EINDHOVEN UNIVERSITY OF TECHNOLOGY

Department of Mathematics and Computer Science
Stochastic Operations Research

# Applications of Distributionally Robust Optimisation in Heavy-Tailed Environments

*Master's Thesis*

Simcha van Helvoort

Supervisors:
prof. dr. Bert Zwart
prof. dr. ir. Sem Borst
dr. rer. nat. Christopher Hojny

Final version

Eindhoven, June 2026

# Voorwoord

Beste lezer,

Voor u ligt mijn masterscriptie die ik heb geschreven voor de masteropleiding Industrial and Applied Mathematics aan de Technische Universiteit Eindhoven. Na het afronden van HBO Toegepaste Wiskunde aan Fontys, en even genoten te hebben van het werkveld, miste ik de verdieping in wiskundige materie voor praktische toepassingen. Aangezien ik niet wist of de overstap van mijn HBO-bachelor naar deze WO-master mogelijk was, heb ik met Alessandro Di Bucchianico contact opgenomen. In dit gesprek heeft Alessandro mij het vertrouwen gegeven, dat ik deze overstap kon maken. Later heeft hij zich ook ingezet om onze HBO- en WO-wiskundeopleidingen dichter bij elkaar te brengen. Ik wil hem hiervoor, ter gedachtenis, enorm bedanken.

Verder wil ik de commissieleden bedanken voor het nemen van de tijd om mijn scriptie te beoordelen. Specifiek wil ik Bert Zwart bedanken voor zijn waardevolle begeleiding in dit voor mij nieuwe vakgebied. Ook wil ik graag mijn collega's bij Fontys Toegepaste Wiskunde bedanken voor de flexibiliteit en steun die een parttime studerende collega nodig heeft. Uiteraard wil ik ook mijn lieve familie en vrienden bedanken voor de aanmoediging en goede woorden. En uiteraard wil ik ook Vicky Hüfken bedanken voor de steun; motivatie; correcties op mijn uitwerking; het luisteren naar mijn, in halve zinnen verwoorde, wiskundige gebrabbel; het vieren toen het allemaal goed ging en het kalmeren toen de stress even hoog zat. En ook voor al het andere wat hoort bij de achtbaan van de afgelopen 6 jaar.

Ik wens u, de lezer, net zoveel plezier toe bij het doornemen van mijn scriptie als ik heb gehad bij het afronden van mijn opleiding.

Simcha van Helvoort

's-Hertogenbosch, 24 juni 2026

# Abstract

In stochastic optimisation, a decision has to be made that minimises the costs, where the costs are subject to uncertainty due to the influence of external random variables. If a decision is made based on the average costs, then it is likely that the realised costs end up being greater, which is undesirable. The difference between the average and realised costs can be extreme when the source of this uncertainty comes from heavy-tailed distributions.

In distributionally robust optimisation (DRO), a worst-case distribution is determined based on the data. Under this worst-case distribution, the mean costs should be pessimistic to avoid disappointment. Two popular methods to determine the worst-case distribution are the Kullback-Leibler (KL) divergence and the Wasserstein metric. To determine which of these two methods works better when the costs follow a heavy-tailed distribution, three applications are studied: mean estimation, wind electricity nomination, and portfolio selection. Based on the results in these applications, we conclude which method works best in a heavy-tailed environment.

In mean estimation, we analyse different robust estimation methods on their probability of over- and underestimating the mean profit of a shifted Pareto distribution. The results show that both the Wasserstein DRO and sample mean with tolerance overestimate the mean profit, which would lead to disappointment. Variance regularisation and Kullback-Leibler DRO rarely overestimate the mean profit, but KL is less conservative than variance regularisation when the tails are extremely heavy.

In wind nomination, the decision-maker should determine how much wind power they expect to deliver the next day. When less power is supplied than nominated, a fine must be paid for the difference, called the up-regulation price. These prices can be extreme and follow a possibly heavy-tailed distribution. We use Finnish wind data to calculate the profit over a period of 4 years, where one season had particularly highly varying fines. In this season, both Wasserstein and KL outperform the Sample Average Approximation (SAA) by more than 20%. KL provided more protection in the season with extreme fines, while Wasserstein was more profitable in the remaining seasons. A third method, Holistic Robust (HR) DRO, is used as a mix between KL and Wasserstein with similar protection against extreme fines but without losing as much profit in the rest of the year.

In the final application, the goal is to select assets such that the probability of earning more than the risk-free rate is maximised. A series of assets are generated from two heavy-tailed stochastic processes called Jump Diffusion and Variance Gamma. The Fourier cosine expansion is used to estimate the density, which keeps the mathematical program continuous instead of a mixed integer program. For different values of the risk-free rate, the KL DRO showed higher or equal probability of achieving more than the risk-free rate than SAA and Wasserstein. From the dual formulation, it becomes clear that the Wasserstein method is penalised for high weights, which is undesirable when a certain asset returns more than the risk-free rate. Where KL puts 100% of the resources in the most profitable stock, Wasserstein only commits 42% and therefore has both lower returns and the probability to exceed the risk-free returns.

These experiments suggest that KL works better than Wasserstein when the data is directly generated from a heavy-tailed distribution, such as in mean estimation or portfolio selection. When the data is generated by the real world, the data is susceptible to noise and can distort the sample before we actually see it. In that case, Wasserstein works well, but other methods like HR DRO are designed to work optimally in that scenario and numerical results confirm this.

# Contents

**Voorwoord** i

**Abstract** ii

**Contents** iii

**1 Introduction** 1
- 1.1 Background . . . 1
- 1.2 Problem . . . 1
- 1.3 Outline . . . 2
- 1.4 Related literature . . . 2
  - 1.4.1 Distributionally Robust Optimisation . . . 3
  - 1.4.2 Heavy-tailed distributions . . . 7

**2 Robust Mean Estimation** 10
- 2.1 Introduction . . . 10
- 2.2 Methodology . . . 10
  - 2.2.1 Disappointment rate and conservative probability . . . 10
  - 2.2.2 Robust estimation . . . 11
  - 2.2.3 Distributionally Robust Estimation . . . 12
  - 2.2.4 Method discussion . . . 13
- 2.3 Results . . . 14
- 2.4 Conclusion . . . 17

**3 Robust Wind Electricity Nomination** 19
- 3.1 Introduction . . . 19
- 3.2 Modelling . . . 19
- 3.3 Methodology . . . 20
  - 3.3.1 Wasserstein method . . . 20
  - 3.3.2 Kullback-Leibler method . . . 22
  - 3.3.3 Holistic Robust method . . . 22
  - 3.3.4 Method discussion . . . 22
- 3.4 Results . . . 23
- 3.5 Conclusion . . . 25

**4 Robust Portfolio Selection** 26
- 4.1 Introduction . . . 26
  - 4.1.1 Preliminaries . . . 27
- 4.2 Modelling . . . 27
  - 4.2.1 Jump Diffusion Process . . . 27
  - 4.2.2 Variance Gamma Process . . . 29
- 4.3 Methodology . . . 29
  - 4.3.1 Correlated stocks . . . 29
  - 4.3.2 Density recovery . . . 30
  - 4.3.3 Wasserstein and KL DRO . . . 31
  - 4.3.4 Experiments . . . 32
  - 4.3.5 Method discussion . . . 34
- 4.4 Results . . . 35
- 4.5 Conclusion . . . 37

**5 Conclusions and recommendations** **39**
5.1 Conclusions . . . 39
5.2 Future research . . . 40

**Bibliography** **41**

**Appendix** **43**

**A Derivations from Chapter 2** **43**
A.1 Derivation of Wasserstein dual . . . 43
A.2 Derivation closed form Wasserstein estimator . . . 44
A.3 Derivation KL dual . . . 46

**B Derivations from Chapter 4** **47**
B.1 Derivation of Wasserstein dual . . . 47
B.2 Derivation of KL dual . . . 48

**C Statement of AI use** **50**

# Chapter 1

# Introduction

## 1.1 Background

Stochastic optimisation is used as a framework to solve problems where decisions have to be made under uncertainty. The type of uncertainty should determine which stochastic optimisation method is appropriate. One type of uncertainty is the inability to understand dynamic and interconnected systems. Such systems can lead to black swan events, a term coined by Nassim Taleb [1] for the inherently unpredictable, low-probability, but highly impactful events. These events are sometimes mathematically modelled as the outliers of a heavy-tailed distribution.

A decision-maker has to be prepared for such outliers if they are operating with heavy-tailed distributions. This anticipation is a challenge since the distribution is unknown and outliers are, by definition, rare. In this research, we study distributionally robust optimisation (DRO), which assumes no knowledge about the distribution. This makes it a natural framework in heavy-tailed environments and is therefore the topic of this research.

In this research, we work on problems that can be formulated as a classical stochastic optimisation problem

$$\min_{x \in \mathcal{X}} \mathbb{E}_{\mathbb{P}}[\ell(x, \xi)]. \tag{1.1}$$

In this problem, $x$ is the decision variable with feasible set $\mathcal{X}$, we call $x^* = \arg\min\{\mathbb{E}_{\mathbb{P}}[\ell(x, \xi)] : x \in \mathcal{X}\}$ its optimiser, and $\xi$ is a random variable that models a possible scenario with support $\Xi \subseteq \mathbb{R}^d$. The loss function $\ell : \mathcal{X} \times \Xi \to \mathbb{R}$, when evaluated, is random due to the input of $\xi$. In (1.1), the expectation is evaluated under an unknown measure $\mathbb{P}$, possibly corresponding to a heavy-tailed distribution.

A naive approach would be to replace the unknown $\mathbb{P}$ with the empirical distribution

$$\hat{\mathbb{P}}_N = \frac{1}{N} \sum_{i=1}^{N} \delta_{\xi_i}, \tag{1.2}$$

with $\delta_z$ the Dirac measure with mass at $z$ and $\xi_1, ..., \xi_N$ are $N$ samples from the original distribution. On average, this approach leads to the disappointment of underestimating the costs, since

$$\mathbb{E}\left[\min_{x \in \mathcal{X}} \mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(x, \xi)]\right] \leq \mathbb{E}\left[\mathbb{E}_{\hat{\mathbb{P}}_N}[\ell(x^*, \xi)]\right] = \mathbb{E}_{\mathbb{P}}[\ell(x^*, \xi)], \tag{1.3}$$

which is known as the *optimiser's curse* [2]. This disappointment is a result of in-sample optimisation and evaluation. In this case, we ignore out-of-sample scenarios to evaluate the loss. The optimiser's curse is not a phenomenon that occurs only for heavy-tailed distributions, but finite samples of these distributions often fail to include extreme events.

## 1.2 Problem

Instead of optimising over $\hat{\mathbb{P}}_N$, we can choose a worst-case probability measure $\mathbb{Q}$ to evaluate the expectation. This measure should be similar to $\hat{\mathbb{P}}_N$, but with more mass on the pessimistic scenarios. We quantify this similarity with the function $D : \mathcal{P} \times \mathcal{P} \to [0, \infty)$, which can be informally interpreted as the

distance between two probability measures. Using $D$, we define the DRO problem with radius $\varepsilon > 0$

$$\begin{aligned} \min_{x \in \mathcal{X}} \max_{\mathbb{Q} \in \mathcal{P}} \quad & \mathbb{E}_{\mathbb{Q}}\left[\ell(x, \xi)\right], \\ \text{s.t.} \quad & D(\hat{\mathbb{P}}_N, \mathbb{Q}) \leq \varepsilon. \end{aligned} \tag{1.4}$$

In this problem, we do not minimise w.r.t. $\hat{\mathbb{P}}_N$. Instead, we find $\mathbb{Q}$ that maximises the expected loss, but is at most $\varepsilon$ away from $\hat{\mathbb{P}}_N$.

Different choices of $D$ lead to different worst-case distributions. Van Parys et al. [3] show that the worst-case distribution based on the Kullback-Leibler (KL) divergence is less pessimistic than any other alternative measure, under the condition of a minimal disappointment probability. In the literature, however, most of the applications are focused on the worst-case distribution based on the Wasserstein metric [4]. One of the reasons for choosing the Wasserstein metric is provided in Aolaritei et al. [5]: *"KL divergence requires absolute continuity between distributions, which precludes modelling black swan events lying outside the support of $\hat{\mathbb{P}}_N$ that represents normal operating conditions"*. In other words, they argue that the Wasserstein method allows for possible worst-case distributions where the support of $\hat{\mathbb{P}}_N$ is expanded and therefore might include black swan events, while this is not possible for KL divergence methods.

Now the question arises: which DRO variant is more effective when the scenarios are from a heavy-tailed distribution? Does the inability of KL to put mass at black-swan events hold? In practical applications, does the ability to enlarge the support outweigh the optimal disappointment probability, as described in Aolaritei et al. [5] and Van Parys et al. [3] respectively?

## 1.3 Outline

The rest of this chapter is dedicated to the related literature on DRO and heavy-tailed distributions, specifically regularly varying distributions and some of their properties. In the DRO section, we cover three DRO variations: Wasserstein, Kullback-Leibler, and Holistic Robust. The first two methods are used in Chapter 2 and Chapter 4, and all of them are used in Chapter 3.

The remaining parts of the report are structured as follows. In Chapter 2, we use various robust estimation methods to find a mean estimator that does not overestimate the mean of a heavy-tailed random variable. In that chapter, we present a closed-form solution to the Wasserstein robust estimation problem, which was developed independently of Van Parys and Zwart [6]. In Chapter 3, we cover the nomination problem for wind electricity generation, which is similar to the news-vendor problem. In Chapter 4, we will study a robust portfolio selection problem with DRO. When writing these application chapters, the aim was to make them self-contained with references to the related literature as substantiation. If the reader is interested in a specific application, it is possible to read the related chapter in isolation.

Finally, in Chapter 5 we summarise the results and answer the question stated in the previous section. We also make some recommendations for future research.

## 1.4 Related literature

The goal of this section is to give the reader a summary of the related literature where the theory is discussed in a generic way. The application of this theory is discussed in later chapters with references to this section as described in the Outline.

**Notation** We write $\mathbb{R}$ as the real line with $\mathbb{R}_+$ as the positive real numbers. We denote the natural logarithm of $z$ as $\log(z)$. For any $N \in \mathbb{N}$, we denote the set $\{1, ..., N\} = [N]$. We use $\xi$ as the random variable that describes the scenarios. We write $N$ random samples from $\xi$ as $\xi_1, ..., \xi_N$ and use a tilde $\tilde{\xi}_i$ to denote the outcome of the $i^{\text{th}}$ sample. Any arbitrary probability measure is denoted as $\mathbb{Q}$, which can vary in meaning. The measures $\mathbb{P}$ and $\hat{\mathbb{P}}_N$ are reserved for the true underlying unknown probability measure and the empirical measure, respectively. We denote the expectation as $\mathbb{E}_{\mathbb{Q}}[\xi] = \int z \, \mathrm{d}\mathbb{Q}(z)$ and the vari-

ance as $\mathbb{V}_{\mathbb{Q}}(\xi)$. Any set of probability measures is written calligraphically; for example $\mathcal{P}_\Xi$ is the set of probability measures with support $\Xi$. Furthermore, a function $f(x) = \mathcal{O}(g(x))$ if $\limsup_{x\to\infty} f(x)/g(x) < \infty$ and $f(x) = o(g(x))$ if $\lim_{x\to\infty} f(x)/g(x) \to 0$.

### 1.4.1 Distributionally Robust Optimisation

We assume that the probability measure $\mathbb{P}$ of the scenario $\xi$ is unknown, and previous research has been devoted to coming up with a suitable alternative that does not underestimate the expected loss. As described before, the empirical distribution does not suffice since this will lead to systematic underestimation of the costs in out-of-sample scenarios. However, we can utilise that, in the space of probability distributions $\mathcal{P}_\Xi$ with support $\Xi$, the location of the empirical distribution will be 'close' to the location of $\mathbb{P}$.

The exact definition of 'close' depends on how we define the similarity between two distributions. For example, we may define 'close' with the Scarf similarity between two probability measures

$$\mathrm{D}_{\text{Scarf}}(\mathbb{Q}_1, \mathbb{Q}_2) = \begin{cases} 0 & \text{if } \mathbb{E}_{\mathbb{Q}_1}[\xi] = \mathbb{E}_{\mathbb{Q}_2}[\xi] \text{ and } \mathbb{E}_{\mathbb{Q}_1}[\xi^2] = \mathbb{E}_{\mathbb{Q}_2}[\xi^2] \\ \infty & \text{otherwise,} \end{cases} \tag{1.5}$$

where $\mathrm{D}_{\text{Scarf}}(\mathbb{Q}_1, \mathbb{Q}_2) = 0$ means the two distributions, $\mathbb{Q}_1$ and $\mathbb{Q}_2$, have exactly the same first two moments. This similarity of probability measures was presented in the 1957 article of Scarf et al. [7], which is credited as the first DRO paper. In this paper, the problem

$$\min_{x\in\mathcal{X}} \max_{\mathbb{Q}\in\mathcal{P}_\Xi} \left\{ \mathbb{E}_{\mathbb{Q}}[\ell(x,\xi)] : D_{\text{Scarf}}(\hat{\mathbb{P}}_N, \mathbb{Q}) = 0 \right\} \tag{1.6}$$

is defined. Here, for all $x \in \mathcal{X}$, the expected loss is maximised using a worst-case distribution $\mathbb{Q}$. However, we cannot choose any distribution for $\mathbb{Q}$, but only distributions that have the same first two moments as the reference distribution $\hat{\mathbb{P}}_N$ (i.e., the mean and variance between $\mathbb{Q}$ and $\hat{\mathbb{P}}_N$ should be equal). This restricted set of possible probability distributions $\{\mathbb{Q} \in \mathcal{P}_\Xi : D_{\text{Scarf}}(\hat{\mathbb{P}}_N, \mathbb{Q}) = 0\}$ is called the ambiguity set.

Other DRO methods have followed the same structure as Scarf's method. Each method has a corresponding similarity score $\mathrm{D} : \mathcal{P}_\Xi \times \mathcal{P}_\Xi \to [0, \infty)$ that defines, for all reference probability measures $\mathbb{Q}_{\text{ref}} \in \mathcal{P}_\Xi$ and $\varepsilon \geq 0$, an ambiguity set $\mathcal{A}_D(\mathbb{Q}_{\text{ref}}) = \{\mathbb{Q} \in \mathcal{P}_\Xi : D(\mathbb{Q}_{\text{ref}}, \mathbb{Q}) \leq \varepsilon\}$ where the parameter $\varepsilon$ determines the size of the ambiguity set.

An example of this structure is visualised in Figure 1.1, where $\hat{\mathbb{P}}_N$ is the reference distribution. All probability measures with similarity $\varepsilon$ from $\hat{\mathbb{P}}_N$ are worst-case distribution candidates. In this figure, the actual probability measure $\mathbb{P}$ is in this set.

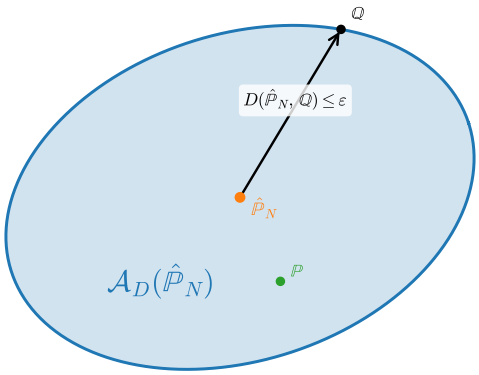



Figure 1.1: An example of the ambiguity set $\mathcal{A}_D(\hat{\mathbb{P}}_N)$[1]

[1]This figure was made with help from Claude Sonnet 4.6. The description of the image was created fully by the author. Claude provided the code to create the image. The code was later tweaked by the author.

Since we have covered the concept of the ambiguity set, we can define the general DRO problem as

$$\min_{x\in\mathcal{X}} \max_{\mathbb{Q}\in\mathcal{A}_D(\mathbb{Q}_{\text{ref}})} \mathbb{E}_{\mathbb{Q}}[\ell(x,\xi)], \tag{1.7}$$

with $\mathbb{Q}_{\text{ref}}$ as reference distribution and $\mathbb{Q}$ the worst-case distribution. This min-max problem means that we find the minimum under the worst-case distribution. Now suppose $\mathbb{P}\in\mathcal{A}_D(\mathbb{Q}_{\text{ref}})$, then we know, by the definition of the inner maximisation problem, that for all $x\in\mathcal{X}$,

$$\mathbb{E}_{\mathbb{P}}[\ell(x,\xi)] \le \max_{\mathbb{Q}\in\mathcal{A}_D(\mathbb{Q}_{\text{ref}})} \mathbb{E}_{\mathbb{Q}}[\ell(x,\xi)] \tag{1.8}$$

and therefore the estimated loss cannot disappoint. However, the event that $\mathbb{P}\in\mathcal{A}_D(\mathbb{Q}_{\text{ref}})$ depends highly on the similarity score $D$. In the next sections, we will discuss three different variations of DRO that use different similarity scores, namely Wasserstein, Kullback-Leibler and the Holistic Robust DRO.

#### 1.4.1.1 Wasserstein DRO

One of the metrics that can be used is the Wasserstein distance. This distance between two probability distributions $\mathbb{Q}_1,\mathbb{Q}_2\in\mathcal{P}_\Xi$ describes the minimum cost to transport (probability) mass within the distribution of $\mathbb{Q}_1$ to create the distribution of $\mathbb{Q}_2$. Formally, the $p$-Wasserstein distance is defined as

$$\mathrm{D}_{W_p}(\mathbb{Q}_1,\mathbb{Q}_2) = \left(\inf_{\gamma\in\Gamma(\mathbb{Q}_1,\mathbb{Q}_2)} \int_{\Xi^2} \|z_1-z_2\|_p^p \, d\gamma(z_1,z_2)\right)^{1/p} \tag{1.9}$$

with corresponding ambiguity set $\mathcal{A}_{W_p}(\mathbb{Q}_{\text{ref}}) = \{\mathbb{Q}\in\mathcal{P}_\Xi : \mathrm{D}_{W_p}(\mathbb{Q}_{\text{ref}},\mathbb{Q})\le\varepsilon\}$ [8]. The cost function to transport the probability mass is the $p$-norm, for all $z\in\mathbb{R}^d$, $\|z\|_p^p = \sum_{i=1}^d |z_i|^p$, and $\Gamma(\mathbb{Q}_1,\mathbb{Q}_2)$ is the set of probability distributions with marginals $\mathbb{Q}_1$ and $\mathbb{Q}_2$. The hyperparameter $\varepsilon$ is called the budget, which restricts the ambiguity set to probability distributions that can be created with less than $\varepsilon$ total cost. If $\varepsilon=0$, then the ambiguity set will reduce to the singleton $\mathcal{A}_{W_p}(\mathbb{Q}_{\text{ref}}) = \{\mathbb{Q}_{\text{ref}}\}$. If we increase $\varepsilon$, then the size of $\mathcal{A}_{W_p}(\mathbb{Q}_{\text{ref}})$ increases as well.

Concretely, for all $x\in\mathcal{X}$, sample realisations $\{\tilde{\xi}_1,\ldots,\tilde{\xi}_N\}$, and using $\hat{\mathbb{P}}_N$ as reference distribution, then the inner maximisation problem of (1.7) can be rewritten as the dual

$$\max_{\mathbb{Q}\in\mathcal{A}_{W_p}(\hat{\mathbb{P}}_N)} \mathbb{E}_{\mathbb{Q}}[\ell(x,\xi)] = \begin{cases} \inf\limits_{\lambda\ge0,\, s\in\mathbb{R}^N} & \lambda\varepsilon + \frac{1}{N}\sum\limits_{i=1}^N s_i, \\ \text{s.t.} & \ell(x,\xi') - \lambda\|\xi'-\tilde{\xi}_i\|_p \le s_i \quad \forall \xi'\in\Xi \quad \forall i\in[N]. \end{cases} \tag{1.10}$$

If we assume that $\xi\mapsto-\ell(x,\xi)$ is a proper[2], convex, and lower semicontinuous function and $\Xi\subseteq\mathbb{R}^d$ is convex and closed, then Esfahani and Kuhn [8] have shown that the semi-infinite program from (1.10) can be rewritten into the tractable program

$$\begin{aligned} &\inf\ \lambda\varepsilon + \frac{1}{N}\sum_{i=1}^N s_i, \\ &\text{s.t. } [-\ell(x,\cdot)]^*(\mathbf{z}_i-\mathbf{v}_i) + \sigma_\Xi(\mathbf{v}_i) - \mathbf{z}_i^\top\tilde{\xi}_i \le s_i \quad \forall i\in[N], \\ &\quad \|\mathbf{z}_i\|_* \le \lambda \quad \forall i\in[N], \\ &\quad \lambda\ge0, \mathbf{s}\in\mathbb{R}^N, \mathbf{z}\in\mathbb{R}^{N\times d}, \mathbf{v}\in\mathbb{R}^{N\times d}, \end{aligned} \tag{1.11}$$

[2] A reminder of some definitions:

- Proper function: a function $f:\mathbb{R}^d\to\bar{\mathbb{R}}$ is proper if $f(z)<\infty$ for at least one $z$ and $f(z)>-\infty$ for all $z\in\mathbb{R}^d$.
- Convex set: A set $K\subset\mathbb{R}^d$ is convex if $\forall z_1,z_2\in K$ and any $\theta\in[0,1]$, $\theta z_1+(1-\theta)z_2\in K$.
- Convex function: Let $K\subset\mathbb{R}^d$ be convex. Then $f:\mathbb{R}^d\to\bar{\mathbb{R}}$ is convex if $\forall z_1,z_2\in K$ and $\theta\in[0,1]$,
$$f(\theta z_1+(1-\theta)z_2) \le \theta f(z_1)+(1-\theta)f(z_2).$$
- Lower semicontinuous function: a function $f:\mathbb{R}^d\to\bar{\mathbb{R}}$ is lower semicontinuous if for all $z_0\in\mathbb{R}^d$,
$$f(z_0) \le \liminf_{z\to z_0} f(z).$$

with convex conjugate $[-\ell(x,\cdot)]^*(z) = \sup_{\xi'\in\Xi}\langle z,\xi'\rangle - (-\ell(x,\xi')) = \sup_{\xi'\in\Xi}\langle z,\xi'\rangle + \ell(x,\xi')$, support function $\sigma_\Xi(z) = \sup_{\xi'\in\Xi}\langle z,\xi'\rangle$ and dual norm $\|z\|_* = \sup_{\|\xi'\|_p\leq 1}\langle z,\xi'\rangle$. Under these assumptions, the authors also show that strong duality holds.

An advantage of the Wasserstein method is the ability to expand the support of the reference distribution [8]. As described in Rahimian and Mehrotra [4], the reference distribution is often chosen as $\hat{\mathbb{P}}_N$ which has support $\hat{\Xi}_N = \{\tilde{\xi}_1,...,\tilde{\xi}_N\}$. If the decision maker would choose $\mathcal{P}_\Xi := \mathcal{P}_{\hat{\Xi}}$ in a practical setting, then the ambiguity set would consist of distributions that only have mass at the points in $\hat{\Xi}_N$. However, one may define the support as an interval $\bar{\Xi}_N := [\min\hat{\Xi}_N, \max\hat{\Xi}_N]$ and use $\mathcal{P}_\Xi := \mathcal{P}_{\bar{\Xi}}$ for the set of probability distributions, which is called 'Wasserstein relaxation'. Then clearly $\hat{\mathbb{P}}_N \in \mathcal{P}_{\bar{\Xi}}$, but the resulting worst-case distribution is not restricted to the same support as $\hat{\mathbb{P}}_N$. This interval $\bar{\Xi}_N$ could be made even larger by decreasing and increasing the lower and upper bounds of $\bar{\Xi}_N$, respectively. Increasing the size of $\bar{\Xi}_N$ increases the possibility that the actual distribution $\mathbb{P} \in \mathcal{A}_{W_p}(\hat{\mathbb{P}}_N)$ and therefore reduces the possibility of disappointment.

A disadvantage of the Wasserstein method is the lack of statistical guarantees for heavy-tailed distributions. Both Esfahani and Kuhn [8] and Blanchet et al. [9] rely on thin-tailed distributions for their analysis. As shown in Van Parys and Zwart [6], for mean estimation of a heavy-tailed distribution, Wasserstein DRO has the same statistical guarantee as the sample mean (see Section 2.2).

#### 1.4.1.2 Kullback-Leibler DRO

Instead of the Wasserstein distance, we can use the relative entropy as a similarity score. The relative entropy is also known as the Kullback-Leibler divergence [10] and is defined as

$$\mathrm{D}_{\mathrm{KL}}(\mathbb{Q}_1,\mathbb{Q}_2) = \int \log\left(\frac{\mathrm{d}\mathbb{Q}_1}{\mathrm{d}\mathbb{Q}_2}\right)\mathrm{d}\mathbb{Q}_1, \tag{1.12}$$

for all $\mathbb{Q}_1 \ll \mathbb{Q}_2$ and $+\infty$ otherwise. Unlike the Wasserstein distance, the KL divergence is not a distance or a metric since KL is not symmetric. For this reason, we talk about the 'similarity' of measures instead of a metric distance. The KL ambiguity set is defined as $\mathcal{A}_{\mathrm{KL}}(\mathbb{Q}_{\mathrm{ref}}) = \{\mathbb{Q}\in\mathcal{P}_\Xi : \mathrm{D}_{\mathrm{KL}}(\mathbb{Q}_{\mathrm{ref}},\mathbb{Q}) \leq \varepsilon\}$ where hyperparameter $\varepsilon \geq 0$ determines the size of the ambiguity set. When $\varepsilon = 0$, then $\mathcal{A}_{\mathrm{KL}}(\mathbb{Q}_{\mathrm{ref}}) = \{\mathbb{Q}_{\mathrm{ref}}\}$ since $\mathrm{D}_{\mathrm{KL}}(\mathbb{Q}_{\mathrm{ref}},\mathbb{Q}) = 0$ if and only if $\mathbb{Q}_{\mathrm{ref}} = \mathbb{Q}$. Now we define the KL inner-maximisation problem from (1.7), using $\hat{\mathbb{P}}_N$ as reference distribution, as

$$\begin{array}{ll} \max\limits_{\mathbb{Q}\in\mathcal{P}_\Xi} & \mathbb{E}_{\mathbb{Q}}[\ell(x,\xi)] \\ \text{s.t.} & \mathrm{D}_{\mathrm{KL}}(\hat{\mathbb{P}}_N,\mathbb{Q}) \leq \varepsilon. \end{array} \tag{1.13}$$

There are several dual representations of this maximum problem [3, 4, 11]. We will use the formulation from Bennouna et al. [12]

$$\begin{aligned} \inf\ & \sum_{i\in[N]} w_i\,\mathbb{Q}_{\mathrm{ref}}\left(\tilde{\xi}_i\right) + \lambda(\varepsilon-1) + \eta, \\ \text{s.t. } & w_i \geq \lambda\log\left(\frac{\lambda}{\eta - \ell\left(x,\tilde{\xi}_i\right)}\right) \qquad \forall i\in[N], \\ & \eta \geq \max_{\xi'\in\Xi}\ell(x,\xi'), \\ & w\in\mathbb{R}^N,\ \lambda\geq 0,\ \eta\in\mathbb{R}. \end{aligned} \tag{1.14}$$

Note that in (1.13), the reference distribution in the first argument stays fixed, and we optimise over the second argument. This direction is in contrast to other sources [4, 11] where the second argument stays fixed. This leads to a different worst-case distribution as $\mathrm{D}_{\mathrm{KL}}$ is not symmetric. Both Rahimian and Mehrotra [4] and Sun and Conejo [11] call the formulation of (1.13) the Burg Entropy.

Furthermore, when evaluating $D_{KL}(\hat{\mathbb{P}}_N, \mathbb{Q})$ for any $\mathbb{Q} \in \mathcal{P}_\Xi$, we need absolute continuity $\hat{\mathbb{P}}_N \ll \mathbb{Q}$, which means that $\mathbb{Q}$ must have mass whenever $\hat{\mathbb{P}}_N$ has mass. However, $\mathbb{Q}$ is allowed to have mass outside of the support of $\hat{\mathbb{P}}_N$. This is in contrast to what is stated in other literature [5, 11]. They have possibly reversed the inputs of KL, with the assumption $\mathbb{Q} \ll \hat{\mathbb{P}}_N$. In that case, $\mathbb{Q}$ is dominated by $\hat{\mathbb{P}}_N$: whenever $\hat{\mathbb{P}}_N$ has 0 mass, $\mathbb{Q}$ must also have 0 mass. For this reason, the support can not be extended like Wasserstein when evaluating $D_{KL}(\mathbb{Q}, \hat{\mathbb{P}}_N)$, but the support can be extended when we consider $D_{KL}(\hat{\mathbb{P}}_N, \mathbb{Q})$.

The reason to use the formulation in (1.13), is the relation with large deviation theory. With large deviation theory, this direction of KL DRO is proved to be optimal under the condition of an exponentially decaying disappointment probability [3]. Let $\boldsymbol{\xi} = (\xi_1, ..., \xi_N)$ be a random vector of $N$ i.i.d. samples following the same distribution as $\xi$. Let $\mathbb{P}^N$ be the probability measure of random vector $\boldsymbol{\xi}$ such that $\mathbb{P}^N(\boldsymbol{\xi} = \mathbf{z}) = \prod_{i=1}^N \mathbb{P}(\xi_i = z_i)$. Then Sanov's theorem tells us that for every Borel set $\mathcal{B} \subseteq \mathcal{P}$, we have the asymptotic upper bound

$$\limsup_{N\to\infty} \frac{1}{N} \log \mathbb{P}^N \left(\hat{\mathbb{P}}_N \in \mathcal{B}\right) \leq - \inf_{\mathbb{Q}\in\mathcal{B}} D_{KL}(\mathbb{Q}, \mathbb{P}). \tag{1.15}$$

This bound means that the probability of $\hat{\mathbb{P}}_N \in \mathcal{B}$ decreases exponentially with rate $-N \inf D_{KL}(\mathbb{Q}, \mathbb{P})$. By using Sanov's theorem, the following exponentially decaying upper bound can be constructed

$$\mathbb{P}^N(\hat{\mathbb{P}}_N \in \mathcal{B}) \leq \exp\left(-N \inf_{\mathbb{Q}\in\mathcal{B}} D_{KL}(\mathbb{Q}, \mathbb{P}) + o(N)\right), \tag{1.16}$$

when $N$ is large. By choosing $\mathcal{B}$ carefully, this upper bound can be used as a disappointment probability upper bound, which is exactly what Van Parys et al. [3] have done.

Let $\varepsilon > 0$ and define $\hat{\mathbb{Q}}(x, \mathbb{Q}_{\text{ref}}) = \arg\max\{\mathbb{E}_{\mathbb{Q}}[\ell(x, \xi)] : \mathbb{Q} \in \mathcal{P}_\Xi, D_{KL}(\mathbb{Q}_{\text{ref}}, \mathbb{Q}) \leq \varepsilon\}$ as the worst-case probability distribution based on $x \in \mathcal{X}$ and reference distribution $\mathbb{Q}_{\text{ref}}$. Then the set of disappointing probability distributions is defined as $\Pi(x, \mathbb{P}) = \{\mathbb{Q}_{\text{ref}} \in \mathcal{P}_\Xi : \mathbb{E}_{\mathbb{P}}[\ell(x, \xi)] > \mathbb{E}_{\hat{\mathbb{Q}}(x, \mathbb{Q}_{\text{ref}})}[\ell(x, \xi)]\}$. This set can be interpreted as the set of all reference distributions that would lead to an underestimation of the loss if they would be used as reference distributions. It would be undesirable if $\hat{\mathbb{P}}_N \in \Pi$, since it is a natural reference distribution. Now, the authors of [3] show that, selecting $\mathcal{B} = \Pi(x, \mathbb{P})$ for (1.15), results in

$$\limsup_{N\to\infty} \frac{1}{N} \log \mathbb{P}^N \left(\hat{\mathbb{P}}_N \in \Pi(x, \mathbb{P})\right) \leq -\varepsilon, \tag{1.17}$$

which leads to an exponentially decaying disappointment probability as $N$ increases. Furthermore, any other worst-case distribution that results in an estimate that is less conservative than the KL DRO leads to an increase in disappointment for out-of-sample predictions [3]. This makes KL DRO optimal and $\hat{\mathbb{P}}_N$ the natural reference distribution.

#### 1.4.1.3 Holistic Robust DRO

The KL DRO protects optimally against the error that comes from the uncertainty of working with a sample, also known as statistical error. However, if the empirical distribution has been shifted away from $\mathbb{P}$, then this optimilaity will not hold. For example, if the distribution is shifted such that $\hat{\mathbb{P}}_N$ has mass where $\mathbb{P}$ does not, then $\hat{\mathbb{P}}_N$ would be biased. A way to solve this problem is by using the Holistic Robust (HR) DRO [12]. The HR DRO is designed to be robust to three different types of error: statistical error, noise error, and misspecification error. HR DRO assumes that a sample taken from the true distribution $\mathbb{P}$ gets corrupted by noise and misspecification. For example, this noise could be white noise on top of a signal, and the misspecification could be an incorrect registration of a data point that was manually entered.

Consider the process visualised in Figure 1.2. Let $\tilde{\mathbb{P}}_N$ be the distribution after sampling from the unknown $\mathbb{P}$. Then this distribution gets corrupted by applying noise $\mathcal{N}$ (for example the ball $\mathcal{N} = \{z : \|z\| \leq \varepsilon\}$ with $\varepsilon \geq 0$), and misspecifying at most $\alpha \in [0, 1]$ of the data, which results in the observed $\hat{\mathbb{P}}_N$. We use

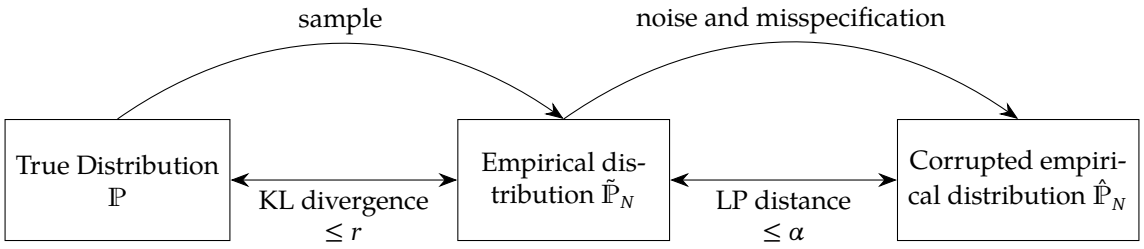



Figure 1.2: A schematic overview that shows the process from the true distribution to the distribution observed by the decision maker. Source: Bennouna et al. [12].

KL to protect against sampling error and we will use the Lévy-Prokhorov (LP) metric to protect against noise and misspecification. The LP metric is defined as

$$\mathrm{LP}_{\mathcal{N}}(\mathbb{Q}_1, \mathbb{Q}_2) = \inf_{\gamma \in \Gamma(\mathbb{Q}_1, \mathbb{Q}_2)} \int \mathbb{1}\{z_1 - z_2 \notin \mathcal{N}\} d\gamma(z_1, z_2) \quad \forall\, \mathbb{Q}_1, \mathbb{Q}_2 \in \mathcal{P}_\Xi. \tag{1.18}$$

The ambiguity set is defined as $\mathcal{A}_{\mathrm{HR}}(\mathbb{Q}_{\mathrm{ref}}) = \{\mathbb{Q} \in \mathcal{P}_\Xi : \forall \tilde{\mathbb{P}}_N \in \mathcal{P}_\Xi,\ \mathrm{LP}_{\mathcal{N}}(\mathbb{Q}_{\mathrm{ref}}, \tilde{\mathbb{P}}_N) \leq \alpha,\ \mathrm{D}_{\mathrm{KL}}(\tilde{\mathbb{P}}_N, \mathbb{Q}) \leq r\}$ for $r \geq 0$. Note that $\mathrm{LP}_{\mathcal{N}}$, just like $W_p$, is an optimal transport problem where the cost function $\|z_1 - z_2\|_p$ is replaced by $\mathbb{1}\{z_1 - z_2 \notin \mathcal{N}\}$, which is 1 if the mass is moved outside of noise set $\mathcal{N}$ and 0 otherwise. In the case where $\mathcal{N} = \{z : \|z\|_1 \leq \varepsilon\}$, at most $\alpha$ mass can be moved more than $\varepsilon$ away and at least $1 - \alpha$ of the mass will be moved within range $\varepsilon$.

Using $\mathcal{A}_{\mathrm{HR}}$ in the inner maximisation problem of (1.7), we have the following strong dual [12]

$$\begin{aligned}
&\inf \sum_{i \in [N]} w_i\, \mathbb{Q}_{\mathrm{ref}}\left(\tilde{\xi}_i\right) + \lambda(r-1) + \eta + \beta\alpha, \\
&\text{s.t. } w_i \geq \lambda \log\left(\frac{\lambda}{\eta - \ell^{\mathcal{N}}\left(x, \tilde{\xi}_i\right)}\right) && \forall i \in [N], \\
&\qquad w_i \geq \lambda \log\left(\frac{\lambda}{\eta - \max_{\xi' \in \Xi} \ell(x, \xi')}\right) - \beta && \forall i \in [N], \\
&\qquad \eta \geq \max_{\xi' \in \Xi} \ell(x, \xi'), \\
&\qquad w \in \mathbb{R}^N,\ \lambda \geq 0,\ \beta \geq 0,\ \eta \in \mathbb{R},
\end{aligned} \tag{1.19}$$

with inflated loss function $\ell^{\mathcal{N}}(x, \xi) = \max\{\ell(x, \xi - z) : z \in \mathcal{N}, \xi - z \in \Xi\}$. Note that this formulation is similar to Equation 1.14, since HR is an extension of KL.

The HR DRO protects holistically against multiple types of errors; however, specific parameter choices allow us to reduce the model to other DRO methods. When $r > 0$, $\mathcal{N} = \{0\}$ and $\alpha = 0$, HR DRO reduces to KL DRO, and the dual from (1.19) becomes the dual formulated in (1.14) as $\ell^{\mathcal{N}}$ becomes $\ell$. Furthermore, choosing $r = 0$ with arbitrary $\mathcal{N}$ and $\alpha > 0$ gives LP DRO, and choosing $\mathcal{N} = \{z : \|z\| \leq \varepsilon\}$ with both $r = 0$ and $\alpha = 0$ results in $W_\infty$ DRO [12]. For this reason, we can see HR DRO as a combination of KL, LP and $W_\infty$ DRO.

### 1.4.2 Heavy-tailed distributions

Heavy-tailed distributions are a common class of distributions that arise in a variety of fields. These are distributions whose tail decreases slower than the tail of a normal distribution. Unlike the normal distribution, where all moments are finite, the heavy-tailed distribution moments can be infinite due to the presence of large outliers, sometimes referred to as black swans [1]. If certain moments are infinite, then assumptions of standard statistical inference methods, such as the finite variance for the Central Limit

Theorem (CLT) and the Gauss-Markov theorem[3], are violated. Even estimating the mean of a distribution, the most basic inference there is, will be over- or underestimated by naive methods. Furthermore, non-parametric models will be unreliable for finite samples since data about the outliers is, by definition, rare [13]. Nevertheless, heavy-tailed distributions themselves are not an outlier. Many mechanisms can result in heavy-tailed distributions, such as the generalised CLT when the second moment is infinite [14].

One subset of heavy-tailed distributions is called *regularly varying distributions*. We define a regularly varying distribution $F$ by means of its right tail function $\bar{F}(z) = 1 - F(z) = \mathbb{P}(\zeta > z)$ where $\zeta \sim F$. Now, if $F$ is a regularly varying distribution and $\zeta$ is a regularly varying random variable, then there exists a $\rho > 0$, such that for all $c > 0$,

$$\lim_{z\to\infty} \frac{\bar{F}(cz)}{\bar{F}(z)} = c^{-\rho}. \tag{1.20}$$

We call $\rho$ the tail index. Equivalently, we can write $\bar{F}$ in terms of a function $L$

$$\bar{F}(z) = L(z)z^{-\rho} \tag{1.21}$$

where $L$ is slowly varying, which means that $L(cz)/L(z) \to 1$ as $z \to \infty$.

A property of regularly varying distributions is that most of the moments are infinite (some are even undefined). To be exact, suppose $\zeta$ is a regularly varying random variable with tail index $\rho > 0$, then for all $k \in \mathbb{R}_+$,

$$\begin{aligned} \mathbb{E}\left[\zeta^k\right] &< \infty \quad \text{if } 0 \le k < \rho, \\ \mathbb{E}\left[\zeta^k\right] &= \infty \quad \text{if } k > \rho, \end{aligned} \tag{1.22}$$

which is Theorem 2.14 from Nair et al. [14]. The case were $\rho = k$ is omitted since the result depends on the slowly-varying function $L$. For large $\rho$, this will not have a significant effect. But when $2 < \rho < 4$, we already have infinite kurtosis and therefore the variance $\mathbb{V}(\hat{\sigma}_N^2) = \infty$ with $\hat{\sigma}_N^2$ the sample variance estimate

$$\hat{\sigma}_N^2 = \frac{1}{N}\sum_{i=1}^{N}(\zeta_i - \hat{\mu}_N)^2, \tag{1.23}$$

with $\hat{\mu}_N = \mathbb{E}_{\hat{\mathbb{P}}_N}[\zeta]$. As a consequence, the sample variance is still asymptotically consistent, but with a slower convergence rate than $\mathcal{O}\left(N^{-1/2}\right)$. Decreasing $\rho$ gives more problems. When $\rho < 3$, we have that the Berry-Esséen[4] bound does not apply [15]. When $\rho < 2$, the CLT fails to converge to a normal distribution. And when $\rho < 1$, the mean is uninformative, and therefore the Law of Large Numbers is inapplicable.

A characteristic of regularly varying distributions is the overestimation of the sample mean. Suppose $\rho > 1$, $c > 0$ and $\zeta$ is a nonnegative random variable, then from Van Parys and Zwart [6], we have

$$\mathbb{P}^N(\hat{\mu}_N > \mathbb{E}[\zeta] + c) = N\,\mathbb{P}(\zeta \ge cN)(1 + o(1)). \tag{1.24}$$

This overestimation comes from the fact that the sample is likely to contain outliers. A single outlier greater than $cN$ will pull the sample mean away from the theoretical mean, which causes an overestimation of the mean.

[3] The Gauss-Markov theorem states that the ordinary least squares estimator is the estimator with the lowest variance from the class of unbiased linear estimators under the assumption that the residuals have equal and finite variance.

[4] The Berry-Esséen Theorem states the following. Let $\hat{F}_N$ be the distribution function of the normalised sample mean of i.i.d. samples and $\Phi$ the standard Gaussian distribution function. There exists a $C < \infty$ such that for all $x$ and $N$, $|\hat{F}_N(x) - \Phi(x)| \le C\kappa(\sigma^3\sqrt{N})^{-1}$ with $\kappa < \infty$ the absolute third moment.

#### 1.4.2.1 Consequences for DRO

Clearly, the DRO proposed by Scarf et al. [7] is not a good variation for heavy-tailed distributions as the moments can be infinite or undefined. The same can be said for Wasserstein-2 DRO. Villani [16] writes $W_p(\hat{\mathbb{P}}_N \mathbb{Q}) < \infty$ if and only if measure $\mathbb{Q}$ has finite $p$-moments. This makes Wasserstein-2 unsuitable when $\mathbb{P}$ is heavy-tailed with $\rho < 2$. Therefore, when dealing with heavy-tailed distributions, the 1-norm is most suitable as cost function, which only assumes finite mean.

KL is based on Sanov's theorem, which does not require finite moments [3]. Therefore, KL is applicable for heavy-tailed distributions.

#### 1.4.2.2 Pareto distribution

An example of a heavy-tailed distribution is the Pareto distribution, arguably the most famous heavy-tailed distribution due to its connection with the well-known '80/20' Pareto principle. If $\zeta \sim F$ is Pareto distributed, then

$$\bar{F}(z) = \begin{cases} \left(\frac{z_{\min}}{z}\right)^{\rho} & \text{if } z \geq z_{\min}, \\ 1 & \text{if } z < z_{\min}, \end{cases} \tag{1.25}$$

with tail index $\rho > 0$ and positive $z_{\min}$ such that the support $\Xi = [z_{\min}, \infty)$. Clearly this distribution is regularly varying as $L(z) = z_{\min}^{\rho}$ is constant.

For the Pareto distribution we have $k$-th moments

$$\begin{aligned} \mathbb{E}[\zeta^k] &= \tfrac{\rho z_{\min}^k}{\rho - k} \quad \text{if } \rho > k, \\ \mathbb{E}[\zeta^k] &= \infty \quad \text{if } \rho \leq k, \end{aligned} \tag{1.26}$$

where we note the equality in the second case. This means that we have infinite variance for $\rho \leq 2$.

In Chapter 2, we use the Pareto distribution and estimate the mean. In Chapter 3, we have higly varying electricity price data. Lastly, we model stock prices in Chapter 4 with a stochastic process that has heavier tails than the Gaussian distribution.

# Chapter 2

# Robust Mean Estimation

## 2.1 Introduction

One of the simpler statistical inference tasks is the estimation of the mean. In decision making, a common mean estimation task is the estimation of the expected loss $\mathbb{E}_{\mathbb{P}}[\ell(x,\xi)]$ where $\xi$ follows an unknown distribution $\mathbb{P}$. We assume that $x \in \mathcal{X}$ is fixed and define the revenue as $\zeta = -\ell(x,\xi)$ with the assumption that $\zeta \sim F$ is regularly varying. Normally we want to minimise $\mathbb{E}_{\mathbb{P}}[\ell(x,\xi)]$, but now we maximise $\mathbb{E}_{\mathbb{P}}[\zeta]$ because of the sign. We do not want to overestimate the revenue, as this would lead to disappointment when we expect a certain revenue but fall short. On the other hand, we also do not want to be too conservative, which could lead to outperformance by other methods.

In this chapter, we empirically investigate different mean estimation methods that are robust to outliers. The empirical results will be compared to the analytical results in Van Parys and Zwart [6]. Specifically, the disappointment rate and the conservative probability defined in that paper will be estimated. These notions will be presented in Section 2.2 together with the robust estimation methods and their analytical results. In Section 2.3 we present the empirical results alongside the analytical results. Finally, we summarise and conclude this chapter in Section 2.4.

**Notation** We denote $\hat{e}_N = \hat{e}_N(\hat{\mathbb{P}}_N)$ as the estimator for $\mathbb{E}_{\mathbb{P}}[\zeta] = \mu$ where $\zeta$ is a nonnegative and regularly varying random variable with tail index $\rho > 1$. We define $\sigma^2 = \mathbb{V}_{\mathbb{P}}(\zeta) = \mathbb{E}_{\mathbb{P}}[(\zeta - \mathbb{E}_{\mathbb{P}}[\zeta])^2]$ as the variance under the probability measure $\mathbb{P}$. If the support has $T \in \mathbb{N}$ unique values, we assume that the support is $\{\tilde{\zeta}_1, ..., \tilde{\zeta}_T\}$ and write $\mathbb{Q}$ as the vector $[q_1, ..., q_T]$ s.t. $\sum_{i=1}^{T} q_i = 1$ and $\mathbb{Q}(\tilde{\zeta}_i) = q_i$. For the empirical distribution, we denote $\hat{\mathbb{P}}_N(\tilde{\zeta}_i) = \hat{p}_i$. Furthermore, we define the sample estimators $\hat{\mu}_N = \mathbb{E}_{\hat{\mathbb{P}}_N}[\zeta]$ and $\hat{\sigma}_N^2 = \mathbb{V}_{\hat{\mathbb{P}}_N}[\zeta]$. We denote $\mathbb{P}^N$ as the measure of the random vector $(\zeta_1, ..., \zeta_N)$, which are $N$ i.i.d. copies of $\zeta$. Note that, when $\zeta$ follows a continuous distribution, then almost surely $T = N$. Let $\tilde{\zeta}_i$ be the outcome of $\zeta_i$ for all $i \in [N]$.

## 2.2 Methodology

To verify the results from Van Parys and Zwart [6], we will perform a Monte Carlo simulation. We let $\xi$ follow a heavy-tailed distribution; in this case, a shifted Pareto distribution with tail index $\rho > 1$, $z_{\min} = 1$ and shifted by $-1$. By shifting the Pareto by $-1$, we have $\mu = \frac{\rho}{\rho-1} - 1 = \frac{1}{\rho-1}$ with support $\Xi = [0, \infty)$ all nonnegative real numbers.

We will estimate $\mu$ with three different robust estimation and two distributionally robust estimation methods. We compare the methods on two metrics: disappointment and conservative probabilities.

### 2.2.1 Disappointment rate and conservative probability

In Van Parys and Zwart [6], the disappointment rate $\lambda(N)$ for the loss function is defined as the scaling term of the disappointment probability such that

$$\limsup_{N\to\infty} \frac{1}{\lambda(N)} \log \mathbb{P}^N(\hat{e}_N > \mu) \leq -1. \tag{2.1}$$

We can interpret $\lambda(N)$ as the rate of decay with the approximate bound

$$\mathbb{P}^N(\hat{e}_N > \mu) \leq \exp(-\lambda(N)). \tag{2.2}$$

for large $N$. If $\lambda(N) = \mathcal{O}(N)$, then this will result in an exponential decay of the disappointment probability as the samples get larger. If $\lambda(N) = o(N)$, then the decay is subexponential. We define $r(N) = \lambda(N)/N$.

An obvious candidate for (2.1) is $\hat{e}_N = 0$ regardless of the sample. However, we need an estimator that not only does not disappoint, but an estimator that also is not too conservative. We define the conservative term $b \in (0, \mu)$ and the conservative probability as

$$\mathbb{P}^N(\hat{e}_N < \mu - b), \tag{2.3}$$

which should decay to 0 as $N$ increases. Later in Section 2.3, we consider a reformulation of the conservative probability, namely $\mathbb{P}^N(\hat{e}_N/\mu < 1 - c)$. This reformulation makes the conservative term $c = b/\mu \in (0, 1)$ dimensionless.

### 2.2.2 Robust estimation

#### 2.2.2.1 Tolerance level

A simple method is to take the sample mean $\hat{\mu}_N$ and apply a tolerance level $\Delta \geq 0$. Then define the robust estimator $\hat{e}_{N,\Delta} = \hat{\mu}_N - \Delta$ where the simple idea is to correct for the overestimation by a constant $\Delta$. However, by using (1.24) the disappointment probability can be estimated as [6]

$$\mathbb{P}^N(\hat{e}_{N,\Delta} > \mu) = N\,\mathbb{P}(\zeta > \Delta N)(1 + o(1)) = \exp(-(1 + o(1))(\rho - 1)\log N), \tag{2.4}$$

with decay rate remarkably not depending on $\Delta$. In this case, $\lambda(N) = \mathcal{O}(\log(N))$ which results in a subexponential decay.

Since $\hat{e}_{N,\Delta}$ is biased, it will be conservative for any $\Delta > 0$. Now suppose $\Delta = 0$, then the conservative probability can be calculated as

$$-\frac{1}{N}\log \mathbb{P}^N(\hat{e}_{N,0} \leq \mu - b) \to I(b) \tag{2.5}$$

with $I(b) = \sup_{s>0}(b - \mu)s - \log \mathbb{E}_{\mathbb{P}}[\exp(-s\zeta)]$ the Legendre transform [6].

Suppose $\zeta'$ is the unshifted Pareto distribution. Then the Laplace transform $\mathbb{E}_{\mathbb{P}}[\exp(-s\zeta')]$ can be expressed in terms of the incomplete Gamma function $\Gamma(a, x)$, such that $\Gamma(a, 0) = \Gamma(a) = (a-1)!$ if $a \in \mathbb{N}$,

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}}[\exp(-s\zeta')] &= \int_1^\infty \exp(-sx)\rho x^{-\rho-1}\,\mathrm{d}x && (2.6)\\
&= \rho \int_1^\infty \exp(-sx)x^{-\rho-1}\,\mathrm{d}x = \rho\,\mathcal{E}_{\rho+1}(s) && (2.7)\\
&= \rho s^\rho \Gamma(-\rho, s), && (2.8)
\end{aligned}$$

Here we used that the generalized exponential integral $\mathcal{E}_N(x) = x^{N-1}\Gamma(1-N, x)$ [17]. Now since $\zeta = \zeta' - 1$, we have

$$\begin{aligned}
\log \mathbb{E}_{\mathbb{P}}[\exp(-s\zeta)] &= \log \mathbb{E}_{\mathbb{P}}[\exp(-s(\zeta' - 1))] && (2.9)\\
&= \log \mathbb{E}_{\mathbb{P}}[\exp(-s\zeta' + s)] = \log\left(\mathbb{E}_{\mathbb{P}}[\exp(-s\zeta')]\right) + s && (2.10)
\end{aligned}$$

Therefore $I(b) = \sup_{s>0}(b - \mu - 1)s - \log(\rho s^\rho \Gamma(-\rho, s))$.

#### 2.2.2.2 Truncated mean

If the sample mean is sensitive to outliers, then one option is to truncate the outliers by a smaller value. This idea is exactly what the truncated mean (also known as trimmed mean) does. Let $a \in (1,2]$ and $\mathbb{E}_{\mathbb{P}}[\zeta^a] \leq A < \infty$. Take $\lambda(N) = o(N)$ s.t. $r(N) \to 0$ but $\lambda(N) \to \infty$. Then the truncated mean estimator is defined as

$$\hat{e}_N^{a,A} = \frac{1}{N}\sum_{i=1}^{N} \min\{\zeta_i, r(N)^{-1/a}\} - c_a r(N)^{(a-1)/a}, \tag{2.11}$$

where $c_a = (a-1)^{-(a-1)/a} a C^{1/a}$, and $C = \max\{(a-1)^{-1}a^{-a}, 2Ae^a(2+a)^{-1}\}$. For $\rho > 1.25$, we use $a = \min\{\rho - 0.25, 2\}$ and $\mathbb{E}_{\mathbb{P}}[\zeta^a] = A$. For $\rho \leq 1.25$, we decide not to use this estimator.

Note that this estimator uses the theoretical moment $\mathbb{E}_{\mathbb{P}}[\zeta^a]$, which is generally not available.

From Van Parys and Zwart [6], we know that

$$\mathbb{P}^N(\hat{e}_N^{a,A} > \mu) \leq \exp(-\lambda(N)), \tag{2.12}$$

which is subexponential, and that $\mathbb{P}^N(\hat{e}_N^{a,A} \leq \mu - b)$ decays exponentially with rate $\lambda(N)$.

#### 2.2.2.3 Variance regularisation

Suppose $\rho > 2$, such that $\mathbb{V}_{\mathbb{P}}(\zeta) < \infty$. Then we can use the variance as a tolerance level similar to $\hat{e}_{N,\Delta}$. In the case of an outlier, the sample variance will increase and correct for this outlier. The exact tolerance $\Delta$ is based on the CLT [6]. Using this argument we define the variance regularisation (VR) estimator as

$$\hat{e}_N^{\mathrm{VR}} = \hat{\mu}_N - \sqrt{2r(N)}\hat{\sigma}_N. \tag{2.13}$$

The disappointment rate decays with rate $\lambda(N) = \mathcal{O}(N)$ [6]. However, note that the sample standard deviation is also heavily influenced by outliers and can be unreliable for $\rho < 4$ as discussed in Section 1.4.2. If $\hat{\sigma}_N > \frac{\hat{\mu}}{\sqrt{2r(N)}}$, then this estimator is negative. For this reason, the estimator can be extremely conservative, with lower conservative probability bound [6]

$$\mathbb{P}^N(\hat{e}_N^{\mathrm{VR}} \leq \mu - b) \geq (1 + o(1))N\,\mathbb{P}\left(\zeta > b\sqrt{\frac{N}{2r(N)}}\right) = (1 + o(1))N\,\mathbb{P}\left(\zeta > b\frac{N}{\sqrt{2\lambda(N)}}\right). \tag{2.14}$$

We will still test this estimator for $\rho \leq 2$ to compare it with the other estimators.

### 2.2.3 Distributionally Robust Estimation

#### 2.2.3.1 Wasserstein

The Wasserstein estimator is defined as $\hat{e}_{N,\varepsilon}^{W} = \min_{\mathbb{Q} \in \mathcal{A}_{W_1}(\hat{\mathbb{P}}_N)} \mathbb{E}_{\mathbb{Q}}[\zeta]$ where $\mathcal{A}_{W_1}$ is the ambiguity set based on the Wasserstein-1 distance from (1.9). We use $\varepsilon := r(N) = \log(N)/N$ as radius of $\mathcal{A}_{W_1}$. The estimator can be calculated with the dual formulation

$$\hat{e}_{N,\varepsilon}^{W} = \max_{\alpha \geq 0}\left\{-\alpha\varepsilon + \sum_{i=1}^{N} \hat{p}_i \min_{j \in [N]}\left\{\tilde{\zeta}_j + \alpha|\tilde{\zeta}_i - \tilde{\zeta}_j|\right\}\right\}. \tag{2.15}$$

The derivation of the dual can be found in Section A.1 and is based on Lagrange duality. This maximisation problem has a closed form, which is given as

$$\hat{e}_{N,\varepsilon}^{W} = \sum_{i=1}^{N} \hat{p}_i \tilde{\zeta}_i - \varepsilon = \hat{\mu}_N - \varepsilon, \tag{2.16}$$

under the assumption that $\varepsilon < \hat{\mu}_N$. This is a reasonable assumption since the estimator would otherwise be negative. The derivation of (2.16) can be found in Section A.2, which uses the Wasserstein relaxation. If the Wasserstein relaxation was not used, then this would lead to the same estimator as (2.16), but with the condition that $\varepsilon + \min_{i\in[N]} \tilde{\zeta}_i < \hat{\mu}_N$. Since $\min_{i\in[N]} \tilde{\zeta}_i \to 0$ as $N \to \infty$, the Wasserstein relaxation does not lead to a drastically different estimator.

The closed form shows that the Wasserstein estimator is actually a sample mean with a shrinking tolerance level $\Delta(N) = \varepsilon = r(N)$ in disguise. However, since $r(N) = \log(N)/N$, we know $r(N) \to 0$ as $N \to \infty$, which makes $\hat{e}^W_{N,\varepsilon}$ asymptotically unbiased. For a shrinking tolerance, no theoretical disappointment probability has been derived.

#### 2.2.3.2 Kullback-Leibler

The Kullback-Leibler estimator is defined as $\hat{e}^{\mathrm{KL}}_{N,\varepsilon} = \min_{\mathbb{Q}\in\mathcal{A}_{\mathrm{KL}}(\hat{\mathbb{P}}_N)} \mathbb{E}_{\mathbb{Q}}[\zeta]$ with again $\varepsilon := r(N)$ as radius of $\mathcal{A}_{\mathrm{KL}}$. The KL DRO estimator can be worked out as

$$\hat{e}^{\mathrm{KL}}_{N,\varepsilon} = \begin{cases} \min\limits_{\mathbb{Q}\in\mathcal{P}_\Xi} & \mathbb{E}_{\mathbb{Q}}[\zeta], \\ \text{s.t.} & \sum_{i=1}^N \log\left(\frac{\hat{p}_i}{q_i}\right)\hat{p}_i \le \varepsilon, \\ & \sum_{i=1}^N q_i = 1, \\ & q_i \ge 0 \quad \forall i \in [N]. \end{cases} \tag{2.17}$$

Just like the Wasserstein mean estimator, we can define this estimator in terms of its strong dual

$$\hat{e}^{\mathrm{KL}}_{N,\varepsilon} = -\min\left\{\alpha(\varepsilon-1) + \nu + \sum_{i=1}^N \alpha\hat{p}_i \log\frac{\alpha}{\tilde{\zeta}_i + \nu} : \alpha \ge 0, \nu > -\min_{i\in[N]} \tilde{\zeta}_i\right\}. \tag{2.18}$$

The derivation can be found in Section A.3, which is also based on the Lagrange dual.

As $N \to \infty$, we have the disappointment probability bound

$$\mathbb{P}^N(\hat{e}^{\mathrm{KL}}_{N,\varepsilon} > \mu) \le \exp[-\lambda(N)(1 + o(1))], \tag{2.19}$$

with $\lambda(N) = \mathcal{O}(N)$, and conservative probability

$$\mathbb{P}^N(\hat{e}^{\mathrm{KL}}_{N,\varepsilon} < \mu - b) \le \exp[-NI(b)(1 + o(1))], \tag{2.20}$$

which is similar to the conservative probability of the sample mean. Van Parys and Zwart also derived a non-asymptotic disappointment probability bound

$$\mathbb{P}^N(\hat{e}^{\mathrm{KL}}_{N,\varepsilon} > \mu) \le \exp[-\lambda(N)](e\lambda(N)\log(N) + e^2), \tag{2.21}$$

which can be used to choose $\lambda(N)$ such that $\mathbb{P}^N(\hat{e}^{\mathrm{KL}}_{N,\varepsilon} > \mu) \le p$ with $p \in [0,1]$. There is no closed-form solution of $\exp[-\lambda(N)](e\lambda(N)\log(N)+e^2) - p = 0$, but it can be found easily with a root-finding algorithm like Newton-Raphson.

### 2.2.4 Method discussion

Before discussing the empirical results, a few limitations of this research should be addressed. We present this discussion here such that the reader is aware of these limitations and considers them when reading the results.

For this research, we compare the different methods for estimating the mean of a Pareto distribution. However, it could be that some methods work better for other heavy-tailed distributions with positive support where the body of the distribution does not follow a power law, unlike the Pareto. For this distribution, more of the data will be centred around the mean, and the mean will be easier to estimate.

Furthermore, we only compare two DRO methods since these methods were the methods discussed in [6]. However, other DRO methods like the HR DRO could also be used. Other robust methods like median-of-means or $M$-estimators were also mentioned in [6]. Without going into further detail of these methods, they could be tested in future research.

Lastly, note that the formulation of (2.17) is restricting the worst-case distribution to the same support as $\hat{\mathbb{P}}_N$. In order to not restrict $\mathbb{Q}$, the conditions should be changed to $\int_{[0,\infty)} \log(\frac{d\hat{\mathbb{P}}(x)}{d\mathbb{Q}(x)})\, d\hat{\mathbb{P}}(x)$ and $\int_{[0,\infty)} d\mathbb{Q}(x) = 1$. However, we chose this formulation as it gives a good and simple introduction of KL DRO without having to go into technicalities.

## 2.3 Results

We ran two Monte Carlo experiments. The first experiment is to examine what happens for various values of $\rho$. By varying $\rho$, we get insight into how the methods behave for different levels of heavy-tailedness. For the second experiment, we want to analyse the behaviour when $N$ increases.

The first experiment is a Monte Carlo simulation with in total $B = 100{,}000$ runs. For each run, we sampled $N = 1000$ i.i.d. samples from a shifted Pareto distribution with tail index $\rho$ and support $[0, \infty)$. We used $\rho \in \{1.1, 1.5, 2, 2.5, 4, 5\}$, which means infinite variance for the first three tail indices and finite variance for the last three. Furthermore, for KL DRO, we take $\lambda(N)$ as the solution of setting (2.21) equal to a precision of $p = 0.01$. For the Wasserstein, truncated mean and variance regularisation, we use $\lambda(N) = \log(N)$.

The second experiment is a Monte Carlo simulation with $B = 200{,}000$ runs. We vary the sample size $N = \{10, 50, 100, 500, 1000, 5000\}$ for two tail indices $\rho \in \{1.5, 2.5\}$. We use the same $\lambda(N)$ as in the first experiment, except that we use a precision of $p = 0.1$ for KL.

### Experiment 1

In Figure 2.1, we can see a sample of the mean estimates from the first experiment. The continuous line represents the actual mean, and an overestimation would mean that a data point is above this line. The first thing that stands out is that the truncated mean estimates are unrealistic and deviate from the rest of the estimates. Since $a = \rho - 0.25$, the $a$-th moment for $\rho \leq 1.25$ is infinite and therefore will not be used. For $\rho = 1.5$ and $\rho = 2$, the truncated mean yields negative mean estimates for a nonnegative random variable, which is not feasible. For later values of $\rho$, the truncated mean is sometimes positive, but much more pessimistic than the other methods.

These poor results from the truncated mean are a result of the correction term $c_a r(N)^{(a-1)/a}$. We give two examples, one for $\rho = 1.5$ and one for $\rho = 5$, to illustrate the effect of this correction term. When $\rho = 1.5$, we get $c_a r(N)^{(a-1)/a} = 4.7133$ with $\mathbb{P}(\zeta < c_a r(N)^{(a-1)/a}) \approx 0.9268$, which means that we can expect 92.68% of the data below this correction term. We find that $\mathbb{E}_{\mathbb{P}}[\zeta - c_a r(N)^{(a-1)/a}] = -2.7132$, which roughly where most of the truncated means are concentrated in the figure. The truncation term is not very effective since $\mathbb{P}(\zeta > r(N)^{-1/a}) = 0.0025$.

Now for $\rho = 5$, we have that $c_a r(N)^{(a-1)/a} = 0.4124$ with $\mathbb{P}(\zeta < c_a r(N)^{(a-1)/a}) \approx 0.8221$. So most of the data points are still below the correction term. However, $\mathbb{E}_{\mathbb{P}}[\zeta - c_a r(N)^{(a-1)/a}] = -0.1625$ is much less drastic than for the $\rho = 1.5$ case, albeit still negative. The truncation term is not contributing with $\mathbb{P}(\zeta > r(N)^{-1/a}) = 2.6 \times 10^{-6}$. In the rest of the analysis, we omit the truncated mean.

Now we will analyse the remaining methods. We can see that both the sample mean (without tolerance) and the Wasserstein mean data points are frequently above the black line, meaning they are overestimating the theoretical mean. The Wasserstein estimates are below the sample mean by a small margin, corresponding to $\varepsilon = \log(N)/N$. Since $N$ stays fixed, and therefore also $\varepsilon$, the margin between Wasserstein and the sample mean increases relatively to the magnitude of the theoretical mean.

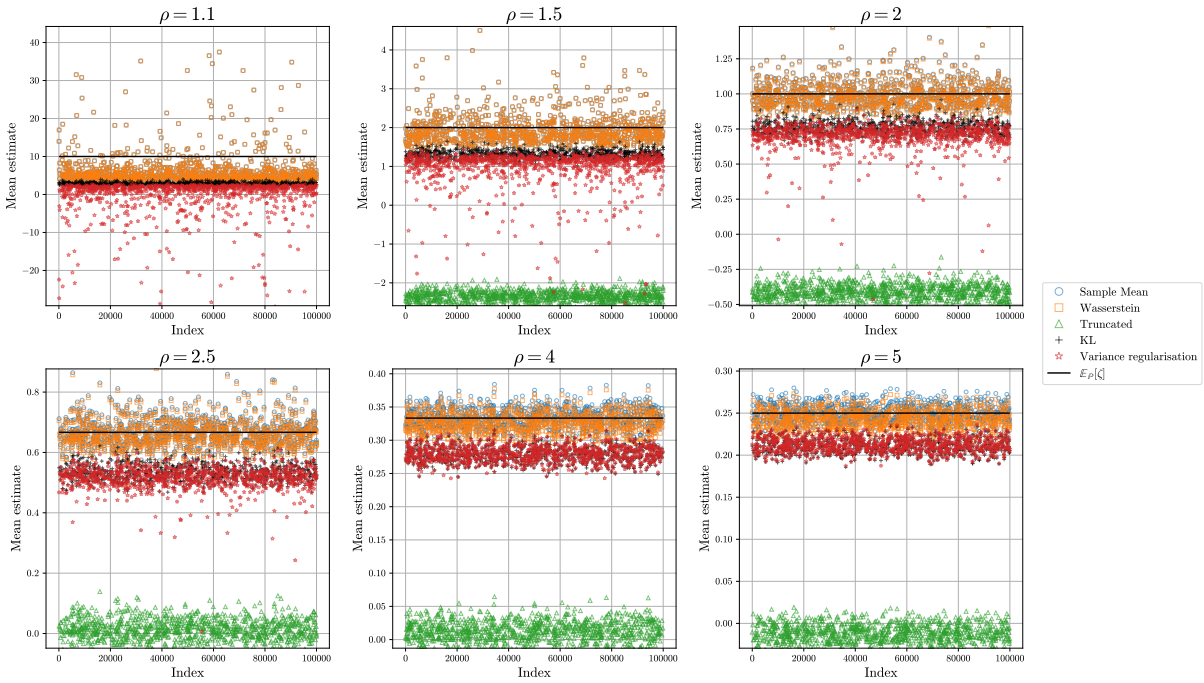



Figure 2.1: Random 1000 samples from = 100,000 Monte Carlo replications where the mean is estimated of $\zeta \sim \text{Pareto}(\rho) - 1$

Table 2.1: Fraction of data points that overestimated the mean in $B = 100{,}000$ trials. KL and VR are omitted since we did not observe overestimations.

| | $\rho = 1.1$ | $\rho = 1.5$ | $\rho = 2$ | $\rho = 2.5$ | $\rho = 4$ | $\rho = 5$ |
|---|---|---|---|---|---|---|
| $\hat{\mu}$ | 0.100 | 0.305 | 0.426 | 0.475 | 0.503 | 0.497 |
| Wasserstein | 0.100 | 0.294 | 0.384 | 0.403 | 0.311 | 0.238 |

Both KL and VR do not overestimate the theoretical mean in these samples. Moreover, for low $\rho$, notably the infinite variance cases, we see that KL is less conservative than VR. For VR, we see some negative mean estimates with an order of magnitude equal to the overestimations by the sample mean and Wasserstein. This can be explained by extreme outliers in the sample. When such an extreme outlier is present, both the sample mean and variance increase drastically. The latter causes VR to overcorrect.

From $\rho = 2.5$ onwards, we have finite variance and the outliers have less effect on the VR estimator, which performs as well as KL and even better when $\rho = 5$. However, only in $\rho = 5$ do we see that the VR estimates are well concentrated without any outliers. This can be explained by the fact that Pareto distributions with $\rho \leq 4$ has infinite kurtosis and therefore the $\mathbb{V}_{\mathbb{P}}(\hat{\sigma}^2) = \infty$.

In the entire sample of $B = 100{,}000$ for all values of $\rho$, the KL and VR mean estimations did not overestimate the theoretical mean for the sample mean and Wasserstein mean. In Table 2.1, we can see the fraction of mean estimates that overestimated the theoretical mean. We can see that for low values of $\rho$, the sample mean overestimates less frequently than for larger values of $\rho$. For Wasserstein, we can see that the frequency of overestimation diverges from the sample mean. As $\rho$ increases, the magnitude of the theoretical mean decreases, and $\varepsilon = \log(N)/N$ plays a larger role as the tolerance level since $N$ stays fixed. We will discuss how the variation of $N$ influences the mean estimation in the next section.

Lastly, we analyse the methods on their conservative probability. In Figure 2.2, we can see for all values of $\rho$, the empirical probability that the estimator was less than $100(1-c)\%$ of the mean. The solid black line is the bound shown in (2.5). Overall, the sample mean and the Wasserstein are the least conservative,

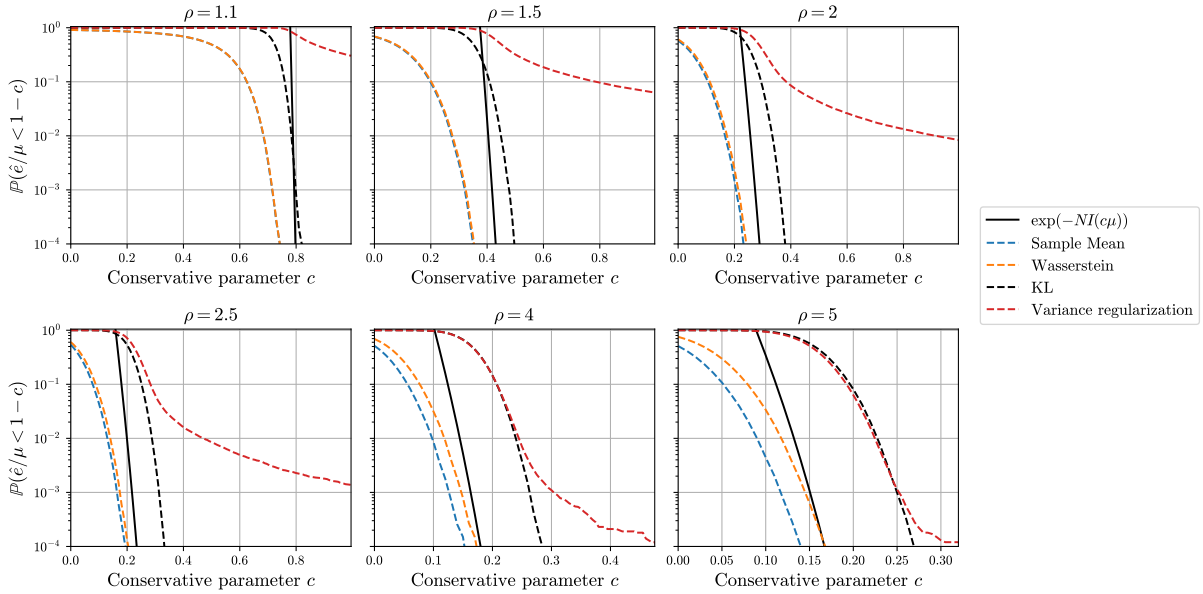



Figure 2.2: The conservative probability for different methods compared to the bound presented in [6].

but at the cost of frequent overestimation. For KL and VR, where no overestimation happened, we can see that for KL the conservative probability decays exponentially with the same rate as the theoretical bound. However, this is not the case for VR, where the decay flattens as $c$ increases.

## Experiment 2

In Figure 2.3, we can see the disappointment probabilities as the sample size $N$ varies. We have left out the disappointment probabilities of the variance regularisation method, as it had no overestimations. Similarly, KL only had 1 and 7 overestimations for $\rho = 1.5$ and $\rho = 2.5$ out of $B = 200{,}000$ trials. This rate is well below the targeted precision of 10% indicating that the theoretical bound in (2.21) is loose for these parametric choices.

For the sample mean, we can see that the overestimation probability sharply follows the asymptotic bound of (2.4) for both values of $\rho$. For the Wasserstein estimates, we see an increasing overestimation probability as $N$ increases. This can be caused by the fact that, for Wasserstein, $r(N) = \log(N)/N$ decreases as the sample size increases. At the same time, the probability of an outlier increases as the sample size

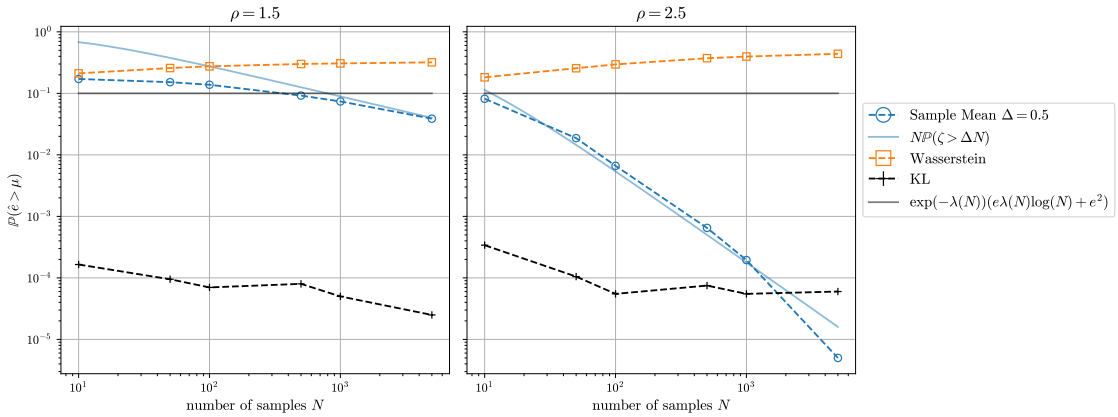



Figure 2.3: The disappointment probability as $N$ increases on log-log scale. The dotted lines are the empirical results while the solid lines are the theoretical bounds.

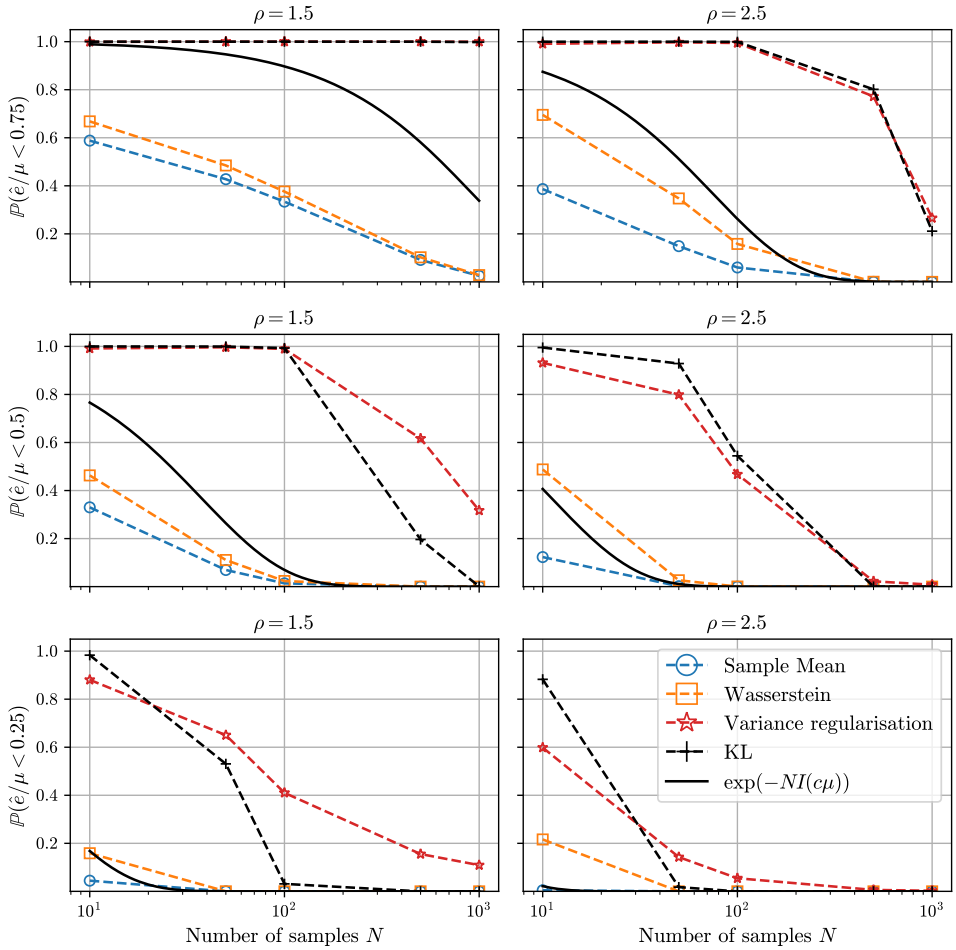



Figure 2.4: The conservative probability as $N$ increases. The dotted lines are the empirical results. The solid line is the asymptotic bound for both the sample mean and KL.

grows.

We can see in Figure 2.4 the conservative probability as $N$ grows for $c = 0.25$, 0.5, and 0.75. For these values of $c$, we can interpret $\mathbb{P}(\hat{e}/\mu < 1-c)$ as the probability that the estimator underestimated the mean by more than 25%, 50% and 75%, respectively. For the sample mean and Wasserstein, these methods are mostly below or on the asymptotic bound. We can see that there is a discrepancy between the rate at which the KL method decreases and the asymptotic bound $\exp(-NI(\mu c))$ as the KL line is always above the asymptotic bound.

The KL method is less conservative than variance regularisation for $\rho = 1.5$, except for low sample sizes where the probability of an extreme outlier is still small and no overcorrection happens. For greater sample sizes, KL is always less conservative than VR, and the difference gets bigger for greater values of $c$. Furthermore, for $\rho = 2.5$, KL and VR decrease almost identically, with VR less conservative for lower $N$ and KL less conservative for greater $N$. Again, we see that for greater $c$, KL is less conservative than VR.

## 2.4 Conclusion

In this chapter, we have studied the mean estimation of a Pareto distribution. We used both robust and distributionally robust estimation techniques to estimate the mean. We compared the bounds derived in Van Parys and Zwart [6] to the results of a Monte Carlo simulation. In these results, we have found

that both the sample mean and Wasserstein mean overestimate the theoretical mean with a significant probability, even though Wasserstein Relaxation was used. The truncated mean was found to be unreliable due to the correction term. The Kullback-Leibler and variance regularisation estimators did not overestimate the mean. Empirically, we confirmed that the Kullback-Leibler mean was less conservative than the variance regularisation estimates for all tested tail indices and especially for large sample sizes.

Finally, we recommend a direction for future research. Both the Wasserstein estimator and the KL estimator were derived by directly using the empirical distribution $\hat{\mathbb{P}}_N$ and its support $\hat{\Xi} = \{\tilde{\zeta}_1, ..., \tilde{\zeta}_N\}$. For future research, one could try to fit a kernel density estimator on the data where the kernel itself is heavy-tailed. Then the continuous equivalent Wasserstein and KL DRO could be applied. This would better represent the Pareto distribution as it will be continuous.

# Chapter 3

# Robust Wind Electricity Nomination

## 3.1 Introduction

In this chapter, we discuss and build upon the work of Aolaritei et al. [5]. This paper covers the nomination problem for wind electricity producers in the day-ahead market. In this nomination problem, producers must submit the amount of electricity they will produce for the day ahead (called the *nomination*), without knowing the exact electricity production or market prices of the next day. The producer receives the spot price (the price after the market is cleared) for the nominated production and has to settle the difference between the actual and nominated production with the regulation prices. If the producer promised more electricity than generated, the producer has to cover the difference according to the up-regulation prices. On the other hand, if more electricity is generated than nominated, then the difference is bought from the producer according to the down-regulation price.

The down-regulation prices are typically lower than the spot price. This means that the producer is most profitable when they nominate exactly what they generate. If the producer nominates less than what they generate, then the surplus of electricity will be sold for less. The up-regulation prices are generally higher than the spot price. For example, as can be seen in Figure 3.1, in May 2020, the up-regulation price was around 60 times the spot price. As a consequence, nominating more energy than will be generated can be extremely costly.

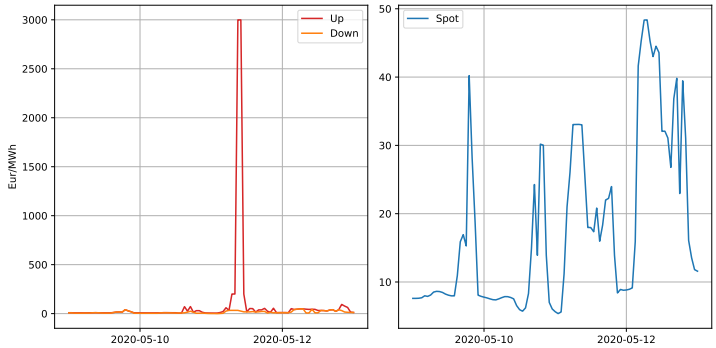



Figure 3.1: The up- and down-regulation prices in the crash of May 2020. The up-regulation price peaks at €3000 per MWh while the spot price was between €10 and €50 per MWh.

Aolaritei et al. argue that these prices follow a heavy-tailed distribution, albeit informally. In this chapter we will not test their claim since the objective is not to determine the exact tail index, but to be robust against extreme outliers. Furthermore, the asymmetry between up- and down-regulation prices is a good reason to apply robust optimisation. Aolaritei et al. show that Wasserstein DRO can be used to determine a more conservative nomination than the forecast, while still being profitable. In this chapter, we show that a more robust Kullback-Leibler-based DRO model gives better protection against market crashes.

## 3.2 Modelling

Our objective is to minimise the loss, and therefore maximise the profit. Formally, we define the profit function

$$\max_{x\in\mathbb{R}} \mathbb{E}_{\mathbb{P}}[\pi(x,\xi)] = -\min_{x\in\mathbb{R}} \mathbb{E}_{\mathbb{P}}[\ell(x,\xi)], \tag{3.1}$$

where we define $\pi : \mathbb{R} \times \Xi \to \mathbb{R}$ as the profit function, $x$ the nomination value and $\xi = (g, s, r^-, r^+)$ as a random vector containing the total energy generation $g$, the spot price $s$, the down-regulation price $r^-$, and the up-regulation price $r^+$. The loss function is determined as

$$\ell(x, \xi) = -\pi(x, \xi) = -(xs + (g - x)_+ r^- - (x - g)_+ r^+) \tag{3.2}$$

where $(t)_+ = \max\{t, 0\}$. In the first term $xs$, we quantify that the producer will receive the spot price for the nominated $x$. In the second term $(g-x)_+$, if more energy is generated than nominated, i.e. $g - x > 0$, then the producer receives $r^-$. In the occasional event that $r^- < 0$, the producer has to pay for the overproduction. For the last term $-(x - g)_+ r^+$, if less energy is generated than nominated, $g - x < 0$, then the producer must pay for the difference at price $r^+$.

We assume that $\xi \sim \mathbb{P}$ and the hourly samples $\xi_1, ..., \xi_N$ are i.i.d. copies of $\xi$. This assumption is debatable as dependence between hourly samples is observed. However, the DRO methods we study assume independence even though the statistical guarantees discussed in Section 1.4 might not apply.

Lastly, we have 52 wind energy generation forecast per hour, denoted as $\mathbf{f}_i \in \mathbb{R}^{52}$ for $i \in [N]$, that complete the training dataset $\{(\mathbf{f}_i, \xi_i)\}_{i=1}^N$. The price and wind energy data are obtained from Finland's transmission system operator Fingrid[1]. We used the same forecasts as Aolaritei et al. [5], which are 52 hourly forecasts obtained for different weather scenarios. The mean of the 52 forecasts is used as the Sample Average Approximation (SAA), which will be the benchmark nomination strategy.

## 3.3 Methodology

We follow the methodology of Aolaritei et al. [5], as this should give a good indication of how KL and HR DRO perform as a simple 'plug-in' method. The test data consists of one left-out season, with the remaining data used for training. However, not all training samples are used for calibration. A $\theta \in [0, 1]$ fraction of the training dataset is selected based on a new day-ahead forecast $\hat{\mathbf{f}}$ from the test data. The selected samples are the $\lceil \theta N \rceil$ samples whose minimum *historical* forecast $\min_{k\in[52]} f_{i,k}$ is closest in absolute distance to the minimum *day-ahead* forecast $\min_{k\in[52]} \hat{f}_k$. The index set of the selected samples is denoted as $\mathcal{I}$ such that $|\mathcal{I}| = \lceil \theta N \rceil$.

As reference distribution, given the day-ahead forecast $\hat{\mathbf{f}} \in \mathbb{R}^{52}$, we use the weighted empirical distribution

$$\hat{\mathbb{P}}_{N,\mathbf{w}} = \sum_{i\in\mathcal{I}} w_i W \delta_{\xi_i}, \tag{3.3}$$

where $w_i = \left(1 - \frac{|\min_{k\in[52]} \hat{f}_k - \min_{k\in[52]} f_{i,k}|}{d_{\max}}\right)^\beta$ are the raw weights, $d_{\max} = \max_{i\in\mathcal{I}} |\min_{k\in[52]} \hat{f}_k - \min_{k\in[52]} f_{i,k}|$, and $W = \left(\sum_{i\in\mathcal{I}} w_i\right)^{-1}$ is the normalising constant. This distribution will be used as a reference distribution for Wasserstein, Kullback-Leibler, and Holistic Robust DRO methods. In the remainder of this section, we show how the methods covered in Section 1.4 are be applied to the nomination problem.

### 3.3.1 Wasserstein method

The full description of the Wasserstein method is covered in Aolaritei et al. [5]. In this section, a summary of the steps is described.

As the transport cost function, the 1-norm $\|\cdot\|_1$ is chosen, which means we have a $W_1$ distance. Three choices of the radius are chosen as $\varepsilon \in \{0.5, 1, 1.5\}$. These values coincide with a low, medium and high level of robustness, respectively. The problem is reformulated with the deterministic reparametrisation

[1]The data is available at https://data.fingrid.fi/en/datasets. Specifically, the records are "The price of consumption imbalance electricity", "The sales price of production imbalance electricity", "The buying price of production imbalance electricity", and "Wind power generation - 15 min data". The main period of interest is between January 2017 and January 2021.

$\bar{\xi} = (g, s, r^-, r^+, t^-, t^+)$ with $t^- = gr^-$ and $t^+ = gr^+$, where $\bar{\xi}_i$ has the same weights as $\xi_i$ in the weighted empirical distribution. The support of $\bar{\xi}$ is defined as $\bar{\Xi} = \{\xi \in \mathbb{R}^d : C\xi \leq d\}$ with

$$C = \begin{pmatrix} -1 & & & & & \\ 1 & \ddots & & & & \\ & \ddots & \ddots & & & \\ & & \ddots & -1 & & \\ & & & 1 & & \end{pmatrix} \in \mathbb{R}^{12\times 6}, \quad d = \begin{pmatrix} -0.8\min\{g_i\} \\ 1.2\max\{g_i\} \\ \vdots \\ 1.2\max\{g_i r_i^+\} \end{pmatrix} \in \mathbb{R}^{12}, \tag{3.4}$$

which allows for an increase in the empirical support of 20%. This margin means that, for example, scenarios with 20% less than the minimum wind production and 20% more than the maximum wind production were considered. The same holds for the other variables. Because $r^-$ can be negative, the empirical minimum of both $r^-$ and $t^-$ is multiplied by $-1.2$, instead of $-0.8$.

By defining two auxiliary functions $a_1(x) = (0, -x, 0, x, 0, -1)$ and $a_2(x) = (0, -x, x, 0, -1, 0)$, the loss function from (3.2) can be defined as piecewise affine

$$\begin{aligned} \ell(x, \xi') &= \max\{xr^+ - t^+ - xs, xr^- - t^- - xs\} \\ &= \max\{a_1(x)^\top \xi', a_2(x)^\top \xi'\} \\ &= \max\{\ell_1(x, \xi'), \ell_2(x, \xi')\}. \end{aligned} \tag{3.5}$$

Now let $k \in \{1, 2\}$ and use (1.11), with the convex conjugate expressed as its dual

$$[-\ell_k(x, \cdot)]^*(\mathbf{z}_i - \mathbf{v}_i) = \sup_{C\xi' \leq d} (\mathbf{z}_i - \mathbf{v}_i + a_k(x))^\top \xi', \tag{3.6}$$

$$= \inf_{\eta_i \geq 0} \sup_{\xi'} \{(\mathbf{z}_i - \mathbf{v}_i + a_k(x))^\top \xi' - \eta_i^\top (C\xi' - d)\}, \tag{3.7}$$

$$= \inf_{\eta_i \geq 0} \eta_i^\top d + \underbrace{\sup_{\xi'} \{(\mathbf{z}_i - \mathbf{v}_i + a_k(x) - C^\top \eta_i)^\top \xi'\}}_{=0 \text{ if } \mathbf{z}_i - \mathbf{v}_i + a_k(x) = C^\top \eta_i \text{ and } +\infty \text{ otherwise}}, \tag{3.8}$$

$$= \inf_{\eta_i \geq 0} \{\eta_i^\top d : \mathbf{z}_i = C^\top \eta_i + \mathbf{v}_i - a_k(x)\}. \tag{3.9}$$

Similarly, $\sigma_{\bar{\Xi}}(\mathbf{v}_i) = \sup_{\xi' : C\xi' \leq d} \langle \mathbf{v}_i, \xi' \rangle = \inf_{\kappa_i \geq 0} \sup_{\xi'} \{\mathbf{v}_i^\top \xi' - \kappa_i^\top (C\xi' - d)\} = \inf_{\kappa_i \geq 0} \{\kappa_i^\top d : C^\top \kappa_i = \mathbf{v}_i\}$. Then, since $\mathbf{z}_i = C^\top \eta_i + \mathbf{v}_i - a_k(x)$ and $\mathbf{v}_i = C^\top \kappa_i$, we have $-\mathbf{z}_i^\top \xi' = -(\eta_i + \kappa_i)^\top C\xi' + a_k(x)^\top \xi'$ and $\|\mathbf{z}_i\|_* = \|C^\top(\kappa_i + \eta_i) - a_k(x)\|_*$. Since the 1-norm was chosen, we have that the dual norm $\|\cdot\|_* = \|\cdot\|_\infty$.

All the conditions in the dual are expressed in terms of $\kappa_i$ and $\eta_i$. We can simplify the problem by defining $\gamma_i = \kappa_i + \eta_i$. Furthermore, we can linearise by splitting up the conditions into the part where $\ell_1(x, \xi') > \ell_2(x, \xi')$ and vice versa. This results in the linear program presented in Aolaritei et al. [5]

$$\begin{aligned} \inf_{x, \lambda, s, \gamma_1, \gamma_2} \quad & \lambda\varepsilon + W \sum_{i \in \mathcal{I}} w_i s_i, \\ \text{s.t.} \quad & a_1(x)^\top \xi_i + \gamma_{i1}^\top (d - C\xi_i) \leq s_i \quad \forall i \in \mathcal{I}, \\ & a_2(x)^\top \xi_i + \gamma_{i2}^\top (d - C\xi_i) \leq s_i \quad \forall i \in \mathcal{I}, \\ & \left\| C^\top \gamma_{i1} - a_1(x) \right\|_\infty \leq \lambda \quad \forall i \in \mathcal{I}, \\ & \left\| C^\top \gamma_{i2} - a_2(x) \right\|_\infty \leq \lambda \quad \forall i \in \mathcal{I}, \\ & \boldsymbol{\gamma}_1 \in \mathbb{R}_+^{|\mathcal{I}| \times 12}, \boldsymbol{\gamma}_2 \in \mathbb{R}_+^{|\mathcal{I}| \times 12}, \lambda \geq 0. \end{aligned} \tag{3.10}$$

Lastly, the parameters $\theta = 1/3$ and $\beta = 2$ are copied from Aolaritei et al. [5].

### 3.3.2 Kullback-Leibler method

The KL DRO is implemented as described in (1.14). We increase $\theta$ from 1/3 to 0.5 since the KL method benefits from a larger sample, which increases the empirical support. We keep $\beta = 2$. We choose the radius $\varepsilon$ that solves (2.21)$= p$. We call this solution $\varepsilon_p$. Since only a limited amount of money can be earned, we know that $\ell$ is bounded from below. For this reason, even though the bound requires nonnegative losses, the bound is still valid. We choose $p = \frac{1}{25\times24\times365.25}$ for the most conservative KL DRO. Since the data is hourly, we interpret $p$ as the probability that we overestimate the expected profit once every 25 years. Then we also choose two levels of robustness besides the most conservative $\varepsilon_p$. For middle level of robustness, we choose $\frac{1}{2}\varepsilon_p$ and for low level of robustness we choose $\frac{1}{4}\varepsilon_p$.

### 3.3.3 Holistic Robust method

For the HR DRO, we decide to use the same radius $r \in \{\varepsilon_p, \frac{1}{2}\varepsilon_p, \frac{1}{4}\varepsilon_p\}$ as for the KL method[2]. However, since the upper bound in (2.21) is only applicable for the KL worst-case distributions, there is no connection between the chosen values for $r$ and the disappointment probability. Similar to KL, we use $\theta = 0.5$ and $\beta = 2$ for the empirical distribution.

We choose $\alpha = 0$ such that no misspecification is assumed. For the noise set, we choose the 1-norm ball $B_1 = \{z : \|z\|_1 \leq \varepsilon\}$. For any $\varepsilon > 0$, we know that HR DRO will be more conservative than KL DRO. Therefore, we choose the smallest $\varepsilon = 0.5$ that Aolaritei et al. [5] selected for their Wasserstein-1 DRO. This will allow the support to be expanded while still benefiting from the KL optimality. Then, by definition of the noise loss and the formulation of (3.5), we get

$$\begin{aligned}\ell^{\mathcal{N}}(x,\xi) &= \max_{z\in B_1} \ell(x,\xi - z)\\ &= \max_{z\in B_1} \max\{a_1(x)^\top \xi - a_1(x)^\top z, a_2(x)^\top \xi - a_2(x)^\top z\}\\ &= \max_{z\in B_1} \max\{\ell_1(x,\xi-z), \ell_2(x,\xi-z)\} = \max_{z\in B_1} \ell(x,\xi-z)\end{aligned}$$

If $\ell_1 > \ell_2$, then $\mathrm{d}\ell/\mathrm{d}z = a_1(x)$ and if $\ell_2 > \ell_1$, then $\mathrm{d}\ell/\mathrm{d}z = a_2(x)$. In both cases, if $x \geq 1$, then it suffices that $z = (0,\varepsilon,0,0,0,0)^\top = \varepsilon e_2$ with unit vector $e_2 = (0,1,0,0,0,0)^\top$. If $x < 1$, then $\varepsilon$ needs to be distributed over the $t^+$ and $t^-$ elements, which results in $z = (0,0,0,0,\varepsilon/2,\varepsilon/2)^\top$. For practical purposes, since $x \gg 1$, we define

$$\ell^{\mathcal{N}}(x,\xi) = \ell(x,(g,s-\varepsilon,r^-,r^+)) = \ell(x,\xi-\varepsilon e_2) \tag{3.11}$$

and ignore the $x < 1$ case. Lastly, we need to determine $\max_{\xi'\in\Xi} \ell(x,\xi')$. In the extreme case, we have nominated $x$ MWh, but there was no production ($g = 0$), and we would not get any money for it ($s = 0$). Then the loss would be $\max\{xr^+, xr^-\}$, and since typically $r^+ > r^-$, we have $\max_{\xi'\in\Xi} \ell(x,\xi') = \max_{r^+\in\{r_1^+,\ldots,r_N^+\}} xr^+$ with $r_i^+$ is the $r^+$ component of $\xi_i$ for all $i \in [N]$.

### 3.3.4 Method discussion

The reference distribution $\hat{\mathbb{P}}_{N,\mathbf{w}}$ was used in all three DRO methods, notably also in KL DRO. The disappointment probability of Van Parys et al. [3] uses the unweighted $\hat{\mathbb{P}}_N$ as reference distribution to determine the optimal robust optimiser with Sanov's theorem. Nevertheless, we used the weighted empirical distribution to match the methodology from the original paper as closely as possible. This way, we can evaluate the other DRO methods as a 'plug-in' method.

For HR DRO, the largest up-regulation price in the sample was selected as $\max_{r^+} r^+$. In future research, this value can also be better estimated with extreme value theory. If we assume that $r^+$ follows a power law with $\rho = 1$ tail index, then the $\max_{r^+} r^+$ can be approximated by a Fréchet distribution. However, this tail index should be carefully estimated. In [5], the authors describe that $r^+$ follows a power law

[2]Note that we use $\varepsilon$ for the LP distance and $r$ for the KL distance.

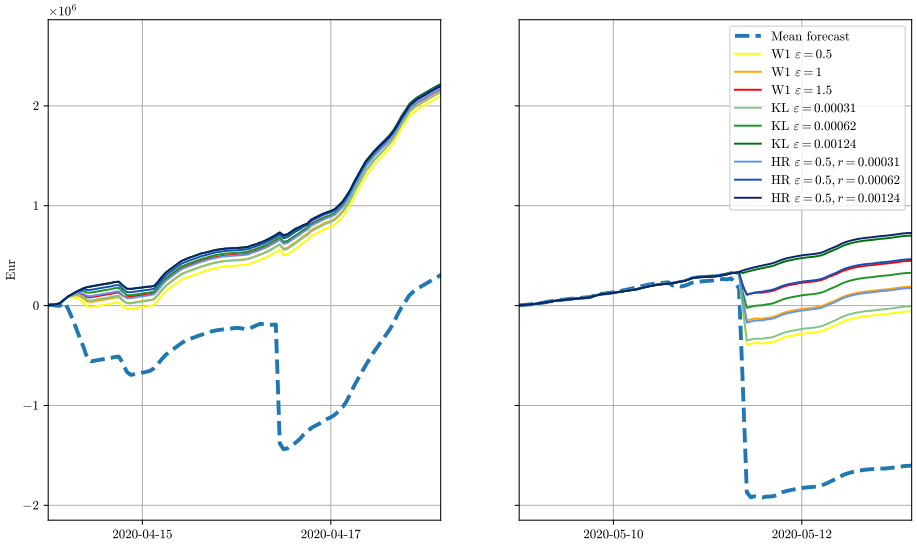



Figure 3.2: The nomination profit for the three different DRO methods during April 2020 (left) and May 2020 (right)

distribution based on different values for $q$ in $\mathbb{P}(r^+ \geq q)$, and then make an estimation of $\varepsilon$ based on this. However, to find a more exact estimation of $\mathbb{P}(r^+ \geq q)$, the Hill estimator could be used [14]. This would give a more exact estimation of the tail index $\rho$.

A similar strategy can be used to determine the extension of the support in the Wasserstein DRO method. More extensive research could be done on the effect of this margin. An extreme value theory approach as described previously could be beneficial to link the extended size of the support to statistical properties.

Finally, we used one season as a test set and the remaining seasons as a training set, including data after the test season. Of course, as a decision-maker, if you have to make decisions for a season, then the data after that season is not available. It could be that using the data from summer, fall and winter of 2020 leaks information into the spring of 2020 evaluation. For example, if there was less economic activity in Finland due to the COVID-19 pandemic, then this would reflect in the entire 2020 year data. If more data was available, then a moving window approach could be researched. In that case, for the day-ahead nomination, you use the previous four years of data.

## 3.4 Results

We first present the results on the 2020 spring crashes before looking at the performance of other seasons. In Figure 3.2, we can see how the three different DRO methods perform in the crashes of April and May 2020. For the crash in April, we can see that all DRO methods have similar profit, with KL having slightly more profit compared to the other types. For the second crash in May, we can see that KL and HR are

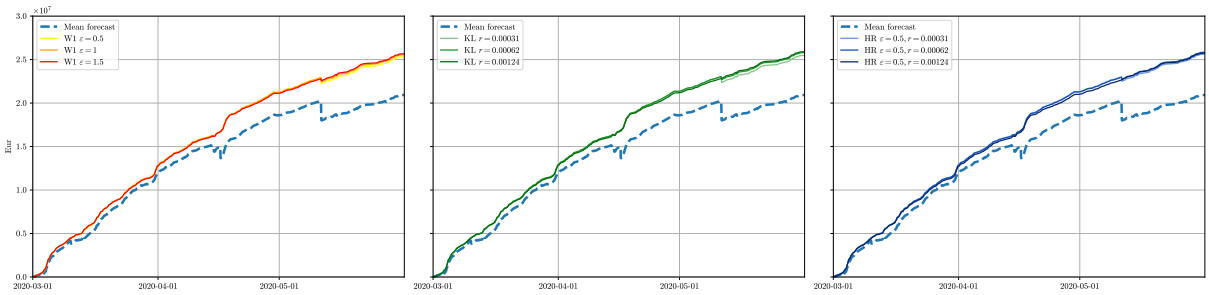



Figure 3.3: The nomination profit for the three different DRO methods during 2020 spring

Table 3.1: The profit ($\times$ €1 million) of the DRO methods at the end of 2020 spring.

| Model | Profit |
|---|---|
| $W_1$ $\varepsilon = 0.5$ | 25.35 |
| $W_1$ $\varepsilon = 1$ | 25.57 |
| $W_1$ $\varepsilon = 1.5$ | 25.69 |
| KL $\varepsilon = 0.00031$ | 25.49 |
| KL $\varepsilon = 0.00062$ | 25.81 |
| **KL** $\varepsilon = 0.00124$ | **25.91** |
| $HR_{\varepsilon=0.5}$ r=0.00031 | 25.64 |
| $HR_{\varepsilon=0.5}$ r=0.00062 | 25.85 |
| $HR_{\varepsilon=0.5}$ r=0.00124 | 25.75 |

significantly more robust than the Wasserstein DRO. The most robust configuration of HR DRO does not lose money during the crash, while the most robust KL only slightly does. The middle robust HR matches the crash of the Wasserstein DRO highest robust configuration, but is slightly more profitable in April.

In Figure 3.3, we can see that all methods largely follow the same pattern during the 2020 spring season. All methods vastly outperform the mean forecast nomination strategy. Furthermore, the more robust configurations of the methods (high for Wasserstein and KL, and middle for HR) end up being more profitable at the end of the season, even though these configurations were less profitable before the crash in May.

If we look at the final profit of 2020 spring in Table 3.1, we can see that KL with $\varepsilon_p = 0.00124$ is the most profitable, albeit with a small margin (0.86% compared to Wasserstein and 0.23% compared to HR). Then we have the middle and high robust HR and middle robust KL being similar with only 0.04 difference between middle robust HR and middle robust KL. Lastly, we have the low robust KL and HR with the Wasserstein methods.

In Table 3.2, we can see how the DRO methods have performed on all the seasons. A couple of patterns can be inferred. First of all, as expected, HR DRO is always more conservative than KL DRO when comparing the same robust level configurations. This results in a lower profit across the seasons except

Table 3.2: The % of profit compared to the SAA strategy for different DRO methods. The colouring is done per row.

| | SAA | Wasserstein | | | KL | | | HR $\varepsilon = 0.5$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Season | x1 million | 0.5 | 1 | 1.5 | 0.00031 | 0.00062 | 0.00124 | 0.00031 | 0.00062 | 0.00124 |
| 2016 winter | € 26.01 | -0.53% | -0.87% | -1.25% | -0.91% | -1.50% | -2.58% | -1.21% | -1.80% | -2.78% |
| 2017 spring | € 33.45 | -1.68% | -2.07% | -2.49% | -1.99% | -2.83% | -4.39% | -2.41% | -3.25% | -4.71% |
| 2017 summer | € 27.74 | 0.82% | 0.84% | 0.73% | 0.58% | 0.32% | -0.58% | 0.49% | 0.14% | -0.75% |
| 2017 autumn | € 40.71 | -0.19% | -0.20% | -0.29% | -0.70% | -1.14% | -2.44% | -0.82% | -1.40% | -2.59% |
| 2017 winter | € 50.99 | -0.37% | -0.57% | -0.81% | -0.75% | -1.34% | -2.70% | -1.01% | -1.66% | -2.88% |
| 2018 spring | € 45.20 | -1.69% | -2.08% | -2.41% | -2.80% | -3.73% | -7.00% | -3.20% | -4.29% | -7.91% |
| 2018 summer | € 61.27 | -0.10% | -0.21% | -0.35% | -0.39% | -0.72% | -1.49% | -0.53% | -0.90% | -1.58% |
| 2018 autumn | € 19.18 | -0.12% | -0.20% | -0.30% | -0.26% | -0.50% | -1.14% | -0.36% | -0.66% | -1.22% |
| 2018 winter | € 71.27 | 0.16% | 0.05% | -0.10% | -0.02% | -0.35% | -1.18% | -0.18% | -0.55% | -1.29% |
| 2019 spring | € 60.00 | -0.16% | -0.33% | -0.54% | -0.54% | -0.99% | -2.01% | -0.73% | -1.24% | -2.12% |
| 2019 summer | € 41.73 | -0.32% | -0.54% | -0.83% | -0.69% | -1.23% | -2.42% | -0.92% | -1.52% | -2.59% |
| 2019 autumn | € 52.82 | 0.78% | 0.65% | 0.44% | 0.62% | -0.04% | -1.51% | 0.34% | -0.40% | -1.68% |
| 2019 winter | € 59.30 | -0.20% | -0.55% | -0.99% | -0.96% | -2.12% | -4.61% | -1.47% | -2.78% | -4.85% |
| 2020 spring | € 20.94 | 21.04% | 22.07% | 22.65% | 21.71% | 23.25% | 23.73% | 22.40% | 23.41% | 22.95% |
| 2020 summer | € 21.73 | 0.60% | 0.31% | -0.17% | -0.08% | -1.08% | -3.44% | -0.49% | -1.66% | -3.77% |
| 2020 autumn | € 20.75 | 0.10% | -0.15% | -0.54% | -0.88% | -2.02% | -4.75% | -1.37% | -2.78% | -4.97% |
| Total | € 653.09 | 0.48% | 0.32% | 0.10% | 0.08% | -0.47% | -1.89% | -0.16% | -0.80% | -2.13% |

for 2020 spring. However, while the middle robust HR DRO has slightly less profit in the 2020 spring season compared to the high robust KL DRO, it is less conservative in the rest of the year, showing a profit much closer to the SAA strategy.

Similarly, we see that Wasserstein is less conservative than KL except in 2020 spring. In the regular seasons, we see that the Wasserstein profit is close to that of SAA, while this is only the case for low robust KL. Middle and high robust KL deviate more from the SAA strategy, although they are still profitable.

Finally, we can see in Table 3.2 that the total profit of high robust KL and medium robust HR was not enough to be equally profitable to SAA while this is the case for the high robust Wasserstein. One reason for this is that the total profit made in 2020 spring is the third-lowest in the entire period of interest, which is also partially because of the crashes in that season.

## 3.5 Conclusion

In this chapter, we have studied how DRO can be used for the nomination problem in day-ahead energy markets. We used Wasserstein as followed in Aolaritei et al. [5] and applied the same methodology, but with two different DRO strategies: Kullback-Leibler and Holistic Robust. This gives a good indicator of how well the KL and HR DRO perform when we simply 'plug in' these two methods.

From the empirical results, we find that the KL and HR DRO strategies have protected better against shocks of large up-regulation costs than Wasserstein and the naive SAA. Especially during the shock of May 2020, the most robust configuration of KL and HR DRO barely lost money during the crashes.

In the other seasons during the period of interest, KL and HR DRO are still profitable strategies, albeit less profitable than Wasserstein and SAA. Even though KL and HR were robust during a period with large shocks, these methods lost performance to SAA and Wasserstein in the absence of these shocks. Empirically, over a period of 4 years, the cost of being insured against a crash was 1.89% for the most robust KL DRO and 0.8% for the middle robust HR DRO compared to the SAA strategy.

Based on the fact that; HR with medium robustness is more profitable than KL with high robustness; and HR shows similar protection in volatile seasons, a hypothesis can be stated that the price and generation data not only contain sampling errors, but also noise. This is the reason why HR was developed, but a more extensive study can be done to validate this. The possibility that the data contains noise can also explain why Wasserstein ends up being more profitable than SAA over the 4 years, since Wasserstein protects against noise. If that is the case, then hedging against black swan events is a positive side effect.

For further research, we recommend a similar suggestion as in Chapter 2. The weighted empirical distribution has support on $\{\xi_1, ..., \xi_N\}$, which limits the possibilities for the worst-case distribution. By fitting a kernel density estimator with heavy-tail kernels, it could be possible to use a continuous version of the DRO methods.

Furthermore, other similarity score combinations could be researched instead of the Lévy-Prokhorov metric with KL in the HR DRO method. Since we do not assume any contamination of the sample (since $\alpha = 0$), other metrics could be applied that focus more on the noise aspect. A class of DRO that has been studied is the smoothed $\phi$-divergence DRO in Liu et al. [18], where KL is not only combined with LP but also with the Wasserstein-$p$ distance.

Finally, we recommend modelling the electricity prices and power generation. These models can be calibrated on the Finnish data and used to generate synthetic data, potentially with added noise. For example, the up-regulation prices can be generated from a regularly varying distribution. By varying the tail-index $\rho$, the three DRO methods can be tested for robustness on different degrees of variability. In such research, a more complete picture of the three DRO variants in electricity markets could be established.

# Chapter 4

# Robust Portfolio Selection

## 4.1 Introduction

In this chapter, we will look at the portfolio selection problem. In this problem, we have $d$ assets and must choose how we want to distribute the available funds among those assets. Specifically, let our decision variable $x$ be a vector with nonnegative entries and $\|x\|_1 = 1$. Then for all $k \in [d]$, $x_k$ can be interpreted as the fraction of total funds that was allocated to asset $k$. For simplicity, we assume that the possible assets consist of $d-1$ stocks that have a stochastic yearly return and one savings account that returns $R^*$ risk-free.

We choose to approach the portfolio selection in a novel way. The goal of this portfolio selection problem is not to maximise the returns but to maximise the probability of achieving $R^+$ more returns than the risk-free rate. We will call this probability the 'exceeding probability'. Mathematically, let $\xi$ be the yearly returns of our $d$ assets under unknown probability measure $\mathbb{P}$ and $\mathcal{X} = \{x \in \mathbb{R}^d_+ : \|x\|_1 = 1\}$. Then the optimisation problem is defined as

$$\max_{x \in \mathcal{X}} \mathbb{P}\left(x^\top \xi \geq R^*(1+R^+)\right). \tag{4.1}$$

Since $\mathbb{P}$ is unknown, we will replace it by a distributionally robust worst-case measure $\mathbb{Q}$. Therefore we get the DRO problem

$$\max_{x \in \mathcal{X}} \min_{\mathbb{Q} \in \mathcal{A}} \mathbb{Q}\left(x^\top \xi \geq R^*(1+R^+)\right). \tag{4.2}$$

Here $\mathcal{A}$ is the ambiguity set as described in Section 1.4, which differs based on the similarity function that is applied. Note that this problem is still of the standard stochastic programming form when we consider the expected reward $\mathbb{Q}(x^\top \xi \geq R^*(1+R^+)) = \mathbb{E}_{\mathbb{Q}}(r(x,\xi))$ with $r(x,\xi) = \mathbb{1}\left\{x^\top \xi \geq R^*(1+R^+)\right\}$.

Clearly, the decision-maker only observes one annual return per year, which limits the amount of information the decision-maker has. For this reason, we will not directly work on the problem defined in (4.2), but rather on a proxy problem. This proxy will be based on the *daily log-returns* instead of the annual returns under model assumptions. However, since our goal is to guarantee a certain annual return, we will evaluate it as such.

In the remainder of this chapter, we will describe how the experiments are set up and the results are evaluated. For instance, in Section 4.2, we describe how the stock data is synthetically generated with two heavy-tailed stochastic processes. In Section 4.3, we describe how the min-max problem in (4.2) is solved for both the Wasserstein and Kullback-Leibler methods. We also describe the proxy problem in more detail and how we will evaluate the results. Then in Section 4.4 we show and discuss the empirical results from the experiments. Lastly, in Section 4.5, we conclude our findings and this chapter.

**Notation** We use $\mu$ as the drift parameter and $\sigma$ as a scaling parameter[1]. The vector equivalents are denoted by $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ such that for all $i \in [d]$, we have that $\boldsymbol{\mu}_i$ and $\boldsymbol{\sigma}_i$ are the drift and scaling parameters of stock $X^{(i)}$ from random vector $\mathbf{X} = [X^{(1)}, ..., X^{(d)}]$. The superscripts of the stochastic process $X_t$ indicate which model is being used. We use $X \overset{d}{=} Y$ to state that $X$ and $Y$ share the same underlying distribution.

[1] We avoid using the term "volatility" as this parameter does not solely determine the volatility of a stock.

### 4.1.1 Preliminaries

Here we define a couple of stochastic processes that are used in the modelling section: Lévy process, Brownian motion, Poisson Process and Gamma Process. However, to keep the modelling section brief, we will define these processes in this section. The reader may skip this section if desired.

First, we define the Lévy process. A Lévy process $\{X_t\}_{t\geq 0}$ is a stochastic process if the process has the following properties:

- *Starts at the origin*: $X_0 = 0$ almost surely,
- *Independent and stationary increments*: For any $0 \leq t_1 < t_2$ and $0 \leq s_1 < s_2$ such that $[t_1, t_2] \cap [s_1, s_2] = \varnothing$, we have $X_{t_2} - X_{t_1}$ is independent of $X_{s_2} - X_{s_1}$. Furthermore, $X_{t_2} - X_{t_1} \overset{d}{=} X_{t_2 - t_1}$.
- *Continuous in probability*: for all $\varepsilon > 0$, for all $t \geq 0, \mathbb{P}(|X_{t+\delta} - X_t| > \varepsilon) \to 0$ as $\delta \to 0$.
- *Càdlàg sample paths*: Sample paths are, almost surely, right-continuous with left limits.

The Brownian motion (BM), Poisson process and Gamma process are all Lévy processes. For BM $X_t$ with drift $\mu$, we have $X_{t_2} - X_{t_1} \sim \mathcal{N}(\mu(t_2 - t_1), t_2 - t_1)$. We write $W_t$ as a Standard Brownian motion when $\mu = 0$. For the Poisson process $X_t^{\mathrm{P}}$ we assume that the increments follow a Poisson distribution

$$\mathbb{P}(X_{t+\Delta t}^{\mathrm{P}} - X_t^{\mathrm{P}} = k) = \frac{(\mu_{\mathrm{P}} \Delta t)^k \exp(-\mu_{\mathrm{P}} \Delta t)}{k!}, \quad k \in \mathbb{N}, \tag{4.3}$$

and for a Gamma process we assume that the increments follow a Gamma distribution with density for all $z > 0$

$$f_{\Gamma(c_1, c_2)}(z) = \frac{1}{c_2^{c_1} \Gamma(c_1)} z^{c_1 - 1} e^{-z/c_2}, \tag{4.4}$$

and 0 otherwise.

## 4.2 Modelling

We have stock $S_t = S_0 \exp(X_t)$ which is defined on the time interval $[0, T]$ with time steps $\Delta t \ll T$. Without loss of generality, since we can take $\Delta t$ arbitrarily small, we take $T = 1$ year. Then we define the log-returns as $\log(S_T / S_0) = X_T$, which will follow a heavy-tailed distribution. We define $\mathbf{S}_t = [S_t^{(1)}, ..., S_t^{(d-1)}]$ as a vector of $d-1$ stocks which are assumed not to be independent. Specifically, we define $\bar{\sigma}_{i,j} = \mathrm{corr}(X^{(i)}, X^{(j)})$ as the correlation coefficient between the log-returns of stock $i$ and $j$, where $X^{(i)} = [X_0^{(i)}, X_{\Delta t}^{(i)}, ..., X_T^{(i)}]$. The final dataset will be the daily log-returns of one year $\{\mathbf{X}_0, \mathbf{X}_{\Delta t} - \mathbf{X}_0, ..., \mathbf{X}_T - \mathbf{X}_{T - \Delta t}\}$ with $\Delta t = \frac{1}{260}$ (assuming 260 trading days in a year).

We consider two stochastic processes for $X_t$ called Brownian motion with Poisson Jumps $X_t^{\mathrm{J}}$ (also called Jump Diffusion Process, abbreviated as JDP) and Variance Gamma Process (VGP) $X_t^{\mathrm{VG}}$. Both of these processes are Lévy processes, where the JDP has finite activity, and the VGP has infinite activity. Finite and infinite activity means that the processes typically have, respectively, a finite and infinite number of jumps in a finite time interval [19].

### 4.2.1 Jump Diffusion Process

The Jump Diffusion Process (JDP) is a Brownian motion extended with jumps driven by a Poisson process with mean $\mu_{\mathrm{P}} t \geq 0$ [19]. We can interpret $\mu_{\mathrm{P}}$ as the average number of jumps in a year. Clearly, since $\mu_{\mathrm{P}} < \infty$, we have finite activity. Now, the JDP can be defined as

$$X_{t+\Delta t}^{\mathrm{J}} - X_t^{\mathrm{J}} = \mu \Delta t + \sigma(W_{t+\Delta t} - W_t) + J(X_{t+\Delta t}^{\mathrm{P}} - X_t^{\mathrm{P}}), \tag{4.5}$$

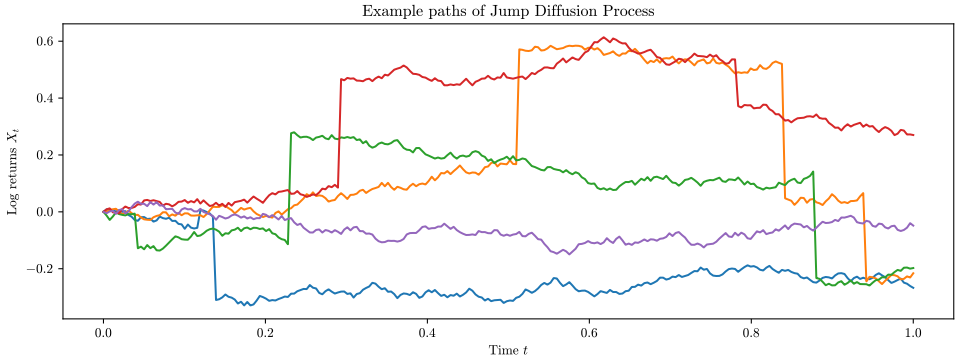



Figure 4.1: An example of 5 JDP paths. We can see Brownian motion samples between periods of jumps. There are more downward jumps as $p_K = 0.4 < 0.5$. But the upward jumps have greater magnitude since $\eta_1 < \eta_2$.

where $W_t$ is a SBM independent of $X_t^{\mathrm{P}}$ and $J$ is a random jump magnitude. This definition is a discretised version of the continuous version that can be found in Oosterlee and Grzelak [19]; however, to avoid unnecessary technicalities, we stick to the discretised version for both discussed processes.

An example of JDP paths can be found in Figure 4.1. We can clearly see the jumps in the stock process. Between the jumps, the process is light-tailed since the increments are from a Brownian motion. By adding jumps to a Brownian motion, we increase the probability mass to the tails. Since $W_t$ has a normal distribution, the resulting JDP will have heavier tails than a normal distribution, which makes it heavy-tailed but not necessarily regularly varying. The distribution of jump magnitude $J$ and the mean of the Poisson process $\mu_{\mathrm{P}}$ determine how much more mass is transferred to the tails. If $J \equiv 0$ or $\mu_{\mathrm{P}} \equiv 0$, then $X_t^{\mathrm{J}}$ is a regular Brownian motion. As discussed before, if we increase $\mu_{\mathrm{P}}$, we get more jumps of size $J$ and therefore increasingly heavier tails. By choosing a model for the jumps, we can determine the behaviour of $J$ and therefore also the behaviour of $X_t^{\mathrm{J}}$.

A popular choice for $J$ is Kou's model [19, 20]. Let $p_K \in [0,1]$, $\eta_1 > 1$, and $\eta_2 > 0$. Then in Kou's model, we have that $J$ has probability density function

$$f_J(z) = p_K \eta_1 \exp(-\eta_1 z)\, \mathbb{1}\{z \geq 0\} + (1-p_K)\eta_2 \exp(\eta_2 z)\, \mathbb{1}\{z < 0\}, \tag{4.6}$$

which is a non-symmetric double exponential distribution. We favour this model over Merton's model [21] where $J$ is Gaussian, because of this non-symmetric property. This non-symmetric property matches the empirical phenomena of asymmetric tail behaviour [20]. We can interpret $p_K$ as the probability of an upward jump with mean upward jump magnitude $1/\eta_1$. Then similarly, $1-p_K$ is the probability of a downward jump with mean downward jump $1/\eta_2$.

The closed form distribution function of $X_t^{\mathrm{J}}$ is not known. However, the characteristic function $\phi_{X_t^{\mathrm{J}}}(u) = \mathbb{E}[\exp(iuX_t^{\mathrm{J}})]$ is known and can be used to numerically estimate the distribution function. The characteristic function of the JDP is given by [19]

$$\phi_{X_t^{\mathrm{J}}}(u) = \exp\left(iu\mu t - \frac{1}{2}\sigma^2 u^2 t\right) \exp\left(\mu_{\mathrm{P}} t\, \mathbb{E}\left[e^{iuJ} - 1\right]\right), \tag{4.7}$$

with

$$\mathbb{E}\left[e^{iuJ}\right] = p_K \frac{\eta_1}{\eta_1 - iu} + (1-p_K)\frac{\eta_2}{\eta_2 + iu}. \tag{4.8}$$

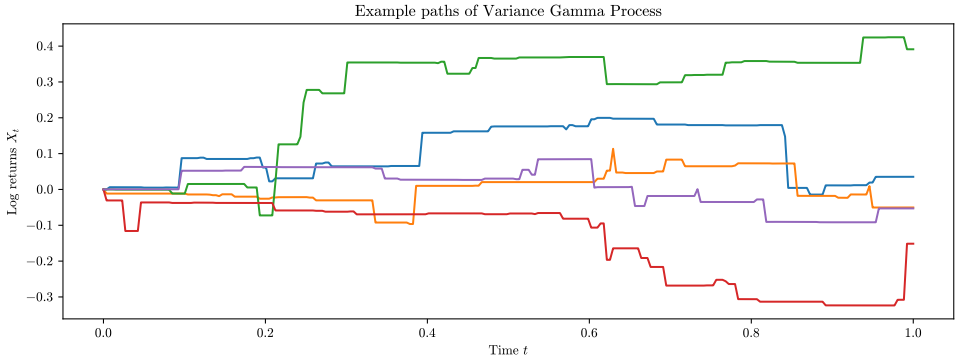



Figure 4.2: An example of 5 VGP paths. Unlike JDP, there is no clear Brownian motion behaviour. The paths are mostly constant between jumps. This is because VGP is a pure jump process.

### 4.2.2 Variance Gamma Process

Similar to JDP, the Variance Gamma Process is also an adaptation to the Brownian motion that modifies the BM into a stochastic process with a heavy-tailed distribution. The idea of VGP is to evaluate a Brownian motion (with drift) at stochastic times following a Gamma Process $\Gamma_{t;1,\beta}$. This adaptation results in infinite jump activity [22] and where we normally have $\mathbb{V}(W_t) = t$, we now have that $\mathbb{V}(W_{\Gamma_{t;1,\beta}} \mid \Gamma_{t;1,\beta}) = \Gamma_{t;1,\beta}$ is stochastic. Since the time steps can be greater than $\Delta t$, larger steps than BM are possible, and therefore VGP has heavier tails than a Gaussian [23]. However, VGP is not regularly varying since all moments exist.

First we define the Gamma Process as a Lévy process with increments $\Gamma_{t+\Delta t;1,\beta} - \Gamma_{t;1,\beta}$ distributed according to a Gamma distribution with mean $\Delta t$ and variance $\beta \Delta t$ with $\beta > 0$. Now substitute the time variable $t$ with $\Gamma_{t;1,\beta}$ in Brownian motion

$$t \mapsto \theta t + \sigma W_t \tag{4.9}$$

with drift $\theta \in \mathbb{R}$, scaling $\sigma > 0$ and SBM $W_t$. We assume $W_t$ and $\Gamma_{t;1,\beta}$ are independent. Then by adding a deterministic drift term to the process, we get that the VGP is defined as

$$X_t^{\mathrm{VG}} = \mu t + \theta \Gamma_{t;1,\beta} + \sigma W_{\Gamma_{t;1,\beta}}. \tag{4.10}$$

Since $\beta$ controls the variance of the Gamma Process, we have that $\beta$ controls the variance of $\mathbb{V}(W_{\Gamma_{t;1,\beta}} \mid \Gamma_{t;1,\beta})$, and therefore the kurtosis of $W_{\Gamma_{t;1,\beta}}$.

A couple of paths are presented in Figure 4.2. We can see much less variation between the large jumps compared to JDP. This is because VGP is a pure jump process [19].

Similar to the JDP, there is no closed-form distribution function, but we have the characteristic function [19]

$$\phi_{X_t^{\mathrm{VG}}}(u) = \exp(iu\mu t)\left(1 - i\beta\theta u + \frac{1}{2}\beta\sigma^2 u^2\right)^{-t/\beta}. \tag{4.11}$$

## 4.3 Methodology

### 4.3.1 Correlated stocks

To create a multivariate stock process with correlations, we can define the processes without the drift term as $\tilde{\boldsymbol{X}}_t$. Then $\tilde{\boldsymbol{X}}_t$ is a vector of independent stock processes that can be correlated with a Cholesky decomposition $\boldsymbol{C} = \boldsymbol{L}\boldsymbol{L}^\top$ where $\boldsymbol{C} \in \mathbb{R}^{d-1 \times d-1}$ is the covariance matrix of $\tilde{\boldsymbol{X}}_t$ such that $\boldsymbol{C}_{i,j} = \bar{\sigma}_{i,j}\sigma_i\sigma_j$ for

all $i, j \in [d-1]$. For example, for all $i \in [d-1]$, we define an independent JDP as

$$\tilde{X}^{\mathrm{J},(\mathrm{i})}_{t+\Delta t} - \tilde{X}^{\mathrm{J},(\mathrm{i})}_{t} = \sigma(W^{(i)}_{t+\Delta t} - W^{(i)}_{t}) + J(X^{\mathrm{P},(\mathrm{i})}_{t+\Delta t} - X^{\mathrm{P},(\mathrm{i})}_{t}) \tag{4.12}$$

where $W^{(i)}_t$ and $X^{\mathrm{P},(\mathrm{i})}_t$ are independent processes. Then given the Cholesky decomposition of the covariance matrix $\boldsymbol{C} = \boldsymbol{L}\boldsymbol{L}^\top$ and drift vector $\boldsymbol{\mu} \in \mathbb{R}^{d-1}$, we can calculate the JDP stock process as

$$\boldsymbol{X}^{\mathrm{J}}_{t+\Delta t} - \boldsymbol{X}^{\mathrm{J}}_{t} = \boldsymbol{\mu}\Delta t + \mathbf{L}^\top \left(\tilde{\boldsymbol{X}}^{\mathrm{J}}_{t+\Delta t} - \tilde{\boldsymbol{X}}^{\mathrm{J}}_{t}\right). \tag{4.13}$$

We fix the correlation coefficients $\bar{\boldsymbol{C}}$ by generating a random correlation matrix following the algorithm of Davies and Higham [24]. This algorithm needs the eigenvalues $\Lambda \in \mathbb{R}^{d-1}_{+}$ of $\bar{\boldsymbol{C}}$ as input. The eigenvalues have been generated such that $\|\Lambda\|_1 = d-1$ and for all $i \in [d-1]$, $\Lambda_i \propto \tilde{\Lambda}_i$ where $\tilde{\Lambda}_i$ is evenly spaced such that $\Lambda_1 = \frac{1}{d-1}$ and $\Lambda_{d-1} = 0.5$.

### 4.3.2 Density recovery

In this section, we give a short description of the COS-method as described in more detail in Oosterlee and Grzelak [19]. The COS-method allows us to recover the probability density functions of $X^{\mathrm{J}}_t$ and $X^{\mathrm{VG}}_t$ with their characteristic functions. To recover the density from the characteristic function, we use the relation

$$\phi_X(u) = \int_{\mathbb{R}} e^{iux} f_X(x)\,\mathrm{d}x \quad \Leftrightarrow \quad f_X(x) = \frac{1}{2\pi}\int_{\mathbb{R}} e^{-iux}\phi_X(u)\,\mathrm{d}u. \tag{4.14}$$

for any random variable $X$.

The COS-method is based on the Fourier cosine expansion defined as

$$f_X(x) = \frac{1}{2}A_0 + \sum_{k=1}^{\infty} A_k \cos\left(k\pi\frac{x-a}{b-a}\right) \tag{4.15}$$

with coefficients

$$A_k = \frac{2}{b-a}\int_a^b \cos\left(k\pi\frac{x-a}{b-a}\right) f_X(x)\,\mathrm{d}x. \tag{4.16}$$

The idea of the COS-method is to replace the $A_k$ term with an expression that is not dependent on the density. First, the COS-method evaluates the truncated integral for all $[a,b] \subseteq \mathbb{R}$

$$\hat{\phi}_X(u) = \int_a^b e^{iux} f_X(x)\,\mathrm{d}x. \tag{4.17}$$

For a large enough interval $[a,b]$, we have that $\hat{\phi}_X(u) \approx \phi_X(u)$. Then, $\hat{\phi}$ is evaluated at $u = \frac{k\pi}{b-a}$ and scaled with $\exp\left(-i\frac{ka\pi}{b-a}\right)$ to get

$$\hat{\phi}_X\left(\frac{k\pi}{b-a}\right)\exp\left(-i\frac{ka\pi}{b-a}\right) = \int_a^b \exp\left(ix\frac{k\pi}{b-a} - i\frac{ka\pi}{b-a}\right) f_X(x)\,\mathrm{d}x. \tag{4.18}$$

Finally, we will conclude that, if we take the real part of the above integral, we will get an expression similar to (4.16). First, we write $ix\frac{k\pi}{b-a} - i\frac{ka\pi}{b-a} = ik\pi\frac{x-a}{b-a}$. Then, because taking the real part of the integral is equal to the integral of the real part, and by Euler's formula $e^{i\theta} = \cos\theta + i\sin\theta$, we have

$$\mathfrak{Re}\left\{\exp\left(ik\pi\frac{x-a}{b-a}\right)\right\} = \cos\left(k\pi\frac{x-a}{b-a}\right), \tag{4.19}$$

and therefore,

$$\mathfrak{Re}\left\{\hat{\phi}_X\left(\frac{k\pi}{b-a}\right)\exp\left(-i\frac{ka\pi}{b-a}\right)\right\} = \int_a^b \cos\left(k\pi\frac{x-a}{b-a}\right) f_X(x)\,\mathrm{d}x. \tag{4.20}$$

And now since $A_k$ is proportional to (4.20), we have that

$$f_X(x) = \frac{1}{b-a}\mathfrak{Re}\{\hat{\phi}_X(0)\} + \frac{2}{b-a}\sum_{k=1}^{\infty}\mathfrak{Re}\left\{\hat{\phi}_X\left(\frac{k\pi}{b-a}\right)\exp\left(-i\frac{ka\pi}{b-a}\right)\right\}\cos\left(k\pi\frac{x-a}{b-a}\right). \tag{4.21}$$

By definition of the characteristic function, we know that $\phi(0) = 1$, and therefore we substitute $\hat{\phi}_X(0)$ with $\phi(0) = 1$. Then, the density can be numerically approximated if we truncate the sum to $K-1$ terms

$$\hat{f}_{X,K}(x) = \frac{1}{b-a} + \frac{2}{b-a}\sum_{k=1}^{K-1}\mathfrak{Re}\left\{\hat{\phi}_X\left(\frac{k\pi}{b-a}\right)\exp\left(-i\frac{ka\pi}{b-a}\right)\right\}\cos\left(k\pi\frac{x-a}{b-a}\right). \tag{4.22}$$

Then, including the $\hat{\phi}_X(0)$ term, we have $K$ terms to approximate the density with for all $x \in [a,b]$, $\hat{f}_{X,K}(x) \to f_X(x)$ as $K \to \infty$ with exponential convergence when the density is smooth (i.e. no jumps and on convex support) [19].

For later calculations, we will need the distribution function $F_X$, which we can approximate by integrating the recovered density

$$\hat{F}_{X,K}(z) = \int_a^z \hat{f}_{X,K}(x)\,\mathrm{d}x \tag{4.23}$$

$$= \int_a^z \frac{1}{b-a}\,\mathrm{d}x + \int_a^z \frac{2}{b-a}\sum_{k=1}^{K-1}\mathfrak{Re}\left\{\hat{\phi}_X\left(\frac{k\pi}{b-a}\right)\exp\left(-i\frac{ka\pi}{b-a}\right)\right\}\cos\left(k\pi\frac{x-a}{b-a}\right)\mathrm{d}x. \tag{4.24}$$

Then by Fubini's theorem, we can interchange the sum and integral to get

$$\hat{F}_{X,K}(z) = \frac{z-a}{b-a} + \frac{2}{\pi}\sum_{k=1}^{K-1}\mathfrak{Re}\left\{\hat{\phi}_X\left(\frac{k\pi}{b-a}\right)\exp\left(-i\frac{ka\pi}{b-a}\right)\right\}\frac{\sin\left(k\pi\frac{z-a}{b-a}\right)}{k}. \tag{4.25}$$

We will use the COS-method to recover the distribution function of the portfolio log-returns $x^\top\xi$. This distribution function could also be approximated by the empirical distribution function $\hat{\mathbb{P}}_N(x^\top\xi \le t)$, which is a sum of indicator functions. However, if we are optimising over $x$, then the indicators are adding an integer component to the mathematical program. Instead we use $\hat{F}_{x^\top\xi,K}$, which is continuous.

### 4.3.3 Wasserstein and KL DRO

First, just like the other applications, we choose the Wasserstein-1 DRO, whose derivation follows the same steps as discussed in Chapter 2. In Appendix B, we derived the dual of the inner minimisation problem for all $x \in \mathcal{X}$ and $\varepsilon > 0$,

$$\max_{\lambda\ge 0}\left\{-\lambda\varepsilon + \sum_{i=1}^{N}\hat{p}_i \min\left\{1, \lambda\frac{x^\top\tilde{\xi}_i - R^*(1+R^+)}{\|x\|_\infty}\right\}\mathbb{1}\{x^\top\tilde{\xi}_i \ge R^*(1+R^+)\}\right\} = \max_{\lambda\ge 0} h(\lambda), \tag{4.26}$$

with slope

$$h'(\lambda) = -\varepsilon + \sum_{i=1}^{N}\hat{p}_i\frac{x^\top\tilde{\xi}_i - R^*(1+R^+)}{\|x\|_\infty}\mathbb{1}\left\{0 \le \lambda\frac{x^\top\tilde{\xi}_i - R^*(1+R^+)}{\|x\|_\infty} \le 1\right\}. \tag{4.27}$$

If $\varepsilon$ is greater than the sum in (4.27), in particular when $x^\top \tilde{\xi}_i < R^*(1+R^+)$ for all $i \in [N]$, then the slope is always negative and the maximum is attained at $\lambda = 0$. In all other cases, we prove in Appendix B that the maximum is attained at $\lambda = \frac{\|x\|_\infty}{x^\top \tilde{\xi}_{i^*} - R^*(1+R^+)}$, where index $i^* = \arg\min_{i\in[N]} \{x^\top \xi_i - R^*(1+R^+) : x^\top \xi_i > R^*(1+R^+)\}$ can be interpreted as the $i^{\text{th}}$ scenario that yields the smallest log-return greater than $R^*(1+R^+)$[2].

This results in the following closed-form Wasserstein worst-case probability

$$\min_{\mathbb{Q}\in\mathcal{A}_{W_1}(\hat{\mathbb{P}}_N)} \mathbb{Q}(x^\top \xi \geq R^*(1+R^+)) = \max\left\{0, \hat{\mathbb{P}}_N\left(x^\top \xi \geq R^*(1+R^+)\right) - \frac{\varepsilon\|x\|_\infty}{x^\top \xi_{i^*} - R^*(1+R^+)}\right\}. \tag{4.28}$$

Now since the empirical distribution is discontinuous and depends on the decision variable $x$, we approximate the problem in terms of the smooth COS-method recovered distribution function from (4.25). Using this method, we have a continuous (nonconvex) program instead of an integer program, which is tractable in practice. We then get that

$$\max_{x\in\mathcal{X}} \max\left\{0, 1 - \hat{\mathbb{P}}_N\left(x^\top \xi \leq R^*(1+R^+)\right) - \frac{\varepsilon\|x\|_\infty}{x^\top \xi_{i^*} - R^*(1+R^+)}\right\} \tag{4.29}$$

$$= \max\left\{0, \max_{x\in\mathcal{X}}\left\{1 - \hat{\mathbb{P}}_N\left(x^\top \xi \leq R^*(1+R^+)\right) - \frac{\varepsilon\|x\|_\infty}{x^\top \xi_{i^*} - R^*(1+R^+)}\right\}\right\} \tag{4.30}$$

$$\approx \max\left\{0, 1 - \min_{x\in\mathcal{X}}\left\{\hat{F}_{x^\top \xi, K}\left(R^*(1+R^+)\right) + \frac{\varepsilon\|x\|_\infty}{x^\top \xi_{i^*} - R^*(1+R^+)}\right\}\right\}. \tag{4.31}$$

To determine $\varepsilon$, we performed cross-validation on 10 JDP and 10 VGP paths for different values of $\varepsilon$. On average, $\varepsilon = 2^{-9}N^{-1}$ gave the highest probability to beat the $R^*$ returns by $R^+$%.

We use $\varepsilon = 0$ as benchmark as this gives us the SAA strategy

$$\min_{x\in\mathcal{X}} \hat{F}_{x^\top \xi, K}\left(R^*(1+R^+)\right). \tag{4.32}$$

For KL DRO, we have the dual problem

$$-\min_{\substack{x\in\mathcal{X},\\ \lambda\geq 0,\mu>0}} \lambda(\varepsilon - 1) + \mu + \lambda\left(\log\left(\frac{\lambda}{\mu+1}\right)\hat{\mathbb{P}}_N(x^T\xi > t) + \log\left(\frac{\lambda}{\mu}\right)\hat{\mathbb{P}}_N(x^T\xi \leq t)\right), \tag{4.33}$$

which is also derived in Appendix B. Again, we approximate the empirical distribution with the COS-method approximation to avoid working with a sum of indicator functions. This results in

$$-\min_{\substack{x\in\mathcal{X},\\ \lambda\geq 0,\mu>0}} \lambda(\varepsilon - 1) + \mu + \lambda\left(\log\lambda - \log(\mu+1)\left[1 - \hat{F}_{x^\top \xi, K}\left(R^*(1+R^+)\right)\right] - \log(\mu)\hat{F}_{x^\top \xi, K}\left(R^*(1+R^+)\right)\right). \tag{4.34}$$

We determine $\varepsilon$ by finding the root of (2.21) $-p$. The value of $p$ was determined with cross-validation on 10 JDP and 10 VGP paths. On average, all candidates gave roughly the same results, but $p = 2^{-10}N^{-1}$ gave slightly better results. Note that this formulation is a approximation of the KL that was proved to be optimal since we now approximate the empirical distribution.

The condition that $\|x\|_1 = 1$ and $x_i \geq 0$ for all $i \in [d]$ are given as strong conditions in the implementation.

### 4.3.4 Experiments

We generate 100,000 paths of 9 stocks with both asset models for $T = 1$ year and time step $\Delta t = \frac{1}{260}$ being one day (assuming 260 trading days in a year). These asset models have random parameters, but

[2]We use strictly greater than $R^*(1+R^+)$ to avoid division by 0 and it does not affect the outcome.

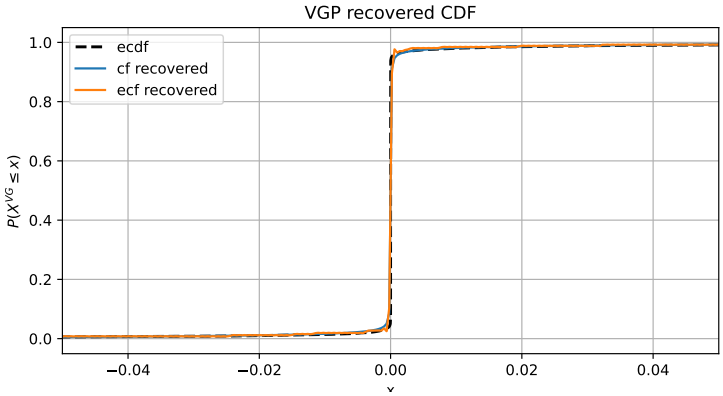



Figure 4.3: Example of COS-method for daily log-return data with characteristic function (cf), empirical cf and empirical cumulative distribution function (ecdf). For COS, we use $b = 1$ and $K = 5000$.

the random parameters have been chosen such that a JDP asset with the average parameter set has the same mean, variance and kurtosis as the VGP asset with an average parameter set. To be more precise, for the JDP we sampled $\sigma_{JDP} \sim \text{Unif}[0, 0.254]$, $\mu_P \sim \text{Unif}[1, 3]$, $p_K \sim \text{Unif}[0.4, 0.776]$, $\eta_1 \sim \mathcal{N}(7, 1.5)$ and $\eta_2 \sim \mathcal{N}(7, 1.5)$[3]. For VGP, we selected $\sigma_{VGP} \sim \text{Unif}[0, 0.632]$, $\theta \sim \mathcal{N}(0.11, 0.05)$ and $\beta \sim \text{Unif}[0.3, 0.7]$. Once sampled, these parameters stay fixed for the entire experiment. The drift terms are deterministic with for $i \in [d]$, we have $\mu_i = -1 + (i-1)\frac{2}{d-1}$ such that $\mu_1 = -1$ and $\mu_{d-1} = 1$. Furthermore, we have one risk-free savings account where $R^*\%$ is the yearly returns ranging from 0.1% to 1.5%. We want to surpass this savings with $R^+ = 0.2$.

For each simulation, we only have one annual return. Therefore, we scale back the exceeding probability to daily returns to get 260 data points to fit a distribution on. This means we will optimise the following

$$\max_x \mathbb{P}(x^\top(\mathbf{X}_{t+\Delta t} - \mathbf{X}_t) \geq R^*(1 + R^+)\Delta t). \tag{4.35}$$

We scale $R^*(1 + R^+)$ to *daily* log-returns by multiplying by the time step as a heuristic. By stationarity of Lévy processes, we have $\mathbf{X}_{t+\Delta t} - \mathbf{X}_t \overset{d}{=} \mathbf{X}_{\Delta t}$, which we will use in the calculations of the empirical characteristic function, which is defined as

$$\hat{\phi}_{\mathbf{X}_{\Delta t},N}(u) = \mathbb{E}_{\hat{\mathbb{P}}_N}[\exp(iu^\top \mathbf{X}_{\Delta t})], \qquad \forall u \in \mathbb{R}^d. \tag{4.36}$$

We assume that the actual characteristic function of the asset process is unknown to make the method completely data-driven and therefore rely on the empirical characteristic function (ECF). We use the ECF in the COS-method for the Wasserstein and KL DRO. In Figure 4.3, we can see an example of the COS-method applied on a VGP path. The recovered cdf with the known characteristic function fits well to the empirical distribution function. The recovered distribution function with the ECF has more error, especially close to $x = 0$, but this error is acceptable. Note the distribution function is no longer monotonically increasing since we approximate it with cosine functions.

We will evaluate the SAA, $W_1$ and KL DRO portfolios on the average exceeding probability and the average log-returns. Even though we only optimise for the exceeding probability, the average returns should give an indication of the return magnitude. These two metrics are calculated using the distribution function of $(x^*)^\top \xi$, which is retrieved using the COS-method and the characteristic function under $\mathbb{P}$. This can be done by rewriting $\phi_{x^\top \xi}$ into the product of characteristic functions, as we will show now.

First we simplify the problem by splitting up the savings and stocks $x^\top \xi = x_1 R^* + \tilde{x}^\top \boldsymbol{Y}_T$ where $\boldsymbol{Y}_T = \mu T + L^\top \tilde{\boldsymbol{X}}_T$ are the correlated stocks and $\tilde{x} = [x_2, \ldots, x_d]^\top$ the stock weights. For ease of notation, write

[3]We need $\eta_1 > 1$ and $\eta_2 > 0$, but sampling from a normal gives a non-zero probability that these conditions do not hold. However, this probability is around $10^{-5}$.

$\tilde{\boldsymbol{X}}_T = \tilde{\boldsymbol{X}}$. Then by expansion of the exponents, we get, for $u \in \mathbb{R}$,

$$\phi_{x^\top \xi}(u) = \mathbb{E}_{\mathbb{P}}\left[\exp(iux^\top \xi)\right] = \exp(iux_1 R^*)\exp(iu\tilde{x}^\top \mu T)\,\mathbb{E}_{\mathbb{P}}\left[\exp(iu\tilde{x}^\top (L^\top \tilde{\boldsymbol{X}})\right]. \tag{4.37}$$

Now since $\tilde{\boldsymbol{X}}$ consists of independent stock processes, we can use this independence to write the expectation in terms of the product of expectations. This is done by the following calculations

$$\mathbb{E}_{\mathbb{P}}\left[\exp(iu\tilde{x}^\top (L^\top \tilde{\boldsymbol{X}}))\right] = \mathbb{E}_{\mathbb{P}}\left[\exp\left(iu\sum_{j=1}^{d-1}\tilde{x}_j\left(\sum_{k=1}^{d-1}L_{jk}\tilde{X}_k\right)\right)\right] = \mathbb{E}_{\mathbb{P}}\left[\exp\left(iu\sum_{k=1}^{d-1}\sum_{j=1}^{d-1}\tilde{x}_j L_{jk}\tilde{X}_k\right)\right], \tag{4.38}$$

$$= \mathbb{E}_{\mathbb{P}}\left[\prod_{k=1}^{d-1}\exp\left(iu\tilde{x}^\top L_{\cdot k}\tilde{X}_k\right)\right], \tag{4.39}$$

$$= \prod_{k=1}^{d-1}\mathbb{E}_{\mathbb{P}}\left[\exp\left(iu\tilde{x}^\top L_{\cdot k}\tilde{X}_k\right)\right] = \prod_{k=1}^{d-1}\phi_{\tilde{X}_k}(u\tilde{x}^\top L_{\cdot k}), \tag{4.40}$$

where $L_{\cdot k}$ is the $k$-th column vector of $L$, and where (4.40) is valid since we are taking the expectation of a product of independent random variables. We can use this result in (4.37) to get

$$\phi_{x^\top \xi}(u) = \exp(iux_1 R^*)\exp(iu\tilde{x}^\top \mu T)\prod_{k=1}^{d-1}\phi_{\tilde{X}_k}(u\tilde{x}^\top L_{\cdot k}), \tag{4.41}$$

where we use the actual characteristic function of $\tilde{X}_k$ to estimate the actual exceeding probability to evaluate the data-driven methods.

### 4.3.5 Method discussion

The recovered density and distribution function have two types of approximation errors. The first approximation error is from (4.18) since the integral is taken over a subset $[a, b]$ of the real line, and the second approximation error comes from (4.22) where only the first $K$ terms of the infinite series is calculated. We determined $K = 5000$, $b = 1$ and $a = -b$ by checking if the distribution function matches the empirical distribution of a couple samples. However, one could determine $K$ and $b$ by solving

$$\min_{K\in\mathbb{N},\, b>0}\int |\hat{F}_{X,K}(x) - \tilde{F}_{X,N}(x)|\,\mathrm{d}x, \tag{4.42}$$

where $\tilde{F}_{X,N}(x)$ is the empirical distribution function. Then, given a sample, the COS-recovered distribution function will be fit per sample. This would make sure that each COS-recovery would be tailored to the sample, but it does increase the computation time.

Another approximation error in this application is the scaling from annual to daily returns. In (4.35), we scale the term $R^*(1+R^+)$ by the time step. However, this limits the portfolio evaluation in two ways. First we discuss the effect of a heavy left tail. On one day, we can earn the target return of $R^*(1+R^+)\Delta t$, but lose more than twice the target return on another day. On daily evaluation, this leads to one success event and one non-success event. However, two non-success events of earning $\frac{1}{2}R*(1+R^+)\Delta t$ can get you closer to the target $R*(1+R^+)$ on annual bases. The same argument holds for a heavy right tail. This causes the model to be greedy for short-term returns.

We analysed the methods on synthetic data which allows us to calculate the highly accurate exceeding probability since we know the characteristic function. However, there is no guarantee that the results of this research can directly be applied as a real-world trading strategy. The method is applicable to real world data, but careful analysis on the jump size and activity should be done in advance.

Lastly, the parameters for the JDP and VGP samples have been sampled once from a distribution per asset. Sampling from such distributions allows for independent selection of the parameters and the result

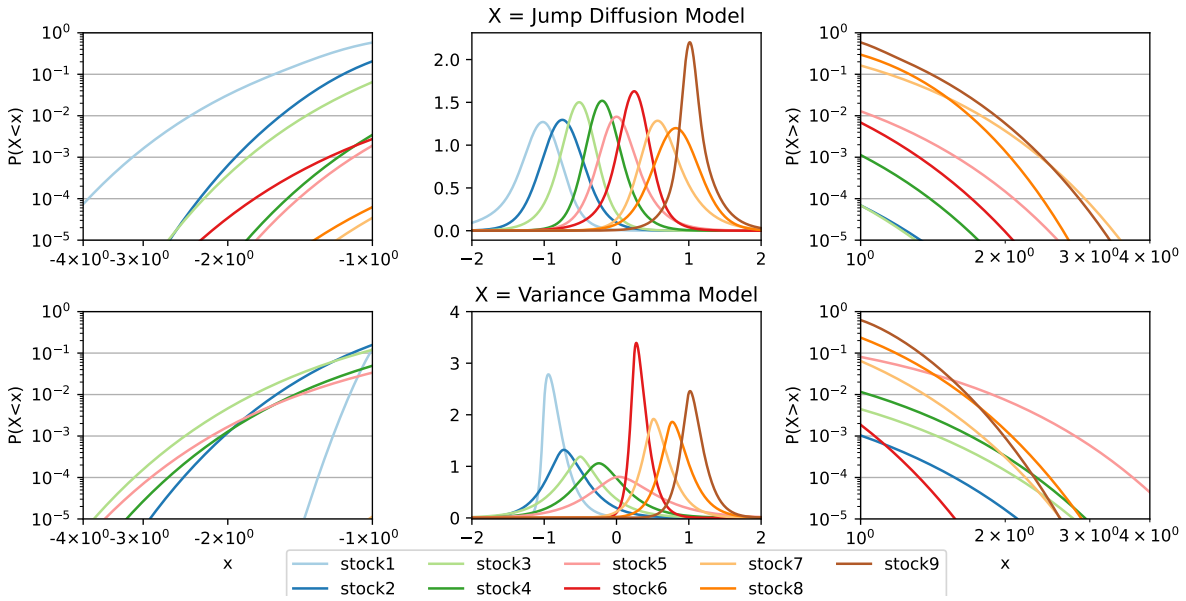



Figure 4.4: The distributions of 9 different assets from 3 perspectives for both the JDP (top) and VGP (bottom) processes. Left: the left tail on a symmetric log plot. Middle: the densities of the 9 assets. Right: the right tail on a loglog plot.

outcome. However, the resulting assets could lead to implausible relationships between assets. There are two different ways to select the parameters. One is by designing the assets such that they would match a class of assets similar to those in the real world, ranging from conservative to speculative. The other way is by selecting certain assets from the real world and matching the parameters from option implied volatility curves, as described in Oosterlee and Grzelak [19]. Both methods would have the benefit that the assets would better reflect the real world and would be less likely to lead to arbitrage opportunities if done correctly.

## 4.4 Results

First of all, we recovered the distribution function of the assets. In Figure 4.4, we can see two sets of distributions, the top for JDP and the bottom for VGP. By design, the assets have drift ranging from $-1$ to 1 in equal steps. The other parameters can be found in Table 4.1.

The first thing we can note from Figure 4.4 is that the distributions for JDP are similar in shape up to a difference in variance. The magnitude of the tail probabilities of the stocks are roughly in order of $\mu$. This is in contrast with the distributions of VGP where the distributions vary a lot more in skewness and tails. For example, the distributions of stocks 4 and 5 are much wider than the distribution of stocks 1 and 6. Furthermore, stocks 1 and 5 are right-skewed, while the others are more or less symmetric.

Table 4.1: The parameters for each asset visualised in Figure 4.4

| | | JDP | | | | | VGP | | |
|---|---|---|---|---|---|---|---|---|---|
| **stocks** | $\mu$ | $\sigma$ | $p_K$ | $\mu_P$ | $\eta_1$ | $\eta_2$ | $\sigma$ | $\theta$ | $\beta$ |
| 1 | -1 | 0.227 | 0.424 | 1.166 | 7.520 | 3.804 | 0.133 | 0.184 | 0.572 |
| 2 | -0.75 | 0.245 | 0.491 | 1.479 | 7.112 | 8.045 | 0.382 | 0.092 | 0.541 |
| 3 | -0.5 | 0.196 | 0.432 | 1.443 | 7.733 | 6.849 | 0.473 | -0.002 | 0.699 |
| 4 | -0.25 | 0.193 | 0.704 | 1.727 | 8.338 | 8.453 | 0.478 | 0.025 | 0.524 |
| 5 | 0 | 0.180 | 0.464 | 2.621 | 6.189 | 9.763 | 0.628 | 0.125 | 0.519 |
| 6 | 0.25 | 0.178 | 0.473 | 1.120 | 7.285 | 5.309 | 0.138 | 0.090 | 0.557 |
| 7 | 0.5 | 0.195 | 0.706 | 1.899 | 5.618 | 7.842 | 0.269 | 0.072 | 0.591 |
| 8 | 0.75 | 0.247 | 0.705 | 2.626 | 10.199 | 7.050 | 0.268 | 0.090 | 0.546 |
| 9 | 1 | 0.095 | 0.622 | 1.528 | 6.429 | 6.975 | 0.201 | 0.069 | 0.535 |

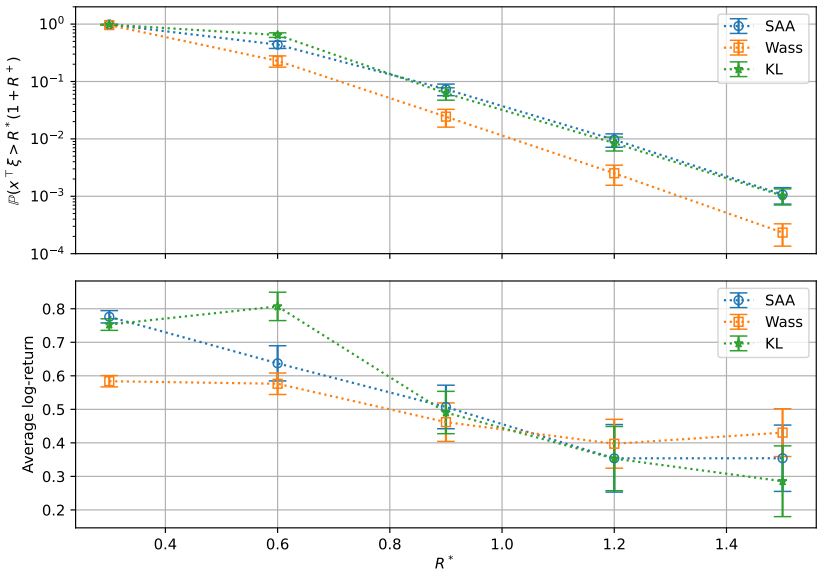



Figure 4.5: The average exceeding probability and log-return of 100,000 simulations for the JDP assets. The top of the error bars are the mean absolute deviation to the mean probability/return.

In Figure 4.5, we can see the results for the JDP assets. For all values of $R^*$, SAA and KL outperform Wasserstein on the exceeding probability. Considering the margins, there is little difference between the exceeding probability of SAA and KL. For the returns, again SAA and KL are similar, except for $R^* = 0.9$ where KL had a higher average return. Taking the margin into account, Wasserstein has a lower average return for $R^* = 0.3$, but not much else can be concluded since the error bars are overlapping.

In Figure 4.6, we can see that KL has a better exceeding probability than Wasserstein by a slight margin. Furthermore, we see that KL has a higher or equal average exceeding probability than SAA and that Wasserstein has a lower or equal average exceeding probability than SAA. Regarding the average returns, we see a great variation for the lower $R^*$ values. KL has a greater or equal average return than the other strategies, where again SAA is in between KL and Wasserstein. The average returns are equal for $R^* \geq 0.9$.

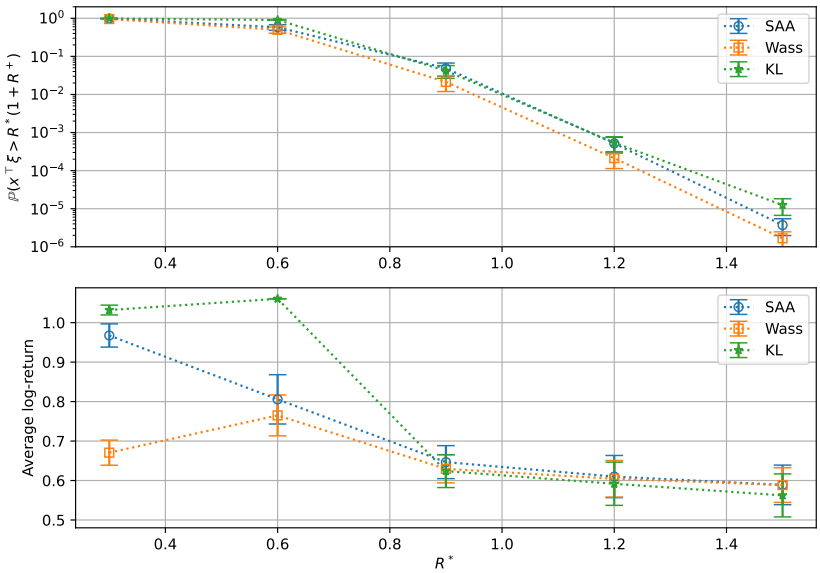



Figure 4.6: The average exceeding probability and log-return of 100,000 simulations for the VGP assets. The error bars are constructed in the same way as Figure 4.5.

An interesting phenomenon is the similarity between SAA and KL. In Figure 4.7 we can see the weights for the assets generated by JDP and VGP. For each of these figures, we can see the weights of the assets as $R^*$ increases. For all strategies, we can see a pattern where there is interest in stock 9 for low values of $R^*$. For JDP, the weights shift to stock 6, 7 and 8 as $R^*$ increases. For VGP, the weights become more uniform with the highest weight to the risk-free asset. Interestingly, both SAA and KL show fractional interest in stock 1 and 2 even though they have, on average, negative returns and heavy left tails.

We can see that the Wasserstein weights stay below 0.3 and 0.45 for JDP and VGP respectively. When KL puts 100% of the resources in stock 9, Wasserstein only commits 42%. Wasserstein has, just like SAA and KL, higher weights for stock 9 for $R^* = 0.3$ and 0.6 and then the weights for stocks 6, 7 and 8 start to increase for JDP and for the risk-free for VGP. Even though this pattern is similar, we see clearly that the highest stock weight is dampened. This dampening can be partially explained by the $\|x\|_\infty$ term in the Wasserstein dual since it penalises the highest weights of all assets.

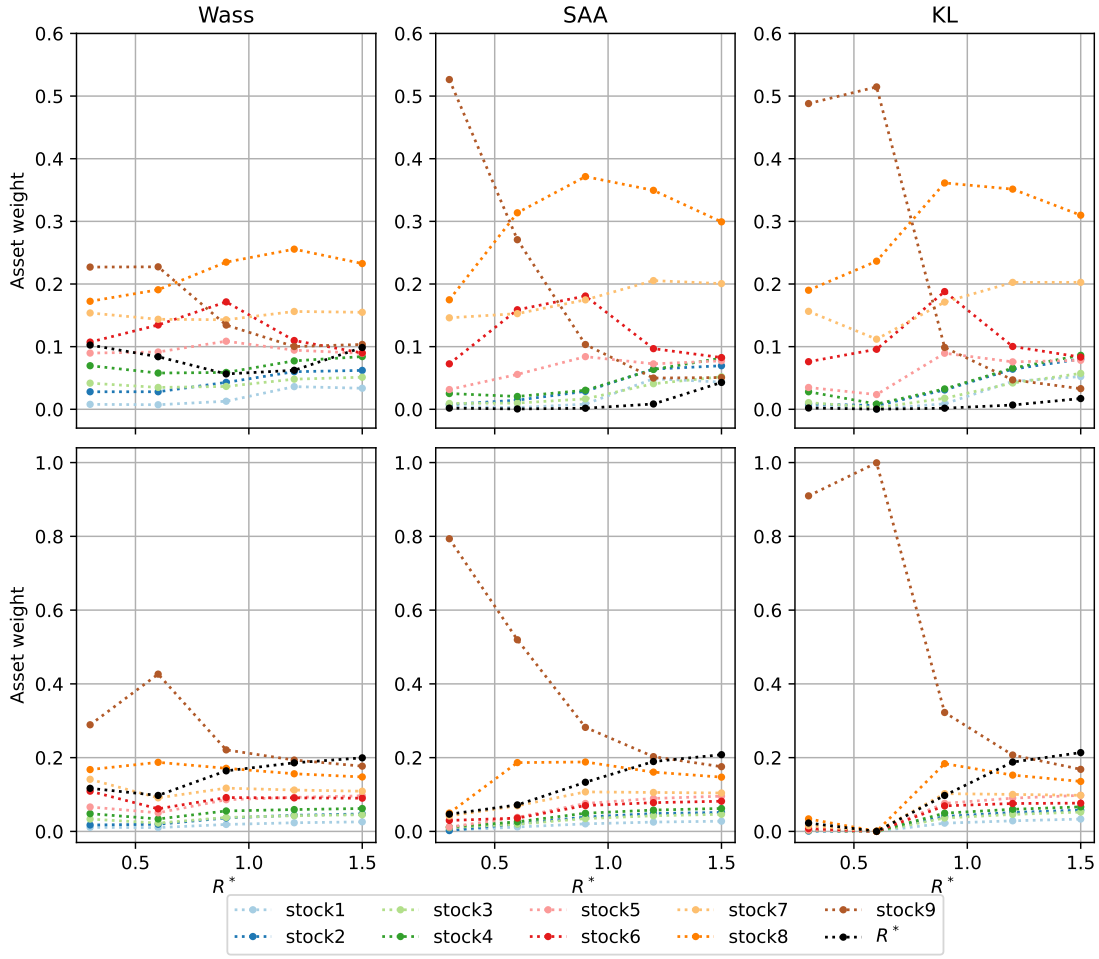



Figure 4.7: Average stock weights of the JDP assets (top) and VGP assets (bottom)

## 4.5 Conclusion

In this chapter, we approached the portfolio selection problem in a novel way. Instead of optimising the returns, we were optimising for the probability that we would achieve a return greater than the risk-free rate by margin $R^+$.

Since the decision variable determined the distribution of the yearly return, we replaced the empirical

distribution with the density approximated with the COS-method and the empirical characteristic function. This way, we avoided writing our optimisation problem as an integer program while still keeping the approach data-driven.

We found that the SAA and KL DRO were better methods to determine the portfolio weights than Wasserstein when the assets are generated from JDP and VGP. When investigating the portfolio weights, we found that the Wasserstein DRO had similar patterns as SAA and KL, but the weights are dampened because of the penalty term in the dual.

Furthermore, for both JDP and VGP, we found that KL had both higher or equal exceeding probabilities and returns than SAA. This means that, even though the performance of the two methods is similar, KL gives a higher probability of exceeding the $R^*$ risk-free rate by at least $R^+$%.

# Chapter 5

# Conclusions and recommendations

In this research, we show for three different applications how KL DRO and Wasserstein DRO compare to each other and to other (robust) strategies. We were mostly interested in the performance for these two methods and how the two arguments, optimality for KL versus support expansion for Wasserstein, translate in practical applications. In the remainder of this conclusion, we discuss what we have learned from the literature and from the three applications. We close this chapter with further recommendations for research.

## 5.1 Conclusions

From the literature, we have learned that DRO is a general method where a worst-case distribution is used to evaluate the expected loss function. This worst-case distribution is chosen from an ambiguity set, which is defined based on a similarity score between two probability measures. We have focused on the Wasserstein, Kullback-Leibler and also Holistic Robust similarity scores where each of the scores has its own advantages and disadvantages. All methods are tractable through their strong dual formulations, and the expected loss functions under their worst-case distributions do not disappoint if the unknown measure $\mathbb{P}$ is in the ambiguity set.

Specifically, the Wasserstein distance is able to expand the ambiguity set by not only increasing the radius, but also by expanding the support of the reference distribution. A common understanding in the literature is that KL does not share this property. However, depending on the direction of KL, it is possible to put probability mass outside the support of $\hat{\mathbb{P}}_N$. This direction is also proven to be the least conservative under the condition of an exponentially decaying disappointment probability. This statistical guarantee is without the assumption of light-tailed distributions that are used in the analysis for Wasserstein DRO.

The KL method has optimal statistical guarantees for sampling errors, but the sample can also contain noise and contamination. By using the Lévy-Prokhorov metric, the decision-maker can also protect against noise and contamination. Combining Lévy-Prokhorov and KL gives you HR DRO which protects against all three errors.

If we look at the first application, then there is a clear outcome. Wasserstein DRO does not benefit from expanding the support and overestimates the mean frequently. KL DRO does rarely overestimate the mean, even less than the theoretical bound suggested. In this case where we only have sampling error, KL DRO does not overestimate, unlike Wasserstein, and is less conservative than other robust mean estimators.

In the second application, we can see that Wasserstein works well and is more profitable than the SAA strategy. During the four years, KL and HR are less profitable than SAA and Wasserstein but protect better against the up-regulation price shocks. Note that, the fact that KL and HR are less profitable does not indicate that KL and HR estimate the loss too conservatively. Besides the expected loss, the profit also depends on the final nomination $x^*$.

In the last application, we see that KL has a higher exceeding probability than Wasserstein. This was to be expected since the data was synthetic, which means there was only sampling errors and no noise or misspecification. Furthermore, KL DRO had the advantage that the portfolio weights were not penalised proportional to the highest weight, which was the case for Wasserstein. The Wasserstein weights showed

a similar weight pattern but were dampened as a consequence of the regularisation term proportional to the maximum of the weights. This property is undesirable in the context of portfolio selection.

In the case where we only have sampling errors, as seen in the mean estimation and portfolio selection application, we can see that KL gave better results than Wasserstein and other methods. In the case of wind power nomination, we can see that the KL DRO is less profitable than SAA and Wasserstein in normal operating conditions. For middle robust configuration, this was less the case for HR DRO. This suggests that the data not only contains sampling error but also has added noise. This shows that Wasserstein can be valuable if there is a reason to believe that the data contains noise, however other methods like HR DRO exist that also take this noise into account. Nevertheless, KL DRO has demonstrated that it protects against black swans efficiently.

## 5.2 Future research

In each chapter, we have recommended several directions for future research for that application specifically. Here, we want to conclude with some more general recommendations.

First of all, for any application, it is recommended to test whether the loss function or scenario random variable follows a heavy-tailed distribution. For example, we can test if a variable is regularly varying using a Hill estimator. In this research, we did not test for heavy-tailed distributions in the mean estimation and portfolio selection applications, as the data we created came from heavy-tailed distributions. The data of the wind power nomination application had outliers, but we did not test for heavy tails ourselves. However, in Aolaritei et al. [5], a power-law behaviour was argued, albeit not with conventional methods as described in Nair et al. [14] and Embrechts et al. [15].

Furthermore, in the second and third applications, we chose to use a different reference distribution than the empirical distribution even though this measure was used to prove that KL DRO is optimal. In future research, it would be interesting to analyse the difference in results if we had used the empirical distribution (when possible).

Lastly, all our applications have used continuous probability distributions. This causes a discrepancy between the nature of the underlying distribution and the empirical distribution, which is always discrete. An interesting application would be an application with a discrete heavy-tailed distribution, for example, in network science or epidemiology [25].

# Bibliography


[1] N. N. Taleb, *The Black Swan*. Random House, 2007.

[2] J. E. Smith and R. L. Winkler, "The optimizer's curse: Skepticism and postdecision surprise in decision analysis," *Management Science*, vol. 52, no. 3, pp. 311–322, Mar. 2006, issn: 0025-1909, 1526-5501. doi: 10.1287/mnsc.1050.0451.

[3] B. P. G. Van Parys, P. M. Esfahani, and D. Kuhn, "From data to decisions: Distributionally robust optimization is optimal," Dec. 2019. doi: 10.48550/arXiv.1704.04118.

[4] H. Rahimian and S. Mehrotra, "Distributionally robust optimization: A review," *Open Journal of Mathematical Optimization*, vol. 3, pp. 1–85, 2022, issn: 2777-5860. doi: 10.5802/ojmo.15.

[5] L. Aolaritei, B. Bangoura, S. Bolognani, N. Lanzetti, and F. Dörfler, "Hedging against black swans in day-ahead energy markets," Oct. 2025. doi: 10.48550/arXiv.2510.14328.

[6] B. P. G. Van Parys and B. Zwart, "Robust mean estimation for optimization: The impact of heavy tails," Mar. 2025. doi: 10.48550/arXiv.2503.21421.

[7] H. E. Scarf, K. Arrow, and S. Karlin, "A min-max solution of an inventory problem," Rand Corporation Santa Monica, Tech. Rep., 1957.

[8] P. M. Esfahani and D. Kuhn, "Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations," Jun. 2017. doi: 10.48550/arXiv.1505.05116.

[9] J. Blanchet, K. Murthy, and V. A. Nguyen, "Statistical analysis of Wasserstein distributionally robust estimators," Aug. 2021. doi: 10.48550/arXiv.2108.02120.

[10] S. Kullback and R. A. Leibler, "On information and sufficiency," *Annals of Mathematical Statistics*, vol. 22, pp. 79–86, 1951.

[11] X. A. Sun and A. J. Conejo, *Robust Optimization in Electric Energy Systems.* (International Series in Operations Research & Management Science). Cham: Springer International Publishing, 2021, vol. 313, isbn: 978-3-030-85127-9. doi: 10.1007/978-3-030-85128-6.

[12] A. Bennouna, B. van Parys, and R. Lucas, "Holistic robust data-driven decisions," Feb. 2025. doi: 10.48550/arXiv.2207.09560.

[13] N. N. Taleb, "Statistical consequences of fat tails: Real world preasymptotics, epistemology, and applications," 2025. doi: 10.48550/arXiv.2001.10488.

[14] J. Nair, A. Wierman, and B. Zwart, *The Fundamentals of Heavy Tails: Properties, Emergence, and Estimation*, 1st ed. Cambridge University Press, 2022, isbn: 978-1-009-05373-0. doi: 10.1017/9781009053730.

[15] P. Embrechts, C. Klüppelberg, and T. Mikosch, *Modelling Extremal Events*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1997, isbn: 978-3-642-08242-9. doi: 10.1007/978-3-642-33483-2.

[16] C. Villani, "The Wasserstein distances," in *Optimal Transport* (Grundlehren der mathematischen Wissenschaften), Grundlehren der mathematischen Wissenschaften. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, vol. 338, pp. 93–111, isbn: 978-3-540-71049-3. doi: 10.1007/978-3-540-71050-9_6.

[17] M. Abramowitz and I. Stegun, *Handbook of Mathematical Functions*. United States: United States National Bureau of Standards, Jun. 1964, isbn: 0-486-61272-4.

[18] Z. Liu, B. P. G. V. Parys, and H. Lam, "Smoothed $f$-divergence distributionally robust optimization," Oct. 2023. doi: 10.48550/arXiv.2306.14041.

[19] C. W. Oosterlee and L. A. Grzelak, *Mathematical Modeling and Computation in Finance: With Exercises and Python and MATLAB Computer Codes*. WORLD SCIENTIFIC (EUROPE), Nov. 2019, isbn: 978-1-78634-794-7.

[20] S. Kou, "A jump-diffusion model for option pricing," *Management Science*, vol. 48, no. 8, pp. 1086–1101, Aug. 2002.

[21] R. C. Merton, "Option pricing when underlying stock returns are discontinuous," *Journal of Financial Economics*, vol. 3, no. 1-2, pp. 125–144, Jan. 1976, issn: 0304405X. doi: 10.1016/0304-405X(76)90022-2.

[22] D. B. Madan, P. P. Carr, and E. C. Chang, "The variance gamma process and option pricing," *Review of Finance*, vol. 2, no. 1, pp. 79–105, Apr. 1998, issn: 1572-3097, 1573-692X. doi: 10.1023/A:1009703431535.

[23] D. B. Madan and E. Seneta, "The variance gamma (v.g.) model for share market returns," *The Journal of Business*, vol. 63, no. 4, p. 511, Jan. 1990, issn: 0021-9398, 1537-5374. doi: 10.1086/296519.

[24] P. I. Davies and N. J. Higham, "Numerically stable generation of correlation matrices and their factors," *BIT Numerical Mathematics*, vol. 40, no. 4, pp. 640–651, Dec. 2000, issn: 0006-3835, 1572-9125. doi: 10.1023/A:1022384216930.

[25] R. Van Der Hofstad, *Random graphs and complex networks*. Cambridge University Press, 2024, vol. 2.

[26] J. Blanchet, L. Chen, and X. Y. Zhou, *Distributionally robust mean-variance portfolio selection with Wasserstein distances*, 2018. doi: 10.48550/arXiv.1802.04885.

# Appendix A

# Derivations from Chapter 2

## A.1 Derivation of Wasserstein dual

Both the derivation of the Wasserstein dual and the KL dual were inspired by Sun and Conejo [11] (section 4.2.2.4 and section 4.2.3.3).

We define the optimal transport program as

$$\min_{\pi} \sum_{i=1}^{N}\sum_{j=1}^{N} c_{ij}\pi_{ij}, \tag{OT 1}$$

$$\text{s.t.} \sum_{j=1}^{N} \pi_{ij} = \hat{p}_i \quad \forall i \in [N], \tag{OT 2}$$

$$\sum_{i=1}^{N} \pi_{ij} = q_j \quad \forall j \in [N], \tag{OT 3}$$

$$\pi_{ij} \geq 0 \quad \forall i, j \in [N]. \tag{OT 4}$$

In the case of estimation, we take $c_{ij} := |\tilde{\zeta}_i - \tilde{\zeta}_j|$ where $\tilde{\zeta}_i$ is the $i$-th observation. We interpret $\mathbb{Q} = \{q_1, ..., q_N\}$ as the adversary probability distribution and $\hat{\mathbb{P}}_N = \{\hat{p}_1, ..., \hat{p}_N\}$ as the empirical probability distribution.

The Wasserstein mean estimator is defined as

$$\hat{e}^W_{N,r} = \begin{cases} \min\limits_{\mathbb{Q}\in\mathcal{P}} & \sum_{j=1}^{N} q_j\tilde{\zeta}_j, \\ \text{s.t.} & \mathrm{W}_1(\mathbb{Q}, \hat{\mathbb{P}}) \leq \varepsilon. \end{cases} \tag{A.1}$$

Here, $\mathcal{P}$ is the collection of probability measures with $N$ points with positive probability mass.

By taking the Lagrange dual, we find

$$\hat{e}^W_{N,\varepsilon} = \max_{\alpha\geq 0} \min_{\mathbb{Q}} \left\{ \sum_{j=1}^{N} q_j\tilde{\zeta}_j - \alpha\varepsilon + \alpha\,\mathrm{W}_1(\mathbb{Q}, \hat{\mathbb{P}}) \right\} \tag{A.2}$$

$$= \max_{\alpha\geq 0} \left\{ -\alpha\varepsilon + \min_{\mathbb{Q}} \left\{ \alpha\,\mathrm{W}_1(\mathbb{Q}, \hat{\mathbb{P}}) + \sum_{j=1}^{N} q_j\tilde{\zeta}_j \right\} \right\} \tag{A.3}$$

$$= \max_{\alpha\geq 0} \left\{ -\alpha\varepsilon + \min_{\mathbb{Q},\,\pi\in\Pi} \left\{ \sum_{i=1}^{N}\sum_{j=1}^{N} \alpha|\tilde{\zeta}_i - \tilde{\zeta}_j|\pi_{ij} + \sum_{j=1}^{N} q_j\tilde{\zeta}_j \right\} \right\} \tag{A.4}$$

where $\Pi = \{\pi : \text{(OT 2)}, \text{(OT 3)}, \text{(OT 4)}\}$. Now by (OT 3), we substitute $q_j$ to get

$$= \max_{\alpha\geq 0} \left\{ -\alpha\varepsilon + \min_{\pi\in\Pi^*} \left\{ \sum_{i=1}^{N}\sum_{j=1}^{N} (\alpha|\tilde{\zeta}_i - \tilde{\zeta}_j| + \tilde{\zeta}_j)\pi_{ij} \right\} \right\} \tag{A.5}$$

where $\Pi^* = \{\pi : (\text{OT } 2), (\text{OT } 4)\}$. Now we focus on the inner minimum and again take the Lagrange dual

$$g(\alpha) = \begin{cases} \min\limits_{\pi \geq 0} & \sum\limits_{i=1}^{N}\sum\limits_{j=1}^{N}(\alpha|\tilde{\zeta}_i - \tilde{\zeta}_j| + \tilde{\zeta}_j)\pi_{ij}, \\ \text{s.t.} & \sum\limits_{j=1}^{N}\pi_{ij} = \hat{p}_i \quad \forall i \in [N] \end{cases} \tag{A.6}$$

$$= \max_{\mu \in \mathbb{R}^N} \min_{\pi \geq 0} \left\{ \sum_{i=1}^{N}\sum_{j=1}^{N}(\alpha|\tilde{\zeta}_i - \tilde{\zeta}_j| + \tilde{\zeta}_j)\pi_{ij} + \sum_{i=1}^{N}\left(\hat{p}_i - \sum_{j=1}^{N}\pi_{ij}\right)\mu_i \right\} \tag{A.7}$$

$$= \max_{\mu \in \mathbb{R}^N} \left\{ \sum_{i=1}^{N}\hat{p}_i\mu_i + \min_{\pi \geq 0}\sum_{i=1}^{N}\sum_{j=1}^{N}(\alpha|\tilde{\zeta}_i - \tilde{\zeta}_j| + \tilde{\zeta}_j - \mu_i)\pi_{ij} \right\}. \tag{A.8}$$

In order to maximise this w.r.t. $\boldsymbol{\mu}$, we need the second term to be nonnegative. This can be achieved by requiring $\alpha|\tilde{\zeta}_i - \tilde{\zeta}_j| + \tilde{\zeta}_j - \mu_i$ to be nonnegative. This will result in the second term being equal to 0 if we want to minimise that term w.r.t. $\pi$.

$$= \begin{cases} \max\limits_{\mu \in \mathbb{R}^N} & \sum_{i=1}^{N}\hat{p}_i\mu_i, \\ \text{s.t.} & \alpha|\tilde{\zeta}_i - \tilde{\zeta}_j| + \tilde{\zeta}_j - \mu_i \geq 0 \quad \forall i,j \in [N] \end{cases} \tag{A.9}$$

$$= \begin{cases} \max\limits_{\mu \in \mathbb{R}^N} & \sum_{i=1}^{N}\hat{p}_i\mu_i, \\ \text{s.t.} & \alpha|\tilde{\zeta}_i - \tilde{\zeta}_j| + \tilde{\zeta}_j \geq \mu_i \quad \forall i,j \in [N] \end{cases} \tag{A.10}$$

$$= \begin{cases} \max\limits_{\mu \in \mathbb{R}^N} & \sum_{i=1}^{N}\hat{p}_i\mu_i, \\ \text{s.t.} & \min\limits_{j \in [N]} \alpha|\tilde{\zeta}_i - \tilde{\zeta}_j| + \tilde{\zeta}_j \geq \mu_i \quad \forall i \in [N] \end{cases} \tag{A.11}$$

$$= \sum_{i=1}^{N}\hat{p}_i \min_{j \in [N]}\{\alpha|\tilde{\zeta}_i - \tilde{\zeta}_j| + \tilde{\zeta}_j\} \tag{A.12}$$

Then combine this with (2.15) to get the estimator

$$\hat{e}^W_{N,\varepsilon} = \max_{\alpha \geq 0}\left\{ -\alpha\varepsilon + \sum_{i=1}^{N}\hat{p}_i \min_{j \in [N]}\{\tilde{\zeta}_j + \alpha|\tilde{\zeta}_i - \tilde{\zeta}_j|\} \right\}. \tag{A.13}$$

## A.2 Derivation closed form Wasserstein estimator

We assume $r < \sum_{i=1}^{N}\hat{p}_i\tilde{\zeta}_i$. Write the Wasserstein estimator from (A.13) as

$$\hat{e}^W_{N,\varepsilon} = \max_{\alpha \geq 0}\left\{ -\alpha\varepsilon + \sum_{i=1}^{N}\hat{p}_i m_i \right\} \tag{A.14}$$

with $m_i = \min_j\{\alpha|\tilde{\zeta}_i - \tilde{\zeta}_j| + \tilde{\zeta}_j\}$ for all $i \in [N]$. Now take $\Delta_{ij} = \tilde{\zeta}_j - \tilde{\zeta}_i$ to get

$$m_i = \min_{\Delta_{ij}}\{\alpha|\Delta_{ij}| + \Delta_{ij} + \tilde{\zeta}_i\} \tag{A.15}$$

$$= \min_{\Delta_{ij}}\{\alpha|\Delta_{ij}| + \Delta_{ij}\} + \tilde{\zeta}_i \tag{A.16}$$

$$= \alpha|\Delta_i^*| + \Delta_i^* + \tilde{\zeta}_i, \tag{A.17}$$

with $\Delta_i^*$ the optimal $\Delta_{ij}$.

In order to find $\Delta_i^*$, we analyse the function $f(x) = \alpha|x| + x$. Note that $f$ is piecewise affine in $x$ with a breakpoint at 0 and its behaviour depends on three cases $\alpha = 1$, $\alpha < 1$ and $\alpha > 1$.

**Case** $\alpha = 1$: we can rewrite $f(x) = 2(x)_+ = 2\max\{0, x\}$ with a minimum of 0 attained for all $x \leq 0$.
**Case** $\alpha < 1$: we can rewrite the function as $f(x) = (1+\alpha)x\,\mathbb{1}\{x > 0\} + (1-\alpha)x\,\mathbb{1}\{x \leq 0\}$. Now suppose $A \subseteq \mathbb{R}$ is bounded, then by monotonicity of $f$,

$$\inf_{z \in A} f(z) = f(\inf A). \tag{A.18}$$

**Case** $\alpha > 1$: using $f(x) = (1+\alpha)x\,\mathbb{1}\{x > 0\} - (1-\alpha)x\,\mathbb{1}\{x \leq 0\}$, now $f$ attains a minimum of 0 at $x = 0$.

With these insights, we know that $\Delta_i^* = 0$ when $\alpha \geq 1$. We know $\Delta_i^*$ attains 0 since $\Delta_i^* = \Delta_{ii} = \tilde{\zeta}_i - \tilde{\zeta}_i = 0$. For $\alpha < 1$, we have that the set $\{\tilde{\zeta}_1 - \tilde{\zeta}_i, ..., \tilde{\zeta}_N - \tilde{\zeta}_i\}$ is bounded. If we fix $\tilde{\zeta}_i$, we can even change this set by the Wasserstein Relaxation where we consider the entire support of $\zeta$, $[0, \infty)$, to get the set $[0 - \tilde{\zeta}_i, \infty)$. Therefore we have $\Delta_i^* = \min[-\tilde{\zeta}_i, \infty) = -\tilde{\zeta}_i$.

We can now rewrite the function maximised in (A.13). We write

$$g(\alpha) = \alpha\varepsilon - \sum_{i=1}^N \hat{p}_i m_i, \tag{A.19}$$

such that $\hat{e}_{N,\varepsilon}^W = \max_{\alpha \geq 0} -g(\alpha) = -\min_{\alpha \geq 0} g(\alpha)$. Then by (A.17),

$$\begin{aligned} g(\alpha) &= \alpha\varepsilon - \sum_{i=1}^N \hat{p}_i \tilde{\zeta}_i - \sum_{i=1}^N \hat{p}_i(\alpha|\Delta_i^*| + \Delta_i^*) && \text{(A.20)} \\ &= \alpha\varepsilon - \sum_{i=1}^N \hat{p}_i \tilde{\zeta}_i - \mathbb{1}\{\alpha < 1\}\left(\sum_{i=1}^N \hat{p}_i \alpha|-\tilde{\zeta}_i| - \sum_{i=1}^N \hat{p}_i \tilde{\zeta}_i\right) && \text{(A.21)} \\ &= \alpha\varepsilon - \sum_{i=1}^N \hat{p}_i \tilde{\zeta}_i - \mathbb{1}\{\alpha < 1\}\left((\alpha-1)\sum_{i=1}^N \hat{p}_i \tilde{\zeta}_i\right). && \text{(A.22)} \end{aligned}$$

Now we want to minimise $g(\alpha)$. We split up in two cases $\alpha \in [0, 1)$ and $\alpha \geq 1$ and analyse $\inf g(\alpha)$ in both cases.

**Case** $\alpha \in [0, 1)$: now we have

$$g(\alpha) = \alpha\varepsilon - \alpha \sum_{i=1}^N \hat{p}_i \tilde{\zeta}_i, \tag{A.23}$$

which is affine in $\alpha$. The slope of $g(\alpha)$ is negative if $\varepsilon < \sum_{i=1}^N \hat{p}_i \tilde{\zeta}_i$ and we have a positive slope otherwise. By assumption, we have $\varepsilon < \sum_{i=1}^N \hat{p}_i \tilde{\zeta}_i$ and therefore a negative slope. This means that $\inf_{\alpha \in [0,1)} g(\alpha)$ is approached as $\alpha \uparrow 1$.

**Case** $\alpha \geq 1$: now

$$g(\alpha) = \alpha\varepsilon - \sum_{i=1}^N \hat{p}_i \tilde{\zeta}_i \tag{A.24}$$

and $\inf_{\alpha \geq 1} g(\alpha)$ is attained at $\alpha = 1$, as $\varepsilon > 0$.

If we combine the two cases, we have a global optimiser $\alpha^* = 1$. This means $\hat{e}_{N,\varepsilon}^W = -g(1) = \sum_{i=1}^N \hat{p}_i \tilde{\zeta}_i - \varepsilon$.

## A.3 Derivation KL dual

The KL DRO estimator is defined as

$$\hat{e}^{KL}_{N,\varepsilon} = \begin{cases} \min\limits_{\boldsymbol{q}\in[0,1]^N} & \sum_{i=1}^N q_i\tilde{\zeta}_i, \\ \text{s.t.} & \sum_{i=1}^N \log\left(\frac{\hat{p}_i}{q_i}\right)\hat{p}_i \le \varepsilon, \\ & \sum_{i=1}^N q_i = 1. \end{cases} \tag{A.25}$$

Then by taking the Lagrange dual, we find

$$\hat{e}^{KL}_{N,\varepsilon} = \max_{\alpha\ge 0,\nu\in\mathbb{R}} \min_{\boldsymbol{q}\in[0,1]^N} \left\{ \sum_{i=1}^N q_i\tilde{\zeta}_i + \alpha\left(\sum_{i=1}^N \log\left(\frac{\hat{p}_i}{q_i}\right)\hat{p}_i - \varepsilon\right) + \nu\left(\sum_{i=1}^N q_i - 1\right)\right\} \tag{A.26}$$

$$= \max_{\alpha\ge 0,\nu\in\mathbb{R}} \left\{ -\alpha\varepsilon - \nu + \min_{\boldsymbol{q}\in[0,1]^N} \sum_{i=1}^N q_i(\tilde{\zeta}_i + \nu) + \alpha\hat{p}_i(\log\hat{p}_i - \log q_i)\right\} \tag{A.27}$$

Since there are no restrictions on each element of $\boldsymbol{q}$ any more, it suffices to minimise $g(q_i) = q_i(\tilde{\zeta}_i + \nu) + \alpha\hat{p}_i(\log\hat{p}_i - \log q_i)$ for each $i \in [N]$. Let $i \in [N]$ and define $q = q_i$, $\hat{p} = \hat{p}_i$, $c = \tilde{\zeta}_i + \nu$. Then

$$g'(q) = c - \alpha\frac{\hat{p}}{q} = 0, \tag{A.28}$$

$$q^* = \frac{\alpha\hat{p}}{c}. \tag{A.29}$$

This is a minimiser since $g''(q) = \alpha\hat{p}/q^2 > 0$. Plugging this back into $g$ gives

$$g(q_i^*) = \alpha\hat{p}_i + \alpha\hat{p}_i\left(\log\hat{p}_i - \left(\log\hat{p}_i + \log\frac{\alpha}{\tilde{\zeta}_i + \nu}\right)\right). \tag{A.30}$$

Using this result in the Lagrangian, and using $\sum_{i=1}^N \hat{p}_i = 1$, gives

$$\hat{e}^{KL}_{N,\varepsilon} = \max\left\{ -\alpha(\varepsilon - 1) - \nu - \sum_{i=1}^N \alpha\hat{p}_i \log\frac{\alpha}{\tilde{\zeta}_i + \nu} : \alpha \ge 0,\ \nu > -\min_{i\in[N]}\tilde{\zeta}_i\right\} \tag{A.31}$$

$$= -\min\left\{ \alpha(\varepsilon - 1) + \nu + \sum_{i=1}^N \alpha\hat{p}_i \log\frac{\alpha}{\tilde{\zeta}_i + \nu} : \alpha \ge 0,\ \nu > -\min_{i\in[N]}\tilde{\zeta}_i\right\} \tag{A.32}$$

Finally, note that the program in (A.25) is convex since the KL divergence is convex and the remaining statement is affine. Since choosing $q_i = \hat{p}_i$ for all $i \in [N]$ is feasible and $\mathrm{KL}(\hat{\mathbb{P}}_N, \mathbb{Q}) - \varepsilon = 0 - \varepsilon < 0$, we satisfy Slater's condition under the assumption $\varepsilon > 0$. Therefore, we have strong duality.

# Appendix B

# Derivations from Chapter 4

## B.1 Derivation of Wasserstein dual

We continue from (A.13) where we substitute the reward function of the portfolio problem $r(x,\xi) = \mathbb{1}\{x^\top \xi \geq t\}$

$$\max_{\lambda \geq 0} \left\{ -\lambda\varepsilon + \sum_{i=1}^{N} \hat{p}_i \min_{j\in[N]} \left\{ \mathbb{1}\{x^\top \tilde{\xi}_j \geq t\} + \lambda \|\tilde{\xi}_j - \tilde{\xi}_i\|_1 \right\} \right\} \tag{B.1}$$

Just like the closed form calculations, fix $i \in [N]$ and take

$$\begin{aligned} m_i &= \min_j \mathbb{1}\{x^\top \tilde{\xi}_j \geq t\} + \lambda \|\tilde{\xi}_j - \tilde{\xi}_i\|_1, && \text{(B.2)} \\ &= \min_{\Delta_{ij}} \mathbb{1}\{x^\top (\tilde{\xi}_i + \Delta_{ij}) \geq t\} + \lambda \|\Delta_{ij}\|_1, && \text{(B.3)} \\ &= \min_{\Delta_{ij}} r_j - r_i + r_i + \lambda \|\Delta_{ij}\|_1, && \text{(B.4)} \end{aligned}$$

where $r_i = \mathbb{1}\{x^\top \tilde{\xi}_i \geq t\} = \mathbb{1}\{x^\top (\tilde{\xi}_j - \Delta_{ij}) \geq t\}$. Now call $\Delta r_{ij} = r_j - r_i$ as the difference between the indicators. Now assume that $\Delta'_j$ is a candidate for the minimiser of (B.4). Then we have to condition on $\Delta r_{ij} \in \{0, 1, -1\}$.

**Case 1** $\Delta r_{ij} = 0$**:** Then $r_i = r_j$. Now we have to find a $\Delta_{ij}$ that also fits this condition has $\|\Delta_{ij}\|_1 < \|\Delta'_j\|_1$. Clearly $\Delta_{ij} = 0$ fits as $\tilde{\xi}_i = \tilde{\xi}_j$ satisfies $r_i = r_j$, and norms are always nonnegative. Therefore $\Delta_i^* = 0$.

**Case 2** $\Delta r_{ij} = 1$**:** Then $r_i = 0$ and $r_j = 1$. In that case, $m_i \geq 1$ and any choice of $\Delta_{ij}$ will lead to a greater value than the candidate of Case 1. Therefore, choose $j^* = i$ such that we have minimizer $\Delta_i^* = 0$.

**Case 3** $\Delta r_{ij} = -1$**:** Then $r_i = 1$ and $r_j = 0$. This is the important case as we now cannot simply choose a $j$ such that $\Delta_{ij} = 0$, since there might exist a $j$ such that $r_j = 0$ and $\lambda\|\Delta_{ij}\|_1 < 1$. So in this case, we have $m_i = \min\left\{1, \lambda \min_j\{\|\Delta_{ij}\|_1 : x^\top(\tilde{\xi}_i + \Delta_{ij}) < t\}\right\}$.

We use the Wasserstein relaxation to approach this problem with a (continuous) linear program where $\tilde{\xi}_i$ stays fixed

$$\begin{aligned} \lambda \inf_{\Delta} \quad & \|\Delta\|_1 \\ \text{s.t.} \quad & x^\top(\tilde{\xi}_i + \Delta) < t. \end{aligned} \tag{B.5}$$

The smallest value of $\|\Delta\|_1$ in the feasible region is on the boundary $x^\top(\tilde{\xi}_i + \Delta) = t$. By solving LP $\min_\Delta\{\sum_k \Delta_k : \sum_k x_k \Delta_k = t - x^\top \tilde{\xi}_i\}$, we want $\Delta_k = 0$ for all $k$ except where $x_{k^*} = \max_l x_l$ since this has the steepest slope to $t - x^\top \tilde{\xi}_i$. Then by the boundary, we have $\Delta_{k^*} = (t - x^\top \tilde{\xi}_i)/x_{k^*}$. Then the $\|\Delta\|_1 = (x^\top \tilde{\xi}_i - t)/\|x\|_\infty$ since we assumed $r_i = 1$.

Combining these cases, we have

$$m_i = \begin{cases} \min\left\{1, \lambda \frac{x^\top \tilde{\xi}_i - t}{\|x\|_\infty}\right\} & \text{if } r_i = 1 \\ 0 & \text{if } r_i = 0. \end{cases} \tag{B.6}$$

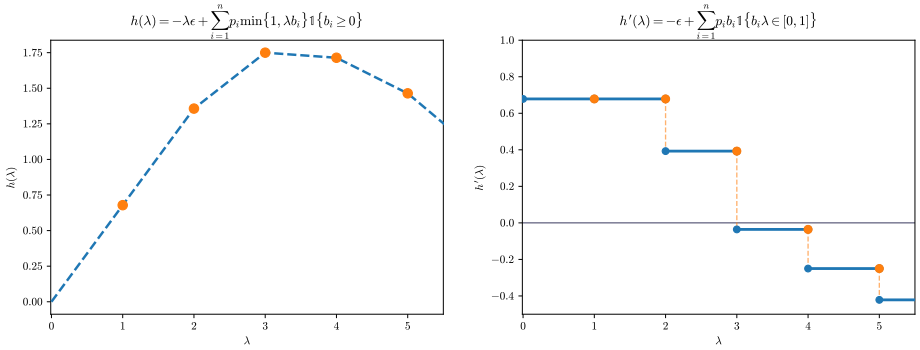



Figure B.1: Example of how $h(\lambda)$ could look like. In this figure, $b_i$ are the breakpoints for the piecewise affine function.

We can substitute this back into the original dual to get

$$\max_{\lambda \geq 0} \left\{ -\lambda\varepsilon + \sum_{i=1}^{N} \hat{p}_i \min\left\{1, \lambda \frac{x^\top \tilde{\xi}_i - t}{\|x\|_\infty}\right\} \mathbb{1}\{x^\top \tilde{\xi}_i \geq t\} \right\} = \max_{\lambda \geq 0} h(\lambda). \tag{B.7}$$

This function $h$ has slope

$$h'(\lambda) = -\varepsilon + \sum_{i=1}^{N} \hat{p}_i \frac{x^\top \tilde{\xi}_i - t}{\|x\|_\infty} \mathbb{1}\left\{0 \leq \lambda \frac{x^\top \tilde{\xi}_i - t}{\|x\|_\infty} \leq 1\right\}. \tag{B.8}$$

Now we assume that $h'(0) > 0$. In that case, all indicator terms are 1 and the full sum is greater than $\varepsilon$. As $\lambda$ increases, more summands in the derivative will be 0 and the derivative will decrease. An example of how $h(\lambda)$ could look like is shown in Figure B.1. We can see that the function starts linear as $x \mapsto \min\{1, x\}$ does not have any effect yet. Once terms get substituted by 1, the derivative will start to decrease and the $-\lambda\varepsilon$ term will take over. Then there will be a point $i \in [N]$ such that $h'\left(\frac{\|x\|_\infty}{x^\top \tilde{\xi}_i - t}\right) \geq 0$ and $h'\left(\frac{\|x\|_\infty}{x^\top \tilde{\xi}_{i+1} - t}\right) < 0$. Take $i^* = i$ and optimiser $\lambda^* = \frac{\|x\|_\infty}{x^\top \tilde{\xi}_i - t}$. Note that $\lambda^*$ indeed optimises since it is the largest value with $\lambda^* \frac{x^\top \tilde{\xi}_{i^*} - t}{\|x\|_\infty} = 1$ and still having positive derivative.

**Statement of AI use:** For this derivation, mostly the strategy of Blanchet et al. [26] was used. However, since they use $W_2$, the proof went in a different direction. I used Claude Sonnet 4.6 to solve the relaxed minimisation problem for $\Delta$ and the maximisation of $h(\lambda)$. Claude recognised that the minimisation problem is equivalent to finding the distance to the closest point on a hyperplane and from there I could continue. For the maximisation of $h$, Claude recognised that the terms in the sum of the derivative were set to 0 and that $h'$ should approach 0.

## B.2 Derivation of KL dual

The primal program is defined as

$$\begin{aligned} &\max_{x} \min_{\hat{\mathbb{P}}_N \ll \mathbb{Q}} \mathbb{Q}\left(x^\top \xi \geq t\right) \\ &\text{s.t.} \qquad D_{KL}(\hat{\mathbb{P}}_N, \mathbb{Q}) \leq \varepsilon, \end{aligned} \tag{B.9}$$

where we use $t = R^*(1 + R^+)$ to save notation. Note that we can rewrite the tail probability as $\mathbb{Q}(x^\top \xi \geq t) = \mathbb{E}_{\mathbb{Q}}[r(x, \xi)]$ where the reward $r(x, \xi) = \mathbb{1}\{x^\top \xi \geq t\}$. We can also rewrite the KL-divergence in terms of the densities $q$ and $p$ if we assume absolute continuity of the measures $\mathbb{Q}$ and $\hat{\mathbb{P}}_N$

$$D_{KL}(\hat{\mathbb{P}}_N, \mathbb{Q}) = \int \log\left(\frac{\hat{p}(z)}{q(z)}\right) \hat{p}(z)\, dz, \tag{B.10}$$

This gives us the Lagrangian

$$\max_{\lambda>0,\mu}\min_{q}\left[-\lambda\varepsilon+\lambda\left(\int\hat{p}(z)\log\hat{p}(z)\,\mathrm{d}z-\int\hat{p}(z)\log q(z)\,\mathrm{d}z\right)+\int r(x,z)q(z)\,\mathrm{d}z+\mu\left(\int q(z)\,\mathrm{d}z-1\right)\right], \tag{B.11}$$

where the last term is added to enforce $\int q(z)\,\mathrm{d}z=1$. Rearranging terms gives us

$$\max_{\lambda>0,\mu}-\lambda\varepsilon-\mu+\lambda\int\hat{p}(z)\log\hat{p}(z)\,\mathrm{d}z+\min_{q}\int\left[r(x,z)+\mu-\lambda\frac{\hat{p}(z)}{q(z)}\log q(z)\right]q(z)\,\mathrm{d}z \tag{B.12}$$

Since the inner integrand depending on $q$ is convex, we will optimise this part pointwise. We assume $\mu>0$ and $q(z)>0$ for all $z\in\Xi$. We omit the inputs of the functions to save notation and write the inner integrand as

$$h(q)=\left(r+\mu-\lambda\frac{\hat{p}}{q}\log(q)\right)q=(r+\mu)q-\lambda\hat{p}\log q, \tag{B.13}$$

which will differentiate and solve for $q$. This results in

$$h'(q)=r+\mu-\lambda\frac{\hat{p}}{q}=0, \tag{B.14}$$

$$-\lambda\hat{p}=-q(r+\mu), \tag{B.15}$$

$$q^*=\frac{\lambda\hat{p}}{r+\mu}. \tag{B.16}$$

Plugging in the optimal solution gives

$$h(q^*)=\lambda\hat{p}-\hat{p}\lambda\log\left(\frac{\lambda\hat{p}}{r+\mu}\right). \tag{B.17}$$

If we use $h(q^*)$ in the Lagrangian, we can cancel out some terms such that

$$\begin{aligned}&\max_{\lambda\geq0,\mu>0}-\lambda\varepsilon-\mu+\int\lambda\hat{p}(z)\log\hat{p}(z)\,\mathrm{d}z\\&\qquad+\int\left[\lambda\hat{p}(z)-\lambda\hat{p}(z)\log(\hat{p}(z))-\lambda\hat{p}(z)\log\left(\frac{\lambda}{\mu+r(x,z)}\right)\right]\mathrm{d}z\end{aligned} \tag{B.18}$$

$$=\max_{\lambda\geq0,\mu>0}-\lambda\varepsilon-\mu+\lambda-\lambda\int\hat{p}(z)\log\left(\frac{\lambda}{\mu+r(x,z)}\right)\mathrm{d}z. \tag{B.19}$$

Now we simplify and take the minus out to change the dual to a minimisation problem. This leaves us with the generic dual problem

$$-\min_{\lambda\geq0,\mu>0}\lambda(\varepsilon-1)+\mu+\lambda\int\hat{p}(z)\log\left(\frac{\lambda}{\mu+r(x,z)}\right)\mathrm{d}z. \tag{B.20}$$

Now we can simplify further by substituting our reward function, which gives us our final dual

$$-\min_{\lambda\geq0,\mu>0}\lambda(\varepsilon-1)+\mu+\lambda\,\mathbb{E}_{\hat{\mathbb{P}}_N}\left[\log\left(\frac{\lambda}{\mu+\mathbb{1}\{x^Tz\geq t\}}\right)\right] \tag{B.21}$$

$$=-\min_{\lambda\geq0,\mu>0}\lambda(\varepsilon-1)+\mu+\lambda\left(\log\left(\frac{\lambda}{\mu+1}\right)\hat{\mathbb{P}}_N(x^T\xi\geq t)+\log\left(\frac{\lambda}{\mu}\right)\hat{\mathbb{P}}_N(x^T\xi\leq t)\right) \tag{B.22}$$

$$=-\min_{\lambda\geq0,\mu>0}\lambda(\varepsilon-1)+\mu+\lambda\left(\log\lambda-\log(\mu+1)(1-\hat{F}(t))-\log(\mu)\,\hat{F}(t)\right), \tag{B.23}$$

which is the same bound as (4.34).

# Appendix C

# Statement of AI use

Besides the mentioned use of AI in the report for Figure 1.1 and parts of the proof in Appendix B, LLMs like ChatGPT Pro and Claude Opus 4.8 were used to check the 'almost-final' report for spelling/grammar mistakes, mathematical correctness, and overall presentation. The author critically assessed the suggested improvements before implementing them.