Distributionally Robust Performance Analysis: Data, Dependence and Extremes

Fei He

Submitted in partial fulfillment of the
requirements for the degree
of Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2018

ABSTRACT

Distributionally Robust Performance Analysis: Data, Dependence and Extremes

Fei He

This dissertation focuses on distributionally robust performance analysis, which is an area of applied probability whose aim is to quantify the impact of model errors. Stochastic models are built to describe phenomena of interest with the intent of gaining insights or making informed decisions. Typically, however, the fidelity of these models (i.e. how closely they describe the underlying reality) may be compromised due to either the lack of information available or tractability considerations. The goal of distributionally robust performance analysis is then to quantify, and potentially mitigate, the impact of errors or model misspecifications. As such, distributionally robust performance analysis affects virtually any area in which stochastic modelling is used for analysis or decision making.

This dissertation studies various aspects of distributionally robust performance analysis. For example, we are concerned with quantifying the impact of model error in tail estimation using extreme value theory. We are also concerned with the impact of the dependence structure in risk analysis when marginal distributions of risk factors are known. In addition, we also are interested in connections recently found to machine learning and other statistical estimators which are based on distributionally robust optimization.

The first problem that we consider consists in studying the impact of model specification in the context of extreme quantiles and tail probabilities. There is a rich statistical theory that allows to extrapolate tail behavior based on limited information. This body of theory is known as extreme value theory and it has been successfully applied to a wide range of settings, including building physical infrastructure to withstand extreme environmental events and also guiding the capital requirements of insurance companies to ensure their financial solvency. Not surprisingly, attempting to extrapolate out into the tail of a

distribution from limited observations requires imposing assumptions which are impossible to verify. The assumptions imposed in extreme value theory imply that a parametric family of models (known as generalized extreme value distributions) can be used to perform tail estimation. Because such assumptions are so difficult (or impossible) to be verified, we use distributionally robust optimization to enhance extreme value statistical analysis. Our approach results in a procedure which can be easily applied in conjunction with standard extreme value analysis and we show that our estimators enjoy correct coverage even in settings in which the assumptions imposed by extreme value theory fail to hold.

In addition to extreme value estimation, which is associated to risk analysis via extreme events, another feature which often plays a role in the risk analysis is the impact of dependence structure among risk factors. In the second chapter we study the question of evaluating the worst-case expected cost involving two sources of uncertainty, each of them with a specific marginal probability distribution. The worst-case expectation is optimized over all joint probability distributions which are consistent with the marginal distributions specified for each source of uncertainty. So, our formulation allows to capture the impact of the dependence structure of the risk factors. This formulation is equivalent to the so-called Monge-Kantorovich problem studied in optimal transport theory, whose theoretical properties have been studied in the literature substantially. However, rates of convergence of computational algorithms for this problem have been studied only recently. We show that if one of the random variables takes finitely many values, a direct Monte Carlo approach allows to evaluate such worst case expectation with $O(n^{-1/2})$ convergence rate as the number of Monte Carlo samples, n, increases to infinity.

Next, we continue our investigation of worst-case expectations in the context of multiple risk factors, not only two of them, assuming that their marginal probability distributions are fixed. This problem does not fit the mold of standard optimal transport (or Monge-Kantorovich) problems. We consider, however, cost functions which are separable in the sense of being a sum of functions which depend on adjacent pairs of risk factors (think of the factors indexed by time). In this setting, we are able to reduce the problem to the study of several separate Monge-Kantorovich problems. Moreover, we explain how we can even include martingale constraints which are often natural to consider in settings such as financial applications.

While in the previous chapters we focused on the impact of tail modeling or dependence, in the later parts of the dissertation we take a broader view by studying decisions which are made based on empirical observations. So, we focus on so-called distributionally robust optimization formulations. We use optimal transport theory to model the degree of distributional uncertainty or model misspecification. Distributionally robust optimization based on optimal transport has been a very active research topic in recent years, our contribution consists in studying how to specify the optimal transport metric in a data-driven way. We explain our procedure in the context of classification, which is of substantial importance in machine learning applications.

Table of Contents

List of Figures

List of Tables

Acknowledgments

First and foremost, I would like to thank my advisor, Professor Jose Blanchet, for his continued generous support. His broad knowledge, great insight and unwavering dedication to research have always encouraged and inspired me to pursue interesting research topics and helped me overcome many challenges and difficulties.

Besides my advisor, I would like to thank the rest of my committee members: Prof. David Yao, Prof. Karl Sigman, Prof. Henry Lam and Prof. Jing Dong, for offering their invaluable insight and expertise to improve my dissertation. I have learned much from Prof. Yao and Prof. Sigman's probability courses. Henry and Jing are also my research collaborators and I really appreciate their guidance and patience during the collaboration.

I am also grateful to many IEOR professors, especially Prof. Garud Iyengar, Prof. Martin Haugh, Prof. Agostino Capponi, Prof. Jay Sethuraman, Prof. Tim Leung and Prof. Marcel Nutz, as well as many doctoral students and friends, especially Zhipeng Liu, Yanan Pei, Anran Li, Yang Kang, Kartheyek Murthy, Ni Ma, Xinshang Wang, Mingxian Zhong, Chaoxu Zhou, Francois Fagan, Octavio Ruiz Lacedelli, Di Xiao, Kevin Guo, Brian Ward, Bin Qi, Yang Zhang, Yangbin Li, Xin Li, Zheng Wang, Jianshu Wu, Linan Yang, Juan Li, Yin Lu, Yunhao Tang, Fan Zhang, Fengpei Li and Raghav Singal.

I would also like to thank Prof. Mark Podolskij, Prof. Rainer Dahlhaus, Prof. Tilmann Gneiting, Yu Hao, Stefan Richter, Brandon Williams, Yuhong Dai, Nopporn Thamrongrat, Cike Peng, Julie Yuan Merten, Shuxin Chen and Claudio Heinrich for their support and help during my study in Germany. I am also much obliged to many friends in China: Chengbin Peng, Hao Zhang, Junhui Zhang, Lu Chen, Huajie Mao, Hongyu Zhu, Wei Lin, Minjie Zhang, Fang Lv, Hongyu Zhu, Bo Zhang, Wei Shi, Xiyuan Chen, Dingsheng Lin and Yu Shi.

Last but not least, I am deeply indebted to my parents and Chris for their unconditional love and support over the years. No words can express my gratitude to them.

Meinen Eltern und meinem Schatz

Chapter 1

Introduction

This dissertation focuses on distributionally robust performance analysis, which is an area of applied probability whose aim is to quantify the impact of model errors. Stochastic models are built to describe phenomena of interest with the intent of gaining insights or making informed decisions. Typically, however, the fidelity of these models (i.e. how closely they describe the underlying reality) may be compromised by either the lack of information available or by tractability considerations. The goal of distributionally robust performance analysis is then to quantify, and potentially mitigate, the impact of errors or model misspecifications. As such, distributionally robust performance analysis affects virtually any area in which stochastic modelling is used for analysis or decision making.

More specifically, in a stochastic model, the performance evaluation can be represented as $\mathbb{E}_P[h(X)]$ for a given probability measure P, a random variable X and a function h. A modeler faces the task of choosing a probability model P which is not only close to the reality but is also tractable. However, this procedure will often suffer from model errors, either due to the lack of data or due to the estimation errors.

A popular approach to address this problem is by considering the distributionally robust bound as the optimal value of the optimization problem

$$\sup_{P\in\mathcal{U}} \mathbb{E}_P[h(X)],$$

over a family of plausible alternative probability models $\mathcal{U}$. A natural way to specify the family $\mathcal{U}$ is by defining an uncertainty neighborhood $\{P : d(P, P_{ref}) \leq \delta\}$, where P_{ref} is the chosen reference model and δ is a tolerance level. Here d is a metric which measures the discrepancy between two probability measures. Popular choices for d are the KL-divergence (Breuer and Csiszar (2013a), H.Lam (2013), Glasserman and Xu (2014a)) and the Wasserstein distances (Esfahani and Kuhn (2015), Wozabal (2012), Blanchet and Murthy (2016)) to quantify the model uncertainty. Despite the fact that KL-divergence is not a true metric, KL-divergence is a popular choice due to its tractability. This approach provides a bound for the performance evaluation regardless of the probability measure used as long as such measures stay within a prescribed tolerance δ of an appropriate reference model.

□ In Chapter 2 we study the distributional robustness in the context of the extreme value theory (EVT). Our focus is closer in spirit to distributionally robust optimizations as in, for instance, Dupuis *et al.* (2000), Hansen and Sargent (2001), Ben-Tal *et al.* (2013), Breuer and Csiszár (2013b). However, in contrast to the literature on robust optimization, the emphasis here is on understanding the implications of distributional uncertainty regions in the context of EVT. As far as we know this is the first paper that studies distributional robustness in the context of EVT. Here, our objective is to provide a robust bound for the estimate of the value at risk of a risk factor X,

$$\mathrm{VaR}_p(X) = F^{\leftarrow}(p) := \inf\{x : P\{X \leq x\} \geq p\}, \text{ for } p \in (0,1).$$

EVT provides reasonable statistical principles which can be used to extrapolate tail distributions and then estimate this extreme quantiles. In particular, we focus on the classical block maxima approach for the extrapolation, that is, we divide the i.i.d. data X_i into several blocks, where each block contains n data points. Then we pick the maximum value M_n from each block. The Fisher-Tippett-Gnedenko theorem ensures that under certain assumptions of the underlying distribution of the X_i, the maximum M_n has some types of limiting distribution P_{GEV}, the so-called generalized extreme value distribution, and produces $P_{GEV}^{-1}(p^n)$ as an estimate for the quantile $\mathrm{VaR}_p(X)$. However, as with any form for extrapolation, extreme value analysis rests on assumptions that are rather difficult (or impossible) to verify. Therefore, it makes sense to provide a mechanism to robustify the inference obtained via EVT. Similarly we formulate the robust estimate through an uncertainty neighborhood of the limiting distribution with radius δ and then give a robust estimate of $\mathrm{VaR}_p(X)$ by

$$\sup\{G^{\leftarrow}(p^n) : d(G, P_{GEV}) \leq \delta\}.$$

Here, we choose d as the Rényi divergence, also called the α-divergence, which includes KL-divergence as a special case for $\alpha = 1$. We show that using KL-divergence to form the uncertainty set around P_{GEV} would include a probability measure whose tail probabilities decay at an unrealistically slow rate and the parameter α gives modeler the freedom to tune the uncertainty set and include distributions with tails are heavier than the reference model but not prohibitively heavy. We give concrete algorithms to calculate this robust estimate and we also provide some practical ways to specify the hyperparameters α and the radius of the uncertainty set δ. We also give some examples where the standard EVT can significantly underestimate the quantiles of interest while our estimator is quite robust and at the same time not too conservative.

In addition to extreme value estimation, which is associated to risk analysis via extreme events, another feature which often plays a role in the risk analysis is the impact of dependence structure among risk factors. Chapter 3 and Chapter 4 are devoted to find the lower or upper bounds among any dependence structure with two sources of uncertainty or multiple sources of uncertainty, that is, measuring the impact of the joint distribution with two or multiple fixed marginals.

□ In Chapter 3 we study a direct Monte-Carlo-based approach for computing lower and upper bounds among any dependence structure for a function of two random vectors whose marginal distributions are assumed to be known.

More precisely, suppose that $X \in \mathbb{R}^d$ follows distribution μ and $Y \in \mathbb{R}^l$ follows distribution ν. We define $\Pi(\mu, \nu)$ to be the set of joint distributions π in $\mathbb{R}^{d \times l}$ such that the marginal of the first d entries coincides with μ and the marginal of the last l entries coincides with ν. In other words, for any probability measure π in $\mathbb{R}^{d \times l}$ (endowed with the Borel σ-field), if we let $\pi_X(A) = \pi\left(A \times \mathbb{R}^l\right)$ for any Borel measurable set $A \in \mathbb{R}^d$, and $\pi_Y(B) = \pi\left(\mathbb{R}^d \times B\right)$ for any Borel measurable set $B \in \mathbb{R}^l$, then $\pi \in \Pi(\mu, \nu)$ if and only if $\pi_X = \mu$ and $\pi_Y = \nu$. We are interested in the quantity (focusing on minimization)

$$V = \min\{\mathbb{E}_\pi [c(X, Y)] : \pi \in \Pi(\mu, \nu)\} \tag{1.1}$$

where $c(\cdot,\cdot) \in \mathbb{R}$ is some cost function. Formulation (1.1) is well-defined as the class $\Pi(\mu,\nu)$ is non-empty, because the product measure $\pi = \mu \times \nu$ belongs to $\Pi(\mu,\nu)$. The worst-case expectation is optimized over all joint probability distributions which are consistent with the marginal distributions specified for each source of uncertainty. So, our formulation allows to capture the impact of the dependence structure of the risk factors. This formulation is equivalent to the so-called Monge-Kantorovich problem studied in optimal transport theory, whose theoretical properties have been studied in the literature substantially (Villani (2003), Villani (2008)).

We focus on the setting where one of the marginals, say Y, has a distribution ν with finite support $\{y_1,...,y_m\} \subset \mathbb{R}^l$ and another, say X, has a multi-dimensional distribution μ that can be continuous. Suppose we can i.i.d. sample $X_i, i = 1,\ldots,n$ from the distribution μ then we approximate V by

$$V_n = \min\{\mathbb{E}_\pi\left[c\left(X,Y\right)\right] : \pi \in \Pi\left(\mu_n,\nu\right)\} \tag{1.2}$$

where μ_n is the empirical distribution of X constructed from the X_i's, i.e.,

$$\mu_n(A) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \in A)$$

for any Borel measurable A.

Our main result shows that the error of our procedure is $O(n^{-1/2})$ where n is the sample size, independent of the dimension d or l. We also identify the limiting distribution in the associated CLT. The closest work to our results, as far as we know, is the recent work of Sommerfeld and Munk (2016), which derives a CLT when both marginal distributions are finitely discrete.

On the other hand, it is difficult to further generalize our procedure to the case when both X and Y are continuous. The study on the rate of convergence in Wasserstein distance of the empirical measure gives ideas that in this general case the convergence rate fail to retain $O(n^{-1/2})$ (Fournier and Guillin (2015)). For instance, suppose both $X, Y \sim U[0,1]^d$, i.e. $\mu = \nu$ are d-dim uniform distributions, and $c(x,y) = \|x - y\|$, the optimal value V corresponds to the Wasserstein distance (of order 1) between X and Y, which is of course 0. It is well-known that sampling X and keeping Y continuous will give, for $d \geq 3$, an expected optimal value of

$$V_n = \min\{\mathbb{E}_\pi\left[c\left(X,Y\right)\right] : \pi \in \Pi\left(\mu_n, \mu\right)\} \qquad \mu_n(\cdot) := \frac{1}{n}\sum_{i=1}^{n} I(X_i \in \cdot)$$

is of order $n^{-1/d}$, i.e., $C_1 n^{-1/d} \leq \mathbb{E}V_n \leq C_2 n^{-1/d}$ for all n for some $C_1, C_2 > 0$ (see e.g.van Handel (2014)).

□ In Chapter 4 we study a discretization approach for computing lower and upper bounds among any dependence structure for a function of multiple random vectors whose marginal distributions are assumed to be known. Given d marginal distributions $\mu_1, \ldots, \mu_d$ on a common compact metric space $\mathscr{X}$, we focus on the lower bound

$$\inf_{\pi \in \Pi(\mu_1, \ldots, \mu_d)} \mathbb{E}_\pi[c(X_1, \ldots, X_d)], \tag{1.3}$$

where $\Pi(\mu_1, \ldots, \mu_d)$ is the set of all joint distributions with marginals $X_1 \sim \mu_1, \ldots, X_d \sim \mu_d$, and c is a cost function. Note that when $d = 2$, the problem (1.3) is the standard optimal transport problem. For $d > 2$, this problem has been studied by Gangbo and Swiech (1998) and G.Carlier *et al.* (2008). Such problems often arise from risk management, where the performance depends on d risk factors, and the marginal distributions of each risk factor is known but the dependence structure is ambiguous.

We approach this problem by first create a partition of the compact space $\mathscr{X}$ with $\mathscr{X} = \sum_{k=1}^{n} A_k$ such that the diameter of every A_k does not exceed δ, with $\delta = O(n^{-1})$. Then we choose a representative $x_k \in A_k$ for each k and form a discrete set $\mathscr{X}_\delta = \{x_k : k = 1, \ldots, n\}$ with an associated quantization map

$$\begin{aligned} T : \mathscr{X} &\rightarrow \mathscr{X}_\delta \\ x &\mapsto \sum_{k=1}^{n} x_k \mathbb{I}(x \in A_k). \end{aligned}$$

In addition, we define the corresponding quantized measures as

$$\mu_{1,\delta}(x_k) = \mu_1(A_k), \; \cdots \; , \mu_{d,\delta}(x_k) = \mu_d(A_k), \text{ for } k = 1, \cdots, n \tag{1.4}$$

Then the discretized approximate version of the problem is as follows:

$$\min_{\pi} \sum_{i_1,\cdots,i_d=1}^{n} c(x_{i_1},\cdots,x_{i_d})\pi(x_{i_1},\cdots,x_{i_d})$$

s.t.

$$\sum_{i_2=1,\cdots,i_d=1}^{n} \pi(x_{i_1},\cdots,x_{i_d}) = \mu_{1,\delta}(x_{i_1}),\ i_1 = 1,\cdots,n,$$

$$\cdots$$

$$\sum_{i_1=1,\cdots,i_{d-1}=1}^{n} \pi(x_{i_1},\cdots,x_{i_d}) = \mu_{d,\delta}(x_{i_d}),\ i_d = 1,\cdots,n,$$

$$\sum_{i_1=1,\cdots,i_d=1}^{n} \pi(x_{i_1},\cdots,x_{i_d}) = 1,\ \pi(x_{i_1},\cdots,x_{i_d}) \geq 0,$$

For $d = 2$, it is an assignment problem, which can be solved by various network algorithms that are much faster than the general LP algorithms. For instance, with the successive shortest path algorithm (see R.K.Ahuja *et al.* (2000) p.320) one can achieve $O(n^2\log(n))$. We will also quantify the error bounds for the difference between the true optimal value and the optimal value of the discretized version. For $d > 2$ we can in general not transform it to assignment problems except when the cost function c is separable, that is, c takes the form of $c(X_1,\cdots,X_d) = \sum_{k=1}^{d-1} c_t(X_t, X_{t+1})$, where $c_t,\ t = 1,\cdots,d-1$ are cost functions depending only on the two adjacent marginals X_t and X_{t+1}. Then the above discretized version can be decomposed into $d-1$ assignment problems and hence can be solved efficiently by using network algorithms. In fact, with this separable cost function, we can apply this discretization approach to the so-called martingale optimal transport problem, which is first studied by Beiglbock *et al.* (2013) and Galichon *et al.* (2014). A general form of the martingale optimal transport problem looks as follows:

$$\inf_{\pi\in\mathscr{M}(\mu_1,\ldots,\mu_d)} \mathbb{E}_\pi[c(X_1,\ldots,X_d)], \tag{1.5}$$

where $\mathscr{M}(\mu_1,\ldots,\mu_d)$ is the set of all martingale measures, i.e. the underlying process $(X_t)_{t=1,\ldots,d}$ satisfies $X_t \sim \mu_t, \mathbb{E}_\pi[X_t|\mathcal{F}_{t-1}] = X_{t-1}$. The martingale optimal transport problem is different from the previous one in that in general there exists no easy way to convert it to the discretized version due to the martingale constraint, but we can show that when the cost function c is separable, then the problem can still be discretized to $d-1$ linear programming problems.

A major application of martingale optimal transport problem is in mathematical finance, where it is important to choose a pricing model when evaluating an exotic option; such a model is characterized by a martingale measure while the marginal distributions are the daily underlying prices. Instead of postulating a model, we use (1.5) to give a model-free lower bound for the price of exotics, whose payoff function c depends on the d-marginal distributions of a certain underlying X, indexed by time $t = 1, \cdots, d$. Similarly, by maximization instead of minimization we also obtain an upper bound for the price. This price range is robust against model errors and it complies with market prices of vanilla options, which are liquid and suitable hedging instruments. We provide some examples of financial derivatives whose model-free price ranges can be obtained by our method.

While in the previous chapters we focused on the impact of tail modeling or dependence, in the later parts of the dissertation we take a broader view by studying decisions which are made based on empirical observations. We focus on so-called distributionally robust optimization formulations. The objective of distributionally robust optimization is to choose a decision β that minimizes the worst-case expected loss $\sup_{P\in\mathcal{U}} \mathbb{E}_P[l(X,\beta)]$, where the worst-case is taken over an uncertainty neighborhood $\mathcal{U}$ of an unknown true distribution P^*. Though the true distribution P^* is unknown, we usually have some information or properties about P^*, such as the empirical measure P_n, so in practice we often form the uncertainty neighborhood around P_n. Distributionally robust optimization has two main advantages: one is to improve the out-of-sample performance of stochastic programmings and the other one is that distributionally robust models are often tractable even though the corresponding stochastic models are NP-hard. A good choice of uncertainty neighborhood $\mathcal{U}$ should be rich enough to include the true distribution with high confidence while at the same time it should be small enough to exclude uninteresting distributions so as to avoid too conservative decisions. Previous works usually use moment constraints (J.Goh and M.Sim (2010), Wieseman *et al.* (2014)) and KL-divergence (Breuer and Csiszar (2013a), H.Lam (2013), Glasserman and Xu (2014a)) to quantify model misspecification and model uncertainty. Despite the fact that KL-divergence is not a true metric, it is a popular choice due to its tractability. However, many of these earlier works also acknowledge the shortcomings of KL-divergence, as the absolute continuity requirement rules out many interesting settings. For instance, all the probability measures in the neighborhood of an empirical measure defined by the KL-divergence are just re-weighting of this empirical measure; the neighborhood fails to include any continuous measures. Recently, people start applying Wasserstein distance to distributionally robust optimization and quantify model misspecification (Wozabal (2012), Esfahani and Kuhn (2015), Blanchet and Murthy (2016)). When the cost function c is a metric, i.e. $c(x, y) = d(x, y)$, then the optimal transport problem actually induces a metric called the Wasserstein distance or the optimal transport metric, which characterizes a distance between the two probability measures μ and ν, and in turn we can use it to define a neighborhood of a measure and apply it to the distributionally robust problems. The uncertainty set contains both continuous and discrete distributions that are close to the measure of interest (e.g. the empirical measure) with respect to the Wasserstein distance, which makes it possible to incorporate many tractable surrogate models and offers better out-of-sample performance. However, distributionally robust models with Wasserstein uncertainty neighborhood are generally harder in computations and they are still attractive topics in research.

□ Chapter 5 uses optimal transport theory to model the degree of distributional uncertainty or model misspecification, and extends the following distributionally robust optimization (DRO) model proposed by Blanchet *et al.* (2016a), where they reveal that the DRO models links to several machine learning algorithms such as regularized logistic regression for classification,

$$\min_{\beta} \max_{P \in \mathcal{U}_\delta(P_n)} \mathbb{E}_P[l(X, Y, \beta)] = \min_{\beta} \left(\mathbb{E}_{P_n}[l(X, Y, \beta)] + \delta \left\| \beta \right\|_p \right), \tag{1.6}$$

where l is some loss function, and $\mathcal{U}_\delta\left(P_n\right) = \{P : D_c(P, P_n) \leq \delta\}$ is a neighborhood of the empirical measure P_n defined by the optimal transport distance

$$D_c\left(P, P_n\right) = \inf_{\pi} \left\{ \mathbb{E}_\pi[c(P, P_n)] : \pi \text{ is a joint distribution of } P \text{ and } P_n \right\}$$

and the optimal transport cost function

$$c((x, y), (x', y')) = \left\| x - x' \right\|_q^2 I(y = y') + \infty \cdot I(y \neq y'),$$

where $p^{-1} + q^{-1} = 1$ for $p \in [1, \infty)$, and $\mathbb{E}_{P_n}[l(X, Y, \beta)] := \frac{1}{n} \sum_{i=1}^{n} l(X_i, Y_i, \beta)$. We can interpret the DRO problem on the left hand side of (1.6) as we choose a decision β for minimization, while the adversarial player selects a model P, a perturbation of the data P_n, from $\mathcal{U}_\delta(P_n)$. This interpretation has applications in adversarial training of neural networks, see e.g. Sinha *et al.* (2017). Note that the *shape* of $\mathcal{U}_\delta(P_n)$ is determined by the cost function $c(\cdot)$ in the definition of the optimal transport discrepancy $D_c(P, P_n)$, but so far it has been taken as a given ℓ_q-norm, but not chosen in a data-driven way; this is the starting point of this project to improve the DRO method.

Our contribution consists in studying how to specify the optimal transport metric in a data-driven way. We would propose a data-driven DRO (DD-DRO) model with the cost function c_Λ defined by a local metric $d_\Lambda(x, x') := \sqrt{(x - x')^T \Lambda(x)(x - x')}$, where the matrix $\Lambda(x)$ is trained by metric learning methods, see, e.g. Bellet *et al.* (2013). Note that when we use a data-driven cost function, we may no longer have correspondence as (1.6) but we can still directly solve the DRO problem on the left hand side. We expect that DD-DRO is able to improve the generalization property compared to many other state-of-the-art classifiers on a large number of data sets from UCI machine learning database, because it exploits the side information (the information about the intrinsic metric, the "shape") of the data.

The main methodologies and contributions of this project are the followings:

- We would use DRO as a link that combines k-NN methods with logistic regressions for classification. We use k-NN method to generate the side information of the data and then form the shape of the distributional uncertainty neighborhood by learning a metric from this side information.

- The DD-DRO is able to recover adaptive regularized ridge regression estimator. The DD-DRO provides a novel and interpretable way to select hyper-parameters in adaptive regularized ridge regression (see e.g.Zou (2006)) from a metric learning perspective.

- We would use an approximation algorithm based on stochastic gradient descent to solve DD-DRO. We would reformulate the DRO problem by using the duality representation given in Blanchet and Murthy (2016) and then solve it by smooth approximation and stochastic gradient descend algorithms.

- We would employ the robust metric learning to deal with the noisiness of side information. Since the side information is usually noisy, we borrow the idea from robust optimization (see e.g. Ben-Tal *et al.* (2009)) and build a doubly robust data-driven distributionally robust optimization (DD-R-DRO) model on top of the DD-DRO model to achieve robust metric learning.

Chapter 2

On Distributionally Robust Extreme Value Analysis

2.1 Introduction

Extreme Value Theory (EVT) provides reasonable statistical principles which can be used to extrapolate tail distributions, and, consequently, estimate extreme quantiles. However, as with any form for extrapolation, extreme value analysis rests on assumptions that are rather difficult (or impossible) to verify. Therefore, it makes sense to provide a mechanism to robustify the inference obtained via EVT.

The goal of this paper is to study non-parametric distributional robustness (i.e. finding the worst case distribution within some discrepancy of a natural baseline model) in the context of EVT. We ultimately provide a data-driven method for estimating extreme quantiles in a manner that is robust against possibly incorrect model assumptions. Our objective here is different from standard statistical robustness which is concerned with data contamination only (not model error); see, for example, Tsai *et al.* (2010), for this type of analysis in the setting of EVT.

Our focus in this paper is closer in spirit to distributionally robust optimization as in, for instance, Dupuis *et al.* (2000), Hansen and Sargent (2001), Ben-Tal *et al.* (2013), Breuer and Csiszár (2013b). However, in contrast to the literature on robust optimization, the emphasis here is on understanding the implications of distributional uncertainty regions in the context of EVT. As far as we know this is the first paper that studies distributional robustness in the context of EVT.

We now describe the content of the paper, following the logic which motivates the use of EVT.

2.1.1 Motivation and Standard Approach

In order to provide a more detailed description of the content of this paper, its motivations, the specific contributions, and the methods involved, let us invoke a couple of typical examples which motivate the use of extreme value theory. As a first example, consider the problem of forecasting the necessary strength that is required for a skyscraper in New York City to withstand a wind speed that gets exceeded only about once in 1000 years, using wind speed data that is observed only over the last 200 years. In another instance, given the losses observed during the last few decades, a reinsurance firm may want to compute, as required by Solvency II standard, a capital requirement that is needed to withstand all but about one loss in 200 years.

These tasks, and many others in practice, present a common challenge of extrapolating tail distributions over regions involving unobserved evidence from available observations. There are many reasonable ways of doing these types of extrapolations. One might take advantage of physical principles and additional information, if available, in the windspeed setting; or use economic principles in the reinsurance setting. In the absence of any fundamental principles which inform tail extrapolation of a random variable X, one may opt to use purely statistical considerations.

One such statistical approach entails the application of the popular extremal types theorem (see Section 2.2) to model the distribution of block maxima of a modestly large number of samples of X, by a generalized extreme value (GEV) distribution. Once we have a satisfactory model for the distribution of $M_n = \max\{X_1, \ldots, X_n\}$, evaluation of any desired quantile of X is straighforward because of the relationship that $P(M_n \leq x) = (P(X \leq x))^n$ for any $x \in \mathbb{R}$. Another common approach is to use samples that exceed a certain threshold to model conditional distribution of X exceeding the threshold. The standard texts in extreme value theory (see, for example, Leadbetter *et al.* (1983),de Haan and Ferreira (2006),Resnick (2008)) provide a comprehensive account of such standard statistical approaches.

Regardless of the technique used, various assumptions underlying an application of a result similar to the extremal types theorem might be subject to model error. Consequently, it has been widely accepted that tail risk measures, particularly for high confidence levels, can only be estimated with considerable statistical as well as model uncertainty (see, for example, Jorion (2006)). The following remark due to Coles (2001) holds significance in this discussion: "Though the GEV model is supported by mathematical argument, its use in extrapolation is based on unverifiable assumptions, and measures of uncertainty on return levels should properly be regarded as *lower bounds* that could be much greater if uncertainty due to model correctness were taken into account."

Despite these difficulties, however, EVT is widely used (see, for example, de Haan and Ferreira (2006)) and regarded as a reasonable way of extrapolation to estimate extreme quantiles.

2.1.2 Proposed Approach Based on Infinite Dimensional Optimization

We share the point of view that EVT is a reasonable approach, so we propose a procedure that builds on the use of EVT to provide upper bounds which attempts to address the types of errors discussed in the remark above from Coles (2001). For large values of n, under the assumptions of EVT, the distribution of M_n lies close to, and appears like, a GEV distribution. Therefore, instead of considering only the GEV distribution as a candidate model, we propose a non-parametric approach. In particular, we consider a family of probability models, all of which lie in a "neighborhood" of a GEV model, and compute a conservative worst-case estimate of Value at risk (VaR) over all of these candidate models. For $p \in [0,1]$, the value at risk $\mathrm{VaR}_p(X)$ is defined as

$$\mathrm{VaR}_p(X) = F^{\leftarrow}(p) := \inf\{x : P\{X \leq x\} \geq p\}.$$

Mathematically, given a reference model, P_{ref}, which we consider to be obtained using EVT (using a procedure such as the one outlined in the previous subsection), we consider the optimization problem

$$\sup \left\{ P\{X > x\} : \; d(P, P_{ref}) \leq \delta \right\}. \tag{2.1}$$

Note that the previous problem proposes optimizing over all probability measures that are within a tolerance level δ (in terms of a suitable discrepancy measure d) from the chosen baseline reference model P_{ref}.

There is a wealth of literature that pursues this line of thought (see Dupuis *et al.* (2000), Hansen and Sargent (2001), Ahmadi-Javid (2012), Ben-Tal *et al.* (2013),Breuer and Csiszár (2013b),Glasserman and Xu (2014b)), but, no study has been carried out in the context of EVT. Moreover, while the solvability of problems as in (2.1) have understandably received a great deal of attention, the qualitative differences that arise by using various choices of discrepancy measures, d, has not been explored, and this is an important contribution of this paper. For tractability reasons, the usual choice for discrepancy d in the literature has been KL-divergence. In Section 2.3 we study the solution to infinite dimensional optimization problems such as (2.1) for a large class of discrepancies that includes KL-divergence as a special case, and discuss how such problems can be solved at no significant computational cost.

2.1.3 Choosing Discrepancy and Consistency Results

One of our main contributions in this paper is to systematically demonstrate the qualitative differences that arise by using different choices of discrepancy measures d in (2.1). Since our interest in the paper is limited to robust tail modeling via EVT, this narrow scope, in turn, lets us analyse the qualitative differences that may arise because of different choices of d.

As mentioned earlier, the KL-divergence[1] is the most popular choice for d. In Section 2.4 we show that for any divergence neighborhood $\mathcal{P}$, defined using $d =$ KL-divergence around a baseline reference P_{ref}, there exists a probability measure P in $\mathcal{P}$ that has tails as heavy as

$$P(x, \infty) \geq c \log^{-2} P_{ref}(x, \infty),$$

[1] KL-divergence, and all other relevant divergence measures, are defined in Section 2.3.1

for a suitable constant c, and all large enough x. This means, irrespective of how small δ is (smaller δ corresponds to smaller neighborhood $\mathcal{P}$), a KL-divergence neighborhood around a commonly used distribution (such as exponential, (or) Weibull (or) Pareto) typically contains tail distributions that have infinite mean or variance, and whose tail probabilities decay at an unrealistically slow rate (even logarithmically slow, like $\log^{-2} x$, in the case of reference models that behave like a power-law or Pareto distribution). As a result, computations such as worst-case expected short-fall[2] may turn out to be infinite. Such worst-case analyses are neither useful nor interesting.

For our purposes, we also consider a general family of divergence measures D_α that includes KL-divergence as a special case (when $\alpha = 1$). It turns out that for any $\alpha > 1$, the divergence neighborhoods defined as in $\{P : D_\alpha(P, P_{ref}) \leq \delta\}$ consists of tails that are heavier than P_{ref}, but not prohibitively heavy. More importantly, we prove a "consistency" result in the sense that if the baseline reference model belongs to the maximum domain of attraction of a GEV distribution with shape parameter γ_{ref}, then the corresponding worst-case tail distribution,

$$\bar{F}_\alpha(x) := \sup\{P(x, \infty) : D_\alpha(P, P_{ref}) \leq \delta\}, \tag{2.2}$$

belongs to the maximum domain of attraction of a GEV distribution with shape parameter $\gamma^* = (1 - \alpha^{-1})^{-1}\gamma_{ref}$ (if it exists).

Since our robustification approach is built resting on EVT principles, we see this consistency result as desirable. If a modeler who is familiar with certain type of data expects the EVT inference to result in an estimated shape parameter which is positive, then the robustification procedure should preserve this qualitative property. An analysis of the maximum domain of attraction of the distribution $\bar{F}_\alpha(x)$, depending on α and γ_{ref}, is presented in Section 2.4, along with a summary of the results in Table 1.

Note that the smaller the value of α, the larger the absolute value of shape parameter γ^*, and consecutively, heavier the corresponding worst-case tail is. This indicates a gradation in the rate of decay of worst-case tail probabilities as parameter α decreases to 1, with

[2] Similar to VaR, expected shortfall (or) conditional value at risk (referred as CVaR) is another widely recognized risk measure.

the case $\alpha = 1$ (corresponding to KL-divergence) representing the extreme heavy-tailed behaviour. This gradation, as we shall see, offers a great deal of flexibility in modeling by letting us incorporate domain knowledge (or) expert opinions on the tail behaviour. If a modeler is suspicious about the EVT inference he/she could opt to select $\alpha = 1$, but, as we have mentioned earlier, this selection may result in pessimistic estimates.

The relevance of these results shall become more evident as we introduce the required terminology in the forthcoming sections. Meanwhile, Table 2.1 and Figure 2.1 offer illustrative comparisons of $\bar{F}_\alpha(x)$ for various choices of α.

2.1.4 The Final Estimation Procedure

The framework outlined in the previous subsections yields a data driven procedure for estimating VaR which is presented in Section 2.5. A summary of the overall procedure is given in Algorithm 2. The procedure is applied to various data sets, resulting in different reference models, and we emphasize the choice of different discrepancy measures via the parameter α. The numerical studies expose the salient points discussed in the previous subsections and rigorously studied via our theorems. For instance, Example 3 shows how the use of the KL divergence might lead to rather pessimistic estimates. Moreover, Example 4 illustrates how the direct application of EVT can severely underestimate the quantile of interest, while the procedure that we advocate provides correct coverage for the extreme quantile of interest.

The very last section of the paper, Section 2.6, contains technical proofs of various results invoked in the development.

2.2 Generalized extreme value distributions

The objective of this section is to mainly fix notation and review properties of generalized extreme value (GEV) distributions that are relevant for introducing and proving our main results in Section 2.4. For a thorough introduction to GEV distributions and their applications to modeling extreme quantiles, we refer the readers to the wealth of literature that is available (see, for example, Leadbetter *et al.* (1983), Embrechts *et al.* (1997), de Haan and

Ferreira (2006), Resnick (2008) and references therein).

If we use M_n to denote the maxima of n independent copies of a random variable X with cumulative distribution funtion $F(\cdot)$, then extremal types theorem identifies all non-degenerate distributions $G(\cdot)$ that may occur in the limiting relationship,

$$\lim_{n\to\infty} P\left\{\frac{M_n - b_n}{a_n} \leq x\right\} = \lim_{n\to\infty} F^n\left(a_n x + b_n\right) = G(x), \tag{2.3}$$

for every continuity point x of $G(\cdot)$, with a_n and b_n representing suitable scaling constants. All such distributions $G(x)$ that occur in the right-hand side of (2.3) are called *extreme value distributions.*

Extremal types theorem (Fisher and Tippet (1928), Gnedenko (1943)). The class of extreme value distributions is $G_\gamma(ax + b)$ with $a > 0, b, \gamma \in \mathbb{R}$, and

$$G_\gamma(x) := \exp\Big(-\left(1+\gamma x\right)^{-1/\gamma}\Big), \qquad 1 + \gamma x > 0. \tag{2.4}$$

If $\gamma = 0$, the right-hand side is interpreted as $\exp(-\exp(-x))$.

The extremal types theorem asserts that any $G(x)$ that occurs in the right-hand side of (2.3) must be of the form $G_\gamma(ax + b)$. As a convention, any probability distribution $F(x)$ that gives rise to the limiting distribution $G(x) = G_\gamma(ax + b)$ in (2.3) is said to belong to the maximum domain of attraction of $G_\gamma(x)$. In short, it is written as $F \in \mathcal{D}(G_\gamma)$. The parameters $\gamma, a > 0$ and b are, respectively, called the shape, scale and location parameters. From the above we have

$$P(M_n \leq x) = P\Big(\frac{M_n - b_n}{a_n} \leq \frac{x - b_n}{a_n}\Big) \approx G_{\gamma_0}\Big(\frac{x - b_n}{a_n}\Big) =: G_{\gamma_0}(a_0 x + b_0),$$

where γ_0, a_n, b_n are estimated by a parameter estimation technique such as maximum likelihood and $a_0 := 1/a_n,\ b_0 := -b_n/a_n$. We will use P_{GEV} to denote the distribution $G_{\gamma_0}(a_0 x + b_0)$.

2.2.1 Frechet, Gumbel and Weibull types

Though the limiting distributions $G_\gamma(ax+b)$ seem to constitute a simple parametric family, they include a wide-range of tail behaviours in their maximum domains of attraction, as discussed below: For a distribution F, let $\bar{F}(x) = 1 - F(x)$ denote the corresponding tail probabilities, and $x_F^* = \sup\{x : F(x) < 1\}$ denote the right endpoint of its support.

1) **The Frechet Case** ($\gamma > 0$). A distribution $F \in \mathcal{D}(G_\gamma)$ for some $\gamma > 0$, if and only if right endpoint x_F^* is unbounded, and its tail probabilities satisfy

$$\bar{F}(x) = \frac{L(x)}{x^{1/\gamma}}, \qquad x > 0 \tag{2.5}$$

for a function $L(\cdot)$ slowly varying at ∞[3]. As a consequence, moments greater than or equal to $1/\gamma$ do not exist. Any distribution $F(x)$ that lies in $\mathcal{D}(G_\gamma)$ for some $\gamma > 0$ is also said to belong to the maximum domain of attraction of a Frechet distribution with parameter $1/\gamma$. The Pareto distribution $1 - F(x) = x^{-\alpha} \wedge 1$ is an example for a distribution that belongs to $\mathcal{D}(G_{1/\alpha})$.

2) **The Weibull case** ($\gamma < 0$). Unlike the Frechet case, a distribution $F \in \mathcal{D}(G_\gamma)$ for some $\gamma < 0$, if and only if its right endpoint x_F^* is finite, and its tail probabilities satisfy

$$\bar{F}(x_F^* - \epsilon) = \epsilon^{-1/\gamma} L\left(\frac{1}{\epsilon}\right), \qquad \epsilon > 0 \tag{2.6}$$

for a function $L(\cdot)$ slowly varying at ∞. A distribution that belongs to $\mathcal{D}(G_\gamma)$ for some $\gamma < 0$ is also said to belong to the maximum domain of attraction of Weibull family. The uniform distribution on the interval $[0, 1]$ is an example that belongs to this class of extreme value distributions.

3) **The Gumbel case** ($\gamma = 0$). A distribution $F \in \mathcal{D}(G_0)$ if and only if

$$\lim_{t \uparrow x_F^*} \frac{\bar{F}(t + xf(t))}{\bar{F}(t)} = \exp(-x), \qquad x \in \mathbb{R} \tag{2.7}$$

for a suitable positive function $f(\cdot)$. In general, the members of G_0 have exponentially decaying tails, and consequently, all moments exist. Probability distributions $F(\cdot)$ that give rise to limiting distributions $G_0(ax + b)$ are also said to belong to the Gumbel domain of attraction. Common examples that belong to the Gumbel domain of attraction include exponential and normal distributions.

[3] A function $L : \mathbb{R} \to \mathbb{R}$ is said to be slowly varying at infinity if $\lim_{x\to\infty} L(tx)/L(x) = 1$ for every $t > 0$. Common examples of slowly varying function include $\log x, \log\log x, 1 - \exp(-x)$, constants, etc.

Given a distribution function F, Proposition 2.1 is useful to test to determine its domain of attraction:

Proposition 2.1. Suppose $F''(x)$ exists and $F'(x)$ is positive for all x in some left neighborhood of x_F^*. If

$$\lim_{x \uparrow x_F^*} \left(\frac{1-F}{F'} \right)'(x) = \gamma, \tag{2.8}$$

then F belongs to the domain of attraction of G_γ.

The proof of Proposition 2.1 and further details on the classification of extreme value distributions can be found in any standard text on extreme value theory (see, for example, Leadbetter *et al.* (1983) or de Haan and Ferreira (2006)).

2.2.2 On model errors and robustness

After identifying a suitable GEV model P_{GEV} for the distribution of block maxima M_n, it is common to utilize the relationship $P\{M_n \leq x\} = P\{X \leq x\}^n$, to compute a desired extreme quantile of X. It is useful to remember that $P_{GEV}(-\infty, x]$ is only an approximation for $P\{M_n \leq x\}$, and the quality of the approximation is, in turn, dependent on the unknown distribution function F (see Resnick (2008),de Haan and Ferreira (2006)). Therefore, in practice, one does not know the block-size n for which the GEV model P_{GEV} well-approximates the distribution of M_n. Even if a good choice of n is known, one cannot often employ it in practice, because larger n means smaller m, and consequentially, the inferential errors could be large. Due to the arbitrariness in the estimation procedures and the nature of applications (calculating wind speeds for building sky-scrapers, building dykes for preventing floods, etc.), it is desirable to have, in addition, a data-driven procedure that yields a conservative upper bound for x_p that is robust against model errors. To accomplish this, one can form a collection of competing probability models $\mathcal{P}$, all of which appear plausible as the distribution of M_n, and compute the maximum of p^n-th quantile over all the plausible models in $\mathcal{P}$. This is indeed the objective of the sections that follow.

2.3 A non-parametric framework for addressing model errors

Let $(\Omega, \mathcal{F})$ be a measurable space and $M_1(\mathcal{F})$ denote the set of probability measures on $(\Omega, \mathcal{F})$. Let us assume that a reference probability model $P_{ref} \in M_1(\mathcal{F})$ is inferred by suitable modelling and estimation procedures from historical data. Naturally, this model is not the same as the distribution from which the data has been generated, and is expected only to be close to the data generating distribution. In the context of Section 2.2, the model P_{ref} corresponds to P_{GEV}, and the data generating model corresponds to the true distribution of M_n. With slight perturbations in data, we would, in turn, be working with a slightly different reference model. Therefore, it has been of recent interest to consider a family of probability models $\mathcal{P}$, all of which are plausible, and perform computations over all the models in that family. Following the rich literature of robust optimization, where it is common to describe the set of plausible models using distance measures (see Ben-Tal *et al.* (2013)), we consider the set of plausible models to be of the form

$$\mathcal{P} = \left\{ P \in M_1(\mathcal{F}) : d\left(P, P_{ref}\right) \leq \delta \right\}$$

for some distance functional $d : M_1(\mathcal{F}) \times M_1(\mathcal{F}) \to \mathbb{R}_+ \cup \{+\infty\}$, and a suitable $\delta > 0$. Since $d(P_{ref}, P_{ref}) = 0$ for any reasonable distance functional, P_{ref} lies in $\mathcal{P}$. Therefore, for any random variable X, along with the conventional computation of $E_{P_{ref}}[X]$, one aims to provide “robust” bounds,

$$\inf_{P \in \mathcal{P}} E_P[X] \leq E_{P_{ref}}[X] \leq \sup_{P \in \mathcal{P}} E_P[X].$$

Here, we follow the notation that $E_P[X] = \int X dP$ for any $P \in M_1(\mathcal{F})$. Since the state-space Ω is uncountable, evaluation of the above sup and inf-bounds, in general, are infinite-dimensional problems. However, as it has been shown in the recent works Breuer and Csiszár (2013b),Glasserman and Xu (2014b), it is indeed possible to evaluate these robust bounds for carefully chosen distance functionals d.

2.3.1 Divergence measures

Consider two probability measures P and Q on $(\Omega, \mathcal{F})$ such that P is absolutely continuous with respect to Q. The Radon-Nikodym derivative dP/dQ is then well-defined. The

Kullback-Liebler divergence (or KL-divergence) of P from Q is defined as

$$D_1(P,Q) := E_Q\left[\frac{dP}{dQ}\log\left(\frac{dP}{dQ}\right)\right]. \tag{2.9}$$

This quantity, also referred to as relative entropy (or) information divergence, arises in various contexts in probability theory. For our purposes, it will be useful to consider a general class of divergence measures that includes KL-divergence as a special case. For any $\alpha > 1$, the Rényi divergence of degree α is defined as:

$$D_\alpha(P,Q) := \frac{1}{\alpha-1}\log E_Q\left[\left(\frac{dP}{dQ}\right)^\alpha\right]. \tag{2.10}$$

It is easy to verify that for every α, $D_\alpha(P,Q) = 0$, if and only if $P = Q$. Additionally, the map $\alpha \mapsto D_\alpha$ is nondecreasing, and continuous from the left. Letting $\alpha \to 1$ in (2.10) yields the formula for KL-divergence $D_1(P,Q)$. Thus KL-divergence is a special case of the family of Rényi divergences, when the parameter α equals 1. If the probability measure P is not absolutely continuous with respect to Q, then $D_\alpha(P,Q)$ is taken as ∞. Though none of these divergence measures form a metric on the space of probability measures, they have been used in a variety of scientific disciplines to discriminate between probability measures. For more details on the divergences D_α, see Rényi (1961),Liese and Vajda (1987).

2.3.2 Robust bounds via maximization of convex integral functionals

Recall that P_{ref} is the reference probability measure obtained via standard estimation procedures. Since the model P_{ref} could be misspecified, we consider all models that are not far from P_{ref} in the sense quantified by divergence D_α, for any fixed $\alpha \geq 1$. Given a random variable X, we consider optimization problems of form

$$V_\alpha(\delta) := \sup\left\{E_P[X] : D_\alpha(P, P_{ref}) \leq \delta\right\}. \tag{2.11}$$

Though KL-divergence has been a popular choice in defining sets of plausible probability measures as above, use of divergences D_α, $\alpha \neq 1$ is not new altogether: see Atar *et al.* (2015),Glasserman and Xu (2014b). Due to the Radon-Nikodym theorem, $V_\alpha(\delta)$ can be alternatively written as,

$$V_\alpha(\delta) = \sup\left\{E_{P_{ref}}[LX] : E_{P_{ref}}[\phi_\alpha(L)] \leq \bar{\delta}, E_{P_{ref}}[L] = 1, L \geq 0\right\}, \tag{2.12}$$

where $L = dP/dP_{ref}$ and

$$\phi_\alpha(x) = \begin{cases} x^\alpha & \text{if } \alpha > 1, \\ x \log x & \text{if } \alpha = 1 \end{cases} \quad \text{and} \quad \bar{\delta} = \begin{cases} \exp\left((\alpha-1)\delta\right) & \text{if } \alpha > 1, \\ \delta & \text{if } \alpha = 1. \end{cases} \tag{2.13}$$

A standard approach for solving optimization problems of the above form is to write the corresponding dual problem as below:

$$V_\alpha(\delta) \le \inf_{\substack{\lambda>0,\\ \mu}} \sup_{L \ge 0} E_{P_{ref}} \left[LX - \lambda\left(\phi_\alpha(L) - \bar{\delta}\right) + \mu(L-1) \right].$$

The above dual problem can, in turn, be relaxed by taking the sup inside the expectation:

$$V_\alpha(\delta) \le \inf_{\substack{\lambda>0,\\ \mu}} \left\{ \lambda\bar{\delta} - \mu + \lambda E_{P_{ref}} \left[\sup_{L \ge 0} \left\{ \frac{(X+\mu)}{\lambda} L - \phi_\alpha(L) \right\} \right] \right\}. \tag{2.14}$$

By first order condition the inner supremum is solved by

$$L^*_\alpha(c_1, c_2) := \begin{cases} c_1 \exp(c_2 X), & \text{if } \alpha = 1, \\ (c_1 + c_2 X)_+^{1/(\alpha-1)}, & \text{if } \alpha > 1, \end{cases} \tag{2.15}$$

for some suitable constants $c_1 \in \mathbb{R}, c_2 > 0$ when $\alpha > 1$; and $c_1 \in (0,1)$ and $c_2 > 0$ when $\alpha = 1$. Then the following theorem is intuitive:

Theorem 2.2. *Fix any $\alpha \ge 1$. For $L^*_\alpha(c_1, c_2)$ defined as in (2.15), if there exists constants c_1 and c_2 such that*

$$L^*_\alpha(c_1, c_2) \ge 0,\ E_{P_{ref}} \left[L^*_\alpha(c_1, c_2)\right] = 1 \text{ and } E_{P_{ref}} \left[\phi_\alpha\left(L^*_\alpha(c_1, c_2)\right)\right] = \bar{\delta},$$

then $L^*_\alpha(c_1, c_2)$ solves the optimization problem (2.12). The corresponding optimal value is

$$V_\alpha(\delta) = E_{P_{ref}} \left[L^*_\alpha(c_1, c_2) X\right]. \tag{2.16}$$

Proof. Under the specified assumptions, when we plug $L^*_\alpha(c_1, c_2)$ into the right-hand-side of inequality (2.14), it is simplified to $E_{P_{ref}} \left[L^*_\alpha(c_1, c_2)X\right]$, so we have $V_\alpha(\delta) \le E_{P_{ref}} \left[L^*_\alpha(c_1, c_2)X\right]$. On the other hand, since $L^*_\alpha(c_1, c_2)$ satisfies all the constraints in the problem (2.12), we have $V_\alpha(\delta) \ge E_{P_{ref}} \left[L^*_\alpha(c_1, c_2)X\right]$. □ □

Remark 2.1. Let us say one can determine constants c_1 and c_2 for given X, α and δ. Then, as a consequence of Theorem 2.2, the optimization problem (2.11) involving uncountably many measures can, in turn, be solved by simply simulating X from the original reference measure P_{ref}, and multiplying by corresponding $L^*_\alpha(c_1, c_2)$ to compute the expectation as in (2.16).

A general theory for optimizing convex integral functionals of form (2.12), that includes a bigger class of general divergence measures, can be found in Breuer and Csiszár (2013b). If the random variable X above is an indicator function, then computation of bounds $V_\alpha(\delta)$ turns out to be even simpler, as illustrated in the example below:

Example 2.1. Let P_{ref} be a probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. For a given $\delta > 0$ and $\alpha \geq 1$, let us say we are interested in evaluating the worst-case tail probabilities

$$\bar{F}_{\alpha,\delta}(x) := \sup\{P(x,\infty) : D_\alpha(P, P_{ref}) \leq \delta\}.$$

Consider the canonical mapping $Z(\omega) = \omega,\ \omega \in \mathbb{R}$. Then

$$\bar{F}_{\alpha,\delta}(x) = \sup\left\{E_{P_{ref}}[L\mathbf{1}(Z > x)] : E_{P_{ref}}[\phi_\alpha(L)] \leq \bar{\delta}, E_{P_{ref}}[L] = 1, L \geq 0\right\}.$$

is an optimization problem of the form (2.11). Therefore, due to Theorem 2.2 and equation (2.15), the optimal L^* has the form

$$L^*_\alpha(c_1, c_2) := \begin{cases} c_1 \exp(c_2\mathbf{1}(Z > x)), & \text{if } \alpha = 1, \\ (c_1 + c_2\mathbf{1}(Z > x))_+^{1/(\alpha-1)}, & \text{if } \alpha > 1, \end{cases}$$

When we consider the two cases of $Z > x$ and $Z \leq x$, and combine the range information on c_1, c_2 following equation (2.15), the above formulation of $L^*_\alpha(c_1, c_2)$ can further be simplified to $\theta\mathbf{1}(x,\infty) + \tilde{\theta}\mathbf{1}(-\infty, x]$ for some constants $\theta > 1$ and $\tilde{\theta} \in (0,1)$. Substituting for $L^* = \theta\mathbf{1}(x,\infty) + \tilde{\theta}\mathbf{1}(-\infty, x]$ in the constraints $E_{P_{ref}}[\phi_\alpha(L^*)] = \bar{\delta}$ and $E_{P_{ref}}[L^*] = 1$, we obtain the following conclusion: Given $x > 0$, if there exists a $\theta_x > 1$ such that

$$P_{ref}(x,\infty)\phi_\alpha(\theta_x) + P_{ref}(-\infty, x]\phi_\alpha\left(\frac{1 - \theta_x P_{ref}(x,\infty)}{P_{ref}(-\infty, x]}\right) = \bar{\delta}, \tag{2.17}$$

then $\bar{F}_{\alpha,\delta}(x) = \theta_x P_{ref}(x,\infty)$.

2.4 Asymptotic analysis of robust estimates of tail probabilities

In this section we study the asymptotic behaviour of $\bar{F}_{\alpha,\delta}(x) := \sup\{P(x,\infty) : D_\alpha(P, P_{ref}) \leq \delta\}$, for any $\alpha \geq 1$ and $\delta > 0$, as $x \to \infty$. We first verify in Proposition 2.3 below that $\bar{F}_{\alpha,\delta}(x)$, viewed as a function of x, satisfies the properties of a tail distribution function. A proof of Proposition 2.3 is presented in Section 2.6.

Proposition 2.3. *The function, $F_{\alpha,\delta}(x) := 1 - \bar{F}_{\alpha,\delta}(x)$, viewed as a function of x, satisfies properties of cumulative distribution function of a real-valued random variable.*

Thus from here onwards, we shall refer $\bar{F}_{\alpha,\delta}(\cdot)$ as the *α-family worst-case tail distribution*, and study its qualitative properties such as domain of attraction for the rest of this section. All the probability measures involved, unless explicitly specified, are taken to be defined on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Since $D_\alpha(P_{ref}, P_{ref}) = 0$, it is evident that the worst-case tail estimate $\bar{F}_{\alpha,\delta}(x)$ is at least as large as $P_{ref}(x,\infty)$. While the overall objective has been to provide robust estimates that account for model perturbations, it is certainly not desirable that the worst-case tail distribution $\bar{F}_{\alpha,\delta}(\cdot)$, for example, has unrealistically slow logarithmic decaying tails. Seeing this, our interest in this section is to quantify how heavier the tails of $\bar{F}_{\alpha,\delta}(\cdot)$ are, when compared to that of the reference model.

The bigger the plausible family of measures $\{P : D_\alpha(P, P_{ref}) \leq \delta\}$, the slower the decay of tail $\bar{F}_{\alpha,\delta}(x)$ is, and vice versa. Hence it is conceivable that the parameter δ is influential in determining the rate of decay of $\bar{F}_{\alpha,\delta}(\cdot)$. However, as we shall see below in Theorem 2.5, it is the parameter α (along with the tail properties of the reference model P_{ref}) that solely determines the domain of attraction, and hence the rate of decay, of $\bar{F}_{\alpha,\delta}(\cdot)$.

Since our primary interest in the paper is with respect to reference model P_{ref} being a GEV model, we first state the result in this context:

Theorem 2.4. *Let the reference GEV model P_{GEV} have shape parameter γ_{ref}. Then the distribution F induced by P_{GEV} satisfies the regularity assumptions of Proposition 2.1 with $\gamma = \gamma_{ref}$. For any $\alpha > 1$, let $\bar{F}_{\alpha,\delta}(x) := \sup\{P(x,\infty) : D_\alpha(P, P_{GEV}) \leq \delta\}$, and*

$$\gamma^* := \frac{\alpha}{\alpha - 1}\gamma_{ref}.$$

Then the distribution function $F_{\alpha,\delta}(x) = 1 - \bar{F}_{\alpha,\delta}(x)$ belongs to the domain of attraction of G_{γ^*}.

Theorem 2.4 is, however, a corollary of Theorem 2.5 below.

Theorem 2.5. *Let the reference model P_{ref} belong to the domain of attraction of $G_{\gamma_{\text{ref}}}$. In addition, let P_{ref} induce a distribution F that satisfies the regularity assumptions of Proposition 2.1 with $\gamma = \gamma_{ref}$. For any $\alpha > 1$, let $\bar{F}_{\alpha,\delta}(x) := \sup\{P(x,\infty) : D_\alpha(P, P_{ref}) \leq \delta\}$, and*

$$\gamma^* := \frac{\alpha}{\alpha - 1}\gamma_{ref}.$$

Then the distribution function $F_{\alpha,\delta}(x) = 1 - \bar{F}_{\alpha,\delta}(x)$ belongs to the maximum domain of attraction of G_{γ^*}.

The special case corresponding to $\alpha = 1$ is handled in Propositions 2.6 and 2.7. Proofs of Theorems 2.4 and 2.5 are presented in Section 2.6.

Remark 2.2. First, observe that $P(x,\infty) \leq \bar{F}_{\alpha,\delta}(x)$, for every P in the neighborhood set of measures $\mathcal{P}_{\alpha,\delta} := \{P : D_\alpha(P, P_{ref}) \leq \delta\}$. Therefore, for any $\alpha > 1$, apart from characterizing the domain of attraction of $\bar{F}_{\alpha,\delta}$, Theorem 2.5 offers the following insights on the neighborhood $\mathcal{P}_{\alpha,\delta}$:

1) If the reference model belongs to the domain of attraction of a Frechet distribution (that is, $\gamma_{ref} > 0$), and if P is a probability measure that lies in its neighborhood $\mathcal{P}_{\alpha,\delta}$, then P must satisfy that

$$P(x,\infty) = O\left(x^{-\frac{\alpha-1}{\alpha\gamma_{ref}}+\epsilon}\right), \tag{2.18}$$

as $x \to \infty$, for every $\epsilon > 0$. This conclusion is a consequence of (2.5): $\bar{F}_{\alpha,\delta}$ is in the domain of attraction of G_{γ^*}, then by (2.5) we have

$$\bar{F}_{\alpha,\delta}(x) = L(x)x^{-1/\gamma^*} = L(x)x^{-\frac{\alpha-1}{\alpha\gamma_{ref}}},$$

and the observation that $P(x,\infty) \leq \bar{F}_{\alpha,\delta}(x)$. In addition, as in the proof of Theorem 2.5, one can exhibit a measure $P \in \mathcal{P}_{\alpha,\delta}$ such that $P(x,\infty) \geq cx^{-(\alpha-1)/\alpha\gamma_{ref}}$ for some $c > 0$ and all large enough x.

2) On the other hand, if the reference model belongs to the Gumbel domain of attraction ($\gamma_{ref} = 0$), then every $P \in \mathcal{P}_{\alpha,\delta}$ satisfies $P(x, \infty) = o(x^{-\epsilon})$, as $x \to \infty$, for every $\epsilon > 0$.

3) Now consider the case where $P_{ref} \in \mathcal{D}(G_{\gamma_{\text{ref}}})$ for some $\gamma_{ref} < 0$ (that is, the reference model belongs to the domain of attraction of a Weibull distribution). Let $x^*_F < \infty$ denote the supremum of its bounded support. In that case, any probability measure P that belongs to the neighborhood $\mathcal{P}_{\alpha,\delta}$ must satisfy that $P(-\infty, x^*_F) = 1$ and
$$P(x^*_F - \epsilon, x^*_F) = O\left(\epsilon^{-\frac{\alpha-1}{\alpha\gamma_{ref}} - \epsilon'}\right),$$
as $\epsilon \to 0$, for every $\epsilon' > 0$. In addition, one can exhibit a measure $P \in \mathcal{P}_{\alpha,\delta}$ such that $P(x^*_F - \epsilon, x^*_F) \geq c\epsilon^{-(\alpha-1)/\alpha\gamma_{ref}}$, for some positive constant c and all $\epsilon > 0$ sufficiently small.

It is important to remember that the above properties hold for all $\alpha > 1$, and is not dependent on δ.

For a fixed reference model P_{ref}, it is evident from Remark 2.2 that the neighborhoods $\mathcal{P}_{\alpha,\delta} = \{P : D_\alpha(P, P_{ref}) \leq \delta\}$ include probability distributions with heavier and heavier tails as α approaches 1 from above. This is in line with the observation that $D_\alpha(P, P_{ref})$ is a non-decreasing function in α, and hence larger neighborhoods $\mathcal{P}_{\alpha,\delta}$ for smaller values of α. In particular, when $\alpha = 1$ and shape parameter $\gamma_{ref} = 0$, the quantity $\gamma^* = \gamma_{ref}\alpha/(\alpha - 1)$ defined in Theorem 2.4 is not well-defined. This corresponds to the set of plausible measures $\{P : D_1(P, G_0) \leq \delta\}$ defined using KL-divergence around the reference Gumbel model G_0. The following result describes the tail behaviour of $\bar{F}_{\alpha,\delta}$ in this case:

Proposition 2.6. *Recall the definition of extreme value distributions G_γ in* (2.4). *Let $\bar{F}_{1,\delta}(x) = \sup\{P(x, \infty) : D_1(P, G_0) \leq \delta\}$, and $F_{1,\delta}(x) = 1 - \bar{F}_{1,\delta}(x)$. Then $F_{1,\delta}$ belongs to the domain of attraction of G_1.*

The following result, when contrasted with Remark 2.2, better illustrates the difference between the cases $\alpha > 1$ and $\alpha = 1$.

Proposition 2.7. *Recall the definition of G_γ as in* (2.4). *For every $\delta > 0$, one can find a probability measure P in the neighborhood $\{P : D_1(P, G_{\gamma_{\text{ref}}}) \leq \delta\}$, along with positive constants c_+ or c_- or c_0, and x_+ or x_0 or ϵ_- such that* *a)* $P(x,\infty) \geq c_+ \log^{-3} x$ *for every* $x > x_+$, *if* $\gamma_{_{ref}} > 0$;

b) $P(x,\infty) \geq c_0 x^{-1}$ *for every* $x > x_0$, *if* $\gamma_{_{ref}} = 0$; *and*

c) $P(-\infty, x^*_{_G}) = 1$ *and* $P(x^*_{_G} - \epsilon, x^*_{_G}) \geq c_3 \log^{-3} \frac{1}{\epsilon}$ *for every* $\epsilon < \epsilon_-$, *if* $\gamma_{_{ref}} < 0$. *Here, the right endpoint* $x^*_{_G} = \sup\{x : G_{\gamma_{_{ref}}}(x) < 1\}$ *is finite because* $\gamma_{_{ref}} < 0$.

In addition, it is useful to contrast these tail decay results for neighboring measures with that of the corresponding reference measure $G_{\gamma_{\text{ref}}}$ characterized in (2.5), (2.6) or (2.7). It is evident from this comparison that the worst-case tail probabilities $\bar{F}_{\alpha,\delta}(x)$ decay at a significantly slower rate than the reference measure when $\alpha = 1$ (the KL-divergence case). Table 2.1 below summarizes the rates of decay of worst-case tail probabilities $\bar{F}_{\alpha,\delta}(\cdot)$ over different choices of α when the reference model is a GEV distribution. In addition, Figure 2.1, which compares the worst-case tail distributions $\bar{F}_{\alpha,\delta}(x)$ for three different GEV example models, is illustrative. Proofs of Theorems 2.4 and 2.5, Propositions 2.6 and 2.7 are presented in Section 2.6.

2.5 Robust estimation of VaR

Given independent samples $X_1, \ldots, X_N$ from an unknown distribution F, we consider the problem of estimating $F^{\leftarrow}(p)$ for values of p close to 1. In this section, we develop a data-driven algorithm for estimating robust upper bounds for these extreme quantiles by employing traditional extreme value theory in tandem with the insights derived in Sections 2.3 and 2.4. Our motivation has been to provide conservative estimates for $F^{\leftarrow}(p)$ that are robust against incorrect model assumptions as well as calibration errors.

Naturally, the first step in the estimation procedure is to arrive at a reference model $P_{_{GEV}}(-\infty, x) = G_{\gamma_0}(a_0 x + b_0)$ for the distribution of block-maxima M_n. Once we have a candidate model $P_{_{GEV}}$ for M_n, the p^n-th quantile of the distribution $P_{_{GEV}}$ serves as an estimator for $F^{\leftarrow}(p)$. Instead, if we have a family of candidate models (as in Sections 2.3 and 2.4) for M_n, a corresponding robust alternative to this estimator is to compute the worst-case quantile estimate over all the candidate models as below:

$$\hat{x}_p := \sup\big\{G^{\leftarrow}(p^n) : D_\alpha(G, P_{_{GEV}}) \leq \delta\big\}. \tag{2.19}$$

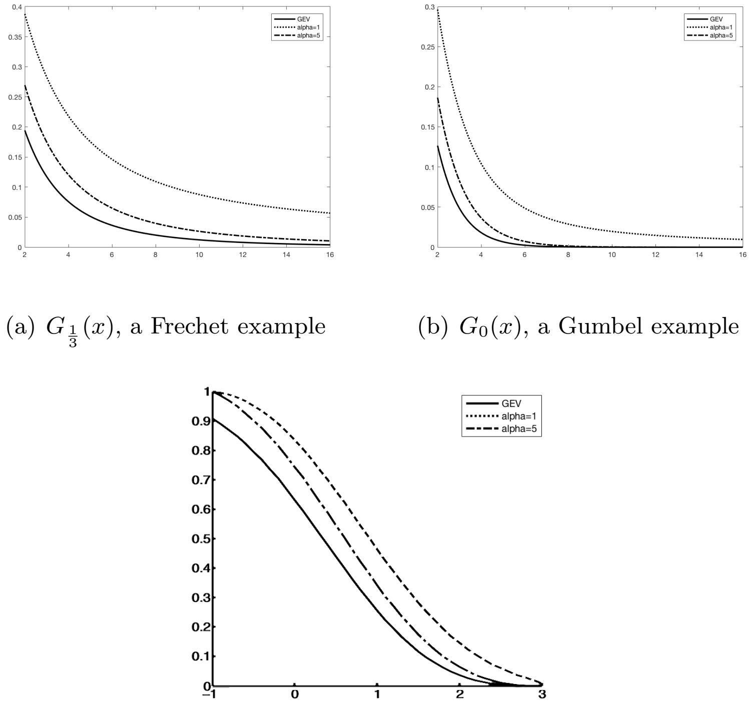
Table 2.1: A summary of domains of attraction of $F_{\alpha,\delta}(x) = 1 - \bar{F}_{\alpha,\delta}(x)$ for GEV models. Throughout the paper, $\gamma^* := \frac{\alpha}{\alpha-1}\gamma_{ref}$

Reference model	Domain of attraction of Worst-case tail $\bar{F}_{\alpha,\delta}(\cdot),\ \alpha > 1$	Domain of attraction of Worst-case tail $\bar{F}_{\alpha,\delta}(\cdot),\ \alpha = 1$ (the KL-divergence case)
G_0 (Gumbel light tails)	G_0 (Gumbel light tails)	G_1 (Frechet heavy tails)
$G_{\gamma_{\text{ref}}}, \gamma_{ref} > 0$ (Frechet heavy tails)	G_{γ^*} (Frechet heavy tails)	– (slow logarithmic decay of $\bar{F}_{\alpha,\delta}(x)$ as $x \to \infty$)
$G_{\gamma_{\text{ref}}},\ \gamma_{ref} < 0$ (Weibull)	G_{γ^*} (Weibull)	– (slow logarithmic decay of $\bar{F}_{\alpha,\delta}(x)$ to 0 at a finite right endpoint x^*)

Here $G^{\leftarrow}$ denotes the usual inverse function $G^{\leftarrow}(u) = \inf\{x : G(x) \geq u\}$ with respect to distribution G. Since the framework of Section 2.3 is limited to optimization over objective functionals in the form of expectations (as in (2.11)), it is immediately not clear whether the supremum in (2.19) can be evaluated using tools developed in Section 2.3. Therefore, let us proceed with the following alternative: First, compute the worst-case tail distribution

$$\bar{F}_{\alpha,\delta}(x) := \sup\left\{G(x,\infty) : D_\alpha(G, P_{GEV}) \leq \delta\right\}, \quad x \in \mathbb{R}$$

Figure 2.1: Comparison of $\bar{F}_{\alpha,\delta}(x)$ for different GEV models: The solid curves represents the reference model $G_{\gamma_{\mathrm{ref}}}(x)$ for $\gamma_{ref} = 1/3$ (top left figure), $\gamma_{ref} = 0$ (top right figure) and $\gamma_{ref} = -1/3$ (bottom figure). Computations of corresponding $\bar{F}_{\alpha,\delta}(x)$ are done for $\alpha = 1$ (the dotted curves), and $\alpha = 5$ (the dash-dot curves) with δ fixed at 0.1. The dotted curves (corresponding to $\alpha = 1$, the KL-divergence case) conform with our reasoning that $\bar{F}_{\alpha,\delta}(x)$ have vastly different tail behaviours from the reference models when KL-divergence is used.



(a) $G_{\frac{1}{3}}(x)$, a Frechet example

(b) $G_0(x)$, a Gumbel example

(c) $G_{-\frac{1}{3}}(x)$, a Weibull example

over all candidate models, and compute the corresponding inverse

$$F^{\leftarrow}_{\alpha,\delta}(p^n) := \inf\{x : 1 - \bar{F}_{\alpha,\delta}(x) \geq p^n\}.$$

The estimate $\hat{x}_p$ (defined as in (2.19)) is indeed equal to $F^{\leftarrow}_{\alpha,\delta}(p^n)$, and this is the content of Lemma 2.1.

Lemma 2.1. *For every $u \in (0,1)$,* $F^{\leftarrow}_{\alpha,\delta}(u) = \sup\left\{G^{\leftarrow}(u) : D_\alpha(G, P_{GEV}) \leq \delta\right\}.$

Proof. For brevity, let $\mathcal{P} = \{G : D_\alpha(G, P_{GEV}) \leq \delta\}$. Then, it follows from the definition of

$\bar{F}_{\alpha,\delta}(\cdot)$ and $F^{\leftarrow}_{\alpha,\delta}(\cdot)$ that

$$\begin{aligned}
F^{\leftarrow}_{\alpha,\delta}(u) &= \inf\left\{x : \sup_{G\in\mathcal{P}} G(x,\infty) \leq 1-u\right\}\\
&= \inf \bigcap_{G\in\mathcal{P}} \left\{x : G(x,\infty) \leq 1-u\right\}\\
&= \inf \bigcap_{G\in\mathcal{P}} \big[G^{\leftarrow}(u), \infty\big) = \sup_{G\in\mathcal{P}} G^{\leftarrow}(u).
\end{aligned}$$

This completes the proof of Lemma 2.1. □ □

Now that we know $\hat{x}_p = F^{\leftarrow}_{\alpha,\delta}(p^n)$ is the desired upper bound, let us recall from Example 2.1 how to evaluate $\bar{F}_{\alpha,\delta}(x)$ for any x of interest. If $\theta_x > 1$ solves

$$P_{GEV}(x,\infty)\phi_\alpha(\theta_x) + P_{GEV}(-\infty,x)\phi_\alpha\left(\frac{1-\theta_x P_{GEV}(x,\infty)}{P_{GEV}(-\infty,x)}\right) = \bar{\delta},$$

then $\bar{F}_{\alpha,\delta}(x) = \theta_x P_{GEV}(x,\infty)$. Though θ_x cannot be obtained in closed-form, given any $x > 0$, one can numerically solve for θ_x, and compute $\bar{F}_{\alpha,\delta}(x)$ to a desired level of precision. On the other hand, given a level $u \in (0,1)$, it is similarly possible to compute $F^{\leftarrow}_{\alpha,\delta}(u)$ by solving for x that satisfies $P_{GEV}(x,\infty) < 1-u$ and

$$P_{GEV}(x,\infty)\phi_\alpha\left(\frac{1-u}{P_{GEV}(x,\infty)}\right) + P_{GEV}(-\infty,x)\phi_\alpha\left(\frac{u}{P_{GEV}(-\infty,x)}\right) = \bar{\delta}. \tag{2.20}$$

Therefore, given α and δ, it is computationally not any more demanding to evaluate the robust estimates $F^{\leftarrow}_{\alpha,\delta}(p^n)$ for $F^{\leftarrow}(p)$.

2.5.1 On specifying the parameter δ.

For a given choice of paramter $\alpha \geq 1$, there are several divergence estimation methods available in the literature to obtain an estimate $\hat{\delta} = D_\alpha(\hat{P}_{M_n}, P_{GEV})$, where $\hat{P}_{M_n}$ is the empirical distribution of M_n. For our examples, we use the k-nearest neighbor (k-NN) algorithm of Póczos and Schneider (2011) and Q.Wang *et al.* (2009). See also Nguyen *et al.* (2009),Nguyen *et al.* (2010),Gupta and Srivastava (2010) for similar divergence estimators. These divergence estimation procedures provide an empirical estimate of the divergence between sample maxima and the calibrated GEV model P_{GEV}.

The specific details of the k-NN divergence estimation procedure we employ from Póczos and Schneider (2011) and Q.Wang *et al.* (2009) are provided in Remark 2.3 below:

Remark 2.3. *Suppose $M_{n,1}, \ldots, M_{n,m}$ are independent samples of M_n, and $L_1, \ldots, L_l$ are samples from P_{GEV}. Define $\rho_k(i)$ to be the Euclidean distance between $M_{n,i}$ and its k-th nearest neighbour among all $M_{n,1}, \ldots, M_{n,m}$ and similarly $\nu_k(i)$ the distance between $M_{n,i}$ and its k-th nearest neighbour among all $L_1, \ldots, L_l$. The k-NN based density estimators are*

$$\hat{p}_k(M_{n,i}) = \frac{k/(m-1)}{|B(\rho_k(i))|} \text{ and } \hat{q}_k(M_{n,i}) = \frac{k/l}{|B(\nu_k(i))|},$$

where $|B(\rho_k(i))|$ denotes the volume of a ball with radius $\rho_k(i)$. Then, for a fixed α, the estimator for $\delta = D_\alpha(P_{M_n}, P_{GEV})$ is given by

$$\hat{\delta} = \frac{1}{\alpha - 1} \log\Big(\frac{1}{m}\sum_{i=1}^{m}\Big(\frac{(m-1)\rho_k(i)}{l\nu_k(i)}\Big)^{1-\alpha} \cdot \frac{\Gamma(k)^2}{\Gamma(k-\alpha+1)\Gamma(k+\alpha-1)}\Big),$$

for $\alpha > 1$, where Γ denotes the gamma function, and

$$\hat{\delta} = \frac{1}{m}\sum_{i=1}^{m} \log\Big(\frac{l\nu_k(i)}{(m-1)\rho_k(i)}\Big),$$

for $\alpha = 1$.

For a fixed choice of $\alpha \geq 1$ and desired p close to 1, the Rob-Estimator(p, α) procedure in Algorithm 1 below provides a summary of the prescribed estimation procedure.

2.5.2 On specifying the parameter α.

To input to the estimation procedure Rob-Estimator(p, α) in Algorithm 1, one can perhaps choose α via one of the three approaches explained below:

1) Choose α so that the corresponding $\gamma^* = \gamma_0\alpha/(\alpha - 1)$ matches with an appropriate confidence interval for the estimate γ_0 : For example, if $\gamma_0 > 0$ and the confidence interval for γ_0, estimated from data, is given by $(\gamma_0 - \epsilon, \gamma_0 + \epsilon)$, then we choose α satisfying

$$\gamma_0 \frac{\alpha}{\alpha - 1} = \gamma_0 + \epsilon. \tag{2.21}$$

See Examples 2.2 and 2.3 for demonstrations of choosing α following this approach.

2) Alternatively, one can choose α based on domain knowledge as well: For example, consider the case where one uses Gaussian distribution to model returns of a portfolio.

In this instance, if a financial expert identifies the returns are instead heavy-tailed, then one can take $\alpha = 1$ to account for the imperfect assumption of Gaussian tails. See Example 2.4 for a demonstration of choosing α based on this approach.

3) One can also adopt the following approach that mimicks the cross-validation proce-

Algorithm 1 To compute a robust upper bound $\hat{x}_p$ for $\mathrm{VaR}_p(X)$

Given: N independent samples $X_1, \ldots, X_N$ of X, a level p close to 1, and a fixed choice $\alpha \geq 1$.

```
procedure Rob-Estimator($p, \alpha$)
    Initialize $n < N$, and let $m = \lfloor \frac{N}{n} \rfloor$.

    Step 1 (Compute block-maxima): Partition $X_1, \ldots, X_N$ into blocks of size $n$, and compute the block maxima for each block to obtain samples $M_{n,1}, \ldots, M_{n,m}$ of maxima $M_n$.

    Step 2 (Calibrate a reference GEV model): Treat the samples $M_{n,1}, \ldots, M_{n,m}$ as independent samples coming from a member of the GEV family and use a parameter estimation technique (for example, maximum-likelihood) to estimate the parameters $a_0, b_0$ and $\gamma_0$, along with suitable confidence intervals.

    Step 3 (Determine the family of candidate models): For chosen $\alpha \geq 1$, determine $\delta$ using a divergence estimation procedure (for an example, see Section 2.5.1). Then the set $\{P : D_\alpha(P, P_{GEV}) \leq \delta\}$ represents the family of candidate models.

    Step 4 (Compute the $p^n$-th quantile for the reference GEV model, and as well as the worst-case estimate over all candidate models):
    Solve for $x$ such that $G_{\gamma_0}(a_0 x + b_0) = p^n$, and let $x_p$ be the corresponding solution.
    Solve for $x > x_p$ in (2.20) and let the solution be $\hat{x}_p$.

    Return $x_p$ and $\hat{x}_p$
```

dure used in machine learning for choosing hyperparameters:

Recall that our objective is to estimate $F^{\leftarrow}(p)$ for some p close to 1. With this approach, we first estimate $F^{\leftarrow}(q)$ as a plug-in estimator from the empirical distribution, for some $q < p$; while it is desirable that q is closer to p, care should be taken in the choice that $F^{\leftarrow}(q)$ should be estimable from the given N samples with high confidence.

Having estimated $F^{\leftarrow}(q)$ directly from the empirical distribution, the idea now is to divide the given N samples, uniformly at random, into K mini-batches, each of which is independently input as samples to the procedure Rob-Estimator(q, α) in Algorithm 1 to yield K different robust estimates of $F^{\leftarrow}(q)$ for an initially chosen value of α (say, $\alpha = 1$). If the mini-batches are of size N/r, then it is reasonable to choose the scale-down factor r to be of the same order of magnitude as $(1-q)/(1-p)$.

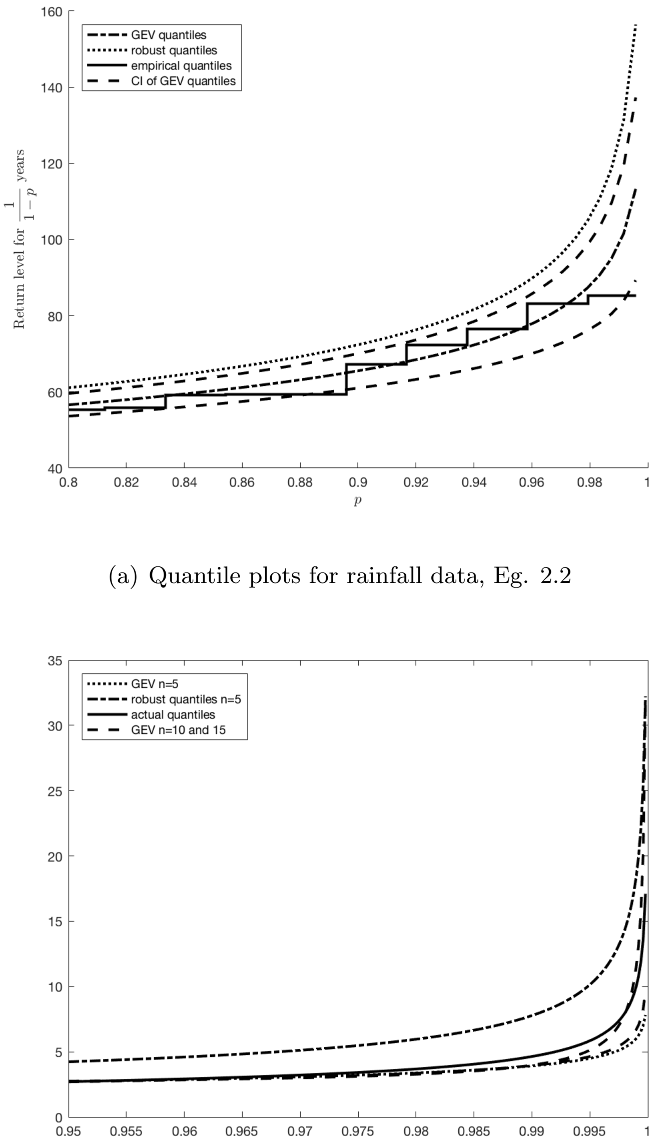
We repeat the above experiment for small increments of α to identify the largest value of α for which the robust estimates obtained from the K sub-problems still cover the plug-in estimate for $F^{\leftarrow}(q)$ obtained initially from the empirical distribution. We utilize this largest value of α that performs well in the scaled-down sub-problems to be the choice of α for robust estimation of $F^{\leftarrow}(p)$.

The third approach avoids using the upper end-point of a confidence interval of γ to pick α. Instead it incorporates a trade-off between the choice of α and δ. Estimating δ requires the estimation of the Rényi divergence, which is typically handled by k-NN methods as explained in Remark 2.3. Large values of α may be desirable because they generate better upper bounds, but since $\alpha \to D_\alpha$ is nondecreasing as mentioned in Section 2.3.1, it also requires large neighborhoods to include the true distribution and hence large values of δ. Further, by Theorem 2.5 if the true distribution has heavier tail than the chosen GEV model, then there does exist a threshold of α over which the neighborhoods will not include the true distribution or any other distributions with the same or more tail heaviness than the true distribution, regardless of how large δ is. Therefore when the chosen α is so large that the true distribution has the tail with an index greater than γ^*, any attempt to estimate such δ will be unstable and underestimated and causes the failure of coverage for true quantile. The above cross-validation-like procedure incorporates this trade-off and picks a suitable pair (α, δ). Example 2.5 gives the corresponding numerical experiments using this approach.

2.5.3 Numerical examples

Example 2.2. For a demonstration of the ideas introduced, we consider the rainfall accumulation data, due to the study of Coles and Tawn (1996), from a location in south-west England (see also Coles (2001) for further extreme value analysis with the dataset). Given annual maxima of daily rainfall accumulations over a period of 48 years (1914-1962), we attempt to compute, for example, the 100-year return level for the daily rainfall data. In other words, we aim to estimate the daily rainfall accumulation level that is exceeded about only once in 100 years. As a first step, we calibrate a GEV model for the annual maxima. Maximum-likelihood estimation of parameters results in the following values for shape, scale and location parameters: $\gamma_0 = 0.1072,\ a_0 = 9.7284$ and $b_0 = 40.7830$. The 100-year return level due to this model yields a point estimate 98.63mm with a standard error of ± 17.67mm (for 95% confidence interval). It is instructive to compare this with the corresponding estimate 106.3 ± 40.7mm obtained by fitting a generalized Pareto distribution (GPD) to the large exceedances (see Example 4.4.1 of Coles (2001)). To illustrate our methodology, we pick $\alpha = 2$, as suggested in (2.21). Next, we obtain $\delta = 0.05$ as an empirical estimate of divergence D_α between the data points representing annual maxima and the calibrated GEV model $P_{_{GEV}} = G_{\gamma_0}(a_0 x + b_0)$. This step is accomplished using a simple k-nearest neighbor estimator (see Póczos and Schneider (2011)). Consequently, the worst-case quantile estimate over all probability measures satisfying $D_\alpha(P, P_{_{GEV}}) \leq \delta$ is computed to be $F_\alpha^{\leftarrow}(1 - 1/100) = 132.24$mm. While not being overly conservative, this worst-case 100 year return level of 132.44mm also acts as an upper bound to estimates obtained due to different modelling assumptions (GEV vs GPD assumptions). To demonstrate the quality of estimates throughout the tail, we plot the return levels for every $1/(1-p)$ years, for values of p close to 1, in Figure 2.2(a). While the return levels predicted by the GEV reference model is plotted in solid line (with the dash-dot lines representing 95% confidence intervals), the dotted curve represents the worst-case estimates $F_\alpha^{\leftarrow}(p)$. The empirical quantiles are drawn in the dashed line.

Figure 2.2: Plots for Examples 2.2 and 2.3



(a) Quantile plots for rainfall data, Eg. 2.2

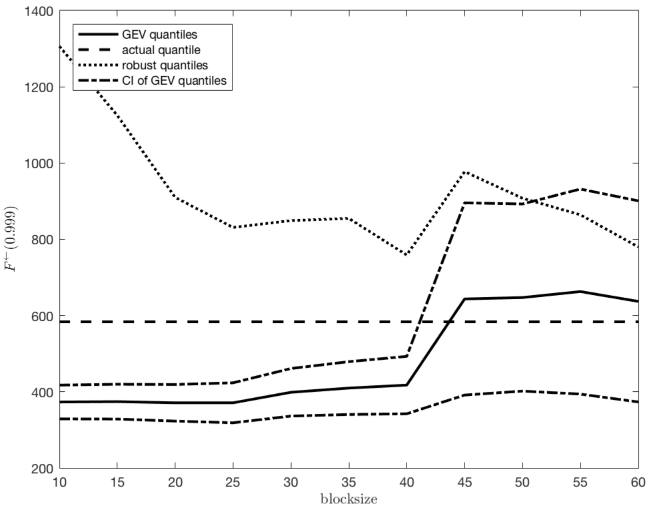
(b) Quantile plots for Pareto data, Eg. 2.3

Example 2.3. In this example, we are provided with 100 independent samples of a Pareto random variable satisfying $P\{X > x\} = 1 - F(x) = 1 \wedge x^{-3}$. As before, the objective is to compute quantiles $F^{\leftarrow}(p)$ for values of p close to 1. As the entire probability distribution is known beforehand, this offers an opportunity to compare the quantile estimates returned by our algorithm with the actual quantiles. Unlike Example 2.2, the data in this example does not present a natural means to choose block sizes. As a first choice, we choose block size $n = 5$ and perform routine computations as in Algorithm 1 to obtain a reference GEV model $P_{_{GEV}}$ with parameters $\gamma_0 = 0.11, a_0 = 0.58, b_0 = 1.88$, and corresponding tolerance parameters $\alpha = 1.5$ and $\delta = 0.8$. Then the worst-case quantile estimate $F_\alpha^{\leftarrow}(p^n) = \sup\{G^{\leftarrow}(p^n) : D_\alpha(G, P_{_{GEV}}) \leq \delta\}$ is immediately calculated for various values of p close to 1, and the result is plotted (in the dotted line) against the true quantiles $F^{\leftarrow}(p) = (1-p)^{-1/3}$ (in the solid line) in Figure 2.2(b). These can, in turn, be compared with the quantile estimates x_p (in the solid line) due to traditional GEV extrapolation with reference model $P_{_{GEV}}$. Recall that the initial choice for block size, $n = 5$, was arbitrary. One can perhaps choose a different block size, which will result in a different model for corresponding block-maximum M_n. For example, if we choose $n = 10$, the respective GEV model for M_{10} has parameters $\gamma_0 = 0.22, a_0 = 0.55$ and $b_0 = 2.3$. Whereas, if we choose $n = 15$, the GEV model for M_{15} has parameters $\gamma_0 = 0.72, a_0 = 0.32$ and $b_0 = 2.66$. When considering the shape parameters, these models are different, and subsequently, the corresponding quantile estimates (plotted using dashed lines in Figure 2.2(b)) are also different. However, as it can be inferred from Figure 2.2(b), the robust quantile estimates (in the dotted line) obtained by running Algorithm 1 forms a good upper bound to the actual quantiles $F^{\leftarrow}(p)$, as well as to the quantile estimates due to different GEV extrapolations from different block sizes $n = 10$ and 15.

Example 2.4. The objective of this example is to demonstrate the applicability of Algorithm 1 in an instance where the traditional extrapolation techniques tend to not yield stable estimates. For this purpose, we use N = 2000 independent samples of the random variable $Y = X + 50\mathbf{1}(X > 5)$ as input to the maximum likelihood based GEV model estimation, with the aim of calculating the extreme quantile $F^{\leftarrow}(0.999)$. Here, F denotes the distribution function of random variable Y, and X is a Pareto random variable with distribution $\max(1 - x^{-1.1}, 0)$. The quantile estimates (and the corresponding 95% confidence intervals) output by this traditional GEV estimation procedure, for various choices of block sizes, is displayed with the solid line in Figure 2.3. Even for modestly large block size choices, it can be observed that the 95% confidence regions obtained from the calibrated

GEV models are far below the true quantile drawn in the dashed line. This underestimation is perhaps because of the sudden shift of samples of block-maxima M_n from a value less than 5 to a value larger than 55 (recall that the distribution F assigns zero probability to the interval $(5, 55)$).

Figure 2.3: Plot for Example 2.4, instability in estimated quantile $F^{\leftarrow}(0.999)$



Next, we use Algorithm 1 to yield an upper bound that is robust against model errors. Unlike previous examples where standard errors are used to calculate the suitable α, in this example, we use the domain knowledge that the samples of Y have finite mean, which means, $\gamma^* \leq 1$. Assuming no additional information, we resort to the conservative choice $\gamma^* = 1$. The dashed curve in Figure 2.3 corresponds to the upper bound on $F^{\leftarrow}(0.999)$ output by Algorithm 1. We note the following observations: First, the worst case estimates output by Algorithm 1 indeed act as an upper bound for the true quantile (drawn in solid line), irrespective of the block-size chosen and the baseline GEV model used. Second, for block-sizes smaller than $n = 45$, it appears that the calibrated baseline GEV models are not representative enough of the distribution of M_n, and hence higher the value of δ for these choices of block sizes. Understandably, this results in a conservative worst case estimate for the smaller choices of block sizes. However, we argue that the overall procedure is not discouragingly conservative, by observing that the spread of 95% confidence region for block size choices $n = 50$ to 60 (where the traditional GEV calibration appears correct) is comparable to the difference between the true quantile and the worst-case estimate produced by Algorithm 1 for majority of block size choices (from $n = 20$ to 60).

Example 2.5. *In this example we consider the St. Petersburg distribution, which is not in the maximum domain of attraction of any GEV distribution (see e.g. Fukker* et al. *(2016)). Recall that X is St.Petersburg distributed if*

$$P\{X = 2^k\} = 2^{-k}, \quad k = 1, 2, \ldots \tag{2.22}$$

Note that the St. Petersburg distribution takes large values with tiny probability. Let B denote a Bernoulli random variable with parameter $1/5$. In addition let W be exponentially distributed with mean 8 and define $Z = B \cdot X + W$. Suppose we have 5000 data points from the distribution of Z. Similar to the previous example, we want to estimate its quantile $F^{\leftarrow}(0.999)$.

Here we demonstrate another approach to choose the parameter α. The idea, as described earlier in Item 3) is to first choose a tail probability level q for which $F^{\leftarrow}(q)$ can be accurately estimated from the whole data set. For our example, we take $q = 0.99$ and compute the plug-in estimate $F^{\leftarrow}(q)$ from the empirical distribution. Then we independently divide the given data set uniformly at random into 10 batches each of size 625 samples (corresponding to a scale-down factor = 8). We employ the procedure Rob-Estimator(q, α) *for various values of α on each of these 10 sub-sampled mini-batches independently, and choose the largest value of α such that the robust estimates from each of the 10 sub-samples cover the earlier plug-in estimate $F^{\leftarrow}(0.99)$. The specific details for this example are as follows:*

1) The plug-in estimate for $F^{\leftarrow}(0.99)$ from the given 5000 samples is 44.9. Note that with 5000 samples, this estimate from empirical distribution is with reasonably high confidence.

2) Resample the data into 10 mini-batches of size 5000/8 = 625 samples. With blocksize = 20 we utilize the procedure Rob-Estimator$(0.99, \alpha)$ *on each of the 10 mini-batches to choose the largest α such that the respective robust estimates from all the 10 sub-sampled mini-batches cover the empirical estimate of $F^{\leftarrow}(0.99)$ obtained from step 1). This approach leads us to the choice of $\alpha = 4.47$. Computing block maxima from blocks of samples with size = 48, the subsequent robust upper bound from the procedure* Rob-Estimator$(0.999, 4.47)$

turns out to be 652.90, *which covers the true quantile,* $F^{\leftarrow}(0.999) = 268.27$. *In contrast, the* 95%*-confidence interval of GEV estimate is* $[93.81, 201.60]$, *which fails to cover the true quantile.*

This approach incorporates the trade-off between the choice of α *and* δ. *Large values of* α *may be desirable because they generate less conservative upper bounds. But Step 2) avoids picking too large values of* α, *because too large values of* α, *combined with the corresponding estimators for* δ *empirically do not lead to good coverage for* $F^{\leftarrow}(0.99)$. *Therefore this cross-validation-like procedure automatically incorporates the trade-off between the choice of hyperparameters* α *and* δ.

2.6 Proofs of main results

In this section, we provide proofs of Theorems 2.4 and 2.5, along with proofs of Propositions 2.3, 2.6 and 2.7.

Proof of Proposition 2.3

By definition, $F_{\alpha,\delta}(x)$ is non-decreasing in x. Since $F_{\alpha,\delta}(x) \leq P_{ref}(-\infty, x)$, we have $\lim_{x\to-\infty} F_{\alpha,\delta}(x) = 0$. In addition, we have from Example 2.1 that $\bar{F}_{\alpha,\delta}(x) = \theta_x P_{ref}(x,\infty)$, where θ_x satisfies (2.17). Since $P_{ref}(x,\infty)\phi_\alpha(\theta_x) \leq \bar{\delta}$ (follows from (2.17)), we have $\theta_x \leq \phi_\alpha^{-1}(\bar{\delta}/P_{ref}(x,\infty))$, where $\phi_\alpha^{-1}(\cdot)$ is the inverse function of $\phi_\alpha(\cdot)$ (recall the defintion of $\phi_\alpha(\cdot)$ in (2.13) to see that the inverse is well-defined for every $\alpha \geq 1$). As a result,

$$\bar{F}_{\alpha,\delta}(x) \leq \phi_\alpha^{-1}\left(\frac{\bar{\delta}}{P_{ref}(x,\infty)}\right) P_{ref}(x,\infty). \tag{2.23}$$

If we let $W(x)$ denote the product log function[4], then $\phi_\alpha^{-1}(u) = u^{-1/\alpha}$ when $\alpha > 1$ and $\phi_\alpha^{-1}(u) = u/W(u)$ when $\alpha = 1$. Consequently for any $\alpha \geq 1$, $\epsilon\phi_\alpha^{-1}(1/\epsilon) \to 0$ as $\epsilon \to 0$. As a result, $\lim_{x\to\infty} \bar{F}_{\alpha,\delta}(x) = 0$ for any choice of $\alpha \geq 1$ and $\delta > 0$. Thus $\lim_{x\to\infty} F_{\alpha,\delta}(x) = 1$.

[4] W is the inverse function of $f(x) = xe^x$

To show that $F_{\alpha,\delta}(x)$ is right-continuous, we first see that

$$\begin{aligned} F_{\alpha,\delta}(x+\epsilon) - F_{\alpha,\delta}(x) &= \sup_{P:D_\alpha(P,P_{ref})\leq\delta} P(x,\infty) - \sup_{P:D_\alpha(P,P_{ref})\leq\delta} P(x+\epsilon,\infty) \\ &\leq \sup_{P:D_\alpha(P,P_{ref})\leq\delta} P(x,x+\epsilon], \end{aligned}$$

for any $\epsilon > 0$, for every choice of $\delta > 0, \alpha \geq 1$ and P_{ref}. Following the same reasoning as in (2.23), we obtain that

$$\sup_{P:D_\alpha(P,P_{ref})\leq\delta} P(x,x+\epsilon] \leq \phi_\alpha^{-1}\left(\frac{\bar{\delta}}{P_{ref}(x,x+\epsilon]}\right) P_{ref}(x,x+\epsilon],$$

for which the right hand side vanishes when $\epsilon \to 0$. As a result, $F_{\alpha,\delta}(x)$ is right-continuous as well, thus verifying all the properties required to prove that $F_{\alpha,\delta}(\cdot)$ is a cumulative distribution function. □

Proof of Theorem 2.5

Our goal is to determine the maximum domain of attraction of $\bar{F}_{\alpha,\delta}(x) = \sup\{P(x,\infty) : D_\alpha(P,P_{_{ref}}) \leq \delta\}$. We already have an upper bound for $\bar{F}_{\alpha,\delta}(x)$ in (2.23) in the proof of Proposition 2.3. To obtain a lower bound for $\bar{F}_{\alpha,\delta}(x)$, first consider a probability measure Q defined by

$$\frac{dQ}{dP_{_{ref}}}(x) = \phi_\alpha^{-1}\left(\frac{c}{P_{_{ref}}(x,\infty)(1-\log P_{_{ref}}(x,\infty))^2}\right),$$

for a suitable positive constant c. Then $D_\alpha(Q,P_{_{ref}}) < \infty$ because of a simple change of variables $u = P_{_{ref}}(x,\infty)$ in the integration

$$\int \phi_\alpha\left(\frac{dQ}{dP_{_{ref}}}\right) dP_{_{ref}} = \int_0^1 \frac{c}{u(1-\log u)^2} du < \infty.$$

Consequently, due to a continuity argument, one can demonstrate a constant $a \in (0,1)$ such that $D_\alpha(aQ + (1-a)P_{_{ref}}, P_{_{ref}}) \leq \delta$. Then, it follows from the definition of $\bar{F}_{\alpha,\delta}(x)$ that

$$\bar{F}_{\alpha,\delta}(x) \geq \big(aQ + (1-a)P_{_{ref}}\big)(x,\infty) = \int_x^\infty \left(a\frac{dQ}{dP_{_{ref}}}(t) + 1 - a\right) P_{_{ref}}(dt)$$

Since $dQ/dP_{_{ref}}(t)$ is eventually increasing, as $t \to \infty$, we have that,

$$\bar{F}_{\alpha,\delta}(x) \geq a\phi_\alpha^{-1}\left(\frac{c}{P_{_{ref}}(x,\infty)(1-\log P_{_{ref}}(x,\infty))^2}\right) P_{_{ref}}(x,\infty),$$

for sufficiently large values of x. For brevity, let

$$A(x) := P_{ref}(x,\infty),\; g(x) := a\phi_\alpha^{-1}(c(1-\log x)^{-2}/x) \text{ and } h(x) := \phi_\alpha^{-1}(\bar{\delta}/x).$$

Then, combining the above lower bound with the upper bound in (2.23), we obtain

$$\bar{F}_{low}(x) := g(A(x))A(x) \leq \bar{F}_{\alpha,\delta}(x) \leq h(A(x))A(x) =: \bar{F}_{up}(x), \tag{2.24}$$

for large values of x. Recall that the reference measure P_{ref} belongs to the maximum domain of attraction of $G_{\gamma_{\text{ref}}}$. The following lemma characterizes the extreme value distributions corresponding to the upper and lower bounds $\bar{F}_{up}$ and $\bar{F}_{low}$.

Lemma 2.2. *Suppose that the quantity* $\gamma^* = \frac{\alpha}{\alpha-1}\gamma_{ref}$ *is well-defined. Additionally, let* $x^* = \sup\{x : A(x) > 0\}$. *Then the following are true:*

$$(a)\; \lim_{x\uparrow x^*} -\left(\frac{\bar{F}_{up}}{\bar{F}'_{up}}\right)'(x) = \gamma^*, \text{ and } (b)\; \lim_{x\uparrow x^*} -\left(\frac{\bar{F}_{low}}{\bar{F}'_{low}}\right)'(x) = -\gamma^*.$$

As a consequence of Proposition 2.1 and Lemma 2.2, if γ^* is finite, both $\bar{F}_{low}$ and $\bar{F}_{up}$ lie in the maximum domain of attraction of G_{γ^*}. As $\bar{F}_{\alpha,\delta}(x)$ is sandwiched between $\bar{F}_{low}(x)$ and $\bar{F}_{up}(x)$ as in (2.24), if at all $\bar{F}_{\alpha,\delta}$ belongs to the maximum domain of attraction of G_γ for some $\gamma \in \mathbb{R}$, then γ must equal γ^*. Since $\bar{F}_{\alpha,\delta}(x) \sim \bar{F}_{\alpha,\delta}(x^-)$ as $x \uparrow x^*$, due to Theorem 1.7.13 of Leadbetter *et al.* (1983), this is indeed the case. Therefore, the α-family worst-case tail distribution $\bar{F}_{\alpha,\delta}$ belongs to the maximum domain of attraction of G_{γ^*}. □

Proof of Lemma 2.2(a). Recall that $\bar{F}_{up}(x) = h(A(x))A(x)$. By repeatedly applying elementary rules of differentiation, it is obtained that

$$\begin{aligned} -\left(\frac{\bar{F}_{up}}{\bar{F}'_{up}}\right)'(x) = &-\left(\frac{A}{A'}\right)'(x)\left(1+\frac{A(x)h'(A(x))}{h(A(x))}\right)^{-1} + \\ &+\left(\frac{A(x)h'(A(x))}{h(A(x))} + A^2(x)\left(\frac{h'}{h}\right)'(A(x))\right)\left(1+\frac{A(x)h'(A(x))}{h(A(x))}\right)^{-2} \end{aligned} \tag{2.25}$$

Case $\alpha > 1$: Since $h(x) = (\bar{\delta}/x)^{1/\alpha}$ and $h'(x)/h(x) = -(\alpha x)^{-1}$, we obtain

$$-\left(\frac{\bar{F}_{up}}{\bar{F}'_{up}}\right)'(x) = -\left(\frac{A}{A'}\right)'(x)\left(1-\frac{1}{\alpha}\right)^{-1} + \left(-\frac{1}{\alpha}+\frac{1}{\alpha}\right)\left(1-\frac{1}{\alpha}\right)^{-2}.$$

In addition, as required in the statement of Theorem 2.5, $A(x) := P_{ref}(x,\infty)$ satisfies $-(A/A')'(x) \to \gamma_{ref}$, as x approaches its right endpoint $x^* = \sup\{x : A(x) > 0\}$. Therefore,

$$\lim_{x \uparrow x^*} -\left(\frac{\bar{F}_{up}}{\bar{F}'_{up}}\right)'(x) = \frac{\alpha}{\alpha - 1} \lim_{x \uparrow x^*} \left[-\left(\frac{A}{A'}\right)'(x)\right] = \frac{\alpha}{\alpha-1}\gamma_{ref}.$$

Case $\alpha = 1$: When α equals 1, $\phi_\alpha^{-1}(x) = x/W(x)$, where $W(x)$ is the product log function. Then the following calculations are simply algebraic:

$$\frac{xh'(x)}{h(x)} = -\left(1 + \frac{1}{W\left(\frac{\bar{\delta}}{x}\right)}\right)^{-1}$$

and

$$x^2\left(\frac{h'}{h}\right)'(x) = \left[1 + \left(1 + W\left(\frac{\bar{\delta}}{x}\right)\right)^{-1}\right]\left(1 + \frac{1}{W\left(\frac{\bar{\delta}}{x}\right)}\right)^{-2}.$$

Substituting these in (2.25), we obtain

$$-\left(\frac{\bar{F}_{up}}{\bar{F}'_{up}}\right)'(x) = \left[-\left(\frac{A}{A'}\right)'(x) W\left(\frac{\bar{\delta}}{A(x)}\right) - 1\right]\left(1 + \frac{1}{W\left(\frac{\bar{\delta}}{A(x)}\right)}\right)^{-1}. \qquad (2.26)$$

Recall that $-(A/A')'(x)$ converges to γ_{ref}, as $x \uparrow x^*$. Letting $x \to x^*$ in the above expression, we obtain

$$-\left(\frac{\bar{F}_{up}}{\bar{F}'_{up}}\right)'(x) = \begin{cases} \infty, & \text{if } \gamma_{ref} > 0, \\ -\infty, & \text{if } \gamma_{ref} < 0, \end{cases}$$

which indeed equals $\frac{\alpha}{\alpha-1}\gamma_{ref}$. This completes the proof of Part (a) of Lemma 2.2. □

Proof of Lemma 2.2(b). First, an expression for $(\bar{F}_{low}/\bar{F}'_{low})'$ similar to (2.25) can be obtained by simply substituting g in place of h in (2.25). Again, the cases $\alpha > 1$ and $\alpha = 1$ are calculated separately:

Case $\alpha > 1$: When $\alpha > 1, \phi_\alpha^{-1}(x) = x^{1/\alpha}$. By applying elementary rules of differentiation, we obtain

$$\frac{xg'(x)}{g(x)} = \frac{1}{\alpha}\frac{1 + \log x}{1 - \log x} \quad \text{and} \quad x^2\left(\frac{g'}{g}\right)'(x) = \frac{1}{\alpha}\frac{1 + \log^2 x}{(1 - \log x)^2}.$$

Letting $x \uparrow x^*$, we obtain $A(x)g'(A(x))/g(A(x)) \to -1/\alpha$ and $A(x)^2(g'/g)'(A(x)) \to 1/\alpha$. Subsequently,

$$\lim_{x\uparrow x^*}\left(\frac{\bar{F}_{low}}{\bar{F}'_{low}}\right)'(x) = \left(1-\frac{1}{\alpha}\right)^{-1}\lim_{x\uparrow x^*}\left(\frac{A}{A'}\right)'(x) + \left(-\frac{1}{\alpha}+\frac{1}{\alpha}\right)\left(1-\frac{1}{\alpha}\right)^2,$$

which equals $\frac{\alpha}{\alpha-1}\gamma_{ref}$, as in the proof of Part (a) of Lemma 2.2. The case $\alpha = 1$ is similar to that of proof of Part (a), but more tedious, and is not presented here in the interest of space and readability. □

Proof of Theorem 2.4

Theorem 2.4 follows as a simple corollary of Theorem 2.5, once we verify that any GEV model $G(x) := P_{GEV}(-\infty, x)$ satisfies $G'(x) > 0$ and $G''(x)$ exists in a left neighborhood of $x^*_G = \sup\{x : G(x) < 1\}$, along with the property that

$$\lim_{x\uparrow x^*_G}\left(\frac{1-G}{G'}\right)'(x) = \gamma_{ref},$$

where γ_{ref} is the shape parameter of G. Such a GEV model satisfies $G(x) = G_{\gamma_{\text{ref}}}(ax+b)$ for some scaling and translation constants a and b. Therefore, it is enough to verify these properties only for $G(x) = G_{\gamma_{\text{ref}}}(x)$. Once we recall the definition of G_γ in (2.4), the desired properties are elementary exercises in calculus. □

Proof of Proposition 2.6

First, we derive a lower bound for $\bar{F}_1(x) = \sup\{P(x,\infty) : D_1(P, G_0) \le \delta\}$. Consider the probability density function $f(x) = c(x\log x)^{-2}\mathbf{1}(x \ge 2)$, where c is a normalizing constant that makes $\int f(x)dx = 1$. In addition, let $g(x) = G_0'(x)$ denote the probability density function corresponding to the distribution G_0. Clearly,

$$\begin{aligned} D_1(f,g) &= \int f(x)\log\left(\frac{f(x)}{g(x)}\right)dx \\ &= c\int_2^\infty (x\log x)^{-2}\log\left(\frac{c(x\log x)^{-2}}{\exp(-\exp(-x)\exp(-x))}\right)dx \\ &\le \int_2^\infty \frac{x+\exp(-x)+\log c}{x^2\log^2 x}dx < \infty. \end{aligned}$$

Now, as in the proof of Theorem 2.5, consider a family of densities $\{af + (1-a)G_0' : a \in (0,1)\}$. Due to the continuity of $D_1(af + (1-a)G_0', G_0')$ with respect to a, there exists an $\bar{a} \in (0,1)$ such that $D_1(\bar{a}f + (1-\bar{a})G_0', G_0') \leq \delta$. Then, according to the definition of $\bar{F}_1$,

$$\begin{aligned}\bar{F}_1(x) &\geq \int_x^\infty \left(\bar{a}f + (1-\bar{a})G_0'\right)(u)du \\ &\geq \bar{a}\int_x^\infty \frac{c}{u^2\log^2 u}du = \frac{\bar{a}c + o(1)}{x\log^2 x},\end{aligned}$$

as $x \to \infty$. The asymptotic equivalence used above is due to Karamata's theorem (see Theorem 1 in Chapter VIII.9 of Feller (1966)). Combining this lower bound with the upper bound in (2.23), we obtain, for large enough x,

$$\frac{\bar{a}c}{2x\log^2 x} \leq \bar{F}_1(x) \leq h\big(1 - G_0(x)\big)\big(1 - G_0(x)\big),$$

where $h(x) = \phi_\alpha^{-1}(\bar{\delta}/x)$. For convenience, let us write $\bar{F}_{up}(x) := h\big(1-G_0(x)\big)\big(1-G_0(x)\big)$ and $\bar{F}_{low}(x) := \bar{a}c/(2x\log^2 x)$. Due to the characterization in (2.5), we have that $\bar{F}_{low} \in \mathcal{D}(G_1)$. On the other hand, following the lines of Proof of Lemma 2.2(a), from (2.26), we obtain that

$$-\left(\frac{\bar{F}_{up}}{\bar{F}_{up}'}\right)'(x) = \left[\left(\frac{1-G_0}{G_0'}\right)'(x) W\left(\frac{\bar{\delta}}{1-G_0(x)}\right) + 1\right]\left(1 + \frac{1}{W\left(\frac{\bar{\delta}}{1-G_0(x)}\right)}\right)^{-1}.$$

Since $G_0(x) = \exp(-e^{-x})$, we obtain

$$\left(\frac{1-G_0}{G_0'}\right)'(x) = e^{e^{-x}}\left(e^x\left(1 - e^{-e^{-x}}\right) - 1\right) = \frac{e^{-x}}{2}(1 + o(1)),$$

as $x \to \infty$. Therefore,

$$-\left(\frac{\bar{F}_{up}}{\bar{F}_{up}'}\right)'(x) \sim \frac{e^{-x}}{2}(1+o(1))W\left(\frac{\bar{\delta}}{e^{-x}(1+o(1))}\right) + 1$$

as $x \to \infty$. Since $tW(1/t) \to 0$ as $t \to 0$, it follows that $-(\bar{F}_{up}/\bar{F}_{up}')(x)$ converges to 1 as $x \to \infty$. Then, due to Proposition 2.1, we have that $\bar{F}_{up}$ also belong to the maximum domain of attraction of G_1. Since both $\bar{F}_{low}$ and $\bar{F}_{up}$ lie in the maximum domain of attraction of G_1, following the same line of reasoning as in the proof of Theorem 2.5, we obtain that $\bar{F}_1(x) \in \mathcal{D}(G_1)$. □

Proof of Proposition 2.7

First, let us consider the case $\gamma_{ref} \neq 0$: Recall the probability measure $aQ + (1-a)P_{ref}$ exhibited for establishing the lower bound in the proof of Theorem 2.5. For proving Proposition 2.7, we take the reference measure P_{ref} as $G_{\gamma_{\text{ref}}}$. Further, if we let $g(t) = a\phi_1^{-1}(c(1-\log t)^{-2}/t)$ and $A(x) := 1 - G_{\gamma_{\text{ref}}}(x)$, then as in the proof of Theorem 2.5, the measure $P := aQ + (1-a)P_{ref}$

1) satisfies $D_1(P, G_{\gamma_{\text{ref}}}) \leq \delta$, and

2) admits a lower bound $P(x,\infty) \geq g(A(x))A(x)$.

To proceed further, observe that $A(x) = 1 - G_{\gamma_{\text{ref}}}(x) \geq \bar{c}(1+\gamma_{ref}x)^{-1/\gamma_{ref}}$ for some constant $\bar{c} < 1$ and all x close enough to the right endpoint $x_G^* := \sup\{x : G_{\gamma_{\text{ref}}}(x) < 1\}$. In addition, $tg(t)$ strictly decreases to 0 as t decreases to 0. Therefore, for all x close to the right endpoint $x_G^* := \sup\{x : G_{\gamma_{\text{ref}}}(x) < 1\}$, it follows that

$$P(x,\infty) \geq g\left(\bar{c}(1+\gamma_{ref}x)^{-1/\gamma_{ref}}\right)\bar{c}(1+\gamma_{ref}x)^{-1/\gamma_{ref}}.$$

Since $\phi_1^{-1}(u) \geq u/\log u$ for large enough u, $g(t) \geq act^{-1}\left(1-\log t\right)^{-2}\log^{-1}\left(c/t\right)$, for all t close to 0. As a result, there exists a constant c' such that $tg(t) \geq c'(1-\log t)^{-3}$ for all t sufficiently close to 0. This allows us to write

$$P(x,\infty) \geq c'(1-\log(\bar{c}(1+\gamma_{ref}x)^{-1/\gamma_{ref}}))^{-3} = c'(1+\log(\bar{c}^{1/\gamma_{ref}}(1+\gamma_{ref}x))/\gamma_{ref})^{-3},$$

for x sufficiently close x_G^*, thus verifying the statement in cases (a) and (b) where $\gamma_{ref} \neq 0$. When $\gamma_{ref} = 0$, see the proof of Proposition 2.6 where we exhibit a measure P such that $D_1(P, G_0) \leq \delta$ and $P(x,\infty) = O\left(x^{-1}\log^{-2}x\right)$. This completes the proof.

Chapter 3

Dependence with two sources of uncertainty: Computing Worst-case Expectations Given Marginals via Simulation

We focus on the problem of computing lower and upper bounds among any dependence structure for a function of two random vectors whose marginal distributions are assumed to be known. This problem is motivated from several applications in risk quantification and statistics. Before discussing its applications, let us first describe it precisely.

Suppose that $X \in \mathbb{R}^d$ follows distribution μ and $Y \in \mathbb{R}^l$ follows distribution ν. We define $\Pi\left(\mu,\nu\right)$ to be the set of joint distributions π in $\mathbb{R}^{d\times l}$ such that the marginal of the first d entries coincides with μ and the marginal of the last l entries coincides with ν. In other words, for any probability measure π in $\mathbb{R}^{d\times l}$ (endowed with the Borel σ-field), if we let $\pi_X\left(A\right) = \pi\left(A\times\mathbb{R}^l\right)$ for any Borel measurable set $A \in \mathbb{R}^d$, and $\pi_Y\left(B\right) = \pi\left(\mathbb{R}^d\times B\right)$ for any Borel measurable set $B \in \mathbb{R}^l$, then $\pi \in \Pi\left(\mu,\nu\right)$ if and only if $\pi_X = \mu$ and $\pi_Y = \nu$. We are interested in the quantity (focusing on minimization)

$$V = \min\{\mathbb{E}_\pi\left[c\left(X,Y\right)\right] : \pi \in \Pi\left(\mu,\nu\right)\} \tag{3.1}$$

where $c(\cdot,\cdot) \in \mathbb{R}$ is some cost function. Formulation (3.1) is well-defined as the class $\Pi\left(\mu,\nu\right)$

is non-empty, because the product measure $\pi = \mu \times \nu$ belongs to $\Pi(\mu, \nu)$.

In operations research contexts, problem (3.1) arises as a means to obtain bounds for performance measures in situations where dependence information is ambiguous. Such situations occur because, in practice, accurately estimating the marginal distributions of random variables is often relatively easy, e.g., by goodness-of-fit against well-chosen parametric distributions. They also occur in scenarios where data from different stochastic sources are collected independently (i.e., rather than in pairs), in which case no dependence information between these sources can be inferred. Indeed, special (i.e., discrete) cases of (3.1) have been analyzed in the distributionally robust optimization literature (e.g., Doan *et al.* (2015)). Variants of (3.1) to risk measures have also been studied, regarding both algorithmic approaches (e.g., Rüschendorf (1983), Embrechts *et al.* (2013)) and sharp bounds over specific geometric classes of marginals (e.g., Wang and Wang (2011),Puccetti (2013),Puccetti and Rüschendorf (2013)).

In statistics and machine learning contexts, the value of (3.1) is the Wasserstein distance (of order 1) between X and Y when $c(\cdot, \cdot)$ is taken as a metric. The optimization can be viewed as the classical Kantorovich relaxation to Monge's problem in optimal transport (e.g., Rachev and Rüschendorf, Villani (1998, 2008)), where solutions based on differential properties have been extensively studied. Wasserstein distance is of central importance in probabilistic analysis (e.g., quantifying model discrepancies in Bayesian settings Minsker *et al.* (2014) and convergence rates of ergodic processes Boissard and Le Gouic (2014), among many others). The estimation of the distance itself is also suggested as a tool for statistical inference, including the use in goodness-of-fit tests Del Barrio *et al.* (1999),Del Barrio *et al.* (2005) and in applications such as image recognition Sommerfeld and Munk (2016). It has also been used to quantify model uncertainty in stochastic optimization problems (e.g., Esfahani and Kuhn (2015),Blanchet and Kang (2016),Blanchet and Murthy (2016),Gao and Kleywegt (2016)) and in the application of distributionally robust optimization in machine learning settings Blanchet *et al.* (2016b). As such, there have been growing studies on the convergence behaviors of its empirical estimation. Central limit theorems (CLTs) on the empirical estimation of (3.1), based on representations using quantile functions, have been investigated in the one-dimensional case (e.g., Bobkov and Ledoux (2014), Del Barrio *et al.*

(1999)). More generally, concentration bounds have been studied in the line of work including Horowitz and Karandikar (1994), Bolley *et al.* (2007), Boissard (2011), Sriperumbudur *et al.* (2012), Trillos and Slepčev (2014) and Fournier and Guillin (2015), so do laws of large numbers in some special cases (e.g., Dobrić and Yukich (1995)).

Since classical methods for solving (1), based for instance on Euler-Lagrange equations, may not yield straightforward computational schemes in general, we resort to Monte Carlo for an easy-to-implement approximation. Our contribution is precisely to quantify the rate of convergence of such Monte Carlo schemes. Our results also add to the literature of empirical Wasserstein estimation when these Monte Carlo samples are viewed as data. We focus on the setting where one of the marginals, say Y, is a finite-support distribution, and another, say X, is a multi-dimensional distribution that can be continuous. To approximate V, we consider the drawn samples from the continuous variable X, and replace the infinite-dimensional linear program (LP) in (3.1) by its sampled counterpart, which can be solved by standard LP solvers.

Our main result shows that the error of our procedure is $O(n^{-1/2})$ where n is the sample size, independent of the dimension d or l. We also identify the limiting distribution in the associated CLT. The closest work to our results, as far as we know, is the recent work of Sommerfeld and Munk (2016), who derive a CLT when both marginal distributions are finitely discrete. Our result here can be viewed as a generalization to theirs when one of the distributions is continuous. We remark that our obtained rate differs from the typical rate of $O(n^{-1/d})$ in high-dimensional empirical Wasserstein estimation where $d \geq 3$ is the dimension of the marginal distributions. As we will see, the finite-support property of one of the marginals plays a crucial role in applying classical results in sample average approximation (SAA) that maintain the standard Monte Carlo rate in our scheme.

In the rest of this paper, we will first describe our algorithm, followed by our main results on the convergence analysis.

3.1 Algorithmic Description

Suppose that the distribution ν for Y has finite support $\{y_1, ...y_m\} \subset \mathbb{R}^l$. Supposing that X can be simulated, we sample n i.i.d. observations $X_1, \ldots, X_n$ from μ, and approximate V by

$$V_n = \min\{\mathbb{E}_\pi \left[c\left(X, Y\right)\right] : \pi \in \Pi\left(\mu_n, \nu\right)\} \tag{3.2}$$

where μ_n is the empirical distribution of X constructed from the X_i's, i.e.,

$$\mu_n(A) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \in A)$$

for any Borel measurable A.

Note that (3.2) is a finite-dimensional LP, which can be written more explicitly as

$$\begin{array}{ll} \min & \sum_{i=1}^{n}\sum_{j=1}^{m} c(X_i, y_j)p_{ij} \\ \text{subject to} & \sum_{j=1}^{m} p_{ij} = \frac{1}{n} \ \ \forall i = 1, \ldots, n \\ & \sum_{i=1}^{n} p_{ij} = \nu\{y_j\} \ \ \forall j = 1, \ldots, m \\ & p_{ij} \geq 0 \ \ \forall i = 1, \ldots, n, \ j = 1, \ldots, m \end{array} \tag{3.3}$$

where the decision variables p_{ij} represent the probability masses on (X_i, y_j), and $\nu\{y_j\}$ denotes the mass on y_j under ν. Problem (3.3) is an assignment problem, which is a special type of minimum cost problem and can be solved by, e.g., successive shortest path algorithms in polynomial time of order $O(n^2m + n(n+m)\log(n+m))$ (see, e.g., R.K.Ahuja *et al.* (2000) pp. 471, 500).

3.2 Convergence Analysis

Our main result is a convergence analysis on V_n to V. We impose the assumptions:

Assumption 1. *For each y_j, $c(., y_j)$ is non-negative and lower semicontinuous.*

Assumption 2. *Suppose that ν has finite support $\{y_1, ..., y_m\} \subset \mathbb{R}^l$. We have*

$$\mathbb{E}_\mu[c(X, y_j)^2] < \infty, \ \forall j = 1, \ldots, m.$$

Denote

$$V' = \max_{\beta_1,\ldots,\beta_m \in \mathbb{R}} \mathbb{E}_\mu \Big[\min_{j=1,\ldots,m} \{ c\,(X, y_j) - \beta_j \} + \sum_{j=1}^{m} \beta_j \nu\{y_j\} \Big] \tag{3.4}$$

which is the dual problem of (3.1) (see Lemma 3.1 for an explanation in the special case of finite-dimensional settings). Under Assumptions 1 and 2, strong duality (known as the Kantorovich duality) holds and $V' = V$; see, e.g., Theorem 5.10 in Villani (2008).

In order to state our main result, we need to introduce a Gaussian random field $G(\cdot) : \mathbb{R}^m \to \mathbb{R}$ with covariance structure given by

$$\mathrm{Cov}(G(\beta), G(\beta')) = \mathrm{Cov}\Big(\min_{j=1,\ldots,m} \big\{ c(X, y_j) - \beta_j \big\}, \min_{j=1,\ldots,m} \big\{ c(X, y_j) - \beta'_j \big\} \Big)$$

for any $\beta = (\beta_j)_{j=1}^m$ and $\beta' = (\beta'_j)_{j=1}^m$. Our main result is the following.

Theorem 3.1. *Under Assumption 2, $V_n \xrightarrow{p} V'$ as $n \to \infty$. Moreover,*

$$n^{1/2} \left(V_n - V' \right) \Rightarrow G^*$$

as $n \to \infty$, where

$$G^* = \max_{\beta = (\beta_1,\ldots,\beta_m) \in \mathbb{S}} G\,(\beta)\,.$$

Here $\mathbb{S}$ is the set of all optimal solutions $\beta = (\beta_j)_{j=1}^m \in \mathbb{R}^m$ for the convex optimization problem

$$\max_{\substack{\beta_1,\ldots,\beta_m \in \mathbb{R} \\ \sum_{j=1}^m \beta_j = 0}} \mathbb{E}_\mu \Big[\min_{j=1,\ldots,m} \{ c\,(X, y_j) - \beta_j \} + \sum_{j=1}^{m} \beta_j \nu\{y_j\} \Big]. \tag{3.5}$$

Remark 3.1. *The significance of this result is that one can approximate worst-case expectations by sampling with a rate of convergence (as measured by the sample size of the continuous distribution) of order $O(n^{-1/2})$. As we mentioned earlier, this might be somewhat surprising given that standard empirical estimators for Wasserstein distances exhibit a degradation which becomes quite drastic in high dimensions.*

3.2.1 Proof of Theorem 3.1

We first note that adding a constant to β_j in the objective function of the dual does not change the objective value. To remove this ambiguity we inroduce the next result.

Lemma 3.1. *Define*

$$\widehat{V_n} := \max_{\substack{\beta_j \in \mathbb{R}, j=1,\ldots,m. \\ \sum_{j=1}^m \beta_j = 0}} \left\{ \frac{1}{n} \sum_{i=1}^n \min_{j=1,\ldots,m} \left\{ c(X_i, y_j) - \beta_j \right\} + \sum_{j=1}^m \beta_j \nu\{y_j\} \right\}. \tag{3.6}$$

We have $V_n = \widehat{V_n}$.

Proof. The dual formulation of V_n, depicted as the LP (3.3), is given by

$$\begin{array}{ll} \max & \frac{1}{n} \sum_{i=1}^n \alpha_i + \sum_{j=1}^m \beta_j \nu\{y_j\} \\ \text{subject to} & \alpha_i + \beta_j \leq c(X_i, y_j) \ \forall i = 1, \ldots, n, \ j = 1, \ldots, m \end{array} \tag{3.7}$$

where $(\alpha_i)_{i=1}^m$, $(\beta_j)_{j=1}^m$ are the dual variables. Note that the constraint in (3.7) can be written as $\alpha_i \leq \min_{j=1,\ldots,m}\{c(X_i, y_j) - \beta_j\}$ $\forall i = 1, \ldots, n$, which implies that (3.7) is equivalent to

$$\max_{\beta_j \in \mathbb{R}, j=1,\ldots,m} \left\{ \frac{1}{n} \sum_{i=1}^n \min_{j=1,\ldots,m} \left\{ c(X_i, y_j) - \beta_j \right\} + \sum_{j=1}^m \beta_j \nu\{y_j\} \right\} \tag{3.8}$$

Since shifting any $(\beta_j)_{j=1}^m$ to $(\beta_j + \lambda)_{j=1}^m$ by an arbitrary constant λ does not affect the objective value of (3.8), we can always set $\lambda = -\frac{1}{m} \sum_{j=1}^m \beta_j$ to enforce the constraint $\sum_{j=1}^m \beta_j = 0$, so that (3.8) is equal to (3.6). Finally, since (3.3) is feasible by choosing an independent distribution, strong duality holds. We therefore conclude the lemma. □

Next we show that $\widehat{V_n}$ can be further reduced to a problem with compact feasible region, which will subsequently facilitate the invocation of classical results in SAA:

Proposition 3.2. *Define*

$$\widehat{V_n^b} := \max_{\substack{\beta_j \in \mathbb{R}, |\beta_j| \leq b, j=1,\ldots,m \\ \sum_{j=1}^m \beta_j = 0}} \left\{ \frac{1}{n} \sum_{i=1}^n \min_{j=1,\ldots,m} \left\{ c(X_i, y_j) - \beta_j \right\} + \sum_{j=1}^m \beta_j \nu\{y_j\} \right\}. \tag{3.9}$$

There exists some large enough constant $b > 0$ *such that*

$$V_n = \widehat{V_n^b} \tag{3.10}$$

eventually, i.e., holds for any $n > N$ *for some* $N < \infty$ *almost surely.*

Proof. By Lemma 3.1, we have

$$\begin{aligned} V_n &= \widehat{V_n} \\ &= \max\Bigg\{ \max_{\substack{\beta_j\in\mathbb{R},|\beta_j|\leq b,j=1,\dots,m \\ \sum_{j=1}^m \beta_j=0}} \Bigg\{ \frac{1}{n}\sum_{i=1}^n \min_{j=1,\dots,m} \big\{c(X_i,y_j)-\beta_j\big\} + \sum_{j=1}^m \beta_j \nu\{y_j\}\Bigg\}, \\ &\quad \max_{\substack{\beta_j\in\mathbb{R},j=1,\dots,m,|\beta_j|>b \text{ for some } j \\ \sum_{j=1}^m \beta_j=0}} \Bigg\{ \frac{1}{n}\sum_{i=1}^n \min_{j=1,\dots,m} \big\{c(X_i,y_j)-\beta_j\big\} + \sum_{j=1}^m \beta_j \nu\{y_j\}\Bigg\}\Bigg\} \end{aligned} \tag{3.11}$$

Note that the first term inside the outer max is $\widehat{V_n^b}$ by our definition (3.9). We will show that there exists a deterministic $b>0$ such that the first term dominates the second term eventually, which will then conclude the proposition.

To this end, consider the second term in (3.11)

$$\begin{aligned} &\max_{\substack{\beta_j\in\mathbb{R},j=1,\dots,m,|\beta_j|>b \text{ for some } j \\ \sum_{j=1}^m \beta_j=0}} \Bigg\{ \frac{1}{n}\sum_{i=1}^n \min_{j=1,\dots,m} \big\{c(X_i,y_j)-\beta_j\big\} + \sum_{j=1}^m \beta_j \nu\{y_j\}\Bigg\} \\ \leq\ &\max_{\substack{\beta_j\in\mathbb{R},j=1,\dots,m,|\beta_j|>b \text{ for some } j \\ \sum_{j=1}^m \beta_j=0}} \Bigg\{ \min_{j=1,\dots,m} \big\{-\beta_j\big\} + \sum_{j=1}^m \beta_j \nu\{y_j\}\Bigg\} + \frac{1}{n}\sum_{i=1}^n \max_{j=1,\dots,m} c(X_i,y_j). \end{aligned}$$

We analyze

$$\max_{\substack{\beta_j\in\mathbb{R},j=1,\dots,m,|\beta_j|>b \text{ for some } j \\ \sum_{j=1}^m \beta_j=0}} \Bigg\{ \min_{j=1,\dots,m} \big\{-\beta_j\big\} + \sum_{j=1}^m \beta_j \nu\{y_j\}\Bigg\}. \tag{3.12}$$

Denote $M=\max_{j=1,\dots,m}|\beta_j|$, so that $M>b$ for any β inside the feasible region. There must exist either a $\beta_{j^*}=M$ or $\beta_{j^*}=-M$. In the first case, we have

$$\begin{aligned} &\max_{\substack{\beta_j\in\mathbb{R},j=1,\dots,m,|\beta_j|>b \text{ for some } j \\ \sum_{j=1}^m \beta_j=0}} \Bigg\{ \min_{j=1,\dots,m} \big\{-\beta_j\big\} + \sum_{j=1}^m \beta_j \nu\{y_j\}\Bigg\} \\ \leq\ & -M + \left\{\begin{array}{ll} \max & \sum_{j=1}^m \beta_j\nu\{y_j\} \\ \text{subject to} & \beta_j \leq M\ \forall j=1,\dots,m \\ & \sum_{j=1}^m \beta_j = 0 \end{array}\right\} \\ =\ & -M + M \times \left\{\begin{array}{ll} \max & \sum_{j=1}^m \beta_j\nu\{y_j\} \\ \text{subject to} & \beta_j \leq 1\ \forall j=1,\dots,m \\ & \sum_{j=1}^m \beta_j = 0 \end{array}\right\} \end{aligned} \tag{3.13}$$

where the last equality follows by a change of variable from β_j to β_j/M in the optimization. Note that the optimal value of

$$\begin{array}{ll} \max & \sum_{j=1}^m \beta_j \nu\{y_j\} \\ \text{subject to} & \beta_j \leq 1 \ \forall j = 1, \dots, m \\ & \sum_{j=1}^m \beta_j = 0 \end{array}$$

is strictly less than 1. To see this, observe that the optimal value is at most 1 by using the first constraint. The value of exactly 1 is attained under the first constraint by the unique solution $\beta_j = 1, j = 1, \dots, m$, which is ruled out because it would violate the second constraint. With this claim, we conclude that (3.13) is equal to θM for some $\theta < 0$, which is bounded from above by θb.

In the second case, we have $\beta_{j^*} = -M$. Let $\tilde{j}^* = \text{argmax}_{j=1,\dots,m}\{\beta_j\}$. By the constraint $\sum_{j=1}^m \beta_j = 0$ in (3.12), we must have $\beta_{\tilde{j}^*} \geq M/(m-1)$. Therefore, applying our argument for the first case gives that (3.12) is bounded from above by $\theta M/(m-1) \leq \theta b/(m-1)$ for the same $\theta < 0$ chosen before.

Therefore, in either case (3.12) is bounded from above by $\theta b/(m-1)$. Note that the first term inside the outer max in (3.11), namely $\widehat{V_n^b}$, satisfies $\widehat{V_n^b} \geq (1/n)\sum_{i=1}^n \min_{j=1,\dots,m} c(X_i, y_j)$ by plugging in the feasible solution given by $\beta_j = 0, j = 1, \dots, m$. Thus, with the law of large numbers, by choosing $b > 0$ large enough such that

$$\frac{\theta b}{m-1} + \mathbb{E}_\mu\left[\max_{j=1,\dots,m} c(X, y_j)\right] < \mathbb{E}_\mu\left[\min_{j=1,\dots,m} c(X, y_j)\right] \tag{3.14}$$

the first term dominates the second term inside the outer max in (3.11) as $n \to \infty$ almost surely.

□

We are now ready to prove Theorem 3.1:

Proof of Theorem 3.1. Note that the function

$$F(X, \beta) := \min_{j=1,\dots,m} \{c(X, y_j) - \beta_j\} + \sum_{j=1}^m \beta_j \nu\{y_j\} \tag{3.15}$$

on $\beta = (\beta_j)_{j=1}^m \in \mathbb{R}^m$ is Lipschitz continuous in the sense that

$$|F(X, \beta) - F(X, \beta')| \leq (1 + \|\nu\|)\|\beta - \beta'\|$$

where $\|\cdot\|$ denotes the L_2-norm, and ν is interpreted as a vector $(\nu\{y_j\})_{j=1}^m$. This follows since

$$\left|\min_{j=1,\ldots,m}\left\{c(X,y_j)-\beta_j\right\}-\min_{j=1,\ldots,m}\left\{c(X,y_j)-\beta_j'\right\}\right|\leq\|\beta-\beta'\|_\infty$$

and

$$\left|\sum_{j=1}^m\beta_j\nu\{y_j\}-\sum_{j=1}^m\beta_j'\nu\{y_j\}\right|\leq\|\nu\|\|\beta-\beta'\|$$

by the Cauchy-Schwarz inequality. Since the set $\mathbb{B}:=\{\beta\in\mathbb{R}^m:\sum_{j=1}^m\beta_j=0,|\beta_j|\leq b,\forall j=1,\ldots,m\}$ is compact and $\mathbb{E}_\mu[F(X,\beta)^2]<\infty$ by Assumption 2, by using Theorem 5.7 in Shapiro *et al.* (2009), we have

$$\widehat{V_n^b}\xrightarrow{p}V^b \tag{3.16}$$

and

$$\sqrt{n}(\widehat{V_n^b}-V^b)\Rightarrow G^{*,b} \tag{3.17}$$

where

$$V^b=\max_{\substack{\beta_j\in\mathbb{R},|\beta_j|\leq b,j=1,\ldots,m\\ \sum_{j=1}^m\beta_j=0}}\mathbb{E}_\mu\Big[\min_{j=1,\ldots,m}\{c\left(X,y_j\right)-\beta_j\}+\sum_{j=1}^m\beta_j\nu\{y_j\}\Big] \tag{3.18}$$

and

$$G^{*,b}=\max_{\beta=(\beta_1,\ldots,\beta_m)\in\mathbb{S}^b}G\left(\beta\right)$$

with $\mathbb{S}^b$ denoting the set of optimal solutions for (3.18) and $G(\cdot)$ is defined as in Theorem 3.1 but restricted to the domain $\mathbb{B}$.

By Proposition 3.2, we have $\sqrt{n}(\widehat{V_n^b}-V_n)\xrightarrow{p}0$ as $n\to\infty$. Thus, together with (3.16), we have

$$V_n\xrightarrow{p}V^b$$

and together with (3.17), we have

$$\sqrt{n}(V_n-V^b)\Rightarrow G^{*,b}$$

by Slutsky's Theorem.

To conclude the theorem, we show that $V^b=V'$, and $\mathbb{S}^b=\mathbb{S}$ so that $G^{*,b}=G^*$. By using essentially the same argument as for Proposition 3.2 (with the empirical expectation replaced by $\mathbb{E}_\mu[\cdot]$) and choosing the same b as in (3.14), we have

$$
\begin{aligned}
V' &= \max_{\beta_j\in\mathbb{R}, j=1,\ldots,m} \mathbb{E}_\mu\Big[\min_{j=1,\ldots,m}\{c(X,y_j)-\beta_j\}+\sum_{j=1}^m \beta_j\nu\{y_j\}\Big]\\
&= \max_{\substack{\beta_j\in\mathbb{R}, j=1,\ldots,m\\ \sum_{j=1}^m\beta_j=0}} \mathbb{E}_\mu\Big[\min_{j=1,\ldots,m}\{c(X,y_j)-\beta_j\}+\sum_{j=1}^m \beta_j\nu\{y_j\}\Big]\\
&\qquad \text{by shifting any } (\beta_j)_{j=1}^m \text{ to } (\beta_j-(1/m)\sum_{k=1}^m\beta_k)_{j=1}^m \text{ which does not affect the objective value and}\\
&\qquad \text{enforces the constraint } \sum_{j=1}^m\beta_j=0\\
&= \max\Bigg\{\max_{\substack{\beta_j\in\mathbb{R}, |\beta_j|\leq b, j=1,\ldots,m\\ \sum_{j=1}^m\beta_j=0}} \mathbb{E}_\mu\Big[\min_{j=1,\ldots,m}\{c(X,y_j)-\beta_j\}+\sum_{j=1}^m \beta_j\nu\{y_j\}\Big],\\
&\qquad \max_{\substack{\beta_j\in\mathbb{R}, j=1,\ldots,m, |\beta_j|>b \text{ for some } j\\ \sum_{j=1}^m\beta_j=0}} \mathbb{E}_\mu\Big[\min_{j=1,\ldots,m}\{c(X,y_j)-\beta_j\}+\sum_{j=1}^m \beta_j\nu\{y_j\}\Big]\Bigg\}
\end{aligned}
$$

where

$$
\begin{aligned}
&V^b\\
&= \max_{\substack{\beta_j\in\mathbb{R}, |\beta_j|\leq b, j=1,\ldots,m\\ \sum_{j=1}^m\beta_j=0}} \mathbb{E}_\mu\Big[\min_{j=1,\ldots,m}\{c(X,y_j)-\beta_j\}+\sum_{j=1}^m \beta_j\nu\{y_j\}\Big]\\
&> \max_{\substack{\beta_j\in\mathbb{R}, j=1,\ldots,m, |\beta_j|>b \text{ for some } j\\ \sum_{j=1}^m\beta_j=0}} \mathbb{E}_\mu\Big[\min_{j=1,\ldots,m}\{c(X,y_j)-\beta_j\}+\sum_{j=1}^m \beta_j\nu\{y_j\}\Big]
\end{aligned}
$$

so that $V'=V^b$ and $\mathbb{S}^b=\mathbb{S}$. □

3.3 Additional Discussion and Extensions

Finally, we briefly discuss the challenge in generalizing our procedure to the case when both X and Y are continuous. Here, one may attempt to sample both variables (assuming both can be simulated) and formulate a sampled program like (3.2) or (3.3). However, the analog of its reformulation in (3.6) and (3.9) will have a growing number of variables β_j and an analogous limit in (3.5) that involves an infinite-dimensional variable, which challenges the use of standard SAA machinery. In fact, consider a special example where $X,Y\sim U[0,1]^d$

and $c(x, y) = \|x - y\|$. In this case, (3.1) corresponds to the Wasserstein distance (of order 1) between X and Y, which is of course 0. It is known that sampling X and keeping Y continuous will give, for $d \geq 3$, an expected optimal value of (3.2) that is of order $n^{-1/d}$, i.e., $C_1 n^{-1/d} \leq EV_n \leq C_2 n^{-1/d}$ for all n for some $C_1, C_2 > 0$ (e.g., Problem 5.11 in van Handel (2014)). Thus, the convergence rate deteriorates with the dimension and the standard Monte Carlo rate $O(n^{-1/2})$ cannot be maintained without assuming additional structure or infomation available to the modeler on the primal problem. It is of interest to investigate reasonable assumptions which are useful in applications and which would mitigate such rate-of-convergence deterioration.

Chapter 4

Dependence with several sources of uncertainty: Martingale Optimal Transport with the Markov Property

4.1 Introduction

In this chapter we study a discretization approach for computing lower and upper bounds among any dependence structure for a function of multiple random variables whose marginal distributions are assumed to be known. More specifically, given $d \geq 2$ marginal distributions $\mu_1, \ldots, \mu_d$ on a common compact metric space $\mathscr{X}$, we focus on the lower bound

$$\inf_{\pi \in \Pi(\mu_1, \ldots, \mu_d)} \mathbb{E}_\pi[c(X_1, \ldots, X_d)], \tag{4.1}$$

where $\Pi(\mu_1, \ldots, \mu_d)$ is the set of all joint distributions with marginals $X_1 \sim \mu_1, \ldots, X_d \sim \mu_d$, and c is a cost function. For instance, in risk management, such situations often occur when the estimation of marginal distributions of each risk factor X_i is relatively easy, but the dependence structure among them is ambiguous. Given loss level $\ell > 0$ and the cost function $c(X_1, \ldots, X_d) = \mathbb{I}(X_1 + \cdots + X_d > \ell)$ with d risk factors $X_1, \ldots, X_d$, the quantity

$\inf_{\pi\in\Pi(\mu_1,\ldots,\mu_d)} \mathbb{E}_\pi[c(X_1,\ldots,X_d)]$ gives a lower bound for the probability of the event that the sum of these risk factors exceed the level ℓ.

Note that when $d = 2$, the problem (4.1) is the standard optimal transport (or Monge-Kantorovich) problem, whose theoretical properties have been studied in the literature substantially(Villani (2003), Villani (2008)). For $d > 2$, this problem has been studied by Gangbo and Swiech (1998) and G.Carlier *et al.* (2008).

Based on the above optimal transport problem (4.1), Beiglbock *et al.* (2013) and Galichon *et al.* (2014) further develop the so-called martingale optimal transport problem, which adds the martingale constraint to the joint distribution. The martingale optimal transport problem has the following form

$$\inf_{\pi\in\mathscr{M}(\mu_1,\ldots,\mu_d)} \mathbb{E}_\pi[c(X_1,\ldots,X_d)],$$

where $\mathscr{M}(\mu_1,\ldots,\mu_d)$ is the set of all martingale measures, i.e. the underlying process $(X_t)_{t=1,\ldots,d}$ satisfies $X_t \sim \mu_t, \mathbb{E}_\pi[X_{t+1}|\mathcal{F}_t] = X_t$ for $t = 1,\ldots,d-1$.

In contrast to optimal transport problem where the product measure is always a feasible solution, here the existence of martingale measure π requires some constraint on marginals: for the feasibility of

$$\{\mathbb{P} : X_1 \sim \mu_1,\ldots,X_d \sim \mu_d; \mathbb{E}_\pi[X_{t+1}|\mathcal{F}_t] = X_t \text{ for } t = 1,\ldots,d-1\},$$

we need the condition that all the marginals satisfy the convex order: $\mu_t \leq \mu_{t+1}$, for $t = 1,\ldots,d-1$ where $\mu_t \leq \mu_{t+1}$ is defined as $\mathbb{E}_{\mu_t}[\psi(X_t)] \leq \mathbb{E}_{\mu_{t+1}}[\psi(X_{t+1})], \forall\psi$ convex. See H.G.Kellerer (1972).

The main motivation of studying martingale optimal transport problem stems from the requirement of financial robustness against model risks. In finance, it is important to choose a pricing model when evaluating an exotic option; such a model is characterized by a martingale measure while the marginal distributions are the daily underlying prices. Instead of postulating a model, (4.6) gives a model-free lower bound for the price of exotics, whose payoff function c depends on the d-marginal distributions of a certain underlying X, indexed by time $t = 1,\ldots,d$. Similarly, by using maximization instead of minimization we also obtain an upper bound. This price range is robust against model errors and it complies with market prices of vanilla options, which are liquid and suitable hedging instruments.

Previous literatures tackle the martingale optimal transport problems by numerically solving the Hamilton-Jacobi-Bellman PDEs (Henry-Labordere and Touzi (2013)), but the computational complexity is hard to track.

Compared to their numerical PDE approach, our discretization approach can obtain an approximate solution within certain error under a much mild assumptions, both on the cost function c and on the marginal distributions. We give the computational complexity for the general optimal transport problem in high dimensions. And we also give a discretization method for the martingale transport problem with a special type of cost functions and provide a practical way of robustly pricing certain financial derivatives.

4.2 Optimal Transport Problems with Two Marginal Distributions and Minimum Cost Problems

4.2.1 Problem Definition

For simplicity and to describe the idea of discretization, in this section we consider the case of two given marginals. Let $\mathscr{X}$ be a compact metric space and $\mathscr{P}(\mathscr{X})$ be the set of probability measures on $\mathscr{X}$. In this section, we only consider two (marginal) probability measures. For $\mu, \nu \in \mathscr{P}(\mathscr{X})$, let $\Pi(\mu, \nu)$ be the set of all joint probability measures with μ and ν as marginals. We are interested in the following optimal transport problem

$$P := \min_{\pi \in \Pi(\mu,\nu)} \left\{ \int c(x,y) d\pi(x,y) \right\} = \min_{\pi \in \Pi(\mu,\nu)} \mathbb{E}_\pi c(X,Y) \tag{4.2}$$

where the cost function $c(\cdot,\cdot) : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ is Lipschitz continuous with Lipschitz constant K, and $\|c\| \leq 1$.

In the following we will try to solve the optimal transport problem by discretizing it to a linear programming problem. We want to understand the computational complexity of computing P within $\epsilon > 0$ precision.

4.2.2 Quantization and Discretization

We first create a partition of $\mathscr{X}$ with $\mathscr{X} = \sum_{k=1}^{n} A_k$ such that the diameter of every A_k does not exceed δ, with $\delta = O(n^{-1})$. Then we choose a representative $x_k \in A_k$ for each k and form a discrete set $\mathscr{X}_\delta = \{x_k : k = 1, \ldots, n\}$ with an associated quantization map

$$\begin{aligned} T : \mathscr{X} &\to \mathscr{X}_\delta \\ x &\mapsto \sum_{k=1}^{n} x_k \mathbb{I}(x \in A_k). \end{aligned}$$

In addition, we define the corresponding quantized measures as

$$\mu_\delta(x_k) = \mu(A_k) \text{ and } \nu_\delta(x_k) = \nu(A_k), \text{ for } k = 1, \ldots, n. \tag{4.3}$$

We then obtain the following discretized approximate version of the optimal transport problem:

$$\begin{aligned} P_\delta := \min_{\pi_{i,j}} & \sum_{i,j=1}^{n} c(x_i, x_j)\pi_{i,j} \\ \text{s.t.} & \\ & \sum_{j=1}^{n} \pi_{i,j} = \mu_\delta(x_i), \ i = 1, \cdots, n \\ & \sum_{i=1}^{n} \pi_{i,j} = \nu_\delta(x_j), \ j = 1, \cdots, n \\ & \sum_{i=1,j=1}^{n} \pi_{i,j} = 1, \ \pi_{i,j} \geq 0. \end{aligned} \tag{4.4}$$

It is actually an assignment problem, which is a special type of minimum cost problem that can be solved by various network algorithms that are much faster than the general LP algorithms. For instance, with the successive shortest path algorithm (see R.K.Ahuja *et al.* (2000) p.320) one can achieve $O(n^2 \log(n))$.

Lemma 4.1.

$$|P - P_\delta| \leq K\delta.$$

Proof. Let $\bar{\pi} \in \Pi(\mu, \nu)$ be an ϵ-optimal coupling such that

$$P > \int c(x, y)\bar{\pi}(x, y) - \epsilon.$$

Since $c \in Lip(K)$, we have that $|c(x,y) - c(T(x),T(y))| \leq K\delta$, which gives that

$$\begin{aligned} P > & \int c(T(x),T(y))\bar{\pi}(x,y) - K\delta - \epsilon \\ = & \sum_{i,j} c(x_i,x_j)\bar{\pi}(A_i,A_j) - K\delta - \epsilon \\ \geq & P_\delta - K\delta - \epsilon, \end{aligned}$$

where the last inequality follows from the observation that the probability mass function defined by $\big(\bar{\pi}(A_i \times A_j)\big)_{i,j=1}^n$ is an element of $\Pi(\mu,\nu)$ and P_δ is the corresponding maximum that can be attained.

To proof the other direction, let $(\pi^*_{i,j})^n_{i,j=1}$ be the solution to the linear programing (4.4). We then consider the following sampling procedure:

1. Draw an index (I,J) from the distribution $(\pi^*_{i,j})^n_{i,j=1}$.
2. Given $(I,J)=(i,j)$ draw a sample $X \sim \mu(\cdot|A_i)$ and $Y \sim \nu(\cdot|A_j)$.

Note that (X,Y) has a joint distribution which belongs to $\Pi(\mu,\nu)$ and conditioned on the realization $(I,J)=(i,j)$, we have $|c(X,Y) - c(x_i,x_j)| \leq K\delta$. Therefore

$$\begin{aligned} P & \leq \mathbb{E}c(X,Y) \\ & \leq \sum_{i,j} c(x_i,x_j)\pi^*_{i,j} + K\delta \\ & \leq P_\delta + K\delta. \end{aligned}$$

Since $\epsilon > 0$ is arbitrary, the claim is proved. □

4.3 Optimal Transport Problems with d-Marginals

Now we consider an optimal transport problems with d-marginals, but all of the marginal distributions are supported on a compact metric space $\mathscr{X}$.

$$\inf_{\pi \in \Pi(\mu_1,\ldots,\mu_d)} \mathbb{E}_\pi[c(X_1,\ldots,X_d)]. \tag{4.5}$$

where the cost function c is Lipschitz continuous with Lipschitz constant K, and $\|c\| \leq 1$.

4.3.1 Discretization and Complexity

Just as Section 4.2 we can similarly approximate the optimal solution to problem (4.5) by the following discretization

$$\min_{\pi} \sum_{i_1,\cdots,i_d=1}^{n} c(x_{i_1},\cdots,x_{i_d})\pi(x_{i_1},\cdots,x_{i_d})$$

s.t.

$$\sum_{i_2=1,\cdots,i_d=1}^{n} \pi(x_{i_1},\cdots,x_{i_d}) = \mu_1(x_{i_1}),\ i_1 = 1,\cdots,n,$$

$$\cdots$$

$$\sum_{i_1=1,\cdots,i_{d-1}=1}^{n} \pi(x_{i_1},\cdots,x_{i_d}) = \mu_d(x_{i_d}),\ i_d = 1,\cdots,n,$$

$$\sum_{i_1=1,\cdots,i_d=1}^{n} \pi(x_{i_1},\cdots,x_{i_d}) = 1,\ \pi(x_{i_1},\cdots,x_{i_d}) \geq 0,$$

where μ_k is the quantized marginal distribution of X_k. Since here, we will omit the δ subscript and use μ_k to denote both the marginal distribution and the quantized marginal distribution of X_k when there's no confusion.

This is an LP problem with n^d unknown variables and $d \cdot n$ equality constrainsts, which is underdetermined in the sense that $n^d >> d \cdot n$. In the papers E.Candes *et al.* (2005) and E.Candes and T.Tao (2014) such underdetermined LP problems with sparse solutions are discussed, and they proved that if the coefficient matrix satisfies the so called restricted orthonormality condition then the (sparse) solution exists and unique. However, these papers just use usual LP algorithm as efficient way to recover the sparse solutions.

Note also that for the general cost function c it is difficult to transform it to a minimum cost problem, since there exist no direct way to put these d-dimensional transport into a plane graph. So in this case we can no longer employ the more efficient network flow algorithms.

In G.Puccetti (2014) they propose the rearrangement algorithms, which only works for the case when the cost c takes the form of the linear combination of unknown variables $c(x_{i_1}, \cdots, x_{i_d}) = \sum_{k=1}^{n} \alpha_k x_{i_k}$.

4.4 Martingale Optimal Transport Problems with Separable Cost Functions and the Markov Property

In this section we are focusing on the following martingale optimal transport problem:

$$\inf_{\pi \in \mathscr{M}(\mu_1,\ldots,\mu_d)} \mathbb{E}_\pi[c(X_1,\ldots,X_d)], \tag{4.6}$$

where $\mathscr{M}(\mu_1,\ldots,\mu_d)$ is the set of all martingale measures, i.e. the underlying process $(X_t)_{t=1,\ldots,d}$ satisfies $X_t \sim \mu_t, \mathbb{E}_\pi[X_t|\mathcal{F}_{t-1}] = X_{t-1}$. In addition, X_t takes values in a compact metric space $\mathscr{X}$ and $(\mu_t)_{t=1}^{d}$ satisfy the convex order condition.

Since the dimension d could be interpreted as d days in financial applications, the transport should be proceeded as from a layer to its next layer. It might make sense to add an assumption that the underlying process X_t is Markovian. We expect that under the Markovian assumption the optimization problem can be simplified a lot, but unfortunately we still need another strict assumption to achieve the simplification.

Assumption 1: The cost function $c(x_1, \cdots, x_d)$ is separable in the sense that it can be decomposed into a sum of

$$c_1(x_1, x_2), \cdots, c_{d-1}(x_{d-1}, x_d).$$

Remark 4.1. *The payoff, such as, $(\frac{1}{d}\sum_{i=1}^{d} X_i - K)^+$ of an Asian option doesn't satisfy this condition.*

Nonetheless, with this assumption at hand, we don't even need to assume that the underlying process is Markovian. Instead, we can not only conclude that the underlying process $(X_t)_{t=1,\ldots,d}$ must be Markovian but we can also decompose the original optimization problem into $d-1$ sub optimization problems. Last but not least, in this case we can also easily add martingale constraint onto the sub problems. This is summarized in the following lemma.

Lemma 4.2. *Under Assumption 1 with the particular cost function c, the associated process* $(X_t)_{t=1,\ldots,d}$ *to the martingale optimal problem* (4.6) *is Markovian.*

Proof. Under Assumption 1 we can decompose the objective function as the following:

$$\begin{aligned}
&\min_{\pi\in\mathscr{M}(\mu_1,\cdots,\mu_d)}\int_{x_1,\cdots,x_d} c(x_1,\cdots,x_d)d\pi(x_1,\cdots,x_d)\\
=&\min_{\pi\in\mathscr{M}(\mu_1,\cdots,\mu_d)}\int_{x_1,\cdots,x_d}\Big\{\sum_{k=1}^{d-1} c(x_1,\cdots,x_d)d\pi(x_1,\cdots,x_d)\Big\}\\
=&\min_{\pi\in\mathscr{M}(\mu_1,\cdots,\mu_d)}\sum_{k=1}^{d-1}\int_{x_k,x_{k+1}} c(x_k,x_{k+1})d\pi_{k,k+1}(x_k,x_{k+1}) \qquad (4.7)\\
=&\sum_{k=1}^{d-1}\Big\{\min_{\pi_{k,k+1}\in\mathscr{M}(\mu_k,\mu_{k+1})}\int_{x_k,x_{k+1}} c(x_k,x_{k+1})d\pi_{k,k+1}(x_k,x_{k+1})\Big\}\\
=&\sum_{k=1}^{d-1}\Big\{\min_{P_{k,k+1}}\int_{x_k,x_{k+1}} c(x_k,x_{k+1})\mu_k(x_k)dP_{k,k+1}(x_k,x_{k+1})\Big\},
\end{aligned}$$

where $\pi_{k,k+1}$ is the "marginal joint" distribution of X_k and X_{k+1}, and $P_{k,k+1}$ is the transition probability which satisfies

$$\pi_{k,k+1}(x_k,x_{k+1}) = \mu_k(x_k)P_{k,k+1}(x_k,x_{k+1})$$

and the martingale constraint

$$\int_{x_{k+1}} x_{k+1}P_{k,k+1}(x_k,dx_{k+1}) = x_k.$$

Note that the above derivation does not need the Markovian assumption at all. From the above we know that the minimization of all joint distribution π is equivalent to the minimization over all $P_{k,k+1}$. The optimal joint distribution is determined by

$$\pi^*(x_{i_1},\cdots,x_{i_d}) = \mu_1(x_{i_1})P^*_{12}(x_{i_1},x_{i_2})\cdots P^*_{d-1,d}(x_{i_{d-1}},x_{i_d}).$$

So the associated optimal process is Markovian. □

We now turn to the corresponding discretized version of the martingale optimal transport problem. The particular form of the cost function $c(x_1, \cdots, x_d) = c(x_1, x_2) + \cdots + c(x_{d-1}, x_d)$ leads to the conclusion that the process associated to the optimal solution is Markovian. In addition, the above proof of Lemma 4.2 shows that the martingale constraint can also be decomposed into martingale constraints on the transition kernal. So in the end, the corresponding discretized martingale optimal transport problem is decomposed into the following $d-1$ LP problems:

$$\min_{P_{12}} \sum_{i_1, i_2 = 1}^{n} c(x_{i_1}, x_{i_2}) \mu_1(x_{i_1}) P_{12}(x_{i_1}, x_{i_2})$$

s.t.

$$\sum_{i_1=1}^{n} \mu_1(x_{i_1}) P_{12}(x_{i_1}, x_{i_2}) = \mu_2(x_{i_2}),\ i_2 = 1, \cdots, n;$$

$$\sum_{i_2=1}^{n} P_{12}(x_{i_1}, x_{i_2}) = 1,\ i_1 = 1, \cdots, n;$$

$$\sum_{i_2=1}^{n} x_{i_2} P_{12}(x_{i_1}, x_{i_2}) = x_{i_1},\ i_1 = 1, \cdots, n;$$

$$P_{12}(x_{i_1}, x_{i_2}) \geq 0,\ i_1, i_2 = 1, \cdots, n;$$

$$\vdots$$

$$\min_{P_{d-1,d}} \sum_{i_{d-1}, i_d = 1}^{n} c(x_{i_{d-1}}, x_{i_d}) \mu_{d-1}(x_{i_1}) P_{d-1,d}(x_{i_{d-1}}, x_{i_d})$$

s.t.

$$\sum_{i_{d-1}=1}^{n} \mu_{d-1}(x_{i_{d-1}}) P_{d-1,d}(x_{i_{d-1}}, x_{i_d}) = \mu_d(x_{i_d}),\ i_d = 1, \cdots, n,$$

$$\sum_{i_d=1}^{n} P_{d-1,d}(x_{i_{d-1}}, x_{i_d}) = 1,\ i_{d-1} = 1, \cdots, n;$$

$$\sum_{i_d=1}^{n} x_{i_d} P_{d-1,d}(x_{i_{d-1}}, x_{i_d}) = x_{i_{d-1}},\ i_{d-1} = 1, \cdots, n;$$

$$P_{d-1,d}(x_{i_{d-1}}, x_{i_d}) \geq 0,\ i_{d-1}, i_d = 1, \cdots, n;$$

Therefore, with this special cost function, the algorithm complexity for the martingale optimal transport problem is $(d-1) \cdot \mathrm{LP}(3n \times n^2)$, where $\mathrm{LP}(3n \times n^2)$ is the cost to solve an LP with a $3n \times n^2$ coefficient matrix.

Remark 4.2. *With Assumption 1 but without the martingale constraints, the problem reduces to a particular case in Section 4.5. In this case, it decomposes then into $n-1$ minimum cost problems, and the total complexity can be further improved to $(d-1) \cdot O(n^2 \log(n))$.*

4.4.1 Applications in Pricing Exotic Options

In practice of financial engineering we can observe traded option prices, but know little or nothing about the model. There are many models which are consistent with the market prices of liquidly traded options but they may give very different prices for the exotic. Ideally one might attempt to characterise a model which is consistent with all the market price of options, but this is a very challenging problem, and a less ambitious one is to characterise a model which can give the bounds to the price of exotic options, such as the maximum or minimum of the price of an exotic option. We need the following two assumptions for our model-free pricing framework:

Assumption 2: There exists a risk-neutral measure in the market.

Assumption 3: We could quite exactly estimate the marginal distribution of the underlying process: $X_1 \sim \mu_1, \cdots, X_d \sim \mu_d$.

Remark 4.3. *In practice, since Vanilla options are very suitable hedge instruments because of high liquidity, so the pricing has to comply with their market prices. The distribution of X_t is obtained from vanilla options at $T = t$ by Breeden-Litzenberger formula, $p_{X_t}(x) = \frac{\partial^2}{\partial K^2} C(T = t, K = x)$, where $C(T, K)$ denotes the price of a vanilla call with maturity T and strike K.*

As said before, the payoff of, such as, Asian options $(\frac{1}{d}\sum_{t=1}^{d} S_t - K)^+$ doesn't satisfy Assumption 1. The following two exotic options are examples where their payoff functions satisfy Assumption 1:

Example 4.1. *A cliquet option is an exotic option consisting of a series of ”pre-purchased” at-the-money options where the total premium is determined in advance. The first is active immediately. The second becomes active when the first expires, etc. Each option is struck at-the-money when it becomes active. The payout on each option can either be paid at the final maturity, or at the end of each reset period. For instance, an d-year cliquet with reset dates each year would have n payoffs, the payoff function f can be written as*

$$\sum_{t=1}^{d}(X_t - X_{t-1})^+,$$

where $X_0 := K$, the initial strike.

For a general d-periods cliquet, we can get the lower bound of the price by solving the following $d-1$ LP problems:

$$\min_{P_{12}} \sum_{i_1,i_2=1}^{n} (x_{i_2} - x_{i_1})^+ \mu_1(x_{i_1}) P_{12}(x_{i_1}, x_{i_2})$$

$s.t.$

$$\sum_{i_1=1}^{n} \mu_1(x_{i_1}) P_{12}(x_{i_1}, x_{i_2}) = \mu_2(x_{i_2}),\ i_2 = 1, \cdots, n;$$

$$\sum_{i_2=1}^{n} P_{12}(x_{i_1}, x_{i_2}) = 1,\ i_1 = 1, \cdots, n;$$

$$\sum_{i_2=1}^{n} x_{i_2} P_{12}(x_{i_1}, x_{i_2}) = x_{i_1},\ i_1 = 1, \cdots, n;$$

$$P_{12}(x_{i_1}, x_{i_2}) \geq 0,\ i_1, i_2 = 1, \cdots, n;$$

$$\vdots$$

$$\min_{P_{d-1,d}} \sum_{i_{d-1},i_d=1}^{n} (x_{i_d} - x_{i_{d-1}})^+ \mu_{d-1}(x_{i_1}) P_{d-1,d}(x_{i_{d-1}}, x_{i_d})$$

s.t.

$$\sum_{i_{d-1}=1}^{n} \mu_{d-1}(x_{i_{d-1}}) P_{d-1,d}(x_{i_{d-1}}, x_{i_d}) = \mu_d(x_{i_d}),\ i_d = 1, \cdots, n,$$

$$\sum_{i_d=1}^{n} P_{d-1,d}(x_{i_{d-1}}, x_{i_d}) = 1,\ i_{d-1} = 1, \cdots, n;$$

$$\sum_{i_d=1}^{n} x_{i_d} P_{d-1,d}(x_{i_{d-1}}, x_{i_d}) = x_{i_{d-1}},\ i_{d-1} = 1, \cdots, n;$$

$$P_{d-1,d}(x_{i_{d-1}}, x_{i_d}) \geq 0,\ i_{d-1}, i_d = 1, \cdots, n;$$

Note that the the first payoff $\mathbb{E}(X_1 - K)^+$ can be directly calculated from the marginal distribution μ_1.

4.5 A Numerical Experiment

Example 4.2. *Another similar example is variance swaps. A variance swaps is an agreement to exchange the realized volatility*

$$\Big(\log\big(\frac{X_1}{X_0}\big)\Big)^2 + \cdots + \Big(\log\big(\frac{X_T}{X_{T-1}}\big)\Big)^2$$

for some prespecified fixed volatility $\hat{V}$ at time T. The market convention is to set $\hat{V}$ so that no money needs to change hands at initiation of the trade:

$$\hat{V} = \mathbb{E}\Big[\Big(\log\big(\frac{X_1}{X_0}\big)\Big)^2 + \cdots + \Big(\log\big(\frac{X_T}{X_{T-1}}\big)\Big)^2\Big].$$

Its payoff satifies Assumption 1 as well, so we can apply the above method to get a model-free pricing.

The standard way of pricing the payoff is using the following approximation (see, e.g.P.Carr and D.Madan (2002) and S.Bossu et al. *(2005)):*

$$\hat{V} \approx \frac{2\exp(rT^*)}{T^*}\Big\{\sum_{i=1}^{N_{put}} \frac{p_0(K_i^{put}, T^*)}{(K_i^{put})^2}(K_i^{put} - K_{i-1}^{put}) + \sum_{i=1}^{N_{call}} \frac{c_0(K_i^{call}, T^*)}{(K_i^{call})^2}(K_i^{call} - K_{i-1}^{call})\Big\},$$

*where K_i^{put} and K_i^{call} are the respective strikes of the i-th put and i-th call, N_{put} and N_{call} are the respective number of puts and calls, and $p_0(K_i^{put}, T^*)$ and $c_0(K_i^{call}, T^*)$ denote the respective time-*0 *price of puts and calls with strike K_i^{put} and K_i^{call} and maturity T^*. We will use the standard pricing as a benchmark.*

Suppose the variance swaps has maturity $T^* = 1$ and its underlying security starts at $S_0 = 100$. The interest rate is $r = 0.02$ and the forward price is $F = S_0 * \exp(rT) = 102.02$. The standard pricing gives that the strike of variance swaps is 4.01%. We then use a series of calls and puts with strike K ranging from $[0.01, 2F]$ and maturity T ranging from $[0.01, T^*]$ to approximate the densities of the marginal distributions S_t by the Breeden-Litzenberger formula. By performing both the minimization and maximization our method gives a model-free robust price interval $[3.95\%, 13.35\%]$, in which the standard pricing lies, indicating that there's no arbitrage in this case.

Chapter 5

Data-driven choice of the aspects over which to robustify: Data-driven Optimal Transport Cost Selection and Doubly Robust Distributionally Robust Optimization

5.1 Introduction

A Distributionally Robust Optimization (DRO) problem takes the form of

$$\min_{\beta} \max_{P \in \mathcal{U}_\delta} \mathbb{E}_P \left[l \left(X, Y, \beta \right) \right], \tag{5.1}$$

where β is a decision variable, (X, Y) is a random element, and $l(x, y, \beta)$ measures a suitable loss incurred when $(X, Y) = (x, y)$ and the decision β is taken. The expectation $\mathbb{E}_P[\cdot]$ is taken under the probability measure P. The set $\mathcal{U}_\delta$ is called the *distributional uncertainty neighborhood* and it is indexed by the parameter $\delta > 0$, which measures the size of the distributional uncertainty.

The DRO problem is said to be *data-driven* if the uncertainty set $\mathcal{U}_\delta$ is informed by empirical observations. One natural way to use this information is by placing the "center" of the uncertainty region at the empirical measure, P_n, induced by the data set $\{X_i, Y_i\}_{i=1}^n$, which represents an empirical sample of realizations of $W = (X, Y)$. In order to emphasize the data-driven nature of a DRO formulation, we write $\mathcal{U}_\delta = \mathcal{U}_\delta(P_n)$ to represent that the uncertainty region is informed by an empirical sample. Recently, Blanchet *et al.* (2016a) showed that many prevailing machine learning estimators can be reformulated as a data-driven DRO of form (5.1). For example, suppose that $X \in \mathbb{R}^d$ and $Y \in \{-1, 1\}$. Let $l(x, y, \beta) = \log(1 + \exp(-y\beta^T x))$ denote the log-exponential loss associated to a logistic regression model where $Y \sim Ber(1/(1 + \exp(-\beta_*^T x))$, and β_* is the underlying parameter to learn. Then, given a set of empirical samples $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$, and a judicious choice of the distributional uncertainty set $\mathcal{U}_\delta(P_n)$, Blanchet *et al.* (2016a) shows that

$$\min_\beta \max_{P \in \mathcal{U}_\delta(P_n)} \mathbb{E}_P[l(X, Y, \beta)] = \min_\beta \left(\mathbb{E}_{P_n}[l(X, Y, \beta)] + \delta \|\beta\|_p \right), \tag{5.2}$$

where $\|\cdot\|_p$ is the ℓ_p−norm in $\mathbb{R}^d$ for $p \in [1, \infty)$ and $\mathbb{E}_{P_n}[l(X, Y, \beta)] = n^{-1} \sum_{i=1}^n l(X_i, Y_i, \beta)$. The definition of $\mathcal{U}_\delta(P_n)$ turns out to be informed by the dual norm $\|\cdot\|_q$ with $1/p + 1/q = 1$. If $p = 1$ we see that (5.2) recovers L_1 regularized logistic regression (see Friedman *et al.* (2001)). Other estimators such as Support Vector Machines and sqrt-Lasso are shown in Blanchet *et al.* (2016a) to have similar DRO representations – provided that the loss function and the uncertainty region are carefully chosen. Note that the parameter δ in $\mathcal{U}_\delta(P_n)$ is precisely the regularization parameter on the right hand side of (5.2). So the data-driven DRO representation (5.2) provides a direct interpretation of the regularization parameter as the size of the probabilistic uncertainty around the empirical evidence.

An important element to all of the DRO representations obtained in Blanchet *et al.* (2016a) is that the design of the distributional uncertainty neighborhood $\mathcal{U}_\delta(P_n)$ is based on optimal transport theory. More specifically, we define the distributional uncertainty neighborhood as

$$\mathcal{U}_\delta(P_n) = \{P : D_c(P, P_n) \le \delta\}, \tag{5.3}$$

where $D_c(P, P_n)$ is the minimal cost of rearranging (i.e. transporting the mass of) the distribution P_n into the distribution P. The rearrangement mechanism has a transportation cost $c(u, w) \geq 0$ for moving a unit of mass from location u in the support of P_n to location w in the support of P. For instance, in the setting of (5.2) we choose the cost as

$$c\big((x, y), (x', y')\big) = \|x - x'\|_q^2 \, I\left(y = y'\right) + \infty \cdot I\left(y \neq y'\right). \tag{5.4}$$

As discussed in Section 5.3, $D_c(P, P_n)$ can be computed as the solution of a linear programming (LP), which is also known as Kantorovich's problem (see Villani (2008)).

Other discrepancy notions between probability models have been explored by a vast number of literatures, especially the Kullback-Leibler divergence and other divergence based notions Hu and Hong (2013). Using divergence (or likelihood ratio) based discrepancies to characterize the uncertainty region $\mathcal{U}_\delta(P_n)$ forces the models $P \in \mathcal{U}_\delta(P_n)$ to have the same support as P_n, which may restrict generalization properties of a DRO-based estimator, and such restriction may further induce overfitting problem (see the discussions in Esfahani and Kuhn (2015) and Blanchet *et al.* (2016a)).

The generalization performance of DRO is primarily affected by the choice of distributionally uncertainty set, specifically by its size and shape. Under the setting of (5.3), choosing the size of the uncertainty neighborhood in DRO is equivalent to choosing a tuning parameter δ for regularization. One way to optimally select δ is based on Robust Wasserstein Profile function, whose assymptotic behavior is comprehensively discussed in Blanchet *et al.* (2016a). In practice, we can also choose δ by cross-validation. The work of Blanchet *et al.* (2016a) compares the asymptotically optimal choice against cross-validation, concluding that the performance is comparable in the experiments performed. In this paper, we use cross validation to choose δ.

Note that the the *shape* of $\mathcal{U}_\delta(P_n)$ is determined by the cost function $c(\cdot)$ in the definition of the optimal transport discrepancy $D_c(P, P_n)$, but so far it has been taken as a given, but not chosen in a data-driven way. This is the starting point of this paper to improve the DRO method and our main goal in this paper is to discuss a data-driven framework to inform the shape of the uncertainty neighborhood. As to selecting the types of cost $c(\cdot)$ to be used in practice, we rely on metric-learning procedures. Ultimately, the choice of $c(\cdot)$ is influenced by the nature of the data and the application problem at hand. For example, in the setting of image recognition, it might be natural to use a cost function related to similarity notions.

In brief, DD-DRO employs metric learning procedures to estimate $c(\cdot)$ by exploiting the side information in the data. Then with this learned cost function $c(\cdot)$ we can further define $D_c(P, P_n)$ and the distributional uncertainty neighborhood $\mathcal{U}_\delta(P_n)$ in (5.3). Finally, we solve the DRO problem (5.1) and use cross-validation to choose a proper δ. Based on DD-DRO, we further propose a DD-R-DRO model, which contains two layers of robustification. The first layer is, instead of minimizing risk with respect to empirical measure defined by the training data, DRO minimizes the maximum risk with repect to all the measures in the distributional uncertainty neighborhood of the empirical measure defined via the distance $D_c(P, P_n)$. The second layer of robustness arises from learning the cost function $c(\cdot)$ of $D_c(P, P_n)$ in a robust way to minimize the effect of noisiness among side information.

We now provide our main contributions in this paper:

• We establish a data-driven framework that combines k-NN methods with logistic regressions for classification.

We propose a Data-driven Distributionally Robust Optimization (DD-DRO) model, which uses k-NN method to generate the side information of the data (the side information contains the information about the intrinsic measure among the data) and then form the shape of the distributional uncertainty neighborhood by learning a metric from this side information. This combination is desirable as logistic regression is a linear classifier which has high bias, while k-NN has high variability, so they complement each other well.

• We reveal the connection between DD-DRO and adaptive regularized ridge regression estimator.

Theorem 5.1 reveals the close relationship between DD-DRO and adaptive regularized ridge regression. Our DD-DRO approach provides a novel and interpretable way that selects hyper-parameters in adaptive regularized ridge regression from a metric learning perspective.

• We propose an appoximation algorithm based on stochastic gradient descent to solve DD-DRO.

Thanks to the duality representation given in Blanchet and Murthy (2016), we are able to reformulate the DD-DRO problem, which is solved by a smoothing approximation and stochastic gradient descent algorithm. The error bound of the smoothing approximation is provided in Lemma 5.1.

• **We employed robust metric learning to take care of the noisiness of side information.**

The side information is usually noisy; it is either given, e.g. by the implicit feedback from the customers, which contains a lot of incorrect information, or it is generated, as we did in this paper, by k-NN method, which suffers from high variability. So we borrow the idea from robust optimization and build a doubly robust data-driven distributionally robust optimization (DD-R-DRO) model on top of the DD-DRO model to deal with noisiness of side information by introduing an additional layer of robustification during metric learning for the construction of distributional uncertainty neighborhood. We use primal-dual deepest descent to achieve robust metric learning.

Figure 5.1 summarizes the various combinations of information and robustness which have been studied in the literature so far and in our DD-DRO and DD-R-DRO models.

The figure consists of four diagrams with various arrows. A wiggly arrow indicates potentially noisy testing error estimates. The straight arrows represent the use of a robustification procedure. A wide arrow represents the use of high degree of information.

Diagram (A) represents the standard empirical risk minimization (ERM); which fully uses the training data but often leads to high variability in testing error and poor out-of-sample performance. Diagram (B) represents DRO where only the center, P_n, and the size of the uncertainty, δ, are data driven; this choice controls out-of-sample performance but does not use the side information among data to shape the type of perturbation (i.e. the cost function), thus it potentially results in pessimistic testing error bounds. Diagram (C) illustrates DD-DRO with data-driven shape of distributional uncertainty neighborhood for perturbation through metric learning techniques; this construction uses the side information in the data and reduces the testing error bounds at the expense of increase in the variability of the testing error estimates. Diagram (D) illustrates DD-R-DRO, the shape of the perturbation allowed for the adversary player is estimated by using robust optimization procedure to account for the noisiness of the side information; this double robustification, as we will show in the numerical experiments, is able to control the variability that presents in the third diagram.

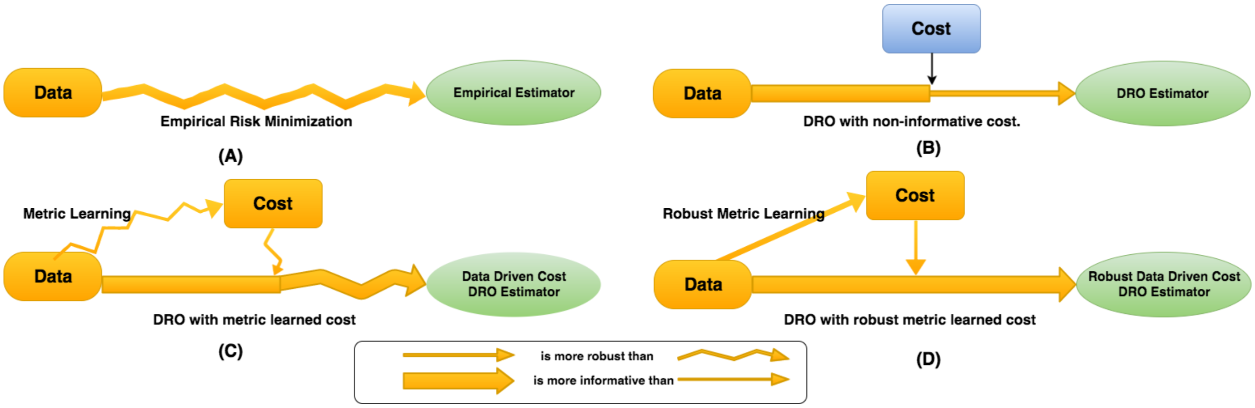

Data
Empirical Risk Minimization
Empirical Estimator
(A)
Cost
Data
DRO with non-informative cost.
DRO Estimator
(B)
Metric Learning
Cost
Data
DRO with metric learned cost
Data Driven Cost DRO Estimator
(C)
Robust Metric Learning
Cost
Data
DRO with robust metric learned cost
Robust Data Driven Cost DRO Estimator
(D)
is more robust than
is more informative than

Figure 5.1: Four diagrams illustrating information on robustness.

The rest of the paper is organized as following. In Section 5.2 we will go into details about the necessity of metric learning in DD-DRO and the intuition behind the improvement of the generalization property. In Section 5.3 we give a quick review of the metric learning and its usage in DD-DRO. In section 5.4, we show the connection between DD-DRO and adaptive regularized ridge regression. In Section 5.5, we introduce an algorithm based on stochastic gradient descent to solve DD-DRO. In Section 5.6, we formulate the robust metric learning problem that is solved by primal-dual steepest descent algorithm. In Section 5.7, we compare the performance of DD-DRO and DD-R-DRO with a number of alternative machine learning methods on various data sets and show that our approach exhibits consistently superior performance.

5.2 Data-Driven DRO: Intuition and Interpretations

One of the main benefits using DRO formulation such as (5.2) is its interpretability. For example, we can readily see from the left hand side of (5.2) that the regularization parameter corresponds precisely to the size of the *data-driven* distributional uncertainty δ, so we can employ statistical thinking to pick it optimally. Additionally, the DRO is appealing as it reveals how to enhance generalization properties. We can interpret (5.1) as a game in which we (the outer player) choose a decision β, while the adversary (the inner player) selects a model which is a perturbation P, of the data (encoded by P_n). The amount of the perturbation is dictated by the size of δ, while the type of perturbation and its measurement is dictated by $D_c(P, P_n)$. Figure 2(a) further explains the necessity of informing $D_c(\cdot)$ in a data-driven way.

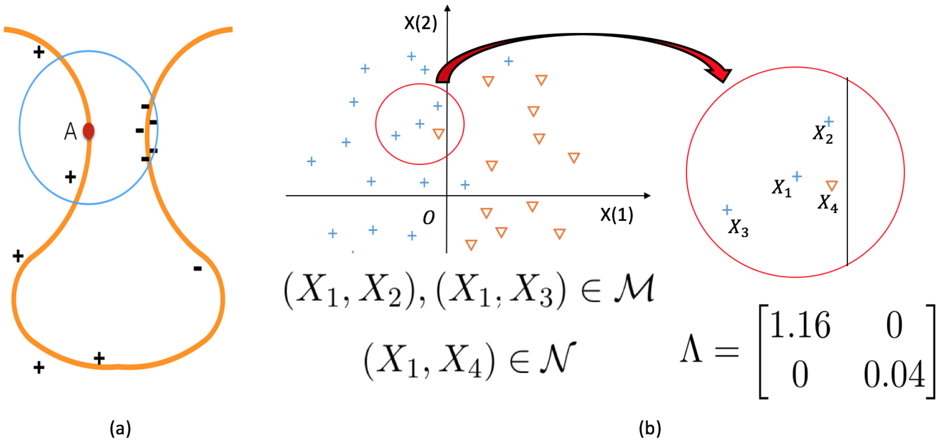

A
+
-
X(2)
X(1)
0
X_1
X_2
X_3
X_4
$(X_1, X_2), (X_1, X_3) \in \mathcal{M}$
$(X_1, X_4) \in \mathcal{N}$
$\Lambda = \begin{bmatrix} 1.16 & 0 \\ 0 & 0.04 \end{bmatrix}$
(a)
(b)

Figure 5.2: Stylized examples illustrating the need for data-driven cost function.

Suppose we have a classification task. The data roughly lies on a lower-dimensional and non-linear manifold. Some data classified as negative are “close” to data classified as positive when one sees the whole space $\mathbb{R}^2$ as the natural ambient domain of the data. However, if we use a distance similar to the geodesic distance intrinsic in the manifold then the negative instances are actually far apart from the positive instances. By learning this intrinsic metric we are able to calibrate a cost function $c(u, w)$ which attaches relatively high transportation costs to (u, w) if transporting mass between these locations has substantial impacts on the response variable and increases the expected risk. This forces the adversary player to carefully choose the data which is to be transported, with a given budget δ. He has to make a compromise; on one hand, he would like to maximize the empirical risk by purturbing the data between locations that has substantial impacts on the response variable, but on the other hand he has to pay higher cost for those purturbations. As a result, this compromise of the DRO procedure leads to the focus on reagions of relevance and hence improves the generalization performance.

The idea can be further explored in the context of a logistic regression shown in Figure 2(b): Suppose that $d = 2$, and that Y depends only on $X(1)$, the first coordinate of X. The metric learning in (5.7) will capture the more informative $X(1)$ in the data and induce a cost function which bears relatively high transportation cost in $X(1)$ direction while relatively low transportation cost along $X(2)$ direction. From the standpoint of the adversarial player, he has to reach a compromise between maximizing the expected loss (which is his objective) by transporting more along the impactful $X(1)$ direction and paying a higher cost for perturbing along $X(1)$ direction with his limited budget δ.

5.3 Background on Optimal Transport and Metric Learning Procedures

This section provides a quick review of basic notions in optimal transport for defining $D_c(P, P_n)$ and in metric learning for calibrating the cost function $c(\cdot)$.

5.3.1 Defining Optimal Transport Distances and Discrepancies

Assume that the cost function $c : \mathbb{R}^{d+1} \times \mathbb{R}^{d+1} \to [0, \infty]$ is lower semicontinuous. We also assume that $c(u, v) = 0$ if and only if $u = v$. Given two distributions P and Q, with supports $\mathcal{S}_P$ and $\mathcal{S}_Q$, respectively, we define the optimal transport discrepancy, D_c, via

$$D_c\left(P, Q\right) = \inf\left\{\mathbb{E}_\pi\left[c(U, V)\right] : \pi \in \mathcal{P}\left(\mathcal{S}_{_P} \times \mathcal{S}_{_Q}\right),\ \pi_{_U} = P,\ \pi_{_V} = Q\right\}, \tag{5.5}$$

where $\mathcal{P}(\mathcal{S}_{_P} \times \mathcal{S}_{_Q})$ is the set of probability distributions π supported on $\mathcal{S}_P \times \mathcal{S}_{_Q}$, and $\pi_{_U}$ and $\pi_{_V}$ denote the two marginals of π. The non-negativeness of $c(\cdot)$ ensures that $D_c(P, Q) \geq 0$ and the condition that $c(u, v) = 0$ if and only if $u = v$ guarantees that $D_c(P, Q) = 0$ if and only $P = Q$. If $c(\cdot)$ is also symmetric (i.e. $c(u, v) = c(v, u)$), and there exists $\varrho \geq 1$ such that $c^{1/\varrho}(u, w) \leq c^{1/\varrho}(u, v) + c^{1/\varrho}(v, w)$ (i.e. $c^{1/\varrho}(\cdot)$ satisfies the triangle inequality) then it can be verified (see Villani (2008)) that $D_c^{1/\varrho}(P,Q)$ is a metric. For example, if $c(u,v) = \|u-v\|_q^{\varrho}$ for $q \geq 1$ (where $\|u-v\|_q$ denotes the l_q norm in $\mathbb{R}^{d+1}$) then $D_c(\cdot)$ is known as the Wasserstein distance of order ϱ. An important observation is that (5.5) is a linear program in the variable π.

5.3.2 On Metric Learning Procedures

In order to keep the discussion focused, we pick only a few metric learning methods for the calibration of cost function in DRO formulation, but we emphasize that our approach can combine with almost any other methods in the metric learning literature. The paper Bellet *et al.* (2013) gives a wide survey of various metric learning procedures. The procedures we employed can already improve significantly upon natural benchmarks, and these metric families can be related to adaptive regularization. This connection will be useful to further enhance the intuition of our procedure.

5.3.2.1 The Mahalanobis Distance

The Mahalanobis metric is defined as

$$d_\Lambda(x,x') = \left((x-x')^T \Lambda (x-x')\right)^{1/2},$$

where Λ is symmetric and positive semi-definite and we write $\Lambda \in PSD$. Note that $d_\Lambda(x,x')$ is the metric induced by the norm $\|x\|_\Lambda = \sqrt{x^T \Lambda x}$.

Suppose our data is of the form $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$ and $Y_i \in \{-1,+1\}$. The prediction variables are assumed to have already been standardized. Motivated by applications such as social networks, where there is a natural graph connecting instances in the data, and the information of connection is summarized in sets $\mathcal{M}$ and $\mathcal{N}$, where $\mathcal{M}$ is the set of the pairs that should be close (so that we can connect them) to each other, while $\mathcal{N}$ characterizes the relations that the pairs should be far away (not connected). They are often called side information of the data. We define them as

$$\mathcal{M} := \{(X_i, X_j) \mid X_i \text{ and } X_j \text{ must connect}\},$$
$$\mathcal{N} := \{(X_i, X_j) \mid X_i \text{ and } X_j \text{ should not connect}\}.$$

While it is typically assumed that $\mathcal{M}$ and $\mathcal{N}$ are given, one may also resort to k-Nearest-Neighbor (k-NN) method to generate these sets if they are not available. This is the approach we follow in our numerical experiments. It is worth noting that the choice of any criterion for the definition of $\mathcal{M}$ and $\mathcal{N}$ should be oriented by the learning task so as to achieve both interpretability and performance.

In our experiments we assign the pair (X_i, X_j) to $\mathcal{M}$ if they are sufficiently close in the k-NN criterion, and with the same label $Y_i = Y_j$. Else if $Y_i \neq Y_j$, we assign them to $\mathcal{N}$. The work of Xing *et al.* (2002), one of pioneer paper on this subject, suggests to solve the following optimization problem

$$\min_{\Lambda \in PSD} \sum_{(X_i,X_j)\in\mathcal{M}} d_\Lambda^2\left(X_i, X_j\right) \tag{5.6}$$

$$s.t. \quad \sum_{(X_i,X_j)\in\mathcal{N}} d_\Lambda^2\left(X_i, X_j\right) \geq \bar{\lambda}. \tag{5.7}$$

to achieve the goal of minimizing the total distance between pairs that should be connect, while keeping the pairs that should not connect well separated. The constant $\bar{\lambda} > 0$ is not essential, since Λ can be normalized by $\bar{\lambda}$ and we can choose $\bar{\lambda} = 1$ without loss of generality.

The optimization problem (5.7) is a typical semidefinite programming and has been widely studied, see, for example, Xing *et al.* (2002) for a projection-based algorithm; and Schultz and Joachims (2004) for a factorization-based procedure; or the survey paper Bellet *et al.* (2013) for comparison between various algorithms.

We have chosen formulation (5.7) to estimate Λ as it is most intuitive, while more advanced metric learning techniques developed recently can also be incorporated into the estimation of cost function $c(\cdot)$ for our DRO formulation. (see Li *et al.* (2016)).

5.3.2.2 Using Mahalanobis Distance in Data-Driven DRO

Just as before assume that the underlying data takes the form $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$, where $X_i \in \mathbb{R}^d$ and $Y_i \in \mathbb{R}$ and the loss function, depending on a decision variable $\beta \in \mathbb{R}^m$, is given by $l(x, y, \beta)$. Note that no linear structure is imposed on the underlying model or on the loss function. Analogous to the cost function $c(\cdot)$ in (5.4), here we define a cost function

$c_\Lambda(\cdot)$ associated with a positive semidefinite matrix Λ as

$$c_{\Lambda}\left((x,y),(x',y')\right) = d_{\Lambda}^{2}\left(x,x'\right) I\left(y=y'\right) + \infty I\left(y \neq y'\right). \tag{5.8}$$

The infinite contribution in the definition of c_{Λ} (i.e. the $\infty \cdot I\left(y \neq y'\right)$ part) indicates that the adversarial player in the DRO formulation is not allowed to perturb the response variable. Since the sets $\mathcal{M}$ and $\mathcal{N}$ depend on $W_i = (X_i, Y_i)$, the cost function $c_{\Lambda}(\cdot)$ will be informed by the Y_i's. With this data-driven cost function at hand we can then estimate β via

$$\min_{\beta} \sup_{P: D_{c_{\Lambda}}(P,P_n) \leq \delta} \mathbb{E}[l(X,Y,\beta)]. \tag{5.9}$$

It is worth noting that Λ comes only into the definition of the cost function.

5.3.2.3 Mahalanobis Metrics on a Non-Linear Feature Space

In this subsection, we consider the case when the cost function is defined on non-linear transformed data. Assume that the data takes the form $\mathcal{D}_n = \{(X_i,Y_i)\}_{i=1}^{n}$, where $X_i \in \mathbb{R}^d$ and $Y_i \in \mathbb{R}$ and the loss function, depending on decision variable $\beta \in R^m$, is given by $l\left(x,y,\beta\right)$. We define

$$c_{\Lambda}^{\Phi}\left((x,y),(x',y')\right) = d_{\Lambda}^{2}\left(\Phi\left(x\right),\Phi\left(x'\right)\right) I\left(y=y'\right) + \infty I\left(y \neq y'\right), \tag{5.10}$$

for $\Lambda \in PSD$. To preserve the properties of a cost function (i.e. non-negativity, lower semicontinuity and $c_{\Lambda}^{\Phi}\left(u,w\right) = 0$ implies $u = w$), we assume that $\Phi\left(\cdot\right)$ is continuous and that $\Phi\left(w\right) = \Phi\left(u\right)$ implies that $w = u$. Then we can apply a metric learning procedure, such as the one described in (5.7), to calibrate Λ.

5.4 Data Driven Cost Selection and Adaptive Regularization

In this section we establish a direct connection between our fully data-driven DRO procedure and adaptive regularization. Our main result also reveals a direct connection between the metric learning and adaptive regularized estimators.

Throughout this section we consider a data set of the form $\mathcal{D}_n = \{(X_i,Y_i)\}_{i=1}^{n}$ with $X_i \in \mathbb{R}^d$

and $Y_i \in \mathbb{R}$ as before. With the cost function $c_{\Lambda}(\cdot)$ defined in (5.8), we have the following results. Its proof is given in the Appendix.

Theorem 5.1 (DRO Representation for Generalized Adaptive Regularization)**.** *Assume that $\Lambda \in \mathbb{R}^{d\times d}$ in (5.8) is positive definite and the loss function is mean squared error, we have the following representation*

$$\min_{\beta} \max_{P:D_{c_\Lambda}(P,P_n)\leq\delta} \mathbb{E}_P^{1/2}\left[\left(Y - X^T\beta\right)^2\right] = \min_{\beta} \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - X_i^T\beta\right)^{1/2}} + \sqrt{\delta}\,\|\beta\|_{\Lambda^{-1}}\,. \quad (5.11)$$

Moreover in the context of adaptive regularized logistic regression with $Y \in \{-1,+1\}$, we have

$$\min_{\beta} \max_{P:D_{c_\Lambda}(P,P_n)\leq\delta} \mathbb{E}\left[\log\left(1+e^{-Y(X^T\beta)}\right)\right] = \min_{\beta}\frac{1}{n}\sum_{i=1}^{n}\log\left(1+e^{-Y_i(X_i^T\beta)}\right) + \delta\,\|\beta\|_{\Lambda^{-1}}(5.12)$$

In particular, when Λ is a diagonal positive definite matrix, we recover a more familiar version of (5.11) for adaptive regularization:

$$\min_{\beta} \max_{P:D_{c_\Lambda}(P,P_n)\leq\delta} \mathbb{E}_P^{1/2}\left[\left(Y - X^T\beta\right)^2\right] = \min_{\beta} \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - X_i^T\beta\right)^2} + \sqrt{\delta}\sqrt{\sum_{i=1}^{d}\beta_i^2/\Lambda_{ii}}(5.13)$$

The adaptive regularization method was initially derived as a generalization of ridge regression in Hoerl and Kennard (1970b) and Hoerl and Kennard (1970a). Recent work shows that adaptive regularization can improve the prediction power of its non-adaptive counterpart, especially in high-dimensional settings (see in Zou (2006) and Ishwaran and Rao (2014)).

In view of (5.13), our discussion in Section 5.3.2.1 uncovers tools which can be used to estimate the coefficients $\{1/\Lambda_{ii} : 1 < i \leq d\}$. To complement the intuition given in Figure 1(b), note that in the adaptive regularization literature one often choose $\Lambda_{ii} \approx 0$ to force $\beta_i \approx 0$ (i.e., there is a high penalty to variables with low explanatory power). This corresponds to the low transport costs along those low explanatory directions in our DRO formulation.

5.5 Solving Data Driven DRO Based on Optimal Transport Discrepancies

To fully take advantage of the synergies between DRO and metric learning it is crucial to have an algorithm efficiently estimating β from (5.1). In the presence of a special representation such as (5.2) or (5.13), we can apply standard stochastic optimization methods (see Lei and Jordan (2016)).

Our objective in this section is to give algorithms which are applicable to more general loss and cost functions, when a simplified representation is not accessible.

Throughout this section the data has the form $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n \subset \mathbb{R}^{d+1}$. The loss function is written as $\{l\left(x, y, \beta\right) : (x, y) \in \mathbb{R}^{d+1}, \beta \in \mathbb{R}^m\}$. We assume that for each (x, y), the function $l\left(x, y, \cdot\right)$ is convex and continuously differentiable. Further, we shall consider cost functions of the form

$$\bar{c}\left(\left(x, y\right), \left(x', y'\right)\right) = c\left(x, x'\right) I\left(y = y'\right) + \infty I\left(y \neq y'\right),$$

as this will simplify the form of the dual representation in the inner optimization of our DRO formulation. To ensure boundedness of the DRO formulation, we impose the following assumption.

Assumption 1. There exists $\Gamma(\beta, y) \in (0, \infty)$ such that $l(u, y, \beta) \leq \Gamma(\beta, y) \cdot (1 + c(u, x))$, for all $(x, y) \in \mathcal{D}_n$, Under Assumption 1, we can guarantee that

$$\max_{P: D_c(P, P_n) \leq \delta} \mathbb{E}_P\left[l\left(X, Y, \beta\right)\right] \leq \left(1 + \delta\right) \max_{i=1,\ldots,n} \Gamma\left(\beta, Y_i\right) < \infty.$$

Using the strong duality theorem for semi-infinity linear programming problem in Appendix B of Blanchet *et al.* (2016a),

$$\max_{P: D_c(P, P_n) \leq \delta} \mathbb{E}_P\left[l\left(X, Y, \beta\right)\right] = \min_{\lambda \geq 0} \frac{1}{n} \sum_{i=1}^{n} \phi\left(X_i, Y_i, \beta, \lambda\right), \tag{5.14}$$

where $\psi(u, X, Y, \beta, \lambda) := l(u, Y, \beta) - \lambda(c(u, X) - \delta)$, $\phi\left(X, Y, \beta, \lambda\right) := \max_{u \in \mathbb{R}^d} \psi(u, X, Y, \beta, \lambda)$. Therefore,

$$\min_{\beta} \max_{P: D_{c_\Lambda}(P, P_n) \leq \delta} \mathbb{E}_P\left[l\left(X, Y, \beta\right)\right] = \min_{\lambda \geq 0, \beta} \left\{\mathbb{E}_{P_n}\left[\phi\left(X, Y, \beta, \lambda\right)\right]\right\}. \tag{5.15}$$

The right hand side of (5.15) is minimized over β and λ, to which we can apply stochastic approximation algorithms if the gradient of $\phi\left(\cdot\right)$ with respect to β and λ exist. However, $\phi\left(\cdot\right)$ itself is given by solving a maximization problem, so its gradient is not readily accessible. Therefore we consider a smoothing approximation technique to remove the maximization problem in $\phi\left(\cdot\right)$.

The smoothing approximation for $\phi\left(\cdot\right)$ is defined as,

$$\phi_{\epsilon,f}\left(X,Y,\beta,\lambda\right)=\epsilon\log\left(\int_{\mathbb{R}^d}\exp\left(\left[\psi\left(u,X,Y,\beta,\lambda\right)\right]/\epsilon\right)f\left(u\right)du\right),$$

where $f\left(\cdot\right)$ is a probability density in $\mathbb{R}^d$; for example, f can be taken as the density of a normal distribution and $\epsilon>0$ is a smoothing parameter.

Theorem 5.1 below quantifies the error due to smoothing approximation.

Lemma 5.1. *Under mild technical assumptions (see Assumption 1-4 in Appendix 5.9.2), there exists* $\epsilon_0>0$ *such that for every* $\epsilon<\epsilon_0$*, we have*

$$\phi(X,Y,\beta,\lambda)\geq\phi_{\epsilon,f}(X,Y,\beta,\lambda)\geq\phi(X,Y,\beta,\lambda)-d\epsilon\log(1/\epsilon)$$

The proof of Lemma 5.1 is given in Appendix 5.9.2.

With the help of smooth approximation we transform the original optimization problem to a standard stochastic optimization problem and we can solve it by mini-batch based stochastic approximation (SA) algorithm. Notice that the gradient of $\phi_{\epsilon,f}\left(\cdot\right)$, as a function and β and λ, satisfies

$$\nabla_\beta\phi_{\epsilon,f}\left(X,Y,\beta,\lambda\right)=\frac{\mathbb{E}_{U\sim f}\left[\exp\left(\psi\left(U,X,Y,\beta,\lambda\right)/\epsilon\right)\nabla_\beta l\left(U,X,Y\right)\right]}{\mathbb{E}_{U\sim f}\left[\exp\left(\psi\left(U,X,Y,\beta,\lambda\right)/\epsilon\right)\right]}, \tag{5.16}$$

$$\nabla_\lambda\phi_{\epsilon,f}\left(X,Y,\beta,\lambda\right)=\frac{\mathbb{E}_{U\sim f}\left[\exp\left(\psi\left(U,X,Y,\beta,\lambda\right)/\epsilon\right)\left(\delta-c_{\Lambda_n}\left(U,X\right)\right)\right]}{\mathbb{E}_{U\sim f}\left[\exp\left(\psi\left(U,X,Y,\beta,\lambda\right)/\epsilon\right)\right]}. \tag{5.17}$$

which is still in the form of expectation, but we can approximate the gradient by a simple Monte Carlo sampling, i.e., we sample U_i's from $f(\cdot)$ and evaluate the numerators and denominators of the gradient using Monte Carlo separately. The details of this SA algorithm are given in Algorithm 2.

Algorithm 2 Stochastic Gradient Descent with Continuous State

```
**Initialize** $\lambda = 0$, and $\beta$ to be empirical risk minimizer, $\epsilon = 0.5$, tracking error $Error = 100$.
**while** $Error > 10^{-3}$ **do**
    **Sample** a mini-batch $\{X_j, Y_j\}_{j=1}^M$ uniformly from $n$ observations , with $M \leq n$.
    For each $j = 1, \ldots, M$, sample i.i.d. $\{U_k^{(j)}\}_{k=1}^L$ from $\mathcal{N}\left(0, \sigma^2 I_{d\times d}\right)$.
    We denote $f_L^j$ as empirical distribution for $U_k^{(j)}$'s, and estimate the batched as
        $$\nabla_\beta \phi_{\epsilon,f} = \frac{1}{M}\sum_{j=1}^M \nabla_\beta \phi_{\epsilon,f_L^j}\left(X_j, Y_j, \beta, \lambda\right), \nabla_\lambda \phi_{\epsilon,f} = \frac{1}{M}\sum_{j=1}^M \nabla_\lambda \phi_{\epsilon,f_L^j}\left(X_j, Y_j, \beta, \lambda\right).$$
    Update $\beta$ and $\lambda$ using $\beta = \beta - \alpha_\beta \nabla_\beta \phi_{\epsilon,f}$ and $\lambda = \lambda - \alpha_\lambda \nabla_\lambda \phi_{\epsilon,f}$.
    Update tracking error $Error$ as the norm of difference between latest parameter and average of last 50 iterations.
**Output** $\beta$.
```

Remark 5.1. *The above optimization problem can be written as a mixed problem in the sense of J.Blanchet* et al. *(2017)*

$$\min_{\lambda\geq 0,\beta} \frac{1}{n}\sum_{i=1}^n \epsilon \log(\mathbb{E}_U[\exp(\psi(U, X_i, Y_i, \beta, \lambda)/\epsilon)]) =: \min_{\lambda\geq 0,\beta} \frac{1}{n}\sum_{i=1}^n \Phi_i(\mathbb{E}_U[\Psi_U(\beta, \lambda)]) =: \min_{\lambda\geq 0,\beta} F(\beta, \lambda), \tag{5.18}$$

where

$$\Psi_U(\beta, \lambda) := \big(\exp(\psi(U, X_1, Y_1, \beta, \lambda)/\epsilon), \cdots, \exp(\psi(U, X_M, Y_M, \beta, \lambda)/\epsilon)\big)^T,$$

and $\Phi_i(\mathbb{E}_U[\Psi_U(\beta, \lambda)]) := \epsilon \log(\mathbb{E}_U[\exp(\psi(U, X_i, Y_i, \beta, \lambda)/\epsilon)]) = \phi_{\epsilon,f}\left(X_i, Y_i, \beta, \lambda\right).$

Let $\theta := (\beta, \lambda)^T$*, then* $\exp\left(\left[\psi\left(u, X, Y, \theta\right)\right]/\epsilon\right) f\left(u\right)$ *is log-convex in* θ *for all* $u \in \mathbb{R}$*, and by Boyd and Vandenberghe (2004) p.106 the function* $\phi_{\epsilon,f}$ *is convex in* θ*, so* F *is also convex.*

Remark 5.2. *Note that*

$$\nabla_\beta \phi_{\epsilon,f}\left(X, Y, \beta, \lambda\right) = \frac{\mathbb{E}_{U\sim f}\left[\exp\left(\psi\left(U, X, Y, \beta, \lambda\right)/\epsilon\right)\nabla_\beta l\left(U, X, Y\right)\right]}{\mathbb{E}_{U\sim f}\left[\exp\left(\psi\left(U, X, Y, \beta, \lambda\right)/\epsilon\right)\right]}$$

relates to the quotient of two expectations. In general,

$$\frac{\sum_{k=1}^L \exp\left(\psi\left(U_k, X, Y, \beta, \lambda\right)/\epsilon\right)\nabla_\beta l\left(U_k, X, Y\right)}{\sum_{k=1}^L \exp\left(\psi\left(U_k, X, Y, \beta, \lambda\right)/\epsilon\right)}$$

is not an unbiased estimator for $\nabla_\beta \phi_{\epsilon,f}\left(X,Y,\beta,\lambda\right)$*. The unbiased estimator* $\nabla_\beta \phi_{\epsilon,f_L^j}\left(X_j,Y_j,\beta,\lambda\right)$ *can be construncted via multi-level randomization, see Blanchet and Glynn (2015).* $\nabla_\lambda \phi_{\epsilon,f}$ *is analogue.*

Remark 5.3. *Let* $\theta = (\beta,\lambda)^T$ *and define*

$$h(X_j,Y_j,(U_k^{(j)})_{k=1}^L,\theta) := \nabla_\theta \phi_{\epsilon,f_L^j}\left(X_j,Y_j,\theta\right)$$

and

$$H(Z,\theta) := \frac{1}{M}\sum_{j=1}^{M} h(X_j,Y_j,(U_k^{(j)})_{k=1}^L,\theta),$$

where $Z = ((X_j,Y_j,(U_k^{(j)})_{k=1}^L)_{j=1}^M)$*. We have* $\mathbb{E}[H(Z,\theta)|\theta] = \mathbb{E}_{P_M \bigotimes P_f}[\nabla_\theta\phi_{\epsilon,f}(X,Y,U,\theta)|\theta]$*, where* P_M *is the empirical measure formed by the* M *data points, and* P_f *denotes the distribution of* U*. Thus, Algorithm 2 can be seen as a special type of stochastic gradient descend algorithm.*

Theorem 5.2. *Let* $\theta^* = (\beta^*,\lambda^*)^T$ *be an optimal value for* $\min_\theta F(\theta)$ *in (5.18). Assume that i) the problem is optimized over a compact and convex set* Θ *which contains* θ^* *and* $(\theta_t)_{t\geq 0}$ *generated by the algorithm, ii) Inside the compact set* Θ *both the loss function* $l(\cdot)$ *and the cost function* $c(\cdot)$ *are twice continously differentiable, Then*
a) If F *is* μ*-strongly convex, when taking step-size* $\alpha_t = \frac{2}{\mu\sqrt{t}}$ *we then have*

$$\mathbb{E}[\|\theta_T - \theta^*\|_2^2] \leq \frac{4C}{\mu^2(T+1)},$$

where C *is some constant.*

b) If F *is just convex but we can find some constant* D *such that* $\mathbb{E}[\|\theta_t - \theta^*\|_2^2] \leq D$ *for all* $t \geq 0$*, when taking step-size* $\alpha_t = \frac{2}{\sqrt{t+1}}$ *we then have*

$$\mathbb{E}[F(\theta_T)] - F(\theta^*) \leq \frac{2\sqrt{2}(C+D)}{\sqrt{T}},$$

where C *is some constant.*

Proof. Since both the loss function $l(\cdot)$ and the cost function $c(\cdot)$ are twice continously differentiable and the set Θ is compact, so the functions Φ_i and Ψ_U are also twice continously differentiable and their gradients are L-Lipschitz continuous for some constant L. Therefore, it satisfies all the assumptions in J.Blanchet *et al.* (2017) and the results follow. □

Remark 5.4. *As shown in Remark 5.1 F is convex, but if F is not strongly convex we can generally add a penalty term $\|\theta\|_2^2$ to make it strongly convex.*

5.6 Robust Metric Learning

The objective of this section is to explore how to learn data-driven cost function c_Λ in a robust way and combine it with our DD-DRO model. We call this new model Doubly Robust Data-Driven Distributionally Robust Optimization (DD-R-DRO). Robust optimization (RO) is a family of optimization techniques that deals with uncertainty or misspecification in the objective function and constraints. It was first proposed in Ben-Tal *et al.* (2009) and has attracted increasing attentions in the recent decades El Ghaoui and Lebret (1997) and Bertsimas *et al.* (2011). It has been applied in machine learning to regularize statistical learning procedures, for example, in Xu *et al.* (2009a) and Xu *et al.* (2009b) robust optimization was employed for SR-Lasso and support vector machines. Note that the classical robust optimization is different from our distributionally robust optimization in the sense that classical robust optimization concerns only deterministic uncertain scenarios and we will apply this classical robust optimization only to the cost function learning procedure in our DD-DRO. The reason that we want to learn the cost function in a robust way is due to the fact that there often exists noisiness or incorrectness in side information (e.g. the training constraint sets $\mathcal{M}$ and $\mathcal{N}$ or the relative constraints $\mathcal{R}$ defined later, are often gained from customers' implicit feedback, and are quite noisy.) This extra layer of robustness will reduce the variability in testing error as it is shown in the later numerical experiments.

5.6.1 Robust Optimization for Relative Metric Learning

The robust metric learning we shall use is based on the work of Huang *et al.* (2012). Consider the relative constraint set $\mathcal{R}$ containing data triplets with relative relation defined as

$$\mathcal{R} = \left\{ (i,j,k) \,|\, d_\Lambda(X_i, X_j) \text{ should be smaller than } d_\Lambda(X_i, X_k) \right\}.$$

and the Relative Metric Learning formulation

$$\min_{\Lambda \succeq 0} \sum_{(i,j,k)\in\mathcal{R}} \left(d_\Lambda^2\left(X_i, X_j\right) - d_\Lambda^2\left(X_i, X_k\right) + 1 \right)_+ . \tag{5.19}$$

Suppose we know that about $1-\alpha \in (0,1]$ of the constraints are noisy (the value of α is usually given by experience or it can also be inferred by cross validation), but we cannot determine exactly which part of them are noisy. Instead of optimizing over all subsets of constraints, we try to minimize the worst case loss function over all possible $\alpha\left|\mathcal{R}\right|$ constraints (where $\left|\cdot\right|$ denote the cardinality of a set) and obtain the following min-max formulation

$$\min_{\Lambda\succeq 0}\max_{q\in\mathcal{T}(\alpha)}\sum_{(i,j,k)\in\mathcal{R}} q_{i,j,k}\left(d_\Lambda^2\left(X_i,X_j\right)-d_\Lambda^2\left(X_i,X_k\right)+1\right)_+, \tag{5.20}$$

where $\mathcal{T}(\alpha)$ is a robust uncertainty set of the form

$$\mathcal{T}\left(\alpha\right)=\left\{q=\{q_{i,j,k}|(i,j,k)\in\mathcal{R}\}\left|0\le q_{i,j,k}\le 1,\sum_{(i,j,k)\in\mathcal{R}} q_{i,j,k}\le \alpha\times\left|\mathcal{R}\right|\right.\right\},$$

which is a convex and compact set.

Firstly we observe that the above minimax problem is equivalent to

$$\min_{\Lambda\succeq 0}\max_{q\in\mathcal{T}(\alpha)}\sum_{(i,j,k)\in\mathcal{R}} q_{i,j,k}\left(d_\Lambda^2\left(X_i,X_j\right)-d_\Lambda^2\left(X_i,X_k\right)+1\right), \tag{5.21}$$

because $q_{i,j,k}=0$ whenever $d_\Lambda^2\left(X_i,X_j\right)-d_\Lambda^2\left(X_i,X_k\right)+1<0$. So our algorithm will be quite different from the original algorithm proposed by Huang *et al.* (2012) in the sense that our alorithm doesn't resort to any smooth technique, instead we will use the primal-dual steepest descent algorithm.

Secondly, the objective function in (5.21) is convex in Λ and concave (linear) in $\tilde{q}$. Define

$$L(\Lambda,q):=\sum_{(i,j,k)\in\mathcal{R}} q_{i,j,k}\left(d_\Lambda^2\left(X_i,X_j\right)-d_\Lambda^2\left(X_i,X_k\right)+1\right)+\frac{\kappa}{2}\|\Lambda\|_F^2, \tag{5.22}$$

where $\kappa>0$ is the tuning parameter and $\|\cdot\|_F$ denote the Frobenius norm. In addition we define

$$f(\Lambda):=\max_{q\in\mathcal{T}(\alpha)} L(\Lambda,q),\quad g(q):=\min_{\Lambda\succeq 0} L(\Lambda,q) \tag{5.23}$$

and since L is strongly convex in Λ and linear in q we also have

$$q(\Lambda):=\arg\max_{q\in\mathcal{T}(\alpha)} L(\Lambda,q),\quad \Lambda(q):=\arg\min_{\Lambda\succeq 0} L(\Lambda,q) \tag{5.24}$$

Our iterative algorithm uses the fact that $f(\Lambda)$ and $q(\Lambda)$ can actually be fast obtained by a simple sorting: for a fixed $\Lambda\succeq 0$, the inner maximization is linear in q, and the optimal

q satisfy $q_{i,j,k} = 1$ whenever $(d_\Lambda(X_i, X_j) - d_\Lambda(X_i, X_k) + 1) \geq 0$ and ranks in the top $\alpha\,|\mathcal{R}|$ largest values and set $q_{i,j,k} = 0$ otherwise. If there are more than one optimal q's, break the tie by choosing $q(\Lambda) := \arg\max_{q\in\mathcal{T}(\alpha)} L(\Lambda, q) - \kappa\|q - q^{(n)}\|_2^2$ at the $(n+1)$-th iteration, where κ is a tuning parameter that is small enough such that it ensures that the $q(\Lambda)$ is chosen among the all the optimal ones. On the other hand, since $L(\Lambda, q)$ is smooth in Λ, we can also obtain $g(q)$ and $\Lambda(q)$ fast by gradient descent.

We summarize the primal-dual steepest descent algorithm as in Algorithm 3.

5.6.2 Robust Optimization for Absolute Metric Learning

The RO formulation of (5.7) that we present here appears to be novel in the literature. First we write (5.7) in its Lagrangian form,

$$\min_{\Lambda\succeq 0} \quad \max_{\lambda\geq 0} \quad \sum_{(i,j)\in\mathcal{M}} d_\Lambda^2(X_i, X_j) + \lambda\big(1 - \sum_{(i,j)\in\mathcal{N}} d_\Lambda^2(X_i, X_j)\big). \tag{5.25}$$

Similar to $\mathcal{R}$, the side information sets $\mathcal{M}$ and $\mathcal{N}$ often suffer from noisiness or inaccuracy as well. Let us assume that about $1-\alpha$ proportion of the constraints in $\mathcal{M}$ and $\mathcal{N}$ are, respectively, inaccurate. We then construct robust uncertainty sets $\mathcal{W}(\alpha)$ and $\mathcal{V}(\alpha)$ from $\mathcal{M}$ and $\mathcal{N}$ as

$$\mathcal{W}(\alpha) = \big\{\tilde{\eta} = \{\eta_{ij} : (i,j) \in \mathcal{M}\}\,|0 \leq \eta_{ij} \leq 1, \sum_{(i,j)\in\mathcal{M}} \eta_{ij} \leq \alpha \times |\mathcal{M}|\big\},$$

$$\mathcal{V}(\alpha) = \big\{\tilde{\xi} = \{\xi_{ij} : (i,j) \in \mathcal{N}\}\,|0 \leq \xi_{ij} \leq 1, \sum_{(i,j)\in\mathcal{N}} \xi_{ij} \geq \alpha \times |\mathcal{N}|\big\}.$$

Then the RO formulation of (5.25) can be written as

$$\min_{\Lambda\succeq 0} \quad \max_{\lambda\geq 0} \quad \max_{\eta\in\mathcal{W}(\alpha),\xi\in\mathcal{V}(\alpha)} \sum_{(i,j)\in\mathcal{M}} \eta_{i,j} d_\Lambda^2(X_i, X_j) + \lambda\big(1 - \sum_{(i,j)\in\mathcal{N}} \xi_{i,j} d_\Lambda^2(X_i, X_j)\big) \tag{5.26}$$

The switch of $\max_\lambda$ with $\max_{(\tilde{\eta},\tilde{\xi})}$ is valid in general. Note also that the Cartesian product $\mathcal{M}(\alpha) \times \mathcal{N}(\alpha)$ is a compact set, and the objective function is convex in Λ and concave (linear) in pair $(\tilde{\eta}, \tilde{\xi})$, so we can further apply Sion's min-max Theorem again (see in Terkelsen (1973)) to switch the order of $\min_\Lambda$-$\max_{(\tilde{\eta},\tilde{\xi})}$. This leads to an iterative algorithm.

Define $q := (\eta, \xi)$ as the dual variable and

$$L(\Lambda, q) := \sum_{(i,j)\in\mathcal{M}} \eta_{i,j} d_\Lambda^2(X_i, X_j) + \lambda\big(1 - \sum_{(i,j)\in\mathcal{N}} \xi_{i,j} d_\Lambda^2(X_i, X_j)\big) + \frac{\kappa}{2}\|\Lambda\|_F^2, \tag{5.27}$$

Algorithm 3 Sequential Coordinate-wise Metric Learning Using Relative Relations

```
1: Initialize Set the iteration counter $n = 0$, the positive definite matrix $\Lambda = I_d$, and
   the tolerance $\epsilon = 10^{-3}$. Then randomly sample $\alpha$ proportion of elements from $\mathcal{R}$ to
   construct $q$.
2: (Optimal test) Terminate if
       $|\min\{f(\Lambda^{(n)}), f(\Lambda(q^{(n)}))\} - \max\{g(q^{(n)}), g(q(\Lambda^{(n)}))\}| \leq \epsilon.$
   and Output
       $\bar{\Lambda} = \arg\min\{f(\Lambda)|\Lambda = \Lambda^{(n)} \text{ or } \Lambda = \Lambda(q^{(n)})\}$
       $\bar{q} = \arg\max\{g(q)|q = q^{(n)} \text{ or } q = q(\Lambda^{(n)})\},$
3: (Line search) Generate the intermediate $\hat{\Lambda}^{(n+1)}, \hat{q}^{(n+1)}$ with perfect line search
       $\hat{\Lambda}^{(n+1)} = (1 - \gamma_n)\Lambda^{(n)} + \gamma_n \Lambda(q(\Lambda^{(n)}))$
       $\hat{q}^{(n+1)} = (1 - \beta_n)q^{(n)} + \beta_n q(\Lambda(q^{(n)})),$
   where
       $\gamma_n = \arg\min_{\gamma \in [0,1]} f\big((1-\gamma)\Lambda^{(n)} + \gamma\Lambda(q(\Lambda^{(n)}))\big)$
       $\beta_n = \arg\max_{\beta \in [0,1]} g\big((1-\beta)q^{(n)} + \beta q(\Lambda(q^{(n)}))\big),$
4: (Update the iterates)
       $\Lambda^{(n+1)} = \arg\min\{f(\Lambda)|\Lambda = \hat{\Lambda}^{(n+1)} \text{ or } \Lambda = \Lambda(q^{(n)})\}$
       $q^{(n+1)} = \arg\max\{g(q)|q = \hat{q}^{(n+1)} \text{ or } q = q(\Lambda^{(n)})\},$
   and then return to Step 2 with counter $n \leftarrow n + 1$.
```

and

$$f(\Lambda) := \max_{q \in \mathcal{T}(\alpha)} L(\Lambda, q), \quad g(q) := \min_{\Lambda \succeq 0} L(\Lambda, q) \tag{5.28}$$

and

$$q(\Lambda) := \arg\max_{q \in \mathcal{T}(\alpha)} L(\Lambda, q), \quad \Lambda(q) := \arg\min_{\Lambda \succeq 0} L(\Lambda, q) \tag{5.29}$$

Similarly, the $f(\Lambda)$ and $q(\Lambda)$ are easy to obtain. At the n-th step, given fixed $\Lambda^{(n-1)} \succeq 0$ and $\lambda > 0$ (it is easy to observe that optimal solution λ is positive, i.e. the constraint is active so we may safely assume $\lambda > 0$), the inner maximization problem looks like,

$$\max_{\eta \in \mathcal{W}(\alpha)} \sum_{(i,j) \in \mathcal{M}} \eta_{i,j} d^2_{\Lambda^{(n-1)}}(X_i, X_j) + \lambda\big(1 - \min_{\xi \in \mathcal{V}(\alpha)} \sum_{(i,j) \in \mathcal{N}} \xi_{i,j} d^2_{\Lambda^{(n-1)}}(X_i, X_j)\big).$$

Analogous to the relative constraints case, the optimal η and ξ satisfy: $\eta_{i,j}$ is 1, if $d^2_{\Lambda^{(n-1)}}(X_i, X_j)$ ranks top α within $\mathcal{M}$ and equals 0 otherwise; while $\xi_{i,j} = 1$ if $d^2_{\Lambda^{(n-1)}}(X_i, X_j)$ ranks bottom α within $\mathcal{N}$ and equals 0 otherwise. So we also define $\mathcal{M}_\alpha(\Lambda^{(n-1)})$ as a subset of $\mathcal{M}$, which contains the constraints with largest α percent of $d_{\Lambda^{(n-1)}}(\cdot)$; and define $\mathcal{N}_\alpha(\Lambda^{(n-1)})$ as a subset of $\mathcal{N}$, which contains the constraints with smallest α percent of $d_{\Lambda^{(n-1)}}(\cdot)$. Then the optimal solution given fixed $\Lambda^{(n-1)}$ can be reformulated as $\eta_{i,j} = 1$ if $(i,j) \in \mathcal{M}_\alpha(\Lambda^{(n-1)})$ and $\xi_{i,j} = 1$ if $(i,j) \in \mathcal{N}_\alpha(\Lambda^{(n-1)})$.

On the other hand, given fixed η and ξ, we can simplify the minimization problem $g(q)$ as

$$\min_{\Lambda \succeq 0} \sum_{(i,j) \in \mathcal{M}_\alpha(\Lambda^{(n-1)})} d^2_\Lambda(X_i, X_j) \quad \text{s.t.} \quad \sum_{(i,j) \in \mathcal{N}_\alpha(\Lambda^{(n-1)})} d^2_\Lambda(X_i, X_j) \geq 1.$$

This formulation of the minimization problem $g(q)$ takes the same form as (5.7) and it thus can be solved by similar SDP algorithms presented in Xing *et al.* (2002). On the whole we solve the minimax problem of (5.26) by using the same primal-dual algorithm as presented in Algorithm 3.

Other robust methods have also been considered in the metric learning literature, see Zha *et al.* (2009) and Lim *et al.* (2013) although the connections to RO are not fully exposed.

Theorem 5.3. *There exists saddle points* $(\bar{\Lambda}, \bar{q})$ *for the minimax problems* (5.22) *and* (5.27) *respectively. The Algorithm 3 converges linearly to the common optimal value* $f(\bar{\Lambda}) = g(\bar{q})$

in the sense that

$$f(\Lambda^{(n+1)}) - f(\bar{\Lambda}) \leq \theta\big(f(\Lambda^{(n)}) - f(\bar{\Lambda})\big)$$
$$g(q^{(n+1)}) - g(\bar{q}) \leq \theta\big(g(q^{(n)}) - g(\bar{q})\big),$$

where $\theta \in (0,1)$ *is some constant and the functions* f *and* g *are define by* (5.23) *and* (5.28) *respectively.*

Proof. In both (5.22) and (5.27) the function L is strongly convex in Λ and q takes value in a bounded set, so it satisfies the condition of existence of saddle points, see Zhu (1994) and R.T.Rockafeller (1970). Note that though L is not strongly concave in q but it is linear in q. At iteration $n+1$ if there are more than one optimal values of q we can use the proximal point algorithm to create a strongly convex-concave Lagrangian, i.e. we maximize $-\|q - q^{(n)}\|_2^2$ to break the tie. The linear convergence follows then from Theorem 3.3 in Zhu (1994). □

5.7 Numerical Experiments

5.7.1 Numerical Experiments for DD-DRO

We validate our data-driven cost function based DRO on 5 real data sets from the UCI machine learning database Lichman (2013). We focus on a DRO formulation for a linear classification model with the log-exponential loss. We use the linear metric learning framework (5.7) to learn a positive semidefinite matrix Λ and then plug it into the cost function c_Λ defined in (5.8). We denote this model DRO-L. In addition, we also fit a cost function c_Λ^Φ to the quadratric transformed data, as explained in (5.10); the model is denoted by DRO-NL. We compare our DRO-L and DRO-NL with logistic regression (LR), and regularized logistic regression (LRL1). For each iteration and each data set, the data is split randomly into training and test sets. We fit the models on the training set and evaluate the performance on test set. The regularization parameter is chosen via $5-$fold cross-validation for LRL1, DRO-L and DRO-NL. For each data set, we perform 200 independent experiments and report the mean and standard deviation for the training error, testing error and testing accuracy. The details of the numerical results and basic information of the data are summarized in Table 5.1.

Table 5.1: Numerical Results for DD-DRO on Real Data Sets.

		breast cancer	qsar	magic	minibone	spambase
	Train	0 ± 0	$.026 \pm .008$	$.213 \pm .153$	0 ± 0	0 ± 0
LR	Test	8.75 ± 4.75	35.5 ± 12.8	17.8 ± 6.77	18.2 ± 10.0	14.5 ± 9.04
	Accur	$.762 \pm .061$	$.701 \pm .040$	$.668 \pm .042$	$.678 \pm .059$	$.789 \pm .035$
	Train	$.185 \pm .123$	$.614 \pm .038$	$.548 \pm .087$	$.401 \pm .167$	$.470 \pm .040$
LRL1	Test	$.428 \pm .338$	$.755 \pm .019$	$.610 \pm .050$	$.910 \pm .131$	$.588 \pm .140$
	Accur	$.929 \pm .023$	$.646 \pm .036$	$.665 \pm .045$	$.717 \pm .041$	$.811 \pm .034$
	Train	$.022 \pm .019$	$.402 \pm .039$	$.469 \pm .064$	$.294 \pm .046$	$.166 \pm .031$
DRO-L	Test	$.126 \pm .034$	$.557 \pm .023$	$.571 \pm .043$	$.613 \pm .053$	$.333 \pm .018$
	Accur	$.954 \pm .015$	$.733 \pm .026$	$.727 \pm .039$	$.714 \pm .032$	$.887 \pm .011$
	Train	$.032 \pm .015$	$.339 \pm .044$	$.381 \pm .084$	$.287 \pm .049$	$.195 \pm .034$
DRO-NL	Test	$.119 \pm .044$	$.554 \pm .032$	$.576 \pm .049$	$.607 \pm .060$	$.332 \pm .015$
	Accur	$.955 \pm .016$	$.736 \pm .027$	$.730 \pm .043$	$.716 \pm .054$	$.889 \pm .009$
Num Predictors		30	30	10	20	56
Train Size		40	80	30	30	150
Test Size		329	475	9990	125034	2951

5.7.2 Numerical Experiments for DD-R-DRO

In this subsection we proceed to verify the further improved performance of our DD-R-DRO method on the same five data sets from UCI machine learning data base.

The side information, i.e. the relative constraint set $\mathcal{R}$ and the absolute constraint sets $\mathcal{M}$ and $\mathcal{N}$ are generated by k-NN method. We then add noisiness to these constraint sets by randomly replacing the correct constraints with wrong constraints with probability $1 - \alpha$.

We consider logistic regression (LR), regularized logistic regression (LRL1), DD-DRO with cost function learned from absolute constraints (DD-DRO (absolute)) and DD-R-DRO(absolute) with cost function learned from absolute constraints at level of $\alpha = 50\%$ and $\alpha = 90\%$; DD-

DRO with cost function learned from relative constraints (DD-DRO (relative)) and DD-R-DRO(relative) with cost function learned from relative constraints at level of $\alpha = 50\%$ and $\alpha = 90\%$. For each data and each experiment, we randomly split the data into training and testing and fit models on training set and evaluate on testing set.

For each data sets, we perform 200 independent experiments and report the mean and standard deviation of training error, testing error, and testing accuracy. The detailed results and data set information are summarized in Table 5.2.

After we obtained the learned cost function, we then apply the smoothing approximation algorithm introduced in Section 5.5 to solve the DRO problem directly, where the size of uncertainty δ is chosen via 5-fold cross-validation.

We observe that DD-R-DRO presents robust improvement comparing to its non-robust counterpart DD-DRO when the cost function is learned from noisy side information at level $\alpha = 90\%$. More important, DD-R-DRO tends to enjoy the variance reduction property due to RO. In addition, as the robust level increases, i.e. $\alpha = 50\%$, where we believe that the side information is highly noisy, we observe that the doubly robust based approach seems to shrink towards to LRL1, and benefits less from the data-driven cost structure.

5.8 Conclusion and Discussion

We have proposed a novel DD-DRO, a fully data-driven DRO procedure, which combines a semiparametric approach (the metric learning) with a parametric procedure (the expected loss minimization) and enhances the generalization performance of the underlying parametric model. A smoothing technique based algorithm is given for solving the DD-DRO problem.

Based on DD-DRO we further take noisiness of the side information into account during the metric learning for the cost function, and introduce robust metric learning method to DD-DRO, which leads to our DD-R-DRO model. The overall method is then doubly robust; one is distributionally robustness around the training data, and the other is robust metric learning of the cost function from the noisy side information. This second layer of robustness not only keeps the improved generalization properties of DD-DRO, but also reduces the variability of the testing errors due to the noise in side information.

We emphasize that our approach is applicable to other DRO formulations and is not restricted to classification tasks. Interesting future research avenues which might be worth considering include the development of a semisupervised framework as in Blanchet and Kang (2017), in which unlabeled data is used to inform the support of the elements in $\mathcal{U}_\delta(P_n)$. Another interesing approach that might be worth exploring is to combine metric learning for domain adaptation with our DRO model.

5.9 Proof of Main Results

5.9.1 Proof of Theorem 5.1

We first state and prove Lemma 5.2 which will be useful for the proof of Theorem 5.1.

Lemma 5.2. *If Λ is a is positive definite matrix and we define $\|x\|_\Lambda = \left(x^T \Lambda x\right)^{1/2}$, then $\|\cdot\|_{\Lambda^{-1}}$ is the dual norm of $\|\cdot\|_\Lambda$. Furthermore, we have*

$$u^T w \leq \|u\|_\Lambda \|w\|_{\Lambda^{-1}},$$

where the equality holds if and only if, there exists non-negative constant τ, s.t $\tau \Lambda u = \Lambda^{-1} w$ or $\tau \Lambda^{-1} w = \Lambda u$.

Proof of Lemma 5.2. This result is a direct generalization of l_2 norm in Euclidean space. Note that

$$u^T w = (\Lambda u)^T \left(\Lambda^{-1} w\right) \leq \|\Lambda u\|_2 \left\|\Lambda^{-1} w\right\|_2 = \|u\|_\Lambda \|w\|_{\Lambda^{-1}}. \tag{5.30}$$

The inequality above is Cauchy-Schwartz inequality for $\mathbb{R}^d$ appling to Λu and $\Lambda^{-1} w$, and the equality holds if and only if there exists nonnegative τ, s.t. $\tau \Lambda u = \Lambda^{-1} w$ or $\tau \Lambda^{-1} w = \Lambda u$. By the definition of the dual norm, we have

$$\|w\|_\Lambda^* = \sup_{u: \|u\|_\Lambda \leq 1} u^T w = \sup_{u: \|u\|_\Lambda \leq 1} \|u\|_\Lambda \|w\|_{\Lambda^{-1}} = \|w\|_{\Lambda^{-1}}.$$

While the first equality follows from the definition of dual norm, the second equality is due to Cauchy-Schwartz inequality (5.30) and the equality condition therein, and the last equality are immediate after maximizing. □

Proof of Theorem 5.1. The technique is a generalization of the method used in proving Theorem 1 in Blanchet *et al.* (2016a). We can apply the strong duality result (see Proposition 6 in Appendix of Blanchet *et al.* (2016a)) to the worst-case expected loss function, and obtain a semi-infinite linear programming problem

$$\sup_{P:D_{c_\Lambda}(P,P_n)\leq\delta} \mathbb{E}_P\left[\left(Y-X^T\beta\right)^2\right] = \min_{\gamma\geq 0}\left\{\gamma\delta - \frac{1}{n}\sum_{i=1}^{n}\sup_{u}\left\{\left(y_i-u^T\beta\right)^2-\gamma\left\|x_i-u\right\|_\Lambda^2\right\}\right\}.$$

For the inner suprema , let us denote $\Delta = u - X_i$ and $e_i = Y_i - X_i^T\beta$ for notation simplicity. The inner optimization problem associated with (X_i, Y_i) becomes,

$$\begin{aligned}
&\sup_{u}\left\{\left(y_i-u^T\beta\right)^2-\gamma\left\|x_i-u\right\|_\Lambda^2\right\}\\
&= e_i^2 + \sup_{\Delta}\left\{\left(\Delta^T\beta\right)^2 - 2e_i\Delta^T\beta - \gamma\left\|\Delta\right\|_\Lambda^2\right\},\\
&= e_i^2 + \sup_{\Delta}\left\{\left(\sum_j\left|\Delta_j\right|\left|\beta_j\right|\right)^2 + 2\left|e_i\right|\sum_j\left|\Delta_j\right|\left|\beta_j\right| - \gamma\left\|\Delta\right\|_\Lambda^2\right\},\\
&= e_i^2 + \sup_{\left\|\Delta\right\|_\Lambda}\left\{\left\|\Delta\right\|_\Lambda^2\left\|\beta\right\|_{\Lambda^{-1}}^2 + 2\left|e_i\right|\left\|\Delta\right\|_\Lambda\left\|\beta\right\|_{\Lambda^{-1}} - \gamma\left\|\Delta\right\|_\Lambda^2\right\},\\
&= \begin{cases} e_i^2\frac{\gamma}{\gamma-\left\|\beta\right\|_{\Lambda^{-1}}^2} & \text{if } \gamma > \left\|\beta\right\|_{\Lambda^{-1}}^2,\\ +\infty & \text{if } \gamma \leq \left\|\beta\right\|_{\Lambda^{-1}}^2. \end{cases}
\end{aligned}$$

While the first equality is due to the change of variable, the second equality follows from the fact that the last term only depends on the magnitude rather than sign of Δ, so the maximization problem will always pick a Δ that satisfies the equality. The third equality follows from the same reason; we can first apply the Cauchy-Schwartz inequality in Lemma 5.2 and the maximization problem will pick a Δ satisfying the equality constraint. The last equality following simply from the first order condition of optimality.

For the outer minimization problem over γ, as the inner suprema equal infinity if $\gamma \leq$

$\|\beta\|_{\Lambda^{-1}}^2$, the worst-case expected loss becomes,

$$\begin{aligned}
&\sup_{P:D_{c_\Lambda}(P,P_n)\le\delta} \mathbb{E}_P\left[\left(Y - X^T\beta\right)^2\right] \qquad (5.31)\\
&= \min_{\gamma>\|\beta\|_{\Lambda^{-1}}^2}\left\{\gamma\delta - \frac{1}{n}\sum_{i=1}^n\left(Y_i - X_i^T\beta\right)\frac{\gamma}{\gamma-\|\beta\|_{\Lambda^{-1}}^2}\right\},\\
&= \left(\sqrt{\frac{1}{n}\sum_{i=1}^n\left(Y_i - X_i^T\beta\right)} + \sqrt{\delta}\,\|\beta\|_{\Lambda^{-1}}\right)^2.
\end{aligned}$$

We see that the objective function on the right hand side of (5.31) is convex and differentiable and the value function will be infinity as $\gamma \to \infty$ and $\gamma \to \|\beta\|_\Lambda^2$. Solving γ through the first order condition of optimality, it is straightforward to obtain the last equality in (5.31). By taking square root on both sides, we proved the claim for the case of mean-squared loss function.

For the log-exponential loss function the proof is analogous. By applying strong duality results of semi-infinity linear programming problem in Blanchet *et al.* (2016a), we can write the worst case expected loss function as

$$\begin{aligned}
&\sup_{P:D_{c_\Lambda}(P,P_n)\le\delta} \mathbb{E}_P\left[\log\left(1+\exp\left(-Y\beta^TX\right)\right)\right]\\
&= \min_{\gamma\ge0}\left\{\gamma\delta - \frac{1}{n}\sum_{i=1}^n\sup_u\left\{\log\left(1+\exp\left(-Y_i\beta^Tu\right)\right) - \gamma\|X_i-u\|_\Lambda\right\}\right\}.
\end{aligned}$$

For each i, we can use Lemma 1 in Shafieezadeh-Abadeh *et al.* (2015) and dual-norm result in Lemma 5.2 to deal with the inner maximization problem and obtain

$$\sup_u\left\{\log\left(1+\exp\left(-Y_i\beta^Tu\right)\right) - \gamma\|X_i-u\|_\Lambda\right\} = \begin{cases} \log\left(1+\exp\left(-Y_i\beta^TX_i\right)\right) & \text{if} \quad \|\beta\|_{\Lambda^{-1}}\le\gamma,\\ \infty & \text{if} \quad \|\beta\|_{\Lambda^{-1}}>\gamma.\end{cases}$$

Moreover, since the outer is minimization problem, following the same discussion for the proof for linear regression case, we can plug-in the result above and get the first equality below,

$$\begin{aligned}
&\min_{\gamma\geq 0}\left\{\gamma\delta-\frac{1}{n}\sum_{i=1}^{n}\sup_{u}\left\{\log\left(1+\exp\left(-Y_i\beta^T u\right)\right)-\gamma\left\|X_i-u\right\|_{\Lambda}\right\}\right\}\\
&\quad=\min_{\gamma\geq\|\beta\|_{\Lambda^{-1}}}\left\{\delta\gamma+\frac{1}{n}\sum_{i=1}^{n}\log\left(1+\exp\left(-Y_i\beta^T X_i\right)\right)\right\}\\
&\quad=\frac{1}{n}\sum_{i=1}^{n}\log\left(1+\exp\left(-Y_i\beta^T X_i\right)\right)+\delta\left\|\beta\right\|_{\Lambda^{-1}}.
\end{aligned}$$

We know that the target function is continuous and monotone increasing in γ, its optimal is $\gamma=\|\beta\|_{\Lambda^{-1}}$, which leads to the second equality above. This proves the theorem for the case of log-exponential loss function. □

5.9.2 Proof of Lemma 5.1

Let us first list all the assumptions required to prove Theorem 5.1. We begin by recalling Assumption 1 from Section 5.5.

Assumption 1. There exists $\Gamma(\beta,y)\in(0,\infty)$ such that $l(u,y,\beta)\leq\Gamma(\beta,y)\cdot(1+c(u,x))$, for all $(x,y)\in\mathcal{D}_n$,

In addition, we introduce the following Assumptions 2-4.

Assumption 2. $\psi\left(\cdot,X,Y,\beta,\lambda\right)$ is twice continuously differentiable and the Hessian of $\psi\left(\cdot,X,Y,\beta,\lambda\right)$ evaluated at u^*, $D_u^2\psi\left(u^*,X,Y,\beta,\lambda\right)$, is positive definite. In particular, we can find $\theta>0$ and $\eta>0$, such that

$$\psi(u,X,Y,\beta,\lambda)\geq\psi\left(u^*,X,Y,\beta,\lambda\right)-\frac{\theta}{2}\|u-u^*\|_2^2,\quad\forall u\text{ with }\left\|u-u^*\right\|_2\leq\eta.$$

Assumption 3. For a constant $\lambda_0>0$ such that $\phi(X,Y,\beta,\lambda_0)<\infty$, let $K=K\left(X,Y,\beta,\lambda_0\right)$ be any upper bound for $\phi(X,Y,\beta,\lambda_0)$.

Assumption 4. In addition to the lower semicontinuity of $c\left(\cdot\right)\geq 0$, we assume that $c\left(\cdot,X\right)$ is coercive in the sense that $c\left(u,X\right)\to\infty$ as $\left\|u\right\|_2\to\infty$.

For any set S, the r-neighborhood of S is defined as the set of all points in $\mathbb{R}^d$ which are at a distance less than r from S, i.e. $S_r=\cup_{u\in S}\{\bar{u}:\left\|\bar{u}-u\right\|_2\leq r\}$.

Proof of Lemma 5.1. The first part of the inequalities is easy to derive. For the second part, we proceed as follows: Under Assumptions 3 and 4, we can define a compact set

$$\mathcal{C}=\mathcal{C}(X,Y,\beta,\lambda)=\{u:c(u,X)\leq l(X,Y,\beta)-K+\lambda_0/(\lambda-\lambda_0)\}.$$

It is easy to check that $\arg\max\{\psi(u, X, Y, \lambda)\} \subset \mathcal{C}$. Owing to optimality of u^* and Assumption 2 that $K \geq \phi(X, Y, \beta, \lambda_0)$, we see that

$$\begin{aligned} l(X,Y) &\leq l(u^*, Y)) - \lambda c(u, X) \\ &= l(u^*, Y) - \lambda_0 c(u^*, X) - (\lambda - \lambda_0) c(u^*, X) \\ &\leq K - \lambda_0 - (\lambda - \lambda_0) c(u^*, X). \end{aligned}$$

By checking the definition of $\mathcal{C} = \mathcal{C}(X, Y, \beta, \lambda)$, one conludes that $u^* \in \mathcal{C}$, which further implies $\{u : \|u - u^*\|_2 \leq \eta\} \subset \mathcal{C}_\eta$. Then we combine the strongly convexity assumption in Assumption 2 and the definition of $\phi_{\epsilon,f}(u, X, Y, \beta, \lambda)$, which yields

$$\begin{aligned} \phi_{\epsilon,f}(X, Y, \beta, \lambda) &\geq \epsilon \log\left(\int_{\|u-u^*\|_2 \leq \eta} \exp\left(\left[\phi(X, Y, \beta, \lambda) - \frac{\theta}{2}\|u - u^*\|_2^2\right]/\epsilon\right) f(u) du\right) \\ &= \epsilon \log\left(\exp\left(\phi(X, Y, \beta, \lambda)/\epsilon\right)\right) \int_{\|u-u^*\|_2 \leq \eta} \exp\left(-\frac{\theta}{2}\|u - u^*\|_2^2/\epsilon\right) f(u) du \\ &= \phi(X, Y, \beta, \lambda) + \epsilon \log \int_{\|u-u^*\|_2 \leq \eta} \exp\left(-\frac{\theta\|u - u^*\|_2^2}{2\epsilon}\right) f(u) du. \end{aligned}$$

As $\{u : \|u - u^*\|_2 \leq \eta\} \subset \mathcal{C}_\eta$, we can use the lower bound of $f(\cdot)$ to deduce that

$$\begin{aligned} \int_{\|u-u^*\|_2 \leq \eta} \exp\left(-\frac{\theta\|u - u^*\|_2^2}{2\epsilon}\right) f(u) du &\geq \inf_{u \in \mathcal{C}_\eta} f(u) \times \int_{\|u-u^*\|_2 \leq \eta} \exp\left(-\frac{\theta\|u - u^*\|_2^2}{2\epsilon}\right) du \\ &= \inf_{u \in \mathcal{C}_\eta} f(u) \times (2\pi\epsilon/\theta)^{d/2} P(Z_d \leq \eta^2\theta/\epsilon), \end{aligned}$$

where Z_d is a chi-squared random variable of d degrees of freedom. To conclude, recall that $\epsilon \in (0, \eta^2\theta\chi_\alpha)$, the lower bound of $\phi_{\epsilon,f}(\cdot)$ can be written as

$$\phi_{\epsilon,f}(X, Y, \beta, \lambda) \geq \phi(X, Y, \beta, \lambda) - \frac{d}{2}\epsilon \log(1/\epsilon) + \frac{d}{2}\epsilon \log\left((2\pi\alpha/\theta) \inf_{u \in \mathcal{C}_\eta} f(u)\right).$$

This completes the proof of Lemma 5.1. □

Table 5.2: Numerical Results for DD-R-DRO on Real Data Sets with Side Information Generated by k-NN Method.

		breast cancer	qsar	magic	minibone	spambase
LR	Train	0 ± 0	$.026 \pm .008$	$.213 \pm .153$	0 ± 0	0 ± 0
	Test	8.75 ± 4.75	35.5 ± 12.8	17.8 ± 6.77	18.2 ± 10.0	14.5 ± 9.04
	Accur	$.762 \pm .061$	$.701 \pm .040$	$.668 \pm .042$	$.678 \pm .059$	$.789 \pm .035$
LRL1	Train	$.185 \pm .123$	$.614 \pm .038$	$.548 \pm .087$	$.401 \pm .167$	$.470 \pm .040$
	Test	$.428 \pm .338$	$.755 \pm .019$	$.610 \pm .050$	$.910 \pm .131$	$.588 \pm .140$
	Accur	$.929 \pm .023$	$.646 \pm .036$	$.665 \pm .045$	$.717 \pm .041$	$.811 \pm .034$
DD-DRO (absolute)	Train	$.022 \pm .019$	$.402 \pm .039$	$.469 \pm .064$	$.294 \pm .046$	$.166 \pm .031$
	Test	$.126 \pm .034$	$.557 \pm .023$	$.571 \pm .043$	$.613 \pm .053$	$.333 \pm .023$
	Accur	$.954 \pm .015$	$.733 \pm .0.026$	$.727 \pm .039$	$.714 \pm .032$	$.887 \pm .011$
DD-R-DRO	Train	$.029 \pm .013$	$.397 \pm .036$	$.420 \pm .063$	$.249 \pm .055$	$.194 \pm .031$
(absolute)	Test	$.126 \pm .023$	$.554 \pm .019$	$.561 \pm .035$	$.609 \pm .044$	$.331 \pm .018$
$\alpha = 90\%$	Accur	$.954 \pm .012$	$.736 \pm .025$	$.729 \pm .032$	$.709 \pm .025$	$.890 \pm .008$
DD-R-DRO	Train	$.040 \pm .055$	$.448 \pm .032$	$.504 \pm .041$	$.351 \pm .048$	$.166 \pm .030$
(absolute)	Test	$.132 \pm .015$	$.579 \pm .017$	$.590 \pm .029$	$.623 \pm .029$	$.337 \pm .013$
$\alpha = 50\%$	Accur	$.952 \pm .012$	$.733 \pm .025$	$.710 \pm .033$	$.715 \pm .021$	$.888 \pm .008$
DD-DRO (relative)	Train	$.086 \pm .038$	$.392 \pm .040$	$.457 \pm .071$	$.322 \pm .061$	$.181 \pm, 036$
	Test	$.153 \pm .060$	$.559 \pm .025$	$582 \pm .033$	$.613 \pm .031$	$.332 \pm .016$
	Accur	$.946 \pm .018$	$.714 \pm .029$	$.710 \pm, 027$	$.704 \pm .021$	$.890 \pm .008$
DD-R-DRO	Train	$.030 \pm .014$	$.375 \pm .038$	$.452 \pm .067$	$.402 \pm .058$	$.234 \pm .032$
(relative)	Test	$.141 \pm .054$	$.556 \pm .022$	$.577 \pm .032$	$.610 \pm .024$	$.332 \pm .011$
$\alpha = 90\%$	Accur	$.949 \pm .019$	$.729 \pm .023$	$.717 \pm .025$	$.710 \pm .020$	$.892 \pm .007$
DD-R-DRO	Train	$.031 \pm .016$	$.445 \pm .032$	$.544 \pm .057$	$.365 \pm .054$	$.288 \pm .029$
(relative)	Test	$.154 \pm .049$	$.570 \pm .019$	$.594 \pm .018$	$.624 \pm .018$	$.357 \pm .008$
$\alpha = 50\%$	Accur	$.948 \pm .019$	$.705 \pm .023$	$.699 \pm .028$	$.698 \pm .018$	$.881 \pm .005$
Num Predictors		30	30	10	20	56
Train Size		40	80	30	30	150
Test Size		329	475	9990	125034	2951

Bibliography

[1] A. Ahmadi-Javid. Entropic value-at-risk: A new coherent risk measure. *Journal of Optimization Theory and Applications*, 155(3):1105–1123, Dec 2012.

[2] Rami Atar, Kenny Chowdhary, and Paul Dupuis. Robust bounds on risk-sensitive functionals via Rényi divergence. *SIAM/ASA J. Uncertain. Quantif.*, 3(1):18–33, 2015.

[3] M. Beiglbock, Henry-Labordere, and F. P.Penkner. Optimal maps for the multidimensional mongekantorovich problem. *Finance Stoch 17: 477*, 2013.

[4] Aurélien Bellet, Amaury Habrard, and Marc Sebban. A survey on metric learning for feature vectors and structured data. *arXiv preprint arXiv:1306.6709*, 2013.

[5] Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust optimization*. Princeton University Press, 2009.

[6] Aharon Ben-Tal, Dick den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

[7] Dimitris Bertsimas, David B Brown, and Constantine Caramanis. Theory and applications of robust optimization. *SIAM review*, 53(3):464–501, 2011.

[8] Jose H. Blanchet and Peter W. Glynn. Unbiased monte carlo for optimization and functions of expectations via multi-level randomization. In *Proceedings of the 2015 Winter Simulation Conference*, WSC '15, pages 3656–3667, Piscataway, NJ, USA, 2015. IEEE Press.

[9] Jose Blanchet and Yang Kang. Sample out-of-sample inference based on wasserstein distance. *arXiv preprint arXiv:1605.01340*, 2016.

[10] Jose Blanchet and Yang Kang. Distributionally robust semi-supervised learning. *arXiv preprint arXiv:1702.08848*, 2017.

[11] Jose Blanchet and Karthyek Murthy. Quantifying distributional model risk via optimal transport. 2016.

[12] Jose Blanchet, Yang Kang, and Karthyek Murthy. Robust wasserstein profile inference and applications to machine learning. *arXiv preprint arXiv:1610.05627*, 2016.

[13] Jose Blanchet, Yang Kang, and Karthyek Murthy. Robust wasserstein profile inference and applications to machine learning. *arXiv preprint arXiv:1610.05627v2*, 2016.

[14] Sergey Bobkov and Michel Ledoux. One-dimensional empirical measures, order statistics and kantorovich transport distances. *preprint*, 2014.

[15] Emmanuel Boissard and Thibaut Le Gouic. On the mean speed of convergence of empirical and occupation measures in wasserstein distance. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 50, pages 539–563. Institut Henri Poincaré, 2014.

[16] Emmanuel Boissard. Simple bounds for the convergence of empirical and occupation measures in 1-wasserstein distance. *Electronic Journal of Probability*, 16:2296–2333, 2011.

[17] François Bolley, Arnaud Guillin, and Cédric Villani. Quantitative concentration inequalities for empirical measures on non-compact spaces. *Probability Theory and Related Fields*, 137(3):541–593, 2007.

[18] Stephen Boyd and Lieven Vandenberghe. *Convex optimization.* Cambridge university press, 2004.

[19] T. Breuer and I. Csiszar. Measuring distribution model risk. *Mathematical Finance*, 2013.

[20] Thomas Breuer and Imre Csiszár. Measuring distribution model risk. *Mathematical Finance*, 2013.

[21] Stuart G. Coles and Jonathan A. Tawn. A bayesian analysis of extreme rainfall data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 45(4):pp. 463–478, 1996.

[22] Stuart Coles. *An introduction to statistical modeling of extreme values*. Springer Series in Statistics. Springer-Verlag London, Ltd., London, 2001.

[23] Laurens de Haan and Ana Ferreira. *Extreme value theory*. Springer Series in Operations Research and Financial Engineering. Springer, New York, 2006. An introduction.

[24] Eustasio Del Barrio, Juan A Cuesta-Albertos, and Carlos Matrán. Tests of goodness of fit based on the l_2-wasserstein distance. *The Annals of Statistics*, 27(4):1230–1239, 1999.

[25] Eustasio Del Barrio, Evarist Giné, and Frederic Utzet. Asymptotics for l2 functionals of the empirical quantile process, with applications to tests of fit based on weighted wasserstein distances. *Bernoulli*, 11(1):131–189, 2005.

[26] Xuan Vinh Doan, Xiaobo Li, and Karthik Natarajan. Robustness to dependency in portfolio optimization using overlapping marginals. *Operations Research*, 63(6):1468–1488, 2015.

[27] V Dobrić and Joseph E Yukich. Asymptotics for transportation cost in high dimensions. *Journal of Theoretical Probability*, 8(1):97–118, 1995.

[28] Paul Dupuis, Matthew R James, and Ian Petersen. Robust properties of risk-sensitive control. *Mathematics of Control, Signals and Systems*, 13(4):318–332, 2000.

[29] E.Candes and T.Tao. Decoding by linear programming. *preprint*, 2014.

[30] E.Candes, J.Romberg, and T.Tao. Stable signal recovery from incomplete and inaccurate measurements. *Preprint*, 2005.

[31] Laurent El Ghaoui and Hervé Lebret. Robust solutions to least-squares problems with uncertain data. *SIAM Journal on Matrix Analysis and Applications*, 18(4):1035–1064, 1997.

[32] Paul Embrechts, Claudia Klüppelberg, and Thomas Mikosch. *Modelling extremal events*, volume 33 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin, 1997. For insurance and finance.

[33] Paul Embrechts, Giovanni Puccetti, and Ludger Rüschendorf. Model uncertainty and var aggregation. *Journal of Banking & Finance*, 37(8):2750–2764, 2013.

[34] Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *arXiv preprint arXiv:1505.05116*, 2015.

[35] William Feller. *An introduction to probability theory and its applications. Vol. II.* John Wiley & Sons, Inc., New York-London-Sydney, 1966.

[36] Nicolas Fournier and Arnaud Guillin. On the rate of convergence in wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3-4):707–738, 2015.

[37] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin, 2001.

[38] Gabor Fukker, Laszlo Gyorfi, and Peter Kevei. Asymptotic behavior of the generalized st.petersburg sum conditioned on its maximum. *Bernoulli 22(2),1026-1054*, 2016.

[39] A. Galichon, P. Henry-Labordre, and N. Touzi. A stochastic control approach to no-arbitrage bounds given marginals, with an application to lookback options. *Ann. Appl. Probab.*, 24(1):312–336, 02 2014.

[40] W. Gangbo and A. Swiech. Optimal maps for the multidimensional mongekantorovich problem. *Comm. Pure Appl. Math.51: 23-45*, 1998.

[41] Rui Gao and Anton J Kleywegt. Distributionally robust stochastic optimization with wasserstein distance. *arXiv preprint arXiv:1604.02199*, 2016.

[42] G.Carlier, C.Jimenez, and F.Santambrogio. Optimal transportation with traffic congestion and wardrop equilibria. *Acta.Math.*, 2008.

[43] P. Glasserman and X. Xu. Robust risk measurement and model risk. *Quantitative Finance, 14(1):29*, 2014.

[44] Paul Glasserman and Xingbo Xu. Robust risk measurement and model risk. *Quantitative Finance*, 14(1):29–58, 2014.

[45] G.Puccetti. The rearragngement algorithm. *https://sites.google.com/site/rearrangementalgorithm*, 2014.

[46] Maya Gupta and Santosh Srivastava. Parametric bayesian estimation of differential entropy and relative entropy. *Entropy*, 12(4):818, 2010.

[47] Lars Peter Hansen and Thomas J. Sargent. Robust control and model uncertainty. *The American Economic Review*, 91(2):pp. 60–66, 2001.

[48] P. Henry-Labordere and N. Touzi. An explicit martingale version of brenier's theorem. *preprint arXiv:1302.4854*, 2013.

[49] H.G.Kellerer. Marlov-komposition und eine anwendung auf martingale. *Math.Ann*, 1972.

[50] H.Lam. Robust sensitivity analysis for stochastic systems. *arXiv preprint arXiv:1303.0326*, 2013.

[51] Arthur E Hoerl and Robert W Kennard. Ridge regression: applications to nonorthogonal problems. *Technometrics*, 12(1):69–82, 1970.

[52] Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[53] Joseph Horowitz and Rajeeva L Karandikar. Mean rates of convergence of empirical measures in the wasserstein metric. *Journal of Computational and Applied Mathematics*, 55(3):261–273, 1994.

[54] Zhaolin Hu and L Jeff Hong. Kullback-leibler divergence constrained distributionally robust optimization. *Available at Optimization Online*, 2013.

[55] Kaizhu Huang, Rong Jin, Zenglin Xu, and Cheng-Lin Liu. Robust metric learning by smooth optimization. *arXiv preprint arXiv:1203.3461*, 2012.

[56] Hemant Ishwaran and J Sunil Rao. Geometry and properties of generalized ridge regression in high dimensions. *Contemp. Math*, 622:81–93, 2014.

[57] J.Blanchet, D.Goldfarb, G.Iyengar, F.Li, and C.Zhou. Unbiased simulation for optimizing stochastic function compositions. *arXiv:1711.07564*, 2017.

[58] J.Goh and M.Sim. Distributionally robust optimization and its tractable approximations. *Operations Research, 58*, 2010.

[59] P. Jorion. *Value at Risk, 3rd Ed.: The New Benchmark for Managing Financial Risk*. McGraw-Hill Education, 2006.

[60] M. R. Leadbetter, Georg Lindgren, and Holger Rootzén. *Extremes and related properties of random sequences and processes*. Springer Series in Statistics. Springer-Verlag, New York-Berlin, 1983.

[61] Lihua Lei and Michael I Jordan. Less than a single pass: Stochastically controlled stochastic gradient method. *arXiv preprint arXiv:1609.03261*, 2016.

[62] Li Li, Chao Sun, Lianlei Lin, Junbao Li, and Shouda Jiang. A mahalanobis metric learning-based polynomial kernel for classification of hyperspectral images. *Neural Computing and Applications*, pages 1–11, 2016.

[63] M. Lichman. UCI machine learning repository, 2013.

[64] Friedrich Liese and Igor Vajda. *Convex statistical distances*. Teubner Texts in Mathematics. BSB B. G. Teubner Verlagsgesellschaft, Leipzig, 1987.

[65] Daryl Lim, Brian McFee, and Gert R Lanckriet. Robust structural metric learning. In *ICML-13*, pages 615–623, 2013.

[66] Stanislav Minsker, Sanvesh Srivastava, Lizhen Lin, and David Dunson. Scalable and robust bayesian inference via the median posterior. In *International Conference on Machine Learning*, pages 1656–1664, 2014.

[67] XuanLong Nguyen, Martin J. Wainwright, and Michael I. Jordan. On surrogate loss functions and f-divergences. *Ann. Statist.*, 37(2):876–904, 04 2009.

[68] XuanLong Nguyen, M.J. Wainwright, and M.I. Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *Information Theory, IEEE Transactions on*, 56(11):5847–5861, Nov 2010.

[69] P.Carr and D.Madan. Towards a theory of volatility trading. *Reprinted in Option Pricing, Interest Rates, and Risk Management*, 2002.

[70] Barnabás Póczos and Jeff Schneider. On the estimation of alpha-divergences. *In AISTATS*, 2011.

[71] Giovanni Puccetti and Ludger Rüschendorf. Sharp bounds for sums of dependent risks. *Journal of Applied Probability*, 50(01):42–53, 2013.

[72] Giovanni Puccetti. Sharp bounds on the expected shortfall for a sum of dependent random variables. *Statistics & Probability Letters*, 83(4):1227–1232, 2013.

[73] Q.Wang, S.Kulkarni, and S.Verdu. Divergence estimation for multidimensional densities via k-nearest-neighbor distances. *IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 55, NO. 5*, 2009.

[74] Svetlozar T Rachev and Ludger Rüschendorf. *Mass Transportation Problems: Volume I: Theory*, volume 1. Springer Science & Business Media, 1998.

[75] Alfréd Rényi. On measures of entropy and information. In *Proc. 4th Berkeley Sympos. Math. Statist. and Prob., Vol. I*, pages 547–561. Univ. California Press, Berkeley, Calif., 1961.

[76] Sidney I. Resnick. *Extreme values, regular variation and point processes*. Springer Series in Operations Research and Financial Engineering. Springer, New York, 2008. Reprint of the 1987 original.

[77] R.K.Ahuja, T.L.Magnanti, and J.B.Orlin. Network flows. *Prentice Hall, Inc.*, 2000.

[78] R.T.Rockafeller. *Convex Analysis.* Princeton university press, 1970.

[79] Ludger Rüschendorf. Solution of a statistical optimization problem by rearrangement methods. *Metrika*, 30(1):55–61, 1983.

[80] S.Bossu, E.Strasser, and R.Guichard. Just what you need to know about variance swaps. *J.P.Morgan London*, 2005.

[81] Matthew Schultz and Thorsten Joachims. Learning a distance metric from relative comparisons. In *Advances in neural information processing systems*, pages 41–48, 2004.

[82] Soroosh Shafieezadeh-Abadeh, Peyman Mohajerin Esfahani, and Daniel Kuhn. Distributionally robust logistic regression. In *Advances in Neural Information Processing Systems*, pages 1576–1584, 2015.

[83] Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński. *Lectures on stochastic programming: modeling and theory.* SIAM, 2009.

[84] A. Sinha, H. Namkoong, and J. Duchi. Certifiable distributional robustness with principled adversarial training. *arXiv:1710.10571*, 2017.

[85] Max Sommerfeld and Axel Munk. Inference for empirical wasserstein distances on finite spaces. *arXiv preprint arXiv:1610.03287*, 2016.

[86] Bharath K Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert RG Lanckriet. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550–1599, 2012.

[87] Frode Terkelsen. Some minimax theorems. *Mathematica Scandinavica*, 31(2):405–413, 1973.

[88] Nicolás Garcia Trillos and Dejan Slepčev. On the rate of convergence of empirical measures in ∞-transportation distance. *arXiv preprint arXiv:1407.1157*, 2014.

[89] Yu-Ling Tsai, Duncan J. Murdoch, and Debbie J. Dupuis. Influence measures and robust estimators of dependence in multivariate extremes. *Extremes*, 14(4):343–363, 2010.

[90] Ramon van Handel. Probability in high dimension. Technical report, DTIC Document, 2014.

[91] Cédric Villani. *Topics in Optimal transportation.* the American Mathematical Society, 2003.

[92] Cédric Villani. *Optimal transport: old and new.* Springer Science & Business Media, 2008.

[93] Bin Wang and Ruodu Wang. The complete mixability and convex minimization problems with monotone marginal densities. *Journal of Multivariate Analysis*, 102(10):1344–1360, 2011.

[94] W. Wieseman, D. Kuhn, and M. Sim. Distributionally robust convex optimization. *Operations Research, 62, pp. 1358*, 2014.

[95] D. Wozabal. A framework for optimization under ambiguity. *Ann. Oper. Res., 193:21*, 2012.

[96] Eric P Xing, Andrew Y Ng, Michael I Jordan, and Stuart Russell. Distance metric learning with application to clustering with side-information. In *NIPS*, volume 15, page 12, 2002.

[97] Huan Xu, Constantine Caramanis, and Shie Mannor. Robust regression and lasso. In *NIPS-2009*, pages 1801–1808, 2009.

[98] Huan Xu, Constantine Caramanis, and Shie Mannor. Robustness and regularization of support vector machines. *JMLR*, 10(Jul):1485–1510, 2009.

[99] Zheng-Jun Zha, Tao Mei, Meng Wang, Zengfu Wang, and Xian-Sheng Hua. Robust distance metric learning with auxiliary knowledge. In *IJCAI*, pages 1327–1332, 2009.

[100] Ciyou Zhu. Solving large-scale minimax problems with the primaldual steepest descent algorithm. *Mathematical Programming*, 1994.

[101] Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.