department

_____ 2024 _____
Dr. Pascale Sicotte, Dean of Faculty

Abstract

Risk-Averse Policy Gradient for Tail Risk Optimization Using Extreme Value Theory

Parisa Davar

In this work, we develop a risk-averse Policy Gradient algorithm in a tail risk optimization problem. Our objective is to find the optimal policy that minimizes tail risk, given a risk measure such as Conditional Value at Risk (CVaR). We employed Extreme Value Theory, along with the automated threshold method to manage risks associated with extreme events. This paper is the first to integrate EVT within risk-averse policy gradient RL algorithms for sequential decision making. To evaluate our approach, we initially test it on simulated data generated from heavy-tailed distributions, including the Generalized Pareto distribution (GPD) and the Burr distribution. Subsequently, we applied our method to address a hedging problem, aiming to mitigate exceedingly high risks and finding optimal gamma hedging strategies within a highly volatile market where options are notably expensive. This involves identifying the optimal proportion of gamma to hedge, while minimizing costs and risk associated with gamma hedging errors. Also, we utilize the finite difference method to approximate the gradient of the estimated CVaR. The experimental results indicate convergence in the policy, CVaR estimation, and the gradient approximation of estimated CVaR. Moreover, integrating Extreme Value Theory into risk-averse policy gradient methods significantly improves performance, especially in markets characterized by an underlying asset following a Normal Inverse Gaussian distribution (NIG), known for its pure-jump semi-heavy tail distribution.

Acknowledgments

I would like to express my heartfelt and profound appreciation to my supervisors, Professor Jose Garrido and Professor Frédéric Godin, for their invaluable support, guidance, expertise, and encouragement throughout this academic journey. Their insightful feedback, belief in my potential, and willingness to invest time have been instrumental in shaping my research and academic growth. Moreover, I want to extend special appreciation for their continuous encouragement and understanding during moments of uncertainty and challenge.

I extend my deepest gratitude to my parents and my brother for their unwavering love, encouragement, and sacrifices throughout my academic journey. Their endless support has been the cornerstone of my success, and I am truly grateful for everything they have done for me and I am deeply thankful for their belief in me.

Finally, I would like to extend my sincere appreciation to all my friends who have provided me with encouragement, understanding, and support during this arduous yet rewarding endeavor.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Recently, reinforcement learning (RL) has been subject to a lot of attention due to its success in many real-world problems with high complexity such as board games, healthcare (Yu, Liu, Nemati, & Yin, 2021), and finance (Hambly, Xu, & Yang, 2023). RL methods are learning through interactions between an environment, characterized by states, and an agent, defined by actions. This interaction dynamics enables RL to find optimal solutions for sequential decision-making problems via an iterative process of trial and error. RL algorithms can be broadly categorized into two main groups: model-based and model-free approaches. Both categories aim to determine an optimal policy. Model-based algorithms either require an explicit knowledge of the environment, represented by transition probabilities, or learning its dynamics. Conversely, model-free methods do not require such environmental knowledge. For an in-depth exploration of RL algorithms, refer to Sutton and Barto (2018) for a comprehensive review.

In traditional RL, the primary objective is to maximize rewards. However, risk-aware RL shifts its focus toward minimizing risk, incorporating a risk measure into the objective function. Risk-aware RL plays a pivotal role in numerous fields, particularly in finance, where effective risk management is essential.

Risk management plays a crucial role in the financial industry, particularly when it comes to assessing rare events that can lead to significant and potentially catastrophic risks. For financial

institutions, it is essential to identify, measure, and mitigate risks in order to protect their investments. As financial institutions become more complicated, the use of appropriate risk management strategies becomes increasingly important. Furthermore, the 2008 global financial crisis and following economic shocks have highlighted the importance of a proactive and comprehensive risk management strategy.

Advanced risk modeling tools are the key elements of modern risk management in financial institution. By thoroughly assessing investments in extreme scenarios, financial institutions can better prepare for economic downturns and protect the interests of their clients and stakeholders.

Extreme Value Theory (EVT) is a well-known statistical risk management methodology to manage rare and extreme events that can have significant consequences. EVT consists of two main methods: the block maxima method and the peaks-over-threshold (POT) method or threshold exceedances method. In the block maxima method, data is divided into equal-sized blocks, and we focus on identifying and modeling distribution of most significant losses or extreme events of each block. However, this method has a drawback as it can waste data because of considering only one data point from each block. On the other hand, the POT method focuses on modeling rare events that exceed a pre-determined high threshold. EVT provides us with significant insights into how these rare events behave, allowing us to better prepare to reduce the likelihood of catastrophic events.

There are two fundamental theorems in EVT: the Fisher-Tippett theorem and the Pickands-Balkema-de Haan theorem. These theorems establish that the excess distribution converges to a Generalized Pareto Distribution (GPD) for sufficiently large thresholds. This convergence is highly useful because EVT suggests that very extreme events follow GPD distributions. Moreover, Maximum Likelihood Estimation (MLE) and the Method of Moments are two well-known methods for estimating the GPD parameters of excess data.

In this thesis, we introduce a risk-averse RL approach designed to effectively manage tail risk and ultimately mitigate catastrophic risk. To quantify the risk associated with the tail distribution,

we employ CVaR as our risk assessment criterion. CVaR, sometimes referred to as expected shortfall, allows us to evaluate the potential losses beyond a certain threshold with a focus on the most severe outcomes. To address rare events characterized by very large observations, we employ the POT method, a technique within EVT. The POT method, also known as threshold exceedances, focuses on modeling data beyond a high threshold. This method is particularly useful when dealing with rare events. The selection of an appropriate high threshold is a crucial aspect of the POT method, and it often poses a challenge. To navigate this issue, we incorporate an automated threshold selection method drawn from the EVT literature. This ensures that we choose a threshold that optimally represents the tail behavior of our data. Subsequently, by applying the EVT formula for CVaR estimation, we can effectively compute the CVaR associated with the excesses.

By effectively integrating the EVT method with the risk-averse RL framework, we identify optimal policies that minimize risk of extreme events. In our optimization process, we utilize the Adaptive Moment Estimation (ADAM) optimization method, a widely adopted approach known for its effectiveness in training deep learning models.

We assess our proposed method in two parts: initially, on simulated datasets generated from heavy-tailed distributions such as GPD and Burr distributions; subsequently, we apply our model to a hedging portfolio problem, showcasing its application in finance.

In the initial phase, we derived costs from the GPD and Burr distributions, as they hold particular significance for us due to their propensity to generate extremes. Heavy-tailed distributions, like these, leading potential catastrophic risks, hence our objective is to find the optimal policy for risk mitigation.

In the second phase, we suppose the market follows an exponential Normal Inverse Gaussian (NIG)-Lévy model, then we consider a problem of gamma hedging for an at-the-money European call option. Gamma hedging is a method aimed at minimizing potential risks associated with unfavorable price fluctuations in the market. However, achieving a perfect hedge is unfeasible in real markets, as the value of the hedging portfolio cannot precisely match the option's payoff at

expiration, leading to hedging errors. Moreover, when the market model follows a NIG distribution, characterized by pure jump semi-heavy-tailed behavior, markets become highly volatile, making options expensive for hedging strategies. An effective approach to reduce hedging costs and minimize the hedging error is to adjust the proportion of gamma to hedge. In this study, the RL risk-averse policy specifies the optimal proportion of gamma to hedge to minimize risk in a highly volatile market. To address our concerns regarding extremely rare events, we incorporate EVT into the risk-averse policy gradient algorithm as it is a suitable method for this purpose.

The experimental outcomes for both simulated data and addressing hedging problems demonstrate the convergence of our proposed algorithm, the risk-aware policy gradient method using EVT with automated threshold selection. Furthermore, we employ the finite difference method to estimate the gradient of the estimated CVaR during the optimization process. The results indicate that the approximation of the gradient for the estimated CVaR converges to zero, indicating that the policy converges to its optimal value.

Additionally, we evaluate the performance of CVaR estimation using EVT and compare it with the sample averaging (SA) method. The findings reveal that EVT outperforms SA, showing the efficacy of our approach in accurately estimating CVaR.

1.1 Main contribution

The primary objective of this thesis is to develop a risk-averse RL method specifically designed to address tail risk optimization problems, with the aim of mitigating catastrophic risks. In this risk-averse RL approach, we employ the EVT method to model the tail distribution and manage rare events. Our focus centers on the POT approach, a key component of EVT. To tackle the challenge of selecting an appropriate threshold in the POT method, we incorporate an automated threshold selection technique introduced by Bader, Yan, and Zhang (2018) and used in Troop, Godin, and Yu (2022). Additionally, to quantify the risk associated with the tail distribution, we use CVaR as our risk criterion. Since there is no closed form formula for the gradient of CVaR for all distributions,

we utilize the finite difference method to approximate the gradient of the estimated CVaR during the optimization process.

Our work involves integrating the EVT method into the risk-averse RL framework. To the best of our knowledge this thesis is the first to integrate EVT within risk-averse policy gradient RL algorithms for sequential decision making. It bridges the gap between RL and the field of EVT. The proposed algorithm can be applied in the realm of portfolio management, where portfolios are intricately composed of diverse assets such as options on stocks. Given the stochastic nature of stock prices, which are subject to constant fluctuation, our aim is to effectively hedge the portfolio against very extreme potential losses. Specifically, we seek to minimize the worst-case losses while optimizing the associated costs. This necessitates a strategic approach to risk management that can navigate the uncertainties inherent in financial markets.

1.2 Related work

RL methods have achieved remarkable results in different areas of finance such as portfolio optimization (Du, Zhai, & Lv, 2016), option pricing and hedging (Cao, Chen, Hull, & Poulos, 2020), and robo-advising or automated investment (Capponi, Olafsson, & Zariphopoulou, 2022). A comprehensive survey of recent developments in the application of RL to finance can be found in Hambly et al. (2023).

In traditional RL, the objective is to maximize the expected value of returns regardless of the associated risks. This approach is recognized as risk-neutral RL. By contrast, risk-sensitive RL methods have emerged to incorporate risk within the RL framework. An overview of these advancements can be found in the survey conducted by Prashanth, Fu, et al. (2022). This survey classifies risk-sensitive RL techniques into two distinct settings. The first involves the goal of maximizing returns while taking risk into account as a constraint. The second setting directly incorporates risk as an objective function within the optimization process. In the latter one, the agent prefers to minimize the risk which is an inherent risk due to the stochastic nature of the

environment. It is referred to as risk-averse RL method.

In the literature, a variety of risk measurement methods have been explored such as mean-variance, exponential utility formulation (Pratt, 1978), cumulative prospect theory (Tversky & Kahneman, 1992), percentile performance criteria (Wu & Lin, 1999), Value at Risk (VaR), and coherent risk measures like CVaR (Artzner, Delbaen, Eber, & Heath, 1999; Rockafellar, Uryasev, et al., 2000). These risk criteria are also used in the context of risk-sensitive RL. For example, Borkar (2001) uses an exponential utility function in risk sensitive RL. More recently, Tamar, Di Castro, and Mannor (2012) and La and Ghavamzadeh (2013) consider a mean-variance risk measure. Moreover, Prashanth, Jie, Fu, Marcus, and Szepesvári (2016) and Jie, Prashanth, Fu, Marcus, and Szepesvári (2018) present risk-sensitive RL algorithms relying on the Cumulative Prospect Theory (CPT) and Chow, Ghavamzadeh, Janson, and Pavone (2017) uses percentile performance as a risk criterion. Two popular risk measures are VaR and CVaR. While VaR is not a coherent risk measure, it is not commonly used as an objective in risk-sensitive RL (Vijayan & Prashanth, 2023). Tamar, Glassner, and Mannor (2015) introduces a policy gradient method for any coherent risk measure and derive a policy gradient theorem. According to Greenberg, Chow, Ghavamzadeh, and Mannor (2022) study, policy gradient emerges as the most widely adopted approach for CVaR optimization in RL.

EVT is a well known method in modeling the tail behavior of a distribution, focusing on events that are rare and more extreme in financial data to avoid the catastrophic risks. Assessing these risks, commonly referred to as catastrophic risks, holds significant importance in finance, particularly in portfolio management within highly volatile markets. According to Godin, Mayoral, and Morales (2012), it is widely acknowledged that markets exhibit semi-heavy-tailed distributions. The Normal Inverse Gaussian (NIG) distribution, belonging to the family of semi-heavy-tailed distributions, represents a pure jump process contributing to a highly volatile market. Thus, the implementation of hedging strategies is essential since it enables investors to protect their portfolios from price fluctuations.

POT is a common method in EVT. The estimation of CVaR using the POT method has been applied across numerous studies, including Gilli and Këllezi (2006), Gkillas and Katsiampa (2018), and Szubzda and Chlebus (2019). A notable challenge within the POT method is the optimal threshold selection. To address this challenge, Bader et al. (2018) introduced an EVT with an automated threshold selection method by ordered goodness-of-fit tests. Later Troop et al. (2022) have further modified the automated threshold selection method within the context of risk-aware multi-armed bandit problem. Their findings indicate that the EVT approach with automated threshold outperforms the sample average method in term of RMSE of estimated CVaR.

In this thesis, we focus on tail risk optimization problem, using a risk averse policy gradient using CVaR as an objective function and EVT with automated threshold selection method for estimating CVaR. This is a bridge between risk averse policy gradient and EVT with automated threshold selection method and there is no previous work to combine these two concepts to avoid the catastrophic risk.

1.3 Thesis structure

This thesis is organized as follows. Chapter 2 gives an overview of the NIG - levy financial framework and hedging strategy that will be considered in the application section of this thesis. Moreover, the dynamics of stock price process, the pricing of European Call options as well as delta gamma hedging formula based on NIG process are presented in this chapter.

Chapter 3 provides a review of two popular risk measures: VaR and CVaR. Also, it delves into the methods of estimating these measures, focusing on both sample averaging and EVT methods. Furthermore, we discuss the POT approach with an automated threshold selection method.

Chapter 4 includes an introduction to RL. We begin by presenting a review of Markov decision processes. Then we introduce models proposed to tackle the sequential decision making problem, with a specific focus on model-free RL methods.

Chapter 5 contains the details of our proposed algorithm. Our approach involves a risk-averse

policy gradient framework for tail risk optimization problem. This chapter introduces the risk-aware policy gradient method, and then provides a more precise definition of the problem being addressed.

Chapter 6 describes our experimental findings. This chapter is divided into two parts: the first shows results derived from simulated data of simple distributions, including the GPD and the Burr distribution, while the second section highlights outcomes obtained from the application of our proposed model in finance, specially in hedging portfolio optimization.

Finally, we conclude the thesis by summarizing the contribution of this thesis and highlighting our main achievements. Additionally, we outline potential works for future research and development in this domain.

Chapter 2

Financial Concepts

In this chapter, I will present definitions of key financial concepts essential for the application section of the experimental results. Firstly, I will introduce the NIG distribution and delve into the option pricing formula within the framework of NIG. Then, I will explain the delta and gamma hedging strategies, along with the formula for delta-gamma hedging under the NIG assumption.

2.1 Normal Inverse Gaussian distribution

The NIG distribution, introduced by Barndorff-Nielsen (1977), is a non-Gaussian distribution and a subclass of Generalized Hyperbolic distributions. The NIG distribution has four parameters that allow for both skewness and higher kurtosis than the Normal distribution, making it more flexible in modeling data with asymmetry and heavy tails. This flexibility makes the NIG distribution particularly applicable in finance, where many financial assets exhibit semi-heavy tails Godin et al. (2012).

A random variable X follows a NIG distribution, denoted as $X \sim NIG(\alpha, \beta, \mu, \delta)$, where α is the tail heaviness, β is the skewness, μ is the location and δ is the scale. The Probability Density Function (PDF) of a NIG distribution is expressed as:

$$\phi^{NIG}(x;\alpha,\beta,\mu,\delta) = \frac{\alpha\delta e^{\delta\gamma}}{\pi}\frac{K_1(\alpha\sqrt{\delta^2+(x-\mu)^2}}{\sqrt{\delta^2+(x-\mu^2)}}e^{\beta(x-\mu)}, x \in R. \tag{1}$$

Here, $K_\lambda(x)$ represents the modified Bessel function of the third kind with index λ, defined as:

$$K_\lambda(x) = \frac{1}{2}\int_0^\infty u^{\lambda-1}e^{-\frac{1}{2}x(u^{-1}+u)}du, x > 0, \tag{2}$$

Also, $\mu \in R$ represents the location, $0 \leq |\beta| < \alpha$ indicates the skewness, $\alpha > 0$ is the shape and $\delta > 0$ is the scale parameter.

Moreover, the CDF and survival function of a NIG are respectively as follow:

$$\Phi^{NIG}(x;\alpha,\beta,\mu,\delta) = \int_{-\infty}^{x}\phi^{NIG}(y;\alpha,\beta,\mu,\delta)dy, \tag{3}$$

$$\overline{\Phi^{NIG}}(x;\alpha,\beta,\mu,\delta) = \int_{x}^{\infty}\phi^{NIG}(y;\alpha,\beta,\mu,\delta)dy. \tag{4}$$

Additionally, the mean and variance are given by:

$$E(x) = \mu + \frac{\delta\beta}{\sqrt{\alpha^2-\beta^2}}, \tag{5}$$

$$Var(x) = \mu + \frac{\delta\alpha^2}{(\alpha^2-\beta^2)^{\frac{3}{2}}}. \tag{6}$$

The NIG is a normal variance-mean mixture distribution. If X has a NIG distribution, then $X \sim N(\mu+\beta w, w)$ where w is an Inverse Gaussian distribution. In other words, the NIG can be constructed by the following formula:

$$\mu + \beta w + \sqrt{w}Z \sim NIG, \tag{7}$$

where Z follows a normal distribution. To simulate an Inverse Gaussian random variable with

parameters a and b, denoted as $IG(a, b)$, according to Schoutens (2003), we can use the following algorithm of Michael, Schucany, and Haas (1976):

(1) Generate a standard Normal random number v.

(2) Set $y = v^2$.

(3) Set $x = \frac{a}{b} + \frac{y}{2b^2} - \frac{\sqrt{4aby+y^2}}{2b^2}$.

(4) Generate a uniform random number u.

(5) If $u \leq a/(a+xb)$, then return the number x as the $IG(a, b)$ random number, otherwise return $\frac{a^2}{b^2 x}$ as the IG(a, b) random number.

This is a useful property of NIG to simulate NIG random variables.

2.2 Option price under NIG assumption

When logarithmic stock returns follow a NIG distribution, the dynamics of both risky asset S_t and riskless assets B_t can be expressed as follows:

$$S_t = S_0 e^{\sum_{k=1}^{t} Z_k}, \tag{8}$$

$$B_t = e^{rt}, \tag{9}$$

where Z_k is a NIG Lévy process defined as follows.

Definition 2.1 (NIG Lévy process) *Let us consider a filtered probability space denoted by* $(\Omega, F, (F_t)_{t\geq 0}, P)$. *An adapted cadlag R-valued process* $X = \{X(t)\}_{t\geq 0}$ *with* $X(0) = 0$ *is a NIG Lévy process if the increments of X are independent, stationary and distributed as*

$$\phi^{NIG}(\alpha, \beta, \delta t, \mu t). \tag{10}$$

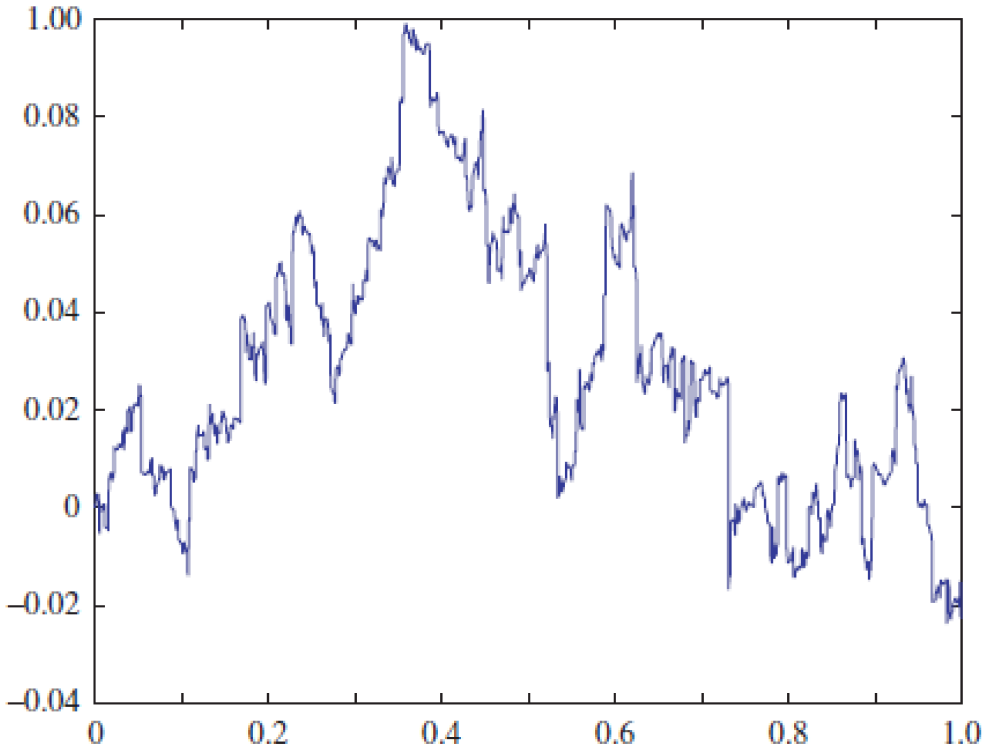


Figure 2.1: A sample path of a NIG process, Schoutens (2003).

Figure 2.1 shows a simulated path of a NIG process.

Under the Black-Scholes model based on the normal distribution, the market is complete. So according to the Fundamental Theorem of Asset Pricing, there exists a unique arbitrage measure, leading to the derivation of the Black-Scholes option pricing formula.

By contrast, the model referenced as (8) is not complete, resulting in an infinite number of martingale measures. To address this, a common approach is the use of the mean-correction risk-neutral measure, as proposed by Schoutens (2003). Under such measure, the original parameter μ is replaced with $\mu + \theta^*$, where θ^* is determined as follows:

$$\theta^* = r - \mu + \delta(\sqrt{\alpha^2 - (\beta+1)^2} - \sqrt{\alpha^2 - \beta^2}). \tag{11}$$

Subsequently, a non-Gaussian Black-Scholes-type formula for a European call with strike K is derived, as presented by Godin et al. (2012):

$$C(t,S_t) = S_t\left(1-\Phi^{NIG}\left(\ln\left(\frac{K}{S_t}\right);\ \alpha,\ \beta+1,\ \delta(T-t),\ [\mu+\theta^*](T-t)\right)\right)$$
$$- Ke^{-r(T-t)}\left(1-\Phi^{NIG}\left(\ln\left(\frac{K}{S_t}\right);\ \alpha,\ \beta,\ \delta(T-t),\ [\mu+\theta^*](T-t)\right)\right), \quad (12)$$

where θ^* is defined as (11) and Φ^{NIG} is the CDF of NIG distribution as expressed in (3).

2.3 Hedging strategy

Hedging in finance is a risk management strategy used to offset potential losses from adverse price movements in the market. For instance, consider holding a short position (seller) in an option contract on an underlying asset, with a specified expiration date and strike price. If the option holder (buyer) decides to exercise the option, you are obliged to fulfill the contract, which could result in losses if the market moves unfavorably. To mitigate this risk, a hedging approach involves taking a long position (buyer) in other options and in the same underlying asset, essentially purchasing the actual asset on which the option is based. This establishes a balancing position that helps reduce potential losses associated with the short option position.

A replication portfolio is a collection of investments designed to replicate the initial position and it will be re-balanced through time. Greeks are mathematical metrics representing the sensitivity of option prices with respect to risk factors, and can be used to construct a replication portfolio. The most commonly used Greeks are Delta and Gamma.

Definition 2.2 (Delta) *Delta is the sensitivity of the option price with respect to the price of the underlying asset,*

$$\Delta_t = \frac{\partial C_t}{\partial S_t}. \quad (13)$$

Definition 2.3 (Gamma) *Gamma is the second order sensitivity of the option price with respect*

to the price of the underlying asset,

$$\Gamma_t = \frac{\partial^2 C_t}{\partial S_t^2} = \frac{\partial \Delta_t}{\partial S_t}. \tag{14}$$

Delta-hedging is highly effective in the presence of minor fluctuations in the underlying stock price. Conversely, a Gamma-hedging strategy is better suited to offset large fluctuations of the underlying asset.

Now, let us consider constructing a replication portfolio using the Delta and Gamma Greeks. Imagine you are holding a short position in call option C on stock S and intend to hedge your position by employing a delta-gamma hedging strategy. Delta-gamma hedging requires utilizing two distinct assets to construct a replicating portfolio: stocks (S) and additional options on the same underlying stock (S), denoted as D. We will determine the quantities of these two assets needed to hedge the position on option C over time.

A replicating portfolio consists of (θ^s) shares of the underlying S, (θ^D) of option D on the same underlying stock S, and a certain amount of cash in the money market account at each time t.

To determine the quantities of both assets, we aim to construct Delta-Gamma neutral portfolios as follows:

$$\begin{cases} 0 = \Delta^{(pf)} = \theta^C \frac{\partial C_t}{\partial S_t} + \theta^D \frac{\partial D_t}{\partial S_t} + \theta^S \frac{\partial S_t}{\partial S_t}, \\ 0 = \Gamma^{(pf)} = \theta^C \dfrac{\partial^2 C_t}{\partial S_t^2} + \theta^D \dfrac{\partial^2 D_t}{\partial S_t^2} + \theta^S \dfrac{\partial^2 S_t}{\partial S_t^2}, \end{cases} \tag{15}$$

where $\Delta^{(pf)}$ and $\Gamma^{(pf)}$ represent the delta and gamma of the replication portfolio, respectively. θ^C denotes the number of shares of option C.

Since $\frac{\partial S_t}{\partial S_t} = 1$ and $\dfrac{\partial^2 S_t}{\partial S_t^2} = 0$, and by definition of delta and gamma, we have:

$$\begin{cases} 0 = \Delta^{(pf)} = \theta^C \Delta^C + \theta^D \Delta^D + \theta^S, \\ 0 = \Gamma^{(pf)} = \theta^C \Gamma^C + \theta^D \Gamma^D. \end{cases} \tag{16}$$

If we short one call option C, then $\theta^C = -1$ and the second equation gives

$$\theta^D = \frac{\Gamma^C}{\Gamma^D}. \tag{17}$$

The first equation then becomes

$$\theta^S = \Delta^C - \frac{\Gamma^C}{\Gamma^D}\Delta^D. \tag{18}$$

2.4 NIG delta-gamma hedging

The Delta of a European call option C_t based on the mean-correcting risk-neutral measure (see Eq 12) is provided by Godin (2016) as follows:

$$\Delta_t = \frac{\partial C_t}{\partial S_t} = 1 - \Phi^{NIG}\left(ln\left(\frac{K}{S_t}\right);\ \alpha,\ \beta+1,\ \delta(T-t),\ [\mu+\theta^*](T-t)\right). \tag{19}$$

Then, it is straightforward to derive its NIG gamma formula, expressed as follows:

$$\Gamma_t = \frac{\partial^2 C_t}{\partial S_t^2} = \frac{1}{S_t}\phi^{NIG}\left(ln\left(\frac{K}{S_t}\right);\ \alpha,\ \beta+1,\ \delta(T-t),\ [\mu+\theta^*](T-t)\right). \tag{20}$$

2.4.1 Hedging error

Consider the previous example in which you hold a short position in a call option with a strike price K on the underlying stock S at maturity time T. If the stock price rises, resulting in $S_T > K$ at maturity, the option holder may choose to exercise the option, purchasing the stock at the predetermined strike price K. In this situation, a loss of $S_T - K$ is incurred. To offset this potential loss, a hedging portfolio is constructed as outlined above.

In a complete market, perfect hedging is achievable. This implies that the value of the hedging portfolio should exactly align with the option's payoff at the terminal time. However, in reality, this equivalence is often not achievable, leading to a hedging error. In mathematical terms, if V_T

denotes the value of the hedging portfolio at the terminal time, the hedging error can be defined as:

$$\text{Hedging Error} = \max(S_T - K, 0) - V_T. \tag{21}$$

Minimizing the hedging error plays an important role in financial risk management, as it directly impacts the effectiveness of hedging strategies and the overall stability of investment portfolios. One key metric used to assess and mitigate this error is Conditional Value at Risk (CVaR), which provides a comprehensive measure of risk by quantifying the expected losses beyond a certain threshold. In the next chapter, we will provide a detailed explanation of CVaR and estimation methods for it.

Chapter 3

Estimating Risk Measures

This chapter provides the necessary background on the main concepts of risk measures. First, we begin by giving the definition of two well-known risk measures, VaR and CVaR. Then we revisit the sample average and extreme value theory (EVT) estimation approaches and review some heavy tail distributions. Finally, we explore sample averaging and extreme value theory (EVT) estimation approaches.

3.1 Risk measures: VaR and CVaR

Risk management plays a vital role in investment decision-making and two widely used risk measures for this purpose are VaR and CVaR. VaR represents the maximum loss of a portfolio at a given confidence level denoted by α (e.g., 0.95, 0.99, or 0.999). By using VaR, investors can gain insight into the potential downside risks associated with their investments. More formally, the definition of VaR at confidence level $\alpha \in (0, 1)$ is expressed as follows.

Definition 3.1 (VaR) *Let X denote a random loss. The VaR at confidence level α is calculated as:*

$$\mathit{VaR}_\alpha(X) = \inf\{x \in \mathbb{R} | F_X(x) \geq \alpha\}, \tag{22}$$

where F_X is the cumulative distribution function (CDF) of X.

In statistical terms, VaR represents the quantile at level α of the cumulative distribution function F_X. However, VaR has some limitations, including its lack of coherence within the framework of risk measures proposed by Artzner et al. (1999). Here we review four desirable properties for risk measures. Those possessing all of them are defined as coherent risk measures. Let ρ be a risk measure and X be the usual loss random variable. The risk measure ρ is considered coherent if it satisfies the following four axioms:

(1) Monotonicity: For all X_1 and X_2, if $X_1 \leq X_2$, then $\rho(X_2) \geq \rho(X_1)$.

(2) Convexity: For all X_1, X_2 and $\lambda \in [0, 1]$, $\rho(\lambda X_1 + (1 - \lambda)X_2) \leq \lambda\rho(X_1) + (1 - \lambda)\rho(X_2)$.

(3) Positive homogeneity: For all $\lambda \geq 0$ and all X, $\rho(\lambda X) = \lambda\rho(X)$.

(4) Translation invariance: For all X and all real numbers α, $\rho(X + \alpha) = \rho(X) + \alpha$.

However VaR is non-subadditive, non-convex, and discontinuous, as pointed out by Sarykalin et al. (2008). Furthermore, it is not responsive to large losses, beyond the specified threshold. CVaR, also known as expected tail risk, or expected shortfall, is an alternative risk measure, introduced by Acerbi and Tasche (2002). CVaR is coherent and provides a more comprehensive view of risks by calculating the average of all losses that exceed the VaR threshold; it is widely used in practice.

Definition 3.2 (CVaR) *Assume X is absolutely continuous. The CVaR of X at confidence level α is*

$$CVaR_\alpha(X) = \mathbb{E}[X|X \geq VaR_\alpha(X)] = \frac{1}{1-\alpha}\int_\alpha^1 VaR_\gamma(X)d\gamma. \tag{23}$$

Moreover, Figure 3.1 shows the relationship between VaR and CVaR.

3.2 Estimation methods for CVaR

In this section, we investigate two significant approaches to estimate the CVaR: sample averaging (SA) and extreme value theory (EVT).

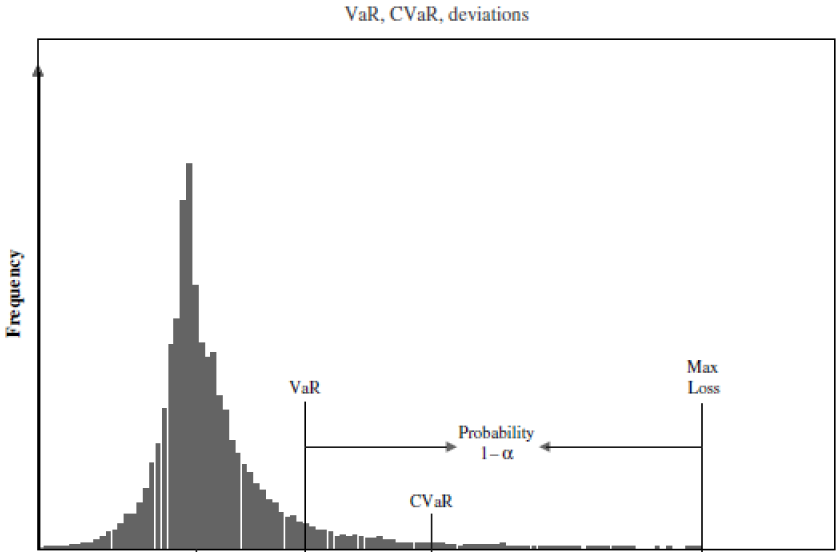


Figure 3.1: VaR and CVaR representation (Sarykalin et al., 2008).

3.2.1 Estimation by sample averaging

The SA method computes the empirical average exceedance of the sample observations above a specific threshold. Let $X_1, X_2, ..., X_n$ be a sequence of identical, independent random variables with a CDF F. The SA method proceeds as follows:

$$\widehat{CVaR}_{\alpha,n}(x) = \frac{\sum_{i=1}^{n} X_i \mathbb{1}_{\{X_i \geq \hat{q}_{\alpha,n}\}}}{\sum_{j=1}^{n} \mathbb{1}_{\{X_j \geq \hat{q}_{\alpha,n}\}}}, \tag{24}$$

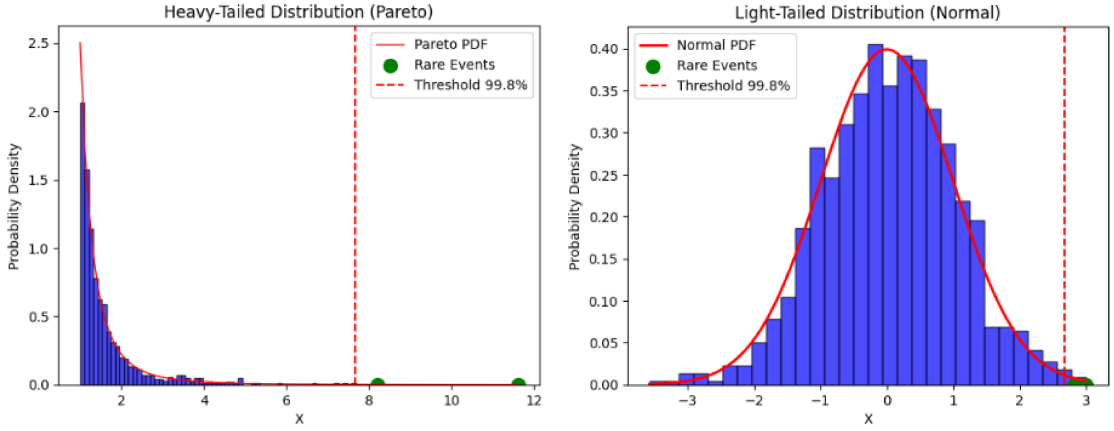
Here, $\hat{q}_{\alpha,n}(X)$ represents the empirical distribution quantiles given by:

$$\hat{q}_{\alpha,n}(X) = \widehat{VaR}_{\alpha,n}(X) = \inf\{x \in \mathbb{R} | \hat{F}_X^n(x) \geq \alpha\} = \min_i \{X_{(i)} | \hat{F}_X^n(X_{(i)}) \geq \alpha\} = X_{(\lceil \alpha n \rceil)}, \tag{25}$$

where $\widehat{F}_X^n(x)$ is the empirical cumulative distribution function of X given by:

$$\widehat{F}_X^n(x) = \frac{1}{n} \sum_{s=1}^{n} \mathbb{1}_{\{X_s \leq x\}}. \tag{26}$$

The values $X_{(1)}, ..., X_{(n)}$ represent the order statistics, which are the sample observations arranged in non-decreasing order. However, this method encounters an issue when the value of the confidence level approaches one. At this point, the number of observations above the quantile of level α becomes scarce, leading to imprecise estimates and a lack of accuracy. This issue is more prevalent in heavy-tailed distributions, where there are more data in the tail of the distribution. See Figure 3.2, which illustrates the difference between GPD and Normal distributions in the number of samples in the tails when α is large. As depicted in Figure 3.2, for a heavy-tailed distribution such as the Pareto distribution, extreme events are not easily observed when $\alpha = 0.998$. This stands in stark contrast to light-tailed distributions like the Normal distribution, where similar low-probability events are close to the mean and do not result in as catastrophic outcomes.



(a) Pareto distribution (b) Normal distribution

Figure 3.2: Heavy tailed (Pareto) vs light tailed (Normal) distributions.

To address this issue, Troop et al. (2022) proposed an extreme value theory (EVT) estimator with automated threshold selection method, which will be elaborated upon in the subsequent section.

3.2.2 Extreme value theory (EVT)

The focus in this section is on the relevant theorems, definitions and concepts of EVT. There are two fundamental theorems in EVT: the Fisher–Tippett–Gnedenko theorem (see Haan & Ferreira,

2006) and the Pickands–Balkema–de Haan theorem (Balkema & De Haan, 1974; Pickands III, 1975). The first theorem focuses on the maximum of a sample, while the second one centers on the asymptotic tail distribution, above a specific threshold. The Fisher–Tippett–Gnedenko theorem is as follows.

Theorem 3.1 (Fisher–Tippett–Gnedenko) *Consider a sequence of independent and identically distributed (i.i.d.) random variables $X_1, X_2, ..., X_n$ with a common CDF denoted by F. The CDF of the running maximum is given by $P(X_{max} \leq x) = F^n(x)$, where $X_{max} = max(X_1, ..., X_n)$. As the number $n \in \mathbb{N}$ of observations increases, there exists a sequence of real-valued constants $a_n > 0$ and b_n, such that the limiting distribution for the sample maximum evaluated at the sequence $a_n x + b_n$ converges to a non-degenerate distribution function $G(x)$ for each continuity point x of G. This is expressed as:*

$$\lim_{n\to\infty} F^n(a_n x + b_n) = G(x). \tag{27}$$

The limiting distribution G must be a generalized extreme value distribution (for some parameter ξ) and the class of distributions F that satisfy condition (27) *are called the maximum domain of attraction of G, denoted as $F \in MDA(G_\xi)$.*

Definition 3.3 (Generalized Extreme Value Distributions (GEVD)) *The generalized extreme value distribution with shape parameter $\xi \in \mathbb{R}$ is defined as follows:*

$$G_\xi(x) = \begin{cases} \exp(-(1+\xi x)^{\frac{-1}{\xi}}), & \text{if } \xi \neq 0, \\ \exp(-e^{-x}), & \text{if } \xi = 0, \end{cases} \tag{28}$$

with the support being all x such that $1 + \xi x > 0$, where ξ is known as a shape parameter. The quantity $\frac{1}{\xi}$ is known as the rate of tail decay.

GEVDs include three specific distributions: if $\xi > 0$, GEVD is a Fréchet distribution. In this case, F is a heavy-tailed distribution and its tail decays like a power function. For example the generalized Pareto distribution (GPD) and the Burr are well-known distributions that are in the

maximum domain of attraction of the heavy tail Fréchet distribution ($F \in MDA(G_\xi)$), see the Appendix for more details about GPD and Burr distributions.

If $\xi = 0$, we have a Gumbel distribution and F is a medium tailed distribution, with an exponential tail decay, such as the normal, exponential, and gamma distributions. If $\xi < 0$, then GEVD gives the Weibull family of distribution and F is short-tailed. The right endpoint of G_ξ is finite; this includes the uniform and beta distributions (see McNeil, 1997).

In the risk management area, heavy-tailed distributions ($\xi > 0$) are of interest since more rare events occur under such distributions. It is important to note that the expectation is not defined for $\xi > 1$, therefore to estimate CVaR we need to assume that $\xi \in [0, 1)$.

When the distribution F belongs to the maximum domain of attraction of G_ξ, a valuable approximation of the distribution of sample extremes, above a high threshold, becomes available.

Definition 3.4 (Excess distribution function) *The excess distribution function, denoted as F_u, represents the distribution of values above a high threshold u. Let X be a random variable with CDF F and let Y represent the magnitude of the exceedance above u, that is $Y = X - u$. We can express F_u as follow:*

$$F_u(y) = P(X - u \leq y | X > u) = P(X \leq y + u | X > u) = \frac{F(y + u) - F(u)}{1 - F(u)}. \tag{29}$$

Based on the Pickands–Balkema–de Haan theorem, the limiting excess distribution converges to a Generalized Pareto Distribution for large values of u.

Theorem 3.2 (Pickands–Balkema–de Haan theorem) *Under the MDA condition in* (27)*, we can find a positive measurable function $\sigma(u)$ such that*

$$\lim_{u \to y_0} \sup_{y_0 \in [0, y_0 - u]} |F_u(y) - G_{\xi, \sigma(u)}(y)|, \tag{30}$$

where $y_0 = sup\{y \in \mathbb{R}; F(y) < 1\} \leq \infty$ and $G_{\xi,\sigma(u)}(y)$ represents the CDF of the GPD with parameters (ξ, σ), as detailed in the next definition.

Definition 3.5 (Generalized Pareto Distribution(GPD)) *The GPD has three parameters: a location parameter μ, a scale parameter σ, and a shape parameter ξ. If $\mu = 0$, the CDF and PDF of a GPD are given by, respectively:*

$$G_{\xi,\sigma}(x) = \begin{cases} 1 - (1 + \frac{\xi x}{\sigma})^{\frac{-1}{\xi}}, & \text{if } \xi \neq 0, \\ 1 - e^{\frac{-x}{\sigma}}, & \text{if } \xi = 0, \end{cases} \tag{31}$$

where the support is $x \geq 0$, for $\xi \geq 0$, and $0 \leq x \leq -\frac{\sigma}{\xi}$, for $\xi \leq 0$, and the PDF is:

$$g_{\xi,\sigma}(x) = \begin{cases} \frac{1}{\sigma}(1 + \frac{\xi x}{\sigma})^{\frac{-1}{\xi}-1}, & \text{if } \xi \neq 0, \\ \frac{1}{\sigma}e^{\frac{-x}{\sigma}}, & \text{if } \xi = 0. \end{cases} \tag{32}$$

It is important to note that if $X \sim GPD(\xi, \sigma)$, then the conditional exceedance $X - u|X > u \sim GPD(\xi, \sigma + \xi u)$, meaning that the excess distribution of a GPD random variable is a GPD with the same shape parameter, and a scaling parameter that grows linearly with the threshold u, please see Proposition A.1 in the Appendix.

Next, consider a key corollary of EVT that provides a practical method for approximating the tail behavior, especially in estimating VaR and CVaR, see (McNeil, Frey, & Embrechts, 2015, Section 7.2).

Corollary 3.1 (EVT approximation) *If $\xi \in [0, 1)$ and $\sigma = \sigma(u)$ satisfy the conditions of Theorem 3.2, then the VaR and CVaR of a random variable $X \in MDA(G_\xi)$ can be approximated using EVT, respectively, as follows:*

$$q_{u,\alpha} = \begin{cases} u + \frac{\sigma}{\xi}(s_{u,\alpha}^{\xi} - 1), & \text{if } \xi \neq 0, \\ u + \sigma \log s_{u,\alpha}, & \text{if } \xi = 0, \end{cases} \tag{33}$$

and

$$c_{u,\alpha} = \begin{cases} u + \frac{\sigma}{1-\xi}(1 + \frac{s_{u,\alpha}^{\xi} - 1}{\xi}), & \text{if } \xi \neq 0, \\ u + \sigma(\log s_{u,\alpha} + 1), & \text{if } \xi = 0, \end{cases} \tag{34}$$

where $s_{u,\alpha} = \frac{1-F_X(u)}{1-\alpha}$. *Here* $q_{u,\alpha}$ *and* $c_{u,\alpha}$ *represent the approximations of VaR and CVaR, respectively, relying on threshold* u.

3.3 EVT estimation of CVaR with automated threshold

Consider now the POT estimation of CVaR using Corollary 3.1. Let $X_1, ..., X_n$ represent the sample set composed of i.i.d. observations. To approximate CVaR through POT, the initial step involves defining an appropriate threshold denoted as u. Subsequently, identify the excess sample values $Y_1, ..., Y_k$, where $Y_i = X_i - u; X_i > u$. Based on EVT, we know that the exceedances over a high threshold follow the GPD with parameters (ξ, σ), asymptotically, and are independent. Maximum likelihood estimation (MLE) is one approach to estimate the parameter values of an assumed probability distribution, based on available data. This method involves maximizing the log-likelihood function with respect to the parameters. Hence, by using MLE we can find an approximation of GPD parameters as follow:

$$(\hat{\xi}_u^{(n)}, \hat{\sigma}_u^{(n)}) = \arg\max_{\xi,\sigma} \sum_{i=1}^{n} log\, g_{\xi,\sigma}(Y_i). \tag{35}$$

In the above equation, $g_{\xi,\sigma}$ denotes the PDF of the GPD, as defined in Equation (32), where (ξ, σ) lie within the range $(-\infty, \infty) \times (0, \infty)$. For deeper insights, see Haan and Ferreira (2006, Section 3.4). It is important to note that a closed-form solution to the optimization problem (35) is not available. As a result, one common approach to estimate these parameters is through numerical methods using standard software packages such as `scipy.stats.genpareto` in Python.

Choosing an appropriate threshold is a crucial step in the POT method. It greatly influences the accuracy of the EVT estimation. A recent study by Troop et al. (2022), introduces a novel approach to threshold selection within the context of CVaR estimation with EVT. They use an automated method based on the paper of Bader et al. (2018) to select the threshold for the POT method, as follows. Consider a fixed set of $u_1 < ... < u_k$ as a candidate threshold where each u_i is the quantile of percentiles $q_1, ..., q_k$, that is $u_i = F_n^{-1}(q_i)$ where F_n^{-1} is the empirical quantile

function. Then there are k_i excess samples over each threshold u_i, $i = 1, ..., k$. We want to measure how closely the exceedances follows a GPD. To do so, the Anderson-Darling (AD) statistic can help us. The null hypotheses for the AD test for each respective test $i = 1, ..., k$ is as follows:

$$H_0^{(i)} : \text{The distribution of the } n_i \text{ exceedances above } u_i \text{ follows a GPD.}$$

Let $\hat{\theta}_i = (\hat{\xi}_{u_i}^{(n)}, \hat{\sigma}_{u_i}^{(n)})$ be the estimated MLE parameters for each threshold. Then the AD test statistic is used to compare the empirical distribution over each threshold with the GPD. Assume that $y_{(1)} < ... < y_{(k_i)}$ are the ordered threshold excesses for test i. Apply the transformation $z_{(j)} = G_{\hat{\theta}_i}(y_{(j)})$, for $j = 1, ..., k_i$ where G represents the CDF of the GPD. The AD statistic is given by:

$$A_n^2 = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)[\log(z_{(i)}) + \log(1 - z_{(n+1-i)})]. \tag{36}$$

Then we should use the corresponding p-value for each test i if the exceedances come from GPD, where $i = 1, ..., l$. Refer to Choulakian and Stephens (2001) to find the lookup table for p-values, or also it can be computed. Then apply the Forward-Stop rule of Troop et al. (2022), which is a modification version of G'Sell, Wager, Chouldechova, and Tibshirani (2016) to choose the threshold. Let $p_1, ..., p_l$ be the p-values for each test. Then the Forward-Stop rule calculates $\hat{w}_F$ as follow:

$$\hat{w_F} = \max\{w \in I| - \frac{1}{w}\sum_{i=1}^{w}\log(1 - p_i) \leq \gamma\}, \tag{37}$$

where γ is a specific level and $I \subset \{1, ..., l\}, I \neq \varnothing$.

Since CVaR is not defined for $\xi \geq 1$, we set a cutoff value ξ_{max}, where $\xi_{\text{max}} < 1$. This ensures reliable CVaR estimation within reasonable bounds. To determine the CVaR estimate, we compare the estimated ξ for each test with ξ_{max}. If ξ exceeds ξ_{max}, the estimation and its corresponding threshold are discarded. In this scenario, the SA method in Equation (24) above is employed to provide a fallback solution. However, if the estimated ξ is within the acceptable range ($\xi < \xi_{\text{max}}$), further steps are taken. The process involves the AD statistic as defined by Equation (36), along with the corresponding p-values for each test. Then if there is no existing $\hat{w_F}$, no rejection occurs

and the threshold corresponding to $u_{\min(I)}$ is selected. But if $\hat{w_F}$ = max(I) then $u_{\max(I)}$ is selected (see Troop et al., 2022, for more details).

Now we have estimates of the GPD parameters and we have selected the threshold u. Then we have the following definition for the CVaR estimate which stems from (34).

Definition 3.6 (POT CVaR estimator) *The CVaR estimator based on the POT method is*

$$\hat{c}_{u,\alpha}^{(n)} = \begin{cases} u + \frac{\hat{\sigma}_u^{(n)}}{1-\hat{\xi}_u^{(n)}} \left(1 + \frac{1}{\hat{\xi}_u^{(n)}} \left[\left(\frac{1-\hat{F}_n(u)}{1-\alpha} \right)^{\hat{\xi}_u^{(n)}} - 1 \right] \right), & \textit{if } \xi \neq 0, \\ u + \hat{\sigma}_u^{(n)} \left[\log \left(\frac{1-\hat{F}_n(u)}{1-\alpha} \right) + 1 \right], & \textit{if } \xi = 0, \end{cases} \tag{38}$$

where $(\hat{\xi}, \hat{\sigma})$ represents the MLE parameter estimates, and α denotes the confidence level, satisfying the condition $\alpha > \hat{F}_n(u)$.

Chapter 4

Reinforcement Learning

This section begins by examining the concept of Markov decision processes, a fundamental framework in sequential decision-making problems. Subsequently, we dive into reinforcement learning techniques employed for addressing Markov decision problems, including methods like value approximation and policy approximation. Additionally, we compare reinforcement learning and dynamic programming in this chapter (for more details see Sutton & Barto, 2018).

4.1 Markov decision processes

Sequential decision making problems are modelled within the formulation of Markov decision processes (MDP). This framework originates from the pioneering research conducted by Bellman (1954). In the MDP framework, an agent makes a sequence of decisions to maximize (minimize) its performance measure. The definiton of MDP refers to Puterman (2014) as follows.

Definition 4.1 (Markov Decision Process (MDP)) *MDP involves a tuple (S, A, R, P, γ) where*

(1) S is a state space,

(2) A is an action space,

(3) R is the set of rewards,

(4) P is the matrix of transition probabilities between states characterizing the evolution of states and rewards:

$$P : S \times R \times S \times A \to [0, 1],$$

(5) γ is a discount factor.

Figure 4.1 illustrates how a MDP works. As shown, in an MDP problem, there is an agent (also called a decision maker or controller) and an environment (also known as a system). At a specific time t, the system is in a state $S_t \in S$. The agent then selects an action $A_t \in A$, resulting in an immediate reward R_{t+1} (or cost). Then the system randomly transitions to state $S_{t+1} \in S$ at time $t + 1$. This leads to the following sequence:

$$S_0, A_0, R_1, S_1, A_1, R_2, S_2, A_2, \ldots$$

.

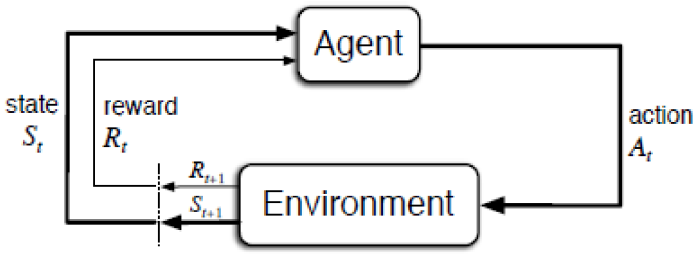


Figure 4.1: The agent–environment interaction in a Markov decision process (Sutton & Barto, 2018).

A MDP is considered finite if the sets S, A, and R contain a finite number of elements. For a finite MDP, the transition probabilities are given by

$$\mathbb{P}[S_{t+1} = s', R_{t+1} = r | S_t = s, A_t = a], \tag{39}$$

for all $s, s' \in S$, $a \in A$ and $r \in R$. These transition probabilities define the dynamics of the MDP. Note that transition probabilities in (39) only depend on the most recent state s and action a, and

they are not affected by the history of states and actions. This feature is known as the Markov property. Next, we explain another important concept which is *Policy*.

4.2 Policy

At each time step, the agent chooses its actions based on a specific strategy known as a policy, denoted as π. A policy indicates the probability of choosing action a when the system is in the state s. There exist two categories of policies: stochastic and deterministic. In a stochastic policy, represented as $\pi(a|s)$, the agent selects actions based on a probability distribution over the action space for a given state. This means that the same state could lead to different actions in different instances of interaction. On the other hand, deterministic policies select a specific action for a given state, which can be expressed as $\pi(s) = a$, where $\pi(s)$ illustrates the action chosen in state s, and a is a deterministic action.

4.3 Reinforcement learning problem

The previous section provides an overview of the MDP. This section is dedicated to examining the methods applied to solve MDP.

We can identify two primary approaches for tackling a MDP: reinforcement learning and dynamic programming (DP) methods. However, DP methods come with some limitations. They require precise knowledge of the environment, specifically involving transition probabilities in their calculations, which are often unavailable in practical scenarios. Additionally, if the number of states and rewards is extensive, solving the MDP becomes computationally impractical.

RL, a significant branch of machine learning, has emerged as a set of learning methods that confront these issues. In contrast to DP, RL is a model-free approach that relies on experimentation rather than requiring a complete model. Furthermore, RL algorithms can be applied to large-scale problems effectively. RL involves a dynamic interplay between an agent and its environment and learns through trial and error.

The agent selects actions, which in turn result in receiving rewards or penalties based on a performance metric. This mechanism serves as a way to obtain feedback from the environment, allowing the agent to assess the efficacy of its chosen actions. In terms of action selection, there is an exploration and exploitation trade-off. Exploration entails selecting greedily actions that maximize discounted rewards or minimize penalties. On the other hand, exploitation involves selecting actions that have been chosen before re-evaluating them. The next section discusses the RL algorithms.

4.4 Model-free RL methods

This section focuses on model-free RL algorithms. The main goal of RL is to find the optimal policy that optimizes the objective function of the problem. In the cases where a perfect environment model is unknown, it becomes necessary to interact with the environment, obtaining experiences from the MDP in order to gather statistical insights about the unknown model. Model-free RL methods can be categorized into two classes:

(1) Value approximation,

(2) Policy approximation.

It is worth noting that there are also hybrid models, which combine the above-mentioned methodologies. In the following sections we discuss these two classes of algorithm.

4.4.1 Value approximation

Value approximation, often referred to as *value learning algorithms*, is a class of reinforcement learning methods that estimate value functions (either state values or action values) to optimize a policy in a reinforcement learning setting. These algorithms can be categorized into on-policy and off-policy methods based on whether the policy being evaluated or improved on is the same as the one used to generate the data.

On-policy methods involve the evaluation or improvement of the policy that is currently in use for decision-making during data generation. In other words, in these methods, the agent interacts with the environment, collects data, and then uses this data to evaluate or improve the same policy. An example of an on-policy method is SARSA, which stands for State-action-reward-state-action.

Off-policy methods, on the other hand, involve the evaluation or improvement of a policy that is different from the one used to generate the data. In these methods, the agent collects experiences using one policy, often referred to as the *behavior policy*, and then leverages this data to enhance a different policy, known as the *target policy*. A well-known example of an off-policy method is Q-learning,

4.4.2 Policy approximation

The main goal of policy approximation methods is to learn directly from various policies in order to maximize a performance measure, with the ultimate aim of reaching the optimal policy. In other words, this method can select actions without consulting a value function.

In some cases, these methods have advantages over value approximation methods. One advantage is that, in many applications, the policy function might be simpler, compared to the value function, then it is easier to approximate the policy function. Another advantage is that these methods make it easy to incorporate prior knowledge about the shape of the optimal policy. In addition to the practical advantages, these methods also have the theoretical advantage that they are guaranteed to converge, at least to a local optimum, which may be good enough in practice. One of the extensively studied methods for policy approximation is the policy gradient algorithm, which is explored in the next chapter.

Chapter 5

Risk-Aware Policy Gradient Method

This chapter presents the *Risk-Aware Policy Gradient algorithm*, the fundamental algorithm employed to address the risk-aware Markov decision problem. To begin, we provide an overview of the policy gradient method. Subsequently, we dive into the Risk-Aware Policy Gradient algorithm, followed by an exploration of policy optimization techniques. Finally, we provide a comprehensive explanation of the gradient estimation method utilized for optimizing the objective function.

5.1 Policy gradient methods

As mentioned in the previous chapter, policy gradient methods directly learn policies to find the optimal policy θ^*. In these methods, the optimal policy is approximated using a parameterized policy.

A parameterized policy refers to a policy that is defined by a vector of parameters $\theta \in R^{d'}$. These parameters determine the behavior of the policy, influencing how the agent selects actions based on the observed states and the goal is to optimize the policy parameters which maximize or minimize a certain objective function $J : \theta \to R$ such that

$$\theta^* = \arg\max_{\theta \in \Theta} J(\theta). \tag{40}$$

5.2 Risk aware policy gradient method

Making decisions while considering potential risks is extremely important in various fields, including finance. In the context of standard (also known as *risk-neural*) reinforcement learning problems, the agent is often designed to maximize a cumulative discounted reward without being influenced by concerns about risk. However, the risk aware policy gradient method takes into account potential risks associated with actions and decisions made by a RL agent. In this method, the objective function involves a risk measure of the random cost, hence, the goal is to seek the optimal policy θ^* that optimizes (minimize) a performance measure $J : \theta \to R$ of Equation (40).

5.2.1 Policy optimization

Problems (40) can be addressed using optimization methods such as stochastic gradient ascent or ADAM algorithm, both of which are powerful approaches to find the minimum value of a function.

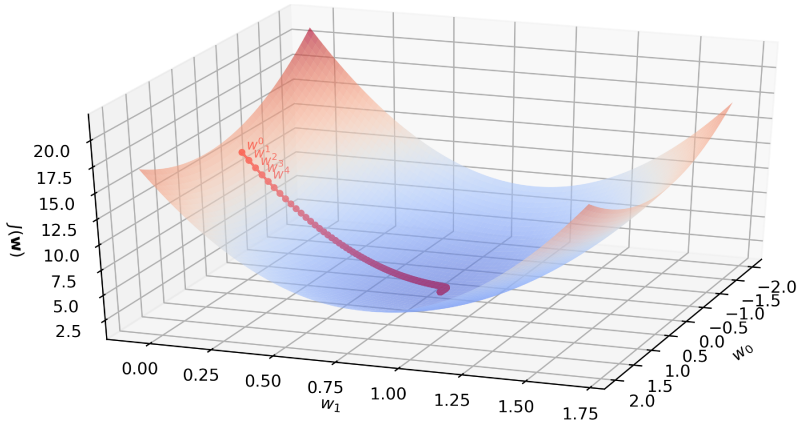


Figure 5.1: Path of converging optimization algorithm.

In this section, we provide a brief description of the ADAM algorithm. ADAM introduced by Kingma and Ba (2014) stands for *Adaptive Moment Estimation*. It combines concepts from both, momentum-based optimization and RMSProp, resulting in an efficient optimization method.

ADAM adjusts the learning rate at each iteration by computing the gradient of the objective function, along with the first and second moment estimates of the gradient and updates the parameters. The algorithm for ADAM is provided in Algorithm 1.

Algorithm 1: ADAM algorithm from paper Kingma and Ba (2014).

g_t^2 indicates the elementwise square $g_t \odot g_t$. Good default settings for the tested machine learning problems are $\alpha = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. All operations on vectors are elementwise. With β_1^t and β_2^t we denote β_1 and β_2 to the power t.

Require: α: stepsize

Require: $\beta_1, \beta_2 \in [0, 1)$: Exponential decay rates for the moment estimates

Require: $f(\theta)$: Stochastic objective function with parameters θ

Require: θ_0: Initial parameter vector

$m_0 \leftarrow 0$ (Initialize 1st moment vector)

$v_0 \leftarrow 0$ (Initialize 2nd moment vector)

$t \leftarrow 0$ (Initialize timestep)

while θ_t *not converged* **do**

$t \leftarrow t + 1$

$g_t \leftarrow \nabla_\theta f_t(\theta_{t-1})$ Get gradients w.r.t. stochastic objective at timestep t,

$m_t \leftarrow \beta_1.m_{t-1} + (1 - \beta_1).g_t$ (Update biased first moment estimate)

$v_t \leftarrow \beta_2.v_{t-1} + (1 - \beta_2).g_t^2$ (Update biased second raw moment estimate)

$\hat{m}_t \leftarrow \frac{m_t}{1-\beta_1^t}$ (Compute bias-corrected first moment estimate)

$\hat{v}_t \leftarrow \frac{v_t}{1-\beta_2^t}$ (Compute bias-corrected second raw moment estimate)

$\theta_t \leftarrow \theta_{t-1} - \alpha.\frac{\hat{m}_t}{\sqrt{\hat{v}_t}+\epsilon}$ (Update parameters)

end

Return: θ_t (Resulting parameters)

The ADAM algorithm utilizes the gradient of the objective function at each iteration. However, computing the gradient directly is not always feasible for all functions. Therefore, it becomes necessary to approximate it, a topic that will be discussed in the next section.

5.3 Estimating the gradient of the objective function

Finite differences is one of the common approaches to compute the gradient of the objective function. It involves approximating the gradient by computing the function's values at nearby points and taking their differences. There are three different types of methods in finite difference approximation such as forward differences, backward differences, and central differences. For example, forward difference approximation method for Equation (40) is as follow:

$$\widehat{\nabla J(\theta)} \approx \frac{\widehat{J}(\theta + \epsilon) - \widehat{J}(\theta)}{\epsilon}, \tag{41}$$

where $\epsilon > 0$ is a small step size and θ is the parameter. This is a useful approximation method when the gradient of a function does not exist.

Chapter 6

Experimental Results

This chapter introduces our proposed algorithm designed to address a risk-aware Markov decision problem using the risk aware policy gradient algorithm discussed in Chapter 5. This chapter is divided into two main parts. In the first part, the algorithm is implemented on a simulated dataset generated from GPD and Burr, which are two heavy-tailed distributions. This allows us to observe how EVT enhances the accuracy of a policy gradient algorithm. Once the behavior of the algorithm is confirmed within this controlled environment, in the second part we explore the application of the risk-aware policy gradient algorithm integrated by EVT in finance, specifically focusing on the NIG Lévy process. We define a numerical example of a hedging problem with the objective of identifying optimal hedging solutions in a highly volatile market.

6.1 Simulation analysis

In this section, the focus is on identifying the optimal policy within a Markov decision process, aimed at minimizing a risk measure, specifically CVaR. We start by defining the key elements of the core of the problem.

For simplicity, we consider the special case with no state and one-dimension action. Here the policy is θ, and actions are defined by $a = \tilde{a}(\theta)$, where $\tilde{a}(\theta) = \theta$. In this case, the action taken is directly determined by the parameter θ. So, for each possible value of θ, there corresponds a

unique action a. The cost, denoted as R, is associated with the chosen action and is a function of θ, generated from a distribution.

At each time step, the agent following the policy, selects an action that incurs a cost, subsequently contributing to the overall risk. We consider CVaR as the risk measure and the goal of algorithm is to minimize this risk. To achieve this objective, we need to identify the optimal policy that guides the agent towards actions leading to minimized costs. In mathematical terms, the optimization problem can be formulated as:

$$\min_{\theta} J(\theta) = \min_{\theta} \text{CVaR}_{\alpha}(R). \tag{42}$$

where θ represents the policy that needs to be learned.

We employ a deterministic policy gradient, utilizing a sequence of policies $\{\theta^{(n)}\}_{n\geq 1}$ that are updated iteratively via ADAM Algorithm 1. The Monte Carlo algorithm is employed to estimate the CVaR for the objective function. Thus, to obtain an estimate of a single point of the objective function CVaR, we simulate a substantial number of independent and identically distributed (i.i.d.) data points from GPD and Burr heavy-tailed distributions, which are of particular interest to us. Heavy-tailed distributions often produce extreme values, which can lead to potentially catastrophic risks, a significant concern in financial institutions.

In this case, since there is no closed form of the derivative of CVaR for all distributions, then we need to compute an approximation of derivative of CVaR with respect to the policy, i.e.

$$\widehat{\nabla J(\theta^{(n)})} = \nabla \widehat{CVaR(\theta^{(n)})}. \tag{43}$$

This approximation is obtained through a finite-difference method. We begin with an initial policy denoted as $\theta^{(0)}$. Then, at each iterations, we have a new policy and for each policy, first we estimates CVaR function derived from Monte-Carlo simulations and EVT or SA methods. After

that, we employ the finite-difference method as follows:

$$\nabla \widehat{CVaR}(\theta^{(n)}) \approx \frac{\widehat{CVaR}(\theta^{(n)} + \epsilon) - \widehat{CVaR}(\theta^{(n)})}{\epsilon}. \tag{44}$$

where $\epsilon > 0$ and $\theta^{(n)}$ is the policy parameter at iterations n.

6.1.1 GPD

GPD is a class of heavy-tailed distributions characterized by three parameters: shape (ξ), scale (σ), and location (μ), as discussed in Section 3.5. Figure 6.1 shows PDF of the GPD for a fixed value of ξ and varying values of σ. As indicated by the figure, as σ decreases, the density becomes lighter, resulting in a decrease in CVaR. Consequently, we consider σ as the policy parameter and set the location parameter μ to 0, while fixing the shape parameter at $\xi > 0$.

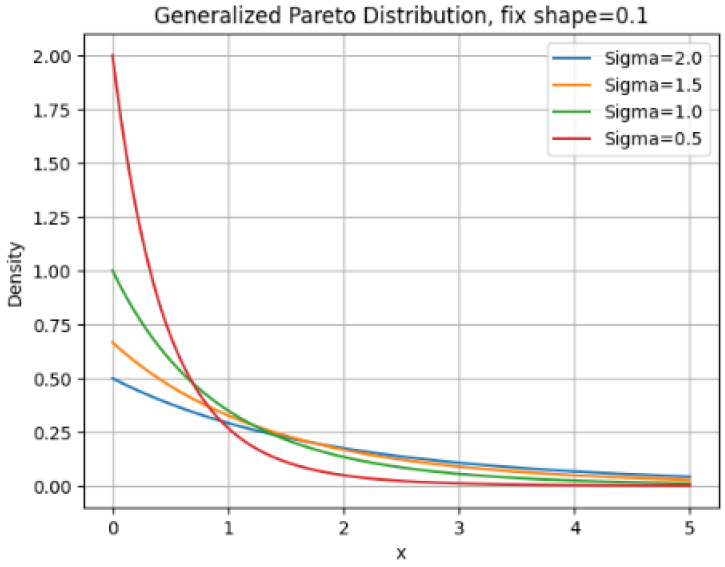


Figure 6.1: Generalized Pareto Distribution for a fixed value of shape and different values of sigma.

Additionally, we introduce the following assumption to incorporate prior knowledge about policy behavior:

$$\sigma = (\theta - \vartheta)^2 + b, \tag{45}$$

where b and $\vartheta > 0$ are predefined values. Now the objective becomes identifying the optimal policy θ that minimizes CVaR, to effectively manage the risk generating from the GPD.

6.1.2 Burr distribution

The Burr distribution characterized by two parameters c and d and is another instance of a heavy-tailed distribution. Let X follows a Burr distribution, the CDF and PDF are, respectively, given by:

$$F_{c,d}(x) = 1 - (1 + x^c)^{-d}, c, d, x > 0. \tag{46}$$

$$f_{c,d}(x) = cd\frac{(x^{c-1})}{(1 + x^c)^{d+1}}. \tag{47}$$

Figure 6.2 illustrates PDF of the Burr distribution for various values of c with a fixed value of d. As observed, a decrease in c results in a lighter density, expecting to a decrease in CVaR. Thus, for our analysis, we take c as a policy parameter expressed as $c = (\theta - \vartheta)^2 + a$, where ϑ and a are pre-defined and θ is the policy.

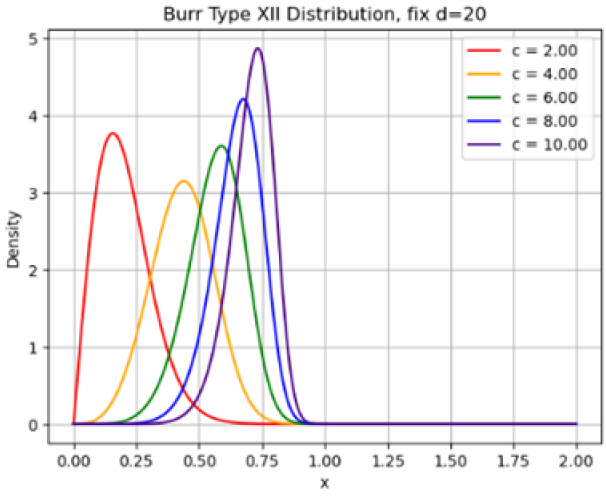


Figure 6.2: Burr distribution for a fixed value of d and different values of c.

6.1.3 Experimental setup

The experiments consist of performing $N = 50$ independent runs to ensure the reliability and robustness of the results. Each run consists of sampling with size $n = 2000$ independent costs R, from GPD and Burr distribution. We employed two methods to estimate the $\text{CVaR}_\alpha(R)$, the first one being the EVT with automated threshold selection and the other being sample average method. We refer to the estimates obtained from these methods as CVaR-EVT and CVaR-SA respectively, following Troop et al. (2022) and keeping same setting of setup for CVaR-EVT. The confidence level was set to $\alpha = 0.998$, indicating an extreme risk level.

We utilize the ADAM optimization algorithm with a step size of 0.01, determining the magnitude of parameter updates in each iterations to optimize the policy. The ADAM algorithm executes for 500 iterations to find the optimal policy and at each iterations, we have 2000 independent outcomes and we employ the finite-difference approximation method (44) to estimate the derivative of CVaR during optimization. iterations means the number of times that ADAM algorithm updates the parameter (policy in this case). Moreover, after selecting a threshold u in each iterations, we use MLE for estimating the GPD parameters using the threshold excesses above u. The procedure of the algorithm is summarized in Algorithm 2.

Remark 1 *In each iterations, we employ the same seed for generating costs from policy $\theta^{(n)}$ and $\theta^{(n)} + \epsilon$ in equation* (44) *to ensure stability.*

Remark 2 *Additionally, in the CVaR-EVT estimation method, we utilize the same threshold u selected for policy $\theta^{(n)}$, also for policy $\theta^{(n)} + \epsilon$ in equation* (44)*.*

Algorithm 2: Risk-Averse Policy Gradient using CVaR-EVT, cost generated from GPD

Require: ϵ, ξ, ϑ, a, Number of paths , Number of iterations , Initial theta

for *j in (number of paths)* **do**

$\theta \leftarrow$ Initial theta :

Starting seed $\leftarrow$ j * Number of iterations

for *i in (number of iterations)* **do**

Seed value $\leftarrow$ starting seed + i

$\sigma \leftarrow (\theta - \vartheta)^2 + b$ (Define policy σ)

$R_1 \leftarrow GPD(\xi, \sigma)$ (Generate cost from GPD with policy σ)

Select threshold u using threshold selection procedure of Troop et al. (2022),

Compute $(\hat{\xi}, \hat{\varsigma})$ of exceedances over threshold u,

Compute $\widehat{CVaR}_{u,\alpha}^{(\theta)}(R_1)$ (CVaR estimation by EVT using automated threshold selection,

Fix same seed value and same threshold,

$\sigma' \leftarrow ((\theta + \epsilon) - \vartheta)^2 + b$ (Define policy σ')

$R_2 \leftarrow GPD(\xi, \sigma')$ (Generate cost from GPD with policy σ')

Estimate $(\hat{\xi}, \hat{\varsigma'})$ of exceedances over the threshold u,

Compute CVaR estimates $\widehat{CVaR}_{u,\alpha}^{(\theta+\epsilon)}(R_2)$ by EVT using the threshold u,

Compute $\nabla \widehat{CVaR}(\theta^{(n)}) \approx \frac{\widehat{CVaR}(\theta^{(n)}+\epsilon) - \widehat{CVaR}(\theta^{(n)})}{\epsilon}$,

Update θ by ADAM algorithm 1.

end

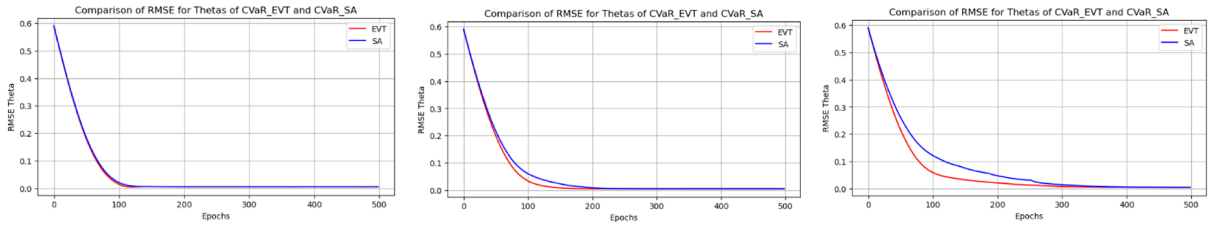
end

6.1.4 Experimental results

This section reports the results of the simulation study. We use the Root Mean Square Error (RMSE) as a metric to assess the performance of the algorithm. Our analysis involves evaluating the convergence of policy θ to solve equation $\min_\theta \text{CVaR}_\alpha(R)$, where R is the cost generated from either the GPD or the Burr distributions. Additionally, we evaluate the accuracy of CVaR and CVaR gradient estimates by using the EVT and SA methods. We fix $\vartheta = 0.4$ and $b = 2$ in assumption

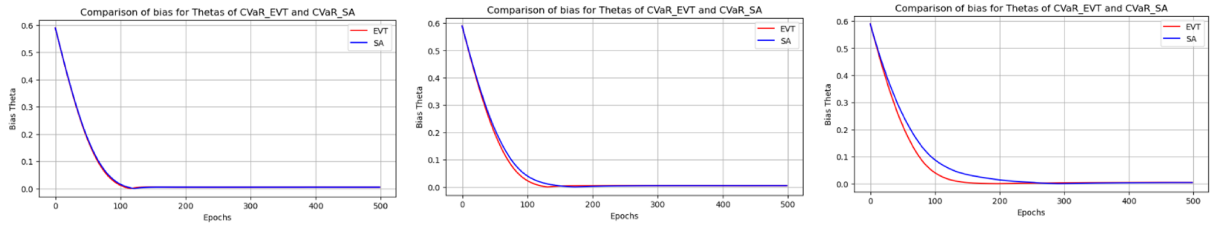
(45).

For the GPD, we benefit from having both the CVaR and its derivative with respect to the policy in closed-form formulas, as detailed in Equation (57) of the Appendix. Based on assumption (45) and formula (57) in the Appendix, we can readily identify the minimum point of CVaR of GPD as $\theta = \vartheta$. Therefore, in our scenario, the optimal value for our policy should be 0.4. The RMSE and bias (calculated as average over the various runs) of the policy in comparison to the optimal one are illustrated in Figures 6.3 and 6.4.



(a) $\xi = 0.4$ (b) $\xi = 0.6$ (c) $\xi = 0.8$

Figure 6.3: RMSE of policy using CVaR-EVT and CVaR-SA for GPD distribution.



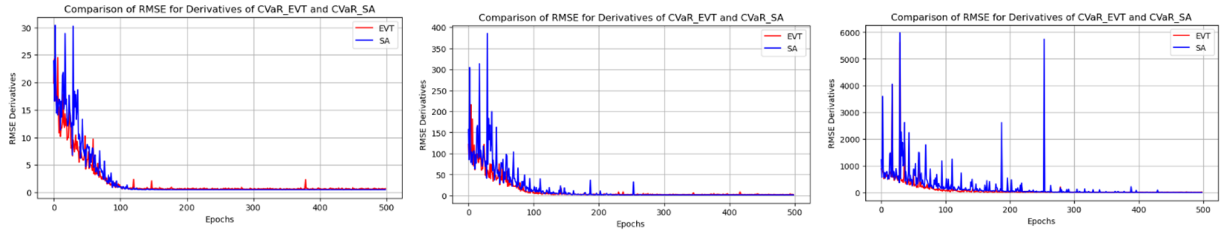
(a) $\xi = 0.4$ (b) $\xi = 0.6$ (c) $\xi = 0.8$

Figure 6.4: Bias of policy using CVaR-EVT and CVaR-SA for GPD distribution.

The x-axis represents the training iterations, while the y-axis shows the RMSE (or bias) of the policy. We examine the convergence for three distinct values of $\xi = 0.4, 0.6, 0.8$ and compare the performance of the CVaR-EVT and CVaR-SA methods. Notably, both methods exhibit convergence. With an increasing number of iterations, the RMSE of the estimated policy decreases and eventually converges to zero. This behavior indicates that our policy estimation aligns closely with the true value.

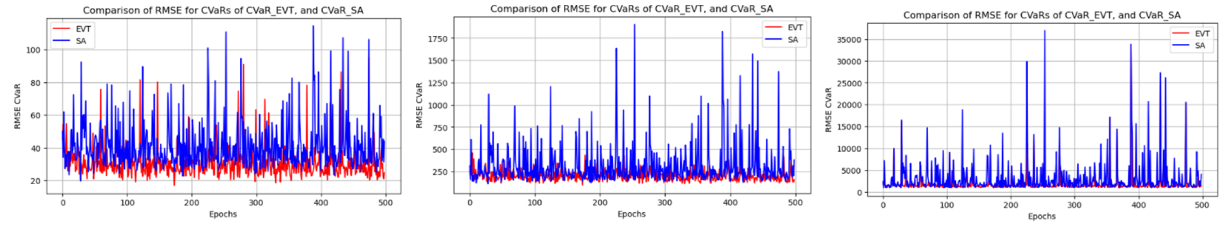
Furthermore, in the GPD framework, as ξ increases for a fixed value, the density of the distribution becomes heavier tailed. Our findings align with these of Troop et al. (2022), in the sense that as the tail density increases, the CVaR estimation based on the EVT theorem with automated threshold exhibits higher excess performance over the CVaR estimation with the SA method.

Moreover, we examine the convergence of the derivative of CVaR to ensure that the policy reaches its minimum value. Figure 6.5 illustrates the RMSE of the objective function gradient calculated across 50 different runs. As depicted in the figure, as the policy converges towards its minimum, the objective function gradient tends to zero. This observation represents the effectiveness of our learning algorithms, which perform well on this specific test case. Additionally, we represent the evolution of the CVaR estimate in Figure 6.6. This figure illustrates that as the distribution becomes heavier (with an increase in ξ), the CVaR estimation based on EVT exhibits less fluctuation compared to the SA method.



(a) $\xi = 0.4$ (b) $\xi = 0.6$ (c) $\xi = 0.8$

Figure 6.5: Objective function gradient convergence for GPD distribution.

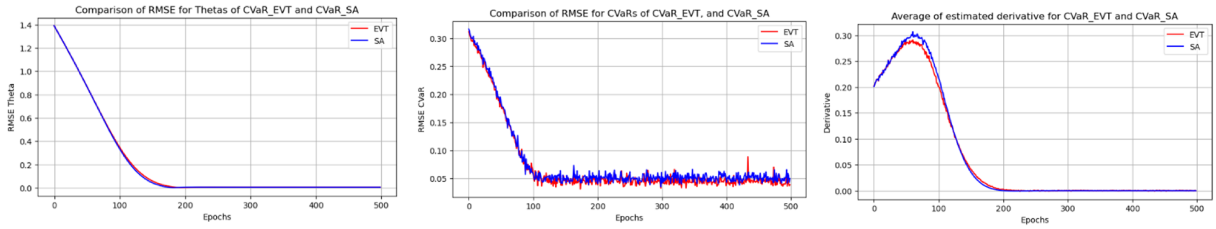


(a) $\xi = 0.4$ (b) $\xi = 0.6$ (c) $\xi = 0.8$

Figure 6.6: CVaR convergence for GPD distribution.

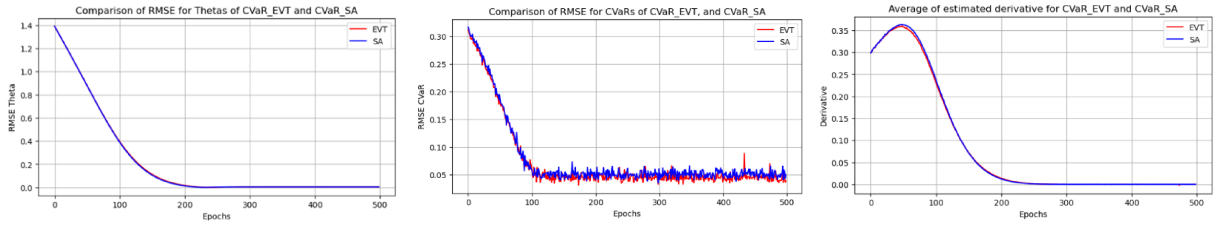
Next, we report the results for the Burr distribution. As detailed in formula (58), we have a closed-form expression for CVaR in the case of the Burr. However, a closed-form expression

for the derivative of CVaR with respect to policy is not available. To tackle this, we employed the `scipy.optimize.minimize package`, which is a widely used optimization tool in Python to determine the minimum value of a function. Then, we are able to seek the minimum value of the CVaR. This approach help us to calculate the RMSE of the policy and the CVaR, subsequently evaluating the effectiveness of our algorithm. We examine the performance of algorithm for Burr distribution for two different values of d = 20 and 40, see Figures 6.7 and 6.8. Similar to the GPD distribution, in this case as well, as the number of iterations increases, the estimated policy converges to its minimum value, the RMSE of the estimated CVaR and the estimated derivative of CVaR tend towards zero. This results demonstrate that our algorithm converges for another heavy-tailed distribution.



(a) (b) (c)

Figure 6.7: Left: Policy convergence, middle: CVaR convergence, right: Derivative convergence for Burr distribution when c considered as policy and d = 20.



(a) (b) (c)

Figure 6.8: Left: Policy convergence, middle: CVaR convergence, right: Derivative convergence for Burr distribution when c considered as policy and d = 40.

6.2 Application in finance

In this section, we present the application of our proposed algorithm in finance. Suppose you hold a short position in a call option with a strike price K. If the market price of the stock rises, the option holder may exercise the option, and due to the absence of a perfect hedge in reality, a hedging error occurs. To mitigate this risk, our objective is to minimize a suitable risk measure of hedging error and enhance our hedging position. In this thesis, we consider CVaR as a risk measure as follows:

$$\min_{\theta} \text{CVaR}_{\alpha}(C_T - V_T^{\theta}), \tag{48}$$

where V_T is the value at maturity of the self-financing portfolio invested in the hedging assets. In high-risk volatile market, gamma hedging is useful to mitigate the impact of large underlying asset variations. But when the variance risk premium is very high, using options for gamma hedging becomes considerably expensive. Trying to fully hedge gamma in these situations is not practical. To overcome this, instead of hedging fully gamma, we focus on hedging a proportion of it. This thesis aims to find the best proportion of gamma to hedge, minimizing the risk associated with gamma hedging errors.

6.2.1 Market set-up and numerical example

We consider the exponential Normal Inverse Gaussian (NIG)-Levy model as described in Chapter 2:

$$S_t = S_0 e^{\sum_{k=1}^{t} Z_k}, \tag{49}$$

$$B_t = e^{rt}. \tag{50}$$

We consider a problem of hedging an at-the-money European call option (C) with strike $K = 1000$ and maturity $T = 0.5 \times 52 = 26$. The option underlying asset is a non-dividend paying stock whose initial price is $S_0 = 1000$. Also, the risk-free rate is $r = 0.02/52$.

Parameters considered are taken from Godin (2016), namely $\mathtt{a}^{\mathbb{P}} = 35.7$, $\beta^{\mathbb{P}} = -10.8$, $\delta^{\mathbb{P}} = 2.04 \times 10^{-2}$ and $\mu^{\mathbb{P}} = 6.7 \times 10^{-3}$.

We consider a market with high volatility risk premium where options are costly; as such we assume risk-neutral parameters and identical to the physical ones, except for the delta parameter driving the returns variance, which is inflated by a factor of 4: $\mathtt{a}^{\mathbb{Q}} = \mathtt{a}^{\mathbb{P}}$, $\beta^{\mathbb{Q}} = \beta^{\mathbb{P}}$, $\delta^{\mathbb{Q}} = 4\delta^{\mathbb{P}}$ and $\mu^{\mathbb{Q}} = \mu^{\mathbb{P}}$. In such market, fully neutralizing the gamma of the option being hedged is most likely sub-optimal due to high option cost, and thus determining the best hedge ratio yield the optimal cost versus risk reduction tradeoff is a non-trivial endeavor. In this case, we refine the formula of Delta-Gamma neutral portfolios provided in Chapter 2 as follows, so as to only hedge a portion of the gamma risk:

$$\begin{cases} 0 = \Delta^{(pf)} = \theta^C \Delta^C + \theta^D \Delta^D + \theta^S, \\ 0 = \Gamma^{(pf)} = \theta^C k \Gamma^C + \theta^D \Gamma^D, \end{cases} \tag{51}$$

where $k \in (0, 1)$. Then, the second equation gives

$$\theta^{D'} = k \frac{\Gamma^C}{\Gamma^D}. \tag{52}$$

By substituting it into the first equation we have

$$\theta^S = \Delta^C - \theta^{D'} \Delta^D. \tag{53}$$

We utilize a rolling-over strategy on an ATM European call option position with maturity $T = 0.1 \times 52$ on the same underlying asset S_t, which allows for the flexibility to open and close positions at the start and end of each period. In this case, at the beginning of each period i, the hedger buys an option whose strike is exactly the same as the current value of the underlying asset, and at the end of the period $i + 1$, the hedger sells that option and buys a new option. Following this strategy, at time i, the value of the portfolio V_i is as follow:

$$\begin{cases} Cash_i = V_i - (\theta_i^s S_i + \theta_i^D D_i^b), \\ V_{i+1} = \theta_i^s S_{i+1} + \theta_i^D D_{i+1}^e + Cash_i \; e^{rdt}, \end{cases} \tag{54}$$

where D_i^b and D_i^e represent the prices of option D at the beginning and end of the period, respectively, and S_i is the stock price at time i. Additionally, cash refers to the amount invested in the risk-free asset within the portfolio. The initial value of the hedging portfolio is equal to the price of option C, $V_0 = C_0$. At each time step, the number of stock shares, θ_i^s, is calculated using formula (53) and the number of option shares, θ_i^D, is determined from formula (52). Moreover, I use weekly time-steps denoted by $dt = 1$ and the hedger rebalances the self-financing hedging portfolio at every discrete time step i for $i = 0, ..., N-1$ where $N = T/dt$.

Algorithm 3: K % of Γ Hedging error using

Require: $eps = 0.05$, NIG parameters, Number of iterations

Simulate S_t (NIG levy stock path)

for i *in (Number of iterations)* **do**

$E_1 = (S_T - strike)^+ - V_T$ (Hedging error by considering K % of Γ in V_T),

Select threshold u using threshold selection procedure of Troop et al. (2022),

Compute $(\hat{\xi}, \hat{\varsigma})$ of exceedances over threshold u,

Compute $\widehat{CVaR}_{u,\alpha}^{(k)}(E_1)$ (CVaR estimation by EVT using automated threshold selection,

$E_2 = (S_T - strike)^+ - V_T$ (Hedging error by considering $(K + \epsilon)$ % of Γ in V_T),

Estimate $(\hat{\xi}, \hat{\varsigma'})$ of exceedances over the same threshold u,

Compute $\widehat{CVaR}_{u,\alpha}^{(k+\epsilon)}(E_2)$ (CVaR estimation by EVT using automated threshold selection,

Compute $\nabla \widehat{CVaR}(k^{(n)}) \approx \frac{\widehat{CVaR}(k^{(n)}+\epsilon) - \widehat{CVaR}(k^{(n)})}{\epsilon}$,

Update k by ADAM Algorithm 1.

end

6.2.2 Risk-averse policy gradient experiment

This section compares the performance of CVaR-EVT and CVaR-SA in a policy gradient method algorithm to solve the hedging problem. Algorithm 3 represents our proposed policy gradient algorithm aimed at finding the optimal policy (i.e. finding k% of gamma in a hedging strategy) to solve equation (48) for a highly volatile market.

To evaluate our algorithm, I simulated 1,000,000 weekly NIG levy paths and plot $\text{CVaR}_\alpha(C_T - V_T)$ for different values of $k \in (0, 1)$ to find an approximation of optimal policy k and CVaR, see Figure 6.9 and Table 6.1 which display the optimal policy k and corresponding minimum CVaR. We set $\alpha = 0.999$, as we concern for very extreme events that can lead to potentially catastrophic risks.

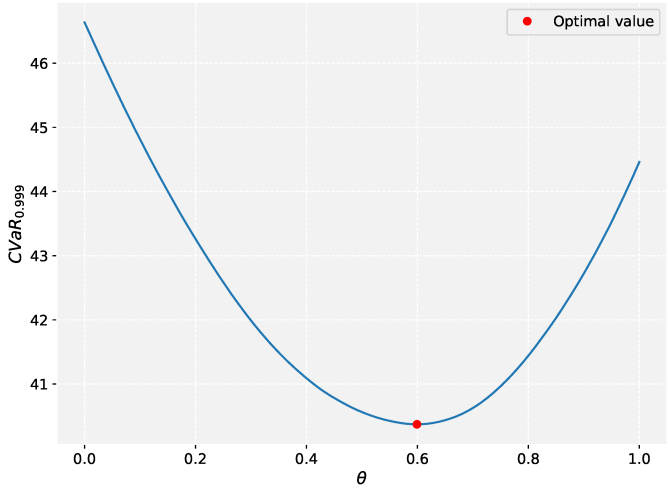


Figure 6.9: CVaR of hedging error with respect to 500 values of $k \in (0, 1)$ for the 1,000,000 weekly paths. Orange point shows the minimum value of CVaR and optimal value of k.

k	CVaR
0.5991	40.37

Table 6.1: Optimal values of policy k and minimum CVaR.

The estimation performance of CVaR-SA and CVaR-EVT on our policy gradient Algorithm 3 are compared via the RMSE. Such methods are applied with either $n = 1{,}000$ or $n = 10{,}000$

simulated paths of weekly stock returns. $R = 100$ independent runs are conducted, each comprised of $M = 500$ iterations for the case $n = 1{,}000$, or $M = 150$ iterations when $n = 10{,}000$. In each run, the initial policy is set to $\theta^{(0)} = 0$. The finite difference shock is $\epsilon = 0.05$. The method of moments is used instead to estimate tail parameters ξ, σ in the CVaR-EVT algorithm since such method exhibited (in unreported tests) greater stability than maximum likelihood estimates in the presented framework. Thus, we replace MLE in the previous section (simulated data from GPD and Burr) with the method to estimate the parameters. First we calculate the sample mean and sample variance of the data, then we substitute them into the equations (63) and (64) of the appendix.

Figure 6.10 reports the performance of the EVT and SA policy gradient algorithms for the hedging problem, by displaying the evolution of the RMSE (across runs) of the estimate of the optimal policy parameter and the corresponding objective function versus the number of iteration conducted. The CVaR-EVT algorithm exhibits materially superior performance by exhibiting much lower errors on estimates for the optimal policy parameter and objective function. The gap in performance between the CVaR-EVT and the benchmark (CVaR-SA) is greater for the lower sample size $n = 1{,}000$, which highlights that our method has more added value in the context of more severe distribution tail data scarcity. Note that none of the two methods have the estimated policy parameter converge to the true optimal value (i.e. RMSE of policy does not converge to zero), which can be explained by the fact that both methods are biased in finite sample n. Nevertheless, we see that higher sample size n increases the precision, with lower RMSEs for the estimates of the policy parameter θ^* and of the objective function.

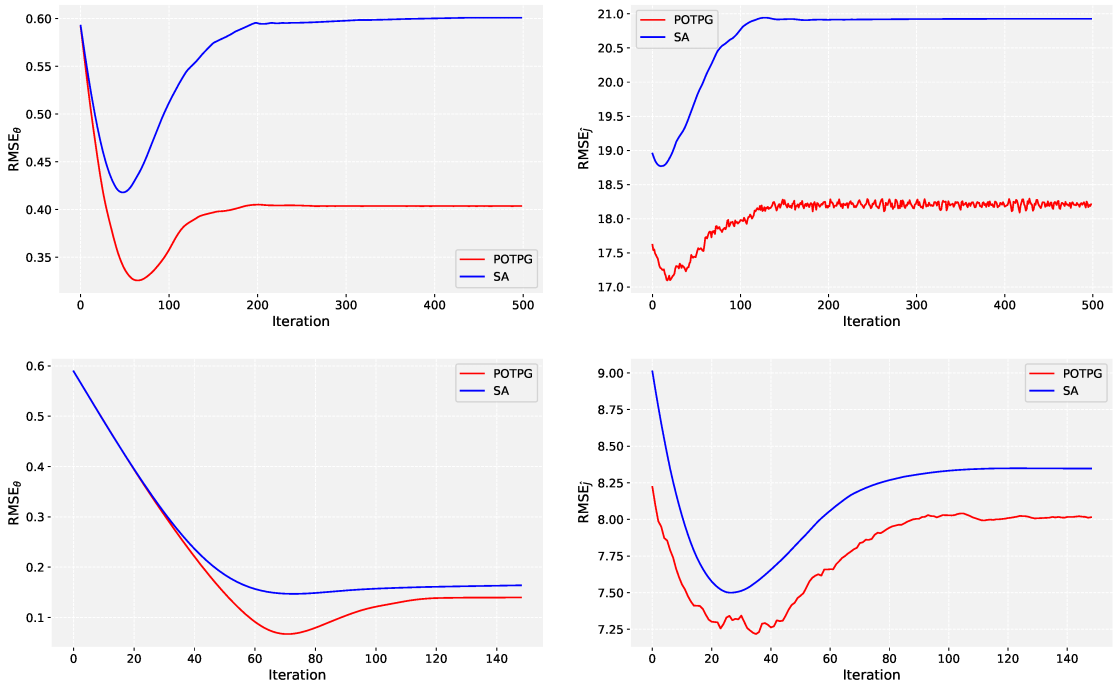


Figure 6.10: Evolution of the RMSE of the estimate of the optimal policy parameter and the corresponding objective function over iterations of our algorithm and the sample averaging (SA) benchmark. Top row: sample size $n = 1{,}000$. Bottom row: $n = 10{,}000$. Left panels: RMSE of policy. Right panels: RMSE of CVaR.

Conclusion

We have integrated risk-aware reinforcement learning and Extreme Value Theory for tail risk optimization to mitigate catastrophic risks. EVT is a well-known method for modeling highly rare risk events. In our approach, we leverage EVT and risk-aware RL to manage extremely rare events. This is the first work to combine EVT into risk-aware RL for a Markov decision problem. We utilize CVaR as our risk measure, employing EVT to estimate CVaR.

Based on EVT, the excess data over a threshold follows the GPD. Therefore, selecting a suitable threshold is crucial in EVT. To tackle this challenge, we employ an automated threshold selection method proposed by Troop et al. (2022) and Bader et al. (2018). Subsequently, we utilize Maximum Likelihood Estimation or the Method of Moments to estimate the GPD parameters of the excess data over the selected threshold. Finally, we apply the EVT formula to estimate CVaR. Heavy tail distributions or semi heavy tail distributions are interested for us in this case, since these distributions lead to more extremely events and EVT concern about extreme events.

The risk-aware policy gradient algorithm aims to find an optimal policy to minimize an objective function, such as CVaR in our case. The algorithm begins by initializing a policy and then generating costs from a heavy-tailed distribution. Subsequently, it estimates CVaR using EVT for the excess data over a selected threshold. The ADAM algorithm is then employed to update the policy. During the optimization process, the algorithm requires the gradient of the estimated CVaR. Since there is no closed form of CVaR for all distributions, we utilize the finite difference method to estimate the derivative of the estimated CVaR.

We assess our proposed algorithm using both simulated data from heavy tail distributions such

as GPD and Burr distributions and semi-heavy tails distributions like the NIG distribution in financial applications.

We also applied our algorithm in numerical experiments within a hedging strategy, presenting its application in finance. We consider the NIG Lévy model as our market setup and address the problem of gamma hedging for an at-the-money European call option. To raise the option price, we adjusted the estimated parameters derived from Godin (2016). Thus, utilizing options for gamma hedging strategies can become prohibitively expensive. To address this challenge, our algorithm introduce a proportion of gamma, denoted as k, to hedge, aiming to minimize extreme risks, such as when $\alpha = 0.999$, while also minimizing costs. We consider k as the policy to be optimized through our proposed algorithm, the risk-aware policy gradient algorithm using EVT with an automated threshold selection method. Experimental results allow analyzing the evolution of the policy parameter and of the estimated CVaR.

Furthermore, we compare the estimated CVaR using EVT and SA methods, and our results demonstrate that EVT outperforms SA. This results is observed in both simulation data from GPD and Burr distributions, as well as in a numerical example for a hedging strategy.

References

Acerbi, C., & Tasche, D. (2002). On the coherence of expected shortfall. *Journal of Banking & Finance*, *26*(7), 1487–1503.

Artzner, P., Delbaen, F., Eber, J.-M., & Heath, D. (1999). Coherent measures of risk. *Mathematical Finance*, *9*(3), 203–228.

Bader, B., Yan, J., & Zhang, X. (2018). Automated threshold selection for extreme value analysis via ordered goodness-of-fit tests with adjustment for false discovery rate. *The Annals of Applied Statistics*, *12*(1), 310–329.

Balkema, A. A., & De Haan, L. (1974). Residual life time at great age. *The Annals of Probability*, *2*(5), 792–804.

Barndorff-Nielsen, O. (1977). Exponentially decreasing distributions for the logarithm of particle size. *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences*, *353*(1674), 401–419.

Bellman, R. (1954). The theory of dynamic programming. *Bulletin of the American Mathematical Society*, *60*(6), 503–515.

Borkar, V. S. (2001). A sensitivity formula for risk-sensitive cost and the actor–critic algorithm. *Systems & Control Letters*, *44*(5), 339–346.

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

Capponi, A., Olafsson, S., & Zariphopoulou, T. (2022). Personalized robo-advising: Enhancing investment through client interaction. *Management Science*, *68*(4), 2485–2512.

Choulakian, V., & Stephens, M. A. (2001). Goodness-of-fit tests for the generalized Pareto distribution. *Technometrics*, *43*(4), 478–484.

Chow, Y., Ghavamzadeh, M., Janson, L., & Pavone, M. (2017). Risk-constrained reinforcement learning with percentile risk criteria. *The Journal of Machine Learning Research*, *18*(1), 6070–6120.

Du, X., Zhai, J., & Lv, K. (2016). Algorithm trading using q-learning and recurrent reinforcement learning. *Positions*, *1*(1), 1–7.

Gilli, M., & Këllezi, E. (2006). An application of extreme value theory for measuring financial risk. *Computational Economics*, *27*, 207–228.

Gkillas, K., & Katsiampa, P. (2018). An application of extreme value theory to cryptocurrencies. *Economics Letters*, *164*, 109–111.

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

Godin, F., Mayoral, S., & Morales, M. (2012). Contingent claim pricing using a normal inverse Gaussian probability distortion operator. *Journal of risk and insurance*, *79*(3), 841–866.

Greenberg, I., Chow, Y., Ghavamzadeh, M., & Mannor, S. (2022). Efficient risk-averse reinforcement learning. *Advances in Neural Information Processing Systems*, *35*, 32639–32652.

G'Sell, M. G., Wager, S., Chouldechova, A., & Tibshirani, R. (2016). Sequential selection procedures and false discovery rate control. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *78*(2), 423–444.

Haan, L., & Ferreira, A. (2006). *Extreme value theory: an introduction* (Vol. 3). Springer.

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

Jie, C., Prashanth, L., Fu, M., Marcus, S., & Szepesvári, C. (2018). Stochastic optimization in a cumulative prospect theory framework. *IEEE Transactions on Automatic Control*, *63*(9), 2867–2882.

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

arXiv:1412.6980.

La, P., & Ghavamzadeh, M. (2013). Actor-critic algorithms for risk-sensitive MDPs. *Advances in Neural Information Processing Systems*, *26*.

McNeil, A. J. (1997). Estimating the tails of loss severity distributions using extreme value theory. *ASTIN Bulletin: The Journal of the IAA*, *27*(1), 117–137.

McNeil, A. J., Frey, R., & Embrechts, P. (2015). *Quantitative risk management: concepts, techniques and tools-revised edition*. Princeton university press.

Michael, J. R., Schucany, W. R., & Haas, R. W. (1976). Generating random variates using transformations with multiple roots. *The American Statistician*, *30*(2), 88–90.

Pickands III, J. (1975). Statistical inference using extreme order statistics. *The Annals of Statistics*, 119–131.

Prashanth, L., Fu, M. C., et al. (2022). Risk-sensitive reinforcement learning via policy gradient search. *Foundations and Trends® in Machine Learning*, *15*(5), 537–693.

Prashanth, L., Jie, C., Fu, M., Marcus, S., & Szepesvári, C. (2016). Cumulative prospect theory meets reinforcement learning: Prediction and control. In *International conference on machine learning* (pp. 1406–1415).

Pratt, J. W. (1978). Risk aversion in the small and in the large. In *Uncertainty in Economics* (pp. 59–79). Elsevier.

Puterman, M. L. (2014). *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

Rockafellar, R. T., Uryasev, S., et al. (2000). Optimization of conditional value-at-risk. *Journal of Risk*, *2*, 21–42.

Sarykalin, S., Serraino, G., & Uryasev, S. (2008). Value-at-risk vs. conditional value-at-risk in risk management and optimization. In *State-of-the-art decision-making tools in the information-intensive age* (pp. 270–294). Informs.

Schoutens, W. (2003). *Lévy processes in finance: pricing financial derivatives*. Wiley Online Library.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Szubzda, F., & Chlebus, M. (2019). Comparison of block maxima and peaks over threshold value-at-risk models for market risk in various economic conditions. *Central European Economic Journal*, *6*(53), 70–85.

Tamar, A., Di Castro, D., & Mannor, S. (2012). Policy gradients with variance related risk criteria. In *Proceedings of the twenty-ninth international conference on machine learning* (pp. 387–396).

Tamar, A., Glassner, Y., & Mannor, S. (2015). Optimizing the CVaR via sampling. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 29).

Troop, D., Godin, F., & Yu, J. Y. (2022). Best-arm identification using extreme value theory estimates of the CVaR. *Journal of Risk and Financial Management*, *15*(4), 172.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*, 297–323.

Vijayan, N., & Prashanth, L. (2023). A policy gradient approach for optimization of smooth risk measures. In *Uncertainty in artificial intelligence* (pp. 2168–2178).

Wu, C., & Lin, Y. (1999). Minimizing risk models in markov decision processes with policies depending on target values. *Journal of Mathematical Analysis and Applications*, *231*(1), 47–67.

Yu, C., Liu, J., Nemati, S., & Yin, G. (2021). Reinforcement learning in healthcare: A survey. *ACM Computing Surveys (CSUR)*, *55*(1), 1–36.

Appendix A

Heavy tailed distributions

A.1 Generalized Pareto Distribution (GPD)

This section recalls the closed form of VaR and CVaR formula for GPD proposed by McNeil et al. (2015). First, we derive the formula for the derivative of CVaR with respect to policy. Then, we present the Burr distribution and its closed form for CVaR.

A.1.1 VaR and excess distribution of GPD:

To further analyze the risk associated with the generalized Pareto distribution (GPD), we can consider two additional concepts: Value at Risk (VaR) and the excess distribution.

For the GPD, the VaR at a confidence level $1-\alpha$ can be calculated as the α-quantile of the distribution. Denoting the α-quantile as q_α, the VaR is given by:

$$VaR(\alpha) = \begin{cases} \frac{\sigma((1-\alpha)^{-\xi}-1)}{\xi}, & \text{if } \xi \neq 0 \\ -\sigma \ln(1-\alpha), & \text{if } \xi = 0 \end{cases} \tag{55}$$

A.1.2 Excess distribution of GPD

The excess distribution represents the distribution of the values above a certain threshold, called the threshold excess. In the context of the GPD, the excess distribution is used to model extreme events above a specified threshold u. To accomplish this, we consider a random variable X with cumulative distribution function (CDF) F. We define a new random variable $Y = X - u$, and then use the conditional probability to find the distribution of Y given $X > u$. That is:

$$P(Y \leq y|X > u) = P(X - u \leq y|X > u) = P(X \leq y + u|X > u) = \frac{F(y+u) - F(u)}{1 - F(u)}.$$

Proposition A.1 *If $X \sim GPD(\xi, \sigma)$, then $X - v|X > v \sim GPD(\xi, \sigma + \xi u)$, meaning that the excess distribution of a GPD random variable is also a GPD with the same shape parameter and scaling parameter that grows linearly with the threshold u. The excess distribution allows us to estimate the probability of extreme events occurring above a given threshold.*

A.1.3 CVaR of GPD and derivative of CVaR with respect to policy θ

We can calculate the CVaR as the expected value of the tail distribution, beyond VaR. Denoting the CVaR as $CVaR_\alpha$, it can be expressed as follows.

Proposition A.2 *Let $X \sim GPD(\xi, \sigma)$ with $-1 < \xi < 1$. If $\sigma = (\theta - \vartheta)^2 + a$, then we have*

$$CVaR(\alpha) = \begin{cases} ((\theta - \vartheta)^2 + a).\left[\frac{(1-\alpha)^{-\xi}}{1-\xi} + \frac{(1-\alpha)^{-\xi}-1}{\xi}\right], & \text{if } \xi \neq 0 \\ ((\theta - \vartheta)^2 + a).[1 - ln(1-\alpha)], & \text{if } \xi = 0 \end{cases}. \tag{56}$$

Moreover, the derivative of the CVaR with respect to θ is given by:

$$\frac{\partial CVaR(\alpha)}{\partial \theta} = \begin{cases} 2.(\theta - \vartheta).\left[\frac{(1-\alpha)^{-\xi}}{1-\xi} + \frac{(1-\alpha)^{-\xi}-1}{\xi}\right], & \text{if } \xi \neq 0 \\ 2.(\theta - \vartheta).\left[1 - ln(1-\alpha)\right], & \text{if } \xi = 0 \end{cases}. \tag{57}$$

So, to easily find the minimum point of CVaR, we set Equation (57) to 0, leading $\theta = \vartheta$.

A.2 Burr distribution

The CVaR formula for the Burr distribution is as follows:

$$CVaR_\alpha(x) = \frac{d[(1/(q_\alpha)^c)]^{d-1/c}}{(1-\alpha)(d-1/c)} {}_2F_1(d - \frac{1}{c}, 1+d, d-\frac{1}{c}+1, \frac{-1}{q_\alpha}). \tag{58}$$

In this formula α represents the α quantile of the Burr distribution, and ${}_2F_1$ is the hypergeometric function.

A.3 Method of moments

Consider formulas for σ and ξ based on the first moment (sample mean) and second moment (sample variance) of the GPD distribution. Given that X follows a GPD distribution with parameters ξ and σ, the expectation and variance of X express as:

$$\begin{cases} E(x) = \frac{\sigma}{1-\xi}, & \text{if } \xi < 1 \\ var(x) = \frac{\sigma^2}{(1-\xi)^2(1-2\xi)}. & \text{if } \xi < \frac{1}{2} \end{cases} \tag{59}$$

Let the first sample moment be as $\bar{x}$ and the second moment as s^2. So we have:

$$\bar{x} = \frac{\sigma}{1-\xi} \longrightarrow \sigma = \bar{x}(1-\xi) \tag{60}$$

$$s^2 = \frac{\sigma^2}{(1-\xi)^2(1-2\xi)} \tag{61}$$

By substituting Equation (60) into equation (61), we can solve for ξ as follows:

$$\xi = \frac{s^2 - \bar{x}^2}{2s^2} \tag{62}$$

Now, we substitute Equation (62) into Equation (60) and simplify to obtain the expressions for σ and ξ:

$$\sigma = \bar{x}(\frac{s^2 + \bar{x}^2}{2s^2}) \tag{63}$$

$$\xi = \frac{s^2 - \bar{x}^2}{2s^2} \tag{64}$$

Now we replace MLE with the above method to estimate the parameters. First we calculate the sample mean and sample variance of the simulated data, then we substitute them into the equations (63) and (64).